



**RADA NAUKOWA DYSZYPLINY INFORMATYKA TECHNICZNA I
TELEKOMUNIKACJA**

ROZPRAWA DOKTORSKA

mgr inż. Krzysztof Pałczyński

**Badania algorytmów sztucznej inteligencji i ich
odpowiednich modyfikacji w procesie modelowania
komórek biologicznych oraz wykrywania wybranych
chorób onkologicznych i kardiologicznych**

*Research on artificial intelligence algorithms and their appropriate
modifications in the process of modelling biological cells and detecting
selected oncological and cardiological diseases*

DZIEDZINA: INŻYNIERYJNO-TECHNICZNA
DYSZYPLINA: INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA

PROMOTOR PRACY

dr hab. inż. Tomasz Talaśka, prof. PBŚ

PROMOTOR POMOCNICZY

dr inż. Tomasz Marciniak, prof. PBŚ

Bydgoszcz, rok 2024

Podziękowania

*Chciałbym podziękować promotorowi **dr hab. inż. Tomaszowi Talaście** oraz promotorowi pomocniczemu **dr inż. Tomaszowi Marciniakowi** za nieocenione wsparcie, celne uwagi i owocną współpracę.*

*Dziękuję serdecznie również **prof. dr hab. inż. Piotrowi Cofcie** za pomoc w znalezieniu drogi do celu, cierpliwość oraz krzewienie ognia pasji naukowej.*

*Dziękuję serdecznie także **prof. dr hab. inż. Dariuszowi Skibickiemu** za wsparcie w realizacji doktoratu, możliwość konsultacji oraz ludzką życzliwość.*

Składam również serdeczne podziękowania dla wszystkich osób, które przyczyniły się do powstania mojej pracy doktorskiej.

Osobne, niemniej ważne, podziękowania składam mojej rodzinie, bez której nie byłoby mnie tutaj.

Spis treści:

| | |
|---|-----------|
| 1. Wykaz artykułów naukowych stanowiących monotematyczny cykl publikacji..... | 4 |
| 2. Krótkie wyjaśnienie spójności dla monotematycznego cyklu publikacji | 6 |
| 3. Wstęp..... | 7 |
| 3.1 Hipoteza badawcza, cel i zakres badań | 10 |
| 4. Metody badań i wyniki modelowania komórek biologicznych..... | 12 |
| 4.1 Opracowanie sposobu modelowania komórek biologicznych | 12 |
| 4.2 Algorytmy sztucznej inteligencji..... | 18 |
| 4.3 Wstępne przetwarzanie wektorów parametryzujących symulacje | 22 |
| 4.4 Zamodelowane szlaki metaboliczne..... | 23 |
| 5. Metodyka badań w procesie wspomagania wykrywania wybranych chorób onkologicznych i kardiologicznych | 43 |
| 5.1 Wykrywanie Ostrej Białaczki Limfoblastycznej..... | 43 |
| 5.2 Wykrywanie wybranych chorób serca na podstawie analizy sygnałów EKG | 50 |
| 6. Podsumowanie i wnioski..... | 55 |
| 7. Literatura | 57 |
| 8. STRESZCZENIE | 63 |
| 9. ABSTRACT..... | 64 |
| 10. Oświadczenie Autora rozprawy doktorskiej | 65 |
| 11. Oświadczenia współautorów artykułów naukowych..... | 71 |
| 12. Kopie artykułów naukowych stanowiących monotematyczny cykl publikacji | 88 |

1. Wykaz artykułów naukowych stanowiących monotematyczny cykl publikacji

| Lp. | Autorzy/Tytuł/ Czasopismo | Liczba pkt. | Współ . IF |
|-----|--|-------------|------------|
| P1 | Sylwester Kloska, Krzysztof Palczyński , Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata Wysocka, Paul Davis, Tadeusz Wysocki, <i>Queueing theory model of Krebs Cycle</i> , 2021, Bioinformatics, 37, 18, 2912-2919, 10.1093/bioinformatics/btab177 | 200 | 6,9 |
| P2 | Sylwester Kloska, Krzysztof Palczyński , Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata Wysocka, Paul Davis, Tadeusz Wysocki, <i>Queueing theory model of pentose phosphate pathway</i> , 2022, Scientific Reports, 12, 1, 10.1038/s41598-022-08463-y | 140 | 4,9 |
| P3 | Sylwester Kloska, Krzysztof Palczyński , Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata Wysocka, Paul Davis, Tadeusz Wysocki, <i>Conversion of fat to cellular fuel-Fatty acids beta-oxidation model</i> , 2023, Computational Biology and Chemistry, 104, 10.1016/j.compbiolchem.2023.107860 | 70 | 3,7 |
| P4 | Sylwester Kloska, Krzysztof Palczyński , Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata Wysocka, Paul Davis, Ghada Soliman, Tadeusz Wysocki, <i>Queueing theory model of mTOR complexes' impact on Akt-mediated adipocytes response to insulin</i> , PLoS One, 2022, 17, 12, 10.1371/journal.pone.0279573 | 100 | 3,7 |
| P5 | Sylwester Kloska, Krzysztof Palczyński , Tomasz Marciniak, Tomasz Talaśka, Beata Wysocka, Paul Davis, Tadeusz Wysocki, <i>Integrating glycolysis, citric acid cycle, pentose phosphate pathway, and fatty acid beta-oxidation into a single computational model</i> , Scientific Reports, 2023, 13, 1, 10.1038/s41598-023-41765-3 | 140 | 4,9 |
| P6 | Krzysztof Palczyński , Sandra Śmigiel, Marta Gackowska, Damian Ledziński, Sławomir Bujnowski, Zbigniew Lutowski, <i>IoT application of transfer learning in hybrid artificial intelligence systems for acute lymphoblastic leukemia classification</i> , Sensors, 2021, 21, 23, 10.3390/s21238025 | 100 | 3,7 |
| P7 | Krzysztof Palczyński , Damian Ledziński, Tomasz Andrysiak, <i>Entropy Measurements for Leukocytes' Surrounding Informativeness Evaluation for Acute Lymphoblastic Leukemia Classification</i> , Entropy, 2022, 21, 11, 10.3390/e24111560 | 100 | 2,7 |
| P8 | Sandra Śmigiel, Krzysztof Palczyński , Damian Ledziński, <i>ECG signal classification using deep learning techniques based on the PTB-XL dataset</i> , Entropy, 2021, 23, 9, 10.3390/e23091121 | 100 | 2,7 |

| | | | |
|-----|--|-------------|-------------|
| P9 | Sandra Śmigiel, Krzysztof Palczyński , Damian Ledziński, <i>Deep learning techniques in the classification of ECG signals using R-peak detection based on the PTB-XL dataset</i> , Sensors, 2021, 21, 24, 10.3390/s21248174 | 100 | 3,7 |
| P10 | Krzysztof Palczyński , Sandra Śmigiel, Damian Ledziński, Sławomir Bujnowski, <i>Study of the few-shot learning for ECG classification based on the PTB-XL dataset</i> , Sensors, 2022, 22, 3, 10.3390/s22030904 | 100 | 3,7 |
| | Łącznie: | 1150 | 40,6 |

Wykaz upublicznonych repozytorium kodu z efektami prac

| Nr. | Artykuł | Link |
|-----|---|---|
| 1 | <i>Queueing theory model of Krebs Cycle</i> [P1] | https://github.com/UTP-WTiE/KrebsCycleUsingQueueingTheory |
| 2 | <i>Queueing theory model of pentose phosphate pathway</i> [P2] | https://github.com/UTP-WTiE/PPPQueueingTheory |
| 3 | <i>Conversion of fat to cellular fuel-Fatty acids beta-oxidation model</i> [P3] | https://github.com/UTP-WTiE/FattyAcidsOxidation |
| 4 | <i>Queueing theory model of mTOR complexes' impact on Akt-mediated adipocytes response to insulin</i> [P4] | https://github.com/UTP-WTiE/IrsMtorcQueuesSimulation |
| 5 | <i>Integrating glycolysis, citric acid cycle, pentose phosphate pathway, and fatty acid beta-oxidation into a single computational model</i> [P5] | https://github.com/UTP-WTiE/CellEnergyMetabolismModel |

2. Krótkie wyjaśnienie spójności dla monotematycznego cyklu publikacji

Tematyka rozprawy doktorskiej związana jest z badaniami algorytmów sztucznej inteligencji (ASI), które wykorzystane były w procesie modelowania komórek biologicznych oraz wykrywania wybranych chorób onkologicznych i kardiologicznych. Badania poprzedzone były szeroką analizą literatury, doбором odpowiednich algorytmów, a następnie ich odpowiednim modyfikacjom. Dodatkowo, w każdym przypadku, wykonano wstępne przetwarzanie wektorów uczących, co bezpośrednio wpłynęło na poprawę interpretacji danych uczących, a w konsekwencji na poprawę uzyskanych wyników.

Modyfikacje ASI pozwoliły na opracowanie nowatorskiej odmiany algorytmu genetycznego, wielomodalnych sieci neuronowych oraz sieci hybrydowych. Wszystkie te działania były realizowane pod kątem rozwiązywania zadań z zakresu szeroko rozumianej medycyny.

W pracach P1-P5 przedstawiono modelowanie procesów metabolicznych występujących w komórkach biologicznych. Modele wykorzystywały tzw. Teorię Kolejek. Zaimplementowane, a następnie przebadane zostały one za pomocą zmodyfikowanego algorytmu genetycznego po odpowiednim, wstępnym przetworzeniu wektorów parametryzujących dane modele.

Proces wspomaganego wykrywania: ostrej białaczki limfoblastycznej został przedstawiony w pracach P6-P7, natomiast wybranych chorób kardiologicznych na podstawie analizy sygnału EKG w pracach P8-P10. W pracach tych, w efekcie zastosowania wstępnego przetwarzania wektorów uczących, uzyskano modele o wyższej dokładności klasyfikacji, przy jednocześnie mniejszej złożoności obliczeniowej.

Prace w ramach prac P1-P5 prowadzone były w ramach grantu NCN nr UMO-2019/33/B/ST6/00875, gdzie Autor niniejszej rozprawy był osobą odpowiedzialną za część informatyczną projektu i odpowiednie przygotowanie modeli z wykorzystaniem zmodyfikowanych algorytmów sztucznej inteligencji, po wcześniejszym odpowiednim przetworzeniu wzorców uczących. Efektem tych prac było opracowanie modeli, które w przyszłości mogą ułatwić i wspomóc wczesne wykrywanie takich schorzeń, jak np. otyłość, cukrzyca, czy nawet nowotwory.

3. Wstęp

Medycyna jest dziedziną nauki stawiającą wiele wyzwań. Do najważniejszych zaliczyć możemy: szybkie wykrywanie różnych chorób (szczególnie onkologicznych, kardiologicznych i genetycznych), opracowanie sposobu ich szybkiego i bezbolesnego leczenia, itp. Ważnym wyzwaniem medycyny jest także próba modelowania zachowania pracy komórek biologicznych, tak aby w przyszłości można było nie tylko leczyć pacjentów, ale przede wszystkim odpowiednio szybko zapobiegać różnym groźnym chorobom cywilizacyjnym (np. nowotworom, cukrzycy, otyłości, itp.). Wyzwań stawianych przed medycyną jest jednak o wiele więcej i z dnia na dzień pojawiają się nowe. Rozwiązanie każdego z nich jest bardzo ważne i cenne dla ludzi, niestety jednocześnie bardzo często związane jest z koniecznością rozwiązania innych problemów. Jednym z nich jest pozyskanie dużej liczby, odpowiednio przygotowanych danych. Wynika to z dużej czasochłonności procesu zbierania takich danych, co dodatkowo często jest „etycznie” problematyczne [P1].

Wykrywanie chorób i odpowiednie ich leczenie jest trudnym zadaniem, które wymaga współpracy różnych specjalistów. Niestety, od wielu lat brakuje wysoko wykwalifikowanej kadry medycznej, co niestety także możemy zauważyć w naszym kraju [11]. Liczba specjalistów jest zbyt mała, a tym samym dostęp do nich mocno ograniczony. Czas diagnozowania chorób często jest zbyt długi, co może prowadzić w skrajnych wypadkach nawet do śmierci pacjentów. Zbyt późne wykrywanie schorzeń ma też negatywny wpływ na rodzaj doboru odpowiednich terapii, co bezpośrednio wydłuża proces ich leczenia. Ważnym zadaniem dla nauki jest zatem próba wspomaganie tego procesu z wykorzystaniem np. nowoczesnych narzędzi informatycznych. W efekcie tego niekorzystnego trendu duże nadzieje w rozwiązywaniu problemów medycznych kieruje się w stronę ASI [12], [13]. Dąży się do tego, aby ASI wspomagały pracę personelu medycznego w rozpoznawaniu chorób, opracowaniu strategii leczenia, czy też doboru odpowiednich leków. Od algorytmów ASI poza szybkością pracy wymaga się także wysokiej precyzji działania (np. rozpoznawania schorzeń). Szybkie i precyzyjne rozpoznanie jest bardzo ważne, wręcz kluczowe, w wielu chorobach, szczególnie natury kardiologicznej i onkologicznej.

Algorytmy sztucznej inteligencji możemy także wykorzystać z powodzeniem do modelowania i symulacji procesów biochemicznych zachodzących w żywych organizmach [P1]. Eksperymenty realizowane na komputerze (tzw. eksperymenty *in silico*) pozwalają na przyspieszenie procesu poznawania badanych zjawisk. To z kolei może się przełożyć na opracowywanie lepszych leków i terapii, co może znacznie poprawić skuteczność leczenia i jakość życia pacjentów. Kolejną zaletą modelowania a następnie symulacji komórek biologicznych z wykorzystaniem ASI jest brak dylematów moralnych związanych z prowadzeniem eksperymentów między innymi na zwierzętach. Eksperymenty *in silico* nie powodują żadnej krzywdy żywym organizmom, przez co prowadzenie ich na szeroką skalę nie wywołuje problemów natury etycznej.

Z każdym kolejnym rokiem obserwuje się rozwój ASI w obszarach nauk medycznych. Powstaje wiele artykułów naukowych o nowych sposobach ich wykorzystania w procesie wspomaganie wykrywania chorób [P6, P7, P8, P9, P10] czy też modelowania procesów biochemicznych, w których użycie klasycznych technik pomiarowych jest często zbyt złożone [P1, P2, P3, P4, P5]. Uwidocznia się przy tym także pilna potrzeba poprawy i korekty wielu wcześniejszych rozwiązań, a także konieczność rozwiązywania coraz to nowych wyzwań i problemów.

Obecnie jednym z najczęściej stosowanych ASI są głębokie sieci neuronowe. Dzieje się tak dlatego, że sieci neuronowe potrafią zamodelować skomplikowane procesy interpretując nieustrukturyzowane dane [P6, P8]. Przez „nieustrukturyzowane” rozumiane są takie dane, których „znaczenie” nie jest zależne od pozycji wartości w wektorze informacji. Przykładem danych ustrukturyzowanych są dane tabelaryczne. W tym typie danych wektor jest wierszem tabeli, a pozycja wartości w wektorze odpowiada kolumnie tabeli. Każda kolumna przechowuje inny typ danych, tak więc pozycja wartości w wektorze ma kluczowe znaczenie. Dane nieustrukturyzowane nie posiadają takiej właściwości. Przykładem jest zdjęcie mikroskopowe białej krwinki. Pozycja krwinki na zdjęciu nie wpływa na wynik interpretacji obrazu.

Umiejętność przetwarzania surowych danych jest bardzo cenna, ponieważ przyspiesza proces prototypowania ASI [15], co wynika bezpośrednio z braku konieczności opracowania technik wstępnego przetwarzania danych w celu osiągnięcia pierwszych oczekiwanych rezultatów głębokiego uczenia.

Opracowanie odpowiednich technik przetwarzania danych jest procesem wymagającym prób i błędów w celu zrozumienia zależności zawartych w wektorach uczących. Z tego powodu kuszące jest wykorzystanie głębokiego uczenia, by pominąć ten krok i od razu zacząć modelować odpowiednie systemy. Mimo, iż modelowanie nieprzetworzonych danych jest możliwe [P6, P8], to zastosowanie wstępnego przetwarzania wektorów uczących skutkuje wyższą precyzją uzyskiwanych wyników [P10] przy jednoczesnej redukcji liczby parametrów wykorzystywanych w implementowanych modelach [P7].

Wstępne przetwarzanie danych uczących potrafi wyekstrahować z nich pożądane informacje jednocześnie filtrując nieprzydatne wartości. Pozwala także na realizację hybrydowych modeli przetwarzających różne rodzaje danych (modeli wielomodalnych) [P9] oraz hybrydowych modeli łączących kilka algorytmów ASI w celu uzyskania lepszych rezultatów [P10].

Ponadto, wstępne przetwarzanie danych pozwala też na wprowadzanie dodatkowych mechanizmów sterujących procesem trenowania w celu uproszczenia treningu, co przedstawiono między innymi w pracy [P1]. Pozwala to na lepsze modelowanie zjawisk w eksperymentach *in silico*. Między innymi z tych powodów wstępne przetwarzanie wektorów uczących z pewnością jest ważną czynnością konieczną do budowy prawidłowo działającego systemu opartego o ASI. Jest to szczególnie ważne w medycynie, gdzie duża dokładność i precyzja są bezwzględnie wymagane, a nawet drobny błąd ASI mógłby potencjalnie narazić na utratę zdrowia, czy na nawet życia pacjenta [16].

W niniejszej pracy skupiono się na badaniu i wykorzystaniu algorytmów sztucznej inteligencji w procesie modelowania komórek biologicznych oraz wspomaganie wykrywania wybranych chorób.

W ramach pracy opracowano i wykonano komputerowy model komórki biologicznej oparty na cyklu Krebsa [P1], szlaku pentozofosforanowego [P2], beta-oksydacji kwasów tłuszczowych [P3] oraz odpowiedzi komórkowej na insulinę [P4]. Model ten został zrealizowany w celu przeprowadzenia eksperymentów *in silico* na danych pochodzących z ludzkich komórek.

W efekcie przeprowadzonych badań powstał model pozwalający na analizowanie procesów biochemicznych w komórce. Analizy te mogą posłużyć między innymi do projektowania leków i terapii w leczeniu nowotworów, cukrzycy, otyłości, itp. W tym celu wykorzystano Teorię Kolejek oraz równania Michaelisa-Menten. Dotychczas modelowanie metabolizmów było realizowane za pomocą równań różniczkowych, czyniąc wykorzystanie Teorii Kolejek

podejściem nowatorskim [P1]. W celu dostosowania modelu komórki do odzwierciedlenia obserwacji wykorzystano algorytm genetyczny. Algorytm ten zmodyfikowano wykorzystując wstępne przetwarzanie wektorów parametryzujących symulację.

Ponadto, w niniejszej rozprawie opisano także proces wykrywania ostrej białaczki limfoblastycznej (ALL – *Acute Lymphoblastic Leukemia*) oraz wybranych chorób serca. Systemy wykrywania chorób mogą być wykorzystane jako wspomagające podczas podejmowania decyzji przez lekarzy. W systemach wykrywania chorób wykorzystano wstępne przetwarzanie wektorów uczących by zaprojektować modele hybrydowe. Modele te łączą zarówno umiejętności głębokich sieci neuronowych do przetwarzania danych nieustrukturyzowanych, jak również zalety algorytmów uczenia maszynowego w interpretacji danych ustrukturyzowanych. Badania rozpoczęto od opracowania modeli głębokiego uczenia do przetwarzania surowych danych, a następnie poprawiano je i redukowano ich złożoność obliczeniową za pomocą odpowiedniego przetwarzania danych. Efektem przetwarzania danych była poprawa jakości klasyfikacji nawet o pięć punktów procentowych [P9] i redukcja liczby parametrów modelu ponad 500 razy [P7].

Wykrywanie ALL zostało zrealizowane poprzez interpretację mikroskopowych zdjęć limfocytów. Zdjęcia zostały uzyskane z publicznie dostępnej bazy danych ALL-IDB [17]. Rozpoznawanie wybranych chorób kardiologicznych zostało natomiast zrealizowane poprzez interpretację sygnału EKG pobranego za pomocą 12 sond. Dane zostały pobrane z publicznie dostępnej bazy danych PTB-XL [18].

W obydwóch przypadkach zrealizowano klasyfikację na nieustrukturyzowanych danych. W przypadku ALL były to obrazy. Natomiast dla chorób kardiologicznych były to sygnały EKG. W toku badań opracowano różne techniki by zaadresować fundamentalne różnice pomiędzy rodzajami danych. Warto zaznaczyć w miejscu tym, że za każdym razem wykorzystanie wstępnego przetwarzania wektorów uczących pozwoliło na uzyskanie wyższej o średnio 2.7 punktu procentowego jakości przetwarzania. Ponadto, interpretacja wyników modeli wykorzystujących wstępnie przetworzone dane pozwoliła na wysunięcie wniosków dotyczących zależności występujących w danych. Przykładowo, w toku badań wykazano, że otoczenie limfocytów u osób chorych na ALL różni się od otoczenia limfocytów osób zdrowych w stopniu wystarczającym do wykrycia tej choroby. Takie sformułowanie wniosku było możliwe dzięki odpowiedniemu przygotowaniu danych, a na jego bazie możliwe były dalsze badania.

Niniejsza rozprawa przedstawia cykl badań składający się z dziesięciu artykułów naukowych [P1-P10]. Efekty prac zostały wyeksportowane do pięciu publicznie dostępnych repozytoriów na platformie GitHub. Badania zrealizowano na potrzeby trzech różnych gałęzi medycyny. W efekcie badań opracowano, zaimplementowano i przebadano modele i techniki mogące wesprzeć proces badania i wykrywania ALL i wybranych chorób kardiologicznych. Dodatkowo, zaproponowane nowe techniki modelowania szlaków metabolicznych w komórkach mogą zostać wykorzystane do poszerzenia wiedzy z zakresu mikrobiologii. Zrealizowane modele szlaków metabolicznych zostały udostępnione publicznie. Na ich podstawie możliwe jest zrealizowanie eksperymentów *in silico* mających na celu znalezienie nowych punktów w metabolizmach podatnych na działanie leków, przy leczeniu różnych chorób.

3.1 Hipoteza badawcza, cel i zakres badań

Głównym celem badań było opracowanie odpowiednich modyfikacji algorytmów sztucznej inteligencji, które będą umożliwiały modelowanie komórek biologicznych oraz wspomagały proces wykrywania wybranych chorób onkologicznych i kardiologicznych.

Przez modyfikację algorytmów sztucznej inteligencji (ASI) Autor niniejszej rozprawy rozumie zarówno wprowadzenie zmian, zaproponowanie nowych algorytmów oraz odpowiednie przetworzenie wektorów uczących w celu poprawy działania technik uczenia maszynowego.

Modelowanie komórek biologicznych zostało zrealizowane poprzez modelowanie osobno szlaków metabolicznych, by na koniec scalić je w jeden duży model. Celem modelowania tych szlaków było umożliwienie symulowania życia komórki biologicznej. W ten sposób możliwe jest realizowanie eksperymentów *in silico* w celu projektowania nowych sposobów leczenia groźnych chorób, takich jak nowotwory, cukrzyca, otyłość, itp.

W skład zamodelowanych szlaków metabolicznych wchodzi cykl Krebsa [P1], szlak pentozofosforanowy [P2], beta-oksydacja kwasów tłuszczowych [P3] oraz odpowiedź komórkowa na insulinę [P4]. Szlaki te zostały wybrane ze względu na ich rolę w produkcji energii w komórce biologicznej. Dodatkowo, modele cyklu Krebsa, szlaku pentozofosforanowego oraz beta-oksydacji kwasów tłuszczowych zostały połączone w jeden model komórkowy [P5]. Do ich zamodelowania wykorzystano równania Michaelisa-Menten oraz Teorię Kolejek. Wykorzystanie Teorii Kolejek jest na czas pisania tej pracy nowością, ponieważ do tej pory najczęściej były wykorzystywane w tym celu równania różniczkowe. W celu poprawnego wytrenowania modelu wykorzystano algorytm genetyczny wraz z jego modyfikacją dotyczącą struktury chromosomu. Modyfikacja ta posłużyła do opracowania mechanizmów przyspieszających ocenę chromosomu oraz zmniejszenie wymiarowości funkcji straty.

Opracowanie nowatorskiej metody wykrywania ALL [P6, P7] oraz wybranych chorób kardiologicznych [P8, P9, P10] miało na celu realizację modeli będących w stanie wspierać pracę lekarzy. W tym celu wykorzystano głębokie sieci neuronowe oraz wstępne przetwarzanie wektorów uczących. Tak zrealizowane przetwarzanie miało na celu poprawienie jakości klasyfikacji oraz zmniejszenie liczby parametrów modelu.

Biorąc pod uwagę główny cel pracy zdefiniowano następujące tezy:

- 1) Teoria Kolejek jest skuteczną alternatywą dla równań różniczkowych do modelowania szlaków metabolicznych.
- 2) Wstępne przetwarzanie wektorów parametryzujących pozwala na wykorzystanie algorytmów sztucznej inteligencji do modelowania, w oparciu o Teorię Kolejek a następnie symulację komputerową, komórek biologicznych.
- 3) Wstępne przetwarzanie wektorów uczących w procesie klasyfikacji pozwala na wydobycie z nich ważnych informacji poprawiających dokładność procesu rozpoznawania wybranych chorób onkologicznych i kardiologicznych.
- 4) Wstępne przetwarzanie wektorów uczących pozwala na użycie modeli wykorzystujących mniejszą liczbę parametrów bez utraty dokładności wykonywanego zadania.
- 5) Wykorzystanie hybrydowych metod uczenia maszynowego pozwala na poprawę procesu klasyfikacji.

Tak zdefiniowane cele badań zostały zrealizowane i opisane w monotematycznym cyklu publikacji P1-P10. Prace te były opublikowane w latach 2020-2023. Pięć repozytoriów z kodem

zostały publicznie udostępnione na platformie GitHub, by inne osoby mogły z nich korzystać w swojej pracy naukowej. We wszystkich wykazanych pracach [P1-P10] Autor niniejszej rozprawy odpowiadał i był odpowiedzialny za: opracowanie i przygotowanie ASI, ich odpowiednich modyfikacji, wstępne przetwarzanie wektorów uczących, implementację systemów i modeli w różnych środowiskach programistycznych, wykonanie szeregu testów, symulacji i badań. Brał też czynny udział w pisaniu i redagowaniu tekstów artykułów na etapie składania publikacji, jak i procesu recenzji.

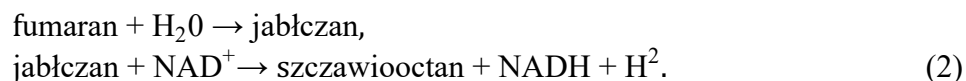
4. Metody badań i wyniki modelowania komórek biologicznych

4.1 Opracowanie sposobu modelowania komórek biologicznych

Biochemiczny stan komórki biologicznej może być przedstawiony poprzez modelowanie jej poszczególnych szlaków metabolicznych (z ang. *metabolic pathway*). Szlak metaboliczny można opisać jako łańcuch reakcji chemicznych [19], w trakcie których substraty przy udziale enzymów przekształcane są w produkty. Przykładem reakcji chemicznej jest hydratacja fumaranu [20]. W czasie tej reakcji fumaran wraz z wodą, dzięki działaniu fumarazy, zostaje przekształcony w jabłczan. Równanie 1 przedstawia tę reakcję:



Przykładem łańcucha reakcji chemicznych jest zamiana fumaranu w szczawiooctan [21]. Fumaran nie może być bezpośrednio przekształcony w szczawiooctan. W tym celu musi być najpierw przekształcony w jabłczan, który to może zostać przekształcony w szczawiooctan. Łańcuch tych reakcji chemicznych jest opisany za pomocą równania 2:



NAD^+ oznacza formę utlenioną dinukleotydu nikotynoamidoadeninowego. NADH jest zredukowaną formą NAD^+ .

Modelowanie szlaków metabolicznych pozwala na symulowanie działania komórki. W ten sposób możliwe jest analizowanie przetwarzania metabolitów w komórce. Akumulacja i niedobór metabolitów pozwalają na diagnozowanie stanu komórki. To z kolei może rzutować na funkcjonowanie całego organizmu. Przykładowo, zakłócenia w szlaku metabolizmu odpowiedzialnym za beta-oksydację kwasów tłuszczowych mogą spowodować nagromadzenie wolnych kwasów tłuszczowych w osoczu [P3]. To z kolei może sugerować występowanie u chorego cukrzycy typu 2, otyłości i niealkoholowe stłuszczenie wątroby [22] [23].

Zrozumienie sposobu działania szlaków metabolicznych za pomocą modelowania umożliwia przeprowadzanie eksperymentów za pomocą symulacji komputerowych (z ang. *in silico*). Tak przeprowadzone modelowanie może zmniejszyć liczbę koniecznych eksperymentów na zwierzętach, jak również przyspieszyć proces testowania nowych leków przed ich wprowadzeniem na rynek [P1]. Eksperymenty *in silico* mają potencjał ułatwić proces identyfikowania nowych punktów w szlakach metabolicznych podatnych na modyfikację z korzyścią dla zdrowia pacjenta [24]. Dzięki temu możliwe jest opracowanie nowych typów leków. Warty podkreślenia jest fakt, że proces ten dzięki nowej metodzie może być znacznie przyspieszony. Cel ten przyświecał modelowaniu metabolizmów komórek biologicznych opisanych w tej pracy.

Modelowanie szlaków metabolicznych zostało zrealizowane poprzez wybór reakcji chemicznych oraz ich substratów i produktów. Decyzja ta ma za zadanie określić jak wiele interakcji między substancjami chemicznymi ma być zamodelowane. Za przykład wzięto część cyklu Krebsa. Ze względu na brak niektórych wartości dotyczących stężeń i właściwości poszczególnych enzymów, możliwe jest bezpośrednio zamodelowanie przejścia cytrynianu w izocytrynian. Możliwe jest też bardziej dokładne zamodelowanie tego procesu poprzez rozbicie go na proces zamiany cytrynianu w cis-akonitan i zamiany cis-akonitanu w izocytrynian. Obydwa podejścia do modelowania zostały opisane równaniem 3:



Wybór dokładności modelowania rzutuje na liczbę symulowanych reakcji oraz zbiór monitorowanych substancji chemicznych. W powyższym przykładzie wykorzystanie uproszczonej reakcji zamiany cytrynianu w izocytrynian wymusza symulowanie tylko jednej reakcji oraz monitorowania stężeń dwóch substancji chemicznych (cytrynianu oraz izocytrynianu). W przypadku modelowania bardziej złożonego procesu konieczne jest symulowanie dwóch reakcji (zamiany cytrynianu w cis-akonitan oraz zamiany cis-akonitanu w izocytrynian) oraz monitorowanie stężeń trzech substancji chemicznych (cytrynianu, cis-akonitanu oraz izocytrynianu). Decyzja ta ma kluczowe znaczenie, ponieważ wraz z zwiększoną złożonością modelu rośnie jego dokładność, a co za tym idzie potencjalna użyteczność, ale także trudność w jego implementacji.

Kolejnym krokiem w procesie modelowania szlaków metabolicznych jest podzielenie substancji biorących udział w reakcjach chemicznych na monitorowane (wartości zmienne) oraz niemonitorowane (wartości stałe). Przykładem tego rozważania jest początkowa reakcja chemiczna zachodząca w szlaku pentozofosforanowym. W tej reakcji glukoza-6-fosforan (*G6P*) zostaje zamieniony w 6-fosfoglukono- δ -lakton (*PGL*). Reakcja ta opisana jest równaniem 4:



W opisanej wyżej reakcji biorą udział *G6P* oraz *PGL*. Poza tym w reakcji znajdują się:

- ester fosforanu dinukleotydu nikotynoamidoadeinowego ($NADP^+$),
- zredukowana forma $NADP^+$ ($NADPH$),
- kation wodorowy (H^+).

Konieczne jest podjęcie decyzji, które z substancji potraktować jako zmienne, a które jako stałe. Wartości zmienne poddawane są modyfikacjom w ramach zachodzących reakcji, podczas gdy wartości stałe pozostają niezmiennie. Przykładowo, jeżeli *G6P* i *PGL* zostaną uznane za jedyne wartości zmienne w tej reakcji, to każda zaistniała reakcja pomniejszy *G6P* i powiększy *PGL*. Wartości $NADP^+$, $NADPH$ i H^+ pozostaną niezmienione pomimo zaistnienia reakcji z ich użyciem. Założenie to można uzasadnić dążeniem komórki do utrzymania stałego poziomu $NADP^+$ oraz $NADPH$, w celu zapewnienia homeostazy i odpowiedniego funkcjonowania procesów życiowych. $NADP^+$ i $NADPH$ są kluczowymi czynnikami w wielu szlakach metabolicznych, gdzie odgrywają istotne role jako nośniki energii w postaci elektronów. Komórka reguluje poziom $NADP^+$ i $NADPH$ poprzez zrównoważenie procesów ich syntezy i degradacji oraz poprzez kontrolę aktywności enzymów, które uczestniczą w ich przemianach.

Oprócz wybrania reakcji chemicznych i wartości zmiennych substancji konieczne jest też opisanie prędkości wspomnianych reakcji. Reakcja chemiczna nie zachodzi natychmiast. Następuje ona w zależności od wielu czynników m.in. stężenia substratów i produktów reakcji. W tej pracy postanowiono zamodelować reakcje chemiczne za pomocą równania kinematyki enzymatycznej Michaelisa-Menten [25]. Za pomocą tego równania możliwe jest obliczenie prędkości każdej reakcji. Przez prędkość reakcji rozumiana jest ilość substratu zamieniająca się w produkt w danej jednostce czasu. Równanie kinetyki enzymatycznej Michaelisa-Menten zostało opisane wzorem 5:

$$v = \frac{V_f \frac{S_1 S_2}{K_{S_1} K_{S_2}} - V_r \frac{P_1 P_2}{K_{P_1} K_{P_2}}}{\left(1 + \frac{S_1}{K_{S_1}} + \frac{P_1}{K_{P_1}}\right) \left(1 + \frac{S_2}{K_{S_2}} + \frac{P_2}{K_{P_2}}\right)}, \quad (5)$$

gdzie:

- v – prędkość reakcji chemicznej,
- V_f – prędkość reakcji zamiany substratów w produkty,
- V_r – prędkość reakcji odwrotnej,
- S_1, S_2, \dots, S_n – wartości stężeń substratów,
- P_1, P_2, \dots, P_n – wartości stężeń produktów,
- $K_{S_1}, K_{S_2}, \dots, K_{S_n}$ – wartości stałe związane z substratami,
- $K_{P_1}, K_{P_2}, \dots, K_{P_n}$ – wartości stałe związane z produktami.

Zdefiniowana w ten sposób prędkość reakcji zależy od dwóch typów wartości:

- Wartości zmiennych – są to wartości stężeń substratów (S_1, S_2, \dots, S_n) i produktów (P_1, P_2, \dots, P_n) monitorowanych w ramach projektowanego modelu. Wartości te ulegają zmianie za każdym razem, gdy reakcja zachodzi. W efekcie przeprowadzenie reakcji wpływa na prędkość reakcji. W przypadku reakcji zamiany G6P na PGL opisanej równaniem 4, wartość S_1 koduje stężenie G6P, a P_1 koduje stężenie PGL.
- Wartości stałych – są to wartości stałych kinetycznych substratów ($K_{S_1}, K_{S_2}, \dots, K_{S_n}$), i produktów ($K_{P_1}, K_{P_2}, \dots, K_{P_n}$). W przypadku reakcji zamiany G6P na PGL opisanej równaniem 4, wartość S_2 oznacza stężenie NADP^+ . Natomiast wartości P_2, P_3 oznaczają stężenia odpowiednio NADPH i H^+ . Wartości stężeń NADP^+ , NADPH i H^+ uznajemy w tym przypadku za stałe.

Kinetyka enzymatyczna Michaelisa-Menten opisana równaniem 5 została wykorzystana do modelowania reakcji chemicznych w szlakach metabolicznych symulowanych w pracach [P1], [P2], [P3], [P4], [P5] wchodzących w skład niniejszego cyklu publikacji. Modelowanie reakcji polegało na:

- Dobraniu stężeń substratów S_1, \dots, S_n oraz produktów P_1, \dots, P_n poprzez analizę literatury.
- Znalezieniu wartości stałych kinetycznych substratów (K_{S_1}, \dots, K_{S_n}) i produktów (K_{P_1}, \dots, K_{P_n}) dla których model działa poprawnie. Poprzez „poprawne działanie modelu” rozumiane jest możliwe wykorzystanie go do przeprowadzenia symulacji *in silico*. Wyniki tych symulacji przeprowadzonych na poprawnie wybranych parametrach są w stanie reprezentować przebieg reakcji biochemicznych żywych komórek.

Wartości stałych substratów (K_{S_1}, \dots, K_{S_n}) oraz stałych produktów (K_{P_1}, \dots, K_{P_n}) są wartościami parametryzującymi jedną reakcję chemiczną. Wektor wartości stałych zawierający parametry K_{S_1}, \dots, K_{S_n} oraz K_{P_1}, \dots, K_{P_n} wszystkich reakcji chemicznej w szlaku metabolicznego jest wektorem parametryzującym szlak metaboliczny. W dalszej części pracy wektor ten będzie nazywany symbolem K . W tej pracy modelowanie szlaku metabolicznego jest tożsame ze znalezieniem optymalnego wektora K .

Wektor K zostaje uznany za optymalny, gdy szlak metaboliczny sparametryzowany jego wartościami umożliwi przeprowadzenie symulacji reprezentującej rzeczywistą komórkę biologiczną. Za pomocą wektora K uzyskiwany jest wektor S opisujący stan komórki

biologicznej w danym kroku czasowym. Wynikiem symulacji jest zbiór wektorów stanu komórki $M = \{S_1, \dots, S_t\}$.

Wektor stężenia substancji S zawiera w sobie wartości stężeń wszystkich substratów i produktów reakcji chemicznych wchodzących w skład szlaku metabolicznego. Znajdują się w nim zarówno stężenia substancji monitorowanych jak i niemonitorowanych. Wartości niemonitorowane nie ulegają zmianom w czasie. Wektor S_t opisuje stan komórki w chwili t . Na początku symulacji (w chwili t_0) wektor S_0 reprezentuje stan komórki przed rozpoczęciem eksperymentu. Wektor S_0 składa się z eksperymentalnie zmierzonych wartości opisanych w literaturze. Zawartość wektora S_0 nie podlega modyfikacji w celu zamodelowania szlaku metabolicznego. W tym celu modyfikowany jest wektor K . Modelowanie szlaku metabolicznego w tej pracy jest tożsame ze znalezieniem odpowiadających wartości wektora K . W tym celu wykorzystane są algorytmy sztucznej inteligencji, które zostaną opisane w dalszej części pracy. Wektor K nie jest modyfikowany w czasie trwania symulacji. Jego wartości podlegają zmianie pomiędzy kolejnymi symulacjami w celu uzyskania wektora pozwalającego na poprawne przeprowadzenie dowolnej liczby symulacji.

Symulowane reakcje chemiczne w modelu szlaku metabolicznego wykorzystują wektory S i K w celu określenia prędkości przejścia substratów w produkty. Wartości prędkości reakcji tworzą wektor V . Wektor V zależy od S_t oraz K zgodnie z równaniami Michaelisa-Menten. Stan komórki w następnym kroku czasowym S_{t+1} zależy od wektora V_t . W celu dokończenia modelu szlaku metabolicznego konieczne jest określenie procesu uzyskania S_{t+1} na podstawie wektorów V_t oraz S_t .

Typowo stosowaną metodą do rozwiązania opisanego problemu jest wykorzystanie równań różniczkowych (ang. *Ordinary Differential Equations*, ODEs) [26] [27] [28] [29] [30]. Modelowanie za pomocą równań różniczkowych posiada jednak pewne wady. Jedną z nich jest brak mechanizmu zapobiegającego osiągnięciu ujemnych wartości stężeń. W ramach przykładu została wykorzystana reakcja chemiczna w szlaku pentozofosforanowym dokonująca przekształcenia rybulozo-5-fosforanu (Ru5P) w rybozo-5-fosforan (R5P). Reakcja ta opisana jest równaniem 6.



Za pomocą równania kinematyki enzymatycznej Michaelisa-Menten możliwe jest opisanie prędkości tej reakcji za pomocą równania 7:

$$v = \frac{V_f \frac{S_1}{K_{S_1}} - V_r \frac{P_1}{K_{P_1}}}{1 + \frac{S_1}{K_{S_1}} + \frac{P_1}{K_{P_1}}}, \quad (7)$$

gdzie:

- S_1 – stężenie Ru5P,
- P_1 – stężenie R5P,
- $V_f, V_r, K_{S_1}, K_{P_1}$ – wartości stałe parametryzujące reakcję.

Założmy, że $S_1 = 1\text{mM}$, $P_1 = 1\text{mM}$ i $v = 2 \frac{\text{mM}}{\text{s}}$. Dodatkowo, założmy, że krok czasowy symulacji wynosi $\Delta t = 1\text{s}$. W takim przypadku S_1 powinno zostać pomniejszone o 2mM , a P_1 powinno zostać powiększone o taką samą wartość. Oznaczałoby to, że nowa wartość substratu reakcji wynosiłaby $S'_1 = S_1 - \Delta t \cdot v = 1\text{mM} - 1\text{s} \cdot 2 \frac{\text{mM}}{\text{s}} = -1\text{mM}$. Jest to błąd symulacji, ponieważ

stężenie metabolitu nie może być mniejsze od zera. Jest to problem typowy dla modelowania szlaków metabolicznych z wykorzystaniem równań różniczkowych. Wykorzystywane są metody mające na celu zapobiegnięciu osiągnięcia wartości ujemnych w równaniach różniczkowych [31]. Metody te mogą jednak wprowadzić błędy obliczeniowe [P1].

Problem ujemnych wartości nie występuje, gdy do modelowania zostanie wykorzystana Teoria Kolejek. Teoria Kolejek jest przede wszystkim wykorzystywana do modelowania zjawisk z dziedziny telekomunikacji i inżynierii. Można ją także wykorzystać do opisanie zmian zachodzących w systemach biologicznych [P1]. Dziedzina ta zakłada modelowanie reakcji chemicznych jako kolejki obsługującej żądanie. W tym systemie reakcja chemiczna ma prawdopodobieństwo „obsłużenia” dyskretnej wartości substratu poprzez zamianę ją w produkt. W tym modelu w każdym kroku czasowym symulacji istnieje pewne prawdopodobieństwo, że określona całkowita wartość substratu zostanie zamieniona w produkt. W ten sposób można zapobiec powstawaniu ujemnych wartości poprzez wyłączenie kolejki do czasu zakumulowania wartości produktu powyżej określonego „żądania” kolejki.

W celu zwizualizowania tego zjawiska odniesiono się do przykładu reakcji chemicznej opisanej wzorem 6. Załóżmy, że $S_1 = 1\text{mM}$, $P_1 = 1\text{mM}$, dyskretna wartość stężenia przekształcana przez kolejkę wynosi $\Delta s = 2\text{mM}$ i prawdopodobieństwo zajścia reakcji chemicznej wynosi $\lambda = 0.5$. „Obsługa żądania” przez kolejkę (reakcję chemiczną) nie następuje, ponieważ $S_1 < \Delta s$. Losowanie wartości prawdopodobieństwa określającego, czy zajdzie reakcja chemiczna, nie miało miejsca, ponieważ nie został spełniony warunek konieczny: wartość stężenia substratów musi być większa lub równa wartości Δs .

Założmy teraz inny przykład; $S_1 = 3\text{mM}$, $P_1 = 1\text{mM}$, $\Delta s = 2\text{mM}$ i $\lambda = 0.5$. W tym przypadku warunek konieczny zajścia reakcji $S_1 \geq \Delta s$ został spełniony. Oznacza to, że można dokonać losowania prawdopodobieństwa, czy kolejka (reakcja chemiczna) obsłuży żądanie (pomniejszy S_1 o Δs i powiększy P_1 o tą samą wartość). W tym celu zostaje wylosowana wartość prawdopodobieństwa x z jednostajnego ciągłego rozkładu prawdopodobieństwa $x \sim U(0,1)$. Do reakcji dochodzi, jeżeli $x < \lambda$. W przeciwnym wypadku reakcja nie następuje. Proces ten został opisany za pomocą równania 8:

$$\Delta s_1 = \begin{cases} 0, & S_1 < \Delta s \\ 0, & x > \lambda \\ \Delta s, & S_1 \geq \Delta s \wedge x < \lambda \end{cases}, \quad (8)$$

$$S_1' = S_1 - \Delta s_1, P_1' = P_1 + \Delta s_1,$$

gdzie Δs_1 jest ilością stężenia substancji „obsłużoną” przez kolejkę (reakcję chemiczną) w danej chwili symulacji.

Podobne obostrzenie może być łatwo zaimplementowane w przypadku otrzymania prędkości ujemnej. Równanie 7 przedstawia wzór na prędkość reakcji w tym przypadku. Ponieważ wszystkie wartości w równaniu muszą być dodatnie, wzór ten może uzyskać wartości ujemne tylko w przypadku opisanym równaniem 9:

$$\begin{aligned}
v &= \frac{V_f \frac{S_1}{K_{S_1}} - V_r \frac{P_1}{K_{P_1}}}{1 + \frac{S_1}{K_{S_1}} + \frac{P_1}{K_{P_1}}}, \\
S_1, P_1, V_f, V_r, K_{S_1}, K_{P_1} &\geq 0, \\
v < 0 &\rightarrow V_f \frac{S_1}{K_{S_1}} - V_r \frac{P_1}{K_{P_1}} < 0, \\
V_f \frac{S_1}{K_{S_1}} &< V_r \frac{P_1}{K_{P_1}}, \\
S_1 &< P_1 \cdot \frac{V_r K_{S_1}}{V_f K_{P_1}},
\end{aligned} \tag{9}$$

Scenariusz opisany równaniem (8) może zaistnieć. Oznacza to, że model musi być w stanie obsłużyć taki przypadek. Niektóre reakcje chemiczne można zamodelować jako reakcje odwracalne. Równania różniczkowe potrzebują dodatkowej metody, by sobie poradzić w tym przypadku. Modelowanie za pomocą Teorii Kolejek wymaga zaledwie dodania dodatkowego warunku koniecznego do przeprowadzenia obsługi żądania przez kolejkę; prawdopodobieństwo obsługi żądania musi być dodatnie. Jest to kolejna zaleta wykorzystania Teorii Kolejek do modelowania szlaków metabolicznych.

Modelowanie za pomocą Teorii Kolejek przedstawione w tej pracy wykorzystuje równania Michaelisa-Menten w celu określenia prawdopodobieństwa zajścia reakcji [P1]. Równania enzymatyczne pozwalają na uzyskanie prędkości reakcji. Prędkość reakcji można uznać za makroskopową reprezentację agregacji wielu mikroskopowych reakcji, które to mogą lub nie, przekształcić dyskretne ilości substratów w produkty. W ten sposób wysokie prawdopodobieństwo zajścia reakcji w mikroskali przekształca się na wysoką prędkość reakcji w makroskali.

Możliwe jest wykorzystanie równań Michaelisa-Menten do implementacji kolejki obsługującej zmianę substratów w produkty. Przybycie obiektów do kolejki opisane jest procesem Poissona. Z kolei rozkład wykładniczy modeluje czas obsługi (odstępów czasowe pomiędzy dwoma kolejnymi zdarzeniami wyjściowymi). Założenia te są zgodne z klasyczną Teorią Kolejek. W efekcie, liczba przybyć w interwale czasu $(t; t + \tau]$ podlega rozkładowi opisanym w równaniu 10:

$$P[(N(t + \tau) - N(t)) = k] = \frac{e^{-\mu(t)\tau} (\mu(t)\tau)^k}{k!}, \tag{10}$$

gdzie:

- $P[(N(t + \tau) - N(t)) = k]$ – prawdopodobieństwo wystąpienia k -żądań do obsługi przez kolejkę w interwale czasowym $(t; t + \tau]$,
- $\mu(t)\tau$ – oczekiwana liczba żądań do obsługi przez kolejkę w interwale czasowym $(t; t + \tau]$.

Czas potrzebny kolejce na przetworzenie żądania (pomniejszenia stężenia substratów o dyskretną wartość i powiększenie stężenia produktów o tą samą wartość) jest opisany za

pomocą rozkładu wykładniczego zmiennej losowej T pod wpływem parametru $\mu(t)$ [P3]. Rozkład ten jest opisany za pomocą równania 11:

$$f(T, \mu(t)) = \begin{cases} \mu(t)e^{-\mu(t)T} & T \geq 0 \\ 0 & T < 0 \end{cases} \quad (11)$$

Kompozycja połączonych ze sobą kolejek (modelujących reakcje chemiczne) opartych o równanie Michaelisa-Menten jest w stanie symulować szlak metaboliczny [P3]. Wewnętrzna struktura równania Michaelisa-Menten sprawia, że akumulacja stężenia substratów zwiększa prawdopodobieństwo zajścia reakcji. W efekcie nadmiar substratu w jednej kolejce może zamienić się w nadmiar produktu, który jest substratem w następnej kolejce w łańcuchu reakcji. W ten sposób sieć połączonych ze sobą kolejek może spełniać rolę realizowaną przez równania różniczkowe [32].

W celu poprawnego zamodelowania szlaku metabolicznego z wykorzystaniem Teorii Kolejek konieczne było znalezienie wartości parametryzujących reakcje chemiczne. Wartości te podstawione pod równanie Michaelisa-Menten wraz z wartościami substratów i produktów miały za zadanie określić wartość prawdopodobieństwa zajścia reakcji. Typowe wartości prawdopodobieństwa zawierają się w przedziale od $[0; 1]$. Jednakże, ze względu na brak ograniczenia maksymalnych wartości stężeń substratów i produktów zrealizowanie obliczeń zgodnie z równaniem Michaelisa-Menten jest w stanie uzyskać dowolne wartości, w tym mniejsze od zera lub większe od 1. Nie oznacza to, że prawdopodobieństwo zajścia reakcji jest ujemne lub większe od 1. Oznacza to, że reakcja albo na pewno zajdzie (wartości większe od 1) albo na pewno nie zajdzie (wartości mniejsze od 0). Z tego powodu należy wprowadzić dodatkowe zasady do przetwarzania wartości spoza zakresu $[0; 1]$.

Jeżeli prawdopodobieństwo reakcji uzyskane z równania Michaelisa-Menten jest ujemne, to dalszy krok zależy od typu modelowanej reakcji chemicznej. Jeżeli reakcja jest odwracalna, to należy zamienić kierunek inkrementacji/dekrementacji substratów i produktów oraz wziąć wartość bezwzględną prawdopodobieństwa zajścia reakcji. Jeżeli natomiast wartość jest większa od 1 to reakcja na pewno zajdzie w tej jednostce czasowej (pod warunkiem, że warunki konieczne zaistnienia reakcji zostały spełnione).

Opracowany na potrzeby tej pracy sposób modelowania komórek biologicznych polega na zaprojektowaniu modeli szlaków metabolicznych. W tym celu wykorzystano Teorię Kolejek do abstrakcyjnego przedstawienia każdej reakcji chemicznej w szlaku za pomocą kolejek. Prawdopodobieństwo zaistnienia reakcji było obliczone w każdym kroku symulacji za pomocą równania Michaelisa-Menten. Zgodnie z tym równaniem prawdopodobieństwo zajścia reakcji jest zależne od wartości stężeń substratów, produktów oraz wartości stałych kinetycznych parametryzujących reakcję. Wektor zawierający wartości stałe kinetyczne parametryzujące wszystkie reakcje chemiczne w modelowanym szlaku jest wektorem parametryzującym model. Znalezienie wartości tego wektora umożliwiających poprawne symulowanie przebiegu szlaków metabolicznych, zostało osiągnięte dzięki wykorzystaniu algorytmów sztucznej inteligencji.

4.2 Algorytmy sztucznej inteligencji

Zadaniem algorytmu sztucznej inteligencji jest wytrenowanie modelu. Przez „wytrenowanie modelu” rozumiane jest znalezienie takich wartości wektora parametryzującego symulację, że wyniki modelu będą poprawnie reprezentowały rzeczywistą realizację szlaku metabolicznego. W tym celu konieczne jest określenie funkcji straty. Funkcja straty jest w stanie przypisać wektorowi parametryzującemu symulację wartość liczbową oceniającą jego rozbieżność z idealnym wektorem cech. Taka funkcja może być wykorzystana do porównywania ze sobą

wektorów pod kątem jakości modelowania szlaków metabolicznych. W ten sposób zadanie algorytmu sztucznej inteligencji może być zdefiniowane jako minimalizacja tej funkcji poprzez modyfikowanie wartości wektora parametryzującego symulację.

ASI wykorzystanym w tej pracy został algorytm genetyczny [33]. Algorytm ten jest heurystyką przeszukiwania przestrzeni wynikowej inspirowaną teorią ewolucji Charlesa Darwina. Algorytm genetyczny wykorzystuje chromosomy. Chromosom jest wektorem wartości parametryzującym model. Każdy chromosom jest potencjalnym rozwiązaniem zadania minimalizacji funkcji straty. Każdy chromosom posiada geny. Gen jest jedną wartością wektora parametryzującego model. W efekcie chromosom składa się z genów. Poniżej znajduje się przykład obrazujący pojęcia chromosomu, genów oraz ich odwzorowania na problem modelowania szlaków metabolicznych.

Założmy modelowany szlak metaboliczny składający się z trzech reakcji chemicznych opisanych wzorami 12:



W celu zamodelowania takiego szlaku metabolicznego potrzebne są cztery substancje chemiczne: A , B , C oraz D . Potrzebne są też trzy reakcje chemiczne. Reakcje te można przedstawić za pomocą modelu Michaelisa-Menten równaniami 13:

$$\begin{aligned} v_{A \rightarrow B} &= \frac{V_{f_{A \rightarrow B}} \frac{A}{K_{S_A}} - V_{r_{A \rightarrow B}} \frac{B}{K_{P_B}}}{1 + \frac{A}{K_{S_A}} + \frac{B}{K_{P_B}}}, \\ v_{B \rightarrow C} &= \frac{V_{f_{B \rightarrow C}} \frac{B}{K_{S_B}} - V_{r_{B \rightarrow C}} \frac{C}{K_{P_C}}}{1 + \frac{B}{K_{S_B}} + \frac{C}{K_{P_C}}}, \\ v_{C \rightarrow D} &= \frac{V_{f_{C \rightarrow D}} \frac{C}{K_{S_C}} - V_{r_{C \rightarrow D}} \frac{D}{K_{P_D}}}{1 + \frac{C}{K_{S_C}} + \frac{D}{K_{P_D}}}. \end{aligned} \tag{13}$$

Tak zamodelowane reakcje przykładowego szlaku metabolicznego posiadają parametry dzielące się na dwie kategorie:

- Wartości zmienne (monitorowane) – ilości stężenia każdej substancji. Wartości początkowe tych substancji są wybrane poprzez analizę literaturową i ulegają zmianie w trakcie realizacji symulacji. Nie są one genami chromosomu. W równaniach 13 są one opisane znakami A , B , C oraz D .
- Wartości stałe – wartości parametryzujące model szlaku metabolicznego. Nie zmieniają się one w trakcie trwania symulacji. Zadaniem algorytmu genetycznego jest znalezienie takich wartości stałych, że wyniki symulacji będą minimalizowały zadaną funkcję straty. Wartości te są genami chromosomu – wektora parametryzującego model szlaku metabolicznego. Za pomocą równań 13 można uzyskać następujący chromosom:

$$\text{Chr} = [V_{f_{A \rightarrow B}} \quad K_{S_A} \quad V_{r_{A \rightarrow B}} \quad K_{P_B} \quad V_{f_{B \rightarrow C}} \quad K_{S_B} \quad V_{r_{B \rightarrow C}} \quad K_{P_C} \quad V_{f_{C \rightarrow D}} \quad K_{S_C} \quad V_{r_{C \rightarrow D}} \quad K_{P_D}].$$

Tak zdefiniowany chromosom Chr może zostać wykorzystany do sparametryzowania modelu przykładowego szlaku metabolicznego. Zdefiniowany został wektor substancji za pomocą równania 14:

$$S = [A \quad B \quad C \quad D]. \quad (14)$$

Wektor S_t przedstawia stan szlaku metabolicznego w czasie t . Oznacza to, że wektor S_0 reprezentuje stężenie substancji na początku symulacji.

Następnie został zdefiniowany wektor prędkości reakcji V oraz funkcja go opisująca. Zostały one zdefiniowane w równaniu 15:

$$V_t = f_V(S_t, \text{Chr}) = [v_{A \rightarrow B} \quad v_{B \rightarrow C} \quad v_{C \rightarrow D}]. \quad (15)$$

Wykorzystując Teorię Kolejek możliwe jest zamodelowanie zamiany wektora V na nowy wektor stężenia substancji w efekcie otrzymana jest funkcja opisana wzorem 16:

$$S_{t+1} = f_K(V_t, S_t). \quad (16)$$

W efekcie model przykładowego szlaku metabolicznego może być reprezentowany następującym pseudokodem.

1. Zdefiniuj $S_0, \text{Chr}, t_{\max}$,
2. $t \leftarrow 0$
3. Dopóki $t < t_{\max}$:
 - $V_t \leftarrow f_V(S_t, \text{Chr})$
 - $S_{t+1} \leftarrow f_K(V_t, S_t)$
 - $t \leftarrow t + 1$
4. Zdefiniuj macierz przebiegu symulacji $M \leftarrow [S_0 \quad \dots \quad S_{t_{\max}-1}]$

Macierz M jest wynikiem symulacji sparametryzowanej za pomocą chromosomu Chr . Chromosom może być oceniony poprzez wykorzystanie funkcji straty f_S analizującej wynik symulacji – macierz M . W efekcie zadanie algorytmu genetycznego może zostać zdefiniowane za pomocą równania 17:

$$\widehat{\text{Chr}} = \arg \min_{\text{Chr}} f_S(M). \quad (17)$$

Chromosom $\widehat{\text{Chr}}$ jest wynikiem końcowym algorytmu genetycznego.

Algorytm genetyczny składa się z następujących kroków:

1. Inicjalizacja populacji poprzez wygenerowanie chromosomów,
2. Ewaluacja chromosomów za pomocą funkcji straty,
3. Selekcja chromosomów na podstawie wyników funkcji straty,
4. Odbudowanie populacji poprzez reprodukcję chromosomów
5. Powrót do punktu 2.

W pierwszym kroku zostaje zainicjalizowana populacja. Populacja to zbiór chromosomów. Inicjalizacja chromosomów jest procesem, w którym tworzone są nowe chromosomy. Liczba chromosomów w populacji jest hiperparametrem (z ang. *hyperparameter*) odpowiadającym za działanie algorytmu genetycznego. Im większa liczba chromosomów w populacji, tym większa szansa na uniknięcie zatrzymania się w minimum lokalnym za cenę zwiększonej liczby

obliczeń. Każdy chromosom jest oceniany przez funkcję straty na podstawie wyników symulacji sparametryzowanej przez ten chromosom. W efekcie ocena przydatności chromosomu wymaga czasu na dokonanie obliczeń, przez co im większa liczba chromosomów w populacji, tym więcej czasu obliczeniowego potrzebuje algorytm.

W następnym kroku każdy chromosom jest oceniany za pomocą funkcji straty. Wynikiem funkcji straty jest wartość liczbowa opisująca jak dobrze chromosom realizuje zadanie. Funkcje straty wykorzystane w tej pracy są opisane w sekcji poświęconej prezentacjom modeli szlaków metabolicznych. Wartość ta jest potrzebna w dwóch celach:

- Ocena znalezionej rozwiązania – przez „znalezioną rozwiązanie” rozumiany jest chromosom posiadający najlepszy wynik funkcji straty. Jeżeli w populacji znajduje się chromosom, którego miara funkcji straty mieści się w przyjętym obszarze akceptowalnych wyników, to dalsze poszukiwanie nie jest konieczne. Algorytm genetyczny kończy w tym momencie działanie, a wynikiem jego działania jest najlepszy chromosom w populacji.
- Możliwość porównania chromosomów między sobą – za pomocą funkcji straty można ocenić, które chromosomy są lepsze, a które gorsze. Pozwala to na odrzucenie nieskutecznych chromosomów zachowując tylko rokujące osobniki.

Jeżeli na etapie ewaluacji chromosomów nie znalazł się osobnik spełniający wymogi akceptowalności rozwiązania, algorytm genetyczny jest dalej realizowany. Następnym krokiem jest selekcja chromosomów. Polega ona na zredukowaniu populacji poprzez usunięcie z niej osobników, które są uznane za nieprzydatne w procesie szukania optymalnego chromosomu. Pozostałe chromosomy są przeznaczone do reprodukcji. Ich liczba jest hiperparametrem algorytmu genetycznego i musi być mniejsza od liczby chromosomów w populacji. Im więcej chromosomów przeznaczonych do reprodukcji, tym większa różnorodność genów zostaje zachowana. Kosztem tego jest wolniejsze przeszukiwanie przestrzeni wynikowej. Modele zrealizowane na potrzeby tej pracy zrealizowały selekcję poprzez wzięcie chromosomów o najniższej wartości funkcji straty.

W ostatnim kroku następuje odbudowa populacji chromosomów poprzez reprodukcję. Reprodukacja chromosomów polega na wygenerowaniu „potomka” (nowego chromosomu) poprzez skrzyżowanie ze sobą dwóch chromosomów i naniesieniu mutacji. Krzyżowanie chromosomu polega na wzięciu dwóch chromosomów i złożenie nowego chromosomu poprzez wybór genów rodziców. Mechanizm ten jest opisany głębiej w dalszej części pracy ze względu na zastosowanie wstępnego przetwarzania wektorów parametryzujących symulację. Operacja krzyżowania pozwala na wykorzystanie genów najlepszych chromosomów do wygenerowania nowego chromosomu. Chromosom potomny ma potencjał przerośnięcia swoich dawców genów poprzez połączenie ich najlepszych cech. Ostatnim etapem jest wprowadzenie mutacji – losowych zmian w genach nowego potomka. Mechanizm ten jest opisany dokładniej w dalszej części pracy. Zadaniem mutacji jest wprowadzenie losowych modyfikacji do chromosomu. W ten sposób nowe potomstwo ma cechy, których „rodzice” nie mają zwiększając potencjał na znalezienie lepszego rozwiązania.

Nowe chromosomy w populacji mają potencjał znalezienia lepszego rozwiązania niż ich „rodzice”. W tym celu konieczna jest ich ocena za pomocą funkcji straty. Algorytm genetyczny przechodzi do punktu drugiego. Procedura ta zakończy swoje działanie w jednym z trzech przypadków:

- algorytm genetyczny znalazł rozwiązanie o dostatecznie niskiej wartości funkcji straty,

- algorytm genetyczny nie znalazł lepszego rozwiązania przez określoną liczbę cykli działania procedury,
- algorytm genetyczny wykonał określoną liczbę cykli.

Powyżej przedstawiono standardową implementację algorytmu genetycznego. Na potrzeby modelowania szlaków metabolicznych zaimplementowano modyfikacje algorytmu w celu poprawienia jego możliwości przeszukiwania przestrzeni wynikowej. Zmiany te opierają się o zastosowane wstępne przetwarzanie wektorów parametryzujących symulacje. Modyfikacje te zostały opisane w następnym sekcji pracy.

4.3 Wstępne przetwarzanie wektorów parametryzujących symulacje

Typowa implementacja algorytmu genetycznego traktuje wszystkie wartości w chromosomach jako zmienne niezależne od siebie. Założenie to bierze się z braku mechanizmu w algorytmie genetycznym przeprowadzającym analizę wpływu zmian wartości genów na wynik. Dokonanie takiej analizy umożliwia zrealizowanie wstępnego przetwarzania chromosomów w celu wyodrębnienia grup znaczeniowych. W takim przypadku możliwe jest przetwarzanie zarówno pojedynczych genów, jak i grup genów. W efekcie uzyskuje się większą kontrolę nad procesem optymalizacji oraz wyższą precyzję uzyskiwanych wyników.

Wyodrębnienie grup znaczeniowych zrealizowane na potrzeby tej pracy zostało wykonane poprzez grupowanie genów w chromosomie odpowiadających za parametryzowanie tej samej reakcji chemicznej. W celu zobrazowania tej koncepcji został wykorzystany przykład szlaku metabolicznego opisanego równaniami 12.

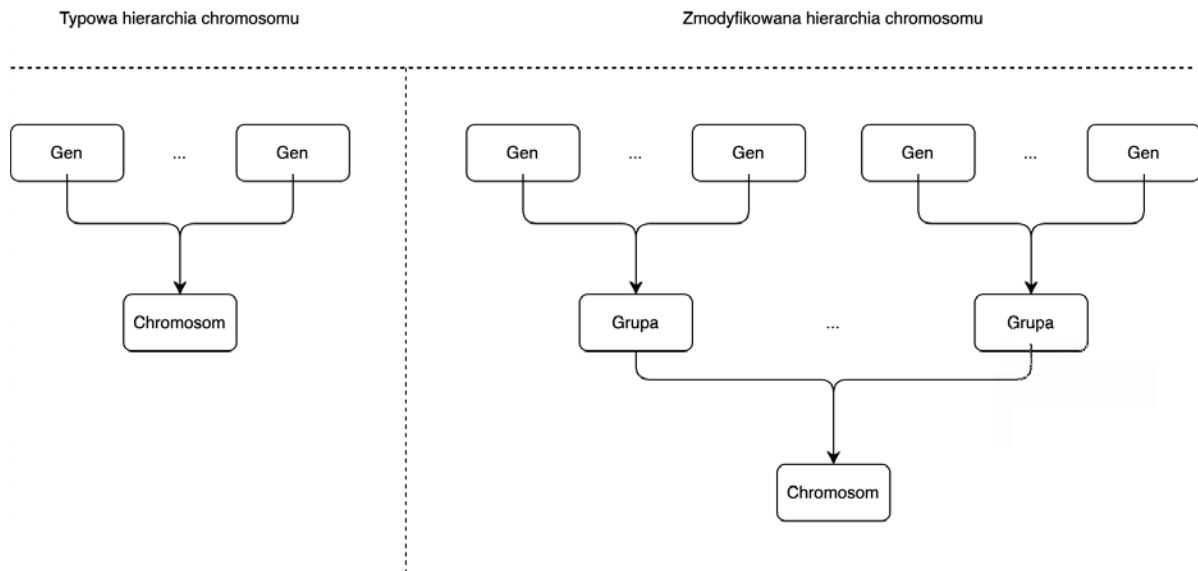
W tym przykładzie szlak metaboliczny składał się z czterech substancji: A, B, C oraz D oraz trzech reakcji chemicznych: $A \rightarrow B$, $B \rightarrow C$ oraz $C \rightarrow D$. Typowa reprezentacja tak zdefiniowanego szlaku metabolicznego za pomocą chromosomu została przedstawiona równaniem 18:

$$\left. \begin{aligned}
 v_{A \rightarrow B} &= \frac{V_{f_{A \rightarrow B}} \frac{A}{K_{S_A}} - V_{r_{A \rightarrow B}} \frac{B}{K_{P_B}}}{1 + \frac{A}{K_{S_A}} + \frac{B}{K_{P_B}}} \\
 v_{B \rightarrow C} &= \frac{V_{f_{B \rightarrow C}} \frac{B}{K_{S_B}} - V_{r_{B \rightarrow C}} \frac{C}{K_{P_C}}}{1 + \frac{B}{K_{S_B}} + \frac{C}{K_{P_C}}} \\
 v_{C \rightarrow D} &= \frac{V_{f_{C \rightarrow D}} \frac{C}{K_{S_C}} - V_{r_{C \rightarrow D}} \frac{D}{K_{P_D}}}{1 + \frac{C}{K_{S_C}} + \frac{D}{K_{P_D}}}
 \end{aligned} \right\} \rightarrow \begin{bmatrix} V_{f_{A \rightarrow B}} \\ K_{S_A} \\ V_{r_{A \rightarrow B}} \\ K_{P_B} \\ V_{f_{B \rightarrow C}} \\ K_{S_B} \\ V_{r_{B \rightarrow C}} \\ K_{P_C} \\ V_{f_{C \rightarrow D}} \\ K_{S_C} \\ V_{r_{C \rightarrow D}} \\ K_{P_D} \end{bmatrix}, \quad (18)$$

Równanie 18 opisuje typową strukturę chromosomu w algorytmie genetycznym. Korzystając z takiego kodowania wektora parametryzującego model, każdy gen jest traktowany osobno. W pracach [P1-5] wykorzystano grupowanie wzajemnie współpracujących ze sobą genów, by zredukować liczbę interakcji pomiędzy wartościami. Tak zrealizowane kodowanie prezentuje równanie 19:

$$\left. \begin{aligned}
 v_{A \rightarrow B} &= \frac{V_{f_{A \rightarrow B}} \frac{A}{K_{S_A}} - V_{r_{A \rightarrow B}} \frac{B}{K_{P_B}}}{1 + \frac{A}{K_{S_A}} + \frac{B}{K_{P_B}}} \\
 v_{B \rightarrow C} &= \frac{V_{f_{B \rightarrow C}} \frac{B}{K_{S_B}} - V_{r_{B \rightarrow C}} \frac{C}{K_{P_C}}}{1 + \frac{B}{K_{S_B}} + \frac{C}{K_{P_C}}} \\
 v_{C \rightarrow D} &= \frac{V_{f_{C \rightarrow D}} \frac{C}{K_{S_C}} - V_{r_{C \rightarrow D}} \frac{D}{K_{P_D}}}{1 + \frac{C}{K_{S_C}} + \frac{D}{K_{P_D}}}
 \end{aligned} \right\} \rightarrow \begin{bmatrix} V_{f_{A \rightarrow B}} \\ K_{S_A} \\ V_{r_{A \rightarrow B}} \\ K_{P_B} \\ V_{f_{B \rightarrow C}} \\ K_{S_B} \\ V_{r_{B \rightarrow C}} \\ K_{P_C} \\ V_{f_{C \rightarrow D}} \\ K_{S_C} \\ V_{r_{C \rightarrow D}} \\ K_{P_D} \end{bmatrix}, \quad (19)$$

Wykonane w ten sposób grupowanie genów w kategorii dodaje poziom złożoności chromosomów. Zamiast typowej hierarchii geny → chromosom wprowadzona modyfikacja ustala nową hierarchię geny → grupy genów → chromosom. Hierarchia ta została wizualnie przedstawiona na Rys 1.



Rys. 1 – Zmodyfikowana hierarchia chromosomu uwzględniająca grupowanie genów.

Wprowadzone grupowanie genów niesie ze sobą możliwość ułatwienia algorytmowi genetycznemu procesu znajdowania wartości optymalnej. Tak zrealizowane przetwarzanie chromosomów pozwoliło na uproszczenie funkcji straty poprzez dokonywanie wstępnej ewaluacji chromosomów jeszcze w trakcie reprodukcji. Zastosowanie tego mechanizmu zostało szczegółowo opisane w dalszej części pracy podczas przedstawienia modelu szlaku metabolicznego cyklu Krebsa.

4.4 Zamodelowane szlaki metaboliczne

W niżej wymienionych pracach opisano modelowanie następujących szlaków metabolicznych i sygnalizacyjnych:

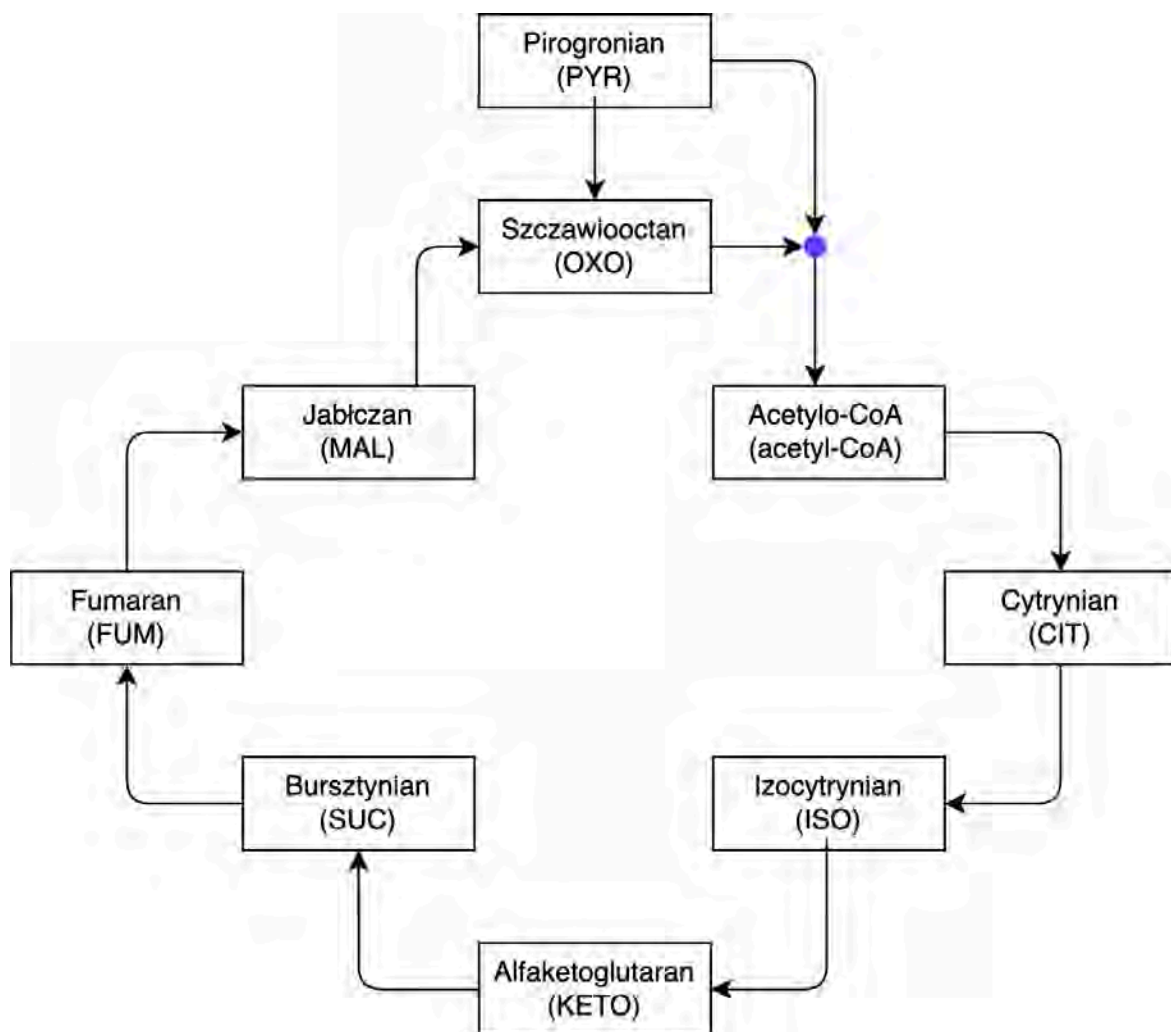
- cykl Krebsa [P1],
- szlak pentozofosforanowy [P2],
- beta-oksydacja kwasów tłuszczowych [P3],
- odpowiedź komórkowa na insulinę [P4].

Każdy z tych szlaków został zamodelowany z wykorzystaniem Teorii Kolejek oraz równań kinetyki enzymatycznej Michaelisa-Menten lub prawa zachowania mas. Ostatnim etapem prowadzonych badań było skomponowanie modelu, który łączył wcześniej opracowane modele oraz nowo wprowadzony szlak obejmujący proces glikolizy [P5].

W kolejnych sekcjach zostały opisane modele każdego ze szlaków. Dodatkowo, w opisie modelowania cyklu Krebsa przedstawiono metody wykorzystania wstępnego przetwarzania wektorów parametryzujących symulację w celu poprawienia jakości optymalizacji parametrów. Mechanizm ten był wykorzystywany także w pozostałych modelach.

Cykl Krebsa

Pierwszym zamodelowanym szlakiem metabolicznym jest cykl Krebsa. Cykl ten znany jest też pod nazwą cyklu kwasów trikarboksylowych. Szlak ten zachodzi w macierzy mitochondrialnej i jego zadaniem jest przetwarzanie metabolitów, dostarczając komórce niezbędnych prekursorów aminokwasów czy związków redukujących, wykorzystywanych w wielu innych reakcjach. Cykl Krebsa dostarcza także nośniki energii pod postacią guanozyno-trifosforanu (GTP), który jest odpowiednikiem trójfosforanu adenozyne (ATP) [34] [35] [36]. Jednym z produktów cyklu jest dwutlenek węgla (CO_2). Rysunek 2 przedstawia model cyklu Krebsa oraz zachodzące w nim reakcje.



Rys. 2 – Schemat modelu cyklu Krebsa (opracowanie na bazie P1).

Tabela 1 przedstawia związki chemiczne uwzględnione podczas modelowania cyklu Krebsa. Zawiera ona wartości stężeń na początku symulacji oraz źródło literaturowe tych wartości. Tabela 2 opisuje zamodelowane reakcje chemiczne.

Tabela 1 - Opis substancji chemicznych wykorzystanych do zamodelowania cyklu Krebsa [P1].

| Metabolit, molekula | Stężenie (mmol/l) | Źródło |
|---|-------------------|--------|
| Koenzym A (CoA) | 0.044 | [37] |
| Pirogronian (PYR) | 0.0586 | [38] |
| Acetylo-CoA (acetyl-CoA) | 0.5 | [39] |
| Cytrynian (CIS) | 0.19 | [26] |
| Cis-Akonitan (Cis-Aco) | 0.0016 | [40] |
| Izocytrynian (ISO) | 0.02 | [39] |
| Alfaketoglutaran (KETO) | 0.54 | [41] |
| Bursztynylo-CoA (Suc-CoA) | 0.66 | [41] |
| Bursztynian (SUC) | 0.07 | [42] |
| Fumaran (FUM) | 0.485 | [37] |
| Jabłczan (MAL) | 0.495 | [41] |
| Szczawiooctan (OXO) | 0.006 | [41] |
| Adenozyno-5'-trifosforan (ATP) | 0.159 | [38] |
| Adenozyno-5'-difosforan (ADP) | 0.0937 | [38] |
| Guanozyno-5'-difosforan (GDP) | 0.0012 | [39] |
| Forma utleniona dinukleotydu nikotynoamidoadeninowego (NAD ⁺) | 0.099 | [39] |
| Forma zredukowana NAD ⁺ (NADH) | 0.025 | [39] |
| Woda (H ₂ O) | 0.170 | [39] |
| Jon wodorowy (H ⁺) | 5.2e-6 | [39] |

Tabela 2 - Reakcje chemiczne wykorzystane do zamodelowania cyklu Krebsa [P1].

| Numer | Wzór reakcji |
|-------|--|
| 1 | $\text{PYR} + \text{CoA} + \text{NAD}^+ \rightarrow \text{acetyl-CoA} + \text{CO}_2 + \text{NADH}$ |
| 2 | $\text{PYR} + \text{HCO}_3^- + \text{ATP} \rightarrow \text{OXO} + \text{ADP} + \text{P}_i$ |
| 3 | $\text{OXO} + \text{acetyl-CoA} + \text{H}_2\text{O} \rightarrow \text{CIT} + \text{CoA} + \text{H}^+$ |
| 4 | $\text{CIT} \rightarrow \text{Cis-Aco} + \text{H}_2\text{O}$ |
| 5 | $\text{Cis-Aco} + \text{H}_2\text{O} \rightarrow \text{ISO}$ |
| 6 | $\text{ISO} + \text{NAD}^+ \rightarrow \text{KETO} + \text{CO}_2 + \text{NADH}$ |
| 7 | $\text{KETO} + \text{NAD}^+ + \text{CoA} \rightarrow \text{Suc-CoA} + \text{P}_i + \text{GDP}$ |
| 8 | $\text{Suc-CoA} + \text{P}_i + \text{GDP} \leftrightarrow \text{SUC} + \text{GTP} + \text{CoA}$ |
| 9 | $\text{SUC} + \text{FAD} \rightarrow \text{FUM} + \text{FADH}_2$ |
| 10 | $\text{FUM} + \text{H}_2\text{O} \rightarrow \text{MAL}$ |
| 11 | $\text{MAL} + \text{NAD}^+ \leftrightarrow \text{OXO} + \text{NADH} + \text{H}_2$ |

W celu uproszczenia modelowania postanowiono połączyć ze sobą izocytrynian z cis-akonitanem i bursztynolo-CoA z kwasem bursztynowym. W efekcie wartości cytrynianu oraz kwasu bursztynowego wyniosły odpowiednio 0.0216 i 0.73 mmol/l. Izocytrynian, cis-akonitan, bursztynolo-CoA oraz bursztynian są substancjami przechodnimi – jak tylko powstaną z reakcji zostają przekształcone w następnej reakcji. Zsumowanie wartości ich stężeń do dwóch metabolitów (izocytrynianu oraz bursztynianu) ułatwiło modelowanie oraz zwiększyło stabilność modelu.

W toku symulowania realizacji cyklu Krebsa monitorowano wartości stężeń następujących metabolitów:

- pirogronian,
- acetylo-CoA,
- cytrynian,
- izocytrynian,
- alfa-ketoglutaran,
- bursztynian,
- fumaran,
- jabłczan,
- szczawiooctan.

W zrealizowanym modelu pirogronian jest substancją wchodzącą do cyklu Krebsa. Z tego powodu wartość stężenia pirogronianu była uzupełniana do wartości początkowej za każdym razem, gdy pirogronian został wykorzystany do tworzenia szczawiooctanu czy acetylo-CoA.

Dodatkowo, wprowadzono tak zwane „odpływy równoważące” w celu zamodelowania odpływu metabolitów z cyklu Krebsa. Odpływy te uzasadnione są koniecznością

wykorzystania tych substancji przez komórkę biologiczną w innych celach, np. związków będących prekursorami innych cząsteczek biologicznie czynnych. Odpływy równoważące zostały wykorzystane do zrównoważenia stężenia szczawiooctanu i cytrynianu. Szczawiooctan jest wykorzystywany przez komórkę w glukoneogenezie, cyklu ornitynowym i syntezie kwasów tłuszczowych. Cytrynian natomiast jest transportowany z mitochondriów do cytoplazmy, gdzie jest wykorzystywany w syntezie kwasów tłuszczowych.

Odpływy równoważące są z punktu widzenia modelu traktowane jak reakcje chemiczne. Oznacza to, że mają również wartości stałe, które muszą zostać zoptymalizowane przez algorytm genetyczny. Prawdopodobieństwo odpływu substancji (szczawiooctanu lub cytrynianu) jest równe stężeniu substancji przemnożonej przez wartość stałą. Wartość stała jest „genem” w „chromosomie”.

Tak zdefiniowany model został wytrenowany za pomocą algorytmu genetycznego. Wartości związków z tabeli 1 zostały wykorzystane w celu inicjalizacji symulacji. Reakcje chemiczne opisane na rysunku 1 zostały przedstawione za pomocą równań Michaelisa-Menten. Wartości stałe parametryzujące reakcje chemiczne zostały scalone do postaci „chromosomu”.

Model cyklu Krebsa został wytrenowany w taki sposób, by wartości monitorowanych substancji ustabilizowały się jak najbliżej wartości początkowych. Korzystając z tego wymagania została zdefiniowana funkcja straty równaniem 20:

$$f(X) = \frac{1}{n} \sum_{i=1}^n \left| X_{i,1} - \frac{1}{100} \sum_{j=1}^{100} X_{i,(T-j)} \right|, \quad (20)$$

gdzie:

- T – liczba kroków czasowych w symulacji,
- X – macierz wektorów stanów szlaku metabolicznego,
- X_i – wektor wartości i -tej monitorowanej substancji w czasie (przykład: $X_{1,100}$ to wartość pirogronianu w setnym kroku czasowym).

Funkcja straty oblicza średnią różnicę pomiędzy początkowymi wartościami monitorowanych substancji, a ich końcową wartością. Końcowa wartość została obliczona poprzez uśrednienie stu ostatnich symulowanych wartości stężeń metabolitów. Uśrednienie to ma za zadanie zredukowanie wpływu szumów typowych dla realizacji procesów probabilistycznych.

Funkcja straty opisana równaniem 20 jest prosta do implementacji. Problemem jednak jest obecność trywialnego rozwiązania polegającego na wyzerowaniu wszystkich wartości V_f oraz V_r w chromosomie. Wszystkie reakcje chemiczne zostały zamodelowane za pomocą równania Michaelisa-Menten opisanym równaniem 7.

Jeżeli współczynniki V_f oraz V_r będą równe zero, to prędkość reakcji v będzie równa zero bez względu na wartości stężeń substancji. W takim przypadku nie dojdzie do żadnej reakcji w modelu. To z kolei sprawi, że symulacja ustabilizuje się osiągając tą samą wartość, co wartość początkowa. W takim przypadku funkcja straty opisana równaniem 20 może osiągnąć minimum globalne równe zero poprzez wyzerowanie parametrów V_f oraz V_r we wszystkich reakcjach chemicznych. Idealnym rozwiązaniem tej funkcji straty jest dobór parametrów, przez które symulacja nie odbędzie się. Oznacza to, że funkcja straty zdefiniowana równaniem 20 umożliwi uzyskanie rozwiązania, które osiąga minimalną wartość (równą zero) jednocześnie nie spełniając wymogów modelowania (wartości substancji nie ulegają zmianie).

W typowym algorytmie genetycznym konieczne byłoby zmodyfikowanie równania 20 o dodatkowe czynniki penalizujące brak postępu w szeregach czasowych stężeń metabolitów. Takie rozwiązanie oznaczałoby bardziej skomplikowaną funkcję straty i przez to trudniejsze przeszukiwanie przestrzeni wynikowej.

W pracy [P1] rozwiązano ten problem za pomocą wstępnego przetwarzania „chromosomów”. „Chromosom” podzielono na grupy „genów” odpowiadające za parametryzację jednej reakcji chemicznej. Korzystając z tego podziału możliwe było wprowadzenie wymagania podczas inicjalizacji oraz reprodukcji „chromosomów”. Wymaganie to wymuszało, by wartości prawdopodobieństwa zajścia każdej reakcji wynosiły pomiędzy 1% a 10%. Warunek ten wymusił, by reakcje zachodziły w symulacji bez konieczności zwiększania złożoności funkcji straty. Podczas gdy funkcja straty wymaga zrealizowania całej symulacji w celu oceny „chromosomu”, warunek ten wymaga jedynie wykonania pierwszego kroku symulacji. Oznacza to, że zapewnienie spełnienia tego warunku nie wymaga takich nakładów obliczeniowych, jakim jest ocena „chromosomu” za pomocą funkcji straty.

Grupowanie „genów” zostało też wykorzystane w procesie krzyżowania „chromosomów”. Zamiast wymiany pojedynczych „genów”, „chromosom potomny” zostaje złożony z grup „genów” swoich rodziców. Oznacza to, że wszystkie wartości parametryzujące jedną reakcję chemiczną pochodzą od tego samego rodzica. Proces ten został opisany za pomocą równania 21:

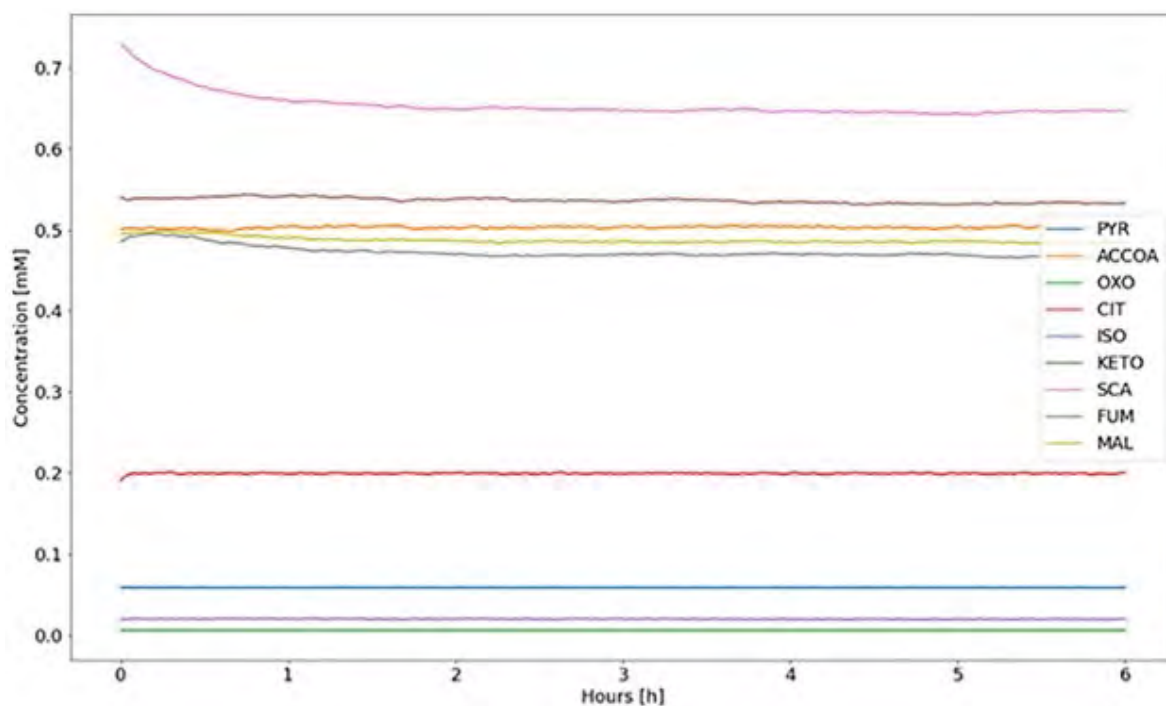
$$\text{Chr}^3 = \left| \left\{ (p = 0) \cdot \text{Chr}_i^1 + (p = 1) \cdot \text{Chr}_i^2 \mid i \in C^+ \cap i < n \right\} \right|, \quad (21)$$

gdzie:

- $\text{Chr}^1, \text{Chr}^2$ – „chromosomy potomne”,
- Chr^3 – „chromosom potomny”,
- Chr_i^1 – i -ta grupa „genów” „chromosomu” pierwszego,
- p – zmienna losowa, której wartość jest określana z binarnej dystrybucji prawdopodobieństwa.

Rozwiązanie to sprawiło, że „chromosomy potomne” są stabilniejsze. Dzieje się tak dlatego, że otrzymują w ten sposób wszystkie wartości parametryzujące reakcję chemiczną od jednego rodzica. Ponieważ tylko „zwycięskie chromosomy” są wykorzystywane do reprodukcji, każda grupa „genów” relatywnie dobrze parametryzuje jedną reakcję chemiczną.

Tak skonfigurowany algorytm genetyczny został wykorzystany do wytrenowania modelu szlaku metabolicznego cyklu Krebsa. Rysunek 3 przedstawia wykresy stężeń monitorowanych metabolitów. Tabela 3 porównuje wartości początkowe stężeń z ustabilizowanymi wartościami końcowymi.



Rys. 3 – Wynik symulacji cyklu Krebsa [P1]. Na wykresie zostały zaprezentowane pierwsze 6 godzin symulacji.

Tabela 3 - Wyniki symulacji cyklu Krebsa [P1]. Wyniki modelu symulacyjnego zostały uśrednione z prób eksperymentu.

| Metabolit | Początkowa Stężenie początkowe (literatura) | Stężenie końcowe (model) | Relatywna różnica |
|-------------------------------------|---|--------------------------|-------------------|
| Pirogronian (PYR) | 0.0586 | 0.0586 | 0.0% |
| Acylokoenzym A (ACCOA) | 0.05 | 0.5028 | +0.55% |
| Szczawiooctan (OXO) | 0.006 | 0.0059 | -1.5% |
| Cytrynian (CIT) | 0.19 | 0.1994 | +4.96% |
| Izocytrynian + cis-Akonitan (ISO) | 0.0216 | 0.0216 | 0.0% |
| Alfaketoglutaran (KETO) | 0.54 | 0.5346 | -1.01% |
| Bursztynylo-CoA + Bursztynian (SCA) | 0.73 | 0.6473 | -11.33% |
| Fumaran (FUM) | 0.485 | 0.467 | -3.72% |
| Jabłczan (MAL) | 0.495 | 0.4847 | -2.08% |

Reakcje cyklu Krebsa były symulowane przez 24 wirtualne godziny. Modelowany system uzyskał stabilność po 5.5 godziny. Rezultaty zostały uznane za akceptowalne, ponieważ największa relatywna różnica została uzyskana przez połączenie bursztynylo-CoA i bursztynianu i wynosiła ona -11.33%.

Tak wytrenowany model został przetestowany poprzez symulację cyklu Krebsa pod działaniem leku hamującego jedną z reakcji cyklu. W tym celu zostało wybrane badanie [43], które dostarczyło wartości stężeń metabolitów po podaniu leków w terapii przeciwnowotworowej. Model cyklu Krebsa został przetestowany poprzez sprawdzenie, czy przy zahamowaniu reakcji chemicznej opisanej w badaniu uzyska on stan szlaku metabolicznego porównywalny z wynikami badania.

Badanie to sprawdziło wpływ Tamoxifenu, leku wykorzystywanego w leczeniu raka piersi, w połączeniu z lekami używanymi przez cukrzyków – Metforminy i Phenforminy. Pomysł na użycie tych leków w terapii przeciwnowotworowej podyktowany był wynikami obserwacji klinicznych [44-46]. Jedną z metod oceny jakości terapii był pomiar stężeń substancji w cyklu Krebsa. Cykl Krebsa jest bardzo ważny dla komórek rakowych ze względu na ich wysokie zapotrzebowanie na energię. W efekcie, zmniejszenie efektywności reakcji cyklu Krebsa i zmniejszenie stężenia metabolitów może potwierdzić efektywność zastosowanej metody leczenia.

W toku pracy udało się poprawnie zamodelować wpływ Tamoxifenu w połączeniu z Metforminą lub Phenforminą. Tabele 4 i 5 przedstawiają porównanie symulowanych wartości stężeń z wartościami rzeczywistymi [43]. Ze względu na trudności w uzyskaniu wartości stężeń metabolitów w badaniu [43] nie wszystkie metabolity cyklu Krebsa zostały uwzględnione w tabelach.

Tabela 4 - Porównanie wartości stężeń podczas terapii z wykorzystaniem Metforminy [P1]. Wyniki modelu symulacyjnego zostały uśrednione z prób eksperymentu.

| Metabolit | Zmiana stężenia po podaniu Metforminy i Tamoxifenu w porównaniu do braku terapii [43] | Zmiany w wynikach symulacji modelu |
|-------------------------|---|------------------------------------|
| Pirogronian (PYR) | -35% | -35% |
| Cytrynian (CIT) | -15% | -16.18% |
| Izocytrynian (ISO) | -40% | -39.17% |
| Alfaketoglutaran (KETO) | -55% | -54.47% |
| Fumaran (FUM) | -37% | -37.03% |
| Jabłczan (MAL) | -39% | -39.43% |

Tabela 5 - Porównanie wartości stężeń podczas terapii z wykorzystaniem Phenforminy [P1]. Wyniki modelu symulacyjnego zostały uśrednione z prób eksperymentu.

| Metabolit | Zmiana stężenia po podaniu Phenforminy i Tamoxifenu w porównaniu do braku terapii [43] | Zmiany w wynikach symulacji modelu |
|-------------------------|--|------------------------------------|
| Pirogronian (PYR) | -65% | -65% |
| Cytrynian (CIT) | -60% | -59.28% |
| Izocytrynian (ISO) | -65% | -63.91% |
| Alfaketoglutaran (KETO) | -80% | -79.38% |
| Fumaran (FUM) | -50% | -50.84% |
| Jabłczan (MAL) | -53% | -53.22% |

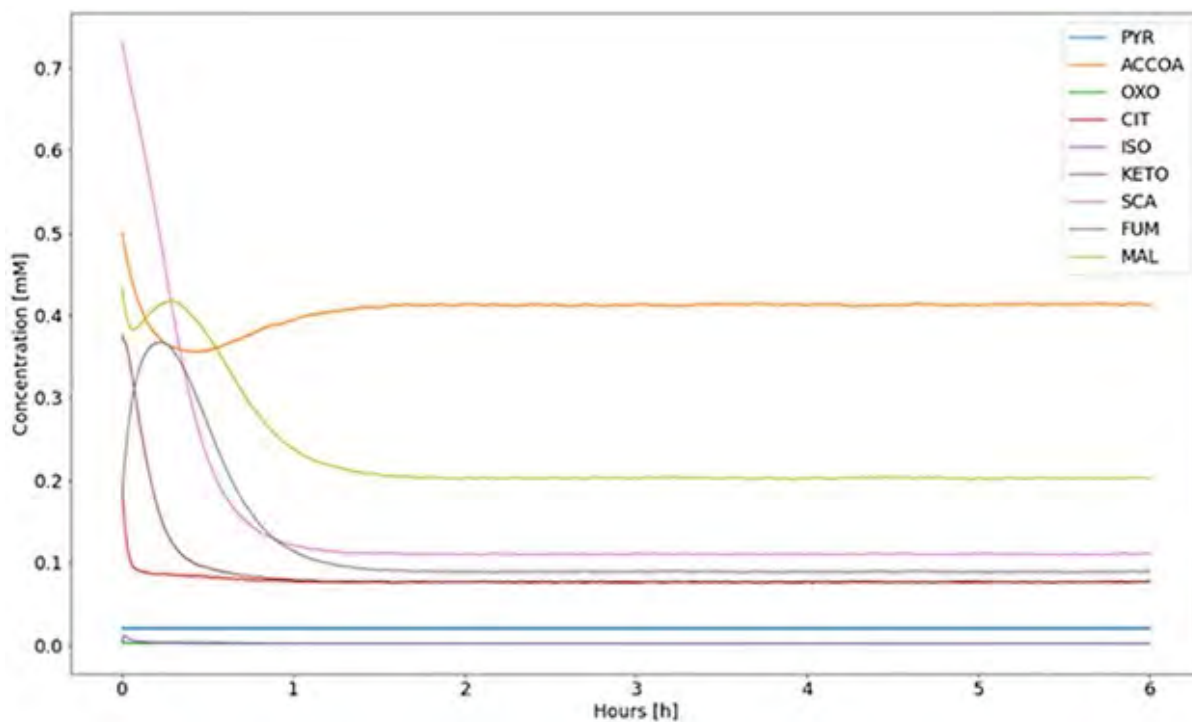
Model cyklu Krebsa został oceniony poprzez sprawdzenie jego wyników w scenariuszu symulowania użycia Phenforminy lub Metforminy i Tamoxifenu. Poprzez porównanie wyników eksperymentalnych z symulacyjnymi uznano, że model działa prawidłowo. Następnie postanowiono wykorzystać wytrenowany i przetestowany model do zasymulowania wartości stężeń pozostałych metabolitów, niemierzonych w badaniu [43]. Wyniki tych połączeń zostały przedstawione w tabeli 6 i 7. Przebieg symulacji został zwizualizowany na rysunkach 4 i 5. Przedstawiają one przebiegi symulacji, w czasie których wartości stężenia substancji ustabilizowały się na poziomach przedstawionych w tabelach 6 i 7. Uzyskano w ten sposób wartości stężeń substancji, które nie zostały pomierzone w badaniu [43]. Największy spadek został zaobserwowany dla połączenia bursztynolo-CoA i bursztynianu wynosząca -84.75%. Biorąc pod uwagę, że według badania [43] wartość alfaketoglutaranu spadła o -80% (według symulacji o -79.38%), to spadek ten jest prawdopodobny.

Tabela 6 - Porównanie wyników symulacji pod wpływem zahamowania reakcji wywołanych użyciem Tamoxifenu i Phenforminy [P1]. Wyniki modelu symulacyjnego zostały uśrednione z prób eksperymentu.

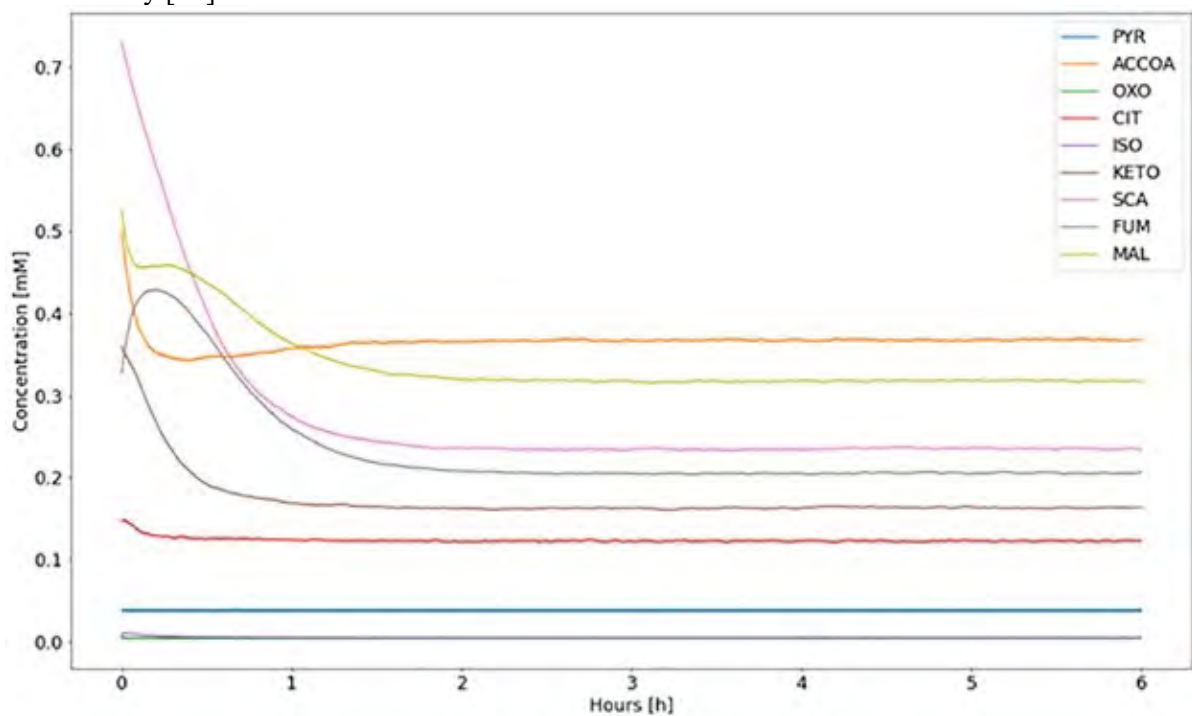
| Metabolit | Stężenie początkowe (wartość literaturowa) | Stężenie końcowe (wartość symulowana) | Relatywna różnica |
|---------------------------------|--|---------------------------------------|-------------------|
| Pirogronian (skrót jeżeli mają) | 0.0205 | 0.0205 | 0.0% |
| Acetylo-CoA | 0.5 | 0.4129 | -17.42% |
| Szczawiooctan | 0.006 | 0.0025 | -58.7% |
| Cytrynian | 0.1899 | 0.0773 | -59.28% |
| Izocytrynian + cis-Akonitan | 0.0056 | 0.002 | -63.91% |
| Alfaketoglutaran | 0.3764 | 0.0776 | -79.38% |
| Bursztynylo-CoA + Bursztynian | 0.73 | 0.1113 | -84.75% |
| Fumaran | 0.1825 | 0.0897 | -50.84% |
| Jabłczan | 0.4335 | 0.2028 | -53.22% |

Tabela 7 - Porównanie wyników symulacji pod wpływem zahamowania reakcji wywołanych użyciem Tamoxifenu i Metforminy [P1]. Wyniki modelu symulacyjnego zostały uśrednione z prób eksperymentu.

| Metabolit | Stężenie początkowe (literaturowa) | Stężenie końcowe (symulowana) | Relatywna różnica |
|-------------------------------|------------------------------------|-------------------------------|-------------------|
| Pirogronian | 0.0381 | 0.0381 | 0.0% |
| Acetylo-CoA | 0.5 | 0.3676 | -26.49% |
| Szczawiooctan | 0.006 | 0.0044 | -27.46% |
| Cytrynian | 0.1466 | 0.1229 | -16.18% |
| Izocytrynian + cis-Akonitan | 0.0081 | 0.0049 | -39.17% |
| Alfaketoglutaran | 0.3596 | 0.1637 | -54.47% |
| Bursztynylo-CoA + Bursztynian | 0.73 | 0.2346 | -67.86% |
| Fumaran | 0.3272 | 0.206 | -37.03% |
| Jabłczan | 0.5235 | 0.3171 | -39.43% |



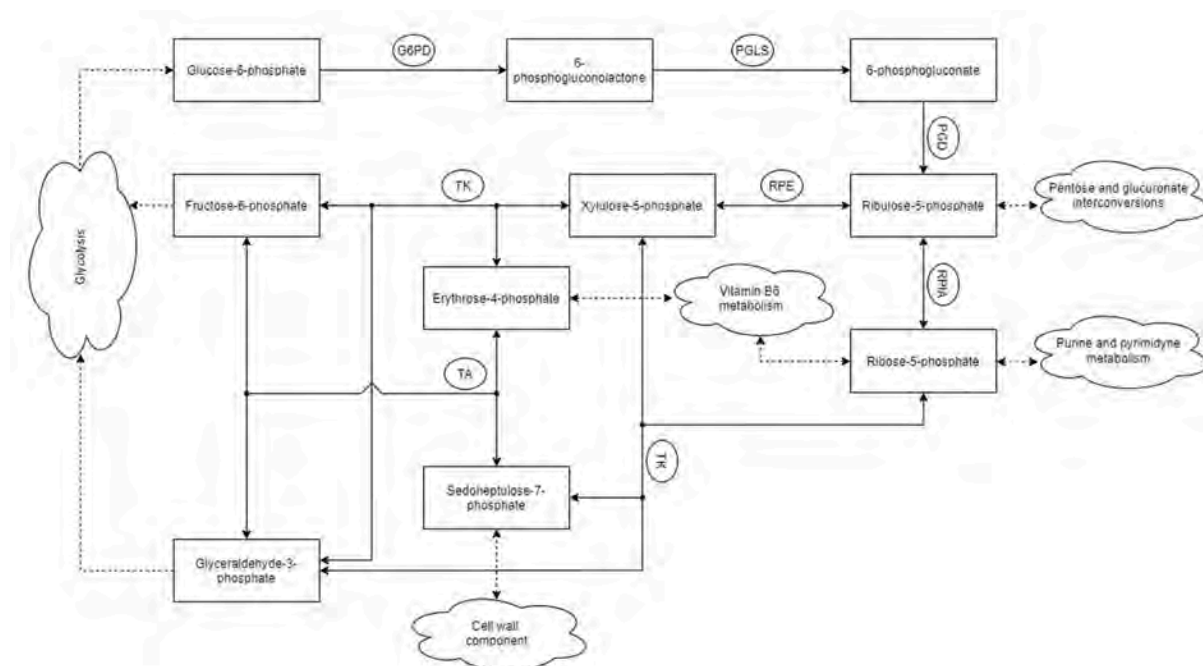
Rys. 4 – Zmiany stężeń substancji w symulacji cyklu Krebsa w czasie zastosowania Tamoxifenu i Phenforminy [P1].



Rys. 5 – Zmiany stężenia substancji w symulacji cyklu Krebsa w czasie zastosowania Tamoxifenu i Metforminy [P1].

Szlak Pentozofosforanowy

Szlak Pentozofosforanowy (PPP) jest szlakiem metabolicznym, którego głównym substratem jest glukoza-6-fosforan (G6P). Produkty PPP są kluczowe w tworzeniu nowych komórek, ze względu na fakt, że dostarcza on cukru, rybozy, stanowiącej element rybonukleozydów i rybonukleotydów. PPP jest jednym ze szlaków metabolicznych, który wykazuje zwiększoną aktywność w komórkach nowotworowych. W porównaniu do zdrowych komórek, aktywność PPP w komórkach nowotworowych jest nawet 8 razy większa [P2]. Z tego powodu konieczne jest opracowanie leków o działaniu przeciwnowotworowym, których zadaniem jest blokowanie określonych reakcji chemicznych w PPP. Rysunek 6 przedstawia poglądowy schemat PPP.



Rys. 6 - Schemat PPP [P2].

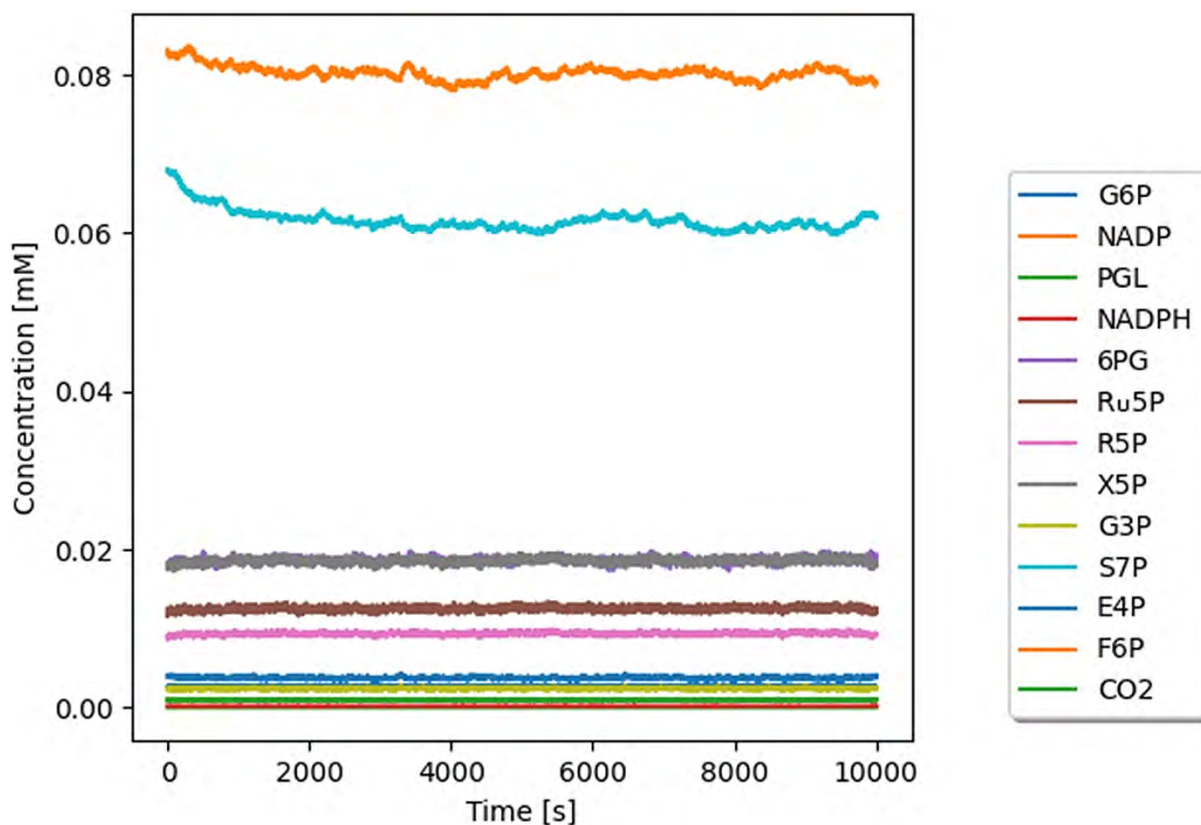
PPP zostało zamodelowane w oparciu o Teorię Kolejek i równania Michaelisa-Menten zgodnie z procedurami określonymi dla cyklu Krebsa. Szlak ten został opisany reakcjami chemicznymi przedstawionymi w tabeli 8. Tak jak w przypadku cyklu Krebsa, model został wytrenowany w celu uzyskania punktu stabilności, jak najbliższej wartości początkowych. Zrealizowano to za pomocą funkcji straty minimalizującej różnicę pomiędzy średnią wartością końcowej fazy symulacji, a wartościami początkowymi. W celu zapobiegnięcia wyzerowania prawdopodobieństwa zajścia reakcji wykorzystano grupowanie „genów” w „chromosomie”, by wymusić odpowiednie prawdopodobieństwo zajścia wszystkich reakcji na początku symulacji. Wykorzystanie tego mechanizmu umożliwiło na prawidłowe wytrenowanie modelu. Efekty trenowania modelu zostały przedstawione w tabeli 9. Przebieg symulacji w warunkach normalnej pracy PPP został przedstawiony na rysunku 7.

Tabela 8 - Reakcje zamodelowane w PPP [P2].

| Numer reakcji | Reakcja |
|---------------|--|
| 1 | $G6P + NADP^+ \rightarrow PGL + NADPH + H^+$ |
| 2 | $PGL + H_2O \rightarrow 6PG + H^+$ |
| 3 | $6PG + NADP^+ \rightarrow Ru5P + NADPH + H^+ + CO_2$ |
| 4A | $Ru5P \rightarrow R5P$ |
| 4B | $Ru5P \rightarrow X5P$ |
| 5 | $R5P + X5P \rightarrow G3P + S7P$ |
| 6 | $X5P + E4P \rightarrow G3P + F6P$ |
| 7 | $G3P + S7P \rightarrow E4P + F6P$ |

Tabela 9 - Wyniki wytrenowanego modelu szlaku pentozofosforanowego uśrednione spośród prób eksperymentu [P2].

| Metabolit | Stężenie początkowe (literaturowa) | Stężenie końcowe (symulowana) | Relatywna różnica |
|-------------------|------------------------------------|-------------------------------|-------------------|
| G6P | 0.0026 | 0.0026 | 0.0% |
| NADP ⁺ | 0.001 | 0.001 | 0.0% |
| NADPH | 0.0002 | 0.0002 | 0.0% |
| PGL | 5.0e-6 | 9.3e-6 | +86.0% |
| 6PG | 0.018 | 0.019 | +5.5% |
| Ru5P | 0.012 | 0.012 | 0.0% |
| R5P | 0.009 | 0.009 | 0.0% |
| X5P | 0.018 | 0.018 | 0.0% |
| G3P | 0.00234 | 0.00242 | +3.4% |
| S7P | 0.068 | 0.062 | -8.8% |
| E4P | 0.004 | 0.004 | 0.0% |
| F6P | 0.083 | 0.079 | -4.8% |



Rys. 7 - Symulacja szlaku pentozofosforanowego w warunkach normalnych [P2].

Trenowanie modelu zakończyło się sukcesem, ponieważ wszystkie substancje poza PGL zmieniły ilość początkowego stężenia poniżej 10%, co pokazano w tabeli 9. Wysoka zmiana PGL jest akceptowalna, ponieważ jest to substancja przechodnia. Jakakolwiek ilość PGL prawie natychmiast zostaje zamieniona w 6PG, przez co jego wysoka wariancja jest zjawiskiem oczekiwanym [P2].

W celu przetestowania modelu efekty symulacji zostały porównane z wynikami badań empirycznych [44]. W artykule opisano efekt redukcji ekspresji genu kodującego enzym PGD w komórkach raka płuc. Redukcja ekspresji genu jest techniką wykorzystywaną w badaniach molekularnych. Prowadzi to do tymczasowego zmniejszenia lub wygaszenia funkcji tego genu w komórkach lub organizmie, a co za tym idzie do obniżenia liczby cząsteczek enzymu PGD w komórkach. Obniżenie ilości enzymu PGD wpływa na obniżenie szybkości reakcji, którą enzym ten katalizuje, co w efekcie spowalnia rozwój guza. Proces ten wpłynął na stężenie niektórych substancji w PPP. Stężenia PGL i 6PG zwiększyły się odpowiednio 7.9 i 11 razy w porównaniu do komórek, w których nie wywołano redukcji ekspresji genu PGD. Natomiast stężenie G3P zmalało 3.8 razy. Warunki te zostały odwzorowane na wytrenowanym modelu poprzez zahamowanie reakcji dehydrogenazy 6-fosfoglukonianowej (PGD). Reakcja ta opisana jest w tabeli 8 pod numerem trzecim. Tabela 10 przedstawia wyniki symulacji i zestawia je z empirycznie uzyskanymi wynikami.

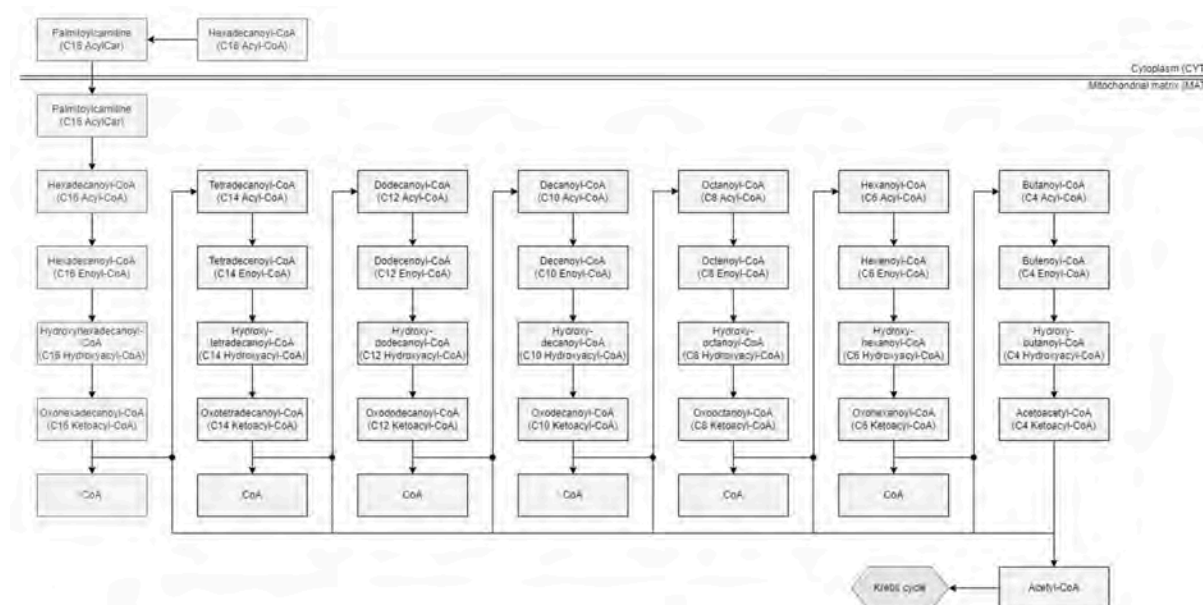
Tabela 10 - porównanie wyników symulacji PPP z badaniem empirycznym. Wyniki modelu symulacyjnego zostały uśrednione z prób eksperymentu [P2].

| Metabolit | Zmiana zaobserwowana w badaniu [44] | Wynik symulacji przy zahamowaniu reakcji 95% | Wynik symulacji przy zahamowaniu reakcji 98% | Wynik symulacji przy zahamowaniu reakcji 100% |
|-----------|-------------------------------------|--|--|---|
| PGL | +7.9x | +5.24x | +6.08x | +7.89x |
| 6PG | +11.0x | +11.88x | +14.56x | +21.8x |
| G3P | -3.8x | -2.63x | -3.85x | -14.29x |

Blokowanie reakcji 3 w modelu pozwoliło na odwzorowanie wyników zaobserwowanych podczas badania empirycznego [44]. Każdą zmianę zaobserwowaną w badaniu [44] udało się odwzorować za pomocą symulacji *in silico* oraz modyfikacji hamowania reakcji. Na tej podstawie stwierdzono poprawność zrealizowanego modelowania. Tak zaprojektowany model może posłużyć do eksperymentowania *in silico* w celu opracowania nowych leków i terapii celujących we wpływ na aktywność PPP.

Beta-oksydacja kwasów tłuszczowych

Beta-oksydacja kwasów tłuszczowych odgrywa istotną rolę w przetwarzaniu energii dostępnej wewnątrz komórki [P3]. Beta-oksydacja prowadzi do skrócenia łańcuchów kwasów tłuszczowych w celu wytworzenia acetylo-CoA, a także NADH oraz FADH₂ [45]. Równania chemiczne oraz ich wzajemne interakcje przedstawiono na rysunku 8. Celem modelowania jest poprawne odzwierciedlenie wartości stężeń metabolitów zawierających łańcuchy węglowe o różnej liczbie atomów węgla (C16, C14, C12, C10, C8, C6, C4) w czasie procesu „rozszczepiania” na krótsze łańcuchy. Wartości te są wynikiem zsumowania stężeń wszystkich metabolitów w modelu o równolicznym łańcuchu.



Rys. 8 – Schemat poglądowy reakcji chemicznych modelowanych w beta-oksydacji kwasów tłuszczowych [P3].

Funkcja straty mająca poprowadzić algorytm genetyczny do znalezienia prawidłowych wartości parametryzujących symulację uległa zmianie. W przeciwieństwie do modeli cyklu Krebsa oraz PPP, beta-oksydacja kwasów tłuszczowych była dostosowana poprzez odwzorowanie punktów w czasie wyznaczonych przez badanie empiryczne [46].

Z badań uzyskanych w [46] wytypowano 10 wektorów stężeń metabolitów biorących udział w beta-oksydacji kwasów tłuszczowych tworzących szereg czasowy. Każdy z wektorów zawierał stężenia metabolitów o różnych długościach łańcuchów węglowych: C16, C14, C12, C10, C8, C6 oraz C4. Z tego badania zostały do trenowania wytypowane trzy wektory, tj.:

- wektor w czasie rozpoczęcia beta-oksydacji kwasów tłuszczowych - w celu inicjalizacji stanu symulacji,
- wektor zmierzony po upływie 1/3 czasu badania – jako wzorzec dla funkcji straty,
- wektor zmierzony pod koniec badania – jako wzorzec dla funkcji straty.

Wektory zmierzone po upływie 1/3 badania oraz pod koniec badania zostały wykorzystane do wytrenowania modelu. Pozostałe wektory posłużyły jako wyznaczniki jakości treningu modelu. Funkcja straty realizująca to zadanie składa się z podfunkcji oceniającej odległość wektora uzyskanego w drodze symulacji od wektora zmierzonego w czasie badania empirycznego. Podfunkcja ta została opisana wzorem 22:

$$f_p(\widehat{X}, X) = \sum_{i=1}^{\{4,6,8,10,12,14,16\}} \left(\frac{\widehat{X}_i - X_i}{X_i} \right)^2 \quad (22)$$

gdzie:

- \widehat{X} – wektor symulowanego stężenia metabolitu o danej długości łańcucha węglowego,
- X – wektor stężenia zaobserwowany podczas badania empirycznego,
- X_i – wartość stężenia metabolitów o i -tej długości łańcucha węglowego. Przykład: dla $i=10$, X_i odpowiada stężeniu metabolitów o długości łańcucha węglowego C10.

Funkcja straty została zdefiniowana z wykorzystaniem opisanej podfunkcji. Wzór 23 przedstawia postać funkcji straty:

$$f(\widehat{X}^1, \widehat{X}^2, X^1, X^2) = f_p(\widehat{X}^1, X^1) + f_p(\widehat{X}^2, X^2) \quad (23)$$

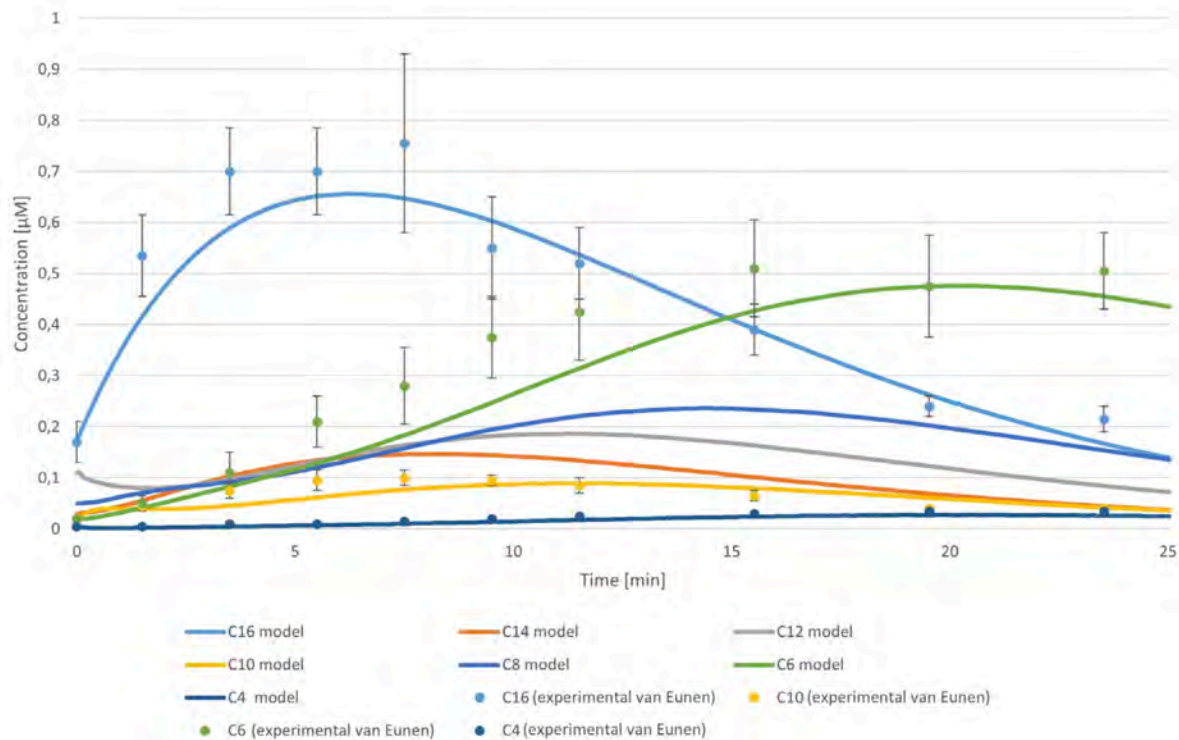
gdzie:

- \widehat{X}^1 – wektor symulowanego stężenia po upływie 1/3 symulacji,
- \widehat{X}^2 – wektor symulowanego stężenia metabolitów pod koniec symulacji,
- X^1 – wektor stężenia zaobserwowany podczas badania empirycznego po upływie 1/3 badania,
- X^2 – wektor stężenia zaobserwowany podczas badania empirycznego pod koniec badania.

Funkcja straty opisana równaniem 23 ma za zadanie pokierować procesem trenowania modelu, by ten odwzorował wyżej wymienione punkty pomiarowe badania empirycznego. Pozostałe wektory nie były wykorzystane w czasie trenowania modelu, lecz służyły za wyznacznik stopnia generalizacji procesu uczenia.

Wynik symulacji został przedstawiony na rysunku 9. Na rysunku zostały przedstawione stężenia metabolitów zależne od czasu. Punkty wraz z zakresami reprezentują wartości

zmierzone w czasie empirycznego badania [46]. Zadaniem modelu było jak najlepsze odwzorowanie przedziałów wartości wyznaczonych przez punkty i zakresy.



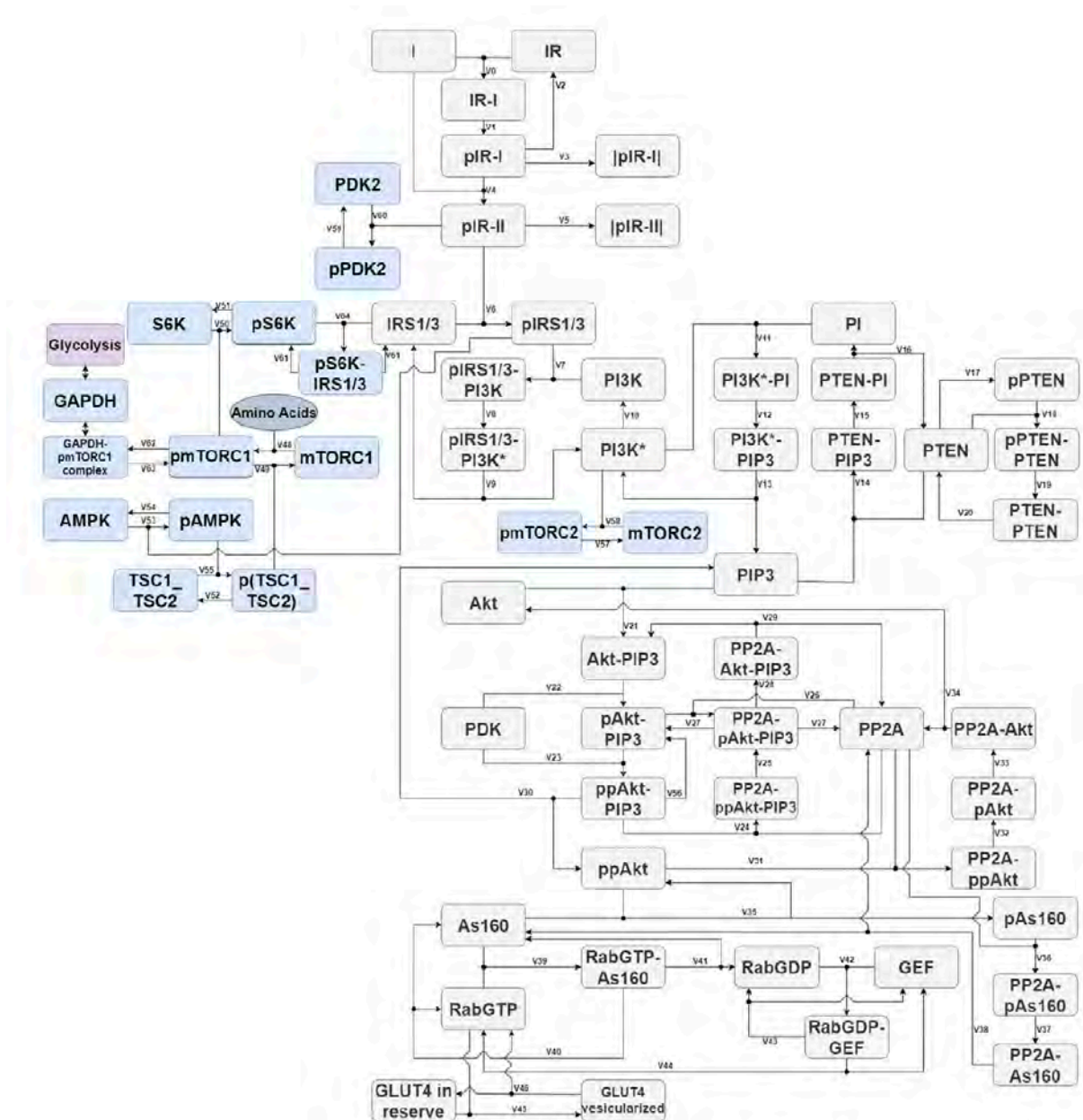
Rys. 9 – Wynik symulacji modelu beta-oksydacji kwasów tłuszczowych [P3].

Model ten został połączony wraz z modelami cyklu Krebsa oraz szlaku pentozofosforanowego w jeden duży model [P5]. W ten sposób powstał model komórki biologicznej, który jest przedmiotem tej pracy. Za jego pomocą możliwa jest identyfikacja kluczowych punktów w szlakach metabolicznych, których zmiany za pomocą leków mogą być wykorzystane w terapii.

Odpowiedź komórkowa na insulinę

Dodatkowo, zamodelowano odpowiedź komórkową na insulinę, która to nadzoruje i utrzymuje odpowiedni poziom glukozy we krwi [P4]. Mechanizm ten wykorzystuje kompleksy mTOR. Kompleksy te regulują metabolizm komórki wpływając na procesy takie jak glikoliza, cykl Krebsa, czy beta-oksydacja kwasów tłuszczowych [47].

mTOR łączy się z innymi proteinami i tworzy dwa kompleksy białkowe: mTORC1 oraz mTORC2. Te kompleksy są odpowiedzialne za regulację wielu ważnych procesów wewnątrzkomórkowych takich jak rozwój komórki, synteza protein, czy transkrypcja DNA [48]. Rysunek 10 przedstawia diagram zamodelowanych reakcji chemicznych kinazy mTOR.

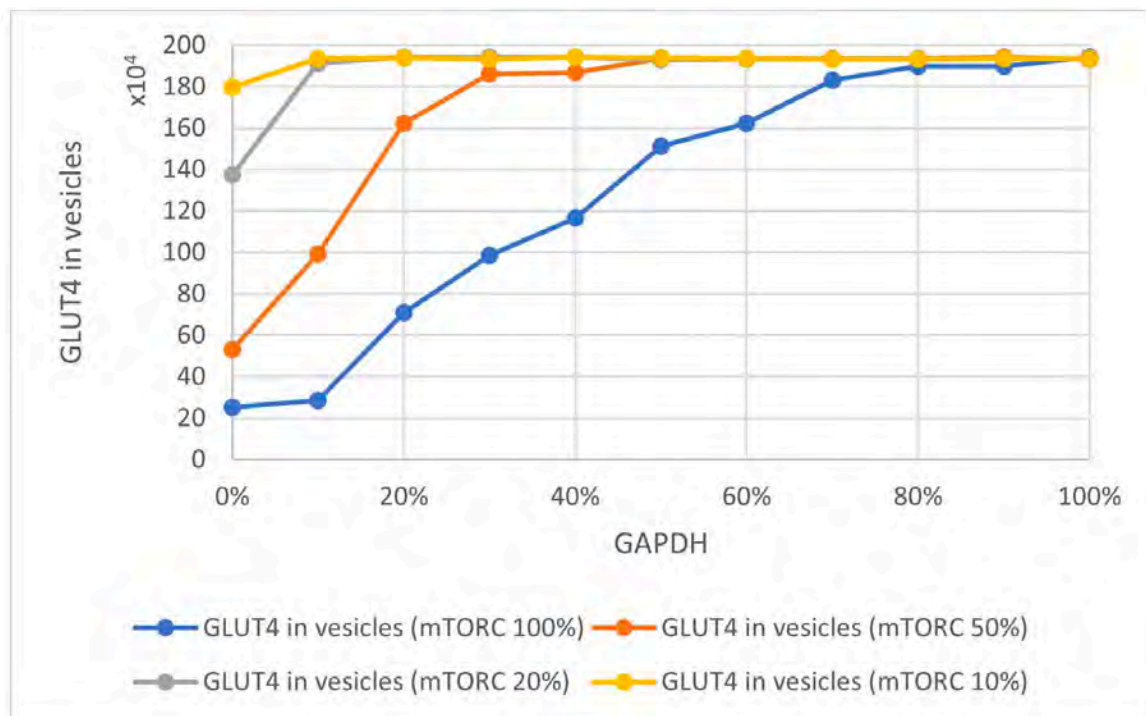


Rys. 10 – Diagram reakcji chemicznych w modelu kinazy mTOR [P4].

Model został wytrenowany w celu odwzorowania szlaku sygnalizacyjnego odpowiedzi komórkowej na insulinę zgodnie z obecnym stanem wiedzy [52, 53]. W badanym modelu przeprowadzono analizę wpływu aktywności kompleksu białkowego mTORC1 na liczbę cząsteczek GLUT4 gotowych do transportu glukozy w komórkach. Na proces ten wpływa również jeden z enzymów glikolitycznych – GAPDH. W zależności od poziomu jego aktywności w glikolizie może on wpływać na liczbę mobilizowanych cząsteczek GLUT4 [P4].

Wyniki modelu zostały przedstawione na rysunku 11. Rysunek przedstawia zależność pomiędzy liczbą molekuł GLUT4 *in vesicles*, a stopniem zajętości GAPDH dla różnej aktywności mTORC1. Kiedy system mTORC jest nieblokowany (aktywność 100%) 200 000 molekuł GLUT4 jest wykorzystywanych do transportu glukozy do komórki. Jednakże, na liczbę tych komórek ma wpływ aktywność komórki, która zależy od „zajętości” GAPDH oraz aktywności mTORC. Rysunek 11 przedstawia efekt zmniejszania aktywności mTORC (osiągalną poprzez podanie odpowiednich leków) na liczbę cząsteczek GLUT4 dla różnych

stopni „zajętości” GAPDH. Wyniki zaprezentowane na rysunku 11 wskazują na to, że leki hamujące aktywność mTORC (o przynajmniej 50%) mają znaczący wpływ na liczbę cząsteczek GLUT4 kierowanych do błony komórkowej w celu transportu glukozy wewnątrz komórki. Podobne wnioski można wysnuć z rezultatów badań opisanych w artykułach [49] [50] potwierdzając działanie zaimplementowanego modelu.



Rys. 11 - wynik modelowania kinazy mTOR. Relacja pomiędzy liczbą molekuł GLUT4 in vesicles oraz poziomem "zajętości" GAPDH [P4].

W artykułach P1 – P4 wykorzystano Teorię Kolejek w celu zamodelowania metabolizmów cyklu Krebsa [P1], szlaku pentozofosforanowego [P2], beta-oksydacji kwasów tłuszczowych [P3] oraz odpowiedzi komórkowej na insulinę [P4]. W artykule [P5] opisano połączenie cyklu Krebsa, szlaku pentozofosforanowego oraz beta-oksydacji kwasów tłuszczowych w jeden model. Udowodniono w ten sposób tezę pierwszą.

Modele metabolizmów wykorzystywały układy równań Michaelisa-Menten w celu obliczenia prędkości reakcji substancji chemicznych. W celu dostrojenia równań wykorzystano wstępne przetwarzanie wektorów parametryzujących symulację do zmodyfikowania algorytmu genetycznego. Przetwarzanie to polegało na pogrupowaniu genów odpowiadających za parametryzację tego samego równania Michaelisa-Menten w grupy znaczeniowe. Tak zdefiniowane grupy znaczeniowe wykorzystano w celu implementacji mechanizmu dziedziczenia. Wprowadzenie tego mechanizmu pozwoliło na zredukowanie wymiarowości funkcji straty oraz uzyskanie lepszych wyników dopasowania modelu. Udowodniono w ten sposób tezę drugą.

5. Metodyka badań w procesie wspomaganie wykrywania wybranych chorób onkologicznych i kardiologicznych

Głębokie sieci neuronowe są jednym z nielicznych algorytmów będących w stanie przetwarzać dane nieustrukturyzowane. Przez dane nieustrukturyzowane rozumiane są dane, które nie można przechowywać w tabelach złożonych z wierszy i kolumn. U podstaw większości algorytmów uczenia maszynowego, takich jak las losowy [51], czy xgboost [52] leży założenie, że dane są ustrukturyzowane. W efekcie głębokie sieci neuronowe zdobyły wielką popularność będąc jednym z niewielu rozwiązań mogących przetwarzać nieustrukturyzowane dane w postaci tekstu, obrazów czy sygnałów.

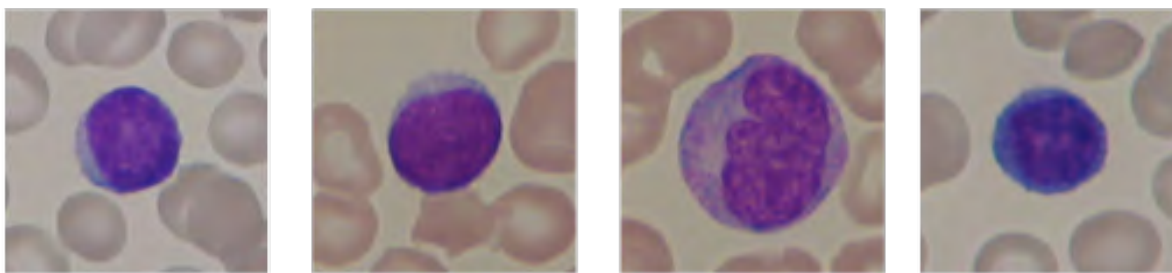
Ta cecha pracy sieci neuronowych jest często wykorzystywana przez różne zespoły badawcze, choć wstępne przetwarzanie wektorów uczących umożliwia zwiększenie dokładności przetwarzania przez sieć neuronową. W niektórych przypadkach odpowiednio zrealizowane przetwarzanie danych pozwala wręcz na uzyskanie wyższej jakości interpretacji danych korzystając z mniej złożonego modelu niż w przypadku podania nieprzetworzonych danych na wejście głębokiej sieci neuronowej [P7].

W tej części pracy zostało przedstawione zastosowanie wstępnego przetwarzania wektorów uczących w celu poprawy procesu wykrywania wybranych chorób onkologicznych i kardiologicznych. Dalsza część pracy została podzielona na dwie sekcje. W sekcji pierwszej zaproponowano system ASI do wspomaganie wykrywania Ostrej Białaczki Limfoblastycznej (ALL – *Acute Lymphoblastic Leukemia*) na podstawie mikroskopowych zdjęć krwi. W sekcji drugiej zaproponowano model systemu do wykrywania i klasyfikacji wybranych chorób kardiologicznych na podstawie sygnałów EKG.

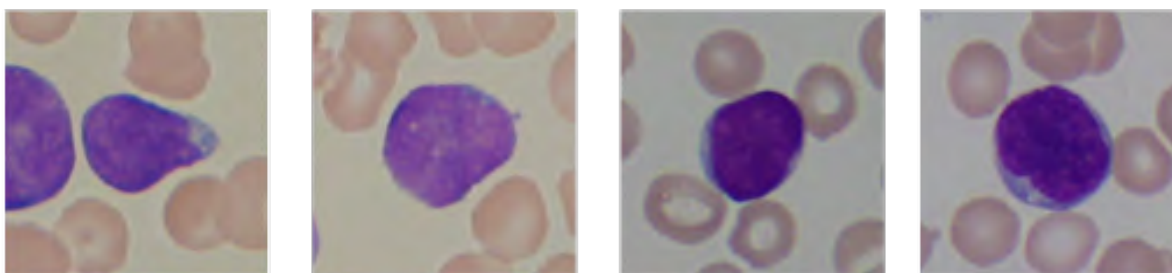
5.1 Wykrywanie Ostrej Białaczki Limfoblastycznej

Ostra białaczka limfoblastyczna (ALL), jest terminem opisującym grupę nowotworów układu limfatycznego. Zachorowania dominują u dzieci wieku od 2 do 5 lat [53]. Diagnozowanie ALL obejmuje przede wszystkim analizę zdjęć mikroskopowych krwi i szpiku kostnego [54].

W celu realizacji badań wykorzystano zdjęcia limfocytów osób zdrowych oraz pacjentów z ALL [55]. Zdjęcia pochodzą ze zbioru ALL-IDB, pobranego za zgodą właścicieli. Baza danych ALL-IDB ma dwie wersje: ALL-IDB1 i ALL-IDB2. W tym cyklu badań przeprowadzono eksperymenty na obrazach z ALL-IDB2. ALL-IDB2 to zestaw wyselekcjonowanych mikroskopowych rejestracji wizualnej krwi pobranych od osób zdrowych i pacjentów z ALL. Zbiór ALL-IDB2 przechowuje 260 zdjęć, na których znajduje się tylko jedna komórka. Rysunek 12 przedstawia przykładowe zdjęcia z bazy ALL-IDB2 prezentujące limfocyty osób zdrowych. Przykładowe limfocyty osób chorych są przedstawione na rysunku 13.



Rys. 22 - prezentacja przykładowych zdjęć mikroskopowych przedstawiających limfocyty osób zdrowych [P6].



Rys. 13 - prezentacja przykładowych zdjęć mikroskopowych przedstawiających limfocyty osób chorych na ALL [P6].

Analizowane są zdjęcia mikroskopowe limfocytów. Każde zdjęcie jest nieposiadającym struktury, trójwymiarowym tensorem posiadającym setki tysięcy bajtów. Tak duża liczba danych utrudnia opracowanie interpretowalnego algorytmu klasyfikującego. Ponadto, przy tak dużej ilości danych zwiększona jest podatność algorytmów ASI na przeuczenie. Z tego powodu wstępne przetwarzanie tensorów uczących jest wskazane w celu poprawienia jakości klasyfikacji. Dodatkowo, prawidłowe zrealizowanie wstępnego przetwarzania danych wejściowych umożliwi uproszczenie systemu, co skutkuje większą odpornością na przeuczenie i poprawioną interpretowalnością zachodzących procesów.

Na początku zostały zrealizowane badania mające na celu określenie jakości wyników generowanych przez analizę surowych danych. W tym celu wykorzystano głębokie konwolucyjne sieci neuronowe (z ang. *Convolutional Neural Networks, CNN*). Najlepszy wynik pracy CNN posłużył za bazę porównawczą do oceny jakości zastosowanych technik przetwarzania danych. Wstępne przetworzenie danych wejściowych ma za zadanie albo poprawić jakość klasyfikacji, albo wielokrotnie zmniejszyć liczbę parametrów modelu koniecznych do osiągnięcia satysfakcjonującego wyniku. Im model ma mniej parametrów, tym bardziej jest odporny na przeuczenie i tym łatwiej można zrozumieć, jak on działa.

W badaniu zaprezentowanym w [P6] wykorzystano niżej opisane sieci neuronowe.

- Prosta sieć CNN o homogenicznej strukturze.
- Sieć MobileNet V2 [56]. Sieć ta ma więcej parametrów od sieci CNN. Dodatkowo, posiada ona specjalnie utworzone techniki w swej strukturze mające na celu poprawienie jakości klasyfikacji. Do tych technik należą: sploty z separacją głębi, liniowe wąskie gardła oraz odwrócone połączenia rezydualne. Sieć MobileNet V2 została udostępniona przez jej autorów i jest szeroko wykorzystywana w pracach naukowych i przemysłowych.

- Sieć MobileNet V2 wstępnie wytrenowana na zbiorze ImageNet [57]. Ten wariant sieci posiada parametry zoptymalizowane pod rozwiązanie zadania będącego częścią konkursu ImageNet. W tym zadaniu sieć realizowała 1000-klasową klasyfikację obiektów występujących w życiu codziennym.

Wyniki sieci neuronowych zostały przedstawione w tabeli 11. Najlepszym rozwiązaniem okazała się sieć MobileNet v2, która uzyskała dokładność klasyfikacji na poziomie 92.8%.

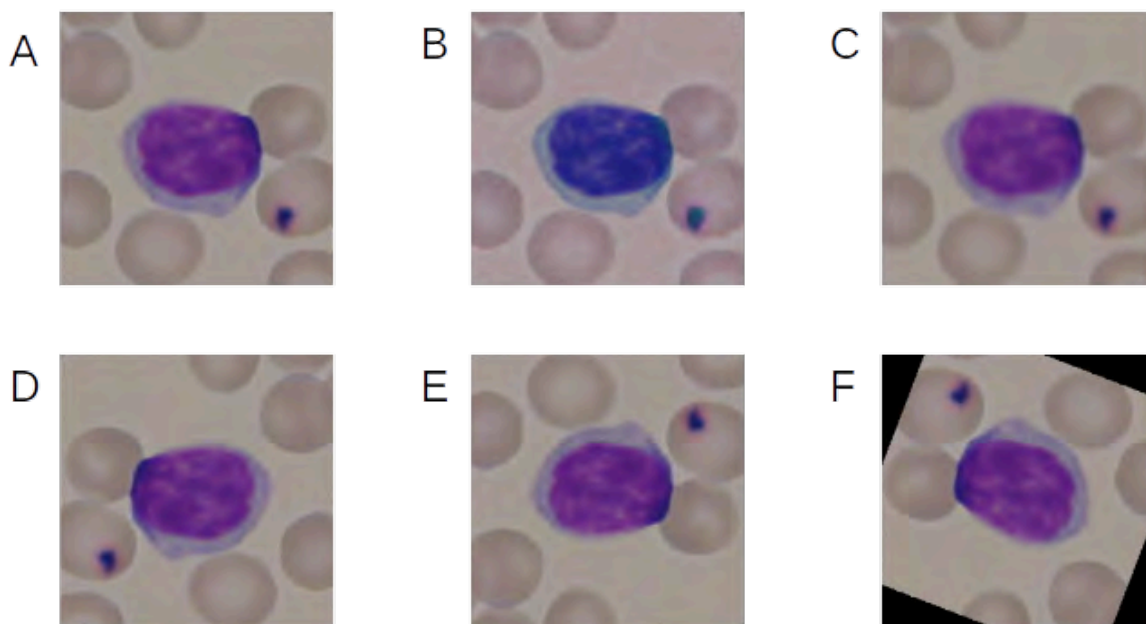
Tabela 11 – porównanie wyników sieci neuronowych przetwarzających surowe dane. Przedstawione wartości zostały uśrednione z prób eksperymentu.

| Nazwa | Przetrenowanie | Dokładność |
|--------------|----------------|------------|
| MobileNet v2 | Tak | 92.8% |
| CNN | Nie | 90.0% |
| MobileNet v2 | Nie | 80.1% |

W celu uzyskania wartości bazowej dla przetwarzania surowych danych postanowiono jeszcze zrealizować *data augmentation*. Procedura ta umożliwi sztuczne zwiększenie liczby przykładów treningowych poprzez wprowadzanie losowych modyfikacji wykorzystując techniki przetwarzania obrazów. W ten sposób można tworzyć różne wersje tego samego obrazu dostarczając modelowi ASI więcej wzorców uczących. Zastosowano następujące techniki *data augmentation*:

- fluktuacja koloru,
- rozmycie gaussowskie,
- odwrócenie w poziomie,
- odwrócenie w pionie,
- obrót pozycji.

Na rysunku 14 przedstawiono przykładowe zastosowanie tych technik. Ze względu na ważność koloru w obrazach biomedycznych wyszczególniono dwa tryby zastosowania *data augmentation*: z fluktuacją koloru i bez niej. Tabela 12 przedstawia wyniki najlepszej architektury sieci neuronowej (MobileNet V2 przetrenowane na zbiorze *ImageNet*) wraz z zastosowaniem *data augmentation*.



Rys. 14 - przykładowe zastosowanie technik *data augmentation*. A. brak modyfikacji, B. fluktuacja koloru, C. rozmycie gaussowskie, D. obrócenie w poziomie, E. odwrócenie w pionie, F. obrót o losowy kąt [P6].

Tabela 12 - przedstawienie wyników sieci neuronowej po zastosowaniu *data augmentation*. Dane zawarte w tabeli pochodzą z [P6]. Wyniki modeli zostały uśrednione z prób eksperymentu.

| Nazwa | <i>Data Augmentation</i> | Dokładność |
|-------------------|-----------------------------|------------|
| MobileNet v2 [P6] | Tak, bez modyfikacji koloru | 94.8% |
| MobileNet v2 [P6] | Tak | 93.8% |
| MobileNet v2 [P6] | Nie | 92.8% |

Wraz z tym badaniem uzyskano bazę porównawczą dla badania efektywności klasyfikacji na podstawie surowych danych wejściowych. Najlepszy wynik osiągnęła sieć MobileNet v2 wstępnie przetrenowana na zbiorze *ImageNet*, dla której dane poddano *data augmentation* bez modyfikacji koloru. Dokładność takiej konfiguracji to 94.8%.

W następnym etapie zastosowano wstępne przetwarzanie wektorów uczących w celu nadania danym wejściowym układu wektorowego. Do klasyfikacji surowych obrazów wykorzystano głębokie sieci neuronowe, ponieważ jest to jeden z niewielu algorytmów będących w stanie przetworzyć nieustrukturyzowane dane (w tym przypadku w postaci obrazu). Zastosowanie wstępnego przetwarzania obrazów w celu ich ustrukturyzowania pozwala dodatkowo na wykorzystanie algorytmów uczenia maszynowego. Algorytmy uczenia maszynowego z reguły mają mniej parametrów od głębokich sieci neuronowych, co zwiększa ich odporność na przetrenowanie i interpretowalność.

W poprzednim kroku badawczym w celu poprawienia dokładności sieci neuronowych wykorzystano w czasie treningu dwa tryby *data augmentation*: uwzględniający fluktuację koloru oraz bez niej. Sieć neuronowa wykorzystująca *data augmentation* bez fluktuacji koloru

uzyskała dokładność o 1 punkt procentowy wyższy od sieci trenowanej ze wszystkimi technikami *data augmentation*. Na tej podstawie uznano, że kolor zawiera informacje pozwalające na klasyfikację występowania ALL. W przeciwnym wypadku wprowadzenie losowych zmian kolorystycznych nie pogorszyłoby efektów uczenia.

Posiłkując się tym wnioskiem postanowiono zakodować obrazy w sposób, który zarówno dokona ich ustrukturyzowania, jak i wzmocni znaczenie wartości koloru. W tym celu postanowiono reprezentować obrazy za pomocą histogramu wartości koloru obrazu. Takie kodowanie ma następujące zalety:

- Wyraża informacje zawarte w dystrybucji wartości kolorów.
- Przedstawia dane w postaci, w której pozycja wartości w wektorze niesie ze sobą informację. Przykładowo, komórka pierwsza wektora zawiera częstość występowania wartości koloru z zakresu pomiędzy 0 a 5. Oznacza to, że wartość 10 na pozycji pierwszej informuje o 10-krotnym zaobserwowaniu wartości koloru z zakresu od 0 do 5. Cecha ta pozwala na wykorzystanie algorytmów uczenia maszynowego takich jak xgboost.
- Redukuje wielkość danych wejściowych. Obraz w zbiorze ALL-IDB2 ma wymiar 257 x 257 pikseli. Oznacza to, że każdy obraz jest trójwymiarowym tensorem posiadającym 198147 wartości. Histogram posiada zaledwie 51 wartości. W efekcie zaproponowane kodowanie redukuje rozmiar danych wejściowych 3800 krotnie.

Każdy histogram został obliczony na podstawie wartości z tylko jednego kanału obrazu. Obraz w formacie RGB posiada trzy kanały: czerwony, zielony i niebieski. Każdy z nich był reprezentowany za pomocą osobnego histogramu. Dodatkowo, wykorzystano też kodowanie HSV do przedstawienia obrazów w postaci bardziej przypominającej ludzkie pojmowanie kolorów [P7].

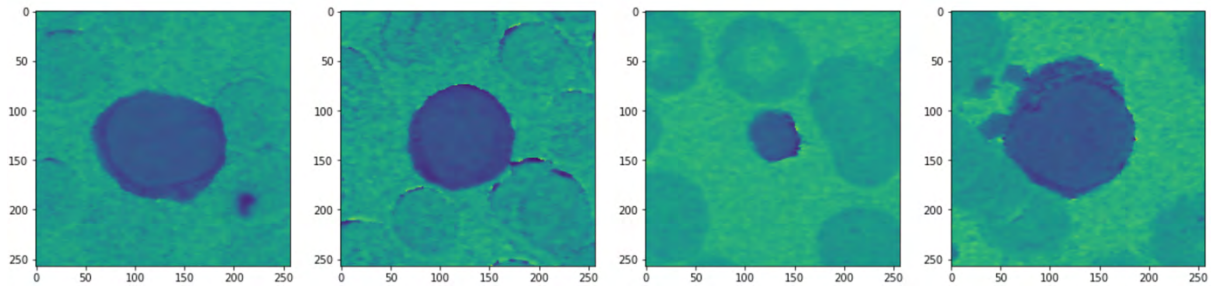
Wektory były interpretowane przez algorytm xgboost. Model ten posiadał 100 drzew, każde mające 64 liście. Oznacza to, że łącznie model posiadał 6400 parametrów. Dla porównania, sieć neuronowa MobileNet v2 wykorzystuje 3.4 miliona parametrów, ponad 500 razy więcej. W toku eksperymentów wykazano, że kanał zielony i kanał odcieni (z ang. *hue*) kodowania HSV uzyskały wysoką dokładność klasyfikacji. Wyniki zostały przedstawione w tabeli 13 wraz z najlepszą siecią neuronową dla porównania.

Tabela 13 – porównanie systemów ASI wykorzystujących wstępne przetwarzanie wektorów w porównaniu do interpretacji surowych danych. Dane zawarte w tabeli pochodzą z [P6][P7]. Wyniki modeli zostały uśrednione z prób eksperymentu.

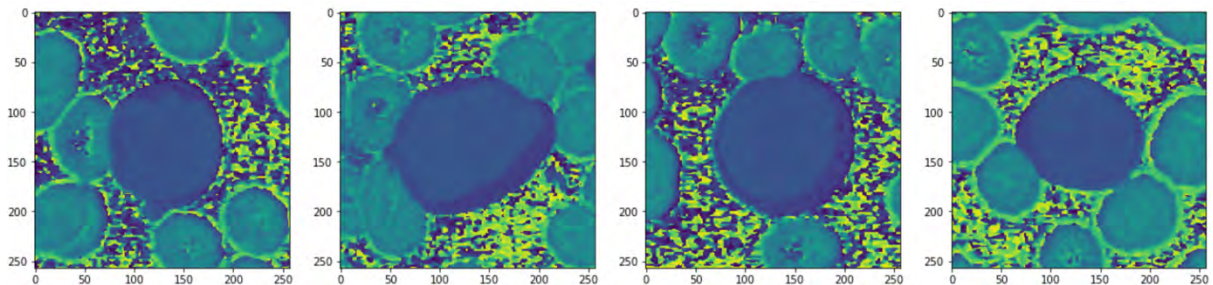
| Model | Kanał | Dokładność | Liczba parametrów |
|-------------------|-------|------------|-------------------|
| Xgboost [P7] | G | 96.0% | 6400 |
| MobileNet v2 [P6] | RGB | 94.8% | 3400000 |
| Xgboost [P7] | H | 94.0% | 6400 |

Algorytm xgboost interpretujący histogramy koloru zielonego uzyskał o 1.2 punktu procentowego lepszy wynik od głębokiej sieci neuronowej. Algorytm ten uzyskał lepszą jakość przetwarzania mając jednocześnie ponad 500 razy mniej parametrów i interpretując dane wejściowe o około 3800 razy mniejszym rozmiarze. Wprowadzenie wstępnego przetwarzania wektorów uczących umożliwiło zarówno zwiększenie dokładności systemu ASI, jak i znacznie zmniejszyło złożoność obliczeniową realizowanych działań.

W toku realizowanych badań przeprowadzono analizę eksploracyjną danych wejściowych w celu wyeksponowania cech. Powodem tej analizy były przesłanki wskazujące na zależność pomiędzy wartościami odcieni (z ang. *hue*) otoczenia limfocytów, a występowaniem ALL. Przesłanka ta została zwizualizowana za pomocą rysunków 15 i 16. Obydwa rysunki przedstawiają kanał odcieni obrazów 4 różnych limfocytów. Rysunek 15 przedstawia limfocyty osób zdrowych. Rysunek 16 prezentuje limfocyty osób chorych.

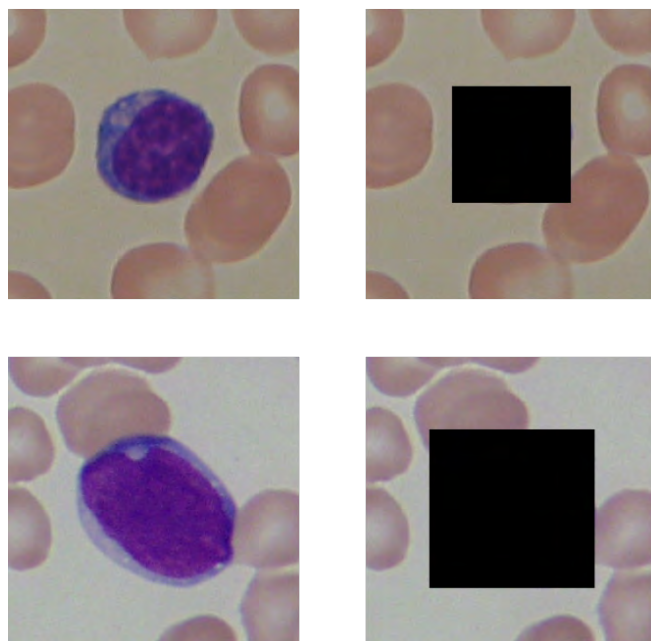


Rys. 15 - wizualizacja kanału odcieni limfocytów osób zdrowych [P7].



Rys. 16 - wizualizacja kanału odcieni limfocytów osób chorych [P7].

W celu zbadania czy różnice w otoczeniu limfocytów u osób chorych i zdrowych przedstawionych na rysunkach 15 i 16 występują w całym zbiorze danych postanowiono usunąć informacje na temat limfocytu, a następnie wytrenować system ASI do rozpoznawania ALL na podstawie danych kodujących wyłącznie otoczenie limfocytu. Informacje dotyczące limfocytu zostały usunięte poprzez zasłonięcie go czarnym prostokątem. Tak zmodyfikowany obraz został wykorzystany w celu obliczenia histogramu wartości kolorów. Proces ten został przedstawiony na rysunku 17. Piksele koloru czarnego wynikające z zasłonięcia limfocytu czarnym prostokątem nie zostały uwzględnione w histogramie.



Rys. 17 – wizualizacja procesu usuwania z obrazu informacji dotyczącej limfocytów [P7]. Zdjęcie u góry przedstawia komórkę osoby zdrowej, a na dole pacjenta chorującego na ALL.

Ostatecznie wytrenowano system ASI w celu rozpoznawania ALL na podstawie histogramów otoczenia limfocytów. Interpretując histogramy odcieni otoczenia limfocytów uzyskano dokładność na poziomie 93%. Jest to wynik niższy o zaledwie jeden punkt procentowy od systemu korzystającego z informacji o limfocycie i otoczeniu. Na tej podstawie można wysnuć wniosek, że otoczenie limfocytów zawiera informacje umożliwiające klasyfikację ALL. Jest to nowe spostrzeżenie, niespotykane wcześniej w literaturze medycznej [P7]. Wyniki uzyskane w toku tego badania przedstawiono w tabeli 14.

Tabela 14 - zestawienie badań przeprowadzonych w celu wykrycia ALL. Dane zawarte w tabeli pochodzą z [P6], [P7]. Wyniki modeli zostały uśrednione z prób eksperymentu.

| Model | Kanał | Limfocyty | Dokładność | Liczba parametrów |
|-------------------|-------|-----------|------------|-------------------|
| Xgboost [P7] | G | Tak | 96.0% | 6400 |
| MobileNet v2 [P6] | RGB | Tak | 94.8% | 3400000 |
| Xgboost [P7] | H | Tak | 94.0% | 6400 |
| Xgboost [P7] | H | Nie | 93.0% | 6400 |

Zastosowanie wstępnego przetwarzania wektorów uczących pozwoliło na wprowadzenie zmian w systemie ASI, których efektem jest zwiększenie dokładności przetwarzania o 1.2 punktu procentowego przy jednoczesnej redukcji parametrów modelu ponad 500 krotnie. Udowodniono w ten sposób tezę czwartą. Dodatkowo, ocena przydatności wstępnego przetwarzania wektorów pozwoliła na przeprowadzenie dodatkowych badań. Efektem tych badań było potwierdzenie, że otoczenie limfocytów także posiada informacje umożliwiające wykrywanie ALL. Jest to wniosek niespotykany w literaturze medycznej [P7].

Wprowadzenie wstępnego przetwarzania wektorów uczących zarówno poprawiło działanie systemu ASI, jak i umożliwiło lepsze zrozumienie zjawisk zachodzących w badanym zadaniu. Znalezione zależności dają nadzieję na zapoczątkowanie nowych badań medycznych mających na celu lepsze zrozumienie mechanizmów rządzących funkcjonowaniem ALL.

5.2 Wykrywanie wybranych chorób serca na podstawie analizy sygnałów EKG

Choroby układu krążenia pozostają główną przyczyną zgonów na całym świecie [58]. Wśród przyczyn chorób sercowo-naczyniowych jedną z najistotniejszych jest arytmia serca. W praktyce istnieje jednak wiele rodzajów nieregularnego bicia serca. Dokładna klasyfikacja różnych typów chorób serca może pomóc w diagnostyce i leczeniu [59].

W tej pracy przedstawiono propozycje pracy systemu do wykrywania wybranych chorób serca na podstawie sygnału EKG. Sygnały EKG zostały wybrane z publicznie dostępnej bazy danych PTB-XL [60]. Baza danych PTB-XL to jeden z największych publicznie dostępnych klinicznych zestawów danych EKG, który jest przystosowany do wykorzystania w celu oceny algorytmów uczenia maszynowego (ML). Zestaw danych EKG PTB-XL zawiera 21837 klinicznych, 12-kanalów EKG od 18885 pacjentów o długości 10 s, próbkowanych przy 500 Hz i 100 Hz z rozdzielczością 16 bitów. Baza danych PTB-XL posiada trzy stopnie dokładności etykiet (tabela 15):

- klasyfikacja binarna – określenie czy osoba cierpi na przypadłość kardiologiczną,
- klasyfikacja 5-klasowa – określenie czy osoba jest zdrowa, czy cierpi na jedną z czterech klas przypadłości kardiologicznych,
- klasyfikacja 20-klasowa – określenie czy osoba jest zdrowa, czy cierpi na jedną z 19 podklas przypadłości kardiologicznych.

Tabela 15 - Opis oraz liczność etykiet w zbiorze PTB-XL o granulacji 20 klas. Opisy zostały zachowane w oryginalnej angielskiej wersji [P8].

| Klasa | Podklasa | Liczba rekordów | Opis |
|-------|-----------|-----------------|---|
| NORM | NORM | 7185 | Prawidłowe EKG |
| CD | LAFB/LPFB | 881 | Blok przedniej wiązki lewej odnogi pęczka Hisa |
| | IRBBB | 798 | Blok prawej odnogi pęczka Hisa |
| | CLBBB | 527 | Kompletny blok lewej odnogi pęczka Hisa |
| | CRBBB | 385 | Kompletny blok prawej odnogi pęczka Hisa |
| | IVCD | 326 | Nieswoiste zaburzenia przewodzenia śródkomorowego |
| | _AVB | 204 | Blok przewodnictwa przedsionkowo-komorowego |
| | WPW | 67 | Zespół Wolffa–Parkinsona–White’a |
| | ILBBB | 44 | Niekompletny blok lewej odnogi pęczka Hisa |
| STTC | STTC | 1713 | Zmiany ST |
| | NST_ | 478 | Niespecyficzne zmiany ST |
| | ISCA | 429 | Niedokrwienie w odprowadzeniach przednich |
| | ISC_ | 297 | Niespecyficzne niedokrwienie |
| | ISCI | 147 | Niedokrwienne w dolnych odprowadzeniach |
| MI | AMI | 1636 | Zawał przedniej ściany serca |
| | IMI | 1272 | Zawał dolnej ściany serca |
| | LMI | 28 | Zawał serca boczny |
| HYP | LVH | 733 | Przerost lewej komory serca |
| | LAO/LAE | 49 | Przeciążenie/powiększenie lewego przedsionka |
| | RAO/RAE | 33 | Przeciążenie/powiększenie prawego przedsionka |

W toku tej pracy zrealizowano zadania binarnej klasyfikacji, pięcioklasowej klasyfikacji oraz dwudziestoklasowej klasyfikacji [P8][P9][P10]. Na początku zrealizowano zadanie poprzez wytrenowanie sztucznej sieci neuronowej CNN do przetwarzania surowych sygnałów EKG. W ten sposób uzyskano bazę porównawczą do oceny przydatności zaproponowanych technik wstępnego przetwarzania wektorów uczących.

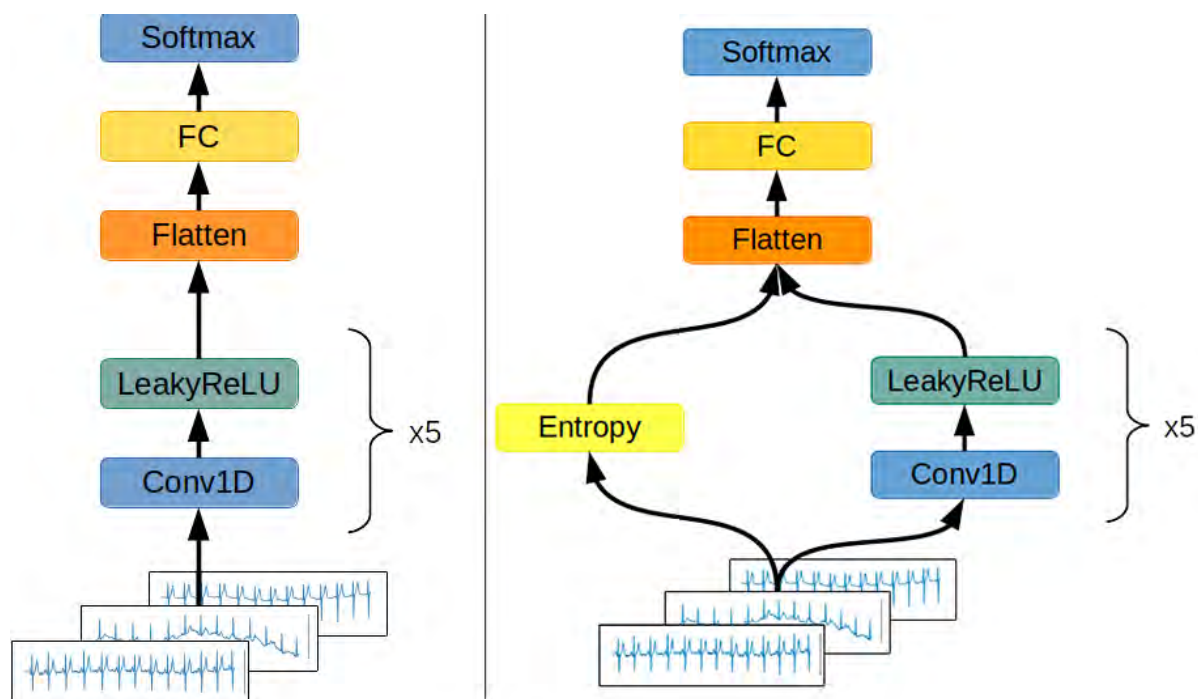
Wyniki otrzymane przez sieć zostały przedstawione w tabeli 16. Rezultaty zostały podzielone względem liczby klas w zadaniu klasyfikacji. Uproszczony schemat sieci neuronowej został zaprezentowany na rysunku 7.

Tabela 16 - zestawienie dokładności przetwarzania sieci CNN w zależności od liczby klas. Dane zawarte w tabeli pochodzą z [P8]. Wyniki modeli zostały uśrednione z prób eksperymentu.

| Model | Liczba klas | Dokładność |
|---------------|-------------|------------|
| Sieć CNN [P8] | 2 | 90.0% |
| Sieć CNN [P8] | 5 | 75.3% |
| Sieć CNN [P8] | 20 | 61.3% |

W celu poprawy dokładności klasyfikacji realizowanej na 20 klasach postanowiono wykorzystać wstępne przetwarzanie wektorów w celu realizacji wielomodalnej sieci neuronowej. Wielomodalna sieć neuronowa jest siecią hybrydową potrafiącą przetwarzać kilka różnych rodzajów danych. Nowo zaprojektowana sieć neuronowa przyjmuje zarówno surowy sygnał EKG, jak również wektor cech wyekstrahowany z sygnału EKG.

W pracy [P8] postanowiono utworzyć wektor cech poprzez obliczenie różnych miar entropii. Entropie zostały wybrane ze względu na ich umiejętność opisanie zależności pomiędzy częstotliwością występowania stanów sygnału. Cecha ta jest trudna do odwzorowania przez sieci neuronowe, które opierają się na operacjach dodawania, mnożenia i rektyfikacji liniowej [61]. Zdefiniowany wektor składający się z następujących entropii sygnału: Shannona [62], aproksymacji [63], próbkowania [63], permutacji [64], widmowej [65], SVD, Rényi'ego [66], Tsallisa [67] oraz Extropy [68]. Rysunek 18 przedstawia wzbogacenie architektury sieci CNN w celu uwzględnienia wprowadzonego wektora cech.



Rys. 18 – wykorzystanie wstępnego przetwarzania wektorów uczących w celu zaprojektowania hybrydowej (dwumodalnej) sieci neuronowej [P8], [P9]. Po lewej przedstawiono sieć CNN interpretującą surowy sygnał EKG. Po prawej przedstawiono dwumodalną sieć neuronową interpretującą zarówno surowy sygnał EKG, jak również miary entropii wyekstrahowane z sygnału EKG.

Ekstrakcję wektora cech entropii zrealizowano w dwóch trybach: z surowego sygnału EKG oraz z sygnału podzielonego na zespoły QRS. Wyniki przedstawiono w tabeli 17.

Tabela 17 – zestawienie wyników zastosowania wstępnego przetwarzania wektorów uczących w celu zaprojektowania hybrydowej (wielomodalnej) sieci neuronowej. Kolumna “Entropie EKG” oznacza, czy sieć wykorzystuje wektor cech obliczony z surowego sygnału EKG. Kolumna “Entropie QRS” oznacza, czy sieć wykorzystuje wektor cech obliczony z sygnału podzielonego na zespoły QRS [P9]. Wyniki modeli zostały uśrednione z prób eksperymentu.

| Model | Entropie EKG | Entropie QRS | Klasy | Dokładność |
|-------------------------|--------------|--------------|-------|------------|
| 2-modalna sieć CNN [P9] | Tak | Tak | 20 | 66.2% |
| 2-modalna sieć CNN [P9] | Nie | Tak | 20 | 65.1% |
| 2-modalna sieć CNN [P9] | Tak | Nie | 20 | 64.3% |
| Sieć CNN [P8] | Nie | Nie | 20 | 61.3% |

Zastosowanie miar entropii w celu implementacji wielomodalnej sieci neuronowej umożliwiło zwiększenie dokładności klasyfikacji sieci CNN o 4.9 punktu procentowego. Natomiast w celu poprawienia jakości klasyfikacji dla 2 i 5 klas zaprojektowano hybrydowe sieci neuronowe łączące możliwość sieci neuronowych do przetwarzania danych nieustrukturyzowanych (sygnału EKG) z algorytmami uczenia maszynowego. W tym celu zaprojektowano sieć CNN do interpretacji sygnału podzielonego na zespoły QRS w trybie *few-shot learning* (FSL). FSL polega na wytrenowaniu sieci neuronowej w taki sposób, że sieć służy za koder sygnałów EKG do postaci abstrakcyjnych wektorów cech. Tak zakodowane dane są ustrukturyzowane, przez co mogą być interpretowane przez algorytmy uczenia maszynowego. Do wytrenowania kodera FSL wykorzystano funkcję straty *triplet margin loss* (TML) przedstawionej za pomocą równania 24:

$$L(a,p,n) = \max(d(a,p) - d(a,n) + m, 0), \quad (24)$$

gdzie:

- a – wektor “kotwiczący” będący bazą porównawczą dla wektorów p i n ,
- p – wektor “pozytywny” należący do tej samej klasy, co wektor “kotwiczący”,
- n – wektor “negatywny” należący do innej klasy niż wektory a i p ,
- m – margines opisujący pożądane rozdzielanie wektorów tej samej klasy od wektorów innych klas,
- d – funkcja mierząca odległość pomiędzy wektorami (przykład – odległość euklidesowa).

Funkcja straty TML wymusza na sieci neuronowej kodowanie danych wejściowych w taki sposób, że wektory reprezentujące obiekty z tej samej klasy są mniej odległe niż wektory z innych klas. W ten sposób możliwe jest zrealizowanie klasyfikacji poprzez pomiar miary odległości między wektorem danych wejściowych a wektorami obiektów o znanych klasach.

Sieć wytrenowana za pomocą TML jest koderem danych. Jej zadaniem było wstępne przetworzenie wektorów uczących w celu ich ustrukturyzowania oraz ekstrakcji cech. Następnie użyto algorytm uczenia maszynowego do klasyfikacji danych na podstawie wektora kodującego dane wejściowe. W tym celu wykorzystano algorytm SVM z funkcją RBF jako

jądrem [69]. Tak zdefiniowano hybrydową sieć neuronową. Jej wyniki oraz porównanie z efektami przetwarzania surowych danych przedstawiono w tabeli 18.

Tabela 18 – prezentacja porównania wyników przetwarzania modeli hybrydowych z normalnymi modelami. Dane zawarte w tabeli pochodzą z [P10]. Wyniki modeli zostały uśrednione z prób eksperymentu.

| Model | Liczba klas | Dokładność |
|----------------------|-------------|------------|
| Sieć FSL + SVM [P10] | 2 | 91.3% |
| Sieć CNN [P8] | 2 | 90.0% |
| Sieć FSL + SVM [P10] | 5 | 79.0% |
| Sieć CNN [P8] | 5 | 75.3% |

Zastosowanie wstępnego przetwarzania wektorów uczących w celu zmodyfikowania systemu ASI zwiększyło dokładność klasyfikacji o 1.3 punktu procentowego dla 2 klas, 3.7 punktu procentowego dla 5 klas oraz 4.9 punktu procentowego dla 20 klas. Udowodniono w ten sposób tezę trzecią i piątą. Hybrydowe modele sieci CNN i algorytmy uczenia maszynowego SVM zwiększyły dokładność sieci neuronowych o 1.3 i 3.7 punktu procentowego. Model ten wykorzystał sieci CNN do wstępnego przetworzenia danych. W ten sposób uzyskano dane ustrukturyzowane, które zostały wykorzystane przez algorytm SVM w celu dokonania klasyfikacji. Tak sformułowany model hybrydowy osiągnął zamierzone rezultaty. W efekcie została potwierdzona teza piąta.

6. Podsumowanie i wnioski

Głównym celem badań niniejszej rozprawy doktorskiej było opracowanie odpowiednich modyfikacji algorytmów sztucznej inteligencji, które będą umożliwiały modelowanie komórek biologicznych oraz wspomagały proces wykrywania wybranych chorób onkologicznych i kardiologicznych.

Cel został zrealizowany, a wyniki zostały opublikowane w monotematycznym zbiorze publikacji naukowych P1-P10 oraz pięciu publicznie dostępnych repozytoriach kodu na platformie GitHub. Wszystkie postawione tezy zostały udowodnione.

Interpretacja równań Michaelisa-Menten jako prawdopodobieństwa przemiany substratów w produkty umożliwiło wykorzystanie Teorii Kolejek do modelowania cyklu Krebsa, szlaku pentozofosforanowego, beta-oksydacji kwasów tłuszczowych i odpowiedzi komórkowej na insulinę. Zastąpienie równań różniczkowych Teorią Kolejek spowodowało rozwiązanie problemów takich jak występowanie ujemnych wartości stężenia substancji. Wykorzystanie równań różniczkowych może prowadzić do uzyskania wartości ujemnych. Jest to błąd, ponieważ stężenia substancji w rzeczywistości nie mogą osiągnąć wartości poniżej zera. To z kolei stwarza problem pod postacią utraty stabilności procesu symulacji. Problem ten nie występuje w podejściu wykorzystującym Teorię Kolejek, przez co technika ta pozwala na uzyskanie stabilniejszych rezultatów symulacji. Udowodniono w ten sposób tezę pierwszą – możliwe jest zastąpienie równań różniczkowych Teorią Kolejek.

Wykorzystanie grupowania genów w chromosomie ze względu na ich przynależność do parametryzowanego równania Michaelisa-Menten pozwoliło na zmodyfikowanie algorytmu genetycznego. Wprowadzono w ten sposób mechanizm oceniania prawdopodobieństwa zajścia reakcji chemicznych bez symulowania eksperymentu. Dzięki temu zarówno przyspieszono poszukiwanie wartości optymalnych parametryzujących symulację, jak również uproszczono funkcję straty oceniającą chromosomy. Korzystając z tej modyfikacji możliwe było przeniesienie z funkcji straty do mechanizmu reprodukcji wymogu uzyskania określonej wariancji w szeregach czasowych obrazujących stężenia substancji. Dzięki temu funkcja straty określa wynik dopasowania chromosomu korzystając z mniejszej liczby warunków. To upraszcza ją czyniąc funkcję straty łatwiejszą do minimalizacji jej wyniku. Tym samym teza druga została udowodniona – wstępne przetwarzanie wektorów parametryzujących symulację pozwala na wykorzystanie sztucznej inteligencji do modelowania, w oparciu o Teorię Kolejek a następnie symulację komputerową, komórek biologicznych.

System wspomagający proces wykrywania ALL został poprawiony dzięki prawidłowo zrealizowanemu wstępnemu przetwarzaniu wektorów uczących. Uzyskano w ten sposób model, który był o 1.2 punktu procentowego dokładniejszy od sieci neuronowej przetwarzającej surowe dane. Ponadto, model ten miał 500 razy mniej parametrów i podejmował decyzje na podstawie tensora danych wejściowych ponad 3000 razy mniejszego. Uzyskano w ten sposób dokładniejszy, szybszy i prostszy model. Potwierdzona została teza czwarta – wstępne przetwarzanie wektorów uczących pozwala na użycie modeli wykorzystujących do trzech rzędów mniejszej liczby parametrów bez utraty dokładności wykonywanego zadania rozpoznawania wybranych chorób onkologicznych i kardiologicznych.

Na podstawie badań nad tym modelem stwierdzono też, że otoczenie limfocytów zawiera informację pozwalającą na wykrycie ALL z dokładnością dochodzącą do 93%. Do tej pory proces wykrywania ALL skupiał się na interpretowaniu limfocytów, tak więc wiedza o związku

pomiędzy otoczeniem limfocytów a występowaniem ALL może posłużyć do lepszej diagnozy tej ciężkiej choroby. Najlepszy opracowany model do wykrywania ALL miał dokładność diagnozy na poziomie 96%.

Proces wykrywania wybranych chorób kardiologicznych może być bardziej dokładny dzięki zastosowaniu wstępnego przetwarzania wektorów uczących. Na ich podstawie opracowano i zaimplementowano hybrydowe sieci neuronowe łączące możliwości interpretacyjne różnych algorytmów ASI oraz hybrydowe sieci neuronowe łączące przetwarzanie różnych typów danych (sieci wielomodalne). Udowodniono w ten sposób tezę piątą – wykorzystanie hybrydowych metod uczenia maszynowego pozwala na poprawę procesu klasyfikacji. Uzyskano w ten sposób zwiększenie dokładności procesu diagnozy o 1.2 punktu procentowego 2-klasowej klasyfikacji, 3.7 punktu procentowego 5-klasowej klasyfikacji oraz 4.9 punktu procentowego 20-klasowej klasyfikacji. Ostatecznie najlepsze opracowane modele do wykrywania wybranych chorób kardiologicznych dla klasyfikacji 2 klas, 5 klas i 20 klas miały odpowiednio 91.3%, 79.0% i 66.2% dokładności. Są to wyniki lepsze niż przy użyciu sieci neuronowych interpretujących surowy sygnał EKG, które uzyskały odpowiednio 90.0%, 75.3% i 61.3%. Udowodniono w ten sposób tezę trzecią – wstępne przetwarzanie wektorów uczących w procesie klasyfikacji pozwala na wydobycie z nich ważnych informacji poprawiających dokładność procesu rozpoznawania wybranych chorób onkologicznych i kardiologicznych.

Na potrzeby tej pracy zrealizowano badania nad szerokim zakresem tematem. Pomimo dużej liczby opublikowanych prac w tej tematyce jest jeszcze wiele do poprawy, zważywszy na pojawiające się coraz większe możliwości bardzo szybkiego przetwarzania wielu różnych danych z wykorzystaniem nowych i zmodyfikowanych ASI.

Wszystkie przedstawione w ramach niniejszej rozprawy prace badawcze mają na celu opracowanie systemu wspomagającego pracę lekarzy specjalistów w wykrywaniu i następnie leczeniu różnych chorób cywilizacyjnych.

7. Literatura

- [P1] S. Kloska, K. Pałczyński, T. Marciniak, T. Talaśka, M. Nitz, B. Wysocki, P. Davis i T. Wysocki, „Queueing theory model of Krebs cycle,” *Bioinformatics*, 2021.
- [P2] S. M. Kloska, K. Pałczyński, T. Marciniak, T. Talaśka, M. Miller, B. J. Wysocki, P. Davis i T. A. Wysocki, „Queueing theory model of pentose phosphate pathway,” *Scientific reports*, tom 12, p. 4601, 2022.
- [P3] S. M. Kloska, K. Pałczyński, T. Marciniak, T. Talaśka, M. Miller, B. J. Wysocki, P. Davis i T. A. Wysocki, „Conversion of fat to cellular fuel—Fatty acids beta-oxidation model,” *Computational Biology and Chemistry*, tom 104, p. 107860, 2023.
- [P4] S. M. Kloska, K. Pałczyński, T. Marciniak, T. Talaśka, M. Miller, B. J. Wysocki, P. H. Davis, G. A. Soliman i T. A. Wysocki, „Queueing theory model of mTOR complexes’ impact on Akt-mediated adipocytes response to insulin,” *PLoS One*, tom 17, 2022.
- [P5] S. M. Kloska, K. Pałczyński, T. Marciniak, T. Talaśka, B. J. Wysocki, P. Davis i T. A. Wysocki, „Integrating glycolysis, citric acid cycle, pentose phosphate pathway, and fatty acid beta-oxidation into a single computational model,” *Scientific Reports*, tom 13, p. 14484, 2023.
- [P6] K. Pałczyński, S. Śmigiel, M. Gackowska, D. Ledziński, S. Bujnowski i Z. Lutowski, „IoT application of transfer learning in hybrid artificial intelligence systems for acute lymphoblastic leukemia classification,” *Sensors*, tom 21, p. 8025, 2021.
- [P7] K. Pałczyński, D. Ledziński i T. Andrysiak, „Entropy Measurements for Leukocytes’ Surrounding Informativeness Evaluation for Acute Lymphoblastic Leukemia Classification,” MDPI, 2022.
- [P8] S. Śmigiel, K. Pałczyński i D. Ledziński, „ECG signal classification using deep learning techniques based on the PTB-XL dataset,” *Entropy*, tom 23, p. 1121, 2021.
- [P9] S. Śmigiel, K. Pałczyński i D. Ledziński, „Deep learning techniques in the classification of ECG signals using R-peak detection based on the PTB-XL dataset,” *Sensors*, tom 21, p. 8174, 2021.
- [P10] K. Pałczyński, S. Śmigiel, D. Ledziński i S. Bujnowski, „Study of the few-shot learning for ECG classification based on the PTB-XL dataset,” MDPI, 2022.
- [11] OECD, „Availability of doctors,” 2022. [Online]. Available: <https://www.oecd-ilibrary.org/sites/d0057e82-en/index.html?itemId=/content/component/d0057e82-en>. [Data uzyskania dostępu: 24 04 2024].
- [12] L. Jiang, Z. Wu, X. Xu, Y. Zhan, X. Jin, L. Wang i Y. Qiu, „Opportunities and challenges of artificial intelligence in the medical field: current application, emerging problems, and problem-solving strategies,” *Journal of International Medical Research*, tom 49, 2021.

- [13] OECD, „AI in Health - Huge Potential, Huge Risks,” OECD, 2024.
- [14] S. Arslan, E. Ozyurek i C. Gunduz-Demir, „A color and shape based algorithm for segmentation of white blood cells in peripheral blood and bone marrow images,” Wiley Online Library, 2014.
- [15] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. Ioannidis, G. S. Collins i M. Maruthappu, „Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies,” *bmj*, tom 368, 2020.
- [16] M. Kiener, „Artificial intelligence in medicine and the disclosure of risks,” *AI & society*, tom 36, pp. 705-713, 2021.
- [17] R. D. Labati, V. Piuri i F. Scotti, „All-IDB: The acute lymphoblastic leukemia image database for image processing,” w *2011 18th IEEE international conference on image processing*, IEEE, 2011, pp. 2045-2048.
- [18] P. Wagner, N. Strodthoff, R.-D. Boussejot, D. Kreiseler, F. I. Lunze, W. Samek i T. Schaeffter, „PTB-XL, a large publicly available electrocardiography dataset,” *Scientific data*, tom 7, pp. 1-15, 2020.
- [19] D. L. Nelson, A. L. Lehninger i M. M. Cox, *Lehninger principles of biochemistry*, Macmillan, 2008.
- [20] M. Mescam, K. C. Vinnakota i D. A. Beard, „Identification of the catalytic mechanism and estimation of kinetic parameters for fumarase,” *Journal of Biological Chemistry*, tom 286, pp. 21100--21109, 2011.
- [21] L. K. Harold, A. Jinich, K. Hards, A. Cordeiro, L. M. Keighley, A. Cross, M. B. McNeil, K. Rhee i G. M. Cook, „Deciphering functional redundancy and energetics of malate oxidation in mycobacteria,” *Journal of Biological Chemistry*, tom 298, 2022.
- [22] S. J. Mihalik, B. H. Goodpaster, D. E. Kelley, D. H. Chace, J. Vockley, F. G. Toledo i J. P. DeLany, „Increased levels of plasma acylcarnitines in obesity and type 2 diabetes and identification of a marker of glucolipotoxicity,” *Obesity*, tom 18, pp. 1695-1700, 2010.
- [23] G. C. Henderson, „Plasma free fatty acid concentration as a modifiable risk factor for metabolic disease,” *Nutrients*, tom 13, p. 2590, 2021.
- [24] V. S. Rao i K. Srinivas, „Modern drug discovery process: An in silico approach,” *Journal of bioinformatics and sequence analysis*, tom 2, pp. 89-84, 2011.
- [25] V. K. Singh i I. Ghosh, „Kinetic modeling of tricarboxylic acid cycle and glyoxylate bypass in *Mycobacterium tuberculosis*, and its application to assessment of drug targets,” *Theoretical Biology and Medical Modelling*, tom 3, pp. 1-11, 2006.

- [26] E. Ahn, P. Kumar, D. Mukha, A. Tzur i T. Shlomi, „Temporal fluxomics reveals oscillations in TCA cycle flux throughout the mammalian cell cycle,” *Molecular systems biology*, tom 13, p. 953, 2017.
- [27] D. M. Cohen i R. N. Bergman, „Estimation of TCA cycle flux, aminotransferase flux, and anaplerosis in heart: validation with syntactic model,” *American Journal of Physiology-Endocrinology and Metabolism*, tom 268, pp. E397--E409, 1995.
- [28] M. Ederer, S. Steinsiek, S. Stagge, M. D. Rolfe, A. T. Beek, D. Knies, M. J. T. d. Mattos, T. Sauter, J. Green i R. K. Poole, „A mathematical model of metabolism and regulation provides a systems-level view of how *Escherichia coli* responds to oxygen,” *Frontiers in microbiology*, tom 5, p. 124, 2014.
- [29] C. J. Foster, S. Gopalakrishnan, M. R. Antoniewicz i C. D. Maranas, „From *Escherichia coli* mutant ¹³C labeling data to a core kinetic model: A kinetic model parameterization pipeline,” *PLoS computational biology*, tom 15, 2019.
- [30] N. Jahan, K. Maeda, Y. Matsuoka, Y. Sugimoto i H. Kurata, „Development of an accurate kinetic model for the central carbon metabolism of *Escherichia coli*,” *Microbial cell factories*, tom 15, pp. 1-19, 2016.
- [31] L. F. Shampine, S. Thompson, J. Kierzenka i G. Byrne, „Non-negative solutions of ODEs,” *Applied mathematics and computation*, tom 170, pp. 556-569, 2005.
- [32] W. A. Massey, „Asymptotic analysis of the time dependent M/M/1 queue,” *Mathematics of Operations Research*, tom 10, pp. 305-327, 1985.
- [33] X.-S. Yang, „Chapter 6 - Genetic Algorithms,” w *Nature-Inspired Optimization Algorithms (Second Edition)*, Academic Press, 2021, pp. 91-100.
- [34] K. Korla, L. Vadlakonda i C. K. Mitra, „Kinetic simulation of malate-aspartate and citrate-pyruvate shuttles in association with Krebs cycle,” *Journal of Biomolecular Structure and Dynamics*, tom 33, pp. 2390-2403, 2015.
- [35] M. Ponizovskiy, „Role of Krebs cycle in mechanism of stability internal medium and internal energy in an organism in norm and in mechanism of cancer pathology,” *Mod. Chem. Appl*, tom 4, 2016.
- [36] A. C. Smith i A. J. Robinson, „A metabolic model of the mitochondrion and its use in modelling diseases of the tricarboxylic acid cycle,” *BMC systems biology*, tom 5, pp. 1-13, BMC systems biology.
- [37] J. O. Park, S. A. Rubin, Y.-F. Xu, D. Amador-Noguez, J. Fan, T. Shlomi i J. D. Rabinowitz, „Metabolite concentrations, fluxes and free energies imply efficient enzyme usage,” *Nature chemical biology*, tom 12, pp. 482-489, 2016.
- [38] E. J. Clement, T. T. Schulze, G. A. Soliman, B. J. Wysocki, P. H. Davis i T. A. Wysocki, „Stochastic simulation of cellular metabolism,” *IEEE Access*, tom 8, pp. 79734-79744, 2020.

- [39] R. Milo, P. Jorgensen, U. Moran, G. Weber i M. Springer, „BioNumbers—the database of key numbers in molecular and cell biology,” *Nucleic acids research*, tom 38, pp. D750-D753, Nucleic acids research.
- [40] B. D. Bennett, E. H. Kimball, M. Gao, R. Osterhout, S. J. V. Dien i J. D. Rabinowitz, „Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*,” *Nature chemical biology*, tom 5, pp. 593-599, 2009.
- [41] E. Mogilevskaya, O. Demin i I. Goryanin, „Kinetic model of mitochondrial Krebs cycle: unraveling the mechanism of salicylate hepatotoxic effects,” *Journal of Biological Physics*, tom 32, pp. 245-271, 2006.
- [42] K. R. Albe, M. H. Butler i B. E. Wright, „Cellular concentrations of enzymes and their substrates,” *Journal of theoretical biology*, tom 143, pp. 163-195, 1990.
- [43] A. Janzer, N. J. German, K. N. Gonzalez-Herrera, J. M. Asara, M. C. Haigis i K. Struhl, „Metformin and phenformin deplete tricarboxylic acid cycle and glycolytic intermediates during cell transformation and NTPs in cancer stem cells,” *Proceedings of the National Academy of Sciences*, tom 111, pp. 10574-10579, 2014.
- [44] V. P. Sukhatme i B. Chan, „Glycolytic cancer cells lacking 6-phosphogluconate dehydrogenase metabolize glucose to induce senescence,” *FEBS letters*, tom 586, pp. 2389-2395, 2012.
- [45] S. M. Houten i R. J. Wanders, „A general introduction to the biochemistry of mitochondrial fatty acid beta-oxidation,” *Journal of inherited metabolic disease*, tom 33, pp. 469-477, 2010.
- [46] K. v. Eunen, S. M. Simons, A. Gerding, A. Bleeker, G. d. Besten, C. M. Touw, S. M. Houten, B. K. Groen, K. Krab i D.-J. Reijngoud, „Biochemical competition makes fatty-acid β -oxidation vulnerable to substrate overload,” *PLoS computational biology*, tom 9, p. e1003186, 2013.
- [47] A. Szwed, E. Kim i E. Jacinto, „Regulation and metabolic functions of mTORC1 and mTORC2,” *Physiological Reviews*, tom 101, pp. 1371-1426, 2021.
- [48] Z. Mao i W. Zhang, „Role of mTOR in glucose and lipid metabolism,” *International journal of molecular sciences*, tom 19, p. 2043, 2018.
- [49] M. R. Rajan, E. Nyman, P. Kjolhede, G. Cedersund i P. Str{\aa}lfors, „Systems-wide experimental and modeling analysis of insulin signaling through forkhead box protein O1 (FOXO1) in human adipocytes, normally and in type 2 diabetes,” *Journal of Biological Chemistry*, tom 291, pp. 15806-15819, 2016.
- [50] A. Veilleux, V. P. Houde, K. Bellmann i A. Marette, „Chronic inhibition of the mTORC1/S6K1 pathway increases insulin-induced PI3K activity but inhibits Akt2 and glucose transport stimulation in 3T3-L1 adipocytes,” *Molecular endocrinology*, tom 24, pp. 766-778, 2010.
- [51] S. J. Rigatti, „Random forest,” *Journal of Insurance Medicine*, tom 47, pp. 31-39, 2017.

- [52] T. Chen i C. Guestrin, „Xgboost: A scalable tree boosting system,” w *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794.
- [53] M. Onciu, „Acute lymphoblastic leukemia,” *Hematology/oncology clinics of North America*, tom 23, pp. 655-674, 2009.
- [54] S. Mohapatra, D. Patra i S. Satpathi, „Image analysis of blood microscopic images for acute leukemia detection,” w *2010 international conference on industrial electronics, control and robotics*, IEEE, 2010, pp. 215-219.
- [55] R. D. Labati, V. Piuri i F. Scotti, „All-IDB: The acute lymphoblastic leukemia image database for image processing,” w *2011 18th IEEE international conference on image processing*, IEEE, 2011, pp. 2045-2048.
- [56] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov i L.-C. Chen, „Mobilenetv2: Inverted residuals and linear bottlenecks,” w *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510-4520.
- [57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li i L. Fei-Fei, „Imagenet: A large-scale hierarchical image database,” w *2009 IEEE conference on computer vision and pattern recognition*, IEEE, 2009, pp. 248-255.
- [58] E. J. Benjamin, S. S. Virani, C. W. Callaway, A. M. Chamberlain, A. R. Chang, S. Cheng, S. E. Chiuve, M. Cushman, F. N. Delling i R. Deo, „Heart disease and stroke statistics—2018 update: a report from the American Heart Association,” *Circulation*, tom 137, pp. e67-e492, 2018.
- [59] G. S. WHO, „Global status report on noncommunicable diseases 2010,” Geneva (Switzerland) WHO, 2014.
- [60] P. Wagner, N. Strodthoff, R.-D. Boussejot, D. Kreiseler, F. I. Lunze, W. Samek i T. Schaeffter, „PTB-XL, a large publicly available electrocardiography dataset,” *Scientific data*, tom 7, pp. 1-15, 2020.
- [61] A. F. Agarap, „Deep learning using rectified linear units (relu),” *arXiv preprint arXiv:1803.08375*, 2018.
- [62] C. E. Shannon, „A mathematical theory of communication,” *The Bell system technical journal*, tom 27, pp. 379-423, 1948.
- [63] J. S. Richman i J. R. Moorman, „Physiological time-series analysis using approximate entropy and sample entropy,” *American journal of physiology-heart and circulatory physiology*, tom 278, pp. H2039-H2049, 2000.
- [64] C. Bandt i B. Pompe, „Permutation entropy: a natural complexity measure for time series,” *Physical review letters*, tom 88, p. 174102, 2002.
- [65] T. Inouye, K. Shinosaki, H. Sakamoto, S. Toi, S. Ukai, A. Iyama, Y. Katsuda i M. Hirano, „Quantification of EEG irregularity by use of the entropy of the power

- spectrum,” *Electroencephalography and clinical neurophysiology*, tom 79, pp. 204-210, 1991.
- [66] A. Renyi, „On measures of entropy and information,” w *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, University of California Press, 1961, pp. 547-562.
- [67] A. Bezerianos, S. Tong i N. Thakor, „Time-dependent entropy estimation of EEG rhythm changes following brain ischemia,” *Annals of biomedical engineering*, tom 31, pp. 221-232, 2003.
- [68] F. Lad, G. Sanfilippo i G. Agro, „Extropy: Complementary dual of entropy,” 2015.
- [69] A. Patle i D. S. Chouhan, „SVM kernel functions for classification,” w *2013 International conference on advances in technology and engineering (ICATE)*, IEEE, 2013, pp. 1-19.
- [70] E. P. R. Service, „Artificial intelligence in healthcare. Applications, risks and ethical and societal impacts.,” 2022.
- [71] E. Ahn, P. Kumar, D. Mukha, A. Tzur i T. Shlomi, „Temporal fluxomics reveals oscillations in TCA cycle flux throughout the mammalian cell cycle,” *Molecular systems biology*, tom 13, p. 953, 2017.
- [72] J. M. Evans, L. A. Donnelly, A. M. Emslie-Smith, D. R. Alessi i A. D. Morris, „Metformin and reduced risk of cancer in diabetic patients,” *Bmj*, tom 330, pp. 1304-1305, 2005.
- [73] S. Jiralerspong, S. L. Palla, S. H. Giordano, F. Meric-Bernstam, C. Liedtke, C. M. Barnett, L. Hsu, M.-C. Hung, G. N. Hortobagyi i A. M. Gonzalez-Angulo, „Metformin and pathologic complete responses to neoadjuvant chemotherapy in diabetic patients with breast cancer,” *Journal of clinical oncology*, tom 27, p. 3297, 2009.
- [74] H. J. Kim, S. Lee, K. H. Chun, J. Y. Jeon, S. J. Han, D. J. Kim, Y. S. Kim, J.-T. Woo, M.-S. Nam i S. H. Baik, „Metformin reduces the risk of cancer in patients with type 2 diabetes: An analysis based on the Korean National Diabetes Program Cohort,” *Medicine*, tom 97, p. e0036, 2018.
- [75] M.-S. Yoon, „The role of mammalian target of rapamycin (mTOR) in insulin signaling,” *Nutrients*, tom 9, p. 1176, 2017.
- [76] F. Tremblay, A. Gagnon, A. Veilleux, A. Sorisky i A. Marette, „Activation of the mammalian target of rapamycin pathway acutely inhibits insulin signaling to Akt and glucose transport in 3T3-L1 and human adipocytes,” *Endocrinology*, tom 146, pp. 1328-1337, 2005.
- [77] F. Tremblay, A. Gagnon, A. Veilleux, A. Sorisky i A. Marette, „Activation of the mammalian target of rapamycin pathway acutely inhibits insulin signaling to Akt and glucose transport in 3T3-L1 and human adipocytes,” *Endocrinology*, tom 146, pp. 1328-1337, 2005.

8. STRESZCZENIE

Badania algorytmów sztucznej inteligencji i ich odpowiednich modyfikacji w procesie modelowania komórek biologicznych oraz wykrywania wybranych chorób onkologicznych i kardiologicznych

mgr inż. Krzysztof Pałczyński

Słowa kluczowe: Hybrydowe Sieci Neuronowe, Teoria Kolejek, Szlaki Metaboliczne, Ostra Białaczka Limfoblastyczna, Choroby Kardiologiczne

Sztuczna inteligencja jest obecnie jedną z najdynamiczniej rozwijanych gałęzi nauki. Z tego powodu pokładane są w niej duże nadzieje. Od algorytmów sztucznej inteligencji (ASI) oczekuje się między innymi bardzo dużej dokładności, a także szybkości działania. Zalety te są szczególnie uwidocznione w zastosowaniach medycznych, np. w procesie wykrywania różnych groźnych chorób.

Niniejsza dysertacja została przedstawiona w formie monotematycznego cyklu publikacji naukowych. Pierwsza grupa publikacji (P1-P5) dotyczy wykorzystania algorytmów sztucznej inteligencji, a także ich odpowiednich modyfikacji do modelowania komórek biologicznych, druga zaś (P6-P10) ich wykorzystania w procesie wykrywania chorób onkologicznych (na przykładzie Ostrej Białaczki Limfoblastycznej (ALL) i wybranych chorób kardiologicznych.

Głównym celem badań było opracowanie odpowiednich modyfikacji algorytmów sztucznej inteligencji, które będą umożliwiały modelowanie komórek biologicznych oraz wspomagały proces wykrywania wybranych chorób onkologicznych i kardiologicznych.

Modelowanie komórek biologicznych zostało zrealizowane poprzez modelowanie osobno szlaków metabolicznych, by na koniec scalić je w jeden duży model. Celem modelowania tych szlaków było umożliwienie symulowania życia komórki biologicznej. W ten sposób możliwe było realizowanie eksperymentów *in silico* w celu projektowania nowych sposobów leczenia groźnych chorób, takich jak nowotwory, cukrzyca, otyłość, itp. W skład zamodelowanych szlaków metabolicznych wchodzi cykl Krebsa [P1], szlak pentozofosforanowy [P2], beta-oksydacja kwasów tłuszczowych [P3] oraz odpowiedź komórkowa na insulinę [P4]. Szlaki te zostały wybrane ze względu na ich rolę w produkcji energii w komórce biologicznej. Dodatkowo, modele cyklu Krebsa, szlaku pentozofosforanowego oraz beta-oksydacji kwasów tłuszczowych zostały połączone w jeden model komórkowy [P5]. Do ich zamodelowania wykorzystano równania Michaelisa-Menten oraz Teorię Kolejek. Opracowanie nowatorskiej metody wykrywania ALL [P6, P7] oraz wybranych chorób kardiologicznych [P8, P9, P10] miało na celu realizację modeli będących w stanie wspierać pracę lekarzy. W tym celu wykorzystano głębokie sieci neuronowe oraz wstępne przetwarzanie wektorów uczących. Tak zrealizowane przetwarzanie miało na celu poprawienie jakości klasyfikacji oraz zmniejszenie liczby parametrów modelu.

Cel pracy został zrealizowany a tezy udowodnione, co można zobaczyć między innymi w niniejszej rozprawie (rozdziału 4-6), a także analizując monotematyczny cykl (P1-P10) publikacji naukowych.

9. ABSTRACT

Research on artificial intelligence algorithms and their appropriate modifications in the process of modelling biological cells and detecting selected oncological and cardiological diseases

M.Eng, Krzysztof Palczyński

Key words: Hybrid Neural Networks, Queuing Theory, Metabolic Pathways, Acute Lymphoblastic Leukemia, Cardiac Diseases

Artificial intelligence is currently one of the most dynamically developed branches of science. For this reason, high hopes are placed on it. Among other things, very high accuracy and speed are expected from artificial intelligence algorithms. These advantages are particularly visible in medical applications, e.g. in the process of detecting various dangerous diseases.

This dissertation is presented in the form of a monothematic series of scientific publications. The first group of publications (P1-P5) presents the usage of artificial intelligence algorithms, as well as their appropriate modifications in the modelling of biological cells, while the second (P6-P10) focuses on their application in the process of detecting oncological diseases (on the example of Acute Lymphoblastic Leukemia (ALL) and selected heart diseases.

The main goal of the research was to develop appropriate modifications of artificial intelligence algorithms that will enable modelling of biological cells and support the process of detecting selected oncological and cardiological diseases.

The modelling of biological cells was achieved by modelling metabolic pathways separately and finally merging them into one large model. The purpose of modelling these pathways was to enable simulation of the life of a biological cell. In this way, it was possible to carry out *in silico* experiments to design new methods of treating dangerous diseases such as cancer, diabetes, obesity, etc. The modelled metabolic pathways include the Krebs cycle [P1], the pentose phosphate pathway [P2], fatty acids beta-acid oxidation [P3] and cellular response to insulin [P4]. These pathways were selected because of their role in energy production in the biological cell. Additionally, models of the Krebs cycle, pentose phosphate pathway and fatty acid beta-oxidation were combined into one cellular model [P5]. The Michaelis-Menten equations and Queuing Theory were used to model them. The development of an innovative method for detecting ALL [P6, P7] and selected cardiac diseases [P8, P9, P10] was aimed at implementing models capable of supporting the work of doctors. For this purpose, deep neural networks and pre-processing of training vectors were used. The processing carried out in this way was aimed at improving the quality of classification and reducing the number of model parameters.

The aim of the work was achieved, and the theses were proven, which can be seen, among others, in this dissertation (chapters 4-6), as well as by analysing the monothematic series (P1-P10) of scientific publications.

10. Oświadczenie Autora rozprawy doktorskiej

Z.16.2021.2022

Załącznik nr 3 do
Instrukcji drukowania, gromadzenia, rejestrowania
i udostępniania rozpraw doktorskich przez rady naukowe
dyscyplin (dyscyplin artystycznych) prowadzących
postępowanie w sprawie nadania stopnia naukowego doktora

Oświadczenie Współautora

Mgr inż. Krzysztof Pałczyński

.....
(tytuł zawodowy, imiona i nazwisko współautora)

Szkoła Doktorska Politechnika Bydgoska

.....
(miejsce pracy/afiliacja)

OŚWIADCZENIE

Oświadczam, iż mój wkład autorski w niżej ~~wymienionym~~/wymienionych ~~artykule~~/artykułach ~~naukowym~~/naukowych był następujący*:

1. Śmigiel, S., Pałczyński, K., & Ledziński, D. (2021). ECG signal classification using deep learning techniques based on the PTB-XL dataset. *Entropy*, 23(9), 1121. DOI: <https://doi.org/10.3390/e23091121> pkt. MNiSW: 100, Impact Factor: 2.738
Wykonane zadania w ramach artykułu:
 - a) koncepcja i implementacja cech sygnału EKG opartych na entropii,
 - b) koncepcja i implementacja zastosowanych modeli sztucznych sieci neuronowych,
 - c) trening sztucznych sieci neuronowych i zebranie wyników,
 - d) opracowanie formuł matematycznych,
 - e) walidacja badań i analiza wyników,
 - f) aktywne uczestnictwo w opracowaniu kolejnych wersji artykułu,
 - g) czynny udział w procesie korekty artykułu po recenzjach.
2. Śmigiel, S., Pałczyński, K., & Ledziński, D. (2021). Deep learning techniques in the classification of ECG signals using R-peak detection based on the PTB-XL dataset. *Sensors*, 21(24), 8174. DOI: <https://doi.org/10.3390/s21248174> pkt. MNiSW: 100, Impact Factor: 3.847
Wykonane zadania w ramach artykułu:
 - a) koncepcja i implementacja cech sygnału EKG opartych na entropii,
 - b) koncepcja i implementacja zastosowanych modeli sztucznych sieci neuronowych,
 - c) trening sztucznych sieci neuronowych i zebranie wyników,

* W przypadku prac dwu- lub wieloautorskich wymagane są oświadczenia kandydata do stopnia doktora oraz współautorów, wskazujące na ich merytoryczny wkład w powstanie każdej pracy (np. twórca hipotezy badawczej, pomysłodawca badań, wykonanie specyficznych badań – np. przeprowadzenie konkretnych doświadczeń, opracowanie i zebranie ankiet itp., wykonanie analizy wyników, przygotowanie manuskryptu artykułu i inne). Określenie wkładu danego autora, w tym kandydata do stopnia doktora, powinno być na tyle precyzyjne, aby umożliwić dokładną ocenę jego udziału i roli w powstaniu każdej pracy.

KP

- d) opracowanie formuł matematycznych,
 - e) walidacja badań i analiza wyników,
 - f) aktywne uczestnictwo w opracowaniu kolejnych wersji artykułu,
 - g) czynny udział w procesie korekty artykułu po recenzjach.
3. Pałczyński, K., Śmigiel, S., Ledziński, D., & Bujnowski, S. (2022). Study of the few-shot learning for ECG classification based on the PTB-XL dataset. *Sensors*, 22(3), 904. DOI: <https://doi.org/10.3390/s22030904> pkt. MNiSW: 100, Impact Factor: 3.847
Wykonane zadania w ramach artykułu:
- a) pomysł na zastosowanie techniki few-shot learning do analizy sygnału EKG,
 - b) koncepcja i implementacja zastosowanych modeli sztucznych sieci neuronowych,
 - c) trening sztucznych sieci neuronowych i zebranie wyników,
 - d) opracowanie formuł matematycznych,
 - e) walidacja badań i analiza wyników,
 - f) aktywne uczestnictwo w opracowaniu kolejnych wersji artykułu,
 - g) czynny udział w procesie korekty artykułu po recenzjach,
 - h) pierwszy autor.
4. Pałczyński, K., Śmigiel, S., Gackowska, M., Ledziński, D., Bujnowski, S., & Lutowski, Z. (2021). IoT application of transfer learning in hybrid artificial intelligence systems for acute lymphoblastic leukemia classification. *Sensors*, 21(23), 8025. DOI: <https://doi.org/10.3390/s21238025> pkt. MNiSW: 100, Impact Factor: 3.847
Wykonane zadania w ramach artykułu:
- a) pomysłodawca zastosowanych metod,
 - b) opracowanie metodologii badawczej,
 - c) koncepcja i implementacja zastosowanych modeli uczenia maszynowego,
 - d) trening uczenia maszynowego i zebranie wyników,
 - e) opracowanie formuł matematycznych,
 - f) walidacja badań i analiza wyników,
 - g) aktywne uczestnictwo w opracowaniu kolejnych wersji artykułu,
 - h) czynny udział w procesie korekty artykułu po recenzjach,
 - i) pierwszy autor.

kp

5. Pałczyński, K., Ledziński, D., & Andrysiak, T. (2022). Entropy Measurements for Leukocytes' Surrounding Informativeness Evaluation for Acute Lymphoblastic Leukemia Classification. *Entropy*, 24(11), 1560. DOI: <https://doi.org/10.3390/e24111560> pkt. MNiSW: 100, Impact Factor: 2.738
Wykonane zadania w ramach artykułu:
- pomysłodawca zastosowanych metod,
 - analiza stanu wiedzy,
 - opracowanie metodologii badawczej,
 - opracowanie i implementacja wykorzystanego algorytmu,
 - opracowanie formuł matematycznych,
 - walidacja badań i analiza wyników,
 - aktywne uczestnictwo w opracowaniu kolejnych wersji artykułu,
 - czynny udział w procesie korekty artykułu po recenzjach,
 - pierwszy autor, autor korespondencyjny.
6. Kloska, S., Pałczyński, K., Marciniak, T., Talaśka, T., Nitz, M., Wysocki, B. J., ... & Wysocki, T. A. (2021). Queueing theory model of Krebs cycle. *Bioinformatics*, 37(18), 2912-2919. DOI: <https://doi.org/10.1093/bioinformatics/btab177> pkt. MNiSW: 200, Impact Factor: 6.9
Wykonane zadania w ramach artykułu:
- Pomysłodawca struktury algorytmu genetycznego wraz z modyfikacjami koniecznymi do wytrenowania modelu,
 - Opracowanie i implementacja metod wstępnego przetwarzania wektorów parametryzujących symulacje,
 - Opracowanie i implementacja modelu od strony informatycznej będącego przedmiotem badań,
 - Implementacja oprogramowania do realizacji eksperymentów,
 - Planowanie oraz realizacja eksperymentów,
 - Eksploracyjna analiza danych oraz analiza wyników modelu,
 - Ocena wyników modelu oraz wizualizacja,
 - Uczestnictwo w pisaniu publikacji

KP

7. Kłoska, S. M., Pałczyński, K., Marciniak, T., Talaśka, T., Miller, M., Wysocki, B. J., ... & Wysocki, T. A. (2022). Queueing theory model of pentose phosphate pathway. *Scientific reports*, 12(1), 4601. DOI: <https://doi.org/10.1038/s41598-022-08463-y> pkt. MNiSW: 140, Impact Factor: 4.9.

Wykonane zadania w ramach artykułu:

- a) Pomysłodawca struktury algorytmu genetycznego wraz z modyfikacjami koniecznymi do wytrenowania modelu,
- b) Opracowanie i implementacja metod wstępnego przetwarzania wektorów parametryzujących symulacje,
- c) Opracowanie i implementacja modelu od strony informatycznej będącego przedmiotem badań,
- d) Implementacja oprogramowania do realizacji eksperymentów,
- e) Planowanie oraz realizacja eksperymentów,
- f) Eksploracyjna analiza danych oraz analiza wyników modelu,
- g) Ocena wyników modelu oraz wizualizacja,
- h) Uczestnictwo w pisaniu publikacji

8. Kłoska, S. M., Pałczyński, K., Marciniak, T., Talaśka, T., Miller, M., Wysocki, B. J., ... & Wysocki, T. A. (2023). Conversion of fat to cellular fuel—Fatty acids β -oxidation model. *Computational Biology and Chemistry*, 104, 107860. DOI: <https://doi.org/10.1016/j.compbiolchem.2023.107860> pkt. MNiSW: 70, Impact Factor: 3.7

Wykonane zadania w ramach artykułu:

- a) Pomysłodawca struktury algorytmu genetycznego wraz z modyfikacjami koniecznymi do wytrenowania modelu,
- b) Opracowanie i implementacja metod wstępnego przetwarzania wektorów parametryzujących symulacje,
- c) Opracowanie i implementacja modelu od strony informatycznej będącego przedmiotem badań,
- d) Implementacja oprogramowania do realizacji eksperymentów,
- e) Planowanie oraz realizacja eksperymentów,
- f) Eksploracyjna analiza danych oraz analiza wyników modelu,
- g) Ocena wyników modelu oraz wizualizacja,

LP

- h) Uczestnictwo w pisaniu publikacji
9. Kłoska, S. M., Pałczyński, K., Marciniak, T., Talaśka, T., Miller, M., Wysocki, B. J., ... & Wysocki, T. A. (2022). Queueing theory model of mTOR complexes' impact on Akt-mediated adipocytes response to insulin. *PLoS One*, 17(12), e0279573. DOI: <https://doi.org/10.1371/journal.pone.0279573> pkt. MNiSW: 100, Impact Factor: 3.7
Wykonane zadania w ramach artykułu:
- a) Pomysłodawca struktury algorytmu genetycznego wraz z modyfikacjami koniecznymi do wytrenowania modelu,
 - b) Opracowanie i implementacja metod wstępnego przetwarzania wektorów parametryzujących symulacje,
 - c) Opracowanie i implementacja modelu od strony informatycznej będącego przedmiotem badań,
 - d) Implementacja oprogramowania do realizacji eksperymentów,
 - e) Planowanie oraz realizacja eksperymentów,
 - f) Eksploracyjna analiza danych oraz analiza wyników modelu,
 - g) Ocena wyników modelu oraz wizualizacja,
 - h) Uczestnictwo w pisaniu publikacji
10. Kłoska, S. M., Pałczyński, K., Marciniak, T., Talaśka, T., Wysocki, B. J., Davis, P., & Wysocki, T. A. (2023). Integrating glycolysis, citric acid cycle, pentose phosphate pathway, and fatty acid beta-oxidation into a single computational model. *Scientific Reports*, 13(1), 14484. DOI: <https://doi.org/10.1038/s41598-023-41765-3> pkt. MNiSW: 140, Impact Factor: 4.9
Wykonane zadania w ramach artykułu:
- a) Pomysłodawca struktury algorytmu genetycznego wraz z modyfikacjami koniecznymi do wytrenowania modelu,
 - b) Opracowanie i implementacja metod wstępnego przetwarzania wektorów parametryzujących symulacje,
 - c) Opracowanie i implementacja modelu od strony informatycznej będącego przedmiotem badań,
 - d) Implementacja oprogramowania do realizacji eksperymentów,
 - e) Planowanie oraz realizacja eksperymentów,
 - f) Eksploracyjna analiza danych oraz analiza wyników modelu,

kp

Z.16.2021.2022

Załącznik nr 3 do
Instrukcji drukowania, gromadzenia, rejestrowania
i udostępniania rozpraw doktorskich przez rady naukowe
dyscyplin (dyscyplin artystycznych) prowadzących
postępowanie w sprawie nadania stopnia naukowego doktora

- g) Ocena wyników modelu oraz wizualizacja,
- h) Uczestnictwo w pisaniu publikacji

29.05.2024
.....
miejsowość, data

Krzysztof Polakowski
.....
podpis Współautora

11. Oświadczenia współautorów artykułów naukowych

Prof. dr hab. inż. Beata J. Wysocki
Department of Biology,
University of Nebraska at Omaha, Omaha, NE 68182, USA

Oświadczenie

Oświadczam, że w artykule „Queueing theory model of Krebs cycle” opublikowanym w czasopiśmie *Bioinformatics* w 2021r.; <https://doi.org/10.1093/bioinformatics/btab177>, którego współautorami są Sylwester Kloska, Krzysztof Palczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata J. Wysocki, Paul Davis oraz Tadeusz A. Wysocki, mój wkład merytoryczny polegał na doradztwie w biochemicznej części badań.

Oświadczam, że w artykule „Queueing theory model of pentose phosphate pathway” opublikowanym w czasopiśmie *Scientific Reports* w 2022r.; <https://doi.org/10.1038/s41598-022-08463-y>, którego współautorami są Sylwester Kloska, Krzysztof Palczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata J. Wysocki, Paul Davis oraz Tadeusz A. Wysocki, mój wkład merytoryczny polegał na doradztwie w biochemicznej części badań.

Oświadczam, że w artykule „Conversion of fat to cellular fuel—Fatty acids beta-oxidation model” opublikowanym w czasopiśmie *Computational Biology and Chemistry* w roku 2023.; <https://doi.org/10.1016/j.cbcb.2023.107860>, którego współautorami są Sylwester Kloska, Krzysztof Palczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis oraz Tadeusz Wysocki, mój wkład merytoryczny polegał na doradztwie w biochemicznej części badań.

Oświadczam, że w artykule „Queueing theory model of mTOR complexes’ impact on Akt-mediated adipocytes response to insulin” opublikowanym w czasopiśmie *PLOS One* w 2022r.; <https://doi.org/10.1371/journal.pone.0279573>, którego współautorami są Sylwester Kloska, Krzysztof Palczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata J. Wysocki, Paul Davis, Ghada A. Soliman oraz Tadeusz A. Wysocki, mój wkład merytoryczny polegał na doradztwie w biochemicznej części badań.

Oświadczam, że w artykule „Integrating glycolysis, citric acid cycle, pentose phosphate pathway, and fatty acid beta-oxidation into a single computational model” opublikowanym w czasopiśmie *Scientific Reports* w 2023r.; <https://doi.org/10.1038/s41598-023-47165-3>, którego współautorami są Sylwester Kloska, Krzysztof Palczyński, Tomasz Marciniak, Tomasz Talaśka, Beata J. Wysocki, Paul Davis oraz Tadeusz Wysocki, mój wkład merytoryczny polegał na doradztwie w biochemicznej części badań.

Jednocześnie wyrażam zgodę na przedłożenie w/w publikacji przez mgr inż. Krzysztofa Palczyńskiego jako część rozprawy doktorskiej w formie spójnego tematycznie zbioru artykułów naukowych opublikowanych w czasopismach naukowych. Oświadczam, że udostępnienie utworów nie będzie naruszało praw autorskich osób trzecich.



(Podpis)

Oświadczenie Współautora

Dr inż. Damian Ledziński
(tytuł zawodowy, imiona i nazwisko współautora)

Wydział Telekomunikacji, Informatyki i Elektrotechniki
Politechnika Bydgoska im. Jana i Jędrzeja Śniadeckich w Bydgoszczy
(miejsce pracy/afiliacja)

OŚWIADCZENIE

Oświadczam, iż mój wkład autorski w niżej wymienionym/wymienionych artykule/artykułach naukowym/naukowych był następujący*:

1. Śmigiel, S., Pałczyński, K., & Ledziński, D. (2021). ECG signal classification using deep learning techniques based on the PTB-XL dataset. *Entropy*, 23(9), 1121. DOI: <https://doi.org/10.3390/e23091121> pkt. MNISW: 100, Impact Factor: 2.738.
Wykonane zadania w ramach artykułu:
 - a) przygotowanie środowiska badawczego w kontekście sprzętowo-software'owym,
 - b) wstępne przetwarzanie danych polegające na konwersji surowych danych do postaci wymaganej przez sieci neuronowej, obejmujące dekodowanie i processing wstępny danych,
 - c) walidacja badań i analiza wyników,
 - d) aktywne uczestnictwo w opracowaniu kolejnych wersji artykułu,
 - e) czynny udział w procesie korekty artykułu po recenzjach.
2. Śmigiel, S., Pałczyński, K., & Ledziński, D. (2021). Deep learning techniques in the classification of ECG signals using R-peak detection based on the PTB-XL dataset. *Sensors*, 21(24), 8174. DOI: <https://doi.org/10.3390/s21248174> pkt. MNISW: 100, Impact Factor: 3.847.
Wykonane zadania w ramach artykułu:
 - a) przygotowanie środowiska badawczego w kontekście sprzętowo-software'owym,
 - b) wstępne przetwarzanie danych polegające na konwersji surowych danych do postaci wymaganej przez sieci neuronowej, obejmujące dekodowanie i processing wstępny danych,
 - c) walidacja badań i analiza wyników,
 - d) aktywne uczestnictwo w opracowaniu kolejnych wersji artykułu,
 - e) czynny udział w procesie korekty artykułu po recenzjach.
3. Pałczyński, K., Śmigiel, S., Ledziński, D., & Bujnowski, S. (2022). Study of the few-shot learning for ECG classification based on the PTB-XL dataset. *Sensors*, 22(3), 904. DOI: <https://doi.org/10.3390/s22030904> pkt. MNISW: 100, Impact Factor: 3.847.
Wykonane zadania w ramach artykułu:
 - a) przygotowanie środowiska badawczego w kontekście sprzętowo-software'owym,
 - b) walidacja badań i analiza wyników,

* W przypadku prac own- lub wieloautorских wymaga się oświadczenia kandydata do stopnia doktora oraz współautorów, wskazującego na ich rzeczywisty wkład w powstanie każdej pracy (np. sformułowanie hipotezy badawczej, pomysłodawca badań, wykonanie specyficznych badań – np. przeprowadzenie konkretnych doświadczeń, opracowanie i zebranie ankiet itp., wykonanie analizy wyników, przygotowanie manuskryptu artykułu i inne). Określenie wkładu danego autora, w tym kandydata do stopnia doktora, powinno być na tyle precyzyjne, aby umożliwić dokładną ocenę jego udziału i roli w powstaniu każdej pracy.

- c) aktywne uczestnictwo w opracowaniu kolejnych wersji artykułu,
 - d) czynny udział w procesie korekty artykułu po recenzjach.
4. Pałczyński, K., Śmigiel, S., Gackowska, M., Ledziński, D., Bujnowski, S., & Lutowski, Z. (2021). IoT application of transfer learning in hybrid artificial intelligence systems for acute lymphoblastic leukemia classification. *Sensors*, 21(23), 8025. DOI: <https://doi.org/10.3390/s21238025> pkt. MNISW: 100, Impact Factor: 3,847.
Wykonane zadania w ramach artykułu:
- a) pomysłodawca idei wykorzystania bazy ALL-IDB w kontekście zastosowania metod uczenia maszynowego,
 - b) przygotowanie środowiska badawczego w kontekście sprzętowo-software'owym,
 - c) dostarczenie i analiza zbioru danych,
 - d) walidacja badań i analiza wyników,
 - e) udział w procesie korekty artykułu po recenzjach.
5. Pałczyński, K., Ledziński, D., & Andrysiak, T. (2022). Entropy Measurements for Leukocytes' Surrounding Informativeness Evaluation for Acute Lymphoblastic Leukemia Classification. *Entropy*, 24(11), 1560. DOI: <https://doi.org/10.3390/e24111560> pkt. MNISW: 100, Impact Factor: 2,738.
Wykonane zadania w ramach artykułu:
- a) przygotowanie środowiska badawczego w kontekście sprzętowo-software'owym,
 - b) dostarczenie i analiza zbioru danych,
 - c) walidacja badań i analiza wyników.

Jednocześnie wyrażam zgodę na przedłożenie wyżej wymienionej/wymienionych pracy/prac przez mgr. inż. Krzysztofa Pałczyńskiego (podać tytuł zawodowy imię i nazwisko kandydata do stopnia doktora) jako część rozprawy doktorskiej opartej na zbiorze opublikowanych i powiązanych tematycznie artykułów naukowych.

Bydgoszcz 28.5.2021
miejscowość; data

Dariusz Lewon
podać: Wzrost/autor

Prof. Ghada A. Soliman

Department of Environmental, Occupational, and Geospatial Health Sciences,
City University of New York, Graduate School of Public Health and Healthy Policy,
New York, NY, United States of America

Statement

I hereby declare that in the paper "Outwiring theory model of mTOR complexes' impact on Akt-mediated adipocytes response to insulin" published in the journal *PLoS One* in 2023; <https://doi.org/10.1371/journal.pone.0279273>, whose co-authors are Sylwester Kloska, Krzysztof Palczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis, Ghada A. Soliman and Tadeusz A. Wysocki, my substantive contribution was to advise on the biochemical part of the study.

At the same time, I consent to the submission of the above-mentioned publication by M. Eng. Krzysztof Palczyński as part of his doctoral dissertation in the form of a thematically coherent set of scientific articles published in scientific journals.



(Signature)

Oświadczenie Współautora

Mgr inż. Marta Gackowska

.....
(tytuł zawodowy, imiona i nazwisko współautora)

Wydział Telekomunikacji, Informatyki i Elektrotechniki

.....
(miejsce pracy/afiliacja)

OŚWIADCZENIE

Oświadczam, iż mój wkład autorski w niżej wymienionym/wymienionych artykule/artykulech naukowym/naukowych był następujący*:

1. Pałczyński, K., Śmigieł, S., Gackowska, M., Ledziński, D., Bujnowski, S., & Lutowski, Z. (2021). IoT application of transfer learning in hybrid artificial intelligence systems for acute lymphoblastic leukemia classification. *Sensors*, 21(23), 8025. DOI: <https://doi.org/10.3390/s21238025>
pkt. MNiSW: 100, Impact Factor: 3.847

Wykonane zadania w ramach artykułu:

- a) analiza stanu wiedzy w zakresie: opisu cech ostrej białaczki limfoblastycznej i potencjału zastosowania modeli uczenia maszynowego, wyników badań innych autorów i określenie luk w wiedzy, które stanowią podstawę do dalszych badań,
- b) walidacja badań i analiza wyników,
- c) aktywne uczestnictwo w opracowaniu kolejnych wersji artykułu,
- d) czynny udział w procesie korekty artykułu po recenzjach.

Jednocześnie wyrażam zgodę na przedłożenie wyżej wymienionej/wymienionych pracy/prac przez mgr. inż. Krzysztofa Pałczyńskiego (podać tytuł zawodowy imię i nazwisko kandydata do stopnia doktora) jako część rozprawy doktorskiej opartej na zbiorze opublikowanych i powiązanych tematycznie artykułów naukowych.

Bydgoszcz, 28.05.24
.....
miejsowość, data

.....
Gackowska Marta
podpis Współautora

* W przypadku prac stworzonych w ramach projektu wymagane są odwołania kandydata do stopnia doktora oraz wypełnienie instrukcji dotyczącej ich merytorycznej oceny w ramach każdej pracy (np. teoria i metody badawcze, porównanie badań, wykazanie specyficznych badań np. przeprowadzenie konkretnych doświadczeń, opracowanie i obrona tezy) oraz: wykonanie analizy wyników, przygotowanie wniosków, artykułu i tzw. Określenie wkładu danego autora, w tym kandydata do stopnia doktora, powołanie listy na jego rzecz, aby umożliwić dokłądną ocenę jego wkładu i roli w powstaniu każdej pracy.

Dr n. med. Sylwester Michał Kloska

Wydział Medyczny

Politechnika Bydgoska im. J i J. Śniadeckich w Bydgoszczy

Oświadczam, że w artykule „Queueing theory model of Krebs cycle” opublikowanym w czasopiśmie *Bioinformatics* w 2021r.; <https://doi.org/10.1093/bioinformatics/btab177>, którego współautorami są Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata J. Wysocki, Paul Davis oraz Tadeusz A. Wysocki, mój wkład w badania zgłoszone w niniejszej pracy polegał na konceptualizacji i realizacji badań. Przeprowadziłem dokładną analizę stanu techniki, zaprojektowałem strukturę algorytmu. Zebrałem niezbędne do badań i przygotowania modelu parametry i dane liczbowe dotyczące stężeń metabolitów występujących w szlaku, a także stałych kinetycznych enzymów katalizujących reakcje zachodzące w modelowanym szlaku. Następnie dane te wykorzystałem do przygotowania równań kinetyki enzymatycznej Michaelisa-Mentena. Przeprowadziłem analizę wyników symulacji i porównałem uzyskane wyniki z danymi literaturowymi. Napisałem oryginalny projekt niniejszej publikacji.

Oświadczam, że w artykule „Queueing theory model of pentose phosphate pathway” opublikowanym w czasopiśmie *Scientific Reports* w 2022r.; <https://doi.org/10.1038/s41598-022-08463-y>, którego współautorami są Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis oraz Tadeusz A. Wysocki, mój wkład w badania zgłoszone w niniejszej pracy polegał na konceptualizacji i realizacji badań. Przeprowadziłem dokładną analizę stanu techniki, zaprojektowałem strukturę algorytmu. Zebrałem niezbędne do badań i przygotowania modelu parametry i dane liczbowe dotyczące stężeń metabolitów występujących w szlaku, a także stałych kinetycznych enzymów katalizujących reakcje zachodzące w modelowanym szlaku. Następnie dane te wykorzystałem do przygotowania równań kinetyki enzymatycznej Michaelisa-Mentena. Przeprowadziłem analizę wyników symulacji i porównałem uzyskane wyniki z danymi literaturowymi. Napisałem oryginalny projekt niniejszej publikacji.

Oświadczam, że w artykule „Conversion of fat to cellular fuel—Fatty acids beta-oxidation model” opublikowanym w czasopiśmie *Computational Biology and Chemistry* w roku 2023.; <https://doi.org/10.1016/j.compbiolchem.2023.107860>, którego współautorami są Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis oraz Tadeusz Wysocki, mój wkład w badania zgłoszone w niniejszej pracy polegał na konceptualizacji i realizacji badań. Przeprowadziłem dokładną analizę stanu techniki, zaprojektowałem strukturę algorytmu. Zebrałem niezbędne do badań i przygotowania modelu parametry i dane liczbowe dotyczące stężeń metabolitów występujących w szlaku, a także stałych kinetycznych enzymów katalizujących reakcje zachodzące w modelowanym szlaku. Następnie dane te wykorzystałem do przygotowania równań kinetyki enzymatycznej Michaelisa-Mentena. Przeprowadziłem analizę wyników symulacji i porównałem uzyskane wyniki z danymi literaturowymi. Napisałem oryginalny projekt niniejszej publikacji.

Oświadczam, że w artykule „Queueing theory model of mTOR complexes’ impact on Akt-mediated adipocytes response to insulin” opublikowanym w czasopiśmie *PLoS One* w 2022r.; <https://doi.org/10.1371/journal.pone.0279573>, którego współautorami są Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata J. Wysocki, Paul Davis, Ghada A. Soliman oraz Tadeusz A. Wysocki, mój wkład w badania zgłoszone w niniejszej pracy polegał na konceptualizacji i realizacji badań. Przeprowadziłem dokładną analizę stanu wiedzy, zaprojektowałem strukturę algorytmu. Zebrałem niezbędne do badań i przygotowania modelu dane liczbowe dotyczące stężeń białek sygnalizacyjnych biorących udział w procesie odpowiedzi komórkowej na insulinę, a także

stałych kinetycznych. Następnie wykorzystałem te parametry do przygotowania równań kinetycznych opartych na prawie oddziaływania mas. Przeprowadziłem analizę wyników symulacji i porównałem uzyskane wyniki z danymi literaturowymi. Napisałem pierwotny projekt niniejszej publikacji.

Oświadczam, że w artykule „Integrating glycolysis, citric acid cycle, pentose phosphate pathway, and fatty acid beta-oxidation into a single computational model” opublikowanym w czasopiśmie *Scientific Reports* w 2023r.; <https://doi.org/10.1038/s41598-023-41765-3>, którego współautorami są Sylwester Klośka, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Beata J. Wysocki, Paul Davis oraz Tadeusz Wysocki, mój wkład w badania zgłoszone w niniejszej pracy polegał na konceptualizacji i realizacji badań. Przeprowadziłem dokładną analizę stanu techniki, zaprojektowałem strukturę algorytmu. Zebrałem niezbędne do badań i przygotowania modelu parametry i dane liczbowe dotyczące stężeń metabolitów występujących w szlaku, a także stałych kinetycznych enzymów katalizujących reakcje zachodzące w modelowanym szlaku. Następnie dane te wykorzystałem do przygotowania równań kinetyki enzymatycznej Michaelisa-Mentena. Przeprowadziłem analizę wyników symulacji i porównałem uzyskane wyniki z danymi literaturowymi. Napisałem oryginalny projekt niniejszej publikacji.

Jednocześnie wyrażam zgodę na przedłożenie w/w publikacji przez mgr inż. Krzysztofa Pałczyńskiego jako część rozprawy doktorskiej w formie spójnego tematycznie zbioru artykułów naukowych opublikowanych w czasopiśmie naukowych. Oświadczam, że udostępnienie utworów nie będzie naruszało praw autorskich osób trzecich.



(Podpis)

Oświadczenie Współautora

Dr inż. Sandra Magdalena Śmigiel
(tytuł zawodowy, imiona i nazwisko współautora)

Wydział Inżynierii Mechanicznej
Politechnika Bydgoska im. Jana i Jędrzeja Śniadeckich w Bydgoszczy
(miejsce pracy/afiliacja)

OŚWIADCZENIE

Oświadczam, iż mój wkład autorski w niżej wymienionym/wymienionych artykule/artykułach naukowym/naukowych był następujący*:

1. Śmigiel, S., Pałczyński, K., & Ledziński, D. (2021). ECG signal classification using deep learning techniques based on the PTB-XL dataset. *Entropy*, 23(9), 1121. DOI: <https://doi.org/10.3390/e23091121> pkt. MNISW: 100, Impact Factor: 2.738.
Wykonane zadania w ramach artykułu:
 - a) pomysłodawca koncepcji badań nad klasyfikacją sygnału EKG z zastosowaniem bazy PTB-XL,
 - b) analiza stanu wiedzy w zakresie: charakterystyki sygnału EKG i potencjału zastosowania modeli uczenia maszynowego, oceny dostępnych baz danych z sygnałami EKG, wyników badań innych autorów i określenie luk w wiedzy, które stanowią podstawę do dalszych badań,
 - c) przygotowanie danych polegające na wstępnej filtracji rekordów zawartych w bazie z wyborem klas i podklas diagnostycznych, istotnych w kontekście klasyfikacji chorób układu sercowo-naczyniowego,
 - d) walidacja badań i analiza wyników,
 - e) aktywne uczestnictwo w opracowaniu kolejnych wersji artykułu,
 - f) czynny udział w procesie korekty artykułu po recenzjach,
 - g) pierwszy autor, autor korespondencyjny.
2. Śmigiel, S., Pałczyński, K., & Ledziński, D. (2021). Deep learning techniques in the classification of ECG signals using R-peak detection based on the PTB-XL dataset. *Sensors*, 21(24), 8174. DOI: <https://doi.org/10.3390/s21248174> pkt. MNISW: 100, Impact Factor: 3.847.
Wykonane zadania w ramach artykułu:
 - a) analiza stanu wiedzy w zakresie: charakterystyki sygnału EKG i potencjału zastosowania modeli uczenia maszynowego, oceny dostępnych baz danych z sygnałami EKG, oceny i identyfikacji dostępnych algorytmów wykrywania załamka R w sygnale EKG, wyników badań innych autorów i określenie luk w wiedzy, które stanowią podstawę do dalszych badań,
 - b) przygotowanie danych polegające na wstępnej filtracji rekordów zawartych w bazie z wyborem klas i podklas diagnostycznych, istotnych w kontekście klasyfikacji chorób układu sercowo-naczyniowego, przygotowanie referencyjnych rekordów sygnałów EKG z manualnym

* W przypadku prac dwu- lub wieloautorских wymagane są oświadczenia kandydata do stopnia doktora oraz współautorów, wskazujące na ich merytoryczny wkład w powstanie każdej pracy (np. sformułowanie hipotezy badawczej, pomysłodawca badań, wykonanie specyficznych badań – np. przeprowadzenie konkretnych doświadczeń, opracowanie i obróbenie ankiet itp., wykonanie analizy wyników, przygotowanie manuskryptu artykułu i tunc). Określenie wkładu danego autora, w tym kandydata do stopnia doktora, powinno być na tyle precyzyjne, aby umożliwić dokładną ocenę jego wkładu i roli w powstaniu każdej pracy.

oznaczeniem załamek R, jako zbiór testowy oraz filtracji rekordów po zastosowaniu algorytmu ekstrakcji odcinków QRS,

- c) pomysłodawca i autor opracowanego algorytmu detekcji załamek R,
- d) pomysłodawca i autor opracowanego algorytmu ekstrakcji odcinków QRS,
- e) walidacja badań i analiza wyników,
- f) aktywne uczestnictwo w opracowaniu kolejnych wersji artykułu,
- g) czynny udział w procesie korekty artykułu po recenzjach,
- h) pierwszy autor, autor korespondencyjny.

3. Pałczyński, K., Śmigiel, S., Ledziński, D., & Bujnowski, S. (2022). Study of the few-shot learning for ECG classification based on the PTB-XL dataset. *Sensors*, 22(3), 904. DOI: <https://doi.org/10.3390/s22030904> pkt. MNiSW: 100, Impact Factor: 3.847.

Wykonane zadania w ramach artykułu:

- a) analiza stanu wiedzy w zakresie: charakterystyki sygnału EKG i potencjału zastosowania modeli uczenia maszynowego, dotychczasowych wyników badań w kontekście wykrywania załamka R w sygnale EKG, obszarów zastosowania Few-Shot Learning w medycynie z przedstawieniem wyników badań innych autorów,
- b) przygotowanie danych polegające na wstępnej filtracji rekordów zawartych w bazie z wyborem klas i podklas diagnostycznych, istotnych w kontekście klasyfikacji chorób układu sercowo-naczyniowego,
- c) walidacja badań i analiza wyników,
- d) aktywne uczestnictwo w opracowaniu kolejnych wersji artykułu,
- e) czynny udział w procesie korekty artykułu po recenzjach,
- f) autor korespondencyjny.

4. Pałczyński, K., Śmigiel, S., Gackowska, M., Ledziński, D., Bujnowski, S., & Lutowski, Z. (2021). IoT application of transfer learning in hybrid artificial intelligence systems for acute lymphoblastic leukemia classification. *Sensors*, 21(23), 8025. DOI: <https://doi.org/10.3390/s21238025> pkt. MNiSW: 100, Impact Factor: 3.847.

Wykonane zadania w ramach artykułu:

- a) pomysłodawca idei wykorzystania bazy ALL-IDB,
- b) analiza bazy danych ALL-IDB i zawartych w niej danych w kontekście medycznym,
- c) walidacja badań i analiza wyników,
- d) udział w procesie korekty artykułu po recenzjach,
- e) autor korespondencyjny.

Jednocześnie wyrażam zgodę na przedłożenie wyżej wymienionej/wymienionych pracy/prac przez mgr inż. Krzysztofa Pałczyńskiego (podać tytuł zawodowy imię i nazwisko kandydata do stopnia doktora) jako część rozprawy doktorskiej opartej na zbiorze opublikowanych i powiązanych tematycznie artykułów naukowych.

Budapest, 28.07.2024
miejscowość, data


podpis Współautora

Marissa Miller, PhD

Department of Electrical and Computer Engineering

University of Nebraska-Lincoln, USA

Statement

I hereby declare that in the paper "Queueing theory model of Krebs cycle" published in the journal *Bioinformatics* in 2021; <https://doi.org/10.1093/bioinformatics/btab177>, whose co-authors are Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata J. Wysocki, Paul Davis and Tadeusz A. Wysocki, my substantive contribution consisted in advising on the programming part of the study.

I hereby declare that in the paper "Queueing theory model of pentose phosphate pathway" published in the journal *Scientific Reports* in 2022; <https://doi.org/10.1038/s41598-022-08463-y>, whose co-authors are Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis, and Tadeusz A. Wysocki, my substantive contribution consisted of advising on the programming part of the study.

I hereby declare that in the paper "Conversion of fat to cellular fuel-Fatty acids beta-oxidation model" published in the journal *Computational Biology and Chemistry* in 2023; <https://doi.org/10.1016/j.compbiolchem.2023.107860>, whose co-authors are Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis, and Tadeusz A. Wysocki, my substantive contribution consisted of advising on the programming part of the study.

I hereby declare that in the paper "Queueing theory model of mTOR complexes' impact on Akt-mediated adipocytes response to insulin" published in the journal *PLOS One* in 2022; <https://doi.org/10.1371/journal.pone.0279573>, whose co-authors are Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis, Ghada A. Soliman and Tadeusz A. Wysocki, my substantive contribution consisted in advising on the programming part of the study.

At the same time, I consent to the submission of the above-mentioned publication by M.Eng. Krzysztof Pałczyński as part of his doctoral dissertation in the form of a thematically coherent set of scientific articles published in scientific journals.

Marissa Miller

.....
(Signature)

Prof. Paul Davis
Department of Biology,
University of Nebraska at Omaha, Omaha, NE 68182, USA

Statement

I hereby declare that in the paper "Queueing theory model of Krebs cycle" published in the journal *Bioinformatics* in 2021; <https://doi.org/10.1093/bioinformatics/btab177>, whose co-authors are Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata J. Wysocki, Paul Davis and Tadeusz A. Wysocki, my substantive contribution was to advise on the biochemical part of the study.

I hereby declare that in the paper "Queueing theory model of pentose phosphate pathway" published in the journal *Scientific Reports* in 2022; <https://doi.org/10.1038/s41598-022-08463-y>, whose co-authors are Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis, and Tadeusz A. Wysocki, my substantive contribution was to advise on the biochemical part of the study.

I hereby declare that in the paper "Conversion of fat to cellular fuel-Fatty acids beta-oxidation model" published in the journal *Computational Biology and Chemistry* in 2023; <https://doi.org/10.1016/j.compbiolchem.2023.107860>, whose co-authors are Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis, and Tadeusz A. Wysocki, my substantive contribution consisted of advising on the biochemical part of the study.

I hereby declare that in the paper "Queueing theory model of mTOR complexes' impact on Akt-mediated adipocytes response to insulin" published in the journal *PLOS One* in 2022; <https://doi.org/10.1371/journal.pone.0279573>, whose co-authors are Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis, Ghada A. Soliman and Tadeusz A. Wysocki, my substantive contribution was to advise on the biochemical part of the study.

I hereby declare that in the paper "Integrating glycolysis, citric acid cycle, pentose phosphate pathway, and fatty acids beta-oxidation into a single computational model" published in the journal *Scientific Reports* in 2023; <https://doi.org/10.1038/s41598-023-41765-3>, whose co-authors are Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Beata J. Wysocki, Paul Davis, and Tadeusz A. Wysocki, my substantive contribution consisted of advising on the biochemical part of the study.

At the same time, I consent to the submission of the above-mentioned publication by M.Eng. Krzysztof Pałczyński as part of his doctoral dissertation in the form of a thematically coherent set of scientific articles published in scientific journals.

.....
(Signature)

Prof. dr hab. inż. Tadeusz A. Wysocki

¹Wydział Telekomunikacji, Informatyki i Elektrotechniki
Politechnika Bydgoska im. J. J. Śniadeckich w Bydgoszczy

²Department of Electrical and Computer Engineering
University of Nebraska-Lincoln, USA

Oświadczenie

Oświadczam, że w artykule „Queueing theory model of Krebs cycle” opublikowanym w czasopiśmie *Bioinformatics* w 2021r.; <https://doi.org/10.1093/bioinformatics/btab177>, którego współautorami są Sylwester Kloska, Krzysztof Palczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata J. Wysocki, Paul Davis oraz Tadeusz A. Wysocki, mój wkład merytoryczny polegał na opiece nad badaniami i współredagowaniu artykułu (senior author).

Oświadczam, że w artykule „Queueing theory model of pentose phosphate pathway” opublikowanym w czasopiśmie *Scientific Reports* w 2022r.; <https://doi.org/10.1038/s41598-022-08463-y>, którego współautorami są Sylwester Kloska, Krzysztof Palczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata J. Wysocki, Paul Davis oraz Tadeusz A. Wysocki, mój wkład merytoryczny polegał na opiece nad badaniami i współredagowaniu artykułu (senior author).

Oświadczam, że w artykule „Conversion of fat to cellular fuel—Fatty acids beta-oxidation model” opublikowanym w czasopiśmie *Computational Biology and Chemistry* w roku 2023.; <https://doi.org/10.1016/j.ccbi.2023.107800>, którego współautorami są Sylwester Kloska, Krzysztof Palczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis oraz Tadeusz Wysocki, mój wkład merytoryczny polegał na opiece nad badaniami i współredagowaniu artykułu (senior author).

Oświadczam, że w artykule „Queueing theory model of mTOR complexes’ impact on Akt-mediated adipocytes response to insulin” opublikowanym w czasopiśmie *PLOS One* w 2022r.; <https://doi.org/10.1371/journal.pone.0275573>, którego współautorami są Sylwester Kloska, Krzysztof Palczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata J. Wysocki, Paul Davis, Ghada A. Soliman oraz Tadeusz A. Wysocki, mój wkład merytoryczny polegał na opiece nad badaniami i współredagowaniu artykułu (senior author).

Oświadczam, że w artykule „Integrating glycolysis, citric acid cycle, pentose phosphate pathway, and fatty acid beta-oxidation into a single computational model” opublikowanym w czasopiśmie *Scientific Reports* w 2023r.; <https://doi.org/10.1038/s41598-023-42765-3>, którego współautorami są Sylwester Kloska, Krzysztof Palczyński, Tomasz Marciniak, Tomasz Talaśka, Beata J. Wysocki, Paul Davis oraz Tadeusz Wysocki, mój wkład merytoryczny polegał na opiece nad badaniami i współredagowaniu artykułu (senior author).

Jednocześnie wyrażam zgodę na przedłożenie w/w publikacji przez mgr inż. Krzysztofa Palczyńskiego jako część rozprawy doktorskiej w formie spójnego tematycznie zbioru artykułów naukowych opublikowanych w czasopiśmie naukowych. Oświadczam, że udostępnienie utworów nie będzie naruszało praw autorskich osób trzecich.



(Podpis)

Oświadczenie Współautora

Dr hab. inż. Tomasz Andrysiak, prof. PBS

.....
(tytuł zawodowy, imię i nazwisko współautora)

Wydział Telekomunikacji, Informatyki i Elektrotechniki

.....
(miejsce pracy/afiliacja)

OŚWIADCZENIE

Oświadczam, iż mój wkład autorski w niżej wymienionym/wymienionych artykule/artykulech naukowym/naukowych był następujący*:

1. Pałczyński, K., Ledziński, D., & Andrysiak, T. (2022). Entropy Measurements for Leukocytes' Surrounding Informativeness Evaluation for Acute Lymphoblastic Leukemia Classification. Entropy, 24(11), 1560. DOI: <https://doi.org/10.3390/e24111560> pkt. MNiSW: 100, Impact Factor: 2.738

Wykonane zadania w ramach artykułu:

- a) walidacja badań i analiza wyników,
- b) aktywne uczestnictwo w opracowaniu kolejnych wersji artykułu,
- c) czynny udział w procesie korekty artykułu po recenzjach.

Jednocześnie wyrażam zgodę na przedłożenie wyżej wymienionej/wymienionych pracy/prac przez mgr. inż. Krzysztofa Pałczyńskiego (podać tytuł zawodowy imię i nazwisko kandydata do stopnia doktora) jako część rozprawy doktorskiej opartej na zbiorze opublikowanych i powiązanych tematycznie artykułów naukowych.

B. Andrysiak 28.05.2021
.....
miejscowość, data

.....
podpis Współautora

* W przypadku prac liter- lub naukowych wymagane są oświadczenia kandydata do stopnia doktora oraz współautorów, wskazujące na ich rzeczywisty wkład w powstanie każdej pracy (np. treść hipotezy badawczej, pomysłów i działań, wykonanie specyficznych badań – np. przeprowadzenie konkretnych doświadczeń, opracowanie i wykonanie układu itp.), wykonanie analogicznych przygotowanie manuskryptu artykułu i inne). Określenie wkładu danego autora, w tym kandydata do stopnia doktora, musi być na tyle precyzyjne, aby umożliwić dokładną ocenę jego udziału i roli w powstaniu każdej pracy.

Dr inż. Tomasz Marciniak, prof. P85

Wydział Telekomunikacji, Informatyki i Elektrotechniki

Politechnika Bydgoska (m. J. i J. Śniadeckich w Bydgoszczy

Oświadczenie

Oświadczam, że w artykule „Queueing theory model of Krebs cycle” opublikowanym w czasopiśmie *Bioinformatics* w 2021r.; <https://doi.org/10.1093/bioinformatics/btab177>, którego współautorami są Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata J. Wysocki, Paul Davis oraz Tadeusz A. Wysocki, mój wkład merytoryczny polegał na doradztwie w części programistycznej badania.

Oświadczam, że w artykule „Queueing theory model of pentose phosphate pathway” opublikowanym w czasopiśmie *Scientific Reports* w 2022r.; <https://doi.org/10.1038/s41598-022-08463-y>, którego współautorami są Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis oraz Tadeusz A. Wysocki, mój wkład merytoryczny polegał na doradztwie w części programistycznej badania.

Oświadczam, że w artykule „Conversion of fat to cellular fuel—Fatty acids beta-oxidation model” opublikowanym w czasopiśmie *Computational Biology and Chemistry* w roku 2023.; <https://doi.org/10.1016/j.cmbiolchem.2023.107960>, którego współautorami są Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis oraz Tadeusz Wysocki, mój wkład merytoryczny polegał na doradztwie w części programistycznej badania.

Oświadczam, że w artykule „Queueing theory model of mTOR complexes’ impact on Akt-mediated adipocytes response to insulin” opublikowanym w czasopiśmie *PLOS One* w 2022r.; <https://doi.org/10.1371/journal.pone.0279573>, którego współautorami są Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata J. Wysocki, Paul Davis, Ghada A. Solliman oraz Tadeusz A. Wysocki, mój wkład merytoryczny polegał na doradztwie w części programistycznej badania.

Oświadczam, że w artykule „Integrating glycolysis, citric acid cycle, pentose phosphate pathway, and fatty acid beta-oxidation into a single computational model” opublikowanym w czasopiśmie *Scientific Reports* w 2023r.; <https://doi.org/10.1038/s41598-023-41765-3>, którego współautorami są Sylwester Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Beata J. Wysocki, Paul Davis oraz Tadeusz Wysocki, mój wkład merytoryczny polegał na doradztwie w części programistycznej badania.

Jednocześnie wyrażam zgodę na przedłożenie w/w publikacji przez mgr inż. Krzysztofa Pałczyńskiego jako część rozprawy doktorskiej w formie spójnego tematycznie zbioru artykułów naukowych opublikowanych w czasopiśmie naukowych. Oświadczam, że udostępnienie utworów nie będzie naruszało praw autorskich osób trzecich.



(Podpis)

dr hab. inż. Tomasz Talaśka, prof. PŚ

Wydział Telekomunikacji, Informatyki i Elektrotechniki

Politechnika Bydgoska im. J. J. Śniadeckich w Bydgoszczy

Oświadczenie

Oświadczam, że w artykule „Queueing theory model of Krebs cycle” opublikowanym w czasopiśmie *Bioinformatics* w 2021r.; <https://doi.org/10.1093/bioinformatics/btab177>, którego współautorami są Sylwester Klośka, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata J. Wysocki, Paul Davis oraz Tadeusz A. Wysocki, mój wkład merytoryczny polegał na doradztwie w części programistycznej badania.

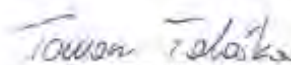
Oświadczam, że w artykule „Queueing theory model of pentose phosphate pathway” opublikowanym w czasopiśmie *Scientific Reports* w 2022r.; <https://doi.org/10.1038/s41598-022-08463-w>, którego współautorami są Sylwester Klośka, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis oraz Tadeusz A. Wysocki, mój wkład merytoryczny polegał na doradztwie w części programistycznej badania.

Oświadczam, że w artykule „Conversion of fat to cellular fuel—Fatty acids beta-oxidation model” opublikowanym w czasopiśmie *Computational Biology and Chemistry* w roku 2023.; <https://doi.org/10.1016/j.campbiochem.2023.107860>, którego współautorami są Sylwester Klośka, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki, Paul Davis oraz Tadeusz Wysocki, mój wkład merytoryczny polegał na doradztwie w części programistycznej badania.

Oświadczam, że w artykule „Queueing theory model of mTOR complexes’ impact on Akt-mediated adipocytes response to insulin” opublikowanym w czasopiśmie *PLOS One* w 2022r.; <https://doi.org/10.1371/journal.pone.0279523>, którego współautorami są Sylwester Klośka, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata J. Wysocki, Paul Davis, Ghada A. Soliman oraz Tadeusz A. Wysocki, mój wkład merytoryczny polegał na doradztwie w części programistycznej badania.

Oświadczam, że w artykule „Integrating glycolysis, citric acid cycle, pentose phosphate pathway, and fatty acid beta-oxidation into a single computational model” opublikowanym w czasopiśmie *Scientific Reports* w 2023r.; <https://doi.org/10.1038/s41598-023-41765-3>, którego współautorami są Sylwester Klośka, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Beata J. Wysocki, Paul Davis oraz Tadeusz Wysocki, mój wkład merytoryczny polegał na doradztwie w części programistycznej badania.

Jednocześnie wyrażam zgodę na przedłożenie w/w publikacji przez mgr inż. Krzysztofa Pałczyńskiego jako część rozprawy doktorskiej w formie spójnego tematycznie zbioru artykułów naukowych opublikowanych w czasopiśmie naukowych. Oświadczam, że udostępnienie utworów nie będzie naruszało praw autorskich osób trzecich.



(Podpis)

Oświadczenie Współautora

Dr inż. Sławomir Bujnowski

.....
(tytuł zawodowy, imiona i nazwisko współautora)

Wydział Telekomunikacji, Informatyki i Elektrotechniki

.....
(miejsce pracy/afiliacja)

OŚWIADCZENIE

Oświadczam, iż mój wkład autorski w niżej wymienionym/wymienionych artykule/artykułach naukowym/naukowych był następujący*:

1. Pałczyński, K., Śmigiel, S., Ledziński, D., & Bujnowski, S. (2022). Study of the few-shot learning for ECG classification based on the PTB-XL dataset. *Sensors*, 22(3), 904. DOI: <https://doi.org/10.3390/s22030904> pkt. MNiSW: 100, Impact Factor: 3.847

Wykonane zadania w ramach artykułu:

- a) przygotowanie danych polegające na konwersji surowych danych do postaci wymaganej przez sieci neuronowej.


2. Pałczyński, K., Śmigiel, S., Gackowska, M., Ledziński, D., Bujnowski, S., & Lutowski, Z. (2021). IoT application of transfer learning in hybrid artificial intelligence systems for acute lymphoblastic leukemia classification. *Sensors*, 21(23), 8025. DOI: <https://doi.org/10.3390/s21238025> pkt. MNiSW: 100, Impact Factor: 3.847

Wykonane zadania w ramach artykułu:

- a) zarządzanie danymi.

Jednocześnie wyrażam zgodę na przedłożenie wyżej wymienionej/wymienionych pracy/prac przez mgr. inż. Krzysztofa Pałczyńskiego (podać tytuł zawodowy imię i nazwisko kandydata do stopnia doktora) jako część rozprawy doktorskiej opartej na zbiorze opublikowanych i powiązanych tematycznie artykułów naukowych.


.....
miejsce, data


.....
podpis współautora

* W przypadku prac druk. lub wizałowanych wymagane są oświadczenia kandydata lub innego doktora oraz współautora, wskazujące na ich rzeczywisty wkład w powstanie każdej pracy (np. wybór hipotezy badawczej, pomysłowości badań, wykonanie specjalnych badań – np. przeprowadzenie konkretnych doświadczeń, opracowanie i wykonanie układu itp.), wykonanie analizy wyników, przygotowanie ostatecznego artykułu i inne). Określenie wkładu drugiego autora, w tym kandydata do stopnia doktora, powinno być nie tyle precyzyjne, aby umożliwić dokładną ocenę jego udziału w roli w powstaniu każdej pracy.

Oświadczenie Współautora

Dr inż. Zbigniew Lutowski

.....
(tytuł zawodowy, imiona i nazwisko współautora)

Wydział Telekomunikacji, Informatyki i Elektrotechniki

.....
(miejsce pracy/afiliacja)

OŚWIADCZENIE

Oświadczam, iż mój wkład autorski w niżej wymienionym/wymienionych artykule/artykułach naukowym/naukowych był następujący*:

1. Pałczyński, K., Śmigiel, S., Gackowska, M., Ledziński, D., Bujnowski, S., & Lutowski, Z. (2021). IoT application of transfer learning in hybrid artificial intelligence systems for acute lymphoblastic leukemia classification. *Sensors*, 21(23), 8025. DOI: <https://doi.org/10.3390/s21238025> pkt. MNiSW: 100, Impact Factor: 3.847
Wykonane zadania w ramach artykułu:

- a) zarządzanie danymi.

Jednocześnie wyrażam zgodę na przedłożenie wyżej wymienionej/wymienionych pracy/prac przez mgr. inż. Krzysztofa Pałczyńskiego (podać tytuł zawodowy imię i nazwisko kandydata do stopnia doktora) jako część rozprawy doktorskiej opartej na zbiorze opublikowanych i powiązanych tematycznie artykułów naukowych.

Brydżowa 28.05.2021
.....
miejscowość, data

.....
podpis Współautora

* W przypadku prac dwu- lub wieloautorskich wymagane są udzieleniem kandydata do stopnia doktora oraz współautorów wskazujące im ich rzeczywisty wkład w wymienionej/wymienionych pracy/pracach (np. wybór literatury badawczej, doposażenie badań, wykonanie konkretnych badań – np. przeprowadzenie konkretnych doświadczeń, opracowanie i wykonanie układu itp., wykonanie analizy wyników, przygotowanie matematycznego artykułu i inne). Określenie wkładu danego autora, w tym kandydata do stopnia doktora, powinno być surowo przestrzegane, aby umożliwić dokładowi oceny jego wkładu i roli w poszczególnych pracach.

12. Kopie artykułów naukowych stanowiących monotematyczny cykl publikacji

| Lp. | Autorzy/Tytuł/ Czasopismo |
|-----|--|
| P1 | Sylwester Kloska, Krzysztof Palczyński , Tomasz Marciniak, Tomasz Talaśka, Marissa Nitz, Beata Wysocka, Paul Davis, Tadeusz Wysocki, <i>Queueing theory model of Krebs Cycle</i> , 2021, Bioinformatics, 37, 18, 2912-2919, 10.1093/bioinformatics/btab177 |
| P2 | Sylwester Kloska, Krzysztof Palczyński , Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata Wysocka, Paul Davis, Tadeusz Wysocki, <i>Queueing theory model of pentose phosphate pathway</i> , 2022, Scientific Reports, 12, 1, 10.1038/s41598-022-08463-y |
| P3 | Sylwester Kloska, Krzysztof Palczyński , Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata Wysocka, Paul Davis, Tadeusz Wysocki, <i>Conversion of fat to cellular fuel-Fatty acids beta-oxidation model</i> , 2023, Computational Biology and Chemistry, 104, 10.1016/j.compbiolchem.2023.107860 |
| P4 | Sylwester Kloska, Krzysztof Palczyński , Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata Wysocka, Paul Davis, Ghada Soliman, Tadeusz Wysocki, <i>Queueing theory model of mTOR complexes' impact on Akt-mediated adipocytes response to insulin</i> , PLoS One, 2022, 17, 12, 10.1371/journal.pone.0279573 |
| P5 | Sylwester Kloska, Krzysztof Palczyński , Tomasz Marciniak, Tomasz Talaśka, Beata Wysocka, Paul Davis, Tadeusz Wysocki, <i>Integrating glycolysis, citric acid cycle, pentose phosphate pathway, and fatty acid beta-oxidation into a single computational model</i> , Scientific Reports, 2023, 13, 1, 10.1038/s41598-023-41765-3 |
| P6 | Krzysztof Palczyński , Sandra Śmigiel, Marta Gackowska, Damian Ledziński, Sławomir Bujnowski, Zbigniew Lutowski, <i>IoT application of transfer learning in hybrid artificial intelligence systems for acute lymphoblastic leukemia classification</i> , Sensors, 2021, 21, 23, 10.3390/s21238025 |
| P7 | Krzysztof Palczyński , Damian Ledziński, Tomasz Andrysiak, <i>Entropy Measurements for Leukocytes' Surrounding Informativeness Evaluation for Acute Lymphoblastic Leukemia Classification</i> , Entropy, 2022, 21, 11, 10.3390/e24111560 |
| P8 | Sandra Śmigiel, Krzysztof Palczyński , Damian Ledziński, <i>ECG signal classification using deep learning techniques based on the PTB-XL dataset</i> , Entropy, 2021, 23, 9, 10.3390/e23091121 |
| P9 | Sandra Śmigiel, Krzysztof Palczyński , Damian Ledziński, <i>Deep learning techniques in the classification of ECG signals using R-peak detection based on the PTB-XL dataset</i> , Sensors, 2021, 21, 24, 10.3390/s21248174 |
| P10 | Krzysztof Palczyński , Sandra Śmigiel, Damian Ledziński, Sławomir Bujnowski, <i>Study of the few-shot learning for ECG classification based on the PTB-XL dataset</i> , Sensors, 2022, 22, 3, 10.3390/s22030904 |

Systems biology

Queueing theory model of Krebs cycle

Sylwester Kloska ^{1,*}, Krzysztof Pałczyński², Tomasz Marciniak², Tomasz Talaśka², Marissa Nitz³, Beata J. Wysocki⁴, Paul Davis⁴ and Tadeusz A. Wysocki^{2,3,*}

¹Faculty of Medicine, Nicolaus Copernicus University Ludwik Rydygier Collegium Medicum, 85-067 Bydgoszcz, Poland, ²Faculty of Telecommunications, Computer Science and Electrical Engineering, UTP University of Science and Technology, 85-796 Bydgoszcz, Poland, ³Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, Omaha, NE 68182, USA and ⁴Department of Biology, University of Nebraska at Omaha, Omaha, NE 68182, USA

*To whom correspondence should be addressed Contact: 503013@stud.umk.pl
Associate Editor: Pier Luigi Martelli

Received on December 29, 2020; revised on March 8, 2021; editorial decision on March 9, 2021; accepted on March 11, 2021

Abstract

Motivation: Queueing theory can be effective in simulating biochemical reactions taking place in living cells, and the article paves a step toward development of a comprehensive model of cell metabolism. Such a model could help to accelerate and reduce costs for developing and testing investigational drugs reducing number of laboratory animals needed to evaluate drugs.

Results: The article presents a Krebs cycle model based on queueing theory. The model allows for tracking of metabolites concentration changes in real time. To validate the model, a drug-induced inhibition affecting activity of enzymes involved in Krebs cycle was simulated and compared with available experimental data.

Availability and implementation: The source code is freely available for download at <https://github.com/UTP-WTliE/KrebsCycleUsingQueueingTheory>, implemented in C# supported in Linux or MS Windows.

Contact: 503013@stud.umk.pl or twysocki2@unl.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Modeling metabolic pathways can be extremely useful in the scientific world (Ederer *et al.*, 2014; Nazaret *et al.*, 2009; Theodosiou *et al.*, 2015; Wu *et al.*, 2007). The ability to predict a cell's response to changes in the surrounding environment, i.e. changes in metabolite levels or external stimulus, would greatly improve the designing of experiments. Therefore, the impact of these changes on the whole cell metabolism should be assessed in the experiment planning stage. Modeling, combined with the ever-evolving metabolomics, could help speed up the diagnosis and treatment of metabolic diseases. Often, drugs are tested on animals, given in various doses and assessed for their impact and effects. Such tests can be lethal to laboratory animals and are responsible for the majority of their deaths (Hajar, 2011; Hawkins *et al.*, 2019; Lynch and Slaughter, 2001). Developing an accurate metabolism model could reduce the need for animal studies and reduce animal cruelty. In addition, before drugs reach medical use, and after *in silico*, *in vitro* and/or *in vivo* testing, they undergo a series of long-term clinical trials during which their impact and long-term effects are carefully assessed. An effective metabolism model could accelerate the process and achieve cost reductions. Animal testing and human clinical trials are both necessary for validating the effectiveness of a drug, however, a long-term goal of our effort is to reduce our reliance on these tests and present *in silico* methods as a sufficient alternative.

Queueing theory is mainly used for issues related to telecommunications and engineering, yet queueing theory is suitable for modeling stochastic changes occurring in biological systems. Until now, the queueing theory has been used, for example, to model insulin levels and the number of insulin receptors needed. Such studies can help to understand insulin-dependent diseases (Cavas and Çavaş, 2007). Interestingly, thanks to the queueing theory, it was also possible to model the impact of ethanol consumption and remove the side effects caused by its consumption (sobering) (Guang, 1998). Such studies indicate a multitude of applications of queueing theory, also in modeling metabolic pathways. Queueing theory has been previously used to model a simple metabolism network and mimic chemical interactions between substrates and products (Evstigneev *et al.*, 2014). Recently, a model of glycolysis based on queueing theory has been presented (Clement *et al.*, 2020). The use of queueing theory is also beneficial from the computational perspective as it requires less computing power, thus accelerating computing time and allows simulations to be carried out in real time. Due to the nature of reactions in Krebs cycle, reaction products become substrates for the next reaction in the cycle. In addition, biological systems have well-organized ways to transform molecules, pass them down the pathway and transport them to where they are needed to maintain normal cell function (Tsitkov *et al.*, 2018), much like transmitting packets in the internet from one node to another one. For this

reason, we decided it would be reasonable to use queueing theory, which has been proven a useful modeling technique in communication systems, to model this metabolic cycle.

In this article, we present the entire process for creating a Krebs cycle simulation model: from a literature review to obtain empirical data on metabolites and enzymes necessary to model the reaction according to the Michaelis–Menten kinetics, through the description of the genetic algorithm used to optimize the kinetic constants found in different sources (Siess *et al.*, 1976; Singh and Ghosh, 2006), and finally the description of the model itself—the obtained concentration results and their comparison against the available data, as well as confirmation of the model’s effectiveness simulating the inhibition of the Krebs cycle induced by drugs.

The Krebs cycle, also known as the tricarboxylic acid cycle (TCA) or the citric acid cycle (CAC), takes place in a mitochondrial matrix. The purpose of the Krebs cycle is to produce energy in a form of guanosine triphosphate (GTP) which also releases carbon dioxide (CO₂) (Fig. 1, Table 1) as a byproduct. GTP is the energetic equivalent of adenosine triphosphate (ATP) (Korla and Mitra, 2014; Ponizovskiy, 2016; Smith and Robinson, 2011). Summarized equation of the Krebs cycle: acetyl-CoA + 3 NAD⁺ + FAD + GDP + P_i + 2 H₂O → 2 CO₂ + 3 NADH + FADH₂ + GTP + 2 H⁺ + CoA. Formulas of Krebs cycle reactions and the enzymes catalyzing these reactions are presented in Table 1. Reaction 1 in Table 1 is the first reaction in which the cell may obtain its energy in the form of acetyl-CoA. The reaction that completes the cycle is the transfer of two acetyl groups from acetyl-CoA to a four-carbon compound—oxaloacetate and the formation of a six-carbon molecule—citrate. Citrate then undergoes a series of reactions, during which energy is produced and CO₂ released. Acetyl-CoA serves as fuel for the Krebs cycle and can be derived from various sources including fats, carbohydrates and proteins, thus connecting the metabolic pathways of these elements. Acetyl-CoA can also be derived from pyruvate, the main product of the glycolysis. Nevertheless, the Krebs cycle is also a source of amino acid precursors, as well as a molecule that is extremely important for metabolism, the reduced form of NAD—NADH, which plays a role in many other reactions in the cell like oxidative phosphorylation (Krebs and Johnson, 1937). Due to the use of individual metabolites as intermediates for the synthesis of further compounds necessary for the proper cell function, these metabolites have the ability to leave the cycle by means of transport mechanisms that move them to the appropriate site. The criticality of the Krebs cycle in mammalian physiology is the primary reason we have sought to undertake the development of the model described herein. Additionally, dysregulation of the Krebs cycle would be deleterious, and could result in large energy losses and overproduction of cofactors like NADH. Cycle regulation is based on the cellular assessment of the amount of available substrates and resulting products. A low concentration of substrates or a high concentration of products will decrease reaction rates. Cellular ADP

availability and its conversion to ATP also affects the speed of reactions. Lower ADP concentrations cause accumulation of NADH, which has inhibitory properties for many enzymes. A high concentration of citrate also affects the course of the cycle because it can inhibit glycolysis reactions, thereby preventing metabolite flow.

2 Methodology

2.1 Obtaining data on metabolite concentrations and characterization of enzymes catalyzing cycle reactions

Interest in metabolomics has grown rapidly due to the development of mass spectrometry (MS), which makes it possible to assess the concentrations of individual metabolites despite the fact that some of them occur in very small amounts, including those in the Krebs cycle (Ahn *et al.*, 2017; Albe *et al.*, 1990; Bennett *et al.*, 2009; Ishii *et al.*, 2007; Milo *et al.*, 2010; Mogilevskaya *et al.*, 2006; Park *et al.*, 2016). The model developed by our team is based on existing knowledge of molecular concentrations as initial values to the model (Table 2). The kinetics of enzymatic reactions are calculated according to Michaelis–Menten kinetics (1) (Singh and Ghosh, 2006). Our model enables tracking the course of reactions—both reaction speed and product growth over time. The speed of the enzymatic reaction depends on factors such as the maximum speed at which the enzyme can convert substrate into a product, the concentration of substrate and the enzymatic constant.

$$v = \frac{V_f \frac{S_1 S_2}{K_{S_1} K_{S_2}} - V_r \frac{P_1 P_2}{K_{P_1} K_{P_2}}}{(1 + \frac{S_1}{K_{S_1}} + \frac{P_1}{K_{P_1}})(1 + \frac{S_2}{K_{S_2}} + \frac{P_2}{K_{P_2}})} \quad (1)$$

where v is the reaction speed, V_f is the forward reaction speed, V_r is the reverse reaction speed, S_1, S_2, \dots, S_x - substrate concentration in mmol/l, P_1, P_2, \dots, P_x is the substrate concentration in mmol/l, $K_{S_1}, K_{S_2}, \dots, K_{S_x}$ is the kinetic constant of substrate and $K_{P_1}, K_{P_2}, \dots, K_{P_x}$ is the kinetic constant of product.

Enzymatic properties of enzymes that are involved in Krebs cycle reactions are presented in Supplementary Table S1 (Singh and Ghosh, 2006). If enzymatic data was unavailable, appropriate assumptions were made. For example, if K_P is unknown, it can be calculated as $10 * K_S$; the reverse speed of reaction is 100x slower than forward reaction ($V_r = \frac{V_f}{100}$). These assumptions are based on previous research and empirical observations (Singh and Ghosh, 2006) and have been adapted in our model. The reaction equations based on the Michaelis–Menten kinetics are presented in Supplementary Table S2. The reaction rates were calculated using kinetic constants and metabolite concentrations available in literature (Ahn *et al.*, 2017; Albe *et al.*, 1990; Bennett *et al.*, 2009; Ishii *et al.*, 2007; Milo *et al.*, 2010; Mogilevskaya *et al.*, 2006; Park *et al.*, 2016; Siess *et al.*, 1976; Singh and Ghosh, 2006). The concentrations of the following metabolites were combined: isocitrate and cis-aconitate, as well as succinyl-CoA and succinate. Combined concentrations were 0.0216 and 0.73 mmol/l, respectively. Isocitrate and cis-aconitate, as well as succinyl-CoA and succinate are transient, and once produced they are immediately used in the next reaction in the cycle. Combining them for simulation purposes with the metabolites adjacent to them in the cycle accelerated the calculation processes of the model and improved its stability.

2.2 Queueing theory

There are many studies examining trials of individual metabolic pathways modeling, but the large amount of interactions between metabolites, enzymes and other biomolecules make modeling metabolic pathways an extremely difficult task. Until now, the preferred method used in systems modeling was ordinary differential equations (ODEs) (Ahn *et al.*, 2017; Cohen and Bergman, 1995; Ederer *et al.*, 2014; Foster *et al.*, 2019; Jahan *et al.*, 2016; Jeffrey *et al.*, 1999; Korla and Mitra, 2014; Kurata and Sugimoto, 2018; Mogilevskaya *et al.*, 2006). Several approaches used scenario-based modeling (SBM) in connection with already existing platforms and tools, like PlayGo (Dräger *et al.*, 2008; Lapid *et al.*, 2019; Nazaret

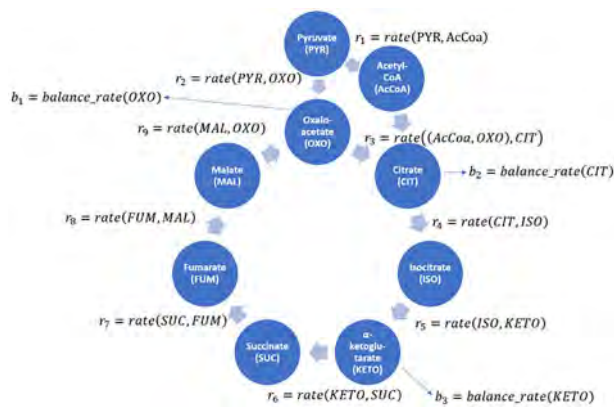


Fig. 1. Krebs cycle scheme using queueing theory. Molecules of the same type are queues. Reactions of the same type are servers. The queues are linked together, and the server output from one queue is connected to the input of the other queue

Table 1. Krebs cycle and related reactions—stoichiometric formulas

| Number | Reaction | Enzyme |
|--------|--|-----------------------------|
| 1 | Pyruvate + CoA + NAD ⁺ → acetyl-CoA + CO ₂ + NADH | Pyruvate dehydrogenase |
| 2 | Pyruvate + HCO ₃ ⁻ + ATP → oxaloacetate + ADP + P _i | Pyruvate carboxylase |
| 3 | Oxaloacetate + acetyl-CoA + H ₂ O → citrate + CoA + H ⁺ | Citrate synthetase |
| 4 | Citrate → cis-aconitate + H ₂ O | Aconitase |
| 5 | Cis-aconitate + H ₂ O → isocitrate | Aconitase |
| 6 | Isocitrate + NAD ⁺ → α-ketoglutarate + CO ₂ + NADH | Isocitrate dehydrogenase |
| 7 | α-Ketoglutarate + NAD ⁺ + CoA → succinyl-CoA + CO ₂ + NADH | Ketoglutarate dehydrogenase |
| 8 | Succinyl-CoA + P _i + GDP ↔ succinate + GTP + CoA | Succinate thiokinase |
| 9 | FAD + succinate → fumarate + FADH ₂ | Succinate dehydrogenase |
| 10 | Fumarate + H ₂ O → malate | Fumarase |
| 11 | Malate + NAD ⁺ ↔ oxaloacetate + NADH + H ⁺ | Malate dehydrogenase |

Table 2. Initial concentration values of Krebs cycle metabolites

| Metabolite, biomolecule | Concentration (mmol/l) | Reference |
|-------------------------------|------------------------|-----------------------------------|
| Coenzyme A | 0.044 | Park <i>et al.</i> (2016) |
| Pyruvate | 0.0586 | Clement <i>et al.</i> (2020) |
| Acetyl-CoA | 0.5 | Milo <i>et al.</i> (2010) |
| Citrate | 0.19 | Ahn <i>et al.</i> (2017) |
| Cis-aconitate | 0.0016 | Bennett <i>et al.</i> (2009) |
| Isocitrate | 0.02 | Milo <i>et al.</i> (2010) |
| α-ketoglutarate | 0.54 | Mogilevskaya <i>et al.</i> (2006) |
| Succinyl-CoA | 0.66 | Mogilevskaya <i>et al.</i> (2006) |
| Succinate | 0.07 | Albe <i>et al.</i> (1990) |
| Fumarate | 0.485 | Park <i>et al.</i> (2016) |
| Malate | 0.495 | Mogilevskaya <i>et al.</i> (2006) |
| Oxaloacetate | 0.006 | Mogilevskaya <i>et al.</i> (2006) |
| ATP | 0.159 | Clement <i>et al.</i> (2020) |
| ADP | 0.0937 | Clement <i>et al.</i> (2020) |
| GDP | 0.0012 | Milo <i>et al.</i> (2010) |
| NAD ⁺ | 0.099 | Milo <i>et al.</i> (2010) |
| NADH | 0.025 | Milo <i>et al.</i> (2010) |
| H ₂ O | 0.170 | Milo <i>et al.</i> (2010) |
| H ⁺ | 5.2 × 10 ⁻⁶ | Milo <i>et al.</i> (2010) |
| P _i | 0.05 | Milo <i>et al.</i> (2010) |
| HCO ₃ ⁻ | 0.003 | Milo <i>et al.</i> (2010) |

et al., 2009; Wu *et al.*, 2007). However, despite many trials and many years of research on metabolism, it has still not been accurately represented in any model. Researchers focus on individual metabolic pathway fragments to understand the metabolites and enzymes that transform them as accurately as possible (Berndt *et al.*, 2012; Iacobazzi and Infantino, 2014; Korla *et al.*, 2015; Tretter and Adam-Vizi, 2005). Thanks to this type of research, existing knowledge can be used in the model we propose. The use of queueing theory together with the grouping of molecules of the same type (queue) and reactions of the same type (server) allow a simpler model than the use of the Gillespie algorithm (Gillespie, 1977; Voit, 2017), where each reaction and each molecule is described by a separate node in Markov chains (Massey, 1985). Queueing networks can be considered and called as hidden Markov chain. As a result,

the mathematical-simulation model is identical to the biological one, as shown in Figure 2. Another advantage of using this approach is that it is impossible to achieve negative results in biological systems, as is sometimes the case with ODE-based models. There are methods forcing the system to obtain non-negative values (Shampine *et al.*, 2005), however, they could cause calculation errors. Usage of queueing theory as the basis for a Krebs cycle simulation model aims at providing a possible realization of stochastic Markovian processes representing variations in the concentration over a given metabolite. The average change in concentration can be achieved by averaging the simulation results for several simulation runs. At the heart of this stochastic model is Michaelis–Menten kinetic equations describing the relationship between quantities of substrate-product pairs and reaction velocities. In this theory, the

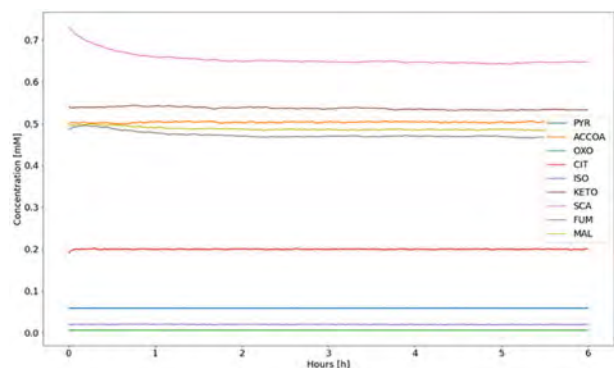


Fig. 2. Concentration level change over 24 h simulation under unperturbed conditions. 'PYR'—pyruvate; 'ACCOA'—acetyl-CoA; 'OXO'—oxaloacetate; 'CIT'—citrate; 'ISO'—isocitrate and cis-aconitate; 'KETO'— α -ketoglutarate; 'SCA'—succinyl-CoA and succinate; 'FUM'—fumarate; 'MAL'—malate

velocity of a reaction is a macroscopic representation of the aggregation of numerous microscopic reactions that may or may not exchange a fixed quantity of substances in a particular period of time. As a result, the speed of a reaction is described as a frequency of the reaction's occurrence and its relationship with the probability of increasing a certain substance by invoking reaction, in which the aforementioned metabolite is a product, and the probability of decreasing the substance by performing reaction, where it is a substrate. By describing the behavior of the Krebs cycle as probabilities of increasing and decreasing each of its substrates and correlating ones' reduction with others' accumulation, we achieved a self-regulating, stochastic process simulating the actual Krebs cycle. Michaelis–Menten kinetic equations are used to calculate the probability that a certain reaction occurs in the current time interval based on the amount of substrate, product and kinetic constants describing the reaction and the duration of the time interval. The results of these equations can be interpreted as the arrival and service rates in the Poisson processes, while an exponential distribution models the service time (the time intervals between two consecutive output events). These assumptions are consistent with classical queueing theory approaches. Therefore, the number of arrivals in any given time interval $(t, t + \tau]$ follows a Poisson distribution with a parameter $(\mu\tau)$ such that:

$$P[(N(t + \tau) - N(t)) = k] = \frac{e^{-\mu\tau} (\mu\tau)^k}{k!} \quad (2)$$

where $P[(N(t + \tau) - N(t)) = k]$ is the probability of k arrivals in the interval $(t, t + \tau]$ and $\mu\tau$ is the expected number of arrivals in a time interval of duration τ .

The time required for the queue to process the metabolite increment is described by the exponential distribution using the probability distribution of random variable X in the terms of the rate parameter μ as follows:

$$f(x; \mu) = \mu e^{-\mu x} \max(x, 0) \quad (3)$$

Therefore, the resulting arrival process at the input of a subsequent queue to which that output of the considered server is connected, follows a Poisson distribution. This a single multivariable stochastic process. All the variables are correlated. The process is described by a queueing network as shown in Figure 1, which consists of a queue describing arrivals and departures of discrete amounts of substances. For example, isocitrate (ISO), which is a product in the reaction citrate \rightarrow isocitrate (CIT \rightarrow ISO) and a substrate in reaction isocitrate \rightarrow α -ketoglutarate (ISO \rightarrow KETO). This means that the increment of ISO produced in the previous time interval adds to the queue of ISO for the ISO \rightarrow KETO reaction, effectively increasing the length of ISO queue. The Krebs cycle is a looped system constructed from queues, with increments of concentration of consecutive metabolites circulating, departing from one queue and arriving at another queue. According to Michaelis–

Menten kinetic equations, the probability of each packet arriving at the metabolite's queue is correlated with the amount of product and inversely correlated with amount of substrate, creating a self-regulating system, reacting to the imbalances of metabolites and equalizing the arrivals and departures from every queue.

2.3 Use of the genetic algorithm to find optimal values of kinetic constants

The genetic algorithm was used to find optimal values of kinetic constants for the Krebs cycle simulation. The genetic algorithm is a heuristic search inspired by Charles Darwin's theory of natural evolution and uses competing 'chromosomes' in order to find optimal parameters that minimize a fitness function (Man et al., 1999). A 'chromosome' in this implementation is the table of constants required for reaction rates calculation. 'The chromosome' is made of 'genes', which are constants used in one reaction. For example, the first gene in the 'chromosome' consists of constant values used in $\text{PYR} \rightarrow \text{AcCoA}$ reaction. There are one hundred 'chromosomes' in the population and each of them is a candidate for table of kinetic constants. The fitness function was designed to force the genetic algorithm to find a table of kinetic constants that allows values of products' concentrations to settle at stable points and to minimize the distance between start values and stable points. The designed fitness function is expressed as:

$$f(X) = \frac{1}{9} \sum_{i=0}^8 |X_{i,0} - \frac{1}{100} \sum_{j=0}^{99} X_{i,(|X|-j)}| \quad (4)$$

where X is the table of values of simulation product concentrations in time.

Evaluation of one 'chromosome' requires running a simulation using its set of genes as a table of kinetic constants. The simulation function returns the values of substrates' concentrations at each second. This table is used by the equation above to output the 'chromosome's' score. The function calculates an average vector of the last 100 recordings and computes absolute difference with initial simulation concentrations. In the last step, there is a calculated average of differences. The 'chromosome' minimizing this function is selected as the optimal table of kinetic constant values. Evaluation of each 'chromosome' is computed by simulating the Krebs cycle through the first one hour. There are 100 'chromosomes' in the population in each step of optimization and after evaluation only the 10 sets of constants that minimize the fitness function are selected for reproduction. The reproductive algorithm is a variation of the standard crossover with additional mechanism preventing the finding of a trivial solution to minimize the loss function problem, which is to zero the probability of every reaction. A step-by-step description of reproduction algorithm is presented in Supplementary Data. The main disadvantage of the fitness function described above is the existence of a trivial solution for its minimization problem. If the 'chromosome' contains only zeros, then no reaction would be performed, so the settling points of concentrations of products in the Krebs cycle would have the same values as initial concentrations, thus finding a global minimum. To prevent the genetic algorithm from converging to this solution, the reproduction mechanism requires that each reaction at $t=0$ has probability of being performed between 1% and 10%. Reaction and balancing flow rates have ranges from 1 to 10% at the beginning of the simulation started from substrates concentration values described in the literature. Applying these constraints to the reaction rates prevent them from being zeroed at the start and also prevents saturation of reactions. The reproduction algorithm has a 10% chance to perform a mutation with the mutation amplitude equal to 1.0.

2.4 Krebs cycle simulation pseudocode

The pseudocode describing the computation process of the Krebs cycle simulation is included in Supplementary Data. This code assumes that:

Table 3. Comparison of concentration data: literature and model (mmol/l)

| Metabolite | Initial concentration (literature) | Final concentration (model) | Standard deviation over mean | Absolute difference | Relative difference (%) |
|----------------------------|------------------------------------|-----------------------------|------------------------------|---------------------|-------------------------|
| Pyruvate | 0.0586 | 0.0586 | 0.0033 | 0.0 | 0.0 |
| Acetyl-CoA | 0.05 | 0.5028 | 0.0145 | 0.0028 | 0.55 |
| Oxaloacetate | 0.006 | 0.0059 | 0.0167 | -0.0001 | -1.5 |
| Citrate | 0.19 | 0.1994 | 0.018 | 0.0094 | 4.96 |
| Isocitrate + cis-aconitate | 0.0216 | 0.0216 | 0.0554 | 0.0 | 0.01 |
| α -ketoglutarate | 0.54 | 0.5346 | 0.0256 | -0.0054 | -1.01 |
| Succinyl-CoA + succinate | 0.73 | 0.6473 | 0.0167 | -0.0827 | -11.33 |
| Fumarate | 0.485 | 0.467 | 0.0161 | -0.018 | -3.72 |
| Malate | 0.495 | 0.4847 | 0.0107 | -0.0103 | -2.08 |

Note: Calculated relative difference shows similarity of obtained results and literature data.

- Kinetic constants are grouped into a table of vectors of constant values. There are 11 vectors in the table corresponding to nine different reactions and two balancing flows. Each of the reactions has a unique, four-dimensional vector and every balancing flow contains a one-dimensional table. The vector describing pyruvate carboxylase and citrate synthetase are the exceptions from this rule as they have eight and six elements, respectively. These exceptions are due to the nature of the reactions catalyzed by these enzymes. While most of the reactions require only four coefficients to be optimized the reactions catalyzed by pyruvate carboxylase and citrate synthetase involve more constants that need to be optimized.
- Concentration increment exchanged during the reactions is called 'delta' and is equal to 0.0001 mmol. 'Delta' is significantly lower than the initial value of the lowest substrate concentration. Delta value must be chosen in a way that it corresponds to a change of more than a single molecule for the rare species, in fact for rare species it should be always chosen to be a positive integer number of molecules.
- Simulation assumes that the concentration of pyruvate in the cycle is varying with 10% Gaussian noise around the constant value of 0.0586 mmol/l. Such signal-to-noise ratio depicts metabolic conditions inside the cell. Due to the various living conditions of the cell, pyruvate is consumed faster or slower. The pyruvate level is dependent on the blood glucose level, which also affects the glucose level in the cell, so the variation of 10% was assumed. It is an arbitrary choice, and the variation range can be changed if there are good reasons to do so, as was done with the drug effect simulation.

The searching for optimal kinetic constants was performed using a PC with IntelTM Core i7-7700HQ @ 2.80 GHz, RAM 16 GB. Code was written in C# 8.0. One search epoch simulating one hour for 100 different tables of kinetic constants using all 8 logic cores took approximately 10 min.

2.5 Inhibition of a specific stage of the cycle and its influence on the concentrations and kinetics of other reactions

To validate the model, we simulated the outcome of various drugs on the Krebs cycle, which are known to affect the concentrations of individual metabolites in the Krebs cycle. This approach may also emulate changes in enzyme activity associated with the progression of various diseases, including metabolic disorders or cancer (Tolstikov et al., 2014); Sutendra and Michelakis, 2013; Zhang et al., 2018). Drugs that affect enzyme reactions in the Krebs cycle are usually competitive inhibitors. Ultimately, the drug slows down the reaction carried out by a particular enzyme because the enzyme

Table 4. Comparison of concentration values during Phenformin treatment: empirical data and model based on the queueing theory

| Metabolite | Concentration change after Phenformin administration in comparison to non-treatment (Janzer et al., 2014) (%) | Model simulation results (%) |
|-------------------------|---|------------------------------|
| Pyruvate | -65 | -65 |
| Citrate | -60 | -59.28 |
| Isocitrate | -65 | -63.91 |
| α -ketoglutarate | -80 | -79.38 |
| Fumarate | -50 | -50.84 |
| Malate | -53 | -53.22 |

processes smaller amounts of substrate than it would under regular conditions, without an inhibitor. Understanding the kinetic properties of inhibitors would be sufficient to predict its effect on cell metabolism. Using the existing research on substances affecting various enzymes involved in the Krebs cycle reactions, an experiment was conducted to reflect the effect of the drug on the rate of enzymatic reaction and the concentration of metabolites.

3 Results

To validate the model, we have tested first its stability. The system becomes stable after approximately 5.5 h of simulation as shown in Figure 2. During this time, for every millisecond of simulation time one simulation step was performed.

In our opinion, these results are satisfactory. The largest relative difference observed in our model in comparison with available biological data is -11.33% in the case of the combined concentrations of succinyl-CoA and succinate (Table 3).

To reflect the kinetics of the Krebs cycle during inhibition induced by a drug that blocks one of the cycle reactions, we selected one of the studies based on the measurement of metabolite concentrations after administration of drugs in anti-cancer therapy (Janzer et al., 2014). This study provided the most detailed information on the concentrations of several metabolites included in the Krebs cycle. Therefore, the study served as the basis for the model to check whether it achieves similar results (Table 4). This study tested the effects of Tamoxifen, already used in the treatment of breast cancer, in combination with the drugs used by diabetics—Metformin and Phenformin. The idea to use these drugs in cancer therapy resulted from clinical observations (Evans et al., 2005; Jiralerspong et al., 2009; Kim et al., 2018; Pollak, 2012). According to these observations, cancer diagnosis incidence rate was lower in patients using the drugs, as well as the mortality rate due to cancer was lower in the diagnosed patients. This observation prompted the idea of combining Metformin and Phenformin together with the standardized Tamoxifen. Metformin and Phenformin doses were 300 and 10 μ M, respectively. One of the methods of assessing the effectiveness of the therapy was the measurement of the concentration of Krebs cycle

metabolites. This metabolic cycle is extremely important for cancer cells due to the high energy demands of these cells. Therefore, a reduction in the Krebs cycle efficiency and lowering the concentration of individual metabolites, may prove the effectiveness of the used treatment method. Simulations present changes in concentration levels of each Krebs cycle metabolite during treatment (Figs 3 and 4).

Due to the difficulty of obtaining measurements, the article (Janzer et al., 2014) did not include all the Krebs cycle metabolites. The data presented in Tables 4 and 5 confirm the accuracy of the proposed solution. The prepared model based on the queueing theory was designed to accurately reflect the stochastic nature of

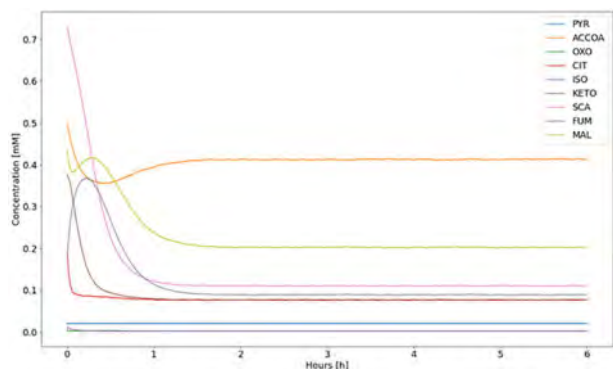


Fig. 3. Concentration levels change during inhibition: reflection of Tamoxifen + Phenformin treatment. 'PYR'—pyruvate; 'ACCOA'—acetyl-CoA; 'OXO'—oxaloacetate; 'CIT'—citrate; 'ISO'—isocitrate and cis-aconitate; 'KETO'— α -ketoglutarate; 'SCA'—succinyl-CoA and succinate; 'FUM'—fumarate; 'MAL'—malate

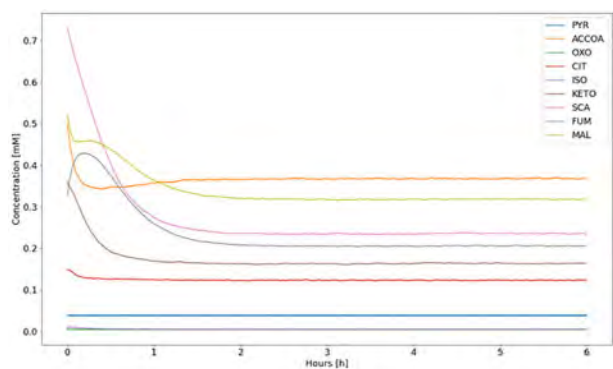


Fig. 4. Concentration levels change during inhibition: reflection of Tamoxifen + Metformin treatment. 'PYR'—pyruvate; 'ACCOA'—acetyl-CoA; 'OXO'—oxaloacetate; 'CIT'—citrate; 'ISO'—isocitrate and cis-aconitate; 'KETO'— α -ketoglutarate; 'SCA'—succinyl-CoA and succinate; 'FUM'—fumarate; 'MAL'—malate

biological reactions. For this reason, training and adapting it to the conditions and phenomena that may occur during biochemical reactions in the cell, also under the influence of pharmaceuticals, allowed to obtain results similar to the experimental values. Tables 6 and 7 present the results of a six-hour simulation of the concentrations (with 10% gaussian noise) of all Krebs cycle metabolites for Tamoxifen + Phenformin and Tamoxifen + Metformin treatment. To obtain the presented results, the so-called 'balancing flow' was used in the model. It imitates the drainage of metabolites due to their various uses in cell functioning (e.g. being precursors for other compounds). Balancing flow was used to stabilize the concentrations of oxaloacetate and citrate. Oxaloacetate is used in gluconeogenesis, in the urea cycle and in the synthesis of fatty acids to create citrate in the form of which it is transported; this happens when there is no demand for energy at the moment. Citrate is transported beyond the mitochondria to the cytoplasm, then broken down into acetyl-CoA and oxaloacetate for the synthesis of fatty acids. Probability rates of balancing flows for oxaloacetate and citrate are linearly correlated with the concentrations of respective metabolites. Such a composition of arriving, departing and balancing rates forms a stochastic representation of the Krebs cycle and provides an accurate and time-efficient model. By comparing the simulation and measurement results, we stipulate that the drugs administration in previously mentioned doses inhibits the reaction catalyzed by pyruvate dehydrogenase by about 30%. α -ketoglutarate and malate have been shown as examples of the metabolites which concentration was measured during the studies on the effects of these drugs (Figs 5 and 6). The model allows for the observation of pharmaceutical influence on the kinetics of the cycle reactions and their influence on metabolite levels.

4 Discussion

The presented results demonstrate the preparation of a model capable of mimicking the conditions of metabolic reactions in living cells. A disadvantage of our model is that the data on metabolite concentrations and enzyme constants come from different sources. Measurements carried out on different measuring devices by different research teams may not be fully compatible. However, to the

Table 5. Comparison of concentration values during Metformin treatment: empirical data and model based on the queueing theory

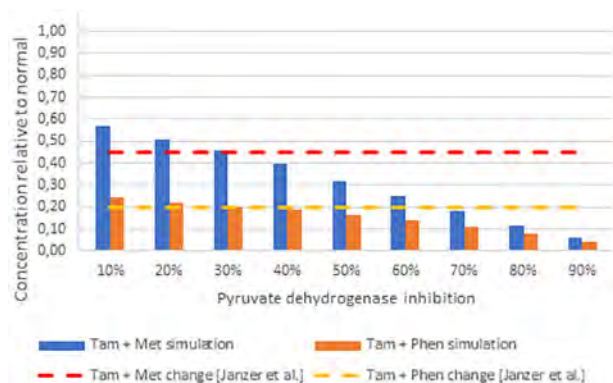
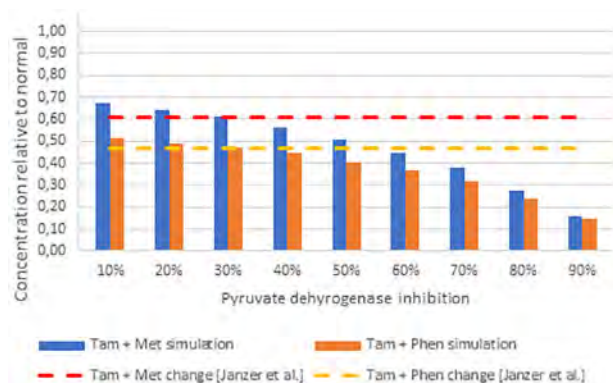
| Metabolite | Concentration change after Metformin administration in comparison to non-treatment (Janzer et al., 2014) (%) | Model simulation results (%) |
|-------------------------|--|------------------------------|
| Pyruvate | -35 | -35 |
| Citrate | -15 | -16.18 |
| Isocitrate | -40 | -39.17 |
| α -ketoglutarate | -55 | -54.47 |
| Fumarate | -37 | -37.03 |
| Malate | -39 | -39.43 |

Table 6. Comparison of concentration data: literature and model (mmol/l) under inhibition caused by Tamoxifen and Phenformin

| Metabolite | Initial concentration (literature) | Final concentration (model) | Standard deviation over mean | Absolute difference | Relative difference (%) |
|----------------------------|------------------------------------|-----------------------------|------------------------------|---------------------|-------------------------|
| Pyruvate | 0.0205 | 0.0205 | 0.0039 | 0.0 | 0.03 |
| Acetyl-CoA | 0.5 | 0.4129 | 0.0161 | -0.0871 | -17.42 |
| Oxaloacetate | 0.006 | 0.0025 | 0.0239 | -0.0035 | -58.7 |
| Citrate | 0.1899 | 0.0773 | 0.0267 | -0.1126 | -59.28 |
| Isocitrate + cis-aconitate | 0.0056 | 0.002 | 0.0636 | -0.0036 | -63.91 |
| α -ketoglutarate | 0.3764 | 0.0776 | 0.042 | -0.2988 | -79.38 |
| Succinyl-CoA + succinate | 0.73 | 0.1113 | 0.0308 | -0.6187 | -84.75 |
| Fumarate | 0.1825 | 0.0897 | 0.0315 | -0.0928 | -50.84 |
| Malate | 0.4335 | 0.2028 | 0.0164 | -0.2307 | -53.22 |

Table 7. Comparison of concentration data: literature and model (mmol/l) under inhibition caused by Tamoxifen and Metformin

| Metabolite | Initial concentration (literature) | Final concentration (model) | Standard deviation over mean | Absolute difference | Relative difference (%) |
|----------------------------|------------------------------------|-----------------------------|------------------------------|---------------------|-------------------------|
| Pyruvate | 0.0381 | 0.0381 | 0.0036 | 0.0 | 0.0 |
| Acetyl-CoA | 0.5 | 0.3676 | 0.0139 | -0.1324 | -26.49 |
| Oxaloacetate | 0.006 | 0.0044 | 0.015 | -0.0016 | -27.46 |
| Citrate | 0.1466 | 0.1229 | 0.0232 | -0.0237 | -16.18 |
| Isocitrate + cis-aconitate | 0.0081 | 0.0049 | 0.0709 | -0.0032 | -39.17 |
| α -ketoglutarate | 0.3596 | 0.1637 | 0.0274 | -0.1959 | -54.47 |
| Succinyl-CoA + succinate | 0.73 | 0.2346 | 0.0228 | -0.4954 | -67.86 |
| Fumarate | 0.3272 | 0.206 | 0.0227 | -0.1212 | -37.03 |
| Malate | 0.5235 | 0.3171 | 0.0139 | -0.2064 | -39.43 |

**Fig. 5.** α -Ketoglutarate concentration change in regards to pyruvate dehydrogenase inhibition. ‘Tam’—Tamoxifen; ‘Met’—Metformin; ‘Phen’—Phenformin**Fig. 6.** Malate concentration change in regard to pyruvate dehydrogenase inhibition. ‘Tam’—Tamoxifen; ‘Met’—Metformin; ‘Phen’—Phenformin

best of our knowledge there is no publication which presents metabolic data for every Krebs cycle metabolite derived from one research group. We have made every effort to ensure that the data used are as accurate and compatible with each other as possible. The models can be a kind of virtual laboratory where one can consider interdependencies between specific substances and metabolites and their influence on basic cellular functions. The developed model may provide knowledge on how chemical compounds obtain their therapeutic efficacy, which may result in improved safety from the early stages of drug development, e.g. setting up experiments before tests on living organisms, and between preclinical testing and clinical trials. The model allows to determine which reactions of the metabolic pathway are the best candidates for disturbing the metabolic pathway. In addition, we can observe what doses of the drug have a significant effect on the metabolic pathway, as well as the dose

above which the effect is imperceptible, something like ‘maximum effective dose’. There are many studies proving the usefulness of various drugs affecting cell metabolism. These drugs are used, in conjunction with others, in bacterial infections and in neoplastic diseases. The available literature data on the kinetic properties of enzymes and the concentration of metabolites can be used in this model. As mentioned before in the inhibition modeling section, Metformin and Phenformin are drugs that lower the level of metabolites in the Krebs cycle. This is associated with a reduction in the supply of pyruvate that could be transformed and enter the Krebs cycle, as well as an increase in the amount of lactate produced under anaerobic conditions. Previous studies have suggested that the Krebs cycle is not inhibited by metformin and changes the source of cellular ‘fuel’ (Janzer et al., 2014). However, the possibility that this apparent difference in inhibition of the Krebs cycle may be due to the analysis of stably transformed cancer cells as opposed to cells at an early stage of transformation was considered. However, biguanide treatment of the stably transformed CAMA-1 breast cancer cell line leads to a reduction in the concentrations of cycle intermediates. This suggests that metabolic reduction of the Krebs cycle by biguanides may be important for inhibiting transformation (Janzer et al., 2014).

5 Conclusions

The combination of knowledge available in the literature and the programming of the model provided a tool capable of mimicking Krebs cycle-related metabolic processes in living cells in real time. We demonstrated that metabolic pathways can be effectively simulated using methods based on queuing theory and affected by simulated application of drug. However, in self-criticism we found a place where the described model could be improved. We assume that access to data obtained under the same conditions on specific cells could potentially improve the obtained results, but due to limited access to such data, our model was prepared based on the most accurate available data. Future research efforts will be devoted to combining the Krebs cycle model with the previously developed glycolysis model (Clement et al., 2020) and adding the pentose phosphate pathway to obtain a comprehensive model of cellular carbohydrate metabolism.

Funding

This work was funded by the National Science Center (NCN) of Poland in terms of Opus-17 Program [2019/33/B/ST6/00875].

Conflict of Interest: none declared.

References

Ahn, E. et al. (2017) Temporal fluxomics reveals oscillations in TCA cycle flux throughout the mammalian cell cycle. *Mol. Syst. Biol.*, 13, 953.

- Albe, K.R. *et al.* (1990) Cellular concentrations of enzymes and their substrates. *J. Theor. Biol.*, **143**, 163–195.
- Bennett, B.D. *et al.* (2009) Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nat. Chem. Biol.*, **5**, 593–599.
- Berndt, N. *et al.* (2012) Kinetic modeling of the mitochondrial energy metabolism of neuronal cells: the impact of reduced-ketoglutarate dehydrogenase activities on ATP production and generation of reactive oxygen species. *Int. J. Cell Biol.*, **2012**, 1–11.
- Cavas, K.L. and Çavaş, L. (2007) An application of queueing theory to the relationship between insulin level and number of insulin receptors. *Türk Biyokimya Dergisi*, **32**, 32–38.
- Clement, E.J. *et al.* (2020) Stochastic simulation of cellular metabolism. *IEEE Access*, **8**, 79734–79744.
- Cohen, D.M. and Bergman, R.N. (1995) Estimation of TCA cycle flux, amino-transferase flux, and anaplerosis in heart: validation with syntactic model. *Am. J. Physiol. Endocrinol. Metab.*, **268**, E397–E409.
- Dräger, A. *et al.* (2008) SbmIsqueezer: a celldesigner plug-in to generate kinetic rate equations for biochemical networks. *BMC Syst. Biol.*, **2**, 39–37.
- Ederer, M. *et al.* (2014) A mathematical model of metabolism and regulation provides a systems-level view of how *Escherichia coli* responds to oxygen. *Front. Microbiol.*, **5**, 124.
- Evans, J.M.M. *et al.* (2005) Metformin and reduced risk of cancer in diabetic patients. *BMJ (Clin. Res. Ed.)*, **330**, 1304–1305.
- Evstigneev, V.P. *et al.* (2014) Theoretical description of metabolism using queueing theory. *Bull. Math. Biol.*, **76**, 2238–2248.
- Foster, C.J. *et al.* (2019) From *Escherichia coli* mutant 13c labeling data to a core kinetic model: a kinetic model parameterization pipeline. *PLoS Comput. Biol.*, **15**, e1007319.
- Gillespie, D.T. (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81**, 2340–2361.
- Guang, W. (1998) Application of queueing theory with Monte Carlo simulation to the study of the intake and adverse effects of ethanol. *Alcohol Alcoholism*, **33**, 519–527.
- Hajar, R. (2011) Animal testing and medicine. *Heart Views Off. J. Gulf Heart Assoc.*, **12**, 42.
- Hawkins, P. *et al.* (2019) Avoiding mortality in animal research and testing. In: *Avoiding Mortality in Animal Research and Testing*. University of Cambridge: RSPCA Research Animals Department, RSPCA Research Animals Department.
- Iacobazzi, V. and Infantino, V. (2014) Citrate—new functions for an old metabolite. *Biol. Chem.*, **395**, 387–399.
- Ishii, N. *et al.* (2007) Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science*, **316**, 593–597.
- Jahan, N. *et al.* (2016) Development of an accurate kinetic model for the central carbon metabolism of *Escherichia coli*. *Microb. Cell Factories*, **15**, 1–19.
- Janzer, A. *et al.* (2014) Metformin and phenformin deplete tricarboxylic acid cycle and glycolytic intermediates during cell transformation and NTPS in cancer stem cells. *Proc. Natl. Acad. Sci. USA*, **111**, 10574–10579.
- Jeffrey, F.M.H. *et al.* (1999) Use of a single ¹³C nmr resonance of glutamate for measuring oxygen consumption in tissue. *Am. J. Physiol. Endocrinol. Metab.*, **277**, E1111–E1121.
- Jiralerspong, S. *et al.* (2009) Metformin and pathologic complete responses to neoadjuvant chemotherapy in diabetic patients with breast cancer. *J. Clin. Oncol.*, **27**, 3297–3302.
- Kim, H.J. *et al.* (2018) Metformin reduces the risk of cancer in patients with type 2 diabetes: an analysis based on the Korean national diabetes program cohort. *Medicine*, **97**, e0036.
- Korla, K. and Mitra, C.K. (2014) Modelling the Krebs cycle and oxidative phosphorylation. *J. Biomol. Struct. Dyn.*, **32**, 242–256.
- Korla, K. *et al.* (2015) Kinetic simulation of malate-aspartate and citrate-pyruvate shuttles in association with Krebs cycle. *J. Biomol. Struct. Dyn.*, **33**, 2390–2403.
- Krebs, H.A. and Johnson, W.A. (1937) Metabolism of ketonic acids in animal tissues. *Biochem. J.*, **31**, 645–660.
- Kurata, H. and Sugimoto, Y. (2018) Improved kinetic model of *Escherichia coli* central carbon metabolism in batch and continuous cultures. *J. Biosci. Bioeng.*, **125**, 251–257.
- Lapid, H. *et al.* (2019). Using reactive-system modeling techniques to create executable models of biochemical pathways. In *Proceedings of the 7th International Conference on Model-Driven Engineering and Software Development, MODELSWARD 2019*. SCITEPRESS – Science and Technology Publications, Lda., Setubal, PRT, pp. 454–464.
- Lynch, J. and Slaughter, B. (2001) Recognizing animal suffering and death in medicine. *Western J. Med.*, **175**, 131–132.
- Man, K.F. *et al.* (1999) Introduction, background and biological inspiration. In: *Genetic Algorithms*. Springer, Springer London, pp. 1–21.
- Massey, W.A. (1985) Asymptotic analysis of the time dependent m/m/1 queue. *Math. Oper. Res.*, **10**, 305–327.
- Milo, R. *et al.* (2010) Bionumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Res.*, **38**, D750–D753.
- Mogilevskaya, E. *et al.* (2006) Kinetic model of mitochondrial Krebs cycle: unraveling the mechanism of salicylate hepatotoxic effects. *J. Biol. Phys.*, **32**, 245–271.
- Nazaret, C. *et al.* (2009) Mitochondrial energetic metabolism: a simplified model of TCA cycle with ATP production. *J. Theor. Biol.*, **258**, 455–464.
- Park, J.O. *et al.* (2016) Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. *Nat. Chem. Biol.*, **12**, 482–489.
- Pollak, M.N. (2012) Investigating metformin for cancer prevention and treatment: the end of the beginning. *Cancer Discov.*, **2**, 778–790.
- Ponizovskiy, M. (2016) Role of Krebs cycle in mechanism of stability internal medium and internal energy in an organism in norm and in mechanism of cancer pathology. *Mod. Chem. Appl.*, **4**, 2.
- Shampine, L.F. *et al.* (2005) Non-negative solutions of odes. *Appl. Math. Comput.*, **170**, 556–569.
- Siess, E. *et al.* (1976) Kinetic and regulatory properties of pyruvate dehydrogenase from Ehrlich ascites tumor cells. *Cancer Res.*, **36**, 55–59.
- Singh, V.K. and Ghosh, I. (2006) Kinetic modeling of tricarboxylic acid cycle and glyoxylate bypass in mycobacterium tuberculosis, and its application to assessment of drug targets. *Theor. Biol. Med. Modell.*, **3**, 27–11.
- Smith, A.C. and Robinson, A.J. (2011) A metabolic model of the mitochondrion and its use in modelling diseases of the tricarboxylic acid cycle. *BMC Syst. Biol.*, **5**, 102–113.
- Sutendra, G. and Michelakis, E.D. (2013) Pyruvate dehydrogenase kinase as a novel therapeutic target in oncology. *Front. Oncol.*, **3**, 38.
- Theodosiou, E. *et al.* (2015) Metabolic network capacity of *Escherichia coli* for Krebs cycle-dependent proline hydroxylation. *Microb. Cell Fact.*, **14**, 1–12.
- Tolstikov, V. *et al.* (2014) Metabolomics analysis of metabolic effects of nicotinamide phosphoribosyltransferase (NAMPT) inhibition on human cancer cells. *PLoS One*, **9**, e114019.
- Tretter, L. and Adam-Vizi, V. (2005) Alpha-ketoglutarate dehydrogenase: a target and generator of oxidative stress. *Philos. Trans. R. Soc. B Biol. Sci.*, **360**, 2335–2345.
- Tsitkov, S. *et al.* (2018) Queueing theory-based perspective of the kinetics of “channeled” enzyme cascade reactions. *ACS Catalysis*, **8**, 10721–10731.
- Voit, E.O. (2017) The best models of metabolism. *Wiley Interdisc. Rev. Syst. Biol. Med.*, **9**, e1391.
- Wu, F. *et al.* (2007) Computer modeling of mitochondrial tricarboxylic acid cycle, oxidative phosphorylation, metabolite transport, and electrophysiology. *J. Biol. Chem.*, **282**, 24525–24537.
- Zhang, W. *et al.* (2018) Liquid chromatography–tandem mass spectrometry method revealed that lung cancer cells exhibited distinct metabolite profiles upon the treatment with different pyruvate dehydrogenase kinase inhibitors. *J. Proteome Res.*, **17**, 3012–3021.



OPEN

Queueing theory model of pentose phosphate pathway

Sylwester M. Kloska¹✉, Krzysztof Pałczyński², Tomasz Marciniak², Tomasz Talaśka², Marissa Miller³, Beata J. Wysocki⁴, Paul Davis⁴ & Tadeusz A. Wysocki^{2,3}✉

Due to its role in maintaining the proper functioning of the cell, the pentose phosphate pathway (PPP) is one of the most important metabolic pathways. It is responsible for regulating the concentration of simple sugars and provides precursors for the synthesis of amino acids and nucleotides. In addition, it plays a critical role in maintaining an adequate level of NADPH, which is necessary for the cell to fight oxidative stress. These reasons prompted the authors to develop a computational model, based on queueing theory, capable of simulating changes in PPP metabolites' concentrations. The model has been validated with empirical data from tumor cells. The obtained results prove the stability and accuracy of the model. By applying queueing theory, this model can be further expanded to include successive metabolic pathways. The use of the model may accelerate research on new drugs, reduce drug costs, and reduce the reliance on laboratory animals necessary for this type of research on which new methods are tested.

In recent years, there has been significant progress in metabolomics. New and improved test methods allow for the measurement of many important biochemical parameters. The acquired data can be used to create simulation models of biochemical reactions and entire metabolic pathways. Queueing theory can successfully model metabolic processes, as demonstrated by the example of the glycolysis pathway¹ and Krebs cycle². The preparation of an accurate model simulating the course of PPP could potentially reduce the time needed for drug testing and reduce the number of laboratory animals on which new drugs are tested³.

The PPP is a metabolic pathway whose main substrate is glucose-6-phosphate (G6P). Throughout the reactions that make up this pathway, numerous molecules are formed, such as: nicotinamide adenine dinucleotide phosphate (NADPH), which is used in the biosynthesis of fatty acids, ribose 5-phosphate (R5P), which is a precursor in the synthesis of nucleotides, and erythrose 4-phosphate (E4P), which is used in the synthesis of aromatic amino acids^{4,5}. Products of the PPP are essential for the formation of new cells. However, under stress, cell growth is slowed down and the PPP is responsible for maintaining cellular levels of NADPH. In fact, such conditions increase the reliance of the PPP in the cell over glycolysis to maintain the needed ratio between NADP⁺ and NADPH⁶. In most living organisms, this pathway takes place in the cell cytosol.

There are two phases in the PPP: the oxidative phase and the non-oxidative phase. During the oxidative phase, NADPH is produced⁷. In the non-oxidative phase, various simple sugars are synthesized. 5-carbon sugars derived from the digestion of nucleic acids can be utilized in the PPP, where their carbon backbones are metabolized into intermediates for glycolysis or gluconeogenesis. In the non-oxidative phase, one of the enzymes—transketolase—is responsible for catalyzing two different reactions, with two sets of substrates. Therefore, these substrates act as inhibitors to each other, since they are competing for the same active site of the enzyme.

It is estimated that as much as 60% of NADPH comes from the PPP⁸. The PPP is most active in the liver, adrenal cortex, and mammary glands. The activity for this pathway is high in red blood cells, making it extremely important in erythrocytes⁹. NADPH formed by the PPP is used in the cell to prevent oxidative stress and the formation of dangerous free radicals that could harm the cell¹⁰. Reactive oxygen species (ROS) can damage cellular lipids, proteins, and nucleic acids, and eventually cause cell death¹¹. It is worth noting that ROS are associated with many diseases^{12–14}. Since erythrocytes do not have mitochondria, they have no other source of reducing oxidative stress other than the PPP. For example, large amounts of NADPH generated in erythrocytes are used to reduce glutathione (GSH). This reduced form of GSH is essential for maintaining the proper state of the cell. If GSH level is lowered in erythrocytes, hemolysis may occur¹⁵.

¹Faculty of Medicine, Nicolaus Copernicus University Ludwik Rydygier Collegium Medicum, 85-094 Bydgoszcz, Poland. ²Faculty of Telecommunications, Computer Science and Electrical Engineering, Bydgoszcz University of Science and Technology, 85-796 Bydgoszcz, Poland. ³Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, Omaha, NE 68182, USA. ⁴Department of Biology, University of Nebraska at Omaha, Omaha, NE 68182, USA. ✉email: 503013@stud.umk.pl; twysocki2@unl.edu

| Metabolite | Initial conc. (literature) | Final conc. (model) | Standard deviation over mean (%) | Absolute difference | Relative difference (%) |
|--------------------------|----------------------------|----------------------|----------------------------------|----------------------|-------------------------|
| Glucose-6-P (G6P) | 0.0026 | 0.0026 | 3 | 0 | 0 |
| NADP ⁺ | 0.001 | 0.001 | 3 | 0 | 0 |
| NADPH | 0.0002 | 0.0002 | 3 | 0 | 0 |
| 6-P-gluconolactone (PGL) | 5×10^{-6} | 9.3×10^{-6} | 36 | 4.3×10^{-6} | 86 |
| 6-P-gluconate (6PG) | 0.018 | 0.019 | 2 | 0.001 | 5.5 |
| Ribulose-5-P (Ru5P) | 0.012 | 0.012 | 2 | 0 | 0 |
| Ribose-5-P (R5P) | 0.009 | 0.009 | 1 | 0 | 0 |
| Xylulose-5-P (X5P) | 0.018 | 0.018 | 1 | 0 | 0 |
| Glyceraldehyde-3-P (G3P) | 0.00234 | 0.00242 | 3 | 0.00008 | 3.4 |
| Sedoheptulose-7-P (S7P) | 0.068 | 0.062 | 1 | 0.006 | 8.8 |
| Erythrose-4-P (E4P) | 0.004 | 0.004 | 3 | 0 | 0 |
| Fructose-6-P (F6P) | 0.083 | 0.079 | 0 | 0.004 | 4.8 |

Table 1. Comparison of concentration data: literature and model (mmol/L). Calculated relative difference shows similarity of obtained results and literature data.

The most common approach used to model metabolic changes in a cell is to use Ordinary Differential Equations (ODE). For metabolic reactions, ODEs provide quantitative information on interactions that occur between metabolites in specific reactions taking place in the cell. Previously, ODEs have been successfully used in simulation studies of biochemical kinetics and biochemical connections^{16–18}. The authors in¹⁹ presented a PPP model based on ODEs. This approach was beneficial because it did not require complicated operations that strained the capabilities of computers in the past, resulting in lower computing power. However, the simplifications and assumptions made when using ODEs in metabolic simulations do not reflect the stochastic nature of cell biochemistry²⁰. The Chemical Master Equation (CME) was another approach used to model the stochasticity of biological reactions²¹. However, due to the complexity and computing requirements, networks based on these models cannot be too extensive. A relatively new approach to computational metabolic modeling is the use of queueing theory. Queueing theory has wide applications in telecommunications, but also in biological and medical science topics, such as modeling drug pharmacokinetics²² or HIV infectivity²³. Using this method, it was possible to accurately model a simple metabolic network and mimic the interactions between metabolites²⁴, as well as the Krebs cycle². A genetic algorithm was used to optimize the kinetic coefficients. A variety of AI methods can be used for this purpose, but genetic algorithm was chosen because it was used with success when modeling the Krebs cycle.

The aim of this work was to prepare a PPP model capable of tracking concentration changes of specific metabolites occurring in this pathway over time. Additionally, the usefulness of the genetic algorithm for finding values of the kinetic constants used in the model was confirmed². A genetic algorithm was used to find values corresponding to those in the literature.

Results

The generated model becomes stable within approximately one hour. Every second, there are 1000 simulations of each pathway reaction (or 1 simulation step per millisecond), averaged over 50 simulated cells. This number has been selected experimentally. However, the model is designed to vary this number depending on the needs of the researcher. Figure 1 shows concentration changes of individual metabolites over time. Due to the various conditions of the living cell, G6P and NADP are consumed faster or slower depending on the blood glucose level, since glucose is phosphorylated to G6P to stay inside the cell and prevent diffusion out of the cell. This affects the glucose level in the cell, so the variation of 10% was assumed. The variation level is an arbitrary choice; meaning it can be changed. The purpose for the use of variation is to represent the concentration fluctuations in the cell. For this model to reflect the flow of metabolites in the cell as accurately as possible, the so-called “balancing flow” was used^{1,2}. This feature allows for proper simulation of metabolite flow depending on the current needs of the cell (Fig. 2). Thus, the level of metabolites that occur in more than one metabolic pathway, e.g. F6P and G3P being part of the PPP and glycolysis, better mimics biological conditions. Table 1 presents the comparison of model generated data and literature data regarding concentration of individual metabolites.

PGL is rapidly hydrolyzed, so the practical equilibrium between G6P and 6PG is directed towards the formation of 6PG²⁵. Any existing PGL is almost immediately converted to 6PG, therefore the variance is very high. The relative difference of PGL is high because it depends on the measurement time. In the future, we intend to combine the PPP with the already developed Krebs cycle and glycolysis models, so the results of the PPP model are likely to be closer to the experimental results.

Due to the high demand of glucose and its metabolites by cancer cells, many drugs are aimed at blocking metabolic pathways that supply cancer cells with substances necessary for proliferation. The PPP is one of the pathways with significantly increased activity in neoplastic cells. Compared to healthy cells, the activity of the PPP

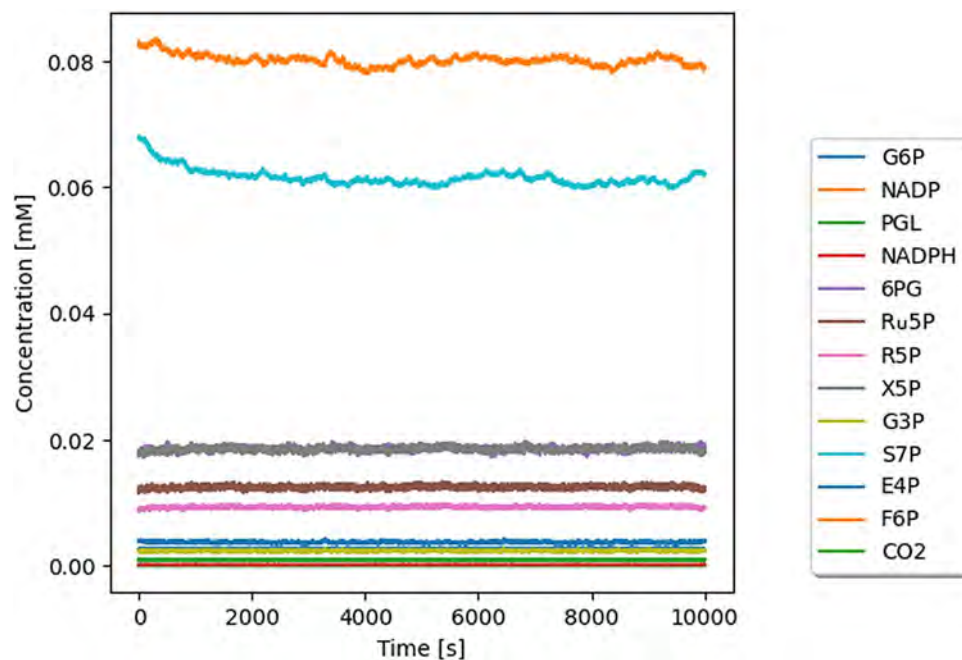


Figure 1. Concentration level change over time under unperturbed conditions. *G6P* glucose-6-phosphate, *NADP* NADP+, *PGL* 6-P-gluconolactone, *6PG* 6-phosphogluconate, *Ru5P* ribulose-5-phosphate, *R5P* ribose-5-phosphate, *X5P* xylulose-5-phosphate, *G3P* glyceraldehyde-3-phosphate, *S7P* sedoheptulose-7-phosphate, *E4P* erythrose-4-phosphate, *F6P* fructose-6-phosphate.

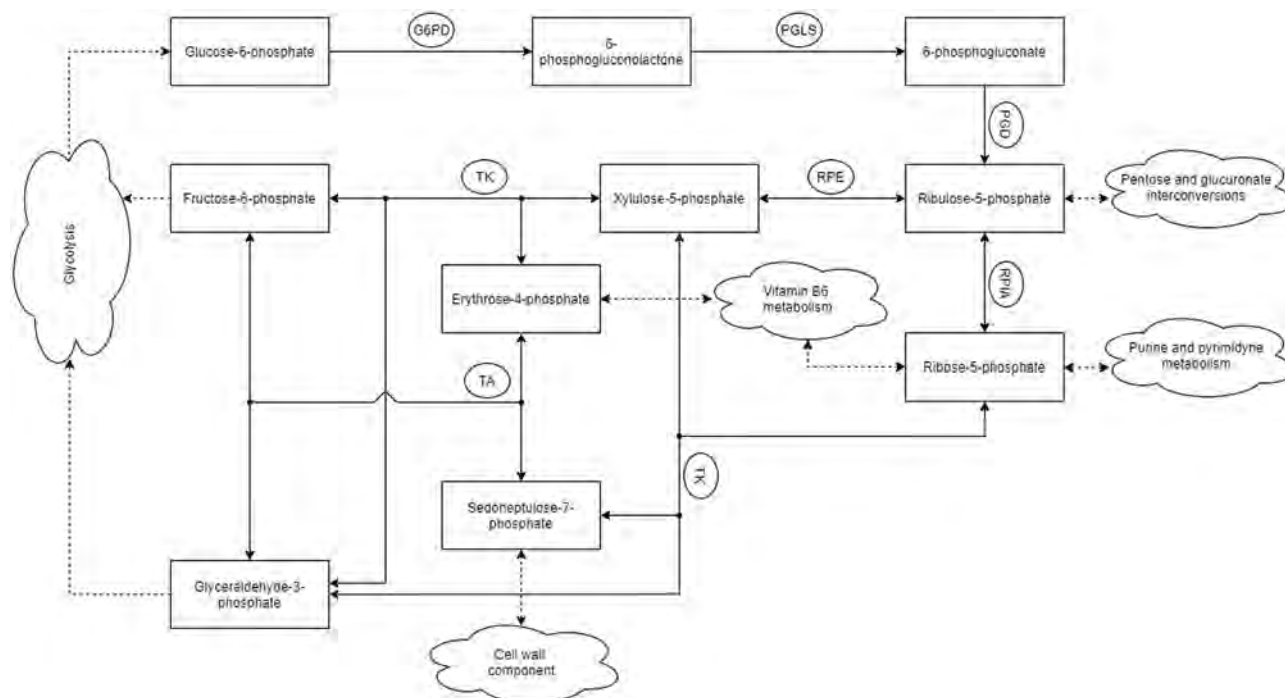


Figure 2. PPP scheme; the graph shows the main carbohydrate products, their relations with other metabolic pathways, and enzymes that catalyze reactions. *G6PD* glucose-6-phosphate dehydrogenase, *PGLS* 6-phosphogluconolactonase, *PGD* 6-phosphogluconate dehydrogenase, *RPIA* ribose-5-phosphate isomerase A, *RPE* ribulose-5-phosphate-3-epimerase, *TA* transaldolase, *TK* transketolase.

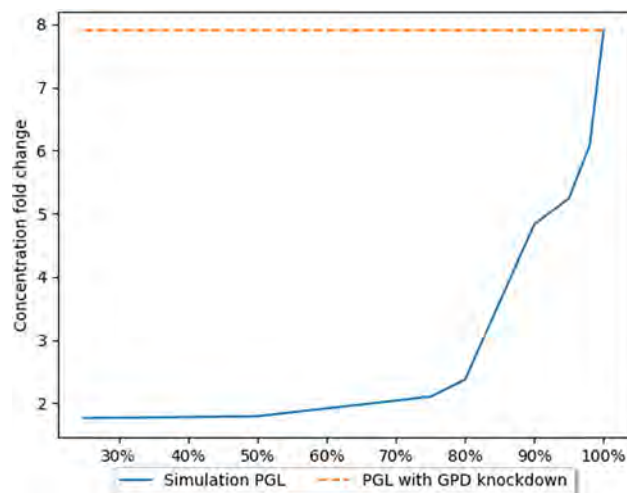


Figure 3. The effects of GPD gene expression knockdown on PGL concentration²⁶. The X axis presents level of simulated GPD inhibition. The Y axis presents fold change in concentration in comparison to the natural state (without inhibition).

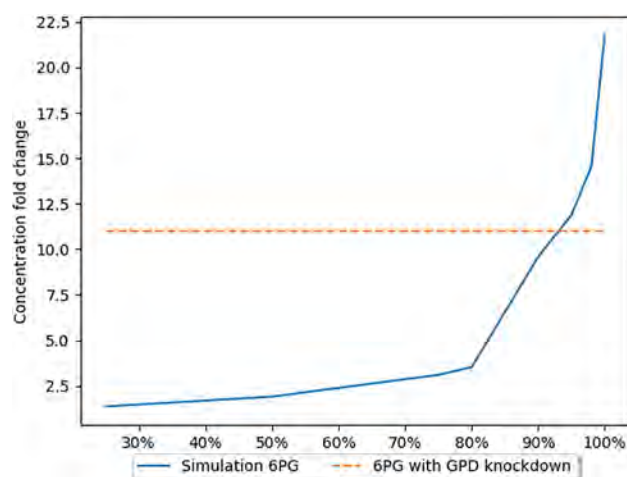


Figure 4. The effects of GPD gene expression knockdown on 6PG concentration²⁶. The X axis presents level of simulated GPD inhibition. The Y axis presents fold change in concentration in comparison to the natural state (without inhibition).

in cancer cells can be increased up to 8 times. The oxidative part of the pathway provides cells with a large amount of NADPH, helping the cell can effectively fight excess oxidative stress. Effects that reduce the effectiveness of the production of NADPH in the cell, in combination with factors that induce this stress, such as radiotherapy or chemotherapy, can kill cancer cells.

To validate the model, model results were compared to those obtained empirically. The paper²⁶ serving as the benchmark for our model described the effect of a third PPP enzyme, PGD, in lung cancer cells. Inhibition of this enzyme's activity does not significantly affect the level of NADPH, but inhibits tumor growth. The gene encoding PGD is characterized by increased expression in neoplastic cells. ShRNA molecules were used to reduce PGD expression. This approach resulted in inhibition of tumor growth, indicating an important role for PGD in cancer cell metabolism. Concentrations of several PPP metabolites were measured, however, not all of them had significant changes. Metabolites of the oxidative phase of the PPP, such as 6-phosphogluconolactone (PGL) and 6-phosphogluconate (6PG) had concentrations 7.9 and 11 times higher than their regular concentrations, respectively (Figs. 3 and 4). These metabolites accumulated due to the absence/decreased activity of PGD. The concentrations of metabolites of the non-oxidative phase of the pathway such as S7P or X5P were not measured, but no significant changes in the concentrations of ribose phosphate and nucleotide triphosphate were detected.

The accumulation of metabolites preceding the blocked reaction is because the expression of the PGD enzyme has been reduced. A bottleneck is created at this stage of the pathway, leading to a reduced efficiency of this stage, as there are not enough protein molecules in the cell to process all metabolite molecules. As a further consequence, a decrease in the concentration of metabolites occurring further down the pathway, e.g., G3P, can be

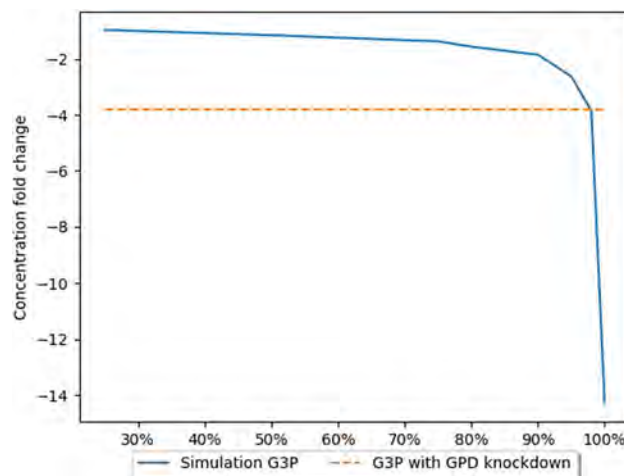


Figure 5. The effects of GPD gene expression knockdown on G3P concentration²⁶. The X axis presents level of simulated GPD inhibition. The Y axis presents fold change in concentration in comparison to the natural state (without inhibition).

| Metabolite | Experimental data concentration change ²⁶ | Model data concentration change using 90% inhibition | Model data concentration change using 95% inhibition | Model data concentration change using 98% inhibition | Model data concentration change using 100% inhibition |
|------------|--|--|--|--|---|
| G6P | 1.8 | 1.8 | 1.8 | 1.8 | 1.8 |
| PGL | 7.9 | 4.83 | 5.24 | 6.08 | 7.89 |
| 6PG | 1 | 9.59 | 11.88 | 14.56 | 21.8 |
| G3P | -3.8 | -1.85 | -2.63 | -3.85 | -14.29 |

Table 2. Comparison of metabolite concentration changes (fold changes) caused by knockdown of the PGD gene.

observed (Fig. 5). For the validation of the model, measurements of the concentrations of metabolites obtained empirically were used. The model makes it possible to simulate and track the changes in the concentrations of the metabolites.

Several measurements were performed to evaluate the level of inhibition of the GPD catalyzed reaction. The obtained results show that the GPD knockdown caused inhibition at the level of 95–98%. These assumptions are based on the results presented in Table 2. The results for these inhibition levels are the closest to the empirical results. The paper²⁶ used shRNA to achieve expression knockdown, which is an incomplete mechanism to reduce (but not eliminate) expression. This form of knockdown is not expected to achieve 100% silencing. Indeed, 80–99% knockdown of expression is normal and expected. The calculated results are comparable to those obtained experimentally and are consistent with current biological knowledge. Another point to consider is that the glucose metabolism of neoplastic cells remains unknown in some aspects and these cells may possibly bypass a blocked reaction in the metabolic pathway. Simulations using 100% inhibition were also performed, but this led to a significant reduction in the concentration of metabolites downstream of the bottleneck of the pathway. However, it can be observed that due to the bidirectional character of reactions of the second phase of the PPP and the flux of metabolites from other pathways, e.g., F6P generated in glycolysis, we do not observe a complete ‘zeroing’ of metabolite concentration.

The results generated in our model (Table 2) follows the trend of changes in concentration observed in vitro, and suggests that knockdown efficiency in vitro was likely near 95%, which is common for shRNA expression knockdown.

Discussion

As mentioned in the introduction, most previous models simulating metabolic pathways, not only PPP, have been based on the use of ODEs. However, due to the advantages offered by queueing theory, it seems reasonable to use this method in modeling. The preparation of a quantitative model of a biological pathway such as the PPP requires the necessary information on starting concentrations and kinetic data of the enzymes that catalyze the pathway reactions. The presented model can be viewed as a ‘virtual laboratory’. This model tracks the relationships between individual metabolites formed at different stages of the pathway. It is possible to observe changes caused by fluctuations in metabolite concentrations and their impact on the entire pathway.

It can also be used to test the effectiveness of new drugs if their influence on the kinetics of the reaction they affect is known. In this way, one can also theoretically get answers to questions such as which reactions are worth blocking to obtain the best possible therapeutic result. Most studies aimed at blocking the PPP pathway

| Number | Reaction | Enzyme |
|--------|--|-----------------------------------|
| 1 | $G6P + NADP^+ \rightarrow PGL + NADPH + H^+$ | Glucose 6-phosphate dehydrogenase |
| 2 | $PGL + H_2O \rightarrow 6PG + H^+$ | 6-Phosphogluconolactonase |
| 3 | $6PG + NADP^+ \rightarrow Ru5P + NADPH + H^+ + CO_2$ | 6-Phosphogluconate dehydrogenase |
| 4A | $Ru5P \rightarrow R5P$ | Ribose-5-phosphate isomerase |
| 4B | $Ru5P \rightarrow X5P$ | Ribulose 5-phosphate 3-epimerase |
| 5 | $R5P + X5P \rightarrow G3P + S7P$ | Transketolase |
| 6 | $X5P + E4P \rightarrow G3P + F6P$ | Transketolase |
| 7 | $G3P + S7P \rightarrow E4P + F6P$ | Transaldolase |

Table 3. Stoichiometric reactions of the PPP. Reactions 1-3 form the oxidative branch of PPP, reactions 4-7 are in the non-oxidative branch.

in cancer patients have focused on blocking the first reaction of the pathway catalyzed by G6PD^{27,28}. However, clinical data indicate that this therapy is not very effective without additional exposure to oxidative stress^{29,30}. For this reason, the results of studies on the knockdown of the gene encoding PGD in this paper were used²⁶. Even though the knockdown of the G6PD gene does not affect the amount of NADPH, which is important for tumor development, the knockdown of this gene alone results in inhibition of tumor growth. Perhaps the metabolites that accumulate in the cell prior to the blocked reaction are responsible for this situation. Their concentration in cells reaches values significantly greater than their natural concentrations. The exact mechanism of tumor growth inhibition is unknown, however, the effect achieved is important.

The proposed model obtained stability based on the data from the above-mentioned paper. We believe that this type of model can be used to predict the impact of therapy, which in turn will lead to an increase in its effectiveness.

Thanks to the use of experimental data together with a computational process based on the queueing theory, a model was obtained that can track the metabolic pathway that takes place in the cells of living organisms. In this paper, we present a separate PPP model without detailed analysis of the relationship between PPP and glycolysis. The metabolites common to both pathways have been identified and several principles have been adopted to create a functional PPP model. In the future, our plan is to connect the existing glycolysis, Krebs cycle, and PPP models together. We believe that such a procedure may also positively affect the consistency of simulation and experimental results.

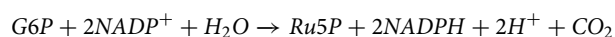
The presented results indicate that the model can be used to predict changes in metabolite concentrations. For this purpose, it is sufficient to enter the concentration value of one of the metabolites. In this way, the entire study can prove to be more cost-effective—no need to determine each metabolite separately, which also saves time.

As demonstrated by the knockdown of one of the genes encoding the enzyme catalyzing the PPP reaction, this model is adapted to follow the trend of metabolite changes. Moreover, it can determine the specific effect of the inhibition of particular reactions on the concentration of metabolites with relatively high accuracy. Further research providing data on how inhibition of a particular pathway step may affect kinetic constants could contribute to an increase in the accuracy of the presented model.

Methods

Obtaining metabolic data and the use of enzymatic reaction kinetics. This work focuses on the reflection of changes in PPP metabolite concentrations over time. For this purpose, a literature review was carried out to provide data on these concentrations (Table 1). Presented concentrations were measured with the use of mass spectrometry³¹. Several kinetic constants, and enzymatic properties, like maximum velocity (V_{max}), necessary for the correct operation of the model were used to calculate the speed of chemical reactions³¹. Reaction rates were calculated using equations based on Michaelis–Menten kinetics (for more information please check Supplementary Information).

NADPH is formed from 2 NADP⁺ molecules in the oxidative phase. The energy generated during the conversion of G6P into ribulose 5-phosphate (Ru5P) is used in the reaction. The overall reaction of the first phase of the pathway is as follows:



Ru5P, which is one of the products of the first phase of the PPP, is the first substrate for the non-oxidative phase. Ribose-5-phosphate isomerase can convert Ru5P to R5P. On the other hand, ribulose 5-phosphate epimerase converts Ru5P to xylulose 5-phosphate (X5P). The next reactions involve changing the length of the carbon chain in the carbohydrates. These two five-carbon sugars then undergo a transketolase-catalyzed reaction. The result is production of glyceraldehyde 3-phosphate (G3P) and sedoheptulose 7-phosphate (S7P). Then G3P and S7P undergo a transaldolase-catalyzed reaction, which produces E4P and fructose 6-phosphate (F6P) (Fig. 2; Table 3).

Queueing theory. The complicated nature of metabolic pathways, in which there are huge amounts of biochemical substances constituting the substrates and reaction products, makes modeling metabolism extremely

challenging. Methods commonly used to model metabolic pathways require supervision and the use of appropriate constraints, like forcing ODEs not to reach negative values. Such treatments may cause small calculation errors which could accumulate in long-term modeling and result in incorrect calculations. Biological systems are organized to pass the products of individual metabolic reactions further down the pathway, so that they become substrates for downstream reactions or are used by the cell to support necessary life processes³². For this reason, the use of queueing theory in metabolic pathway modeling seems to be the right approach.

Queueing networks can be thought of as hidden Markov chains, similar to Gillespie's modelling technique^{20,21}. The advantage of using queueing theory to model metabolic pathways is that they do not require enhanced computing power. Therefore, the results can be obtained close to real time. Networks based on queueing theory can be applied with ease to a significantly greater number of molecules, grouped into the queues representing different molecular species. Due to the nature of this approach, it is capable of combining individual pathways into larger, more complex groups of metabolic pathways.

Averaging the results from several simulation runs provides information on the average changes in the concentrations of the individual pathway metabolites. The proposed model is based on calculations of the kinetics of Michaelis–Menten enzymatic reactions, which focus on the relationship between the concentrations of the substrate and the product, and the velocity of the reaction. According to this theory, the macroscopic concept of enzymatic reaction speed is the sum of many microscopic reactions that can exchange specific amounts of substances per time unit. The description of the PPP as the probability of decreasing and increasing the concentration of each of the substances present in the pathway and the correlation of their reduction with the accumulation of other substrates results in a self-regulating, stochastic process that imitates the actual course of the PPP. The Michaelis–Menten kinetic equation was used to calculate the probability of the reaction. A detailed description of the methodology used is described in the work describing the Krebs cycle model².

The probability of the reaction can be converted to an average amount of arrivals when measured for a significant amount of time. Therefore, the kinetic equations can be used to calculate the adaptive parameter $\mu(t)$ utilized for modelling PPP behavior by a network of inhomogeneous Poisson processes described by equation (1):

$$P[(N(t + \tau) - N(t)) = k, t] = \frac{e^{-\mu(t)\tau} (\mu(t)\tau)^k}{k!} \quad (1)$$

Where:

$$P(N(t + \tau) - N(t)) = k, t] \text{—probability of } k \text{ arrivals in the interval } (t, t + \tau)$$

$$\mu(t)\tau \text{—expected number of arrivals in a time interval duration of } (t, t + \tau)$$

The queue processing time of metabolite increment is described by the exponential distribution of the random variable T in the terms of the rate parameter $\mu(t)$ as follows (2):

$$f(T; \mu(t)) = \begin{cases} \mu(t)e^{-\mu(t)T} & \text{when } T \geq 0 \\ 0 & \text{when } dT < 0 \end{cases} \quad (2)$$

Therefore, the PPP is modelled by the composition of interconnected queues. Departure of substrate's increment from one queue is followed by the arrival at the successive queue. It is worth noting that the network of interconnected queues is equivalent to the set of ODEs as proven by³³.

Probability of substrate's increment departure from each queue depends on the current concentration of the substrates and the kinetic constants of the reaction causing that departure. Every queue uses its individual Michaelis–Menten kinetic equation with kinetic constants normalized according to the method based on the formula described in¹, to determine the likelihood that in this time step the reaction occurs. Since the reaction rates depend on the current concentration of molecules that change from step to step, the resulting inhomogeneous Poisson process implements the feedback loop, which results in a system with memory.

Use of a genetic algorithm to optimize model parameters. Values of enzyme kinetic constants were found with the use of a genetic algorithm starting from literature data. Every 'gene' in the 'chromosome' is a vector of kinetic constants describing each Michaelis–Menten kinetic equation. The new values of kinetic constants are found by randomly selecting from which 'parent' 'offspring' inherits 'gene' (set of kinetic constants for a particular reaction). However, mutation occurs on each kinetic value regardless to which parent it belongs. The loss function optimized by the algorithm is the sum of the squared distances between PPP state described by the literature and the current optimization step of the simulation using kinetic constants that makes an individual 'chromosome'. The formula of loss function is as follows (3):

$$f(X_I; X) = \sum (X_I - X)^2 \quad (3)$$

Where: X_I —vector of substrates described by a literature; X —vector of substrates describing stable state of simulation.

The loss function described above has a trivial solution. If all kinetic constants that are used in Michaelis–Menten reactions as multipliers (instead of dividers) are zeroed, then the results of these equations are equal to zero. As a result, no reactions occur, so the simulation's stable point is equal to the original literature vector. To prevent such a solution, the genetic algorithm sets a constraint on newly generated 'chromosomes'. Each reaction parametrized by values of the 'chromosome' for a literature vector of substrates must have a probability of occurrence between 0.00005 and 0.05.

The first set of ‘chromosomes’ are made of Michaelis–Menten kinetic constants defined in the literature with added gaussian noise. Given the selected starting point, the genetic algorithm is set on finding the optimal value in the proximity of the already established values. This reduces the risk of the algorithm generating an output that minimizes the loss function, but produces kinetic constants significantly different from the literature values.

Pseudocode of the PPP model. The pseudocode describing the computational processes can be found in the Supplementary Information. This code assumes that:

- Kinetic constants are grouped into a table of vectors of constant values. There are 14 vectors in the table corresponding to eight different reactions and six balancing flows. Each of the reactions has a unique vector of dimension equal to the number of kinetic constants used in the reaction rate computation and every balancing flow contains a one-dimensional vector.
- Concentration increment exchanged during the reactions is denoted ‘delta’ and is unique for each reaction. It ranges from 2.3×10^{-6} mM to 5.0×10^{-5} mM. ‘Delta’ is significantly lower than the initial value of the lowest substrate concentration. The ‘delta’ value must be chosen in a way that corresponds to a change of more than a single molecule for the rare species; in fact, for rare species, it should always be a positive integer number of molecules.
- the concentrations of G6P and NADP in the cycle vary with 10% Gaussian noise around the constant values of 0.001 mM and 0.0026 mM, respectively. This signal-to-noise ratio aims to reflect metabolic conditions inside the cell.

The search for optimal kinetic constants was performed using a PC with AMD Ryzen 7 3800X 8-Core Processor, 3900 MHz, RAM 32 GB. Code was written in C# 8.0. One search epoch simulating one hour for 50 different tables of kinetic constants using all 8 logic cores, took approximately 7 hours.

Model validation based on the use of experimental data. G6P dehydrogenase is the enzyme that catalyzes the first reaction of the pathway³⁴. Therefore, it is the enzyme that controls the starting velocity of the pathway. This enzyme is strongly inhibited by NADPH³⁵. Drugs aimed at reducing the intensity of the reaction mainly focus on reducing the activity of this enzyme, which leads to a reduction in the velocity of the entire pathway²⁷. However, clinical results indicate that inhibiting this enzyme is not an effective therapeutic approach²⁶. For this reason, data obtained from the study of knockdown expression of the 6-phosphogluconate dehydrogenase (PGD) enzyme were selected for model validation²⁶.

Data availability

The dataset supporting the conclusions of this article is available in the GitHub repository, <https://github.com/UTP-WTliE/PPPQueueingTheory>.

Received: 30 November 2021; Accepted: 8 March 2022

Published online: 17 March 2022

References

1. Clement, E. J. *et al.* Stochastic simulation of cellular metabolism. *IEEE Access Pract. Innov. Open Solut.* **8**, 79734–79744. <https://doi.org/10.1109/access.2020.2986833> (2020).
2. Kloska, S. *et al.* Queueing theory model of Krebs cycle. *Bioinformatics* **37**, 2912–2919. <https://doi.org/10.1093/bioinformatics/btab177>. <https://academic.oup.com/bioinformatics/article-pdf/37/18/2912/40471271/btab177.pdf> (2021).
3. Hajar, R. Animal testing and medicine. *Heart Views* **12**, 42 (2011).
4. Gonzalez, S. N., Valsecchi, W. M., Maugeri, D., Delfino, J. M. & Cazzulo, J. J. Structure, kinetic characterization and subcellular localization of the two ribulose 5-phosphate epimerase isoenzymes from *Trypanosoma cruzi*. *PLoS One* **12**, 1–27. <https://doi.org/10.1371/journal.pone.0172405> (2017).
5. Gumaa, K. & McLean, P. The pentose phosphate pathway of glucose metabolism. enzyme profiles and transient and steady-state content of intermediates of alternative pathways of glucose metabolism in Krebs ascites cells. *Biochem. J.* **115**, 1009–1029. <https://doi.org/10.1042/bj1151009> (1969).
6. Ralser, M. *et al.* Dynamic rerouting of the carbohydrate flux is key to counteracting oxidative stress. *J. Biol.* **6**, 10–10 (2007).
7. Moritz, B., Striegel, K., De Graaf, A. & Sahm, H. Kinetic properties of the glucose-6-phosphate and 6-phosphogluconate dehydrogenases from *Corynebacterium glutamicum* and their application for predicting pentose phosphate pathway flux in vivo. *Eur. J. Biochem.* **267**, 3442–3452. <https://doi.org/10.1046/j.1432-1327.2000.01354.x> (2000).
8. Caillaud, M. & Paul Quick, W. New insights into plant transaldolase. *Plant J.* **43**, 1–16. <https://doi.org/10.1111/j.1365-313X.2005.02427.x>. <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-313X.2005.02427.x> (2005).
9. Soldin, S. J. & Balinsky, D. Kinetic properties of human erythrocyte glucose 6-phosphate dehydrogenase. *Biochemistry* **7**, 1077–1082 (1968).
10. Zuurbier, C. *et al.* Inhibition of the pentose phosphate pathway decreases ischemia–reperfusion-induced creatine kinase release in the heart. *Cardiovasc. Res.* **62**, 145–153 (2004).
11. Auten, R. L. & Davis, J. M. Oxygen toxicity and reactive oxygen species: The devil is in the details. *Pediatr. Res.* **66**, 121–127 (2009).
12. Alfadda, A. A. & Sallam, R. M. Reactive oxygen species in health and disease. *J. Biomed. Biotechnol.* **2012** (2012).
13. Brieger, K., Schiavone, S., Miller, F. & Krause, K.-H. Reactive oxygen species: From health to disease. *Swiss Med. Wkly.* **142**, w13659 (2012).
14. Görlach, A. *et al.* Reactive oxygen species, nutrition, hypoxia and diseases: Problems solved?. *Redox Biol.* **6**, 372–385 (2015).
15. Giustarini, D., Dalle-Donne, I., Colombo, R., Milzani, A. & Rossi, R. Interference of plasmatic reduced glutathione and hemolysis on glutathione disulfide levels in human blood. *Free Radic. Res.* **38**, 1101–1106 (2004).
16. Ahn, E., Kumar, P., Mukha, D., Tzur, A. & Shlomi, T. Temporal fluxomics reveals oscillations in TCA cycle flux throughout the mammalian cell cycle. *Mol. Syst. Biol.* **13**, 953 (2017).

17. Cohen, D. M. & Bergman, R. N. Estimation of TCA cycle flux, aminotransferase flux, and anaplerosis in heart: Validation with syntactic model. *Am. J. Physiol. Endocrinol. Metab.* **268**, E397–E409 (1995).
18. Mogilevskaya, E., Demin, O. & Goryanin, I. Kinetic model of mitochondrial Krebs cycle: Unraveling the mechanism of salicylate hepatotoxic effects. *J. Biol. Phys.* **32**, 245–271 (2006).
19. Messiha, H. *et al.* Enzyme characterisation and kinetic modelling of the pentose phosphate pathway in yeast. *PeerJ* (2014).
20. Voit, E. The best models of metabolism. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **9** (2017).
21. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).
22. Guo, D. *et al.* Endosomal trafficking of nanoformulated antiretroviral therapy facilitates drug particle carriage and HIV clearance. *J. Virol.* **88**, 9504–9513 (2014).
23. Sharp, A. T., Pannier, A. K., Wysocki, B. J. & Wysocki, T. A. A novel telecommunications-based approach to HIV modeling and simulation. *Nano Commun. Netw.* **3**, 129–137 (2012).
24. Evstigneev, V. P., Holyavka, M. G., Khrapaty, S. V. & Evstigneev, M. P. Theoretical description of metabolism using queueing theory. *Bull. Math. Biol.* **76**, 2238–2248 (2014).
25. Eggleston, L. V. & Krebs, H. A. Regulation of the pentose phosphate cycle. *Biochem. J.* **138**, 425–435 (1974).
26. Sukhatme, V. P. & Chan, B. Glycolytic cancer cells lacking 6-phosphogluconate dehydrogenase metabolize glucose to induce senescence. *FEBS Lett.* **586**, 2389–2395 (2012).
27. Ghergurovich, J. M. *et al.* A small molecule g6pd inhibitor reveals immune dependence on pentose phosphate pathway. *Nat. Chem. Biol.* **16**, 731–739 (2020).
28. Preuss, J. *et al.* Identification and characterization of novel human glucose-6-phosphate dehydrogenase inhibitors. *J. Biomol. Screen.* **18**, 286–297 (2013).
29. Lin, C.-J. *et al.* Impaired dephosphorylation renders g6pd-knockdown hepg2 cells more susceptible to H₂O₂-induced apoptosis. *Free Radic. Biol. Med.* **49**, 361–373 (2010).
30. Polat, I. H. *et al.* Oxidative pentose phosphate pathway enzyme 6-phosphogluconate dehydrogenase plays a key role in breast cancer metabolism. *Biology* **10**, 85 (2021).
31. Sabate, L., Franco, R., Canela, E. L., Centelles, J. J. & Cascante, M. A model of the pentose phosphate pathway in rat liver cells. *Mol. Cell. Biochem.* **142**, 9–17 (1995).
32. Tsitkov, S., Pesenti, T., Palacci, H., Blanchet, J. & Hess, H. Queueing theory-based perspective of the kinetics of “channeled” enzyme cascade reactions. *ACS Catal.* **8**, 10721–10731 (2018).
33. Massey, W. A. Asymptotic analysis of the time dependent M/M/1 queue. *Math. Oper. Res.* **10**, 305–327 (1985).
34. Adediran, S. Kinetic and thermodynamic properties of two electrophoretically similar genetic variants of human erythrocyte glucose-6-phosphate dehydrogenase. *Biochimie* **78**, 165–170 (1996).
35. Kanji, M. I., Toews, M. & Carper, W. A kinetic study of glucose-6-phosphate dehydrogenase. *J. Biol. Chem.* **251**, 2258–2262 (1976).

Acknowledgements

This work was funded by the National Science Center (NCN) of Poland in terms of Opus-17 Program with grant number 2019/33/B/ST6/00875, and NIH GM103427.

Author contributions

S.K., T.W. and B.W. conceived the idea for the model and described the theoretical background. K.P. and M.M. implemented the algorithms, under the supervision of T.T. and T.M. T.W., T.M. and P.D. supervised the project and coordinated the research team. S.K. and K.P. wrote the paper. All authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-08463-y>.

Correspondence and requests for materials should be addressed to S.M.K. or T.A.W.

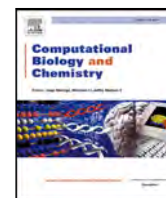
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022



Research Article

Conversion of fat to cellular fuel—Fatty acids β -oxidation model

Sylwester M. Kloska^{a,*}, Krzysztof Pałczyński^b, Tomasz Marciniak^b, Tomasz Talaśka^b,
Marissa Miller^c, Beata J. Wysocki^d, Paul Davis^d, Tadeusz A. Wysocki^{b,c}

^a Department of Forensic Medicine, Nicolaus Copernicus University Ludwik Rydygier Collegium Medicum, Bydgoszcz, Poland

^b Faculty of Telecommunications, Computer Science and Electrical Engineering, Bydgoszcz University of Science and Technology, Bydgoszcz, Poland

^c Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, Omaha, USA

^d Department of Biology, University of Nebraska at Omaha, Omaha, USA

ARTICLE INFO

Dataset link: <https://github.com/UTP-WTiE/FattyAcidsOxidation>

Keywords:

Beta-oxidation
Enzyme kinetics
Fatty acid
Mathematical modeling
Michaelis–Menten
Queueing theory

ABSTRACT

β -oxidation of fatty acids plays a significant role in the energy metabolism of the cell. This paper presents a β -oxidation model of fatty acids based on queueing theory. It uses Michaelis–Menten enzyme kinetics, and literature data on metabolites' concentration and enzymatic constants. A genetic algorithm was used to optimize the parameters for the pathway reactions. The model enables real-time tracking of changes in the concentrations of metabolites with different carbon chain lengths. Another application of the presented model is to predict the changes caused by system disturbance, such as altered enzyme activity or abnormal fatty acid concentration. The model has been validated against experimental data. There are diseases that change the metabolism of fatty acids and the presented model can be used to understand the cause of these changes, analyze metabolites abnormalities, and determine the initial target of treatment.

1. Introduction

1.1. Biological background

β -oxidation of fatty acids plays a major role in the generation of accessible energy within the cell. β -oxidation leads to the shortening of the fatty acid carbon chain to generate acetyl-CoA and energetic coenzymes NADH and FADH₂, which are used in the respiratory chain and lead to adenosine triphosphate (ATP) formation (Houten and Wanders, 2010).

β -oxidation is a complex process involving several steps that can be divided into three phases: (1) transport of fatty acids from adipose tissue to target cells, (2) entry of the fatty acids into the cytoplasm and then into the mitochondria, including activation and transport inside target cell mitochondria, (3) oxidative catabolism inside the mitochondrial matrix (Houten and Wanders, 2010). In our model we focused mostly on the phases 2 and 3.

Activation is a necessary and energy-expensive process in the metabolism of fatty acids. Pre-existing ATP is hydrolyzed, providing the necessary energy to “activate” the fatty acid and form a fatty acyl-CoA molecule. This reaction is catalyzed by fatty acyl-CoA synthetase and prevents fatty acids from leaving the cell; however, this molecule

cannot cross into the mitochondrial membrane. Fatty acyl-carnitine is formed to enable transport into the inner mitochondrial matrix using the carnitine shuttle.

Once inside the mitochondrial matrix, fatty acyl-CoA molecule enter the β -oxidation pathway. Four main stages of β -oxidation can be distinguished: oxidation, hydration, oxidation, and thiolysis (Fig. 1). The two products of β -oxidation are (1) a fatty acyl-CoA molecule, shortened by two carbon atoms, which can re-enter the β -oxidation pathway, and (2) an acetyl-CoA molecule, which is the cell's fuel in the Krebs cycle. The concentration of acyl-CoA intermediates is difficult to measure due to low concentrations of these intermediate molecules, which also share similar chemical structure and mass. Blood acyl-carnitine concentration measurements can be used, which somehow reflect the mitochondrial acyl-CoA concentration (Mihalik et al., 2010; Noland et al., 2009).

1.2. Queueing theory

Advances in experimental techniques have enabled the creation of mathematical models of metabolic pathways. Computational models provide insight into the dynamics of changes taking place in the cell. These models can facilitate the understanding of complex metabolic

* Correspondence to: Curie Skłodowskiej 9 Street, 85-094 Bydgoszcz, Poland.
E-mail address: 503013@stud.umk.pl (S.M. Kloska).

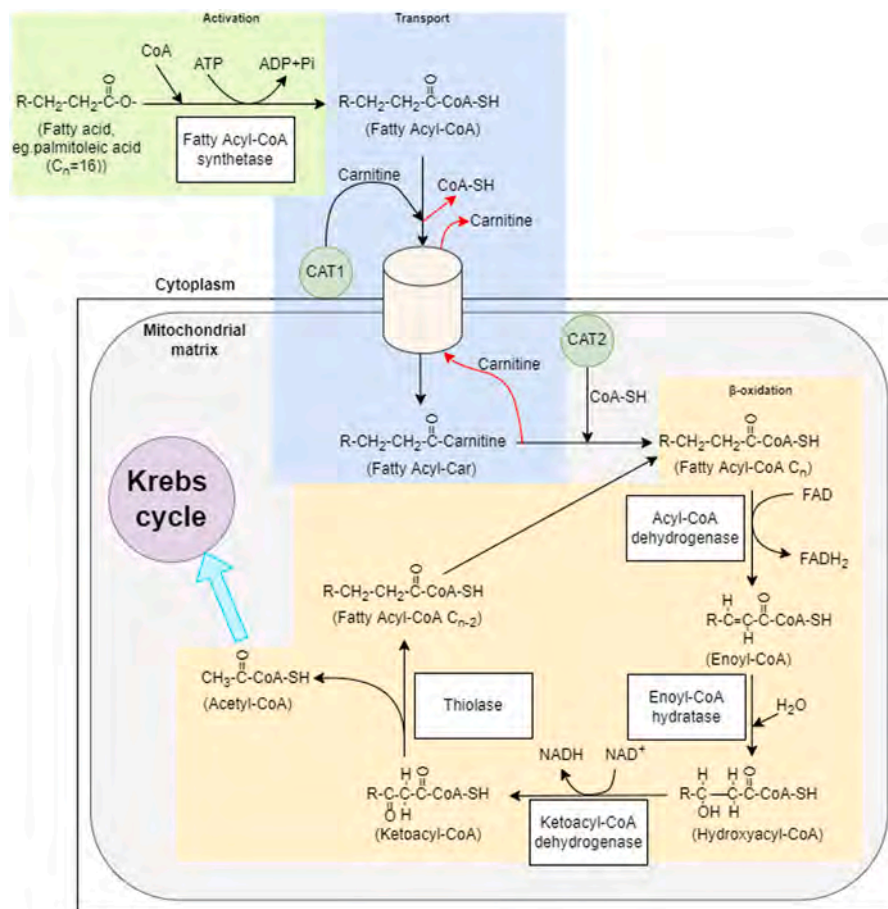


Fig. 1. Diagram illustrating the course of activation, transport into mitochondrial matrix, and β -oxidation.

networks, provide useful information about biological systems' behavior, as well as potentially identify new therapeutic targets which would lead to the development of personalized medicine (Schützhold et al., 2016; van Riel, 2006).

Most of the biological models developed so far are based on Ordinary Differential Equations (ODEs) (van Eunen et al., 2013). ODE-based methods are well known and useful for metabolic pathway modeling but are not without drawbacks. Due to the computational nature of ODEs, during long simulations they may result in the accumulation of numerical errors and require the use of a non-negative ODE solution (Shampine et al., 2005).

The inherent stochasticity of biological processes should be considered when designing computational models of such systems. For this reason, stochastic models can be constructed on the basis of the Chemical Master Equation (CME), which make it possible to demonstrate stochastic processes in biological systems. However, the disadvantage of using the CME in stochastic model construction is that it is not suitable for modeling complex metabolic networks due to computational complexity (Puchałka and Kierzek, 2004). For this reason, it is necessary to look for methods that are lighter in terms of computation, and still capable of modeling stochastic biological processes.

One method is the application of queueing theory. This method allows for the grouping of molecules of the same type into a single queue and reactions of the same type into a single server (Evstigneev, 2014). This modeling approach makes the computations simpler compared to models that use the Gillespie algorithm (Gillespie, 1977; Voit, 2017), where each molecule and reaction is described as a separate node in the Markov chain (Massey, 1985). The model based on queueing theory can be considered a hidden Markov chain model and the resulting

mathematical model is similar to the biological one (Fig. 2). Given the simplified modeling approach and its less complex computations, it has been applied to many diverse biological processes including the HIV infection process (Sharp et al., 2012), pharmacokinetic modeling (Guo et al., 2014), glycolysis (Clement et al., 2020), the Krebs cycle (Kloska et al., 2021), and pentose phosphate pathway (Kloska et al., 2022).

Observing changes in the concentration of individual metabolites in a queueing theory-based model is possible because the results of simulations are averaged over the course of several simulation runs. The model is based on the Michaelis–Menten enzymatic equation, which describes the relationships between reaction rates and substrate–product pairs. In this sense, the speed of a reaction is a macroscopic representation of many microscopic reactions that cause/prevent the exchange of a given amount of a substance from one queue to another in a given time. From this point of view, the reaction rate can be seen as the frequency of the reaction and the probability of increasing the amount of a given molecule, in a reaction in which the amount of substrate is reduced in favor of an increasing amount of product. Using this approach, we were able to obtain a self-regulating model of β -oxidation of fatty acids that reflects the stochastic nature of the process. The Michaelis–Menten equation is used to determine the probability of a reaction occurring based on the kinetic data used in the equation, i.e., metabolite concentration, maximum enzyme velocity. A more detailed description can be found in Kloska et al. (2021).

In addition, we use one of the artificial intelligence (AI) algorithms, specifically the genetic algorithm (GA), to optimize the parameters of the model. We decided to use the GA because it was effective in our previous modeling models.

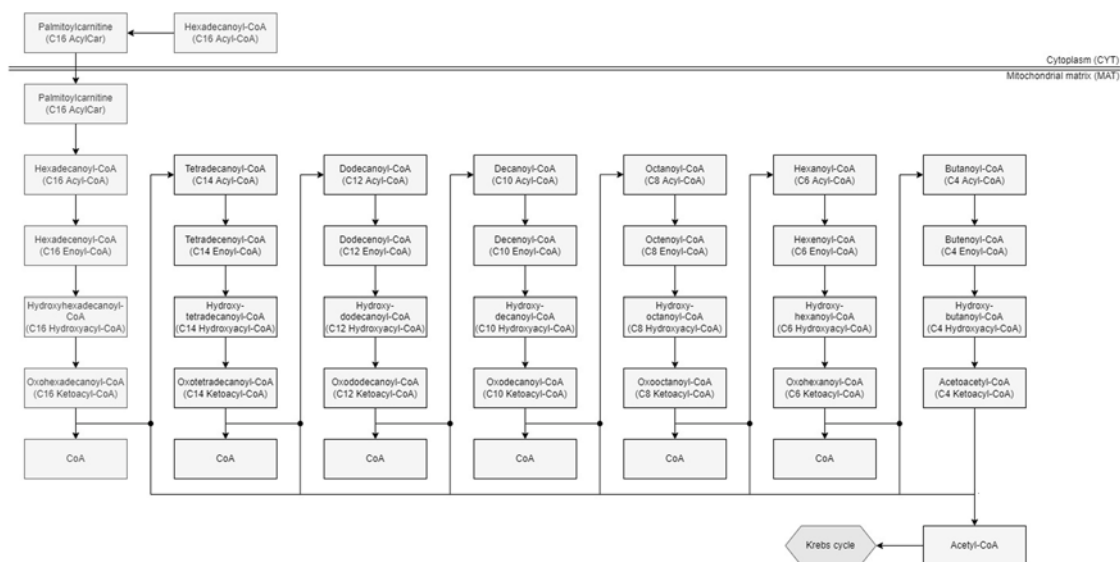


Fig. 2. Diagram showing the reactions in the fatty acids β -oxidation model based on the queuing theory.

2. Materials and methods

2.1. Model construction, data acquisition, enzymatic kinetics

The series of reactions involved in the β -oxidation pathway found from the KEGG Pathway database were prepared (Fig. 2). The kinetics of these reactions are calculated according to the principles of Michaelis–Menten enzymatic kinetics (Eq. (1)). The model is based on available literature data (i.e., kinetic constants, reaction rates, enzyme velocity) of enzymes and metabolites in each reaction. Metabolite concentration data was used as initial values. The values for the initial concentration of each metabolite and kinetic constants are included in the Supporting Information. The proposed model makes it possible to track changes in the concentration of individual metabolites over time.

$$v(t) = \frac{V_f \frac{S_1(t)S_2(t)}{K_{S_1}K_{S_2}} - V_r \frac{P_1(t)P_2(t)}{K_{P_1}K_{P_2}}}{(1 + \frac{S_1(t)}{K_{S_1}} + \frac{P_1(t)}{K_{P_1}})(1 + \frac{S_2(t)}{K_{S_2}} + \frac{P_2(t)}{K_{P_2}})} \quad (1)$$

where:

$v(t)$ - reaction speed

V_f - forward reaction speed

V_r - reverse reaction speed

$S_1(t), S_2(t), \dots, S_x(t)$ - substrate concentration in mmol/L

$P_1(t), P_2(t), \dots, P_x(t)$ - substrate concentration in mmol/L

$K_{S_1}, K_{S_2}, \dots, K_{S_x}$ - kinetic constant of substrate

$K_{P_1}, K_{P_2}, \dots, K_{P_x}$ - kinetic constant of product

It should be noted here that because concentrations of substrates and products change in time, the reaction rates given by Eq. (1) are time dependent, and this dependency introduces implicit feedback in the model.

2.2. Queueing theory

In the proposed model we calculated the adaptive parameter $\mu(t)$ with the use of the Michaelis–Menten equation. The behavior of β -oxidation metabolites and their reactions that appear in the model can be considered as a network of inhomogeneous Poisson processes described by Eq. (2):

$$P[(N(t + \tau) - N(t)) = k, t] = \frac{e^{-\mu(t)\tau} (\mu(t)\tau)^k}{k!} \quad (2)$$

where:

$P[(N(t + \tau) - N(t)) = k, t]$ - probability of k arrivals in the interval $(t, t + \tau]$

$\mu(t)\tau$ - expected number of arrivals in a time interval duration of $(t, t + \tau]$

The exponential distribution of the random variable T in the terms of the rate parameter $\mu(t)$ describes the queue processing time of metabolite increment and is given by Eq. (3):

$$f(T; \mu(t)) = \begin{cases} \mu(t)e^{-\mu(t)T} & T \geq 0 \\ 0 & T < 0 \end{cases} \quad (3)$$

A composition of interconnected queues based on the Michaelis–Menten equation can mimic β -oxidation. After a metabolite outflows from one queue, it will flow into the next queue. Thus, the network of interconnected queues is equivalent to the set of ODEs (Massey, 1985). Substrate's concentration and kinetic constants are factors that influence the likelihood of service to the successive queue. Each Michaelis–Menten equation applies to a particular substrate and affects whether the reaction takes place at the particular time point (Clement et al., 2020).

2.3. Genetic algorithm

To optimize the coefficients used in the Michaelis–Menten equation, we used the genetic algorithm (GA) (Holland, 1984; Katoch et al., 2021). Literature data was used as starting points for optimization. Kinetic constants have been treated as 'genes' in the 'chromosome'. Parameters were selected so that the entire β -oxidation queueing model reflects the experimental results as closely as possible. The β -oxidation state described in the literature and the current optimization stage of the simulations that make up the 'chromosome' are used to calculate the loss function. The loss function is the sum of the squared distances of the above parameters. The loss function formula is presented by Eq. (4):

$$g_p : \hat{X}, X \rightarrow \sum_{i=1}^{|X|} \left(\frac{\hat{X}_i - X_i}{X_i} \right)^2 \quad (4)$$

$$g(\hat{X}', \hat{X}'', X', X'') = g_p(\hat{X}', X') + g_p(\hat{X}'', X'')$$

where:

g_p - subfunction that penalizes the difference between two vectors in relation to second vector

g - loss function formula used for 'chromosomes' evaluation

X' - vector of substrate concentrations described by a literature after $\frac{1}{3}$ simulation time elapsed

X'' - vector of substrate concentrations described by a literature at the end of the simulation

\hat{X}' - vector of substrate concentrations obtained by evaluation of the 'chromosome' after $\frac{1}{3}$ simulation time elapsed

\hat{X}'' - vector of substrate concentrations obtained by evaluation at the end of the simulation

The vectors used in this equation are aggregations of individual substrates of the same carbon chain length. van Eunen et al. (2013) provided the values for literature vectors. The loss function is a summation of two instances of the difference penalty function and its task is to select the 'chromosome' (set of kinetic constants), which generates the simulation results, that resembles the closest time series obtained in the laboratory on living cells.

The points for GA to perform optimization of the simulated concentration values were based on the experimental measurements. van Eunen et al. (2013) measured the concentration values ten times during 24-minutes long experiment. Therefore, we performed optimization at the same time stamps as those measured experimentally.

The penalty subfunction computes the relative square error between subsequent values of the obtained vector and reference vector. The division by the value from the reference vector serves as an enforcement of equal contributions of all substrates in optimization process (it should be noted that at one point the value of C16 concentration is 75 times higher than concentration of C4).

3. Results

The developed queueing theory-based model of fatty acid β -oxidation includes individual steps taking into account the activation reaction catalyzed by fatty acyl-CoA synthetase and transport through the inner mitochondrial membrane with the use of the carnitine shuttle (Figs. 1 and 2). In the presented example, a fatty acid with a length of 16 carbon atoms (palmitoleic acid) was used. The simulation shows the case where acyl-CoA palmitoyl is in excess in the cytoplasm of the cell and its concentration is consistently high throughout the simulation process. After entering the mitochondrion, the newly formed palmitoyl-CoA undergoes an oxidation reaction, forming C16 enoyl-CoA. The enoyl-CoA molecule then hydrates to form hydroxyacyl-CoA, which then undergoes oxidation to form ketoacyl-CoA. The ketoacyl-CoA molecule undergoes thiolysis, resulting in the formation of an acetyl-CoA molecule and an acyl-CoA chain shorter by two carbon atoms. The resulting shorter acyl-CoA re-enters the β -oxidation pathway and the next cycle of reactions takes place.

The diagram shown in Fig. 2 has been transformed into 31 Michaelis–Menten enzymatic kinetics equations with 117 parameters, forming the basis of the presented model. For example, Eq. (5) representing the formation of C10 hydroxyacyl-CoA from C10 enoyl-CoA is as follows:

$$C10HydroxyacylCoAMAT = \frac{v_{crot} \frac{[C10EnoylCoAMAT]}{K_{MC10EnoylCoAMAT}} - \frac{[C10HydroxyacylCoAMAT]}{K_{MC10HydroxyacylCoAMAT}}}{\left(1 + \frac{[C10EnoylCoAMAT]}{K_{MC10EnoylCoAMAT}}\right) \left(1 + \frac{[C10HydroxyacylCoAMAT]}{K_{MC10HydroxyacylCoAMAT}}\right)} \quad (5)$$

The equation includes the speed of the corresponding enzyme involved in the reaction (here particularly crotonase), the substrate (C10 enoyl-CoA) and the product concentrations (C10 hydroxyacyl-CoA) and their kinetic constants. The data used in the equations came from scientific literature (van Eunen et al., 2013). Kinetic constants were optimized with the use of the genetic algorithm (GA). Detailed data on the equations (equation form, kinetic constants, concentrations) and data used in the model (after optimization) are presented in the Supporting Information.

The presented model starts when the C16 acyl-CoA is activated to enter the mitochondria. The model assumes that there are no shorter acyl-CoA molecules in the mitochondria. As the reaction cascade begins, the concentration of shorter carbon chains increases. Fig. 3 shows

the course of the 25 min simulation, taking into account changes in the aggregated concentrations of individual metabolites with the same carbon chain length. It also shows the initial accumulation of molecules with a 16-atom carbon chain, which reproduces the short delay observed in experimental results (van Eunen et al., 2013) for β -oxidation process to start using accumulated C16. This condition persists in the cell over time because the accumulated molecules compete for an enzyme capable of transforming them. The presented model was validated by comparison of the experimental and computational data presented in van Eunen et al. (2013). The results presented by van Eunen et al. (2013) included the experimental measurements of β -oxidation metabolites derived from rat liver mitochondria. Fig. 3 compares the 16-atom long carbon chain (C16) laboratory measurements (van Eunen et al., 2013) and simulation results. The characteristics of the simulation run are similar to laboratory results (van Eunen et al., 2013).

ODE-based model proposed by van Eunen et al. (2013) perfectly reflects qualitative changes of concentration. However, it is not fully precise in the case of quantitative prediction. Therefore, we believe that the queueing theory-based model with improvements proposed by AI algorithms, like GA used in this case, can improve already existing models and tools.

Presented results are the mean of simulated results for 50 independent rat liver cells. The resulting database includes 3 000 measurements (each measurement every half a second), which are the average values of 2 000 measurements. One of the advantages of the model is the speed of its calculations as they can be calculated in real time. Simulations were also conducted taking into account Gaussian noise at levels of 25% and 50% of the metabolites' initial concentrations. The model successfully achieved stability for Gaussian noise of 25% (Fig. 4) and 50% (Fig. 5). The model retains a qualitative reflection of concentration changes (standard deviation over mean results are attached to Supporting Information). In this way, it was shown that this model is resistant to laboratory measurement inaccuracies resulting from extremely low concentrations at which individual pathway metabolites are present. The presented model is computationally efficient, which results from the methodology used. The model enables the tracking of the β -oxidation process and time changes in the concentration of fatty acids with carbon chains of various lengths.

4. Discussion and conclusions

This paper presents a β -oxidation model based on queueing theory. The parameters used in the developed model come from previous literature and were optimized using a GA so that individual reaction equations could be combined into a functional β -oxidation model. The obtained results confirmed the effectiveness of queueing theory in metabolism modeling when compared to the experimental results from rat liver mitochondria, as well as previous models based on ODEs (van Eunen et al., 2013).

The authors in Mc Auley and Mooney (2015) point out that modeling lipid metabolism is difficult due to the fact that the current understanding of the biology of metabolism is very difficult to describe with mathematical equations. Consequently, the simulation results obtained need not always match exactly what is going on in the cell. However, the model we developed, which uses literature data of metabolites and enzymes, showed that queueing theory is capable of reproducing laboratory results.

Schützhold et al. (2016) presented a model of lipid metabolism in yeast. However, the authors focused on other factors of lipid metabolism, such as the distribution of fatty acids in different membranes throughout cell cycle as the cell's requirements change depending on the cycle phase. Therefore, their work provides more opportunities for location-dependent distributions predictions.

Due to the multitude of processes lipids are involved in, disturbances in lipid metabolism are a risk factor for various diseases. Disturbed lipid metabolism occurs in cancer cells (Natter and Kohlwein,

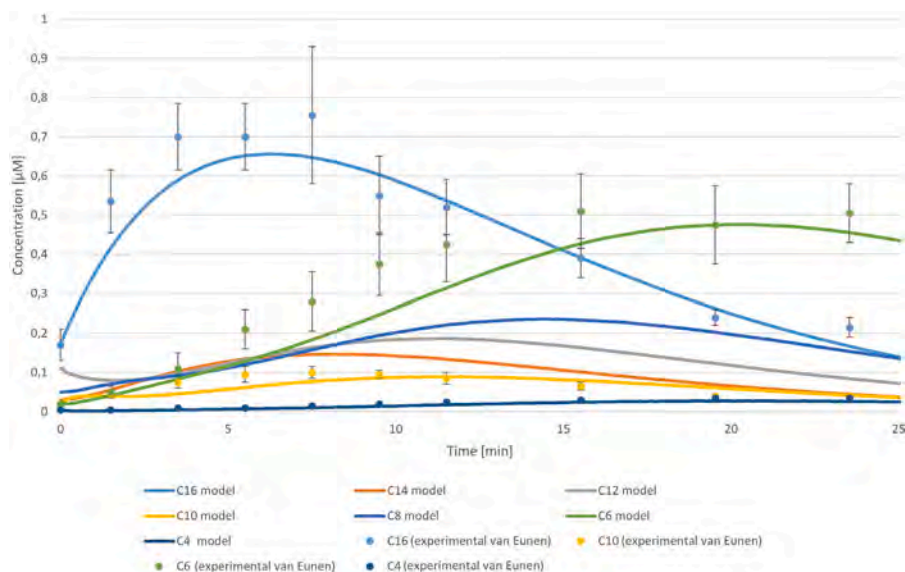


Fig. 3. Concentration level change over 25 min simulation under unperturbed conditions, without Gaussian noise. C16 — metabolites with carbon chain length of 16; C14 — metabolites with carbon chain length of 14; C12 — metabolites with carbon chain length of 12; C10 — metabolites with carbon chain length of 10; C8 — metabolites with carbon chain length of 8; C6 — metabolites with carbon chain length of 6; C4 — metabolites with carbon chain length of 4; experimental results refers to [van Eunen et al. \(2013\)](#).

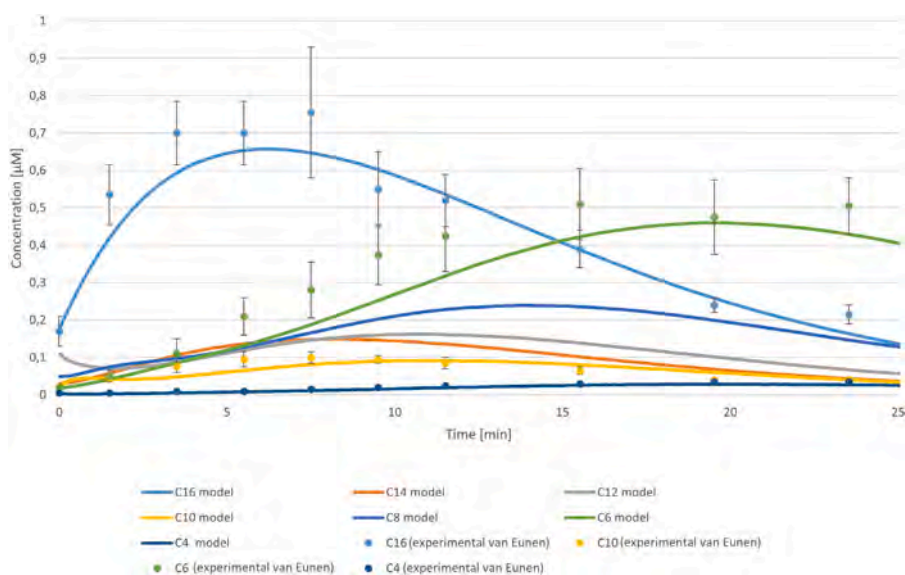


Fig. 4. Concentration level change over 25 min simulation with Gaussian noise of 25%. C16 — metabolites with carbon chain length of 16; C14 — metabolites with carbon chain length of 14; C12 — metabolites with carbon chain length of 12; C10 — metabolites with carbon chain length of 10; C8 — metabolites with carbon chain length of 8; C6 — metabolites with carbon chain length of 6; C4 — metabolites with carbon chain length of 4, experimental results refers to [van Eunen et al. \(2013\)](#).

2013) and in Alzheimer's disease ([Liu et al., 2013](#)). The concentration of free fatty acids in the plasma is enhanced in people with type 2 diabetes, obesity, and in patients with non-alcoholic fatty liver disease ([Mihalik et al., 2010](#); [Henderson, 2021](#)).

Moreover, mutations in the genes encoding the enzymes for the β -oxidation pathway reactions can cause deficiencies and disorders. [Modre-Osprian et al. \(2009\)](#) created a β -oxidation model that allows the percentage of acyl-CoA with different carbon chain lengths to be assessed in relation to the total acyl-CoA concentration. The model can simulate various enzyme deficiencies, and they verified the model's results on neonatal laboratory results. In the case of a deficiency of any particular enzyme, molecules accumulate and are insufficiently processed. Referring to the results of our model ([Fig. 3](#)) in the case of reduction of enzyme efficacy or a shortage of enzymes responsible for the conversion of particles with a 16-atom carbon chain, their accumulation would be longer and lead to an even greater increase in C16

concentration. Excess C16 particles accumulate inside the mitochondria. As the concentration of C16 inside the mitochondrion increases, the rate at which subsequent molecules are delivered decreases, which gives the cell time to start using already accumulated molecules in the first reaction of β -oxidation.

Accumulation of palmitoyl-CoA inside the mitochondria can be inhibited by reducing the activity of enzymes involved in both activation (fatty acyl-CoA synthetase) and transport across the mitochondrial membrane (carnitine palmitoyltransferase I). One possible inhibition mechanism is to lower the concentration of CoA in the cytosol, thus making the activation reaction less frequent.

The limitation of the presented model is that we focused on fatty acids with an even number of carbon atoms. In the case of fatty acids with an odd number of carbon atoms in the chain (e.g., C5), the resulting propionyl-CoA (C3) undergoes a series of reactions that result in the formation of succinyl-CoA, which in this form can enter

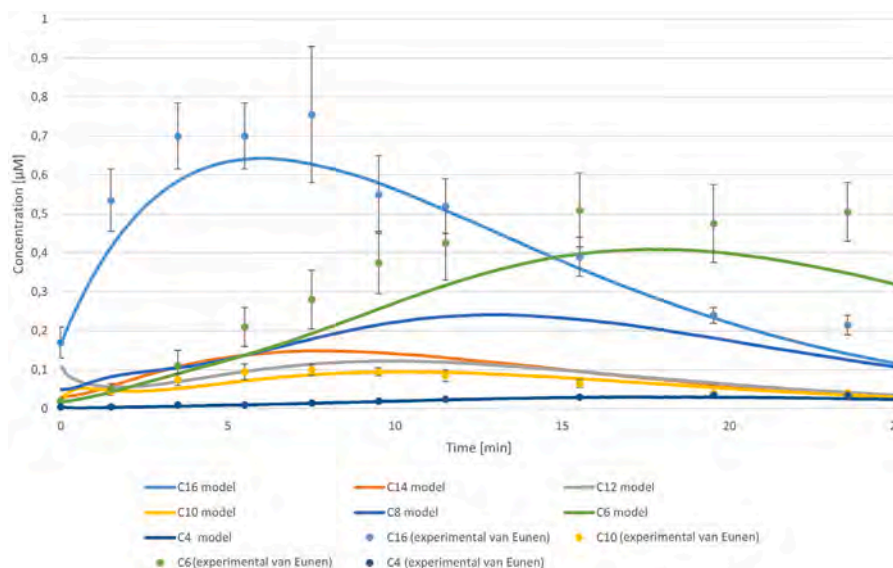


Fig. 5. Concentration level change over 25 min simulation with Gaussian noise of 50%. C16 — metabolites with carbon chain length of 16; C14 — metabolites with carbon chain length of 14; C12 — metabolites with carbon chain length of 12; C10 — metabolites with carbon chain length of 10; C8 — metabolites with carbon chain length of 8; C6 — metabolites with carbon chain length of 6; C4 — metabolites with carbon chain length of 4, experimental results refers to van Eunen et al. (2013).

the Krebs cycle. However, due to the nature of the queuing theory modeling, further reactions can be added at a later stage of the study and combined with other cellular processes such as the Krebs cycle. Another limitation of the model is that we considered only the scenario of palmitoleic acid, which is saturated fatty acid. We did not consider other scenarios of unsaturated fatty acids, which may also have an impact on the β -oxidation kinetics.

A queuing theory-based model of β -oxidation was prepared. The results obtained by the model indicate its accuracy, as these results are similar to experimental results. It can be concluded that the results of this queuing theory model are more accurate than those based on ODEs, and demands less computing power. The less-computationally intensive model can therefore carry out simulations in real time. We believe that the model can be used in research into diseases related to disorders of fatty acid metabolism and contribute to the development of effective therapies. The modeling technique used enables the model to be expanded in the future and to include other metabolic pathways.

Funding

This work was supported by the National Science Center (Narodowe Centrum Nauki, NCN) of Poland in terms of Opus-17 Program [2019/33/B/ST6/00875].

Role of the funding source

The source of funding did not affect the data collection, analysis, or the results described in this manuscript. The authors confirm that they had full access to all of the data in this study and take complete responsibility for the integrity of the data and the accuracy of the data analysis.

CRediT authorship contribution statement

Sylwester M. Kloska: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Krzysztof Pałczyński:** Methodology, Software, Validation, Visualization. **Tomasz Marciniak:** Conceptualization, Software, Supervision, Writing – review & editing. **Tomasz Talaśka:** Software, Validation, Writing – review & editing. **Marissa Miller:** Software, Writing – review & editing. **Beata J. Wysocki:** Conceptualization, Methodology, Writing – review & editing.

Paul Davis: Methodology, Writing – review & editing. **Tadeusz A. Wysocki:** Conceptualization, Software, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability/data sharing plans

The source code for the model is available at the GitHub repository at <https://github.com/UTP-WTliE/FattyAcidsOxidation>.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.compbiolchem.2023.107860>.

References

- Clement, E.J., Schulze, T.T., Soliman, G.A., Wysocki, B.J., Davis, P.H., Wysocki, T.A., 2020. Stochastic simulation of cellular metabolism. *IEEE Access* 8, 79734–79744.
- Evstigneev, M.P., 2014. Vladyslav P. Evstigneev, Marina G. Holyavka, Sergii V. Khrapaty & Bull. Math. Biol. 76, 2238–2248.
- Gillespie, D.T., 1977. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81 (25), 2340–2361.
- Guo, D., Zhang, G., Wysocki, T.A., Wysocki, B.J., Gelbard, H.A., Liu, X.-M., McMillan, J.M., Gendelman, H.E., 2014. Endosomal trafficking of nanoformulated antiretroviral therapy facilitates drug particle carriage and HIV clearance. *J. Virol.* 88 (17), 9504–9513.
- Henderson, G.C., 2021. Plasma free fatty acid concentration as a modifiable risk factor for metabolic disease. *Nutrients* 13 (8), 2590.
- Holland, J.H., 1984. Genetic algorithms and adaptation. In: *Adaptive Control of Ill-Defined Systems*. Springer, pp. 317–333.
- Houten, S.M., Wanders, R.J., 2010. A general introduction to the biochemistry of mitochondrial fatty acid β -oxidation. *J. Inher. Metab. Dis.* 33 (5), 469–477.
- Katoch, S., Chauhan, S.S., Kumar, V., 2021. A review on genetic algorithm: past, present, and future. *Multimedia Tools Appl.* 80 (5), 8091–8126.
- Kloska, S.M., Pałczyński, K., Marciniak, T., Talaśka, T., Miller, M., Wysocki, B.J., Davis, P., Wysocki, T.A., 2022. Queuing theory model of pentose phosphate pathway. *Sci. Rep.* 12 (1), 1–9.
- Kloska, S., Pałczyński, K., Marciniak, T., Talaśka, T., Nitz, M., Wysocki, B.J., Davis, P., Wysocki, T.A., 2021. Queuing theory model of Krebs cycle. *Bioinformatics* 37 (18), 2912–2919.

- Liu, C.-C., Kanekiyo, T., Xu, H., Bu, G., 2013. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature Rev. Neurol.* 9 (2), 106–118.
- Massey, W.A., 1985. Asymptotic analysis of the time dependent M/M/1 queue. *Math. Oper. Res.* 10 (2), 305–327.
- Mc Auley, M.T., Mooney, K.M., 2015. Computationally modeling lipid metabolism and aging: a mini-review. *Comput. Struct. Biotechnol. J.* 13, 38–46.
- Mihalik, S.J., Goodpaster, B.H., Kelley, D.E., Chace, D.H., Vockley, J., Toledo, F.G., DeLany, J.P., 2010. Increased levels of plasma acylcarnitines in obesity and type 2 diabetes and identification of a marker of glucolipotoxicity. *Obesity* 18 (9), 1695–1700.
- Modre-Osprian, R., Osprian, I., Tilg, B., Schreier, G., Weinberger, K.M., Graber, A., 2009. Dynamic simulations on the mitochondrial fatty acid beta-oxidation network. *BMC Syst. Biol.* 3 (1), 1–15.
- Natter, K., Kohlwein, S.D., 2013. Yeast and cancer cells—common principles in lipid metabolism. *Biochim. Biophys. Acta (BBA)-Mol. Cell Biol. Lipids* 1831 (2), 314–326.
- Noland, R.C., Koves, T.R., Seiler, S.E., Lum, H., Lust, R.M., Ilkayeva, O., Stevens, R.D., Hegardt, F.G., Muoio, D.M., 2009. Carnitine insufficiency caused by aging and overnutrition compromises mitochondrial performance and metabolic control. *J. Biol. Chem.* 284 (34), 22840–22852.
- Puchałka, J., Kierzek, A.M., 2004. Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks. *Biophys. J.* 86 (3), 1357–1372.
- Schützhold, V., Hahn, J., Tummler, K., Klipp, E., 2016. Computational modeling of lipid metabolism in yeast. *Front. Mol. Biosci.* 3, 57.
- Shampine, L.F., Thompson, S., Kierzenka, J., Byrne, G., 2005. Non-negative solutions of ODEs. *Appl. Math. Comput.* 170 (1), 556–569.
- Sharp, A.T., Pannier, A.K., Wysocki, B.J., Wysocki, T.A., 2012. A novel telecommunications-based approach to HIV modeling and simulation. *Nano Commun. Netw.* 3 (2), 129–137.
- van Eunen, K., Simons, S.M., Gerding, A., Bleeker, A., den Besten, G., Touw, C.M., Houten, S.M., Groen, B.K., Krab, K., Reijngoud, D.-J., et al., 2013. Biochemical competition makes fatty-acid β -oxidation vulnerable to substrate overload. *PLoS Comput. Biol.* 9 (8), e1003186.
- van Riel, N.A., 2006. Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Brief. Bioinform.* 7 (4), 364–374.
- Voit, E.O., 2017. The best models of metabolism. *Wiley Interdiscipl. Rev.: Syst. Biol. Med.* 9 (6), e1391.

Queueing theory model of mTOR complexes' impact on Akt-mediated adipocytes response to insulin

Sylwester M. Kloska, Krzysztof Pałczyński, Tomasz Marciniak, Tomasz Talaśka, Marissa Miller, Beata J. Wysocki,

Paul H. Davis, Ghada A. Soliman, Tadeusz A. Wysocki

Published: December 27, 2022 • <https://doi.org/10.1371/journal.pone.0279573>

Abstract

A queueing theory based model of mTOR complexes impact on Akt-mediated cell response to insulin is presented in this paper. The model includes several aspects including the effect of insulin on the transport of glucose from the blood into the adipocytes with the participation of GLUT4, and the role of the GAPDH enzyme as a regulator of mTORC1 activity. A genetic algorithm was used to optimize the model parameters. It can be observed that mTORC1 activity is related to the amount of GLUT4 involved in glucose transport. The results show the relationship between the amount of GAPDH in the cell and mTORC1 activity. Moreover, obtained results suggest that mTORC1 inhibitors may be an effective agent in the fight against type 2 diabetes. However, these results are based on theoretical knowledge and appropriate experimental tests should be performed before making firm conclusions.

Citation: Kloska SM, Pałczyński K, Marciniak T, Talaśka T, Miller M, Wysocki BJ, et al. (2022) Queueing theory model of mTOR complexes' impact on Akt-mediated adipocytes response to insulin. PLoS ONE 17(12): e0279573. <https://doi.org/10.1371/journal.pone.0279573>

Editor: Irina U. Agoulnik, Florida International University, UNITED STATES

Received: July 6, 2022; **Accepted:** December 11, 2022; **Published:** December 27, 2022

Copyright: © 2022 Kloska et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability: All relevant data are within the paper and its [Supporting Information](#) files. The source code of the described model can be accessed by: <https://doi.org/10.5281/zenodo.7117138>.

Funding: This work was supported by the National Science Center (Narodowe Centrum Nauki, NCN) of Poland (<https://www.ncn.gov.pl/>) in terms of Opus-17 Program [2019/33/B/ST6/00875 awarded to TAW]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Biological importance

A key hormone that controls blood glucose levels is insulin. This hormone is secreted by the β -cells of pancreatic islets. Insulin facilitates glucose uptake in peripheral tissues including the muscle, and adipose tissue [1]. It inhibits glucose production from non-glucose sources by inhibiting gluconeogenesis and glycogenolysis, while stimulating glycogen synthesis. The hormone with the opposite effect of insulin is glucagon [2]. Both of these hormones together are primarily responsible for the maintenance of glucose homeostasis in mammals.

The attachment of insulin to the insulin receptor starts a cascade of reactions responsible for the absorption of glucose inside the cell [3]. One of the main effects of this cascade is the translocation of glucose transporter 4 (GLUT4) from the center of the cell towards the cell membrane. GLUT4 is a protein that facilitates the diffusion of glucose along a concentration gradient—from a higher concentration in the blood to a lower concentration in the cell. The participation of GLUT4 in the transport of glucose inside the cell increases the amount of transported glucose molecules by 30 times [4, 5].

Adequate management of glucose levels in the cell is crucial to maintain a healthy environment in the cell and its function. One of the mechanisms that supervise the maintenance of adequate blood glucose levels is through mammalian target of rapamycin (mTOR) kinase. mTOR links with other proteins and forms two protein complexes described as mTORC1 and mTORC2. These complexes are responsible for regulation of various important processes inside the cell, including cell growth regulation, cell proliferation, cell motility, cell survival, protein synthesis, autophagy, DNA transcription, and metabolism [6]. The dysregulation and incorrect activity of mTOR complexes can lead to diseases such as obesity, diabetes and even cancer [7, 8]. One of the proteins that regulate the mTORC1 complex is the Rheb protein [9, 10]. It is one of the key mTORC1 activating proteins. However, one enzyme in the glycolytic pathway—glyceraldehyde 3-phosphate dehydrogenase (GAPDH), has a high affinity for the Rheb protein [11]. When GAPDH enzyme molecules are not involved in the reaction that produces 1,3-bisphosphoglycerate (1,3-BPG) from glyceraldehyde 3-phosphate (G3P), they combine with Rheb protein molecules, depriving the mTORC1 complex its key activator, leading to inactivity of the mTORC1. When the cell has normal/high concentrations of G3P, GAPDH molecules are busy processing G3P, so Rheb can freely bind to mTORC1 and activate it. Depending on the above-described mTORC1 activation process, the amount of GLUT4 particles varies. For this reason, we decided to prepare a computational model capable of predicting the number of active GLUT4 particles that are capable of participating in glucose transport.

Queueing theory

Typically, cellular signaling networks have been modeled using a set of ordinary differential equations (ODEs) [12]. Using these equations, it is possible to demonstrate the changes that occur in the cell during rest and in response to external stimuli causing upstream signals. However, when using ODEs, the fluctuations in the cell leading to local changes (e.g., temperature) are not taken into account, which influences the values of the kinetic constants that affect the way the cell responds. To map the intracellular environment more accurately, as well as the random variation, a model based on the queueing theory can be useful. Queueing theory was mainly used in telecommunications and engineering [13–16]. Additionally, it is suitable for modeling stochastic processes in cells. The idea to use a method commonly used in telecommunications comes from the fact that signaling paths, similar to the transmission of internet packets, transmit information from node to node. Likewise, in a cell, signaling molecules are passed on, activating subsequent elements (proteins) of the cascade. To date, the queueing theory approach has been used to model simple metabolic networks [17], metabolic pathways such as glycolysis [18] and the Krebs cycle [19]. The presented model is an extension of the work [20] to include loops related to the regulation of cellular metabolism by mTOR complexes and mTORC1 regulation via GAPDH availability, or more precisely—'occupancy'. In the case of models such as the one presented here, which use a large number of variables, the application of the queueing theory seems to be more optimal than the use of ODEs. The model is capable of achieving stability. Another advantage of using queueing theory to model signaling pathways is that they require significantly less computing power compared to ODE models. For this reason, simulation can be carried out practically in real time. Due to the short duration of the simulation, it can be used to learn about the relationships caused by manipulations of specific kinetic constants or concentrations, which also has its advantages when considering the reactions that are not well studied/established. To confirm the correctness of the obtained simulation results, the simulated data were compared with the results of laboratory experiments [21]. Finally, using the queueing theory gives the possibility of expanding the model with further reactions, without major interference in the course of the previously described, due to the fact that they are based on empirically obtained values. Therefore, the model is adapted to be supplemented with the development of the state of knowledge about the given signaling or metabolic pathways. Moreover, it can be used to theoretically test the kinetic changes brought about by potential mTORC inhibitors [22].

The aim of this work is to present a comprehensive model of cellular response to insulin, which leads to GLUT4 translocation and mTORC activation as a part of processes responsible for maintaining proper cellular glucose concentration. Fig 1 shows the links between molecules involved in the insulin signaling pathway. The research hypothesis of this work is the ability to simulate the cellular response to insulin and track changes in the concentrations of proteins involved in this response using queueing theory based simulation model. The presented model shows the mechanism of mTORC1 influence on mobilization of GLUT4 particles. Since mTORC1 has been reported in literature as having an impact on glucose uptake [23, 24].

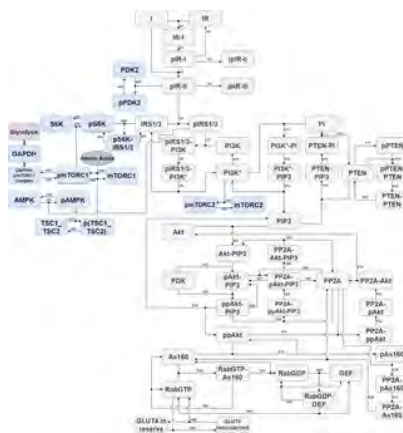


Fig 1. Diagram illustrating the computational model of the insulin signaling pathway.

The cascade of reactions begins when insulin binds to the insulin receptor. The scheme includes the most important proteins in the PI3K/Akt/mTOR pathway which play a role in the cellular response to insulin.

<https://doi.org/10.1371/journal.pone.0279573.g001>

The presented model is an extension of previously described simulation of insulin mediated GLUT4 translocation [20]. Since then the mTORC1 signaling pathway connections with Akt-mediated insulin response has been described [25, 26]. This work presents a model where those connections have been included. Moreover, the paper studies regulation of mTORC1 activity by the glycolytic enzyme GAPDH, which has high affinity for the mTOR activator protein—Rheb. To train the model we have used genetic algorithm (GA) to optimize the kinetic coefficients. The achieved results allow to conclude that artificial intelligence (AI) algorithms, in this case the genetic algorithm, can be effective tools for optimizing computational models. In order to validate the obtained results, we present multiple variants of mTORC1 activity that can be practically obtained through the administration of an mTORC1 inhibitor, such as rapamycin [27].

Methodology

The endpoints of the Akt-mediated insulin signaling pathway are well characterized [28–30]. Therefore, by comparing the experimental and computational results, it can be assessed whether the model works properly. The values of kinetic constants and concentrations of signaling molecules were obtained by searching the PubMed database. Simulations were performed separately for 50 independent cells, which mimic human adipocytes. This type of cell was chosen because of the availability of literature data, which was used in the development of the model. For each of the cells, the concentrations of all molecules participating in the signaling pathway were randomly chosen from the given range, limited by 10% Gaussian noise. According to the queueing theory, the current concentrations of individual molecules in each cell are separate 'stores'—queues [18, 31]. The speed of the response determines the probability of passing from one queue to the next. The simulation results are averaged over the entire cell population. A network based on queues can be used to model reactions whose rates change dynamically and randomly. The simulation was performed in C# 8.0. All the results were obtained using 1ms time increments; however, the simulation allows the choice of any user-selected time increment. While changing the time increment, one should pay attention to the fact that the probabilities of the reaction occurrence are <1. Detailed information on the equations, kinetic constants, and initial concentrations can be found in the S1 File.

The Genetic Algorithm [32, 33] was used to tune the model of interconnected queues realizing Michaelis-Menten equations. Each 'chromosome' consisted of linear coefficients for selected group of queues scaling their probability of reaction occurrence. The population of GA consists of ten 'chromosomes'. In each epoch, every 'chromosome' is evaluated and the two 'chromosomes' with the best scores are chosen. The process of 'chromosome' evaluation consists of performing three simulations with a set of kinetic constants, linear coefficients stored in each 'chromosome', and a value of available GAPDH. Each simulation used a different value of available GAPDH taken from a set {0%, 20%, 50%, 100%}. One simulation was formed emulating 50 cells working in parallel to each other. The evaluation step was added to measure, 1) how many cells reached the maximum value of GLUT4 in vesicles for available GAPDH equal to 100%, 2) how many cells reached the minimum value of GLUT4 in vesicles for available GAPDH equal to 0%, and 3) how distant is the number of cells that reached the maximum value of GLUT4 in vesicles for available GAPDH equal to 50% from aforementioned results for GAPDH equal to 100% and 0%.

To validate the model, theoretical inhibition of mTORC1 was used to test the effects of changes in reduction of its activity. One of the inhibitors of mTOR complexes' activity is rapamycin. Previous studies show that rapamycin causes a number of side effects, including increased risk of infection [34], increased incidence of cancer [35], weight disorders, hyperlipidemia, and diabetes-like metabolic disorders [36]. For this reason, it seems necessary to develop drugs that selectively affect mTORC1 activity, while at the same time not having such significant side effects, like astragaloside IV (As-IV) [37]. As-IV was proven to be effective mTORC1 inhibitor and reduced mTORC1 signaling in mice. The data obtained from the presented model can be used in the study of the kinetics of reactions in the insulin signaling pathway, which will help to select the appropriate place where the influence of therapeutics could have the best effect.

Without insulin activating the cascade and mobilizing GLUT4 to move towards the cell membrane, there are approximately 18,200 GLUT4 molecules proximate to the cell membrane [38], ready to transport glucose inside the cell. This number increases to approximately 195,000 as a result of insulin-stimulated activation [38, 39]. However, these are not total GLUT4 stocks. In fact, the cell has a large reservoir that it can use in extreme cases. The said number 195,000 accounts for approximately 50% of total GLUT4 [21].

To illustrate the changes caused by the influence of GAPDH molecules on mTORC1 activation, two varying scenarios are described below (Fig 2). These scenarios focus on different cellular conditions such as glucose levels and the intensity of glycolysis.

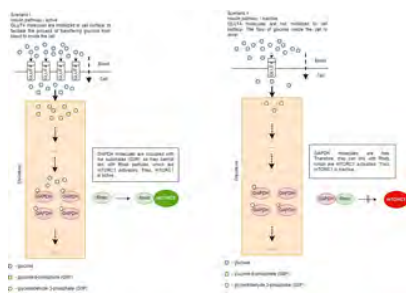


Fig 2. Various scenarios of mTORC1 activity depending on GAPDH 'occupancy'.
<https://doi.org/10.1371/journal.pone.0279573.g002>

Scenario I—the concentration of glucose in the blood is elevated after eating and the insulin signaling pathway works correctly. As a result, GLUT4 molecules are mobilized to migrate to the cell membrane, where they facilitate the flow of glucose from the blood to inside the cell. The glucose level in the blood drops, while the cellular level of glucose rises. To avoid the situation where glucose molecules leave the cell, glucose is phosphorylated and becomes G6P. There are two destinations for G6P molecules: 1) the glycolytic pathway or 2) glycogenesis, the formation of glycogen.

When G6P enters glycolysis, the sequence of reactions takes place and glyceraldehyde 3-phosphate (G3P) molecules are formed. G3P is converted into 1,3-bisphosphoglycerate (1,3BPG) by GAPDH.

GAPDH is particularly important because it is involved in regulation of mTORC1 activity. GAPDH concentration levels in the cell do not change drastically rather they oscillate around the same values. However, what changes is their state—they can be either 'occupied' with processing G3P molecules, or if there are more enzyme molecules than substrate molecules, the excessive amount of enzyme molecules is free. Those free GAPDH molecules connect with Rheb protein and activate mTORC1. It remains unknown how Rheb stimulates the activity of mTORC1.

Scenario II—the organism is in state of prolonged fasting causing a decrease in the supply of extracellular glucose and ceasing insulin secretion. Without the release of insulin from the blood, the reaction remains inactivated and GLUT4 remains stationary and unable to transport glucose. In this situation, the stored amounts of glycogen are hydrolyzed and the basic levels of G6P are maintained. As previously described, glycolysis runs as normal. However, the amount of formed G3P molecules is lower than in Scenario I. In fact, there is larger amount of GAPDH molecules than G3P molecules. Therefore, the free GAPDH molecules can freely bind with Rheb protein, resulting in mTORC1 inactivation.

To conclude, increased extracellular supply of glucose activates insulin signaling. The glycolytic flux is increased and the GAPDH molecules are occupied with processing G3P molecules. As a result, Rheb molecules are floating freely and can bind to and activate mTORC1.

However, the conditions presented in both scenarios are extreme and practically unrealistic in the cell, as the probability of such extreme conditions as 0 or 100% 'occupancy' of GAPDH is low. In a cell, most often intermediate conditions prevail.

Results

Effect of GAPDH and mTORC1 on the amount of GLUT4 involved in glucose transport

A working, stable queueing theory-based model of the insulin signaling pathway was obtained. The presented study was aimed at illustrating the interrelationships between the levels of GLUT4, GAPDH, and mTORC1. These relationships have a significant impact on how the cell responds to insulin and extracellular glucose supply. The results obtained with the use of the model are consistent with the current state of knowledge [10, 40]. The amount of GLUT4 particles ready to take part in the glucose transport process is significantly dependent on the amount of 'occupied' GAPDH. When the system is not inhibited, less than 200,000 GLUT4 molecules are involved in the transport of glucose to the cell. However, depending on the level of activity that is influenced by both GAPDH and indirectly by mTORC1, this number fluctuates. Fig 3 shows the relationship between the level of GLUT4 in the vicinity of the cell membrane and the level of 'occupied' GAPDH. Depending on the condition of the cell, as well as mTORC1 activity, the amount of GLUT4 mobilized can vary considerably (Figs 4 and 5). The greater amount of GAPDH involved in substrate processing allows Rheb to link freely with mTORC1. mTORC1 activity and GLUT4 level are correlated with each other [41, 42]. The same conclusions can be drawn by analyzing the obtained results on the charts.

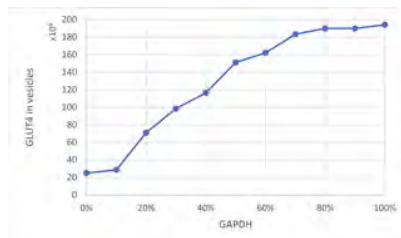


Fig 3. The relationship between the amount of GLUT4 particles in the cell membrane area and the level of 'occupied' GAPDH.
<https://doi.org/10.1371/journal.pone.0279573.g003>

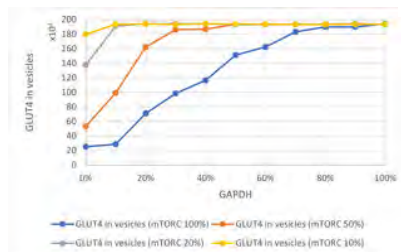


Fig 4. Relationship of GLUT4 in vesicles and 'occupied' GAPDH.
 Colored lines indicate different levels of mTORC1 activity.
<https://doi.org/10.1371/journal.pone.0279573.g004>

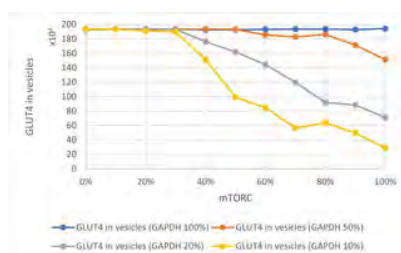


Fig 5. Dependence of GLUT4 level in vesicles in relation to mTORC1 activity.
 The colored lines indicate the different levels of 'occupied' GAPDH.
<https://doi.org/10.1371/journal.pone.0279573.g005>

We also tested the effect of lowering mTORC1 activity, e.g., through the use of drugs, on the amount of GLUT4 particles, while assuming different levels of GAPDH 'occupancy' (Fig 4). Analogous studies were performed for different levels of GAPDH with respect to mTORC1 activity (Fig 5). Both mTORC1 activity and the amount of 'occupied' GAPDH significantly influences the amount of GLUT4 and can contribute to lowering the amount of GLUT4 particles involved in glucose transport (Fig 5). The scenario in which all the GAPDH particles present in the cell are busy processing its substrate so that the mTORC1 can be fully active, keeps the amount of GLUT4 in vesicles at the maximum level (Fig 5). The presented results indicate that drugs that can significantly decrease mTORC1 activity (at least 50% inhibition) are of great importance for the amount of GLUT4 particles directed to the cell membrane for glucose transport inside the cell. Similar conclusions can be drawn from the results presented by Rajan et al. [43] and Veilleux et al. [44] which confirms the validity of the method we presented.

Discussion

Identification of key nodes in insulin signaling

Practical application of the conclusions of the described scenarios for GAPDH and mTORC1 allowed the identification of key nodes for the appropriate cell response to insulin, and confirmed previous experimental results described in [45]. In [45] the authors explained and proved the important role of S6 kinase (S6K). mTORC1 participates in its phosphorylation. S6K is crucial because it

is the link between the mTORC1 loop and the rest of the proteins responsible for insulin signaling. Signaling between mTORC1 and S6K causes a negative-feedback loop which lowers cellular sensitivity for insulin. The activation of the mTORC1/S6K loop leads to increased degradation of insulin receptor substrate 1/3 (IRS1/3) and therefore influences the amount of GLUT4 in vesicles. This entire process affects how many glucose molecules enter the cell from the bloodstream.

The experimental results, as well as those obtained in the presented model, indicate that the insulin response system is very complex and depends on many elements that regulate it. It is characterized by high instability, small changes that can lead to a greatly altered cell response, causing disease such as type 2 diabetes, where the cells become insensitive to insulin. As shown in the above model, there are many elements that can cause glucose malabsorption.

Sonntag et al. [46] focused on determining which of the 'nodes' of the insulin signaling pathway influences AMP-activated protein kinase (AMPK) activity. The equations described in the [46] are based on the mass action law. The obtained results state that IRS1/3 is the 'node' influencing AMPK. The model proposed by Sonntag et al. focused on simplifying the insulin signaling pathway and it does not take into account several 'nodes' that play a significant role in this process. Therefore, the combination of the data and results presented by [46] was a valuable source in the preparation of the model based on the queueing theory. GA was used to find an appropriate scaling of the values so that the model as a whole would work properly.

The presented model has several limitations. It does not take into account other signaling pathways or individual reactions that are also connected to and influence signaling proteins. This is especially true for the Akt protein, which is the central node in the presented signaling model. Moreover, a model based on literature data will only be as good as the available data. However, we do not question the reliability of other research teams and their published results. Another of the limitations is that in queueing theory, each simulation gives one realization of the stochastic process, while ODE gives an averaged solution. Therefore, a limitation is that depending on the number of cells for which one runs simulations and then averages them, this is how accurate the result will be. Therefore, the model presented here is for averaged results for 50 cells.

mTORC1 activity and related treatment strategies

The results of the described model could be used as a suggestion in the process of developing new drugs, including drugs that increase insulin-sensitivity in peripheral tissues such as the muscle and adipose tissue (e.g., Metformin). Identifying key 'nodes' throughout the signaling pathway could guide researchers in helping cells regain their original insulin sensitivity. However, due to the complexity of connections between all signaling molecules, this task is very difficult.

mTORC1 plays an important role in the maintenance of an adequate level of glucose in the blood. When necessary, i.e., in a nourished state, mTORC1 activity stimulates pancreatic β -cells to secrete insulin, thus maintaining adequate glucose tolerance. However, studies in mice [47, 48] show that mTORC1 overactivity may cause a faster deterioration in β -cell function and consequently complications with glucose homeostasis. Therefore, the use of mTORC1 inhibitors to improve glucose tolerance has been considered. Previous studies in mice have shown that S6K knockdown or inhibitors that reduce S6K phosphorylation make cells more insulin sensitive [49, 50]. The results obtained with the use of the queueing theory model confirm earlier reports [45] that mTOR/S6K inhibition could be a therapeutic target in type 2 diabetes.

One of the most common prototype mTOR inhibitors is rapamycin. However, the use of rapamycin has been counterproductive, inducing insulin resistance and disrupting glucose homeostasis in the body [51]. Rapamycin is an effective inhibitor of mTORC1. Most researchers agree that rapamycin does not inhibit mTORC2 at least in the acute stimulation [52]. Few researches suggest that rapamycin inhibits mTORC2 only in some cell types and only with chronic administration due to inhibiting to mTORC2 assembly [53, 54]. Knowing the function and the importance of this complex in signaling pathway, it is no wonder that long-term mTOR inhibition interferes with the body's response to insulin. Due to the fact that rapamycin affects both mTORC1 and mTORC2, it can be concluded that it is worth testing substances that act selectively on only one of these complexes.

Research by Tao et al. [22] provided useful information on the influence of inhibitors on mTORC kinetics and activity. mTORC1 activity can be completely inhibited by ATP competitive inhibitors, like BEZ235 or PI103, while non-competitive ATP inhibitors, like rapamycin, inhibits mTORC1 activity only partially by interacting with the FRB (FKBP-rapamycin-binding) domain. By affecting kinetic properties of mTOR, they influence the process of glucose absorption in the cell. These types of results and information can provide data that can be complemented by the presented model. In this way, it will be possible to characterize changes in the entire signaling pathway induced by the use of mTORC1 inhibitors and evaluate the effect of this inhibition on the amount of GLUT4 in vesicles.

Increased mTORC1 activity has been also reported in many types of cancer [8]. mTOR is one of the factors influencing the development and growth of cells. Its excessive activity encourages cancer cells to further grow, divide and invade other healthy tissues. For this reason, it was decided to test mTORC1 inhibitors in cancer therapy [27], as they appeared to be an effective tool for coercing cancer cells into apoptosis. Although many mTORC inhibitors have been tested, some of them have been approved for therapy, however, their therapeutic capacity is relatively low. For this reason, they are most often used in combination with other anticancer drugs. In addition, their side effects must be considered. Palm et al. [55] demonstrated on mouse model of pancreatic cancer that rapamycin may even promote cell proliferation at poorly vascularized sites of the tumor. In view of all this information, it remains vital to study mTOR more thoroughly because its participation in cancer metabolism is undeniable [56], which is why it seems to be such an important research direction. The presented model can be used for this type of research, during the theoretical phase, where the likely results of their use can be determined using the data on the influence of new drugs on mTOR kinetics.

Conclusions

A queueing theory model of mTORC1 and mTORC2 impact on Akt-mediated cell response to insulin was prepared. The presented results show that queueing theory can effectively model the manipulation of mTORC1 kinase activity influences the amount of GLUT4 used to transport glucose inside the cell, and therefore influences the concentration of glucose in the cell. The work shows suggestions of alternative targets for treating type 2 diabetes. Due to the number of people with diabetes and the existing methods of relieving symptoms, without treating the disease, any new therapeutic target may prove to be crucial. However, it should be noted that due to the nature of the studies performed, our findings must be confirmed in clinical trials.

Supporting information

S1 File. Additional supporting information may be found in the online version of this article.

Supporting Information file contains values of literature concentrations used in the model and reaction equations and kinetic constants used in the model. The source code is freely available for download at <https://github.com/UTP-WTliE/IrsMtorcQueuesSimulation>, implemented in C# supported in Linux or MS Windows. <https://doi.org/10.1371/journal.pone.0279573.s001> (DOCX)

References

1. Saltiel AR, Kahn CR. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature*. 2001;414: 799–806. pmid:11742412
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
2. Watanabe M, Hayasaki H, Tamayama T, Shimada M. Histologic distribution of insulin and glucagon receptors. *Brazilian J Med Biol Res*. 1998;31: 243–256. pmid:9686147
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
3. Cheng Z, Tseng Y, White MF. Insulin signaling meets mitochondria in metabolism. *Trends Endocrinol Metab*. 2010;21: 589–598. pmid:20638297
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
4. Satoh S, Nishimura H, Clark AE, Kozka IJ, Vannucci SJ, Simpson IA, et al. Use of bismannose photolabel to elucidate insulin-regulated GLUT4 subcellular trafficking kinetics in rat adipose cells. Evidence that exocytosis is a critical site of hormone action. *J Biol Chem*. 1993;268: 17820–17829. pmid:8349666
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
5. Yang J, Clarke JF, Ester CJ, Young PW, Kasuga M, Holman GD. Phosphatidylinositol 3-kinase acts at an intracellular membrane site to enhance GLUT4 exocytosis in 3T3-L1 cells. *Biochem J*. 1996;313: 125–131. pmid:8546673
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
6. Mao Z, Zhang W. Role of mTOR in glucose and lipid metabolism. *Int J Mol Sci*. 2018;19: 1–14. pmid:30011848
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
7. Walters HE, Cox LS. mTORC inhibitors as broad-spectrum therapeutics for age-related diseases. *Int J Mol Sci*. 2018;19: 1–33. pmid:30096787
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
8. Zou Z, Tao T, Li H, Zhu X. MTOR signaling pathway and mTOR inhibitors in cancer: Progress and challenges. *Cell Biosci*. 2020;10: 1–11. pmid:32175074
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
9. Goltsov A, Tashkandi G, Langdon SP, Harrison DJ, Bown JL. Kinetic modelling of in vitro data of PI3K, mTOR1, PTEN enzymes and on-target inhibitors Rapamycin, BEZ235, and LY294002. *Eur J Pharm Sci*. 2017;97: 170–181. pmid:27832967
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
10. Yoon MS. The role of mammalian target of rapamycin (mTOR) in insulin signaling. *Nutrients*. 2017;9. pmid:29077002
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
11. Lee MN, Ha SH, Kim J, Koh A, Lee CS, Kim JH, et al. Glycolytic Flux Signals to mTOR through Glyceraldehyde-3-Phosphate Dehydrogenase-Mediated Regulation of Rheb. *Mol Cell Biol*. 2009;29: 3991–4001. pmid:19451232
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
12. Hahl SK, Kremling A. A comparison of deterministic and stochastic modeling approaches for biochemical reaction systems: On fixed points, means, and modes. *Front Genet*. 2016;7. pmid:27630669
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
13. Boxma O, Walraevens J. Computational methods and applications in queueing theory. *Ann Oper Res*. 2017;252: 1–2.
[View Article](#) • [Google Scholar](#)
14. Neuts MF, Chen S-Z. The infinite server queue with semi-Markovian arrivals and negative exponential services. *J Appl Probab*. 1972;9: 178–184.
[View Article](#) • [Google Scholar](#)
15. Sharma AK, Sharma GK. Queueing Theory Approach with queueing model. *Int J Eng Sci Invent*. 2013;2: 1–11.
[View Article](#) • [Google Scholar](#)
16. Qiu T, Xia F, Feng L, Wu G, Jin B. Queueing theory-based path delay analysis of wireless sensor networks. *Adv Electr Comput Eng*. 2011;11: 3–8.
[View Article](#) • [Google Scholar](#)
17. Evstigneev VP, Holyavka MG, Khrapaty S V., Evstigneev MP. Theoretical Description of Metabolism Using Queueing Theory. *Bull Math Biol*. 2014;76: 2238–2248. pmid:25142745
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
18. Clement EJ, Schulze TT, Soliman GA, Wysocki BJ, Davis PH, Wysocki TA. Stochastic simulation of cellular metabolism. *IEEE Access*. 2020;8: 79734–79744. pmid:33747671
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
19. Kloska S, Pałczyński K, Marciniak T, Talaśka T, Nitz M, Wysocki BJ, et al. Queueing theory model of Krebs cycle. *Bioinformatics*. 2021. pmid:33724355

[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)

20. Jezewski AJ, Larson JJ, Wysocki B, Davis PH, Wysocki T. A novel method for simulating insulin mediated GLUT4 translocation. *Biotechnol Bioeng*. 2014;111: 2454–2465. pmid:24917169
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
21. Martin OJ, Lee A, McGraw TE. GLUT4 distribution between the plasma membrane and the intracellular compartments is maintained by an insulin-modulated bipartite dynamic mechanism. *J Biol Chem*. 2006;281: 484–490. pmid:16269413
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
22. Tao Z, Barker J, Shi SDH, Gehring M, Sun S. Steady-state kinetic and inhibition studies of the mammalian target of rapamycin (mTOR) kinase domain and mTOR complexes. *Biochemistry*. 2010;49: 8488–8498. pmid:20804212
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
23. Leprivier G, Rotblat B. How does mTOR sense glucose starvation? AMPK is the usual suspect. *Cell Death Discov*. 2020;6: 0–4. pmid:32351714
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
24. Sangüesa G, Roglans N, Baena M, Velázquez AM, Laguna JC, Alegret M. mTOR is a key protein involved in the metabolic effects of simple sugars. *Int J Mol Sci*. 2019;20. pmid:30841536
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
25. Ardestani A, Lupse B, Kido Y, Leibowitz G, Maedler K. mTORC1 Signaling: A Double-Edged Sword in Diabetic β Cells. *Cell Metab*. 2018;27: 314–331. pmid:29275961
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
26. Vander Haar E, Lee S-I, Bandhakavi S, Griffin TJ, Kim D-H. Insulin signalling to mTOR mediated by the Akt/PKB substrate PRAS40. *Nat Cell Biol*. 2007;9: 316–323. pmid:17277771
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
27. Hosoi H, Dilling MB, Shikata T, Liu LN, Shu L, Ashmun RA, et al. Rapamycin causes poorly reversible inhibition of mTOR and induces p53- independent apoptosis in human rhabdomyosarcoma cells. *Cancer Res*. 1999;59: 886–894. pmid:10029080
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
28. Martin S, Millar CA, Lytle CT, Meerloo T, Marsh BJ, Gould GW, et al. Effects of insulin on intracellular GLUT4 vesicles in adipocytes: Evidence for a secretory mode of regulation. *J Cell Sci*. 2000;113: 3427–3438. pmid:10984434
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
29. Watson RT, Kanzaki M, Pessin JE. Regulated membrane trafficking of the insulin-responsive glucose transporter 4 in adipocytes. *Endocr Rev*. 2004;25: 177–204. pmid:15082519
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
30. Watson RT, Pessin JE. Bridging the GAP between insulin signaling and GLUT4 translocation. *Trends Biochem Sci*. 2006;31: 215–222. pmid:16540333
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
31. Clement EJ, Soliman GA, Wysocki BJ, Davis PH, Wysocki TA. Dynamic Modeling and Stochastic Simulation of Metabolic Networks. *Curr Metabolomics*. 2018;6: 49–56.
[View Article](#) • [Google Scholar](#)
32. Holland JH. Genetic Algorithms and Adaptation. In: Selfridge OG, Rissland EL, Arbib MA, editors. *Adaptive Control of Ill-Defined Systems*. Boston, MA: Springer US; 1984. pp. 317–333. https://doi.org/10.1007/978-1-4684-8941-5_21
33. Katoch S, Chauhan SS, Kumar V. A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*. *Multimedia Tools and Applications*; 2021. pmid:33162782
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
34. Alfonso F, Moreno R, Vergas J. Fatal infection after rapamycin eluting coronary stent implantation. *Heart*. 2005;91: 1–2. pmid:15894752
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
35. Weischer M, Röcken M, Berneburg M. Calcineurin inhibitors and rapamycin: Cancer protection or promotion? *Exp Dermatol*. 2007;16: 385–393. pmid:17437481
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
36. Tataranni T, Biondi G, Cariello M, Mangino M, Colucci G, Rutigliano M, et al. Rapamycin-induced hypophosphatemia and insulin resistance are associated with mTORC2 activation and klotho expression. *Am J Transplant*. 2011;11: 1656–1664. pmid:21672148
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
37. Wu X, Cao Y, Nie J, Liu H, Lu S, Hu X, et al. Genetic and pharmacological inhibition of Rheb1-mTORC1 signaling exerts cardioprotection against adverse cardiac remodeling in mice. *Am J Pathol*. 2013;182: 2005–2014. pmid:23567640
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)

38. Kozka IJ, Clark AE, Reckless JPD, Cushman SW, Gould GW, Holman GD. The effects of insulin on the level and activity of the GLUT4 present in human adipose cells. *Diabetologia*. 1995;38: 661–666. pmid:7672486
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
39. Slot JW, Geuze HJ, Gigengack S, Lienhard GE, James DE. Immuno-localization of the insulin regulatable glucose transporter in brown adipose tissue of the rat. *J Cell Biol*. 1991;113: 123–135. pmid:2007617
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
40. Tremblay F, Gagnon AM, Veilleux A, Sorisky A, Marette A. Activation of the mammalian target of rapamycin pathway acutely inhibits insulin signaling to Akt and glucose transport in 3T3-L1 and human adipocytes. *Endocrinology*. 2005;146: 1328–1337. pmid:15576463
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
41. Kleinert M, Sylow L, Fazakerley DJ, Krycer JR, Thomas KC, Oxbøll AJ, et al. Acute mTOR inhibition induces insulin resistance and alters substrate utilization in vivo. *Mol Metab*. 2014;3: 630–641. pmid:25161886
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
42. Stuart CA, Howell MEA, Baker JD, Dykes RJ, M M, Ramsey MW, et al. Cycle Training Increased GLUT4 and Activation of mTOR in Fast Twitch Muscle Fibers. *Kinesiology*. 2011;42: 423–439.
[View Article](#) • [Google Scholar](#)
43. Rajan MR, Nyman E, Kjølhede P, Cedersund G, Strålfors P. Systems-wide experimental and modeling analysis of insulin signaling through forkhead box protein O1 (FOXO1) in human adipocytes, normally and in type 2 diabetes. *J Biol Chem*. 2016;291: 15806–15819. pmid:27226562
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
44. Veilleux A, Houde VP, Bellmann K, Marette A. Chronic inhibition of the mTORC1/S6K1 pathway increases insulin-induced PI3K activity but inhibits Akt2 and glucose transport stimulation in 3T3-L1 adipocytes. *Mol Endocrinol*. 2010;24: 766–778. pmid:20203102
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
45. Magnan C, Cerasi E, Leibowitz G, Castel J, Fraenkel M, Karaca M, et al. mTOR inhibition by rapamycin prevents beta-cell adaptation to hyperglycemia and exacerbates the metabolic state in type 2 diabetes. *Diabetes*. 2008;57: 945–57. pmid:18174523
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
46. Sonntag AG, Dalle Pezze P, Shanley DP, Thedieck K. A modelling-experimental approach reveals insulin receptor substrate (IRS)-dependent regulation of adenosine monophosphate-dependent kinase (AMPK) by insulin. *FEBS J*. 2012;279: 3314–3328. pmid:22452783
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
47. Mori H, Inoki K, Opland D, Münzberg H, Villanueva EC, Faouzi M, et al. Critical roles for the TSC-mTOR pathway in β -cell function. *Am J Physiol—Endocrinol Metab*. 2009;297: 1013–1022. pmid:19690069
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
48. Shigeyama Y, Kobayashi T, Kido Y, Hashimoto N, Asahara S, Matsuda T, et al. Biphasic Response of Pancreatic β -Cell Mass to Ablation of Tuberosous Sclerosis Complex 2 in Mice. *Mol Cell Biol*. 2008;28: 2971–2979. pmid:18316403
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
49. Khamzina L, Veilleux A, Bergeron S, Marette A. Increased activation of the mammalian target of rapamycin pathway in liver and skeletal muscle of obese rats: Possible involvement in obesity-linked insulin resistance. *Endocrinology*. 2005;146: 1473–1481. pmid:15604215
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
50. Sung U, Frigerio F, Watanabe M, Picard F, Joaquin M, Sticker M, et al. Absence of S6K1 protects against age- and diet-induced obesity while enhancing insulin sensitivity. *Nature*. 2004;431: 200–205. pmid:15306821
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
51. Cunningham JT, Rodgers JT, Arlow DH, Vazquez F, Mootha VK, Puigserver P. mTOR controls mitochondrial oxidative function through a YY1-PGC-1 α transcriptional complex. *Nature*. 2007;450: 736–740. pmid:18046414
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
52. Zeng Z, Sarbassov DD, Samudio IJ, Yee KWL, Munsell MF, Jackson CE, et al. Rapamycin derivatives reduce mTORC2 signaling and inhibit AKT activation in AML. *Blood*. 2007;109: 3509–3512. pmid:17179228
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
53. Sarbassov DD, Ali SM, Sengupta S, Sheen JH, Hsu PP, Bagley AF, et al. Prolonged Rapamycin Treatment Inhibits mTORC2 Assembly and Akt/PKB. *Mol Cell*. 2006;22: 159–168. pmid:16603397
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
54. Schreiber KH, Ortiz D, Academia EC, Anies AC, Liao CY, Kennedy BK. Rapamycin-mediated mTORC2 inhibition is determined by the relative expression of FK506-binding proteins. *Aging Cell*. 2015;14: 265–273. pmid:25652038
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)
55. Palm W, Park Y, Wright K, Pavlova NN, Tuveson DA, Thompson CB. The Utilization of Extracellular Proteins as Nutrients Is Suppressed by mTORC1. *Cell*. 2015;162: 259–270. pmid:26144316
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)

56. Zaytseva YY, Valentino JD, Gulhati P, Mark Evers B. MTOR inhibitors in cancer therapy. *Cancer Lett.* 2012;319: 1–7. pmid:22261336
[View Article](#) • [PubMed/NCBI](#) • [Google Scholar](#)



OPEN

Integrating glycolysis, citric acid cycle, pentose phosphate pathway, and fatty acid beta-oxidation into a single computational model

Sylwester M. Kloska^{1✉}, Krzysztof Pałczyński², Tomasz Marciniak², Tomasz Talaśka², Beata J. Wysocki³, Paul Davis³ & Tadeusz A. Wysocki^{2,4}

The metabolic network of a living cell is highly intricate and involves complex interactions between various pathways. In this study, we propose a computational model that integrates glycolysis, the pentose phosphate pathway (PPP), the fatty acids beta-oxidation, and the tricarboxylic acid cycle (TCA cycle) using queueing theory. The model utilizes literature data on metabolite concentrations and enzyme kinetic constants to calculate the probabilities of individual reactions occurring on a microscopic scale, which can be viewed as the reaction rates on a macroscopic scale. However, it should be noted that the model has some limitations, including not accounting for all the reactions in which the metabolites are involved. Therefore, a genetic algorithm (GA) was used to estimate the impact of these external processes. Despite these limitations, our model achieved high accuracy and stability, providing real-time observation of changes in metabolite concentrations. This type of model can help in better understanding the mechanisms of biochemical reactions in cells, which can ultimately contribute to the prevention and treatment of aging, cancer, metabolic diseases, and neurodegenerative disorders.

Cellular metabolism modeling is an important but difficult task^{1,2}. The difficulty arises from the fact that compounds, which act as substrates and products in the cell's metabolic reactions, are like a system of interconnected vessels. Any change in the concentration of a compound in a cell indirectly affects other, seemingly unrelated compounds, and thus the reactions in which they participate. Many external as well as internal factors affect the course of reactions taking place in the cell, possibly accelerating, inhibiting, or blocking them. Due to the complexity of metabolism during computational modeling, it is necessary to adopt certain start and end points. Therefore, the best target for modeling seems to be those thoroughly studied metabolic pathways that are described in the scientific literature³. Nevertheless, it is necessary to determine certain simplifications that make such modeling possible. These simplifications may include the flow of metabolites between the cytoplasm and mitochondrion depending on the cell's momentary demand, or the flow between different metabolic reactions, since the vast majority of metabolites are used in several different metabolic pathways.

Metabolic models can serve scientists in better planning of experiments. They allow predicting the effects of specific conditions on cell metabolism. Thanks to the ongoing development of metabolomics and computational biology, modeling can speed up the processes of diagnosing metabolic diseases and contribute to development of effective treatment methods⁴. In addition, suitably adapted models can map what happens in a cell under inhibition induced by a specific molecule or gene knockdown. It has long been known that cancer cells reprogram their metabolism and alter activity in pathways that are major sources of energy⁵. However, issues related to metabolism and cancer are still in the orbit of researchers' interest^{6,7}. In cancer, it has been found that many cancer cells have altered metabolism, often characterized by an increased reliance on glucose and a decreased reliance on oxygen for energy production. This metabolic reprogramming allows cancer cells to proliferate and survive in a hostile environment. Targeting the metabolic pathways of cancer cells is becoming a promising new avenue for cancer therapy. Recent research has shown that metabolism plays a key role in many fields

¹Faculty of Medicine, Nicolaus Copernicus University Ludwik Rydygier Collegium Medicum, 85-094 Bydgoszcz, Poland. ²Faculty of Telecommunications, Computer Science and Electrical Engineering, Bydgoszcz University of Science and Technology, 85-796 Bydgoszcz, Poland. ³Department of Biology, University of Nebraska at Omaha, Omaha, NE 68182, USA. ⁴Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, Omaha, NE 68182, USA. ✉email: 503013@stud.umk.pl

that were previously not considered to be related to metabolism, such as aging⁸, and neurodegeneration^{9,10}. In aging and neurodegeneration, it is becoming increasingly clear that metabolic dysfunction plays a key role in the development of these conditions. For example, research has shown that aging is associated with a decline in the function of the mitochondria, the cell's power plants, which can lead to metabolic dysfunction. Similarly, many neurodegenerative diseases, such as Alzheimer's disease, are associated with metabolic dysfunction in the brain. Targeting these metabolic pathways may be a promising new approach for treating these conditions. Overall, the emerging role of metabolism in these fields highlights the importance of understanding the complex interactions between metabolism and disease. Therefore, metabolic modeling and analysis can be used in cancer therapy, as it will contribute to testing the effects of specific molecules as early as at the planning stage of experiments. Another of the advantages of using metabolic models is that they reduce the number of laboratory animals used in research. Many times this type of research can lead to long-term damage to the health of these animals or even their death^{11–13}.

The modeling method used in this work is based on queueing theory. Queueing theory is mainly used in issues related to engineering and telecommunications. However, there is evidence that it can be successfully used to model stochastic biological processes. Examples of applications of queueing theory to model biological processes include studies of signal transduction cascade in the cell¹⁴, insulin-related disorders and diseases¹⁵, glycolysis model¹⁶, tricarboxylic acid cycle (TCA cycle) model¹⁷, and the pentose phosphate pathway model¹⁸. The departure from deterministic models and the incorporation of fluctuations in metabolic simulations represent a significant advancement in our understanding of biological systems. Traditionally, deterministic models have been extensively used to study metabolic processes, assuming precise and predictable behavior. However, it is increasingly recognized that biological systems exhibit inherent stochasticity, where random fluctuations play a fundamental role in shaping cellular behavior. By implementing a flavor of the Kinetic Monte Carlo method¹⁹, similar to the Gillespie algorithm²⁰, in our simulations, we have taken a crucial step towards capturing the effects of these fluctuations. Incorporating fluctuations in metabolic simulations is of utmost importance as it allows us to bridge the gap between the deterministic models and the real-world dynamics of biological systems. Fluctuations arise from various sources such as the discreteness of molecular species, spatial heterogeneity, and the inherent randomness of molecular interactions. Ignoring these fluctuations can lead to an incomplete and biased understanding of cellular processes.

By considering the inherent stochasticity in our model, we gain valuable insights into the behavior of metabolic networks that deterministic models fail to capture. Fluctuations have been shown to influence key aspects of cellular metabolism, including reaction rates, pathway efficiency, and robustness^{21–23}. They can drive cellular decision-making, affect cellular responses to perturbations, and contribute to the emergence of complex phenomena at the system level.

Incorporating fluctuations in metabolic simulations also provides a more accurate representation of biological reality. By acknowledging the stochastic nature of cellular processes, we can better understand and reproduce experimental observations. Fluctuations play a role in generating the observed biological variability, and their inclusion in simulations allows us to better match experimental data and validate the model's predictions. Moreover, by simulating fluctuations, we can explore the effects of different sources of variability, such as noise in gene expression or environmental fluctuations, on metabolic behavior. This information is crucial for understanding how cells respond and adapt to changing conditions and for unraveling the underlying principles governing cellular decision-making²⁴.

It is important to note that incorporating fluctuations in metabolic simulations is not without challenges. Stochastic simulations can be computationally demanding, requiring specialized algorithms and efficient simulation techniques. However, advancements in computational power and the development of efficient algorithms, such as the Kinetic Monte Carlo method¹⁹, have made it increasingly feasible to simulate stochastic models at reasonable timescales. The departure from deterministic models and the incorporation of fluctuations in metabolic simulations represent a significant advancement in computational biology. By embracing the inherent stochasticity of biological systems, we gain deeper insights into the dynamics and behavior of metabolic networks, which would otherwise be overlooked by deterministic models¹⁶. Incorporating fluctuations allows us to better match experimental observations, understand biological variability, and explore the impact of stochasticity on cellular processes. These advancements pave the way for more accurate and comprehensive models of cellular metabolism and contribute to our overall understanding of complex biological systems.

The purpose of the present study was to develop an integrated computational model of the cell's energy metabolism. This model consists of reactions included in important metabolic pathways and cycles, i.e. glycolysis, the pentose-phosphate pathway (PPP), the TCA cycle, and beta-oxidation. These are the pathways that play an important role in energy metabolism of the cell. Glycolysis is a simple metabolic pathway that regulates metabolic functions of various cells²⁵, PPP is a pathway parallel to glycolysis, in which NADPH and 5-carbon sugars are generated²⁶. Beta-oxidation is a series of reactions that break down long carbon chain fatty acids in order to generate acetyl-CoA and co-enzymes used in the electron transport chain, such as FADH₂ and NADH²⁷. TCA cycle is an important metabolic pathway which uses acetyl-CoA produced in catabolic reactions of carbohydrate, fat, and protein metabolism, to generate energy²⁸. TCA cycle is a source of various important biochemical compounds used in many other metabolic reactions in the cell. The presented model enables tracking of changes in the concentrations of individual metabolites of the aforementioned pathways. The innovation of this study is that the model has been based on queueing theory, compared to ODE-based models, which are commonly used for this kind of research. Another innovation is its nature that integrates pathways related to the formation and utilization of acetyl-CoA. In addition, it was showed that artificial intelligence algorithms can be successfully used to tune coefficients of the enzyme equations.

Results

The TCA metabolites' concentration values reported in the literature were either single-number measurements or ranges (Table 1). As a result, it was necessary to select not only the mean and standard deviation of the distribution but also the measurement values from ranges that maximize maximum log-likelihood estimation. This optimization problem was solved using GA.

The last column in the table presents the Z-score of modeled substrates' concentration values regarding the estimated mean and standard deviation of the corresponding Gaussian distribution. All values of substrates, except for α -ketoglutarate, have a Z-score between -2 and 2 . As a result, they are within two sigma distance from the estimated mean. Despite not being within the range of two sigmas for α -ketoglutarate, our data still falls within the range of three sigmas.

During the experiment, glucose consumption in the cell was simulated. At the start of the model, the glucose concentration was fixed at 5 mM. This is a value in the range of normal blood glucose concentration³⁹. In the initial phase of the simulation, the course of glycolysis, PPP, and TCA cycle reactions were modeled. The product of glycolysis, pyruvate, underwent reactions that converted pyruvate to oxaloacetate or acetyl-CoA, which are metabolites of the TCA cycle. Over the course of the simulation, the glucose concentration decreased. As the glucose concentration decreased, the reactions of the glycolysis pathway were extinguished. This was due to a decrease in the probability of occurrence of glycolysis reactions and, consequently, a decrease in the speed of these reactions. As a consequence of the decrease in glycolysis activity, the probability of occurrence of reactions entering the fatty acid beta-oxidation pathway increased, which, after glucose utilization, became the main source of acetyl-CoA used in the TCA cycle. The use of GA allowed combining the reaction of enzymatic kinetics of several energetically important biochemical pathways. Due to the large differences in numerical values between consecutive reactions, as well as influence of the reactions not included in the model on reaction rates, it was necessary to tune the model. GA proved to be an effective tool in this process.

| Metabolite | Reported concentration [mmol/L] | Estimated mean | Estimated SD | Model conc. during glycolysis | Z-score |
|--------------------------|---------------------------------|----------------|--------------|-------------------------------|----------|
| Acetyl-CoA | 0.0288 ²⁹ | 0.3022 | 0.2561 | 0.071 | - 0.9028 |
| | 0.61 ³⁰ | | | | |
| | 0.07 ² | | | | |
| | 0.5 ³¹ | | | | |
| Oxaloacetate | 0.00201 ²⁹ | 0.0036 | 0.0010 | 0.005 | 1.4 |
| | 0.002–0.006 ³² | | | | |
| | 0.005 ² | | | | |
| Citrate | 0.584 ²⁹ | 0.6576 | 0.7103 | 0.184 | - 0.8581 |
| | 2 ³⁰ | | | | |
| | 0.114 ³³ | | | | |
| | 0.4 ² | | | | |
| Isocitrate cis-aconitate | 0.0321 ²⁹ | 0.02604 | 0.0060 | 0.017 | - 1.5067 |
| | 0.002–0.006 ³⁵ | | | | |
| | 0.02 ³¹ | | | | |
| α -ketoglutarate | 0.797 ²⁹ | 0.5067 | 0.1973 | 0.031 | - 2.411 |
| | 0.44 ³⁰ | | | | |
| | 0.25 ² | | | | |
| | 0.54 ³⁶ | | | | |
| | 0.004–0.013 ³⁷ | | | | |
| Succinyl-CoA succinate | 0.23 ³⁰ | 0.2989 | 0.2710 | 0.720 | 1.5538 |
| | 0.0068 ²⁹ | | | | |
| | 0.66 ³⁶ | | | | |
| | 0.36–0.91 ³⁸ | | | | |
| Fumarate | 0.485 ²⁹ | 0.6672 | 0.7496 | 0.488 | - 0.2391 |
| | 0.12 ³⁰ | | | | |
| | 0.124 ²⁹ | | | | |
| | 1.94 ³⁶ | | | | |
| Malate | 1.7 ³⁰ | 1.1137 | 0.4642 | 0.495 | - 1.3328 |
| | 1.39 ²⁹ | | | | |
| | 0.495 ³⁶ | | | | |
| | 0.87 ³⁴ | | | | |

Table 1. Statistical analysis of concentration values from literature. *SD* standard deviation.

In order to check the validity of the results generated by the computational model, they were compared to concentration values measured under experimental conditions. For this purpose, the metabolites concentration values presented in scientific publications (Table 2) on the TCA cycle were used. The TCA cycle was chosen as a reference point due to the fact that it is a well-studied metabolic cycle and represents the final stage of the presented model. Table 2 presents the averaged results for 50 calculated cycles, mimicking 50 liver cells. The simulations covered a time interval of almost three hours (10,000 seconds). Each second, 5 measurements were taken, and their results were averaged and recorded. Changes in metabolite concentrations during the course of the simulations are presented in Fig. 1. The course of changes in the concentration of individual metabolites over time is stable. Compounds whose concentrations change the most over the course of the computational simulation, such as glucose and pyruvate, were expected to behave this way, since the model does not take into account glucose external replenishment over the course of the simulation.

The results presented in Table 2 indicate the high accuracy of the computed results with respect to the concentration values measured under laboratory conditions. The “SD over mean” column shown in Table 2 refers to the 90th percentile SD instead of the maximum SD due to the occurrence of outliers in the time series (e.g. sudden changes). In the case of oxaloacetate, the “SD over mean” value is relatively high, due to the change in the infl w of this compound in the TCA cycle. Oxaloacetate in the initial phase of the simulation is supplied from two sources: (1) it is formed from pyruvate obtained in the glycolysis pathway and (2) it is formed from the acetyl-CoA conversion reaction. In the case of a longer simulation, as in the presented example, the first source related to glycolysis is extinguished, as the glucose concentration decreases, which is not kept constant in the presented model. In this model, the concentration of oxaloacetate in the long-term simulation is kept constant only by acetyl-CoA obtained by beta-oxidation of fatty acids. For the other TCA metabolites, the “SD over mean” value is relatively low, relative to the value of the calculated concentration of these metabolites at specific time points. On this basis, it can be concluded that the model is stable, and the calculated concentration is not subject to sudden, high changes.

The observed discrepancies in the metabolite ranges compared to laboratory data are a significant aspect to address in our research. In order to develop our model, we relied on data obtained from diverse literature sources. It is important to acknowledge that the measurements reported in the literature exhibit considerable variability across different studies and sources. This variability arises from factors such as variations in laboratory setups, measurement techniques, experimental conditions, and potential inter-individual differences. It is crucial to recognize that the data we employed from the literature may not necessarily represent dynamic or steady-state biological measurements. Rather, these measurements often represent snapshots of metabolite concentrations taken under specific experimental conditions that may not precisely align with the steady-state conditions assumed in our model. Consequently, inherent discrepancies can arise between the measured values and the simulated results due to these variations in experimental setups. These factors highlight the need to carefully consider and address the limitations and sources of variability when interpreting and comparing our model outputs with laboratory data.

The results of the sensitivity analysis are presented in Table 3. The impact of the variance of acetyl-CoA and α -ketoglutarate starting concentrations on substrates values at the end of each simulation was measured. It was decided to use these two metabolites as examples due to the fact that there are many various data on concentrations of these metabolites. In order to present sensitivity scores from dozens of different substrates obtained for various starting values of substances above, aggregation was used. Sensitivity scores from different substrates were concatenated into distributions and described by the distribution’s minimum, 5th percentile, median, 95th percentile, and maximum.

The change in activity of individual pathways, clearly depends on changes in glucose concentration. During the simulation run, the model strives to achieve the concentration values presented in the literature, while taking care to maintain the stability of the obtained results. We realize that the concentration of glucose in the cell under real conditions is maintained at a relatively constant level, such as through glycogenolysis or gluconeogenesis. However, due to the complexity and number of connections between biologically active molecules, the presented model does not take into account the maintenance of glucose at a constant level. By designing the model in

| Metabolite | Conc. (literature) [mmol/L] | Model conc. at starting point | Model conc. during glycolysis | Model conc. during β -oxidation | Model conc. at the end of simulation | SD over mean |
|--------------------------|-----------------------------|-------------------------------|-------------------------------|---------------------------------------|--------------------------------------|--------------|
| Acetyl-CoA | 0.07 ² | 0.070 | 0.071 | 0.060 | 0.060 | 0.002 |
| Oxaloacetate | 0.002–0.006 ³² | 0.006 | 0.005 | 0.001 | 0.001 | 0.126 |
| Citrate | 0.114 ³³ | 0.190 | 0.184 | 0.115 | 0.110 | 0.031 |
| Isocitrate cis-aconitate | 0.002–0.006 ³⁵ | 0.020 | 0.017 | 0.010 | 0.009 | 0.091 |
| α -ketoglutarate | 0.004–0.013 ³⁷ | 0.030 | 0.031 | 0.023 | 0.022 | 0.085 |
| Succinyl-CoA succinate | 0.36–0.91 ³⁸ | 0.730 | 0.720 | 0.691 | 0.690 | 0.011 |
| Fumarate | 0.485 ²⁹ | 0.485 | 0.488 | 0.490 | 0.490 | 0.012 |
| Malate | 0.495 ³⁶ | 0.495 | 0.495 | 0.489 | 0.488 | 0.010 |

Table 2. Comparison of concentration data between literature and model (mmol/L) at different time points of the simulation. Comparison of concentration data between literature and model (mmol/L) at different time points of the simulation (20th second during glycolysis and 9000th second during beta-oxidation). SD standard deviation.

| Metabolite | Minimum | 5th percentile | Median | 95th percentile | Maximum |
|-------------------------|---------|----------------|--------|-----------------|---------|
| Acetyl-CoA | 0.02 | 0.51 | 1.0 | 1.51 | 2.32 |
| α -ketoglutarate | 0.24 | 0.66 | 1.0 | 1.64 | 13.23 |

Table 3. Sensitivity analysis of impact generated by varying starting values of Acetyl-CoA and α -ketoglutarate on end values of the substrates of each simulation. The sensitivity scores were concatenated into the sensitivity distributions and described by their minimum, 5th percentile, median, 95th percentile, and maximum.

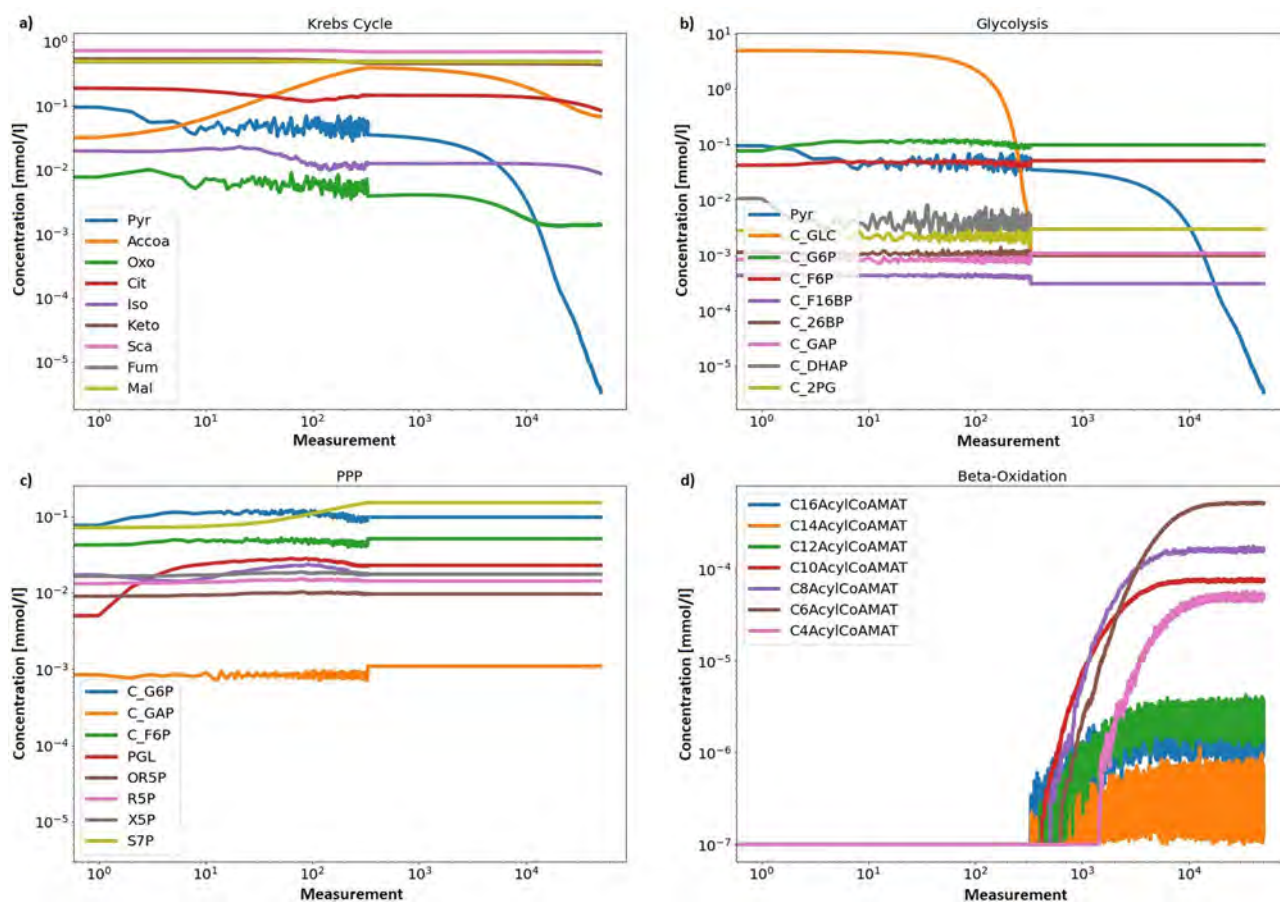


Figure 1. Visualization of the concentration change over the course of simulation in each of the modeled pathways: (a) TCA cycle, (b) Glycolysis, (c) pentose phosphate pathway, (d) beta-oxidation. The X-axis displays number of measurements. During 10,000 seconds of simulation, concentration was measured 50,000 times. The Y-axis displays the concentration of a given metabolite.

this way, a change in the activity of the pathways is highlighted, changing the source of acetyl-CoA used in the TCA cycle. The presented results indicate the stability of the system, which is dependent on the concentration of glucose in the cell. It is also reflected in the sensitivity analysis. For each measured substrate, 90 percent of scores landed in close proximity to 1.0, thus implying the model's robustness on changes in the starting values.

Discussion

This paper introduces a comprehensive model of interconnected metabolic pathways, utilizing queueing theory, with the added benefit of being able to conduct real-time calculations that are not excessively complex. The experiment can be run on a regular desktop computer, as it does not require significant computing power. By examining glucose as a case study, the study illustrated that following carbohydrate depletion, the cell shifts its metabolic activity towards alternative sources of cellular energy, such as beta-oxidation of fatty acids (and potentially protein catabolism). These sources provide the necessary acetyl-CoA for energy conversion in the TCA cycle.

In our analysis, it is important to acknowledge that while the majority of TCA metabolites exhibited Z-scores within two standard deviations of the estimated mean, α -ketoglutarate deviated slightly from this trend. This deviation may be attributed to biological variability or measurement uncertainties associated with α -ketoglutarate in specific experimental conditions. Future research could delve deeper into understanding the factors contributing to this observation and explore potential biological implications. Additionally, our study demonstrates the

effectiveness of the GA approach in optimizing the selection of data from ranges and underscores its utility in handling complex datasets with non-uniform measurement representations. This methodology can be readily applied to similar challenges in metabolomics and bioinformatics to improve the precision and reliability of data analysis, facilitating more accurate interpretations of metabolic pathways and their regulatory mechanisms.

Metabolic modeling plays a crucial role in bridging the gap between theoretical understanding and practical implementations in the field of biology. The ultimate aim of metabolic modeling is to provide insights that can guide the development of effective therapeutic approaches for metabolic disorders. By simulating and analyzing the intricate metabolic pathways within cells, we can unravel the underlying mechanisms and identify potential targets for intervention.

In this context, the queueing methodology utilized in our study offers distinct advantages over traditional methods such as ODEs and FBA. While ODEs assume continuous and deterministic behavior, the queueing methodology embraces the discrete and stochastic nature of biochemical reactions, providing a more realistic representation of cellular processes. By capturing the inherent variability and fluctuations in metabolic networks, the queueing methodology allows for a deeper understanding of the dynamic behavior and robustness of biological systems.

One key advantage of the queueing methodology is its ability to account for queueing delays and waiting times, which are essential factors in cellular processes. These delays reflect the finite capacity of cellular resources and the time required for reactants to interact and traverse various metabolic steps. By considering queueing phenomena, our methodology enables the investigation of how delays impact metabolic fluxes, reaction rates, and overall system behavior. Additionally, the queueing methodology offers unique insights into emergent properties and system-level behaviors that are challenging to capture using other methods. The inherent stochasticity and variability incorporated through the queueing approach allow for the exploration of rare events, transient behaviors, and non-equilibrium phenomena. This capability is particularly relevant in studying metabolic diseases, where small perturbations or rare events can have significant consequences for cellular function and overall health. Furthermore, the queueing methodology facilitates real-time tracking of metabolite concentrations, enabling dynamic simulations that closely mirror the temporal aspects of cellular metabolism. This temporal resolution provides a more comprehensive understanding of metabolic changes and their implications for cellular function.

By highlighting these distinctive features of the queueing methodology, we emphasize its potential in generating insights that cannot be obtained through traditional approaches like ODEs and FBA. The utilization of queueing theory enriches the toolbox of metabolic modeling, expanding the possibilities for practical applications in therapeutic development, personalized medicine, and precision interventions for metabolic disorders.

Metabolic changes are observed in various diseases, including metabolic disorders such as diabetes and obesity, as well as in the aging process⁴⁰. These conditions have garnered significant attention from researchers due to the rising prevalence of metabolic disorders and their impact on health^{41,42}. While regulatory pathways typically maintain metabolite concentrations within narrow bounds⁴³, individual metabolite levels can vary among individuals and deviate from established norms. An example of altered metabolism is found in cancer⁴⁴, where a process known as metabolic reprogramming occurs. Cancer cells exhibit a shift in energy utilization, bypassing the citric acid cycle in mitochondria and relying heavily on glycolysis, followed by lactate fermentation in the cytosol⁴⁵.

In neurodegenerative diseases like Alzheimer's and Parkinson's disease, mounting evidence suggests that mitochondrial dysfunction plays a pivotal role in disease development and progression⁹. Studies have demonstrated reduced activity of the citric acid cycle in the brains of affected individuals⁴⁶. One potential therapeutic approach for neurodegenerative diseases involves targeting the mitochondria and the citric acid cycle to improve their function. This can be achieved through various strategies, including increasing the levels of citric acid cycle enzymes such as citrate synthase or utilizing drugs that target specific enzymes within the cycle. Conversely, another approach involves reducing citric acid cycle activity by inhibiting enzymes, such as isocitrate dehydrogenase, which can help mitigate the production of reactive oxygen species (ROS) within mitochondria and reduce associated cellular damage⁴⁷. Although these approaches are experimental, they hold promise for slowing disease progression and potentially ameliorating symptoms of neurodegenerative diseases. It is important to note that further research is needed to fully comprehend the therapeutic benefits of targeting the citric acid cycle in these conditions.

To date, most scientific publications have focused on modeling macronutrient balance. These studies were focused on different dietary states, so-called intermediate fasting or semi-starvation⁴⁸, or the impact of an unbalanced diet on the development of metabolism-related diseases⁴⁹. The information they contain is extremely valuable and provides a better understanding of metabolic disorders. The purpose of our work, however, was to focus on the changes in metabolism in relation to glucose concentration at the cellular level. By combining these different types of studies, a more comprehensive understanding of metabolism and related processes such as aging can be achieved. Computational modeling of metabolic pathways also holds the potential to expedite the development of effective therapeutic approaches for alleviating metabolic disorders.

There are several limitations to the presented model. Although it is a complex model that includes 68 reactions, it does not take into account numerous other reactions in which the metabolites are involved. The impact of these unaccounted reactions was evaluated using GA (see "Materials and methods"). Another limitation is that the accuracy of the model is dependent on the literature concentrations of metabolites and the kinetic parameters utilized in the model. Therefore, the model is subject to errors that may have arisen during the determination of concentrations and other parameters, such as K_M , K_i , and V_{max} under laboratory conditions. We acknowledged this issue early on in the experiment and recognized that previously published data is something researchers must rely on and trust for the honesty of published outcomes. Consequently, we decided to use literature concentration values as initial values and compare simulation results to these values to evaluate the model's accuracy. It should

also be noted that, while the model results are consistent with literature values, we only observe the end results. This approach has the potential to accumulate errors in the middle phase of the experiment, leading to incorrect outcomes. However, the model's stability, as illustrated in the results section, is in agreement with prior studies on the different pathways incorporated in the model^{16–18}, thus reducing the likelihood of the aforementioned scenario. The presented outcomes demonstrate that the model is useful and appropriate for simulations of alterations in metabolite concentrations with high precision. In the future, we plan to refine the model and continue this research with the objective of creating an application that allows users to input their measured parameters and receive simulation outcomes for the entered values.

Our model is intentionally designed to be generic, incorporating data from various sources, tissues, and organisms due to the limitations in obtaining comprehensive and tissue-specific data from a single organism. However, we recognize the importance of tissue or cell type-specific applications in addressing specific biological questions. The modularity and flexibility of our model allow for the integration of tissue or cell type-specific data in future studies, which can enhance the relevance and applicability of our model to specific biological systems. By leveraging the power of queueing theory in conjunction with more precise and targeted data, we can achieve improved accuracy and gain deeper insights into tissue-specific metabolic dynamics. While our current study focuses on the broader implications of metabolic modeling and the advantages of queueing theory, we appreciate the reviewer's comment as it highlights an important direction for future research, which can further enhance the biological relevance and applicability of our model.

Methods

Queueing theory

While ordinary differential equations (ODEs) have been widely used in computational modeling of biological processes, there are several factors to consider that suggest they may not be the ideal method for biological simulations. One important limitation is that ODEs are deterministic in nature, failing to accurately capture the inherent stochasticity often observed in biological systems. These systems exhibit discrete and random molecular interactions, which are better represented by stochastic simulation methods such as the Gillespie algorithm or agent-based modeling. In addition, negative results can occur in the course of calculations, requiring the use of non-negative ODE solvers⁵⁰ such as in MATLAB. Furthermore, ODE models assume well-mixed conditions and neglect the spatial organization and heterogeneity commonly found in biological systems. However, spatial effects can significantly impact the dynamics of biochemical reactions. Alternative simulation methods, such as partial differential equations (PDEs) or spatial stochastic simulations, take into account the spatial aspects and may yield more accurate results for certain biological phenomena. In addition, ODE models heavily rely on precise knowledge of model parameters, including reaction rate constants and initial conditions. Yet, in many biological systems, these parameters are uncertain and can vary across individuals or experimental conditions. The presence of parameter uncertainty introduces variability and can affect the accuracy of ODE simulations. Alternative approaches like Bayesian inference or sensitivity analysis can help address parameter uncertainty and provide more robust predictions. Another consideration is the computational efficiency of ODE simulations. As mentioned earlier, ODEs can accumulate errors and become computationally demanding, especially for large-scale models or long simulation times. This computational burden restricts the exploration of complex biological systems or extensive parameter sweeps. To overcome these limitations, approximate or alternative simulation methods such as network-free methods or reduced-order modeling can offer more computationally efficient alternatives while still capturing essential dynamics. Moreover, certain biological systems exhibit emergent phenomena, which arise from collective interactions at the system level rather than being solely determined by individual molecular components. ODE models, focusing on the behavior of individual components, may fail to accurately capture these emergent properties. Other modeling techniques such as network models, agent-based modeling, or machine learning approaches can better capture these emergent behaviors and complex system-level dynamics. Considering these factors can provide researchers with a more comprehensive understanding of the limitations of ODEs in biological simulations. Exploring alternative modeling approaches that better suit the specific characteristics of the biological system under investigation will contribute to more accurate and insightful simulations.

Another approach used in the computational biology studies is flux balance analysis (FBA). FBA is a computational method used to study and analyze the metabolic capabilities of biological systems, particularly metabolic networks^{51,52}. By assuming a steady-state condition, where the rates of production and consumption of metabolites within the network are balanced, FBA optimizes an objective function, typically biomass production, while considering mass balance and reaction constraints.

FBA offers several advantages in computational biology studies. Firstly, it demonstrates predictive power by computing the optimal flux distribution that maximizes the production of a specific metabolite or biomass. This enables researchers to make inferences about the metabolic capabilities of an organism under different conditions. Furthermore, FBA is suitable for high-throughput analysis, as it can handle large-scale metabolic networks. It can explore the behavior of thousands of reactions simultaneously, providing a comprehensive understanding of cellular metabolism. This makes it particularly useful for analyzing genome-scale metabolic models and conducting extensive studies. The constraint-based framework utilized by FBA simplifies the representation of complex biochemical networks. By relying on stoichiometric constraints, thermodynamic constraints, and steady-state assumptions, FBA becomes computationally efficient and mathematically tractable. This allows researchers to model and analyze metabolic networks in a practical manner⁵³. However, FBA does have certain limitations. It assumes a steady-state condition, disregarding the temporal dynamics of metabolic networks. This means it cannot capture transient behavior or time-dependent responses of biochemical reactions, limiting its applicability in certain biological processes.

Moreover, FBA relies on several simplifying assumptions that may not hold true in all biological contexts. For instance, it assumes the absence of regulatory mechanisms and optimality of growth. These assumptions can limit the ability of FBA to capture the full complexity of cellular processes and may lead to deviations from real-world observations. Additionally, the accuracy of FBA predictions heavily relies on the completeness and accuracy of the metabolic network model used. Our knowledge of metabolic networks is still incomplete, and the absence of certain reactions or pathways in the model can affect the accuracy of FBA predictions.

In summary, FBA is a powerful computational tool for analyzing metabolic networks. It offers predictive capabilities, high-throughput analysis, and integration with experimental data. However, researchers should consider FBA's assumptions about steady-state conditions, simplified representations of cellular processes, and its inability to capture temporal dynamics. FBA finds extensive application in metabolic engineering, drug target identification, and understanding disease metabolism in computational biology studies.

However, it is important to acknowledge that the methods mentioned above are not without limitations, which prompted us to explore the application of a queueing theory-based approach. Biochemical reactions occur in living organisms in an orderly fashion, and for this reason queueing theory seems well suited for use in such simulation-computing studies. The optimized model has low computational complexity and it is possible to track changes in metabolite concentrations in real time. In addition, using queueing theory, the nature of the simulation is closer to reality, because there is no possibility for negative results to occur, just as in a cell, metabolites cannot reach negative concentrations. Thus, there is no need for artificially forcing non-negative solutions as is the case with ODEs.

The scheme of using queueing theory to model metabolite concentrations is shown in Fig. 2. The concentration of individual metabolites can be seen as a queue. Reactions affecting the increase in concentration of given queue are its inputs, while the reactions that consume the metabolite are its outputs. Processes affecting the concentration of a given metabolite that were not included in the enzyme kinetics equation were reduced to a factor determined using GA.

Various metabolic pathways, which are incorporated in the presented model can be mimicked by a composition of interconnected queues based on the Michaelis-Menten equations. The flow of metabolite concentration from one queue to another is sequential, so that a decrease in concentration in one queue will cause an increase in the next queue. Thus, a network of interrelated queues can be equivalent to a set of differential equations⁵⁴.

The utilization of queueing theory as the foundation for our metabolic simulation model aims to capture the stochastic Markovian processes that represent variations in metabolite concentrations. To obtain the average change in concentration, we average the results from multiple simulation runs. At the core of this stochastic model are the Michaelis-Menten kinetic equations, which describe the relationship between substrate-product pairs and reaction velocities.

By representing a network of interconnected queues and digitizing the concentrations $C_1(t), \dots, C_N(t)$, we can effectively simulate the system. Within this modeling framework, the arrival rates function as queues, while the service rates correspond to the reaction rates $v_{(ij)(C_1(t), \dots, C_N(t)), t}$, normalized with respect to the simulation time step, Δt_i , and the concentration increment, $\Delta(C_i(t))$, reflecting the finite change of $C_i(t)$ within Δt_i .

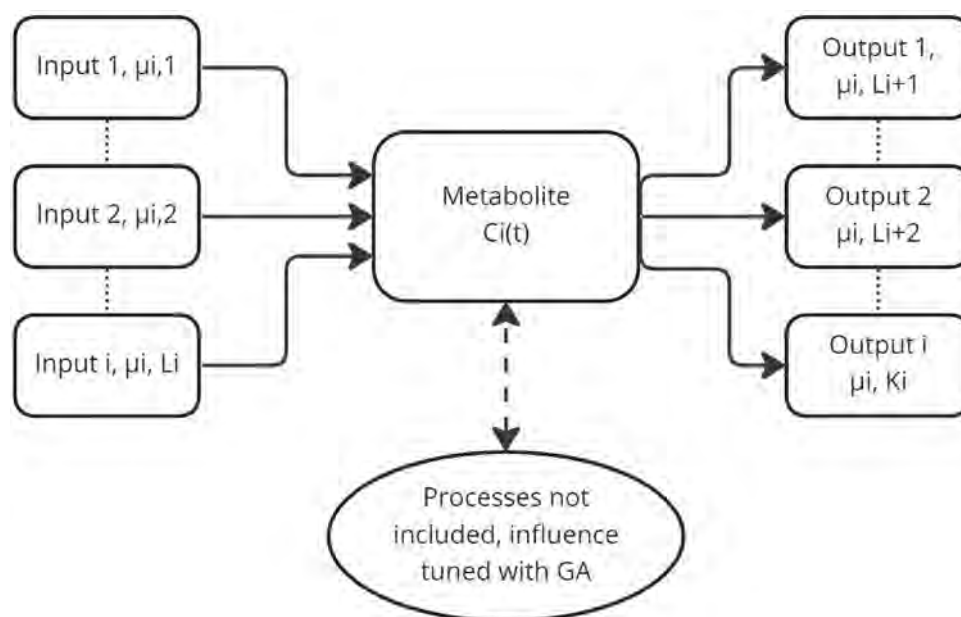


Figure 2. Example queue, which represents concentration $C_i(t)$ of the metabolite. Arrival rates are presented as inputs, while metabolite depleting rates are outputs. Due to the complexity of the metabolic network, some simplifications were adopted (see more details in the “Enzyme kinetics” section). The influence of processes not included in the model were calculated using a genetic algorithm (GA).

It is important to note that we adopt a finite time increment, Δt_i , leading to a finite concentration increment, $\Delta(C_i(t))$. While this discretization introduces quantization error, adjusting the value of $\Delta(C_i(t))$ to minimize the error may entail increased computational time due to a reduced time step, Δt_i , necessitating more simulation steps to reach the desired duration. Thus, striking a balance is essential, and we calculate the normalization of reaction rates to achieve arrival and service rates for the queues using the given formula (Eq. 1).

$$\mu_{ij} = \frac{|v_{ij}(C_1(t), \dots, C_N(t), t)| \Delta t_i}{\Delta(C_i(t))} \quad (1)$$

If the reaction rate $v_{ij}(C_1(t), \dots, C_N(t), t)$ is positive, then its corresponding normalized rate, μ_{ij} , functions as an arrival rate. Conversely, if $v_{ij}(C_1(t), \dots, C_N(t), t)$ is negative, the corresponding normalized rate, μ_{ij} , acts as a service rate. The instantaneous length of each queue embodies a potential realization of a stochastic Markovian process, capturing fluctuations in concentration for a specific metabolite. To obtain the average changes in concentration, we can compute the average of simulation results from multiple simulation runs.

To ensure the accuracy of the simulation, it is crucial to carefully select the simulation time step, Δt_i , and the concentration increment, $\Delta(C_i(t))$, such that all μ_{ij} values are less than one. The arrival and service rates are representative of probabilities for the arrival and service of $\Delta(C_i(t))$ within the given time interval. To guarantee that a single $\Delta(C_i(t))$ is processed in each time interval, the following condition must hold (Eq. 2):

$$\mu_{ij} \ll 1 \quad (2)$$

for $j = 1, \dots, K_i$ and $i = 1, \dots, N$

However, it is not necessary for both the simulation time step, Δt_i , and the concentration increment, $\Delta(C_i(t))$, to be uniform across all $i = 1, \dots, N$. Instead, they can be chosen in a manner that minimizes simulation time while ensuring satisfaction of condition described in Eq. (2). Although dynamic calculation of time increments is feasible within each step, for the present model, we have opted for constant time increments for all reactions. This decision arises from the fact that some reaction rates differ significantly in orders of magnitude, making it impractical to utilize the shortest time increment that satisfies condition described in Eq. (2) for each reaction. By employing cumulative reaction time that remains uniform for all reactions, we can uphold the conservation of molar masses.

In recognition of the stochastic nature of chemical reactions, wherein reaction rates can vary under different environmental conditions, it is possible to introduce randomness by adding Gaussian (or other) noise to the kinetic constants used for computing values of $v_{ij}(C_1(t), \dots, C_N(t), t)$. The same approach can be implemented at time instant, t_0 , for the initial concentrations, $C_{(1)}(t_0), \dots, C_{(N)}(t_0)$. This adaptation allows for a more realistic representation of the inherent fluctuations in chemical reactions, considering their sensitivity to environmental factors.

The reaction velocity serves as a macroscopic representation of numerous microscopic reactions, determining the frequency of reaction occurrence and its connection to the probabilities of increasing or decreasing specific substances. By utilizing these probabilities, we achieve a self-regulating and stochastic process that accurately simulates the behavior of biochemical pathways. The Michaelis-Menten kinetic equations calculate the probability of a reaction occurring based on substrate and product quantities, as well as kinetic constants and the duration of the time interval. These equations provide insights into the arrival and service rates in Poisson processes, where the arrival rate represents the probability of substance production, and the service rate represents the probability of substance consumption. The service time, which represents the interval between consecutive output events, is modeled using an exponential distribution. These assumptions align with classical queueing theory approaches, establishing a framework that integrates probabilities of increasing and decreasing substrates. This enables us to simulate biochemical pathways in a stochastic and self-regulating manner. In our model, the probability of a reaction occurring is determined by the Michaelis-Menten kinetic equations, where the concentration of metabolite-substrate and the kinetic constants play crucial roles. Each Michaelis-Menten equation is associated with a specific substrate and influences whether a reaction occurs at a given time point¹⁶. The reaction probability ranges from 0 to 1, and the reaction speed is considered a macroscopic representation of numerous microscopic reactions, resulting in the conversion of metabolites. The forward and reverse reaction velocities determine whether a metabolite increases or decreases. The probabilities of concentration gain and loss for each metabolite are correlated with the accumulation or increase in concentration of other metabolites. By adopting this approach, we have developed a self-regulating and stochastic model that integrates multiple metabolic pathways. The outcomes of the Michaelis-Menten equations can be interpreted as the arrival frequency and service rate in Poisson processes, with service times modeled using an exponential distribution (the time gaps between two consecutive output events). These suppositions align with traditional queueing theory methods. As a result, the count of arrivals in a specific time period ($t + \tau$) follows a Poisson distribution with a parameter $\mu(t)\tau$ (Eq. 3):

$$P[(N(t + \tau) - N(t)) = k, t] = \frac{e^{-\mu(t)\tau} (\mu(t)\tau)^k}{k!} \quad (3)$$

where

$P[(N(t + \tau) - N(t)) = k, t]$ - probability of k arrivals in the interval $(t, t + \tau)$
 $\mu(t)\tau$ - expected number of arrivals in a time interval duration of $(t, t + \tau)$

The queue processing time of metabolite increment (Eq. 4) is described by the exponential distribution of the random variable T in the terms of the rate parameter $\mu(t)$.

$$f(T; \mu(t)) = \begin{cases} \mu(t)e^{-\mu(t)T} & T \geq 0 \\ 0 & T < 0 \end{cases} \quad (4)$$

Consequently, the arrival process at the beginning of the next queue, which the output of the examined server is linked to, follows a Poisson distribution. This is a complex stochastic process involving multiple variables, which are all connected to each other. As per the Michaelis-Menten kinetic equations, the likelihood of each packet arriving at a metabolite's queue is linked to the quantity of product and inversely related to the quantity of substrate, leading to a self-regulating system that adjusts to the discrepancies of metabolites and ensures balance between arrivals and departures in every queue. One of the advantages of basing the model on queueing theory is the possibility for its further development and addition of more reactions/metabolic pathways without interfering with the previously optimized reactions. This is particularly interesting because the model can be developed with further metabolomics discoveries or combined with pathways not included in this study.

Enzyme kinetics

The data used in the model for the values of metabolite concentrations and kinetic constants: K_M (Michaelis constant), K_i (inhibition constant), V_{max} (maximum velocity), were obtained from scientific publications. It should be noted that these constants are not absolute values, but rely heavily on experimental conditions. In a seminal paper it was shown that modeling of yeast glycolysis requires actual redetermination of kinetic parameters under identical conditions for all enzymes⁵⁵. However, the approach presented in this work aims to demonstrate a model that can be improved with further development of metabolomics, based on new, more accurate data supported by the application of GA. The collected data were used to describe metabolic reactions with Michaelis-Menten equations (Eq. 5) of enzyme kinetics. The model consisted of 68 enzymatic reactions of the form:

$$v(t) = \frac{V_f \frac{S_1(t)S_2(t)}{K_{S_1}K_{S_2}} - V_r \frac{P_1(t)P_2(t)}{K_{P_1}K_{P_2}}}{\left(1 + \frac{S_1(t)}{K_{S_1}} + \frac{P_1(t)}{K_{P_1}}\right)\left(1 + \frac{S_2(t)}{K_{S_2}} + \frac{P_2(t)}{K_{P_2}}\right)} \quad (5)$$

where

- $v(t)$ - reaction speed,
- V_f - forward reaction speed,
- V_r - reverse reaction speed,
- $S_1(t), S_2(t), \dots, S_x(t)$ - substrate concentration in mmol/L,
- $P_1(t), P_2(t), \dots, P_x(t)$ - substrate concentration in mmol/L,
- $K_{S_1}, K_{S_2}, \dots, K_{S_x}$ - kinetic constant of substrate,
- $K_{P_1}, K_{P_2}, \dots, K_{P_x}$ - kinetic constant of product.

It is assumed that all concentration values are sampled from Gaussian distribution specific to the type of examined concentration. The distributions were estimated using maximum log likelihood estimation⁵⁶, given by the following equations (Eqs. 6 and 7):

$$\mu = \arg \max_{\mu} \sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma^2}\right)}\right) \quad (6)$$

$$\sigma = \arg \max_{\sigma} \sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma^2}\right)}\right) \quad (7)$$

Based on the kinetic equations, the probability of occurrence of each reaction was inferred, as described in the [Materials and methods](#) section related to Queueing theory. If the probability indicated that a reaction occurred, there was a decrease in the concentration of the metabolite that acted as a substrate in that reaction, while increasing the concentration of the metabolite that acted as a product. This is how the various reactions of the metabolic pathways included in the model gradually occurred, which were glycolysis, the pentose-phosphate pathway, the TCA cycle, and beta-oxidation. In case of missing literature data on reverse reaction speed, we applied the assumption of⁵⁷ which describes a reverse reaction as 100x slower than the forward reaction. In few cases where the literature review did not provide enzyme kinetic data, the concentrations of two adjacent metabolites were summed and combined into a single metabolite (queue). In such a case, enzyme kinetics data on the second metabolite of the pair were used¹⁷. In the study presented here, such a situation occurred twice, when describing kinetic reactions involving isocitrate and cis-aconitate as well as succinyl-CoA and succinate, which are the metabolites of TCA cycle. To model such a complex metabolic network, it was necessary to establish a specific, rigid framework and scope of model coverage. The influence of cellular processes that are not directly included in the kinetic equations, such as the flow of a metabolite between compartments of a cell, was finetuned using a genetic algorithm (GA)¹⁷.

Genetic algorithm

The equations were supplemented with coefficients selected using a GA. The choice of the algorithm was made arbitrarily, due to its effectiveness in previous similar studies that have been conducted^{17,18}. This procedure was intended to allow combining the reactions of pathways whose computational values of individual reactions can differ markedly. The GA plays a crucial role in our study by optimizing the parameter values within the ranges reported in the literature. It is important to note that these parameter values can vary significantly between

different studies and cell types. In order to achieve system stability and ensure consistency with experimental values, the GA searches for parameter values that allow the model to approximate the observed behavior. By employing the GA, we aim to find parameter values that not only make the system stable but also provide results that are consistent or approximate to experimental values. The algorithm iteratively explores the parameter space, evaluating different combinations of parameter values, and selecting those that best align with the experimental data.

The loss function was calculated with the use of ‘chromosomes’ that consist of two parameters: (1) metabolite’s concentration described in the literature and (2) a current optimization stage of the simulations. There are one hundred ‘chromosomes’ in the population, each of which is a potential solution for the table of kinetic constants. Evaluation of the ‘chromosome’ involves using its ‘genes’ as the values of constants parametrizing Michaelis-Menten equations. In this process, the simulation time series is generated.

The resulting time series is sampled at fixed time stamps in order to compare simulated results with real-life experiments results registered in the literature. The loss function quantifying fitness of the ‘chromosomes’ is the sum of squares of the distances of the sampled points from simulation time series to the literature results (Eq. 8).

$$g_p : \hat{X}, X \rightarrow \sum_{i=1}^{|\hat{X}|} \left(\frac{\hat{X}_i - X_i}{X_i} \right)^2 \quad (8)$$

where

g_p - subfunction that penalizes the difference between two vectors in relation to second vector,

X - vector of substrate concentrations described by a literature,

\hat{X} - vector of substrate concentrations obtained by evaluation.

The loss function is designed to guide the GA to identify a ‘chromosome’ with a table of kinetic constants that leads to stable concentrations of products and minimizes the distance between initial values and stable points, which generates computational results that are closest to those obtained in laboratory measurements. Evaluating one ‘chromosome’ entails running a simulation using its set of genes as the table of kinetic constants. The simulation function returns the values of substrate concentrations at each second. This table is used by the equation to determine the ‘chromosome’s’ score. The function calculates the average vector of the last 100 recordings and computes the absolute difference with the initial simulation concentrations. In the final step, the average of the differences is calculated. The ‘chromosome’ that minimizes this function is selected as the optimal table of kinetic constant values. The evaluation of each ‘chromosome’ is done by simulating the model for the first hour. There are 100 ‘chromosomes’ in the population at each step of optimization, and only the 10 sets of constants that minimize the fitness function are selected for reproduction. The reproductive algorithm is a variation of the standard crossover with an additional mechanism to prevent finding a trivial solution to minimize the loss function problem, which is to zero the probability of every reaction. The main disadvantage of the fitness function described above is the existence of a trivial solution for its minimization problem. If the ‘chromosome’ contains only zeros, then no reaction would occur, so the settling points of concentrations of products in the model would have the same values as initial concentrations, thus finding a global minimum. To prevent the GA from converging to this solution, the reproduction mechanism requires that each reaction at $t = 0$ has a probability of being performed between 1% and 10%. Reaction and balancing flow rates have ranges from 1 to 10% at the beginning of the simulation, which starts from substrate concentration values described in the literature. Applying these constraints to the reaction rates prevents them from being zeroed at the start and also prevents saturation of reactions. The reproduction algorithm has a 10% chance to perform a mutation with the mutation amplitude equal to 1.0. The optimization performed with GA was based on experimental measurements. The relative square error between subsequent values of the obtained vector and reference vector were used to calculate the penalty subfunction. To enforce equal contributions of all substrates in the optimization process, division by the value from the reference vector was performed.

Sensitivity analysis

The resulting simulations of the trained model were subjected to the variance-based sensitivity analysis. It is used to analyze the sensitivity of a model’s output to changes in the input variables. It is based on the idea that the variance of the output of a model can be used to measure the model’s sensitivity to changes in the input variables (Eq. 9):

$$S_j = \left| \frac{V(E_Y|\mu_j)}{V(Y)} \right| \quad (9)$$

where

E_Y - expected value of the signal Y ,

$V(Y)$ - variance of signal Y ,

$V(E_Y|\mu_j)$ - a variance of signal Y generated using input value j .

In this paper, Y represents a set of values of one substrate at the end of each simulation. Each value is a result of a simulation conducted using specific starting values denoted as j . The sensitivity analysis was conducted for each substrate making the cell’s measurable state.

Data availability

All data generated or analysed during this study are included in this published article and its supplementary information files. The datasets generated and/or analysed during the current study are available in the GitHub

repository, <https://github.com/UTP-WTIE/CellEnergyMetabolismModel>, DOI:10.5281/zenodo.7585089, implemented in C# supported in Linux or MS Windows.

Received: 9 May 2023; Accepted: 31 August 2023

Published online: 02 September 2023

References

- Ederer, M. *et al.* A mathematical model of metabolism and regulation provides a systems-level view of how *Escherichia coli* responds to oxygen. *Front. Microbiol.* **5**, 124 (2014).
- Nazaret, C., Heiske, M., Thurley, K. & Mazat, J.-P. Mitochondrial energetic metabolism: A simplified model of TCA cycle with ATP production. *J. Theor. Biol.* **258**, 455–464 (2009).
- Becker, S. A. & Palsson, B. O. Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.* **4**, e1000082 (2008).
- Mardinoglu, A. *et al.* Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Mol. Syst. Biol.* **9**, 649 (2013).
- Phan, L. M., Yeung, S.-C.J. & Lee, M.-H. Cancer metabolic reprogramming: Importance, main features, and potentials for precise targeted anti-cancer therapies. *Cancer Biol. Med.* **11**, 1 (2014).
- Pal, S., Sharma, A., Mathew, S. & Jaganathan, B. Targeting cancer-specific metabolic pathways for developing novel cancer therapeutics. *Front. Immunol.* **13**, 955476 (2022).
- Perri, F. *et al.* Cancer cell metabolism reprogramming and its potential implications on therapy in squamous cell carcinoma of the head and neck: A review. *Cancers* **14**, 3560 (2022).
- Jang, J. Y. *et al.* The role of mitochondria in aging. *J. Clin. Investig.* **128**, 3662–3670 (2018).
- Han, R., Liang, J. & Zhou, B. Glucose metabolic dysfunction in neurodegenerative diseases—new mechanistic insights and the potential of hypoxia as a prospective therapy targeting metabolic reprogramming. *Int. J. Mol. Sci.* **22**, 5887 (2021).
- Muddapu, V. R., Dharshini, S. A. P., Chakravarthy, V. S. & Gromiha, M. M. Neurodegenerative diseases—is metabolic deficiency the root cause?. *Front. Neurosci.* **14**, 213 (2020).
- Hajar, R. Animal testing and medicine. *Heart Views Off. J. Gulf Heart Assoc.* **12**, 42 (2011).
- Hawkins, P. *et al.* *Avoiding Mortality in Animal Research and Testing* (University of Cambridge, RSPCA Research Animals Department, 2019).
- Lynch, J. & Slaughter, B. Recognizing animal suffering and death in medicine. *West. J. Med.* **175**, 131 (2001).
- Tsuruyama, T. Kullback-Leibler divergence of an open-queueing network of a cell-signal-transduction cascade. *Entropy* **25**, 326 (2023).
- Uygulanmasi, İ. An application of queueing theory to the relationship between insulin level and number of insulin receptors. *Türk Biyokimya Dergisi Turk. J. Biochem.* **32**, 32–38 (2007).
- Clement, E. J. *et al.* Stochastic simulation of cellular metabolism. *IEEE Access* **8**, 79734–79744 (2020).
- Kloska, S. *et al.* Queueing theory model of Krebs cycle. *Bioinformatics* **37**, 2912–2919 (2021).
- Kloska, S. M. *et al.* Queueing theory model of pentose phosphate pathway. *Sci. Rep.* **12**, 4601 (2022).
- Guang, W. Application of queueing theory with monte Carlo simulation to the study of the intake and adverse effects of ethanol. *Alcohol Alcohol.* **33**, 519–527 (1998).
- Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).
- Cheong, R., Rhee, A., Wang, C. J., Nemenman, I. & Levchenko, A. Information transduction capacity of noisy biochemical signaling networks. *Science* **334**, 354–358 (2011).
- Kiviet, D. J. *et al.* Stochasticity of metabolism and growth at the single-cell level. *Nature* **514**, 376–379 (2014).
- Selimkhanov, J. *et al.* Accurate information transmission through dynamic biochemical signaling networks. *Science* **346**, 1370–1373 (2014).
- Kitano, H. Systems biology: A brief overview. *Science* **295**, 1662–1664 (2002).
- Guo, X. *et al.* Glycolysis in the control of blood glucose homeostasis. *Acta Pharm. Sin.* **B 2**, 358–367 (2012).
- Alfarouk, K. O. *et al.* The pentose phosphate pathway dynamics in cancer and its dependency on intracellular pH. *Metabolites* **10**, 285 (2020).
- Houten, S. M. & Wanders, R. J. A general introduction to the biochemistry of mitochondrial fatty acid β -oxidation. *J. Inherit. Metab. Dis.* **33**, 469–477 (2010).
- Ponizovskiy, M. Role of Krebs cycle in mechanism of stability internal medium and internal energy in an organism in norm and in mechanism of cancer pathology. *Mod. Chem. Appl.* **4**, 1–8 (2016).
- Park, J. O. *et al.* Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. *Nat. Chem. Biol.* **12**, 482–489 (2016).
- Bennett, B. D. *et al.* Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nat. Chem. Biol.* **5**, 593–599 (2009).
- Milo, R., Jorgensen, P., Moran, U., Weber, G. & Springer, M. Bionumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Res.* **38**, D750–D753 (2010).
- Siess, E. A., Kientsch-Engel, R. I. & Wieland, O. H. Concentration of free oxaloacetate in the mitochondrial compartment of isolated liver cells. *Biochem. J.* **218**, 171–176 (1984).
- Psychogios, N. *et al.* The human serum metabolome. *PLoS One* **6**, e16957 (2011).
- Ahn, E., Kumar, P., Mukha, D., Tzur, A. & Shlomi, T. Temporal fluxomics reveals oscillations in TCA cycle flux throughout the mammalian cell cycle. *Mol. Syst. Biol.* **13**, 953 (2017).
- Hoffmann, G. *et al.* Physiology and pathophysiology of organic acids in cerebrospinal fluid. *J. Inherit. Metab. Dis.* **16**, 648–669 (1993).
- Mogilevskaya, E., Demin, O. & Goryanin, I. Kinetic model of mitochondrial Krebs cycle: Unraveling the mechanism of salicylate hepatotoxic effects. *J. Biol. Phys.* **32**, 245–271 (2006).
- Kohlschütter, A. *et al.* A familial progressive neurodegenerative disease with 2-oxoglutaric aciduria. *Eur. J. Pediatr.* **138**, 32–37 (1982).
- Hansford, R. G. & Johnson, R. N. The steady state concentrations of coenzyme a-sh and coenzyme a thioester, citrate, and isocitrate during tricarboxylate cycle oxidations in rabbit heart mitochondria. *J. Biol. Chem.* **250**, 8361–8375 (1975).
- Saltiel, A. R. & Kahn, C. R. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* **414**, 799–806 (2001).
- Guo, J. *et al.* Aging and aging-related diseases: From molecular mechanisms to interventions and treatments. *Signal Transduct. Target. Ther.* **7**, 391 (2022).
- Ahmad, E., Lim, S., Lamprey, R., Webb, D. R. & Davies, M. J. Type 2 diabetes. *Lancet* **400**, 1803–1820 (2022).
- Nonguierna, E. *et al.* Improving obesogenic dietary behaviors among adolescents: A systematic review of randomized controlled trials. *Nutrients* **14**, 4592 (2022).

43. van Beek, J. H., Kirkwood, T. B. & Bassingthwaighe, J. B. Understanding the physiology of the ageing individual: Computational modelling of changes in metabolism and endurance. *Interface Focus* **6**, 20150079 (2016).
44. Faubert, B., Solmonson, A. & DeBerardinis, R. J. Metabolic reprogramming and cancer progression. *Science* **368**, eaaw5473 (2020).
45. Warburg, O. The metabolism of carcinoma cells. *J. Cancer Res.* **9**, 148–163 (1925).
46. Johri, A. & Beal, M. F. Mitochondrial dysfunction in neurodegenerative diseases. *J. Pharmacol. Exp. Ther.* **342**, 619–630 (2012).
47. Trifunovic, A. & Larsson, N.-G. Mitochondrial dysfunction as a cause of ageing. *J. Intern. Med.* **263**, 167–178 (2008).
48. Hall, K. D. Computational model of in vivo human energy metabolism during semistarvation and refeeding. *Am. J. Physiol. Endocrinol. Metab.* **291**, E23–E37 (2006).
49. Rozendaal, Y. J., Wang, Y., Hilbers, P. A. & van Riel, N. A. Computational modelling of energy balance in individuals with metabolic syndrome. *BMC Syst. Biol.* **13**, 1–14 (2019).
50. Shampine, L. F., Thompson, S., Kierzenka, J. & Byrne, G. Non-negative solutions of odes. *Appl. Math. Comput.* **170**, 556–569 (2005).
51. Lee, J. M., Gianchandani, E. P. & Papin, J. A. Flux balance analysis in the era of metabolomics. *Brief. Bioinform.* **7**, 140–150 (2006).
52. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis?. *Nat. Biotechnol.* **28**, 245–248 (2010).
53. Raman, K. & Chandra, N. Flux balance analysis of biological systems: Applications and challenges. *Brief. Bioinform.* **10**, 435–449 (2009).
54. Massey, W. A. Asymptotic analysis of the time dependent m/m/1 queue. *Math. Oper. Res.* **10**, 305–327 (1985).
55. Teusink, B. *et al.* Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur. J. Biochem.* **267**, 5313–5329 (2000).
56. Rossi, R. J. *Mathematical Statistics: An Introduction to Likelihood Based Inference* (Wiley, 2018).
57. Singh, V. K. & Ghosh, I. Kinetic modeling of tricarboxylic acid cycle and glyoxylate bypass in *Mycobacterium tuberculosis*, and its application to assessment of drug targets. *Theor. Biol. Med. Model.* **3**, 1–11 (2006).

Acknowledgements

This work was funded by the National Science Center (NCN) of Poland in terms of Opus-17 Program with grant number 2019/33/B/ST6/00875 awarded to TAW.

Author contributions

S.M.K., T.W. and B.W. conceived the idea for the model and described the theoretical background. K.P. implemented the algorithms, under the supervision of T.T. and T.M., T.W., T.M. and P.D. supervised the project and coordinated the research team. S.M.K. and K.P. wrote the paper. All authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.


Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-41765-3>.

Correspondence and requests for materials should be addressed to S.M.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023, corrected publication 2023

Article

IoT Application of Transfer Learning in Hybrid Artificial Intelligence Systems for Acute Lymphoblastic Leukemia Classification

Krzysztof Pałczyński ¹, Sandra Śmigiel ^{2,*}, Marta Gackowska ¹, Damian Ledziński ¹,
Sławomir Bujnowski ¹ and Zbigniew Lutowski ¹

- ¹ Faculty of Telecommunications, Computer Science and Electrical Engineering, Bydgoszcz University of Science and Technology, 85-796 Bydgoszcz, Poland; krzysztof@palczynski.com.pl (K.P.); marta.gackowska@pbs.edu.pl (M.G.); damian.ledzinski@pbs.edu.pl (D.L.); slawomir.bujnowski@pbs.edu.pl (S.B.); zbigniew.lutowski@pbs.edu.pl (Z.L.)
- ² Faculty of Mechanical Engineering, Bydgoszcz University of Science and Technology, 85-796 Bydgoszcz, Poland
- * Correspondence: sandra.smigiel@utp.edu.pl; Tel.: +48-52-340-8346

Abstract: Acute lymphoblastic leukemia is the most common cancer in children, and its diagnosis mainly includes microscopic blood tests of the bone marrow. Therefore, there is a need for a correct classification of white blood cells. The approach developed in this article is based on an optimized and small IoT-friendly neural network architecture. The application of learning transfer in hybrid artificial intelligence systems is offered. The hybrid system consisted of a MobileNet v2 encoder pre-trained on the ImageNet dataset and machine learning algorithms performing the role of the head. These were the XGBoost, Random Forest, and Decision Tree algorithms. In this work, the average accuracy was over 90%, reaching 97.4%. This work proves that using hybrid artificial intelligence systems for tasks with a low computational complexity of the processing units demonstrates a high classification accuracy. The methods used in this study, confirmed by the promising results, can be an effective tool in diagnosing other blood diseases, facilitating the work of a network of medical institutions to carry out the correct treatment schedule.

Keywords: hybrid artificial intelligence system; MobileNet v2; IoT; low-resource dataset; lymphocyte cells; leukemia; ALL-IDB database



Citation: Pałczyński, K.; Śmigiel, S.; Gackowska, M.; Ledziński, D.; Bujnowski, S.; Lutowski, Z. IoT Application of Transfer Learning in Hybrid Artificial Intelligence Systems for Acute Lymphoblastic Leukemia Classification. *Sensors* **2021**, *21*, 8025. <https://doi.org/10.3390/s21238025>

Academic Editor: Loris Nanni

Received: 1 November 2021

Accepted: 27 November 2021

Published: 1 December 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Acute lymphoblastic leukemia (ALL) comprises a group of lymphoid neoplasms that morphologically and immunophenotypically resemble B- and T-lineage precursor cells. ALL is the most common neoplasm in children, with a peak incidence between the ages of 2 and 5 years [1], whereas it is scarce in older patients (over 60 years of age) [2]. Diagnosis includes mainly a microscopic examination of blood and bone marrow [3].

The diagnosis of leukemias, including acute myeloid leukemia, requires standardized methods of classification. One of the most commonly used and oldest methods is the French-American-British (FAB) morphological classification [4]. For lymphoblastic leukemia, based on cytological features and the degree of heterogeneity in the distribution of leukemic cells according to the FAB classification, the following types are distinguished: L1, L2, and L3. The characteristics of type L1 are the predominance of small cells, homogeneous nuclear chromatin, and a regular nucleus shape with possible cleavages present. The nuclei are invisible or small and inconspicuous; the amount of cytoplasm is scanty; and the cytoplasmic vacuolization is variable. Deep cytoplasmic basophilia is uncommon.

The L2 type is characterized by a large and heterogeneous cell size and variable heterogeneous nuclear chromatin in each case. The shape of the nucleus is irregular, and

the nucleolus is at least one and is often large. The amount of cytoplasm is variable but is often moderately abundant. The cytoplasmic vacuolization is variable.

The characteristics of the L3 type are a large and uniform cell shape and a finely spotted and uniform chromatin of the nucleus. The nucleus has a regular oval-round shape. Nucleoli are prominent and one or more are vesicular. The amount of cytoplasm is moderately abundant, and the basophilia of the cytoplasm is intense. The challenge in the correct classification is that, as described above, blood cells differ from one another in terms of cytological features and the degree of heterogeneity in their distribution. The features that allow the differentiation of malignant cells and the type of the disease are, among others: the amount of cytoplasm, cell vacuolization, and the shape and size of the cell nucleus or nucleolus. The traditional method is to manually analyze the differences and observe the cells under an electron microscope by an experienced physician. Correct manual classification requires both experience and specialist knowledge. Therefore, there may be some differences between the results obtained. Thanks to artificial intelligence methods, it is possible to speed up the work of medics and increase the effectiveness and repeatability of results.

Artificial intelligence is a widely discussed issue in the world of science and technology for solving engineering problems. However, it is essential to realize that recent research in this area presents advanced applications of artificial intelligence in fields other than medicine, including computer science for developing new methods and algorithms and [5,6] in petroleum engineering [7] or even in civil engineering [8]. In this study, a hybrid artificial intelligence solution was used in medicine, and, at the same time, it is a promising method that can be used in IoT networks.

Some studies focus on image-segmentation methods to locate white blood cells on a microscopic image. In [9], input images were converted from RGB color space to haematoxylin-eosin-DAB (HED) space. Then bilateral filter and canny edge segmentation were used to extract individual lymphocytes. A watershed algorithm was finally used to determine the seed of each region. This method showed an accuracy of over 90%, with low computational complexity and execution time. In turn, the work [10] used the conversion of RGB to CMYK and L^*a^*b and the clustering algorithm K-means; in post-processing, dilation and erosion were used. The results obtained in the experiments had a Kappa index of 0.9306 in the ALL-IDB 2, 0.8603 in the BloodSeg, and 0.9119 in the leukocytes database.

However, the main challenge in diagnosing the disease is the correct classification of malignant lymphocytes. The study [11] proposed the architecture of deep neural networks using the AlexNet model from CNN, and it used softmax to classify acute lymphoblastic leukemia into its subtypes and normal state. The method used also included transfer learning. The segmentation approach based on the simple threshold method was used to prepare the data to distinguish the region of interest. For the developed method and the test set of 330, the accuracy was 97.78%. In turn, the authors in [12], apart from AlexNet, used ImageNet, and, for 33 images from the ALL-IDB database, the system correctly identified 94.1% of lymphoblasts. Convolutional neural network ResNeXt50 with squeeze-and-excitation modules was used in [13] to classify ALL. Initially, the network was pre-trained on ImageNet. An accuracy of 89.7% was achieved.

On the other hand, the authors of the work [14] used the convolutional neural network to classify types of leukemia, such as AML, CLL, CML, and ALL, and healthy patients. In addition, data augmentation was used to diversify the data set. As a result, an 81% efficiency was achieved with 231 test samples in classifying all leukemia subtypes. In addition, cross-validation was used in all experiments.

The authors [15] proposed the Siamese network-based few-shot learning method to classify leukocytes. The Siamese network described in the work contains two convolutional neural subnets with the same structure to know the vector of input images and to share weights. In addition, a two-way one-shot support set was used, which was used as additional information supporting the classification. The average accuracy of the classification of basophil and eosinophil cells using the Siamese network was 89.66%.

The work [16] describes the automatic classification of leukocytes. This method can be divided into three main parts. Initially, white blood cells (WBC) are isolated from the microscopic examination of blood using R–B conversion, threshold segmentation, and binarization. These include eosinophils, basophils, neutrophils, monocytes, and lymphocytes. The PRICoLBP function was then used to reflect the granularity of eosinophils and basophils, which increased their discriminatory power with other WBC types. Then, the stage for which convolutional neural networks were used was the isolation of the constants of three kinds of WBC: neutrophils, monocytes, and lymphocytes. CNN is a special feedforward neural network that consists of several convolutional layers and pooling layers. Finally, with the help of the Random Forest algorithm, the three remaining WBCs were classified. The developed method allowed for an average detection accuracy of 92.8%, while the lowest accuracy was demonstrated when recognizing lymphocytes. In [17], the green component/channel of the RGB microscopic image (input image) was extracted at the beginning. Then, the threshold segmentation, the opening operator, and the border-cleaning techniques were applied to the obtained binary image. Next, the bounding box technique was used to trim each WBC to a single image, while the cosine transform extracted textures and features. Finally, kNN, SVM, and naïve Bayes were used to segregate normal and abnormal cells. The classification of the disease was 97.45%.

This study focused on researching the application of transfer learning by using publicly available pre-trained neural networks as a significant part of hybrid artificial intelligence systems to offset the shortage of domain-specific data and low computational capabilities. In this work, the MobileNet v2 [18] network pre-trained on the ImageNet dataset was used for encoding images into small feature vectors, making them processable for CPU-friendly machine learning models like XGBoost [19] or Random Forest [20]. The MobileNet v2 architecture was employed because it was optimized for small processing units like mobile CPUs or IoT. The results were compared with the bare MobileNet v2 network repurposed for this task and the designed convolutional neural network as a baseline for comparison. Due to the use of advanced artificial intelligence solutions, it was possible to correctly classify and differentiate the disease and the correct state in the diagnosis of ALL. This could effectively diagnose their blood diseases and facilitate a network of medical facilities to undertake the proper treatment schedule. A novelty in this article was the use of hybrid artificial intelligence. The neural network was pre-trained on a vast dataset and then encoded data in a specific set. Then, machine learning models were trained on coded and reduced data. This gave a great advantage for use in IoT networks while demonstrating high classification accuracy.

This article is organized as follows. Section 2 closely describes the methods, the architectures of the hybrid artificial intelligence system, and the previously carried-out image processing. Then, Section 3 presents the results of the research. Then, the discussion is given in Section 4. Finally, Section 5 concludes the article and provides a look at further studies on this topic.

2. Materials and Methods

The methodology of the research described in this paper is depicted in Figure 1. In the first step, the data from the ALL-IDB database were sampled. Then, depending on the model of the experiment, sampled images were either preprocessed using augmentation or left intact. Depending on the selected experiment options, the color modification was not applied during the augmentation process to test the impact of timbre change on overfitting prevention. In the next step, regardless of whether augmentation was used, the z-score normalization was performed into images into the work characteristic of the MobileNet v2 network. At the end of this step, data preprocessing was finished, and the images were ready to be interpreted by the artificial intelligence system.

The prepared data entered the system by being processed by a neural network, serving as an encoder, to extract feature vectors from the input images. Then, data encoded in feature vectors are passed to the classification module, which is explained further in the

article as the “head” for classification. The “head” module is either a one-layer fully-connected neural network or one of three machine learning models: XGBoost, Random Forest, or Decision Tree. This process is described in detail in the section “Hybrid Artificial Intelligence system.” Finally, in the last step, the results of the artificial intelligence system are evaluated.

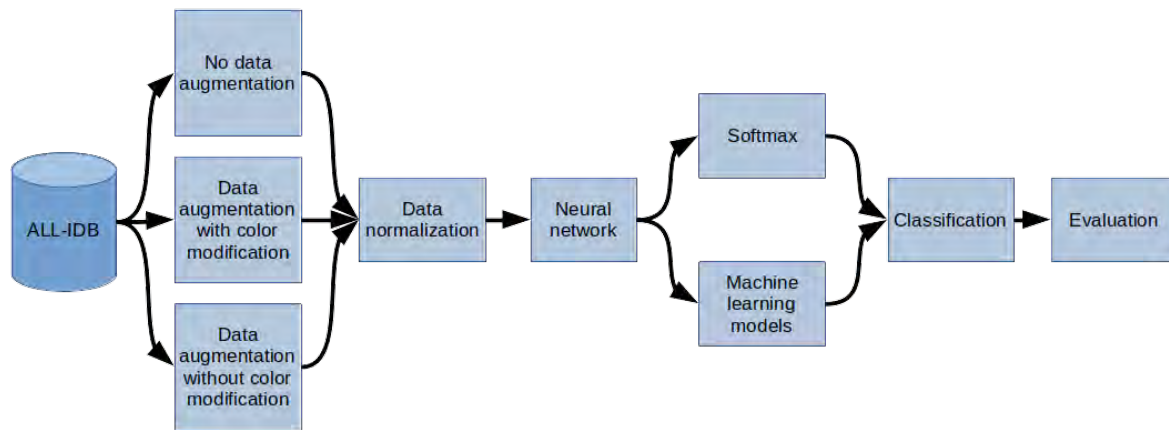


Figure 1. General overview diagram of the method.

2.1. ALL-IDB Database

The study used images of lymphocyte cells, healthy patients, and patients with acute lymphocytic leukemia. The dataset was an ALL-IDB dataset, which was downloaded with the owner’s consent [17,21–23]. The ALL-IDB dataset is a public dataset of microscopic images of peripheral blood cells that have been developed for segmentation, evaluation, and classification. The data contained in the database are considered reliable, as oncologists annotate them. The ALL-IDB database has two distinct versions (ALL-IDB1 and ALL-IDB2). In this study, experiments were performed on images from ALL-IDB2. ALL-IDB2 is a set of excised regions of interest from blood-smear images taken from healthy patients and leukemia patients, who belong to the ALL-IDB1 dataset. The ALL-IDB2 dataset is a subset of 260 segmented images, with 50% containing normal leukocytes and the remaining containing malignant cells (Figures 2 and 3).

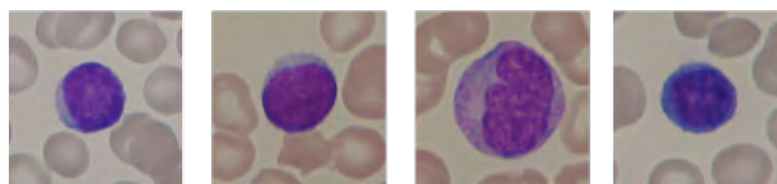


Figure 2. An example of segmented lymphocytes belonging to the non-leukemia class.

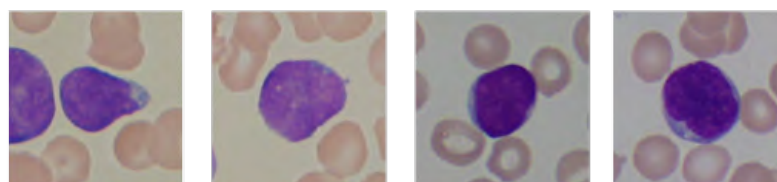


Figure 3. An example of segmented lymphocytes belonging to the leukemia class.

2.2. Image Preprocessing

The following data-augmentation techniques were used to increase the size of the training set:

- color jitter,
- Gaussian blur,
- horizontal flip,
- vertical flip,
- rotation.

Figure 4 shows the example of the effect of the augmentation techniques used.

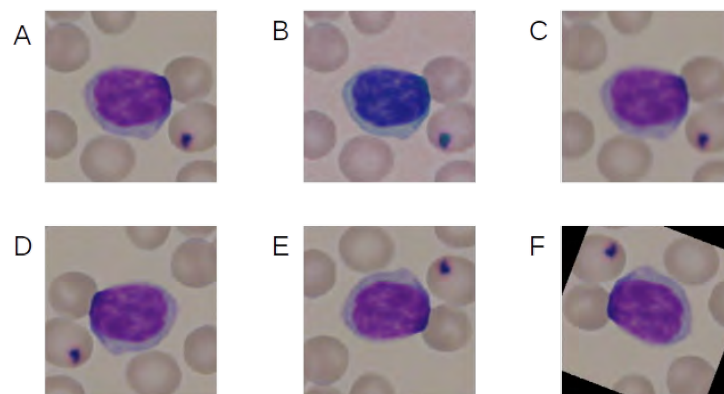


Figure 4. Example of the effect of the augmentation techniques used. (A) No augmentation, (B) color jitter, (C) Gaussian blur, (D) horizontal flip, (E) vertical flip, and (F) rotation.

Two variants of augmentation were used, and color jitter was used only in one of them. After augmentation, all images from the database had been normalized. It consisted of subtracting the means, which were 0.485, 0.456, and 0.406, and dividing by the standard deviations, which were 0.229, 0.224, and 0.225.

2.3. Hybrid Artificial Intelligence System

This research designed artificial intelligence models from two modules: the encoder and the head (Figure 5). The encoder transforms the input image into a fixed-size feature vector, abstractly describing the input data. The head takes the feature vector computed by the encoder and performs the classification; such a division of responsibilities allowed for the modular structure of the artificial intelligence system. Two neural networks were used as encoders and four different machine learning models as heads.

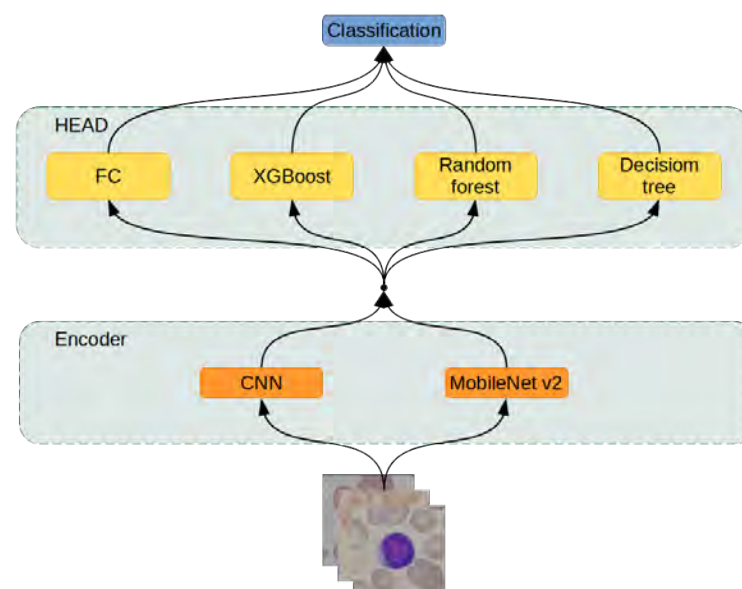


Figure 5. Hybrid artificial intelligence system architecture.

2.3.1. Encoders

The first is the primary deep convolutional neural network with the standard, homogeneous architecture described in Table 1. This network is referred to as the CNN-Encoder further in the article. It was designed without significant structural improvements like residual connections [24] or inception-based [25] layers to serve as a baseline for comparison during model evaluation.

Table 1. Architecture of the deep convolutional neural network.

| Layer | Channels In | Channels Out | Kernel Size | Padding | Stride |
|-----------|-------------|--------------|-------------|---------|--------|
| Conv2d | 3 | 8 | 3 × 3 | 1 × 1 | 1 × 1 |
| MaxPool2d | 8 | 8 | 3 × 3 | 0 × 0 | 3 × 3 |
| Conv2d | 8 | 16 | 3 × 3 | 1 × 1 | 1 × 1 |
| MaxPool2d | 16 | 16 | 3 × 3 | 0 × 0 | 3 × 3 |
| Conv2d | 16 | 32 | 3 × 3 | 1 × 1 | 1 × 1 |
| MaxPool2d | 32 | 32 | 3 × 3 | 0 × 0 | 3 × 3 |
| Conv2d | 32 | 64 | 3 × 3 | 1 × 1 | 1 × 1 |
| MaxPool2d | 64 | 64 | 3 × 3 | 0 × 0 | 3 × 3 |
| Conv2d | 64 | 128 | 3 × 3 | 1 × 1 | 1 × 1 |
| MaxPool2d | 128 | 128 | 3 × 3 | 0 × 0 | 3 × 3 |
| Conv2d | 128 | 2 | 1 × 1 | 0 × 0 | 1 × 1 |

The output value was flattened into a 128-dimensional vector. After each convolutional layer, the Leaky ReLU activation function was applied with a negative slope coefficient $\alpha = 0.01$. Each convolutional layer with kernel 3×3 had a padding value to offset stride and to preserve activation maps' dimensionality. The max pooling layers performed the shrinkage of the above-mentioned activation maps. The last layer performed convolution by applying kernel 1×1 to reduce the number of channels from 128 to 2, resulting in the final encoded feature vector size being reduced by the factor of 64.

The second network used as an encoder was MobileNet V2. This network was selected due to its trade-off between performance and efficiency on mobile CPUs. The MobileNet architecture is more complex than the first one due to the usage of numerous structural improvements like:

- Depthwise separable convolutions—which improves the convolutional layers. In a normal convolutional layer, the equation gives a convolutional layer kernel that has size (w, h, d) where w (width) and h (height) are arbitrarily chosen hyperparameters and d (depth) is equal to the depth of the input tensor. As a result, the amount of weights required to be optimized to train the i -th convolutional layer equation is given by the following equation:

$$|\Theta_i| = n_i \cdot w_i \cdot h_i \cdot d_i \quad (1)$$

where n_i is the number of filters in a layer. This relationship between the input tensor's depth and the number of weights in one filter becomes cumbersome during the stacking of deep convolutional layers. For example, the convolutional layer gets an input tensor of depth 1024 and must preserve the size in the third dimension. These restrictions imply that $d_i = 1024$ and the number of filters is also 1024 $\cdot d_i \cdot w_i \cdot h_i = 1,048,576 \cdot w_i \cdot h_i$. Since $w_i, h_i \in N^+ - 0, 1, 2$, it means that the minimum amount of weights required is equal to 9,437,184. This amount of weights is staggering, taking into consideration that it is merely one convolutional layer. Depthwise separable convolutions reduce this problem by splitting the convolutional layer into two parts: the first one applies one filter of kernel $w_i \cdot h_i$ without depth to every channel instead of having filters interpreting every channel, and the second layer uses a 1D convolution on the output of the first layer, performing a depth-sensitive linear transformation. As a result, the same task is completed, but the equation gives the cost of the weights:

$$|\Theta'_i| = n_i \cdot w_i \cdot h_i + n_i \cdot d_i \cdot 1 \cdot 1 = n_i \cdot (w_i \cdot h_i + d_i) \quad (2)$$

Since $n_i = d_i$ and usually $w_i, h_i \ll d_i$, the maximum reduction obtained from using this method is equal to

$$\lim_{n_i \rightarrow \infty} \frac{n_i^2 + n_i w_i h_i}{n_i^2 w_i h_i} = \frac{1}{w_i h_i} (1 + \lim_{n_i \rightarrow \infty} \frac{w_i h_i}{n_i}) = \frac{1}{w_i h_i} (1 + 0) = \frac{1}{w_i h_i} \quad (3)$$

As a result, this method is able to reduce the number of used weights by the factor of $w_i \cdot h_i$.

- Linear bottlenecks—a newly introduced layer performs a linear transformation of the convolutional layers' activation map, resulting in tensor's depth reduction with minimal information loss and an increasing amount of information stored per channel.
- Inverted residuals—the residual connection between layers bottlenecks instead of connecting normal convolutional layers. Since bottleneck layers are by design depth-reduced transformations of convolutional layers, application of a residual connection by bundling bottleneck layers results in a further computation reduction.

In this research, MobileNet v2 was used in two different versions: pretrained and not pretrained. The pretrained network was optimized to solve tasks from the ImageNet [26] contest consisting of image classification into a thousand different classes. The not-pretrained network started training using weights initialized by the usage of the Kamming He initialization algorithm.

2.3.2. Head

The head module takes as an input a feature vector computed by the encoder and performs its classification. There were four modules chosen for this task:

- Fully connected neural network layer,
- XGBoost,
- Random Forest,
- Decision Tree [27].

The first module is part of the neural networks, and it is able to propagate error gradients further down the network. Because of that, this particular head can be trained together with the encoder. However, the rest of the modules require a fully trained encoder to encode images into small feature vectors.

The hybrid approach allows the network already pre-trained on different tasks (like ImageNet) and uses it as a finished encoder to create a new dataset to translate the original one into the feature space. Then, machine learning algorithms can be trained on the newly created dataset of feature vectors, utilizing their different approach to create a heterogeneous classification system.

2.4. Training

In this research, two different training techniques for neural-network-only systems and hybrid machine learning models were used. In both of these methods, data augmentation was used in one of three modes:

- no augmentation was applied,
- augmentation was applied with all of the available techniques described in the "Image preprocessing" section,
- augmentation was applied with all the techniques except the "Color jitter" method.

The training was performed using the following hardware configurations: dual-Intel Xeon Silver 4210R, 192 GB RAM, and Nvidia Tesla A100 GPU. In this research, PyTorch, Sklearn, Numpy, Pandas, and Jupyter Lab programming solutions were used to implement the neural networks [28].

2.4.1. Neural Network Training

This procedure was employed when a fully connected layer was used as a head of the system. Because all elements in the system can propagate gradient error, both the encoder and the head were trained simultaneously. The models were trained using this procedure:

- CNN-Encoder + fully connected layer,
- not-pretrained MobileNet v2 + fully connected layer,
- pretrained MobileNet v2 + fully connected layer.

Neural networks were trained using the Adam optimizer [29]. Every network was optimized on a training dataset and evaluated on a validation dataset. They were trained for 1000 epochs unless early stopping [30] was performed. If the best result on the validation dataset was not improved in 100 epochs, training was stopped, and another network was created. The learning rate at the beginning was equal to 0.001, and it was reduced by half if the network did not improve its best result on the training dataset within 10 epochs from the last improvement or learning rate reduction. If the learning rate reached 0.000001, then no further reduction was applied. Depending on the augmentation settings selected, each image might be subjected to random augmentation before being put on the input of the neural network.

2.4.2. Hybrid System Training

The hybrid system consisted of a MobileNet v2 encoder pre-trained on the ImageNet dataset and a machine learning algorithm performing the role of the head. These algorithms were:

- XGBoost,
- Random Forest,
- Decision Tree.

To train these head models, a new dataset of encoded images was created. Thus, there were three datasets created. The first one was not subjected to augmentation. In this case, every image in the original dataset was converted into a 1000-dimensional feature vector.

In both the second and the third cases, augmentation was used, resulting in every image being 100 times randomly augmented and its vector added to the dataset. As a result, datasets with an applied boost had a size 100 times greater than the non-augmented one. The difference between the second and third cases was in whether color jitter was used or not.

After datasets creation, each machine learning model was optimized on this set according to its unique training algorithm.

2.5. Metrics

Neural networks were evaluated using the metrics described below [28]. For the purpose of simplicity of equations, certain acronyms were created, as follows: TP—true positive, TN—true negative, FP—false positive, and FN—false negative. The metrics used for network evaluation were:

- Accuracy: $Acc = (TP + TN) / (TP + FP + TN + FN)$,
- Precision = $TP / (TP + FP)$,
- Recall = $TP / (TP + FN)$,
- $F1 = 2 * Precision * Recall / (Precision + Recall)$,
- AUC—area under the receiver operating characteristic (ROC). The ROC is a curve determined by calculating the true-positive rate = $TFP = TP / (TP + FN)$ and the false-positive rate = $FPR = FP / (TN + FP)$. The false-positive rate describes the x-axis and the true-positive rate the y-axis of a coordinate system. By changing the threshold value responsible for classification of an example as belonging to either the positive or negative class, pairs of TFP-FPR were generated, resulting in the creation of the ROC curve. The AUC is a measurement of the area below the ROC curve.

3. Results

Training, validation, and test sets were generated 15 times to evaluate networks to minimize the influence of random dataset division. Each network was trained on a training dataset. During the training, the network was assessed on the validation dataset to select the best, least-overfitted weights set of the network, and to perform early stopping. When such a set of weights was established, the final network's evaluation was performed on the test dataset. Results of the networks were grouped by both architecture selection, whether pre-training was employed or not, application of augmentation, and presence of color modification during the augmentation process. The results are presented in Table 2.

Table 2. Experiment results

| Name | Acc | Acc Avg Std | F1 | F1 Avg Std | AUC | AUC Avg Std |
|--|-------------|---------------|------------|--------------|------------|---------------|
| FC, Mobilenet v2, augmented with no color | 100.0–82.0% | 94.8% 5.3 | 100.0–81.8 | 94.8 5.3 | 100.0–95.5 | 99.2 1.3 |
| FC, Mobilenet v2, augmented | 100.0–87.1% | 93.8% 3.8 | 100.0–86.8 | 93.7 3.9 | 100.0–93.7 | 99.0 1.6 |
| FC, Mobilenet v2, no augmentation | 100.0–76.9% | 92.8% 6.1 | 100.0–76.8 | 92.7 6.1 | 100.0–94.1 | 98.7 1.6 |
| Random Forest, Mobilenet v2 augmented | 97.4–84.6% | 92.1% 4.0 | 97.4–84.5 | 92.0 4.1 | 100.0–95.6 | 98.7 1.3 |
| XGBoost, Mobilenet v2, augmented with no color | 97.4–82.0% | 91.1% 5.1 | 97.4–81.2 | 90.9 5.2 | 100.0–93.2 | 97.8 2.2 |
| XGBoost, Mobilenet v2, augmented | 97.4–76.9% | 91.1% 5.5 | 97.4–76.8 | 91.0 5.5 | 100.0–89.7 | 98.0 2.8 |
| Random Forest, Mobilenet v2, augmented with no color | 94.8–82.0% | 89.9% 4.3 | 94.8–81.6 | 89.8 4.4 | 99.7–92.8 | 97.9 1.9 |
| Decision Tree, Mobilenet v2, augmented | 89.7–64.1% | 80.0% 7.7 | 89.5–63.8 | 79.7 7.8 | 89.5–63.9 | 80.1 8.0 |
| Decision Tree, Mobilenet v2, augmented with no color | 89.7–66.6% | 79.3% 6.6 | 89.6–65.8 | 79.0 6.8 | 90.3–67.7 | 79.8 6.6 |
| Random Forest, Mobilenet v2, no augmentation | 87.1–64.1% | 76.9% 6.9 | 87.1–63.8 | 76.7 7.0 | 94.8–75.6 | 85.2 5.4 |
| XGBoost, Mobilenet v2, no augmentation | 89.7–56.4% | 75.3% 11.6 | 89.7–55.9 | 75.1 11.7 | 95.7–65.7 | 83.0 9.4 |
| Decision Tree, Mobilenet v2, no augmentation | 84.6–46.1% | 62.3% 10.7 | 84.5–44.8 | 62.0 10.8 | 85.0–45.4 | 62.8 10.8 |

4. Discussion

The three best results were obtained from the pre-trained MobileNet v2 repurposed to blast cell detection through learning with a fully connected layer head attached. However, MobileNet v2, not pre-trained with a fully connected layer head, scored much lower despite having the same architecture. It suggests that transfer learning can be used as a regularization technique, preventing overfitting and improving overall performance. In addition, ImageNet contained images depicting objects and animals related to everyday daily life instead of pictures of microbiological phenomena, yet such pre-training proved beneficial. It further suggests that a pre-training network on large datasets from seemingly unrelated domains may improve the results on small, specialized tasks like blast-cell classification in this particular research.

Hybrid systems with MobileNet v2 as an encoder scored the best after the repurposed, pre-trained MobileNet v2 network. Their performance was better than the CNN-Encoder network and the not-pre-trained MobileNet. It suggests that in case of limited access to high-end processing units like GPUs, the strategy described below may have satisfactory results:

- take the available neural network pre-trained on a massive dataset,
- use this network to encode data in the small, domain-specific dataset,
- train a machine learning model on reduced, encoded data.

This strategy may be performed on the CPU. The computational bottleneck in this operation is using a deep neural network on the CPU to encode the dataset. However, this operation must be conducted only once. Its result is sufficient for machine learning model training and is much lighter than the original dataset, making it easier to store. In this research, both machine learning models like XGBoost and the MobileNet v2 network as an encoder were evaluated due to their proficiency in training on the CPU. Such an approach may prove beneficial for systems with reduced computational capabilities like mobile devices or the IoT.

The XGBoost and Random Forest algorithms proved to be capable of extracting abstract information from encoded feature vectors. However, the Decision Tree algorithm scored substantially worse and did not achieve the desired results. Moreover, this algo-

algorithm's simplistic structure was not complex enough to extract the high-level information required for performing classification on a sparse dataset like the one examined in this work.

The pre-trained MobileNet v2 scored better results than CNN-Encoder, which in turn scored better than the not-pre-trained MobileNet v2. These findings suggest that MobileNet v2 was not pre-trained overfitted to the training set due to its more profound and more complicated structure compared with the baseline CNN-Encoder. However, pre-training allowed MobileNet v2 to score better than CNN-Encoder. Thus, it suggests that designing networks seemingly more profound than required and pre-training them may provide better results than applying smaller architectures despite the concern of overfitting an overparameterized model.

The augmentation mode was split into augmentation with the application of color jitter and without it. Because of an a priori assumption, color was an essential factor in cell classification as it is an indicator of biological features. This assumption proved to be correct because top architectures differentiating between themselves only by applying color jitter scored better without this augmentation technique. This proves that augmentation techniques must be challenged to determine whether they are truly beneficial or not for this particular dataset's purposes.

The pre-trained MobileNet v2 network proved its effectiveness in the researched problem despite the training process being conducted on data from different domains. It suggests that the domains of knowledge are not as separated as it seems. However, it is doubtful that understanding the MobileNet v2 network gathered during training on the ImageNet dataset is helpful in this example. It is presumed that only a specific part of this network is useful in this topic. The procedure for such knowledge extraction would be beneficial as it reduces the computation and size of stored weights sets. The authors plan on further investigation of this topic.

5. Conclusions

The proposed strategy of designing hybrid artificial intelligence systems for low-resource, low-computational-complexity processing units' tasks by introducing a pre-trained neural network for data encoding proved beneficial in this particular task. The examined systems using MobileNet v2 as an encoder and XGBoost and Random Forest as classification heads were able to score, on average, an above 90% accuracy, going as high as 97.4%. The system developed in this work can be trained and run on a low-power CPU like a mobile CPU or one dedicated to the IoT. However, the Decision Tree algorithm turned out to be not complex enough to perform meaningful classification. The best results were obtained by repurposing the already trained deep neural network instead of training the same one from scratch or creating the smaller one to reduce overfitting. The regularization benefit of transfer learning was significant during the examination of this dataset.

Author Contributions: Conceptualization, K.P. and S.Š.; methodology, K.P., S.Š. and M.G.; software, K.P. and S.Š.; validation, K.P., S.Š., M.G. and D.L.; formal analysis, K.P.; investigation, K.P., S.Š., M.G. and D.L.; resources, K.P., S.Š., M.G., D.L., S.B. and Z.L.; data curation, S.B. and Z.L.; writing—original draft preparation, K.P., S.Š. and M.G.; writing—review and editing, K.P., S.Š. and M.G.; visualization, K.P., S.Š., D.L. and M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

| | |
|-----------------|--|
| Conv1d | Layer in deep neural networks that performs a convolution on a one-dimensional signal. |
| MaxPool1d | Layer in deep neural networks that performs a pooling operation by selecting the maximum value from the moving window. |
| Fully connected | Layer in deep neural networks that consists of neurons, each of which process the whole input data. |
| Leaky ReLU | Activation function used in deep neural networks. |
| Padding | Parameter used in convolutional layers specifying the amount of zeroed samples added to the start and the end of the processed signal. For example: a padding of 1 means that there is one sample of value zero artificially added on the beginning and at the end of the signal. This operation is conducted in order to mitigate activation-map shrinkage due to the application of convolution. |
| Stride | Parameter used in convolutional layers specifying the shift distance between subsequent windows of convolutions. For example: a stride of 1 means that the next convolution starts right after the the beginning of the previous one, so the windows will overlap (provided that the kernel size is greater than 1). |
| ALL | Acute lymphoblastic (or lymphocytic) leukemia. |
| AML | Acute myeloid (or myelogenous) leukemia. |
| CML | Chronic myeloid (or myelogenous) leukemia. |
| CLL | Chronic lymphocytic leukemia |
| WBC | White blood cells. |

References

- Onciu, M. Acute Lymphoblastic Leukemia. *Hematol./Oncol. Clin. N. Am.* **2009**, *23*, 655–674. [[CrossRef](#)] [[PubMed](#)]
- Aldoss, I.; Forman, S.J.; Pullarkat, V. Acute Lymphoblastic Leukemia in the Older Adult. *J. Oncol. Pract.* **2019**, *15*, 67–75. [[CrossRef](#)] [[PubMed](#)]
- Mohapatra, S.; Patra, D.; Satpathi, S. Image analysis of blood microscopic images for acute leukemia detection. In Proceedings of the 2010 International Conference on Industrial Electronics, Control and Robotics, Rourkela, India, 27–29 December 2010; pp. 215–219.
- Bennett, J.M.; Catovsky, D.; Daniel, M.-T.; Flandrin, G.; Galton, D.A.G.; Gralnick, H.R.; Sultan, C. Proposals for the Classification of the Acute Leukaemias French-American-British (FAB) Co-operative Group. *Br. J. Haematol.* **1976**, *33*, 451–458. [[CrossRef](#)] [[PubMed](#)]
- Amidi, Y.; Nazari, B.; Sadri, S.; Yousefi, A. Parameter Estimation in Multiple Dynamic Synaptic Coupling Model Using Bayesian Point Process State-Space Modeling Framework. *Neural Comput.* **2021**, *33*, 1269–1299. [[CrossRef](#)] [[PubMed](#)]
- Yousefi, A.; Amidi, Y.; Nazari, B.; Eden, U. Assessing Goodness-of-Fit in Marked Point Process Models of Neural Population Coding via Time and Rate Rescaling. *Neural Comput.* **2020**, *32*, 2145–2186. [[CrossRef](#)] [[PubMed](#)]
- Roshani, M.; Phan, G.; Faraj, R.; Phan, N.H.; Roshani, G.; Nazemi, B.; Corniani, E.; Nazemi, E. Proposing a gamma radiation based intelligent system for simultaneous analyzing and detecting type and amount of petroleum by-products. *Nucl. Eng. Technol.* **2021**, *53*, 1277–1283 [[CrossRef](#)]
- Nazemi, B.; Rafiean, M. Forecasting house prices in Iran using GMDH. *Int. J. Hous. Mark. Anal.* **2021**, *14*, 555–568. [[CrossRef](#)]
- Le, D.K.T.; Bui, A.A.; Yu, Z.; Bui, F.M. An automated framework for counting lymphocytes from microscopic images. In Proceedings of the 2015 International Conference and Workshop on Computing and Communication (IEMCON), Vancouver, BC, Canada, 15–17 October 2015; pp. 1–6.
- Vogado, L.H.S.; Veras, R.d.M.S.; Andrade, A.R.; Silva, R.R.V.e.; Araujo, F.H.D.; Medeiros, F.N.S. Unsupervised Leukemia Cells Segmentation Based on Multi-space Color Channels. In Proceedings of the 2016 IEEE International Symposium on Multimedia (ISM), San Jose, CA, USA, 11–13 December 2016. [[CrossRef](#)]
- Rehman, A.; Abbas, N.; Saba, T.; Rahman, S.I.; Mehmood, Z.; Kolivand, H. Classification of acute lymphoblastic leukemia using deep learning. *Microsc. Res. Tech.* **2018**. [[CrossRef](#)] [[PubMed](#)]
- Di Ruberto, C.; Loddo, A.; Puglisi, G. Blob Detection and Deep Learning for Leukemic Blood Image Analysis. *Appl. Sci.* **2020**, *10*. doi.org/10.3390/app10031176 [[CrossRef](#)]
- Prellberg, J.; Kramer, O. Acute lymphoblastic leukemia classification from microscopic images using convolutional neural networks. In *ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 53–61.

14. Ahmed, N.; Yigit, A.; Isik, Z.; Alpkocak, A. Identification of Leukemia Subtypes from Microscopic Images Using Convolutional Neural Network. *Diagnostics* **2019**, *9*, 104. doi.org/10.3390/diagnostics9030104 [[CrossRef](#)] [[PubMed](#)]
15. Guo, Z.; Wang, Y.; Liu, L.; Sun, S.; Feng, B.; Zhao, X. Siamese Network-Based Few-Shot Learning for Classification of Human Peripheral Blood Leukocyte. In Proceedings of the 2021 IEEE 4th International Conference on Electronic Information and Communication Technology (ICEICT), Harbin, China, 20–22 January 2021; pp. 818–822.
16. Zhao, J.; Zhang, M.; Zhou, Z. Automatic detection and classification of leukocytes using convolutional neural networks. *Med. Biol. Eng. Comput.* **2017**, *55*, 1287–1301. doi.org/10.1007/s11517-016-1590-x [[CrossRef](#)] [[PubMed](#)]
17. Labati, R.D.; Piuri, V.; Scotti, F. All-IDB: The acute lymphoblastic leukemia image database for image processing. In Proceedings of the 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011; pp. 2045–2048.
18. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
19. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
20. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
21. Scotti, F. Robust segmentation and measurements techniques of white cells in blood microscope images. In Proceedings of the 2006 IEEE Instrumentation and Measurement Technology Conference Proceedings, Sorrento, Italy, 24–27 April 2006; pp. 43–48.
22. Scotti, F. Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. In Proceedings of the 2005 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, Messian, Italy, 20–22 July 2005; pp. 96–101.
23. Piuri, V.; Scotti, F. Morphological classification of blood leucocytes by microscope images. In Proceedings of the 2004 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, Boston, MA, USA, 14–16 July 2004; pp. 103–108.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
26. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
27. Swain, P.H.; Hauska, H. The decision tree classifier: Design and potential. *IEEE Trans. Geosci. Electron.* **1977**, *15*, 142–147. [[CrossRef](#)]
28. Śmigiel, S.; Pałczyński, K.; Ledziński, D. ECG Signal Classification Using Deep Learning Techniques Based on the PTB-XL Dataset. *Entropy* **2021**, *23*, 1121. [[CrossRef](#)] [[PubMed](#)]
29. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
30. Caruana, R.; Lawrence, S.; Giles, L. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. *Adv. Neural Inf. Process. Syst.* **2001**, 402–408.

Article

Entropy Measurements for Leukocytes' Surrounding Informativeness Evaluation for Acute Lymphoblastic Leukemia Classification

Krzysztof Pałczyński ^{*,†} , Damian Ledziński [†]  and Tomasz Andrysiak 

Faculty of Telecommunications, Computer Science and Electrical Engineering, Bydgoszcz University of Science and Technology, 85-796 Bydgoszcz, Poland

* Correspondence: krzysztof@palczynski.com.pl; Tel.: +48-517-721-327

† These authors contributed equally to this work.

Abstract: The study of leukemia classification using deep learning techniques has been conducted by multiple research teams worldwide. Although deep convolutional neural networks achieved high quality of sick vs. healthy patient discrimination, their inherent lack of human interpretability of the decision-making process hinders the adoption of deep learning techniques in medicine. Research involving deep learning proved that distinguishing between healthy and sick patients using microscopic images of lymphocytes is possible. However, it could not provide information on the intermediate steps in the diagnosis process. As a result, despite numerous examinations, it is still unclear whether the lymphocyte is the only object in the microscopic picture containing leukemia-related information or if the leukocyte's surroundings also contain the desired information. In this work, entropy measures and machine learning models were applied to study the informativeness of both whole images and lymphocytes' surroundings alone for Leukemia classification. This work aims to provide human-interpretable features marking the probability of sickness occurrence. The research stated that the hue distribution of images with lymphocytes obfuscated alone is informative enough to facilitate 93.0% accuracy in healthy vs. sick classification. The research was conducted on the ALL-IDB2 dataset.

Keywords: acute lymphoblastic leukemia classification; image background informativeness; Shannon entropy; cross-entropy; XGBoost



Citation: Pałczyński, K.; Ledziński, D.; Andrysiak, T. Entropy Measurements for Leukocytes' Surrounding Informativeness Evaluation for Acute Lymphoblastic Leukemia Classification. *Entropy* **2022**, *24*, 1560. <https://doi.org/10.3390/e24111560>

Academic Editors: Yoh Iwasa, Su Ruan and Jérôme Lapuyade-Lahorgue

Received: 14 September 2022

Accepted: 26 October 2022

Published: 29 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Acute lymphoblastic leukemia (ALL) diagnosis is closely associated with morphological changes in white blood cells (WBC, or leukocytes). ALL, known in the group of blood diseases, is characterized by the overproduction and continuous proliferation of malignant and immature white blood cells (referred to as lymphoblasts or blasts). Although the number of leukocytes can often be considered an essential indicator of pathological changes in the morphological picture of the blood, it is not always sufficient.

The detection of ALL and its subtypes is often accomplished by examining blood or bone marrow smears. According to the French-American-British (FAB) classification standard, ALL is classified into the L1, L2, and L3 subtypes. Assignment to the correct subtype is carried out according to observation of the nucleus's morphology, including the affected cell's pattern and variation in its shape. This procedure is generally accepted and known from numerous works of authors researching the detection and classification of leukocytes.

In clinical practice, microscopic examination of blood smears to verify ALL is based primarily on counting different types of white blood cells. Equally important is the analysis of the nuclear features of leukocytes, which are often distinguished by their pastel blue coloration. Nevertheless, their further evaluation becomes complicated, as according to the

FAB, features such as the size, shape variation, and texture should be considered. After all, leukocytes can be distinguished by their size, color characteristics, the ratio of the nuclei to the cytoplasm contained in them, etc.

Due to the morphological diversity of white blood cells, the classification into ALL subtypes may not always be realized correctly. The reason for this can perhaps be found in the complexity of the backgrounds of microscopic images. One of the essential aspects is played by the surroundings of white blood cells, which occur in the setting of other morphotic components of the blood such as red blood cells or platelets. However, according to FAB classification, the background does not add significant information from the perspective of classification, although it can significantly hinder the correct classification of ALL. Our work measures background informativeness to determine the influence of the lymphocytes' surroundings on correct Leukemia classification.

In their article [1], Andrade et al. performed an extensive evaluation of the leukocyte segmentation techniques by artificial intelligence systems developed by well-established research teams. The authors performed experiments on five databases: ALL-IDB2 [2], Blood-Seg [3], Leukocytes [4], JTSC Database, and CellaVision [5]. The leukocyte segmentation methods examined by the authors use Otsu threshold [6–9] (Madhloom et al., Arslan et al., Nazlibilek et al., and Prinyakupt et al.), K-means [10–17] (Nasir et al., Mohapatra et al., Madhukar et al., Amin et al., Sarrafzadeh et al., Vincent et al., Vogado et al., and Kumar et al.), region growing [7,10,18] (Nasir et al., Mohammed et al., and Arslan et al.), edge detector [18] (Mohammed et al.), Zack's algorithm [19] (Abdeldaim et al.), and arithmetical image processing operations [3,6,9,14] (Madhloom et al., Mohamed et al., Prinyakupt et al., and Sarrafzadeh et al.). The methods were examined on the images encoded using RGB [9] (Prinyakupt et al.), grayscale [6,18] (Madhloom et al. and Mohammed et al.), HSI [10] (Nasir et al.), $L^*a^*b^*$ [11,12,14,15] (Mohapatra et al., Madhukar et al., Sarrafzadeh et al., and Vincent et al.), HSV [13,17] (Amin et al. and Kumar et al.), CMYK [19] (Abdeldaim et al.), and CMYK + $L^*a^*b^*$ color schemes [16] (Vogado et al.). The authors achieved satisfactory results for all of the datasets, in some cases reaching 97% accuracy. However, none of the methods examined proved to be the best on all datasets. It is also important to note that the method with the best results in this experiment achieved only a 58.44% leukocyte nuclei detection rate. In this article, a leukocyte was considered detected by computing the true positive rate metric TPR_t with a threshold $t = 0.9$.

In [1], Andrade et al. proved that leukocyte nuclei segmentation is a non-trivial task, and none of the well-established methods proved efficient in leukocyte detection. The authors stated in the article that leukocyte nuclei segmentation is performed to classify the presence of leukemia in the cells. The survey of image processing techniques and their results motivated us to perform research focusing on attempting to examine the amount of information contained in non-leukocytes for leukemia classification without performing image segmentation. Furthermore, the plethora of black box-type artificial intelligence systems applied with various evaluation results signified the importance of establishing the features' quality.

The subject of leukemia classification was researched thoroughly using deep learning methods. The authors of [20–24] applied various types of convolutional neural networks (CNNs) [25]. In [20], Rehman et al. applied the AlexNet [26] architecture of CNN networks for leukemia classification, achieving 97.78% accuracy. Similar research was conducted by the authors of [21] (Prellberg et al.) using the ResNeXt50 [27] architecture of CNN networks pretrained on the ImageNet dataset [28]. The researchers achieved 89.7% accuracy. The non-binary classification of leukemia was conducted by Ahmed et al. [22]. The experiments were run to establish the ability of CNNs to discriminate in one vs. many mode against leukemia types such as acute myeloid leukemia (AML), chronic lymphocytic leukemia (CLL), chronic myeloid leukemia (CLM), and acute lymphoblastic leukemia (ALL). They achieved classification with 81% accuracy. The authors of [23] (Guo et al.) used Siamese networks [29] to achieve few-shot learning [30] with 89.96% accuracy. Similar research was conducted by Abhishek et al. [24]. In this work, the authors used transfer learning to

compare the results of deep convolutional neural networks with support vector machines (SVMs) against SVM interpreting features extracted by local binary patterns (LBPs) [31] and the histogram of oriented gradients (HOG) [32]. The deep learning approach obtained 98% accuracy, SVM + LBP resulted in 83% accuracy, and SVM + HOG resulted in 50% accuracy. In their work, Rodrigues et al. [33] also applied deep learning for leukemia classification. What differs them from other works is the optimization of trained neural networks using a genetic algorithm. This procedure improved the results to 98.46%.

The results of leukemia classification using deep learning in [20–23] (Rehman et al., Prellberg et al., Ahmed et al., Guo et al., Abhishek et al., and Rodrigues et al.) provided substantial accuracy for healthy vs. sick discrimination. However, due to the black box nature of deep learning methods, it is unclear what pattern was extracted by neural networks to achieve such efficiency. The knowledge distillation [34] techniques are currently not advanced enough to determine the neural network’s reasoning, leading to classification in a manner humans can understand. Because of this issue, the adoption of deep learning and machine learning in medicine is slow due to the inability to verify the quality of the extracted features. Our work examines the ALL-IDB dataset samples to determine the features understandable by humans, allowing reliable classification and, at the same time, determining the usefulness of the leukocyte’s surroundings in leukemia classification.

1.1. Summary of Surveyed Research Works

The surveyed works are summarized in Table 1 to present various method and image encoding technique combinations applied in a readable format.

Table 1. A summary of investigated research works.

| Ref. | Image Encoding | Methods |
|------|----------------|---|
| [3] | Grayscale | Arithmetical operations |
| [6] | Grayscale | Otsu threshold, arithmetical operations |
| [7] | RGB | Otsu threshold, region growing |
| [8] | Grayscale | Otsu threshold |
| [9] | RGB | Otsu threshold, arithmetical operations |
| [10] | HSI | K-means, region growing |
| [11] | L*a*b* | K-means |
| [12] | L*a*b* | K-means |
| [13] | HSV | K-means |
| [14] | L*a*b* | K-means, arithmetical operations |
| [15] | L*a*b* | K-means |
| [16] | CMYK + L*a*b* | K-means |
| [17] | HSV | K-means |
| [18] | Grayscale | Region growing, edge detectors |
| [19] | CMYK | Zack’s algorithm |
| [20] | RGB | AlexNet |
| [21] | RGB | ResNeXt50 |
| [22] | RGB | CNN |
| [23] | RGB | Siamese networks |
| [24] | RGB | CNN |
| [33] | RGB | ResNet50 v2 |

1.2. The Aim of This Work

This work is a continuation of the research described in [35] (Pałczyński et al.). This article aims to establish the informativeness of a lymphocyte’s surroundings for leukemia classification. The classification is conducted on the features extracted from images with lymphocytes obfuscated using black rectangles. The classification results without information regarding lymphocytes are compared against the quality of discrimination in the unmodified images. The features extracted from the image are deterministic, human-interpretable qualities. The discrimination is performed using both simple, divergence-based clusterization (mean squared error and cross-entropy [36]) and by applying machine

learning algorithms such as logistic regression [37] and the XGBoost algorithm [38]. The images used in this research are encoded using RGB and HSV methods.

This work aims to quantify the amount of information stored in human-interpretable features computed from an image with the lymphocytes obfuscated. This work aims not to achieve the best classification results but to determine how well the classification can be performed using a limited amount of information while remaining interpretable by humans. In our previous article, deep neural networks were applied to the raw images to perform the classification. Although the results were satisfactory, the inherent black box nature of deep neural networks prevented us from acquiring human-interpretable knowledge on the nature of this particular classification problem. This work aims to provide such information.

1.3. Summary of Our Contributions

Our main contributions can be summarized as follows:

1. We examined the influence of lymphocyte obfuscation on acute lymphoblastic leukemia classification to evaluate its surroundings' informativeness. The hue distribution of lymphocytes' surroundings processed by the XGBoost algorithm resulted in classification with 93% accuracy.
2. We evaluated the informativeness of channels' value distributions of both the RGB and HSV color encodings. We determined that the channel encoding color green contained the most information, with an XGBoost classification accuracy of 96%. The same evaluation of red and blue color channels resulted in classification accuracies of 87% and 83%, respectively. The hue, saturation, and value channels obtained classification results of 94%, 94%, and 84%, respectively.
3. The classification results of the XGBoost algorithm interpreting the distributions of individual channel values resulted in a classification quality similar to the effects of deep learning application on raw images performed by other researchers. As a result, we reduced the amount of input information by three orders of magnitude while achieving comparable results.
4. We evaluated the informativeness of the entropy measurements of each channel's values distribution using the Shannon entropy. The Shannon entropy computed for the hue distribution of images with lymphocytes obfuscated resulted in a classification accuracy of 81% and 68% accuracy when using images without the lymphocytes being obfuscated. The results suggest that lymphocytes' surroundings contain essential information for acute lymphoblastic leukemia classification.

1.4. Paper Organization

This work is divided into sections. Section 2 describes the materials and methods used in this research. This section describes the image preprocessing techniques, encoding, feature vectorization, experiments conducted, and metrics. Section 3 presents the results of the experiments described in Section 2. Section 4 provides interpretation of the results presented in Section 3, and Section 5 concludes the paper.

2. Materials and Methods

This section describes the materials and methods used in this research. It describes the data preprocessing, vectorization, and algorithms for generating the experimental results. The experimental procedure involved computing the Shannon entropy, measuring the cross-entropy score between the obtained value distributions, and fitting the machine learning models.

2.1. ALL-IDB Database

The research was conducted on the ALL-IDB database, containing microscopic images of lymphocyte cells documented from healthy people and patients with acute lymphoblastic leukemia. This database is made publicly available by Università degli Studi di Milano,

and it contains annotation of which samples represent cases of leukemia and which were obtained from healthy patients. Oncologists performed the annotation.

The database contains 260 microscopic images. The dataset is balanced between classes, having 130 images of blood smears taken from healthy patients and the same amount from sick ones. The home website of this dataset is accessible at this link: <http://homes.di.unimi.it/scotti/all/> (accessed on 15 October 2021).

2.2. Image Preprocessing

In this section, the techniques of image processing applied during the experiments are presented. The amount of information contained within the background of the image can be examined by removing the lymphocytes from the graphics. Removing lymphocyte information was performed by covering them with black rectangle-shaped bounding boxes. This type of obfuscation was chosen to remove all information from the lymphocytes and information regarding the cells' shapes, which may have interfered with the experiment's results. The resulting shrinkage of the background from such an obfuscation technique was not considered a concern. The results of this operation are presented on the Figure 1. The experiments were conducted using images both unmodified and obfuscated to compare the information stored in background of the image with all of the information contained.

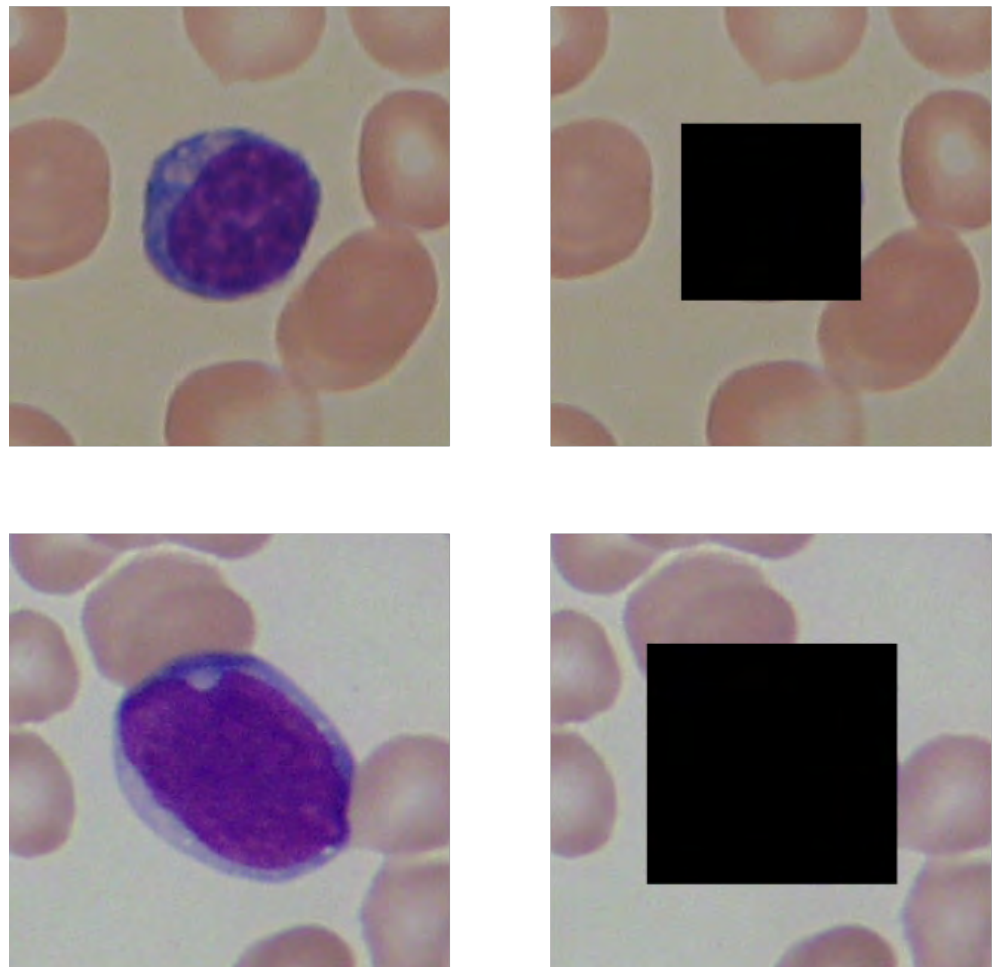


Figure 1. A black rectangle covers an example of lymphocytes in the images in the second column. The first row presents versions of the image of a healthy patient, and the second row depicts versions of the photo taken for a patient sick with leukemia.

The images were also subjected to data augmentation to determine the influence of commonly used image preprocessing techniques on the informativeness of the background.

The data augmentation was performed before lymphocyte obfuscation. The modification methods examined in this research were the following:

- Gaussian blur;
- Median blur;
- Gaussian noise.

Gaussian blur (also known as Gaussian smoothing) is a commonly used data augmentation technique in deep learning for increasing the number of images in the training set. It serves as a low-pass filter, reducing higher frequencies from the image and thus achieving the perceived effect of smoothness. The filter works by convolving the image with a matrix representation of a two-dimensional Gaussian function. Equation (1) presents the method for obtaining the filter matrix:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1)$$

where x and y represent the divergence from the center of the matrix. The bigger the matrix (kernel), the more precise the convolution is. The kernel lengths used in this research were 3, 9, 21, and 51. The results of applying this operation are presented in Figure 2.



Figure 2. Presentation of the Gaussian blurring application on the images. The blurring procedure did not affect the first image in the row, and the rest were convolved with kernel of sizes 3, 9, 21, and 51 pixels, respectively.

Median blur has a similar purpose to Gaussian blur. This filter windows the image and returns the median value from each window. This operation reduces noise and creates a low-pass filter. The sizes used for the windows in this research were 3, 9, 21, and 51. The results of applying this operation are presented in Figure 3.



Figure 3. Presentation of the median blurring application on the images. The blurring procedure did not affect the first image in the row, and the rest were convolved with kernel of sizes 3, 9, 21, and 51 pixels, respectively. The images were 257×257 pixels in size.

The last data augmentation technique used was adding Gaussian noise. Compared with the previous two techniques, this method increases the amount of noise instead of decreasing it. It is also commonly used in deep learning to reduce deep neural networks' overreliance on temporal patterns in favor of a more holistic approach. Each pixel in the output file was computed using Equation (2):

$$x' = \lfloor x \cdot (1 + n) \rfloor, n \sim \mathcal{N}(0, \sigma) \quad (2)$$

Here, x is the current value of the filtered image, and σ is the standard deviation of the distribution. The values of variance used in this research were 0.001, 0.01, and 0.1. The results of applying this technique are presented in Figure 4. The processing data pipeline involving data augmentation and lymphocyte obfuscation is presented in Figure 5.

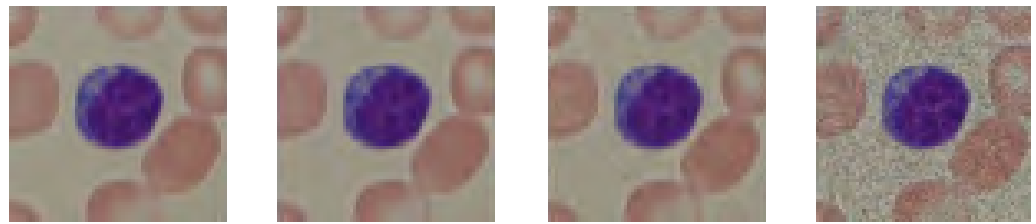


Figure 4. Presentation of the Gaussian noise application on the images. The noise procedure did not affect the first image in the row, and the rest were subjected to the multiplicative noise sampled from the Gaussian distribution, parametrized by the mean equal to zero and variance containing values of 0.001, 0.01, and 0.1, respectively.

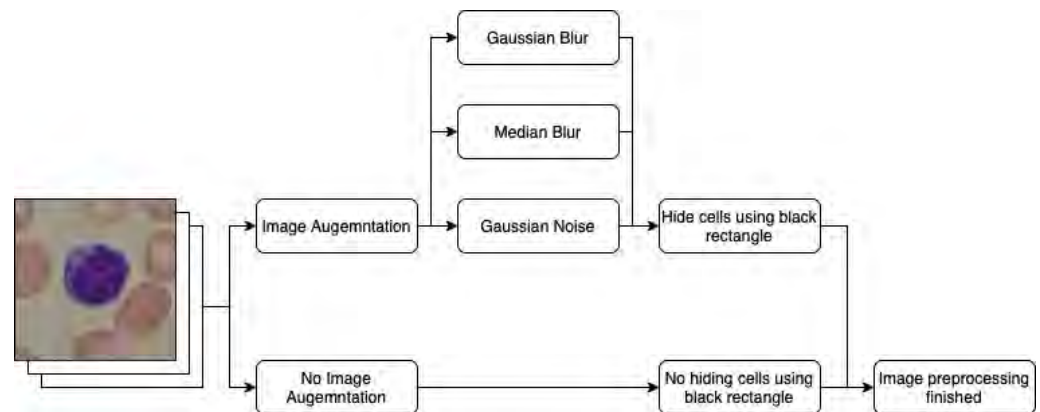


Figure 5. Graphical representation of image processing procedure.

2.3. Image Vectorization

In this section, the preprocessed images are converted into vectors of features ready to be interpreted by the statistical and machine learning models. In this research, background informativeness was measured in regard to one image channel at a time. As a result, each experiment conducted started with selection of the aforementioned channel. The available channels were red (R), green (G), blue (B), hue (H), saturation (S), and value (V).

The first three channels (red, green, and blue) are natural components of RGB-encoded images. However, the hue, saturation, and value metrics are the result of HSV image encoding. HSV is a common type of color expression in an image more akin to the recognition process performed by the human eye.

The hue channel contains information on what color is present in the image. It represents a 360° coordinate of rotation around the circle of colors. Typical representation of the hue coordinates associate the color red with a value of 0° , yellow with 60° , green with 120° , aqua with 180° , blue with 240° , and purple with 300° . It is important to note that the hue represents a rotation angle, so the difference between two hue values is represented by the measurement of the shortest arc connecting two points on the hue circle. For example, the colors red (0°) and purple (300°) are 60° degrees apart instead of 300° . In this research, the OpenCV library was used for image processing, which encoded the hue channels with values from 0 to 180. This behavior was kept for both conducting experiments and presenting the results.

The saturation contains information regarding the color's intensity. A saturation value of 0 results in the color gray, and the maximum value of 255 provides the most intense version of the color. On the other hand, the value channel contains information on the brightness of the color. The numeric value of 0 results in the color black, and the maximum value of 255 generates the brightest version of the color.

Equations (3)–(6) describe the process of image conversion from RGB encoding to HSV encoding:

$$M = \max\{R, G, B\}, m = \min\{R, G, B\}, \quad (3)$$

$$V = M/255 \tag{4}$$

$$S = \begin{cases} \frac{1-m}{M} & M > 0 \\ 0 & M = 0 \end{cases} \tag{5}$$

$$H = \begin{cases} \cos^{-1}\left(\frac{R-\frac{1}{2}G-\frac{1}{2}B}{\sqrt{R^2+B^2+G^2-RB-RG-GB}}\right) & G \geq B \\ 360^\circ - \cos^{-1}\left(\frac{R-\frac{1}{2}G-\frac{1}{2}B}{\sqrt{R^2+B^2+G^2-RB-RG-GB}}\right) & B > G \end{cases} \tag{6}$$

In the next step, the selected channel is vectorized by grouping its individual values into 30 equally spaced numeric bins and computing their distribution in the whole image. Such encoding provides information on what value occurs most frequently in the image. Pixels encoding black rectangles for obfuscation purposes were not included in the computations of color density distribution. An example of such calculations is presented in Figure 6.

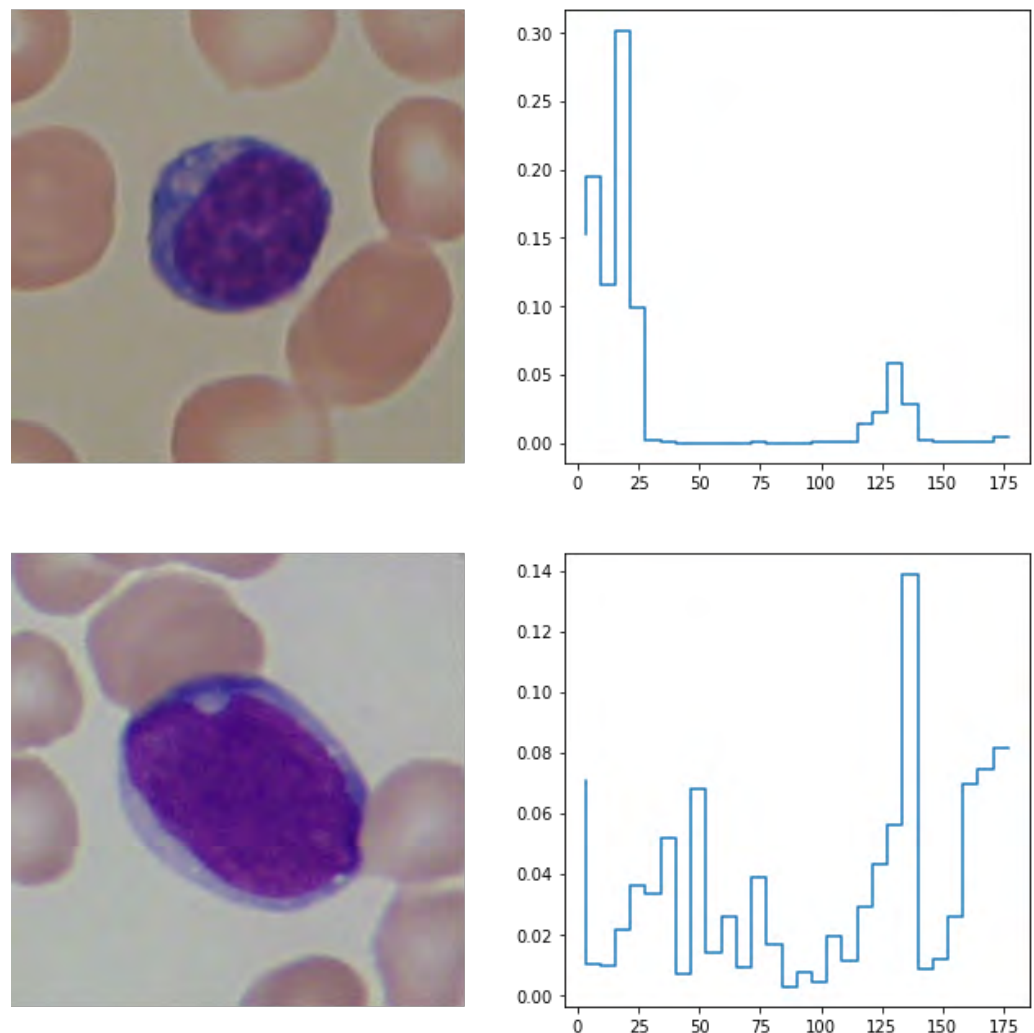


Figure 6. A presentation of images taken from healthy and sick patients (first column) and their corresponding hue distributions. HSV encoding was applied to both of the images.

2.4. Distribution Difference Measurement

In this section, images vectorized into a single channel’s values density distributions are compared to measure the amount of information stored in every channel’s histogram. Distributions from both the obfuscated and unmodified images are examined to establish

the amount of information stored in the color distributions of both the entire images and the backgrounds only. The quality of image classification measures the information stored in color distributions regarding the representation of healthy and sick individuals.

In the first step of the experiment, the set of vectorized images and the corresponding set of labels are randomly split into training and test sets. Then, the distribution vectors from the training set were split into subsets representing each class. Next, the averaged distributions were computed from the subsets, and each computed average distribution represented the class from which the samples were computed. In another step, the divergence metric was chosen to measure how different the samples were from each representative distribution. The divergence metrics chosen in this research were the cross-entropy and mean squared error, and they are explained in detail in Sections 2.4.1 and 2.4.2. The divergence metric was used to compute the divergence of each sample from the training set to each representative distribution, and their values were stored in the respective sets. Next, from each class divergence set, the mean and standard deviation were computed. Then, the statistical model training was finished and ready for performing test evaluation.

During the evaluation, the samples from the test set were subjected to divergence computation for the averaged distributions representing each class. The similarities to each distribution were normalized by subtracting the corresponding mean and dividing by the standard deviation. The normalized similarities of each sample to each distribution were compared. The sample was associated with the class whose representative distribution had the smallest normalized divergence. The evaluation of classification quality measures the amount of information contained in the value distribution. The process is graphically presented in Figure 7 and described in detail in the pseudo-code in Section 2.4.3.

2.4.1. Cross-Entropy

Cross-entropy is a technique from information theory for computing the divergence between two probability distributions. Equation (7) describes the process of metric calculation. The value density distribution can be interpreted as the probability that an individual pixel has a certain value. Such an interpretation allows the usage of cross-entropy in value density distribution classification:

$$H(P, Q) = - \sum_{i=0}^{|P|} P_i \log Q_i, |P| = |Q| \quad (7)$$

where P and Q are the density distributions subjected to the comparison.

2.4.2. Mean Squared Error

The mean squared error (MSE) is a commonly used technique for comparing two vectors. It is described by Equation (8). Due to powering the differences, the MSE is prone to disregarding multiple closely matched dimensions of two vectors in favor of penalizing a few outlying ones. This feature is useful in vector comparison because it forces the algorithms to even out their match functions instead of attuning easily matchable parts of the vector and disregarding the difficult ones:

$$MSE(P, Q) = \sum_{i=0}^{|P|} (P_i - Q_i)^2, |P| = |Q| \quad (8)$$

where P and Q are the vectors subjected to the comparison.

2.4.3. Algorithm

The graphical representation of the experiment process is depicted in Figure 7. The pseudo-code describing this process in detail is presented below Algorithm 1:

Algorithm 1: The mathematical formulation of experimental procedure examining distribution difference measurements.

1. Input the set of samples X ;
 2. Input the set of sample labels Y ;
 3. Input divergence function fn ;
 4. Randomly shuffle the set of indices $I = \{i|i \in N^+ \cap i \in (0; |X| + 1)\}$;
 5. $s = \lfloor \frac{2}{3}|X| \rfloor$;
 6. $I_{train} = I[:s]$;
 7. $I_{test} = I[s:]$;
 8. $X_{train} = \{X_i|i \in I_{train}\}$;
 9. $Y_{train} = \{Y_i|i \in I_{train}\}$;
 10. $X_{test} = \{X_i|i \in I_{test}\}$;
 11. $Y_{test} = \{Y_i|i \in I_{test}\}$;
 12. $X_0 = [\{X_{train}[i]|i \in (0, |X_{train}| + 1) \cap Y_{train}[i] = 0\}]$;
 13. $X_1 = [\{X_{train}[i]|i \in (0, |X_{train}| + 1) \cap Y_{train}[i] = 1\}]$;
 14. $M_0 = [\{\frac{1}{|X_0|} \sum_{j \in X_0} X_0[j, i]|i \in N^+ \cap i \in |X_0[0]|\}]$;
 15. $M_1 = [\{\frac{1}{|X_1|} \sum_{j \in X_1} X_1[j, i]|i \in N^+ \cap i \in |X_1[0]|\}]$;
 16. $D_0 = \{fn(X_{train}[i], M_0)|i \in N^+ \cap i \in (0; |X_{train}| + 1)\}$;
 17. $D_1 = \{fn(X_{train}[i], M_1)|i \in N^+ \cap i \in (0; |X_{train}| + 1)\}$;
 18. $m_0, s_0, m_1, s_1 = mean(D_0), std(D_0), mean(D_1), std(D_1)$;
 19. $Dt_0 = \{\frac{1}{s_0}(fn(X_{test}[i], M_0) - m_0)|i \in N^+ \cap i \in (0, |X_{test}| + 1)\}$;
 20. $Dt_1 = \{\frac{1}{s_1}(fn(X_{test}[i], M_1) - m_1)|i \in N^+ \cap i \in (0, |X_{test}| + 1)\}$;
 21. $P = [\{Dt_1[i] < Dt_0[i]|i \in N^+ \cap i \in (0; |Dt_0| + 1)\}]$;
 22. Compare the prediction vector P with the label vector Y_{test} .
-

2.5. Shannon Entropy

The Shannon entropy is a mathematical tool from the field of information theory that allows measuring the amount of uncertainty the probability distribution contains. The more evenly spaced the probability among the distribution states, the higher the value of the Shannon entropy. The computation of the Shannon entropy is performed using Equation (9):

$$H(P) = - \sum_{i=0}^{|P|} P_i \log P_i \quad (9)$$

where P is the vectorized density distribution subjected to the Shannon entropy computation.

In this research, the Shannon entropy was used to quantify the uncertainty associated with each channel's value distribution and evaluate whether there was a significant difference between the Shannon entropy of the samples from the healthy class and the entropy of the sick class of samples. The significance of the difference in entropy measurements was established by using the Shannon entropy as a single-value determinant in the classification of whether a patient was healthy or sick. The classification was performed by fitting the logistic regression model on randomly split training data and evaluating it on the test data.

2.6. Machine Learning Algorithms

The last experiment aimed to apply machine learning algorithms directly to the channel's value distribution to attempt to perform the classification. The algorithms used for this task were XGBoost and logistic regression. The former is one of the most robust, state-of-the-art machine learning algorithms capable of extracting complex, multi-dimensional patterns. The latter is one of the simplest classification algorithms, providing a basis for comparison.

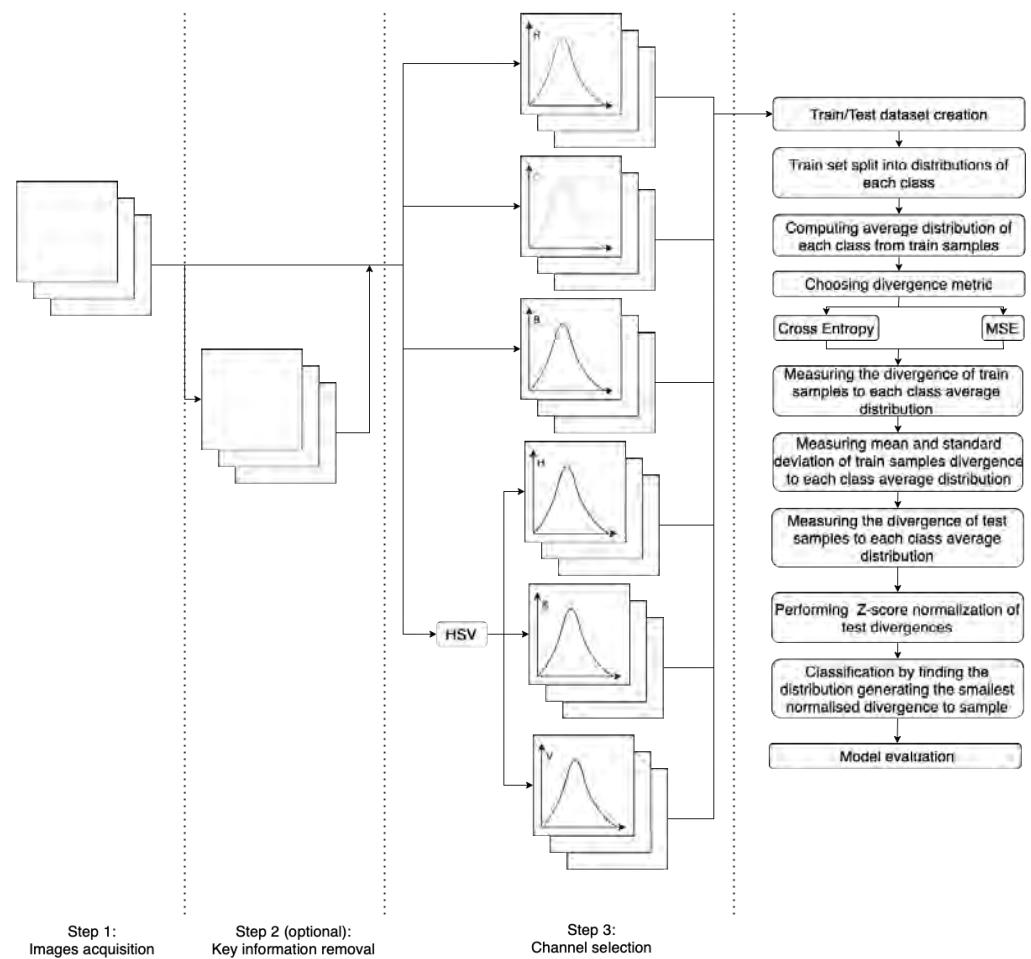


Figure 7. A graphical representation of the classification process involving images’ color distribution comparison, using cross-entropy and mean squared error (MSE) as divergence metrics.

2.7. Metrics

The metrics used for classification quality measurement were *accuracy*, *precision*, *recall*, and F_1 score. The metrics are described by Equations (10)–(13). The following abbreviations are used to simplify these equations:

- TP = true positive;
- TN = true negative;
- FP = false positive;
- FN = false negative.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$F_1 = \frac{2TP}{2TP + FN + FP} \tag{13}$$

The accuracy is a global evaluation metric, and it assesses the model compared to all the data. On the other hand, the precision, recall, and F_1 score are local metrics that evaluate performance regarding the classification of one class vs. all of them.

3. Results

The results of the experiments described in Sections 2.4–2.6 are presented in Tables 2–8. Each experiment was repeated 30 times with randomly selected training and test sets. The tables present the averaged metric values from 30 trials. Section 3.1 presents the values acquired from running experiments on images with and without obfuscation without adding data augmentation. Section 3.3 examines the influence of data augmentation. The experiments were run on the images with and without obfuscation, and data augmentation was applied. The results were presented for the channel, and the experimental results presented in Section 3.1 contained the most information.

3.1. Background Information Measurement

This subsection presents the information measurements in the image’s background in the form of the classification quality and compares it against the information in the whole image. Table 2 contains an evaluation of the distribution difference measurement described in Section 2.4. Table 3 presents the results from evaluating the Shannon entropy of the value distribution as a sole healthy or sick determinant. The experiment procedure is described in Section 2.5. Table 4 contains information on the machine learning algorithm’s performance for the value distribution.

Each table contains the column “Lymphocytes Obfuscated”. The rows with the value “True” in this column contain results from experiments with parts of the image representing lymphocytes covered by a black rectangle. As a result, these rows depict the informativeness of the image background. The ones with “False” in the first column contain results from experiments that used whole images and provide the informativeness measurements of all images. These serve as a basis of comparison for the experiments with obfuscated lymphocytes.

Each table contains the column “Channel” as well. This column presents the information regarding channel selection for the experiment. Each experiment evaluated the informativeness of only one channel at a time to establish whether the background in any channel contained unwanted information.

Table 2. The classification results by the images’ color distribution comparison to the averaged distributions representing their respective classes. The cross-entropy and mean squared error (MSE) metrics were applied. The images were unmodified and had their lymphocytes covered (first column). No image augmentation was applied in this experiment.

| Lymphocytes Obfuscated | Channel | Cross Entropy Acc. | Cross Entropy F_1 (Healthy) | Cross Entropy F_1 (Sick) | MSE Acc. | MSE F_1 (Healthy) | MSE F_1 (Sick) |
|------------------------|------------|--------------------|-------------------------------|----------------------------|----------|---------------------|------------------|
| False | B | 0.50 | 0.56 | 0.41 | 0.68 | 0.68 | 0.68 |
| True | B | 0.55 | 0.59 | 0.49 | 0.69 | 0.68 | 0.70 |
| False | G | 0.81 | 0.81 | 0.81 | 0.62 | 0.63 | 0.61 |
| True | G | 0.47 | 0.53 | 0.38 | 0.60 | 0.62 | 0.58 |
| False | R | 0.53 | 0.51 | 0.54 | 0.46 | 0.49 | 0.43 |
| True | R | 0.37 | 0.49 | 0.18 | 0.45 | 0.48 | 0.42 |
| False | Hue | 0.85 | 0.86 | 0.83 | 0.77 | 0.79 | 0.73 |
| True | Hue | 0.83 | 0.85 | 0.81 | 0.82 | 0.84 | 0.80 |
| False | Saturation | 0.79 | 0.80 | 0.77 | 0.79 | 0.83 | 0.72 |
| True | Saturation | 0.73 | 0.72 | 0.74 | 0.79 | 0.83 | 0.73 |
| False | Value | 0.45 | 0.51 | 0.36 | 0.50 | 0.51 | 0.48 |
| True | Value | 0.38 | 0.45 | 0.28 | 0.52 | 0.53 | 0.49 |

Table 3. The classification results by the images' color distributions' Shannon entropy measurements. The images were unmodified and had their lymphocytes covered (first column). No image augmentation was applied in this experiment.

| Lymphocytes Obfuscated | Channel | Avg. Shannon Entropy (Healthy) | Std. Shannon Entropy (Healthy) | Avg. Shannon Entropy (Sick) | Std. Shannon Entropy (Sick) | Acc | F ₁ (Healthy) | F ₁ (Sick) |
|------------------------|------------|--------------------------------|--------------------------------|-----------------------------|-----------------------------|------|--------------------------|-----------------------|
| False | B | 0.43 | 0.01 | 0.45 | 0.02 | 0.51 | 0.31 | 0.51 |
| True | B | 0.43 | 0.02 | 0.44 | 0.02 | 0.51 | 0.30 | 0.51 |
| False | G | 0.48 | 0.02 | 0.50 | 0.02 | 0.55 | 0.37 | 0.55 |
| True | G | 0.43 | 0.03 | 0.46 | 0.03 | 0.54 | 0.39 | 0.54 |
| False | R | 0.44 | 0.02 | 0.47 | 0.02 | 0.54 | 0.38 | 0.54 |
| True | R | 0.39 | 0.03 | 0.42 | 0.04 | 0.54 | 0.39 | 0.54 |
| False | Hue | 0.63 | 0.03 | 0.71 | 0.07 | 0.68 | 0.70 | 0.65 |
| True | Hue | 0.57 | 0.03 | 0.69 | 0.10 | 0.81 | 0.84 | 0.78 |
| False | Saturation | 0.49 | 0.03 | 0.51 | 0.02 | 0.55 | 0.38 | 0.55 |
| True | Saturation | 0.44 | 0.03 | 0.47 | 0.04 | 0.51 | 0.36 | 0.51 |
| False | Value | 0.43 | 0.02 | 0.44 | 0.02 | 0.51 | 0.30 | 0.51 |
| True | Value | 0.39 | 0.03 | 0.41 | 0.03 | 0.52 | 0.34 | 0.53 |

Table 4. The classification results by images' color distribution interpretation by machine learning algorithms. The XGBoost and logistic regression models were used in this experiment. The images were unmodified and had their lymphocytes covered (first column). No image augmentation was applied in this experiment.

| Lymphocytes Obfuscated | Channel | XGBoost Acc | XGBoost F ₁ (Healthy) | XGBoost F ₁ (Sick) | Logistic Regression Acc. | Logistic Regression F ₁ (Healthy) | Logistic Regression F ₁ (Sick) |
|------------------------|------------|-------------|----------------------------------|-------------------------------|--------------------------|--|---|
| False | B | 0.83 | 0.83 | 0.83 | 0.53 | 0.32 | 0.53 |
| True | B | 0.82 | 0.82 | 0.82 | 0.55 | 0.37 | 0.55 |
| False | G | 0.96 | 0.96 | 0.96 | 0.50 | 0.28 | 0.51 |
| True | G | 0.86 | 0.87 | 0.85 | 0.50 | 0.28 | 0.51 |
| False | R | 0.87 | 0.87 | 0.86 | 0.47 | 0.23 | 0.49 |
| True | R | 0.80 | 0.80 | 0.79 | 0.48 | 0.28 | 0.49 |
| False | Hue | 0.94 | 0.95 | 0.94 | 0.57 | 0.41 | 0.57 |
| True | Hue | 0.93 | 0.94 | 0.93 | 0.60 | 0.46 | 0.59 |
| False | Saturation | 0.94 | 0.94 | 0.94 | 0.54 | 0.32 | 0.54 |
| True | Saturation | 0.88 | 0.89 | 0.88 | 0.55 | 0.36 | 0.55 |
| False | Value | 0.84 | 0.84 | 0.84 | 0.48 | 0.25 | 0.49 |
| True | Value | 0.80 | 0.81 | 0.80 | 0.50 | 0.30 | 0.50 |

Figure 8 represents the averaged value distribution obtained from each channel for each of four states: images of healthy patients without lymphocytes obfuscated, images of healthy patients with lymphocytes obfuscated, images of sick patients without lymphocytes obfuscated, and images of sick patients with lymphocytes obfuscated. These four states are represented in their respective columns. The rows of the chart grid represent each of the six channels: red, green, blue, hue, saturation, and value.

3.2. Comparison with the Literature

A comparison of the most promising models obtained in this research with other works is presented in Table 5. The results were compared against the outcomes of our previous work and work of Rodrigues et al. [33,35], which according to our literature review obtained the best results on the ALL-IDB2 dataset.

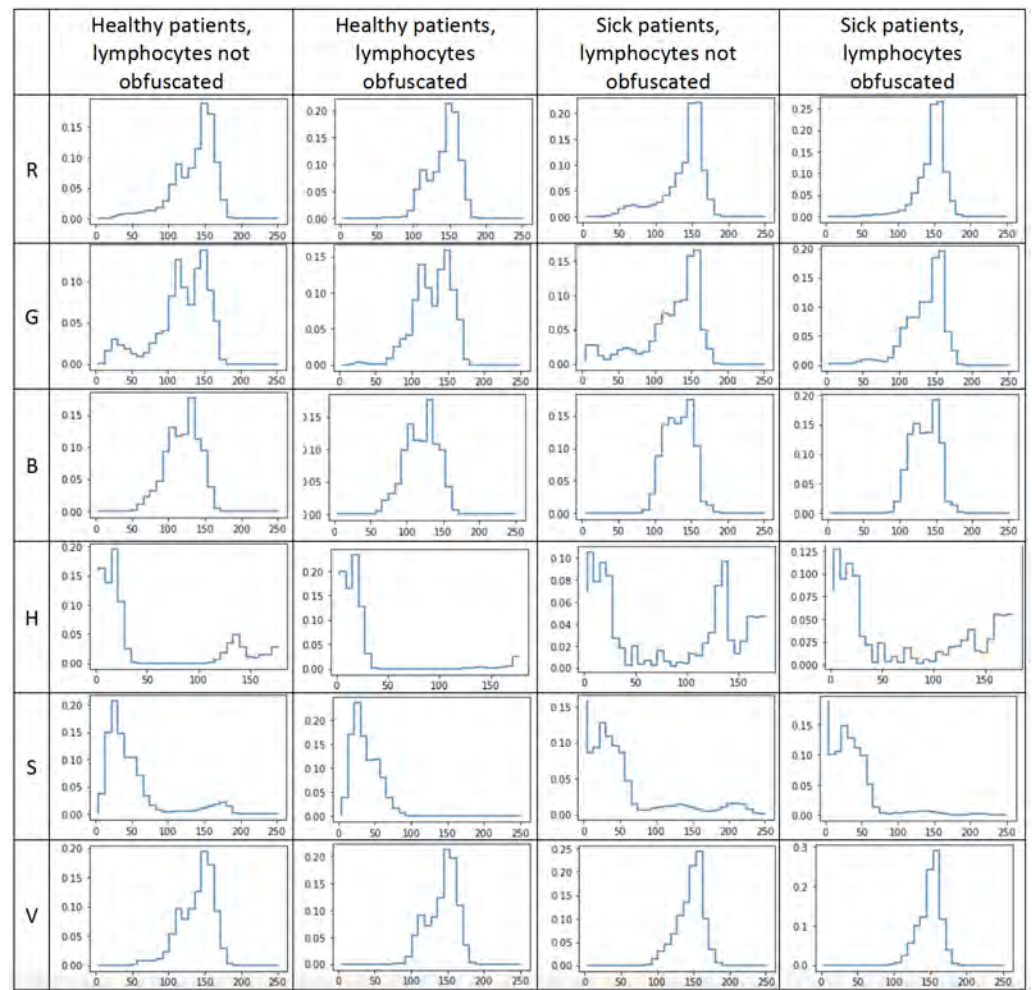


Figure 8. Presentation of averaged distributions of each channel’s values for each combination of images. The columns represent distributions taken from images of healthy patients without lymphocytes covered, images of healthy patients with lymphocytes covered, images of sick patients without lymphocytes covered, and images of sick patients with lymphocytes covered. Each row represents one channel. The presented channels are red (R), green (G), blue (B), hue (H), saturation (S), and value (V).

Table 5. Comparison of the best models with and without lymphocytes obfuscated against other works. Shortcut “dist” stands for “distribution”.

| Article | Input Data | Input Size | Obfuscation | Parameters | Accuracy | Precision | Recall | F_1 |
|-----------|-------------|------------|-------------|------------|----------|-----------|--------|-------|
| This work | Green dist. | 51 | False | 6.4 K | 0.960 | 0.960 | 0.960 | 0.959 |
| This work | Hue dist. | 36 | True | 6.4 K | 0.935 | 0.935 | 0.935 | 0.934 |
| [35] | RGB image | 150 K | False | 3.4 M | 0.948 | 0.950 | 0.951 | 0.948 |
| [33] | RGB image | 150 K | False | 25 M | 0.985 | 0.986 | 0.985 | 0.984 |

3.3. Influence of Data Augmentation

Tables 2–4 prove that the channel containing the most informative image background was the hue channel, with the classification accuracy obtained from merely a background hue value distribution ranging from 82% to 93%. As a result, this channel was subjected to further investigation to determine the data augmentation application’s influence on obfuscated and unmodified images.

Table 6 contains the results of applying Gaussian blur with kernel sizes of 3, 9, 21, and 51. Table 7 presents the results of median blur application with kernel sizes of 3, 9, 21, and 51. Table 8 presents influence of Gaussian noise application on the images with variances of 0.0001, 0.01, and 0.1. The experiments were conducted with and without lymphocyte obfuscation. Each experiment was repeated 30 times, and its values were averaged.

Table 6. Results of Gaussian blur application on the quality of classification.

| Lymphocytes Obfuscated | Kernel | XGBoost Acc. | Logistic Regression Acc. | Cross-Entropy Acc.k | MSE Acc. | Shannon Acc. |
|------------------------|--------|--------------|--------------------------|---------------------|----------|--------------|
| False | 51 | 0.93 | 0.60 | 0.79 | 0.74 | 0.64 |
| False | 21 | 0.93 | 0.58 | 0.79 | 0.77 | 0.69 |
| False | 0 | 0.94 | 0.57 | 0.85 | 0.77 | 0.68 |
| False | 9 | 0.95 | 0.58 | 0.83 | 0.78 | 0.69 |
| False | 3 | 0.95 | 0.57 | 0.83 | 0.77 | 0.69 |
| True | 51 | 0.91 | 0.65 | 0.82 | 0.83 | 0.79 |
| True | 21 | 0.93 | 0.63 | 0.78 | 0.82 | 0.82 |
| True | 9 | 0.93 | 0.62 | 0.83 | 0.81 | 0.82 |
| True | 3 | 0.93 | 0.61 | 0.82 | 0.82 | 0.81 |
| True | 0 | 0.93 | 0.60 | 0.83 | 0.82 | 0.81 |

Table 7. Results of median blur application on the quality of classification.

| Lymphocytes Obfuscated | Kernel | XGBoost Acc. | Logistic Regression Acc. | Cross-Entropy Acc.k | MSE Acc. | Shannon Acc. |
|------------------------|--------|--------------|--------------------------|---------------------|----------|--------------|
| False | 51 | 0.94 | 0.59 | 0.73 | 0.77 | 0.67 |
| False | 0 | 0.94 | 0.57 | 0.85 | 0.77 | 0.68 |
| False | 9 | 0.95 | 0.58 | 0.81 | 0.79 | 0.69 |
| False | 21 | 0.95 | 0.59 | 0.81 | 0.79 | 0.69 |
| False | 3 | 0.95 | 0.57 | 0.83 | 0.77 | 0.69 |
| True | 51 | 0.91 | 0.63 | 0.72 | 0.80 | 0.80 |
| True | 3 | 0.93 | 0.61 | 0.82 | 0.82 | 0.81 |
| True | 9 | 0.93 | 0.63 | 0.81 | 0.81 | 0.82 |
| True | 0 | 0.93 | 0.60 | 0.83 | 0.82 | 0.81 |
| True | 21 | 0.94 | 0.65 | 0.78 | 0.81 | 0.81 |

Table 8. Results of multiplicative Gaussian noise application on the quality of classification.

| Lymphocytes Obfuscated | Kernel | XGBoost Acc. | Logistic Regression Acc. | Cross-Entropy Acc.k | MSE Acc. | Shannon Acc. |
|------------------------|--------|--------------|--------------------------|---------------------|----------|--------------|
| False | 0 | 0.94 | 0.57 | 0.84 | 0.76 | 0.68 |
| False | 0.0001 | 0.95 | 0.57 | 0.84 | 0.76 | 0.68 |
| False | 0.01 | 0.95 | 0.56 | 0.84 | 0.76 | 0.68 |
| False | 0.1 | 0.95 | 0.56 | 0.82 | 0.76 | 0.68 |
| True | 0.01 | 0.93 | 0.59 | 0.82 | 0.82 | 0.80 |
| True | 0.0001 | 0.93 | 0.60 | 0.83 | 0.82 | 0.81 |
| True | 0 | 0.93 | 0.60 | 0.82 | 0.82 | 0.81 |
| True | 0.1 | 0.95 | 0.59 | 0.82 | 0.82 | 0.81 |

4. Discussion

The experimental results presented in Tables 2–4 determined that the hue channel contained the highest amount of image background information. All methods (except for logistic regression) achieved averaged test accuracies above 80%, with the XGBoost model having 93% accuracy. Such scores were obtained merely for the hue distribution of the image background, with the informativeness unconfirmed by academic knowledge.

The methods for background information measurement described in Sections 2.4–2.6 proved to be efficient in determining whether the background contained classification-sensitive information. These methods can be used as training dataset evaluation techniques. Suppose that a supposedly neutral classification-wise background contains the required information. In this case, artificial intelligence models such as deep convolutional neural networks may learn to recognize some unwanted, dataset-related temporal pattern in the background instead of true generalization to real-world scenarios. This method can help in the evaluation of datasets containing high-quality data.

The logistic regression model trained on raw distributions achieved the worst accuracy, and the XGBoost model achieved the best accuracy. This suggests that, although the

information contained in the distribution of the values is substantial, it is not yet obvious. Such background information may not be caught during exploratory data analysis and interfere with machine learning algorithms' training quality. For these reasons, the background informativeness evaluation may prove beneficial in fool-proofing artificial intelligence systems.

Data augmentation techniques reduced the background informativeness extracted by the application of cross-entropy. It had little effect on the Shannon entropy-based classification and did not affect the XGBoost or *MSE* classification quality. This suggests that information in the background of ALL-IDB images may be more complex than just random class-specific noise. The authors plan to investigate this phenomenon further.

The research indicates that information is contained in the hue distribution of ALL-IDB image backgrounds. The XGBoost model achieved 93% accuracy on merely the hue distribution in the background. Such high classification quality has been achieved by just studying the background, which is supposed to be classification-neutral. According to our literature review, the primary indication of acute lymphoblastic leukemia is an examination of the lymphocytes. Academics do not unanimously recognize the lymphocytes' surroundings' informativeness. However, this research proves that this is not the case in this dataset. It is possible that the suspected "classification-neutral background" contains information allowing for healthy and sick discrimination. The authors plan to investigate this phenomenon further.

In our previous work [35], the best combination of artificial neural networks for raw image encoding, classification heads, and image augmentation resulted in an average classification quality of 94.8%. The neural network used for this task was MobileNet v2, the state-of-the-art neural network for numerous image-processing tasks containing 3.4 million parameters. In this work, the XGBoost algorithm alone, which interpreted the green color value distribution, achieved a classification accuracy of 96.0%. A similar result was obtained by the XGBoost algorithm interpreting the hue distributions of images with lymphocytes obfuscated, achieving a classification quality of 93.0%. Much simpler machine learning models operating on limited data obtained results comparable to the state-of-the-art deep learning method. This suggests that the neural networks experienced overfitting during training despite the application of data augmentation techniques. This also suggests that the task of leukemia classification may be performed using much more straightforward and cost-effective methods that also benefit from human interpretability. The authors plan to investigate this phenomenon further. According to our literature review, at the moment of writing this article, the best result was obtained by Rodrigues et al. [33], with an accuracy of 98.5%. This result is 2.5% percentage points higher than our best model. However, we obtained our results using around 3000 times fewer input data and almost 4000 times fewer parameters. As a result, our best model obtained comparable results, requiring much less computational power while remaining interpretable by humans.

The results indicate that hue distribution of a lymphocyte's surroundings contains information supporting leukemia classification. However, a distribution is, by definition, an aggregation of the information stripped from temporal patterns akin to hidden Markov chains. It is possible that more detailed information interpretable by humans can be found during image examination in the hue channel. The authors plan to investigate this claim further.

5. Conclusions

The proposed background informativeness measurement proved its efficiency in dataset quality evaluation. This method based on the Shannon entropy, cross-entropy, and machine learning algorithms provides a comprehensive estimation of value distribution patterns in the background that may cause artificial intelligence models to overfit them instead of finding generalized solutions applicable to real-world problems.

The conducted research on background informativeness on the ALL-IDB dataset found a substantial amount of information in the hue distribution of the image background. The

hue distributions of healthy people and patients who had acute lymphoblastic leukemia differed vastly from each other and by their Shannon entropy measurements. In this research, the lymphocytes were obfuscated with a black rectangle, so this information was contained within the supposedly classification-neutral background. The authors plan to investigate this phenomenon further.

The highest quality of classification was achieved while examining the green channel distribution using the XGBoost model. On average, it achieved 96.0% accuracy. This is a result comparable with deep neural networks while requiring much less computational power and providing a more human-interpretable decision process.

The background hue distribution differences between the images of healthy and sick patients require further investigation. It is unknown whether the differentiable factor is spread uniformly over the whole picture or is concentrated around semantically separable entities. Medical professionals must examine the nature of these changes to understand the features' origins and extrapolate the applicability of this knowledge. The authors plan to investigate this phenomenon further.

Author Contributions: Conceptualization, K.P. and D.L.; methodology, K.P. and D.L., software K.P.; validation K.P. and D.L.; formal analysis, K.P., D.L. and T.A.; investigation K.P., D.L. and T.A.; resources, K.P., D.L. and T.A.; data curation, T.A.; writing—original draft preparation, K.P. and D.L.; writing—review and editing, K.P. and T.A.; visualization, K.P. and D.L.; project administration, K.P. and T.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

| | |
|-----|------------------------------|
| ALL | Acute lymphoblastic leukemia |
| AML | Acute myleoid leukemia |
| CLL | Chronic lymphocytic leukemia |
| CLM | Chronic myleoid leukemia |
| CNN | Convolutional neural network |

References

1. Andrade, A.R.; Vogado, L.H.; Veras, R.D.M.S.; Silva, R.R.; Araujo, F.H.; Medeiros, F.N. Recent computational methods for white blood cell nuclei segmentation: A comparative study. *Comput. Methods Programs Biomed.* **2019**, *173*, 1–14. [[CrossRef](#)] [[PubMed](#)]
2. Labati, R.D.; Piuri, V.; Scotti, F. The Acute Lymphoblastic Leukemia Image Database for Image Processing. In Proceedings of the 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 11–14 September 2011.
3. Mohamed, M.; Far, B.; Guaily, A. An efficient technique for white blood cells nuclei automatic segmentation. In Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Seoul, Korea, 14–17 October 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 220–225.
4. Sarrafzadeh, O.; Rabbani, H.; Talebi, A.; Banaem, H.U. Selection of the best features for leukocytes classification in blood smear microscopic images. In Proceedings of the Medical Imaging 2014: Digital Pathology, San Diego, CA, USA, 15–20 February 2014; SPIE: Bellingham, WA, USA, 2014; Volume 9041, pp. 159–166.
5. Zheng, X.; Wang, Y.; Wang, G.; Liu, J. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron* **2018**, *107*, 55–71. [[CrossRef](#)] [[PubMed](#)]
6. Madhloom, H.; Kareem, S.; Ariffin, H.; Zaidan, A.; Alanazi, H.; Zaidan, B. An automated white blood cell nucleus localization and segmentation using image arithmetic and automatic threshold. *J. Appl. Sci.* **2010**, *10*, 959–966. [[CrossRef](#)]
7. Arslan, S.; Ozyurek, E.; Gunduz-Demir, C. A color and shape based algorithm for segmentation of white blood cells in peripheral blood and bone marrow images. *Cytom. Part A* **2014**, *85*, 480–490. [[CrossRef](#)]

8. Nazlibilek, S.; Karacor, D.; Ercan, T.; Sazli, M.H.; Kalender, O.; Ege, Y. Automatic segmentation, counting, size determination and classification of white blood cells. *Measurement* **2014**, *55*, 58–65. [\[CrossRef\]](#)
9. Prinyakupt, J.; Pluempitiwiriyaewej, C. Segmentation of white blood cells and comparison of cell morphology by linear and naïve Bayes classifiers. *Biomed. Eng. Online* **2015**, *14*, 63. [\[CrossRef\]](#)
10. Nasir, A.A.; Mashor, M.; Rosline, H. Unsupervised colour segmentation of white blood cell for acute leukaemia images. In Proceedings of the 2011 IEEE International Conference on Imaging Systems and Techniques, Penang, Malaysia, 17–18 May 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 142–145.
11. Mohapatra, S.; Samanta, S.S.; Patra, D.; Satpathi, S. Fuzzy based blood image segmentation for automated leukemia detection. In Proceedings of the 2011 International Conference on Devices and Communications (ICDeCom), Ranchi, India, 24–25 February 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 1–5.
12. Madhukar, M.; Agaian, S.; Chronopoulos, A.T. New decision support tool for acute lymphoblastic leukemia classification. In Proceedings of the Image Processing: Algorithms and Systems X; and Parallel Processing for Imaging Applications II, Burlingame, CA, USA, 23–25 January 2012; SPIE: Bellingham, WA, USA, 2012; Volume 8295, pp. 367–378.
13. Amin, M.M.; Kermani, S.; Talebi, A.; Oghli, M.G. Recognition of acute lymphoblastic leukemia cells in microscopic images using k-means clustering and support vector machine classifier. *J. Med. Signals Sens.* **2015**, *5*, 49.
14. Sarrafzadeh, O.; Dehnavi, A.M.; Rabbani, H.; Talebi, A. A simple and accurate method for white blood cells segmentation using K-means algorithm. In Proceedings of the 2015 IEEE Workshop on Signal Processing Systems (SiPS), Hangzhou, China, 14–16 October 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–6.
15. Vincent, I.; Kwon, K.R.; Lee, S.H.; Moon, K.S. Acute lymphoid leukemia classification using two-step neural network classifier. In Proceedings of the 2015 21st Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV), Mokpo, Korea, 28–30 January 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–4.
16. Vogado, L.H.; Veras, R.D.M.S.; Andrade, A.R.; e Silva, R.R.; De Araujo, F.H.; De Medeiros, F.N. Unsupervised leukemia cells segmentation based on multi-space color channels. In Proceedings of the 2016 IEEE International Symposium on Multimedia (ISM), San Jose, CA, USA, 11–13 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 451–456.
17. Kumar, P.; Vasuki, S. Automated diagnosis of acute lymphocytic leukemia and acute myeloid leukemia using multi-SV. *J. Biomed. Imaging Bioeng.* **2017**, *1*, 20–24.
18. Mohammed, E.A.; Mohamed, M.M.; Naugler, C.; Far, B.H. Chronic lymphocytic leukemia cell segmentation from microscopic blood images using watershed algorithm and optimal thresholding. In Proceedings of the 2013 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Regina, SK, Canada, 5–8 May 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 1–5.
19. Abdeldaim, A.M.; Sahlol, A.T.; Elhoseny, M.; Hassanien, A.E. Computer-aided acute lymphoblastic leukemia diagnosis system based on image analysis. In *Advances in Soft Computing and Machine Learning in Image Processing*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 131–147.
20. Rehman, A.; Abbas, N.; Saba, T.; Rahman, S.I.U.; Mehmood, Z.; Kolivand, H. Classification of acute lymphoblastic leukemia using deep learning. *Microsc. Res. Tech.* **2018**, *81*, 1310–1317. [\[CrossRef\]](#)
21. Prellberg, J.; Kramer, O. Acute lymphoblastic leukemia classification from microscopic images using convolutional neural networks. In *ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 53–61.
22. Ahmed, N.; Yigit, A.; Isik, Z.; Alpkocak, A. Identification of leukemia subtypes from microscopic images using convolutional neural network. *Diagnostics* **2019**, *9*, 104. [\[CrossRef\]](#)
23. Guo, Z.; Wang, Y.; Liu, L.; Sun, S.; Feng, B.; Zhao, X. Siamese Network-Based Few-Shot Learning for Classification of Human Peripheral Blood Leukocyte. In Proceedings of the 2021 IEEE 4th International Conference on Electronic Information and Communication Technology (ICEICT), Xi'an, China, 18–20 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 818–822.
24. Abhishek, A.; Jha, R.K.; Sinha, R.; Jha, K. Automated classification of acute leukemia on a heterogeneous dataset using machine learning and deep learning techniques. *Biomed. Signal Process. Control* **2022**, *72*, 103341. [\[CrossRef\]](#)
25. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [\[CrossRef\]](#)
26. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [\[CrossRef\]](#)
27. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. *arXiv* **2016**, arXiv:1611.05431.
28. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [\[CrossRef\]](#)
29. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In Proceedings of the ICML deep learning workshop, Lille, France, 6–11 July 2015; Volume 2.
30. Wang, Y.; Yao, Q.; Kwok, J.T.; Ni, L.M. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.* **2020**, *53*, 63. [\[CrossRef\]](#)
31. Ojala, T.; Pietikainen, M.; Maenpää, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [\[CrossRef\]](#)

32. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 886–893.
33. Rodrigues, L.F.; Backes, A.R.; Travençolo, B.A.N.; de Oliveira, G.M.B. Optimizing a Deep Residual Neural Network with Genetic Algorithm for Acute Lymphoblastic Leukemia Classification. *J. Digit. Imaging* **2022**, *35*, 623–637. [[CrossRef](#)]
34. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531.
35. Pałczyński, K.; Śmigiel, S.; Gackowska, M.; Ledziński, D.; Bujnowski, S.; Lutowski, Z. IoT Application of Transfer Learning in Hybrid Artificial Intelligence Systems for Acute Lymphoblastic Leukemia Classification. *Sensors* **2021**, *21*, 8025. [[CrossRef](#)]
36. Golik, P.; Doetsch, P.; Ney, H. Cross-Entropy vs. Squared Error Training: A Theoretical and Experimental Comparison. In Proceedings of the International Sport and Culture Association, Lyon, France, 25–29 August 2013; pp. 1756–1760. [[CrossRef](#)]
37. Peng, J.; Lee, K.; Ingersoll, G. An Introduction to Logistic Regression Analysis and Reporting. *J. Educ. Res.* **2002**, *96*, 3–14. [[CrossRef](#)]
38. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *arXiv* **2016**, arXiv:1603.02754.

Article

ECG Signal Classification Using Deep Learning Techniques Based on the PTB-XL Dataset

Sandra Śmigiel ^{1,*}, Krzysztof Pałczyński ² and Damian Ledziński ²

¹ Faculty of Mechanical Engineering, UTP University of Science and Technology in Bydgoszcz, 85-796 Bydgoszcz, Poland

² Faculty of Telecommunications, Computer Science and Electrical Engineering, UTP University of Science and Technology in Bydgoszcz, 85-796 Bydgoszcz, Poland; krzysztof@palczynski.com.pl (K.P.); damian.ledzinski@utp.edu.pl (D.L.)

* Correspondence: sandra.smigiel@utp.edu.pl; Tel.: +48-52-340-8346

Abstract: The analysis and processing of ECG signals are a key approach in the diagnosis of cardiovascular diseases. The main field of work in this area is classification, which is increasingly supported by machine learning-based algorithms. In this work, a deep neural network was developed for the automatic classification of primary ECG signals. The research was carried out on the data contained in a PTB-XL database. Three neural network architectures were proposed: the first based on the convolutional network, the second on SincNet, and the third on the convolutional network, but with additional entropy-based features. The dataset was divided into training, validation, and test sets in proportions of 70%, 15%, and 15%, respectively. The studies were conducted for 2, 5, and 20 classes of disease entities. The convolutional network with entropy features obtained the best classification result. The convolutional network without entropy-based features obtained a slightly less successful result, but had the highest computational efficiency, due to the significantly lower number of neurons.

Keywords: ECG signal; classification; PTB-XL; deep learning



Citation: Śmigiel, S.; Pałczyński, K.; Ledziński, D. ECG Signal Classification Using Deep Learning Techniques Based on the PTB-XL Dataset. *Entropy* **2021**, *23*, 1121. <https://doi.org/10.3390/e23091121>

Academic Editor: Ernestina Menasalvas

Received: 5 July 2021

Accepted: 25 August 2021

Published: 28 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

According to publicly available reports, cardiovascular disease remains the leading cause of mortality worldwide [1]. One of the main causes of cardiovascular diseases is cardiac arrhythmia, in which the heartbeat deviates from typical beating patterns [2]. However, there are many types of irregular heartbeat. Accurate classification of heart disease types can aid in diagnosis and treatment [3].

An electrocardiogram (ECG) is a widely used, reliable, noninvasive approach to diagnosing cardiovascular disease. The standard ECG consists of 12 leads [4]. Traditionally, ECG results are manually interpreted by cardiologists based on a set of diagnosis criteria and experience. However, manual interpretation is time consuming and requires skill. Incorrectly interpreted ECG results may give rise to incorrect clinical decisions and lead to a threat to human life and health. With the rapid development of ECG and, at the same time, an insufficient number of cardiologists, the accurate and automatic diagnosis of ECG signals has become an interesting research topic for many scientists.

Over the past decade, numerous attempts have been made to identify a 12-lead clinical ECG, largely on the basis of the availability of large, public, open-source ECG data collections. Previous literature on ECG databases has shown a methodological division: signal processing and machine learning [5,6]. On the one hand, digital signal processing methods mainly include low- or high-pass filters, fast Fourier transform, and wavelet transform [7]. In this area, many algorithms are based on three processes: feature extraction, feature selection, and classification [8]. On the other hand, an alternative method is the application of machine learning methods. Such an application would primarily focus on

the automatic recognition of patterns that classify various disease entities, a method that is gaining greater importance in medical practice.

Algorithms known as deep neural networks have become particularly important in the last five years. Deep learning models have proven to be useful in increasing the effectiveness of diagnoses of cardiovascular diseases using ECG signals. By using the cascade of heterogeneous layers of neural networks to gradually extract increasingly high-level features, they lead to ever-improving neural networks built on their basis. Deep neural networks are reaching their zenith in various areas where artificial intelligence algorithms are applied.

In recent years, machine learning models have given rise to huge innovations in many areas, including image processing, natural language processing, computer games, and medical applications [9]. To date, however, the lack of adequate databases, well-defined assessment procedures, and unambiguous labels identifying signals has limited the possibilities for creating an automatic interpretation algorithm for the ECG signal. Known databases provided by PhysioNet, such as the MIT-BIH Arrhythmia Database and the PTB Diagnostic ECG Database, were deemed insufficient [10,11]. Data from single, small, or relatively homogeneous datasets, further limited by a small number of patients and rhythm episodes, prevented the creation of algorithms in machine learning models.

The work of the PhysioNet/Computing in Cardiology Challenge 2020 project to develop an automated ECG classifier provided an opportunity to address this problem by adding data from a wide variety of sources. Among these, there are numerous works, including the development of a comprehensive deep neural network model for the classification of up to 27 clinical diagnoses from the electrocardiogram. The authors of one of these achieved results, using the ResNet model, at the level of AUC = 0.967 and ACC = 0.43 in their study [12]. A similar approach was proposed [13], using the SE_ResNet model to improve the efficiency of the classification of various ECG abnormalities. Others, focusing on the comparative analysis of the recently published PTB-XL dataset, assessed the possibility of using convolutional neural networks, in particular those based on the ResNet and Inception architectures [14]. A different approach in the classification of cardiovascular diseases was demonstrated by the authors of a work [15] related to the detection of QRS complexes and T & P waves, together with the detection of their boundaries. The ECG classification algorithm was based on 19 classes. Features were extracted from the averaged QRS and from the intervals between the detected points.

The 12-lead ECG deep learning model found its reference mainly to ECG diagnosis in the automatic classification of cardiac arrhythmias. A deep learning model trained on a large ECG dataset was used with a deep neural network [16] based on 1D CNN for automatic multilabel arrhythmia classification with a score of ACC = 0.94 – 0.97. The authors of this study also conducted experiments on single-lead ECG with an analysis of the operation of every single lead. The subject of arrhythmia classification is also of interest to other authors [17], where, with the use of long-short term memory (LSTM), a model with an LSTM score of 0.6 was proposed. The choice of ECG for arrhythmia detection was undertaken by the authors of the paper [18], where they designed a computer-aided diagnosis system for the automatic diagnosis of four types of serious arrhythmias. In this approach, the ECG was analyzed using thirteen nonlinear features, known as entropy. The features extracted in this way were classified using ANOVA and subjected to automated classification using the K-nearest neighbor and decision tree classifiers. The obtained results were for KNN – ACC = 93.3% and DT – ACC = 96.3%. Various deep learning models for the examination of the ECG signal have also been proposed for atrial fibrillation, obtaining the result of ACC = 0.992 [19]. It is worth noting that the presented model successfully detected atrial fibrillation, and the tests were carried out with the use of various ECG signals. Attempts to investigate cardiac arrhythmias and cardiovascular diseases were also carried out in a new convolutional neural network [9] with a nonlocal convolutional block attention module (NCBAM), which focused on representative features along space, time, and channels. For the classification problem of ECG arrhythmia detection, the authors

obtained AUC = 0.93. The approach to convolutional neural networks, the possibilities and usability of tools, and the analysis of biomedical signals were also proposed by the authors of other papers [20]. The research included the implementation of a multilabel classification algorithm with the use of machine learning methods based on a CNN. The work described the details of the algorithm necessary for reconstruction and presented limitations and suggestions for improvement. A different approach to the ECG signal was presented by the authors of [21], where the focus was instead placed on processing the ECG signal, data sampling, feature extraction, and classification. They used a deep learning class model with gated recursive complex (GRU) and extreme learning machine (ELM) to recognize the ECG signal.

The aim of the study was to check the effectiveness of multiclass classification of ECG signals with the use of various neural network architectures. An additional aim was to test the effectiveness of very light nets for classification. A novelty in the article is the combination of a neural network with entropy-based features.

2. Materials and Methods

2.1. PTB-XL Dataset

In this article, data from the PTB-XL ECG database were used [11]. The PTB-XL database is a clinical ECG dataset of unprecedented size, with changes applied to evaluate machine learning algorithms. The PTB-XL ECG dataset contains 21,837 clinical 12-lead ECGs from 18,885 patients of 10 s in length, sampled at 500 Hz and 100 Hz with 16 bit resolution. Figure 1 shows examples of rhythms, consistent with the data contained in Table 1, which were used in the work. Among them there are examples of the following ECG signals: NORM—normal ECG, CD—myocardial infarction, STTC—ST/T change, MI—conduction disturbance, HYP—hypertrophy.

Table 1. The numbers of individual classes.

| Number of Records | Class | Description |
|-------------------|-------|------------------------|
| 7185 | NORM | Normal ECG |
| 3232 | CD | Myocardial Infarction |
| 3064 | STTC | ST/T Change |
| 2936 | MI | Conduction Disturbance |
| 815 | HYP | Hypertrophy |

The PTB-XL database is gender balanced. The data included were derived from 52% males and 48% females, ranging in age from 2 to 95 years (median 62). The data were enriched with additional information about the patient (age, sex, height, weight). Each ECG by the authors of the dataset was classified into one or more of 23 diagnostic subclasses in 5 diagnostic classes, or into classes that are not diagnostic classes. Each class was assigned a probability. Classes are marked according to the standard with the codes SCP_ECG.

The research methodology included classification studies carried out in 3 categories of binary classifications, where the classes were NORM (healthy patient) and all other classes (sick patient), where 5 diagnostic classes were used and where 20 diagnostic subclasses were used.

The research methodology was as follows (Figure 2): Data from the PTB-XL database were filtered and then divided into training, validation, and test groups. These data were then normalized and used as inputs for the neural networks that were examined. The network performed a classification. The signal class was obtained as an output, and this was then subjected to evaluation.

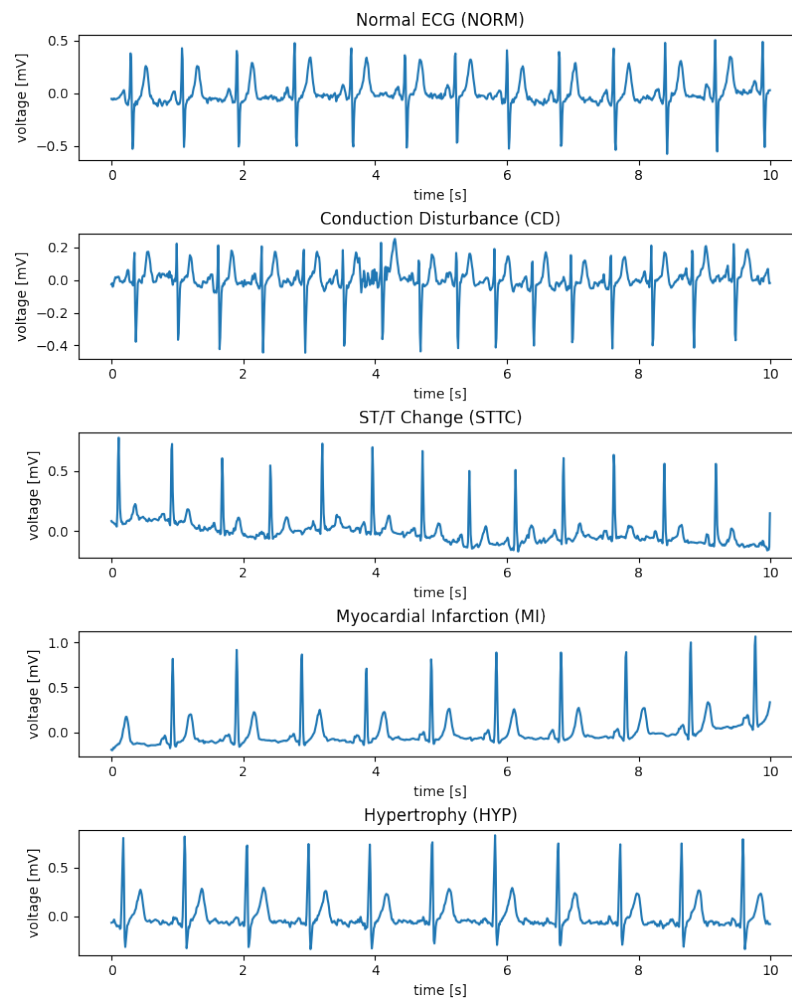


Figure 1. Examples of rhythm ECG signals.



Figure 2. General overview diagram of the method.

During the filtering stage, a set of 21,837 ECG records from the PTB-XL database was included in the simulation. ECGs not classified into diagnostic classes were filtered from the dataset. Subsequently, the ECGs in which the probability of classification was less than 100% were filtered out. In the next stage, ECGs were filtered out of those subclasses whose presence in the dataset was less than 20. A sampling frequency of 100 Hz was selected for the study, with 10 s as the length.

The dataset was divided into training, validation, and test sets in proportions of 70%, 15%, and 15%, respectively. The training set was used to train the network; the validation set was used to select the model; the test set was used to test the network's effectiveness.

As a result of the above activities, a total of 17,232 ECG records were used for the experimental analysis (Figure 3).

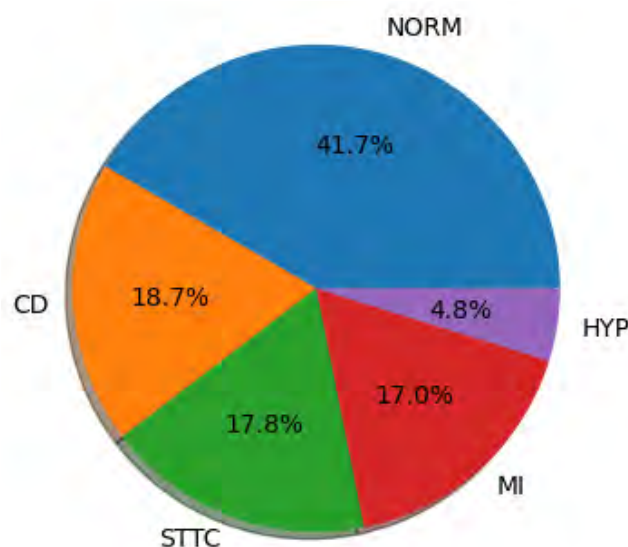


Figure 3. Diagnostic classes used in the study.

A detailed summary of the size of the individual classes used in the study and resulting from the above-described activities on the basis of PTB-XL is presented in Tables 1 and 2. The tables show the number of individual records used in the study, assigned to the appropriate diagnostic classes and subclasses defining cardiovascular diseases sorted by number of records.

Table 2. Numbers of individual subclasses.

| Number of Records | Subclass | Class | Description |
|-------------------|-----------|-------|---|
| 7185 | NORM | NORM | Normal ECG |
| 1713 | STTC | STTC | Non-diagnostic T abnormalities, suggests digitalis effect, long QT interval, ST-T changes compatible with ventricular aneurysm, compatible with electrolyte abnormalities |
| 1636 | AMI | MI | Anterior myocardial infarction, anterolateral myocardial infarction, in anteroseptal leads, in anterolateral leads, in lateral leads |
| 1272 | IMI | MI | Inferior myocardial infarction, inferolateral myocardial infarction, inferoposterolateral myocardial infarction, inferoposterior myocardial infarction, in inferior leads, in inferolateral leads |
| 881 | LAFB/LPFB | CD | Left anterior fascicular block, left posterior fascicular block |
| 798 | IRBBB | CD | Incomplete right bundle branch block |
| 733 | LVH | HYP | Left ventricular hypertrophy |
| 527 | CLBBB | CD | (Complete) left bundle branch block |
| 478 | NST_ | STTC | Nonspecific ST changes |
| 429 | ISCA | STTC | In anterolateral leads, in anteroseptal leads, in lateral leads, in anterior leads |
| 385 | CRBBB | CD | (Complete) right bundle branch block |
| 326 | IVCD | CD | Nonspecific intraventricular conduction disturbance |
| 297 | ISC_ | STTC | Ischemic ST-T changes |
| 204 | _AVB | CD | First-degree AV block, second-degree AV block, third-degree AV block |
| 147 | ISCI | STTC | In inferior leads, in inferolateral leads |
| 67 | WPW | CD | Wolff–Parkinson–White syndrome |
| 49 | LAO/LAE | HYP | Left atrial overload/enlargement |
| 44 | ILBBB | CD | Incomplete left bundle branch block |
| 33 | RAO/RAE | HYP | Right atrial overload/enlargement |
| 28 | LMI | MI | Lateral myocardial infarction |

2.2. Designed Network Architectures

This research compared three neural networks (convolutional network, SincNet, convolutional network with entropy features) in terms of the correct classification of the ECG signal. The research consisted of the implementation and testing of the proposed models of the neural networks. Cross-entropy loss as a loss function was applied to all networks.

The artificial neural networks proposed in this article were based on layers performing one-dimensional convolutions. This is a state-of-the-art solution in signal processing using deep learning due to its ability to extract features based on changes in consecutive samples, while simultaneously being faster and easier to train than recurrent layers such as LSTMs. The convolutional networks described in this article also contain residual connections between convolutional layers as described in [22]. These shortcut connections eliminate the so-called vanishing gradient problem and increase the capacity of models for better representation learning.

The networks were trained using the Adam optimizer as described in [23]. The optimizer trained the neural network using mini-batches of 128 examples in one pass. The learning rate was set at 0.001 at the beginning of the training and was later adjusted to 0.0001 to perform final corrections before ending the training. To prevent overfitting, early stopping was employed as described in [24]. The training of the neural network was stopped as soon as the network was unable to obtain better results on the validation dataset. This was to prevent overfitting. Following testing, the neural network was trained on the test dataset.

The tests were carried out using hardware configurations on a dual-Intel Xeon Silver 4210R, 192 GB RAM, and Nvidia Tesla A100 GPU. In this research, PyTorch and Jupyter Lab programming solutions were used for the implementation of the neural networks.

2.2.1. Convolutional Network

The first network examined is presented in Figure 4. It consists of five layers of one-dimensional convolutions with LeakyReLU activation functions and one fully connected layer with a softmax activation function. The network accepts ECG signals consisting of 12 channels containing 1000 samples each as inputs and outputs a class distribution vector normalized by application of the softmax function. The network determines the class to which an input signal belongs by determining the index of the vector maximum value. The class represented by this index is considered as a class of the input signal.

LeakyReLU was used instead of basic ReLU to preserve gradient loss in neurons outputting negative values. The coefficient describing a negative slope was set to 0.01; thus, the activation function can be described by the equation below:

$$f(x) = \begin{cases} 0.01x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases} \quad (1)$$

This configuration was used in every network proposed in this article.

This architecture was tested on both the normalized signal taken from the dataset without any other transformations and a spectrogram, and the results obtained from the former were better than from the latter. The network computing the spectrogram interpreted each spectrogram as a multichannel one-dimensional signal. Each of the twelve signals' spectrograms was processed by five one-dimensional convolutional blocks with the LeakyReLU activation function. The results of the convolutions were aggregated by performing adaptive average pooling. Afterwards, the results of pooling were flattened to the format of a one-dimensional vector and processed by a fully connected layer with a softmax activation function, and the output was used as a vector describing the probability distribution of the input signals belonging to each of the defined classes.

This is a simplified architecture designed to achieve both better computation time and memory storage efficiency. This network design has only 6 layers and, depending on the number of classes in classification, has just 8882 weights for binary classification and

11,957 weights for detecting 5 different classes of signal. The last segment of the network is a fully connected layer, which has a number of neurons equal to the quantity of possible classes to which the signal may belong. As a result, the more granular the classification process is, the more neurons are required, which increases the number of total weights in the network. The addition of residual connections did not increase the performance of the network significantly, but enlarged the quantity of parameters and computational steps required to process the signal.

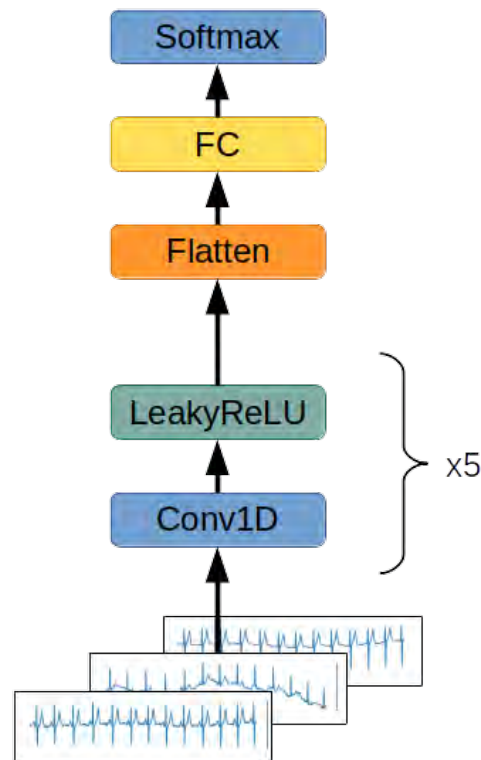


Figure 4. Convolutional network architecture. A twelve-channel ECG signal is passed through five subsequent one-dimensional convolutional layers with the LeakyReLU activation function. The results of the computation are flattened to the format of a one-dimensional vector. The results of the calculation are processed by a fully connected layer with a softmax activation function. The output value is a one-dimensional vector describing the probability distribution of the input signal belonging to each of the defined classes.

2.2.2. SincNet

The second examined network uses the SincNet layers described in [25]. SincNet layers are designed for the extraction of low-level features from a raw signal's data samples. SincNet layers train "wavelets" for feature extraction by performing convolution on the input signal:

$$y[n] = x[n] \cdot g[n, \theta] \quad (2)$$

where n is the index of the probe and θ are the parameters of the wavelets determined during training. The wavelet function g is described with the equation:

$$g[n, f_1, f_2] = 2f_2 \text{sinc}(2\pi f_2 n) - 2f_1 \text{sinc}(2\pi f_1 n) \quad (3)$$

where *sinc* function is defined as:

$$\text{sinc}(x) = \frac{\sin(x)}{x} \quad (4)$$

f_1 and f_2 are the cutoff frequencies determined by the SincNet layer during the training phase and form a set of trainable parameters θ :

$$\theta = \{(f_{i,1}, f_{i,2}) | i \in C^+ \cap i \leq l\} \quad (5)$$

where l is the number of wavelets in the SincNet layer.

The pair of filters (f_1, f_2) are initialized using the frequencies used for calculation of Mel-frequency cepstral coefficients [26].

SincNet layers are designed to interpret only the signal's singular channel at once, so the second network's architecture consists of a subnetwork using a SincNet layer, which encodes each signal's channel separately. The features extracted by the subnetwork are concatenated into one feature vector, which is fed to a block of fully connected layers. The softmax layer serves the role of the output classification layer, while the SincNet subnetwork consists of the SincNet layer adjusting the wavelets to the raw signal, two convolutional layers with LeakyReLU activation functions and layer normalizations, and three fully connected layers with batch normalization and LeakyReLU activation functions (Figure 5).

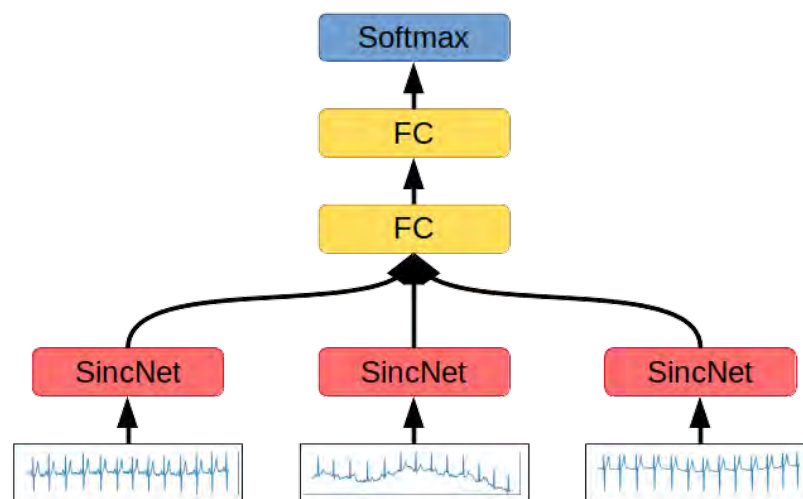


Figure 5. SincNet-based network architecture. Each channel of the 12-channel ECG signal is processed by a dedicated SincNet block. The results of each block are concatenated, flattened to the format of a one-dimensional vector, and used as an input for two subsequent fully connected layers, with LeakyReLU and softmax activation functions, respectively. The output value is a one-dimensional vector describing the probability distribution of the input signal belonging to each of the defined classes.

2.2.3. Convolutional Network with Entropy Features

The third network examined is presented in Figure 6. This network is an extended variant of the convolutional network. The network processes the ECG signal, and the values of the entropies are calculated for every channel of the signal. These entropies are:

- Shannon entropy—the summation of the informativeness of every possible state in the signal by measuring its probability. As a result, Shannon entropy is the measurement of the spread of the data [27];
- Approximate entropy—the measurement of series regularity. It provides information on how much the ECG fluctuates and its predictability [28];
- Sample entropy—an improvement on approximate entropy due to the lack of the signal length's impact on the entropy computations [28];
- Permutation entropy—the measurement of the order relations between ECG samples. This quantifies how regular and deterministic the signal is [29];

- Spectral entropy—the quantification of the energy spread uniformness across the frequency spectrum [30];
- SVD entropy—the measurement of how possible the dimensionality reduction of time series matrix is through factorization using the eigenvector approach;
- Rényi entropy—the generalization of the Shannon entropy by introducing the fractal order of the subsequent informativeness of each signal's state [31];
- Tsallis entropy—the generalization of the Boltzmann–Gibbs entropy, able to detect long-term memory effects on the signal [32];
- Extropy—the measurement of the amount of uncertainty represented by the distribution of the values in the observed ECG signal [33].

Granelo-Belinchon et al., in their article [34], stated that the tools of information theory can be straightforwardly applied to any nonstationary time process when considering small chunks of data spanning a short enough time range, allowing a slow evolution of higher-order moments to be neglected. The augmented Dickey–Fuller test has been conducted on ten-second-long training chunks of signals to determine the momentary stationarity of ECG signals. It turned out that 89.5% of tested signals were deemed stationarity in this small period of time, allowing the use of entropy methods for their interpretation.

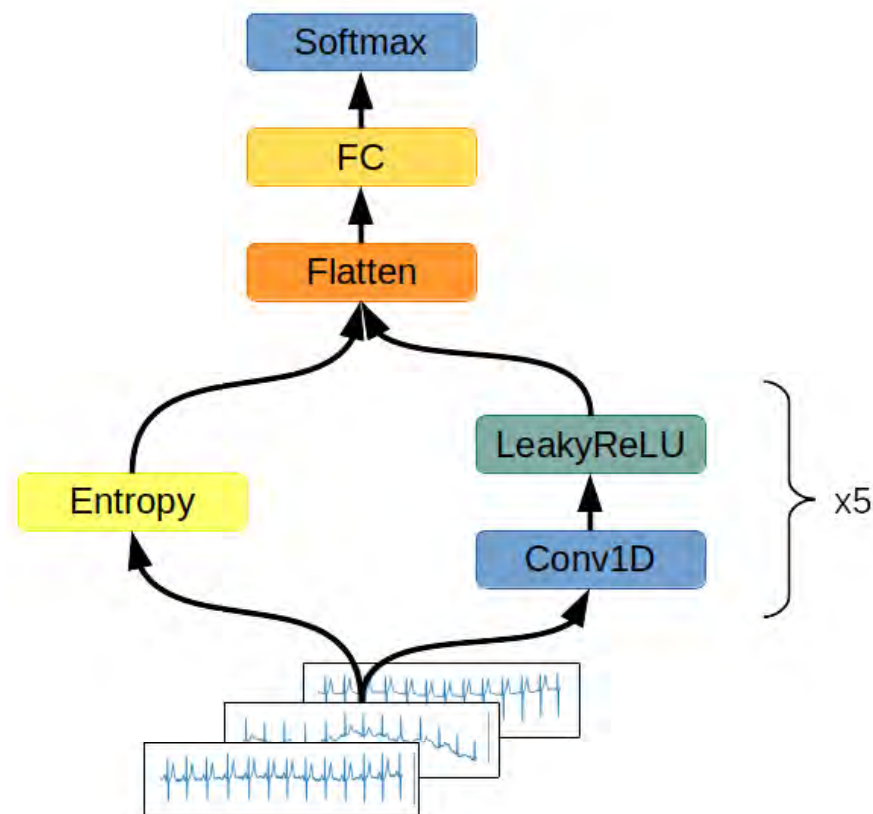


Figure 6. Convolutional network with the entropy features' block architecture. The computational graph of the network is made up of two branches. In the first branch, a twelve-channel ECG signal is passed through five subsequent one-dimensional convolutional layers with the LeakyReLU activation function. In the second branch, the input signal is used to compute the vector of entropies for every channel of the signal. The results of the computations from both branches are concatenated and flattened to the format of a one-dimensional vector. The results of the calculation are processed by a fully connected layer with softmax activation function. The output value is a one-dimensional vector describing the probability distribution of the input signal belonging to each of the defined classes.

The artificial neural network consists of two blocks: convolutional and fully connected. In the first step, a raw ECG signal is encoded by a convolutional block formed by five

one-dimensional convolutional layers with the LeakyReLU activation function. Each layer has a stride parameter equal to 2 to reduce the number of samples representing the time vector. Each layer also has a residual connection with the original, raw signal. Because of the signal's sample reduction due to the applied stride parameter, the ECG signal for each step of the residual connection is shrunk by average pooling with a window size of 2.

The encoded raw ECG signal is concatenated with the values of the entropies of every channel. Such a feature vector is fed to three fully connected layers with LeakyReLU activation functions in the first two and a softmax function in the last layer. The result of the softmax function is the output vector of the network and is used in order to classify the signal. For regularization purposes, there was a dropout with a chance of zeroing the input equal to 20% applied before each layer. The dropout was turned off during the network's evaluation.

2.3. Metrics

The neural networks were evaluated using the metrics described below. For the purpose of the simplicity of the equations, certain acronyms were created, as follows: TP—true positive, TN—true negative, FP—false positive, FN—false negative. The metrics used for the network evaluation are:

- Accuracy: $Acc = (TP + TN)/(TP + FP + TN + FN)$;
- Precision = $TP/(TP + FP)$;
- Recall = $TP/(TP + FN)$;
- $F1 = 2 * precision * recall / (precision + recall)$;
- AUC—area under the curve, ROC—area under the receiver operating characteristic curve. The ROC is a curve determined by calculating TFP = true positive rate = $TP/(TP + FN)$ and FPR = false positive rate = $FP/(TN + FP)$. The false positive rate describes the x-axis and the true positive rate the y-axis of a coordinate system. By changing the threshold value responsible for the classification of an example as belonging to either the positive or negative class, pairs of TFP-FPR are generated, resulting in the creation of the ROC curve. The AUC is a measurement of the area below the ROC curve;
- Total Params—number of neurons in the network. The smaller this number, the better, as less computation is required in order to perform classification.

3. Results

The results of the networks based on the convolutional network, SincNet, and the convolutional network with entropy features are summarized in Tables 3–5. With the recognition of two classes, the network based on the convolutional network achieved 88.2% ACC and with five classes 72.0% ACC. Similarly, the network based on SincNet achieved 85.8% ACC with the recognition of two classes and 73.0% with the recognition of five classes. The network based on the convolutional network with entropy features achieved 89.82% ACC with the recognition of two classes and 76.5% with the recognition of five classes. The network based on the convolutional network turned out to be slightly better than that based on SincNet. The situation changed with the recognition of 20 classes, where SincNet turned out to be slightly more effective. However, the network based on the convolutional network with entropy features turned out to be the best in all cases. It is worth noting that, depending on the number of recognized classes, the convolutional network had 200–600-times less weight than the SincNet-based network, which means it is much lighter. Adding entropy-based features to the convolutional network increases its weight two- to seven-fold. The convolutional neural network with entropy features achieved the highest accuracy in every classification task, scoring 89.2%, 76.5%, and 69.8% for 2, 5, and 20 classes, respectively. The basic convolutional network achieved better accuracy than SincNet during the classification of two classes (healthy/sick), but SincNet performed better on the classification of five and twenty classes. As described by Ravanelli et al. in [25], the neural network was designed to process the human voice without any

data preprocessing and did so successfully according to the authors. However, the results of its usage on ECG signals are far from ideal, as presented in Tables 3–5.

Table 3. The results of the convolutional network.

| Number of Classes | Acc | Avg Precision | Avg Recall | Avg F1 | Avg AUC | Total Params |
|-------------------|-------|---------------|------------|--------|---------|--------------|
| 2 | 0.882 | 0.879 | 0.882 | 0.88 | 0.953 | 8882 |
| 5 | 0.72 | 0.636 | 0.602 | 0.611 | 0.877 | 11,957 |
| 20 | 0.589 | 0.259 | 0.228 | 0.238 | 0.856 | 27,332 |

Table 4. The results of SincNet.

| Number of Classes | Acc | Avg Precision | Avg Recall | Avg F1 | Avg AUC | Total Params |
|-------------------|-------|---------------|------------|--------|---------|--------------|
| 2 | 0.858 | 0.855 | 0.854 | 0.855 | 0.93 | 6,109,922 |
| 5 | 0.73 | 0.666 | 0.589 | 0.6 | 0.884 | 6,109,922 |
| 20 | 0.593 | 0.287 | 0.269 | 0.262 | 0.807 | 6,269,204 |

Table 5. The results of the convolutional network with entropy features.

| Number of Classes | Acc | Avg Precision | Avg Recall | Avg F1 | Avg AUC | Total Params |
|-------------------|-------|---------------|------------|--------|---------|--------------|
| 2 | 0.892 | 0.889 | 0.893 | 0.891 | 0.96 | 58,178 |
| 5 | 0.765 | 0.714 | 0.662 | 0.68 | 0.910 | 58,259 |
| 20 | 0.698 | 0.355 | 0.339 | 0.332 | 0.815 | 58,664 |

Figures 7–15 show the confusion matrices of the results of the evaluated networks.

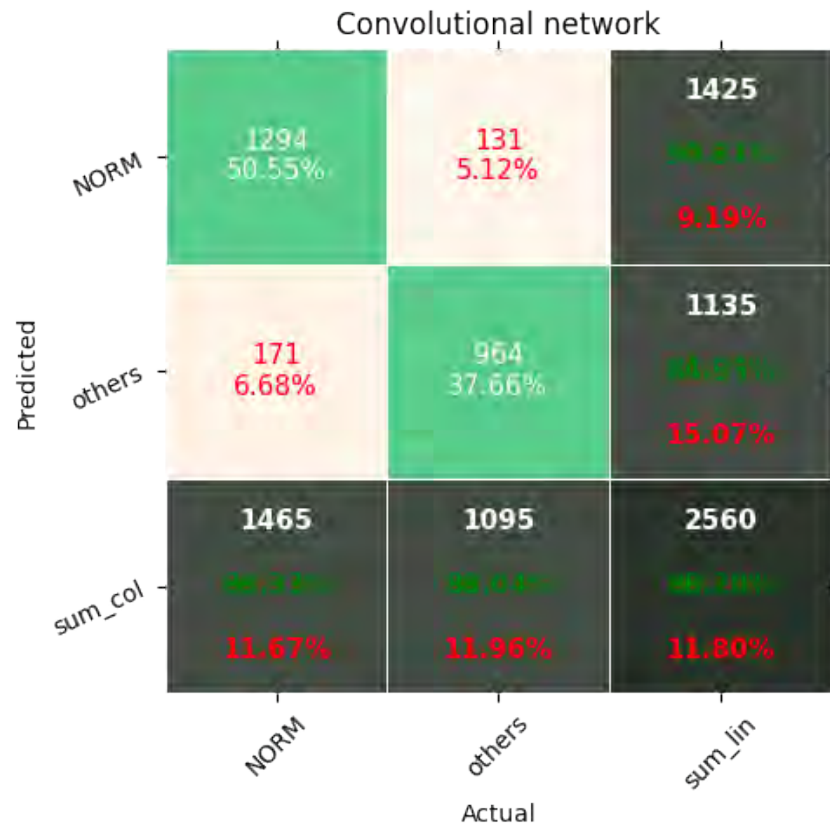


Figure 7. Confusion matrix of results for 2 classes for the convolutional network.

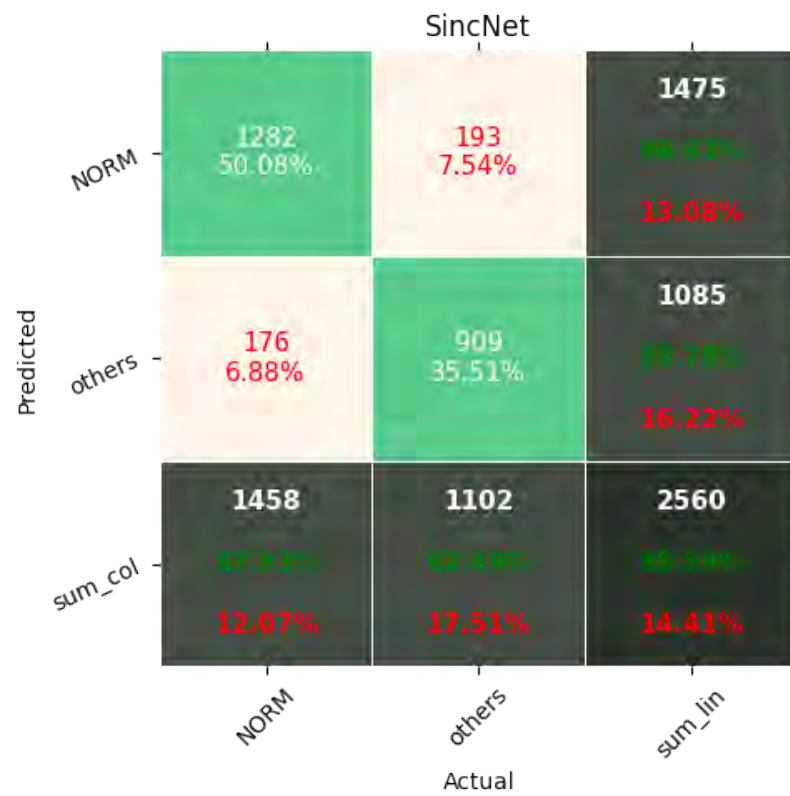


Figure 8. Confusion matrix of results for 2 classes for SincNet.

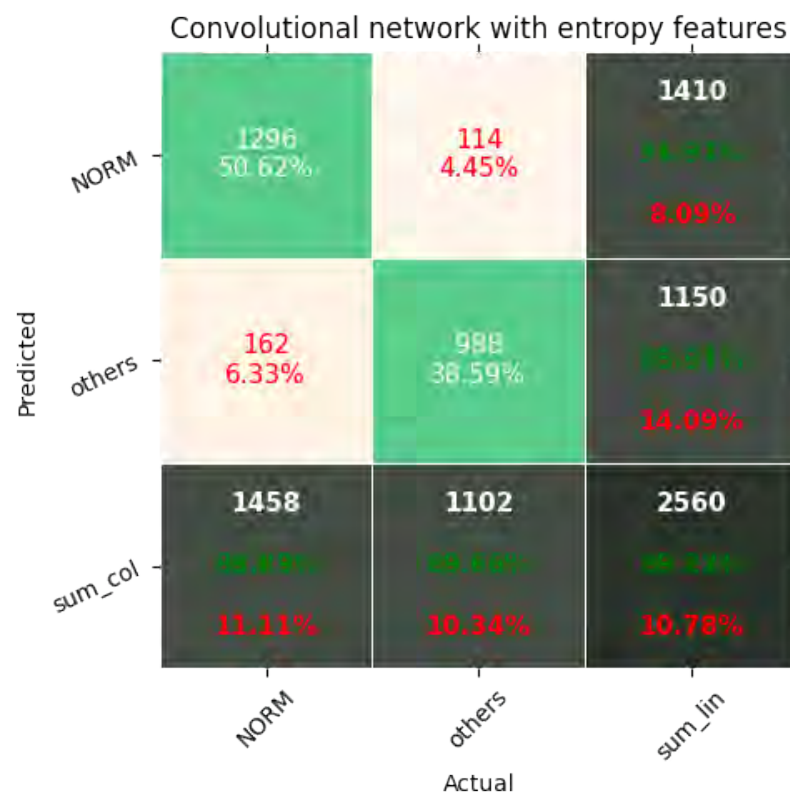


Figure 9. Confusion matrix of results for 2 classes for the convolutional network with entropy features.

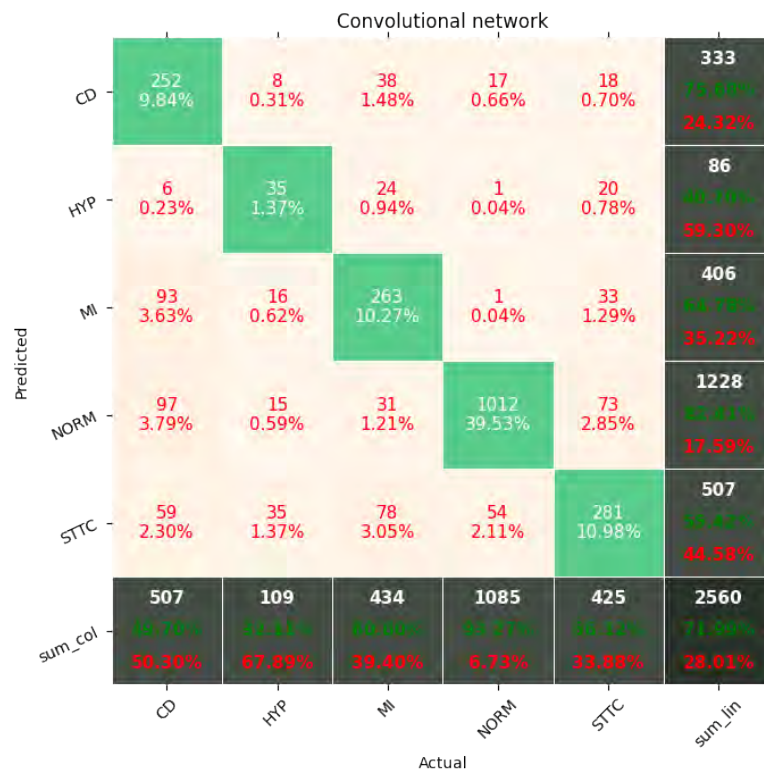


Figure 10. Confusion matrix of results for 5 classes for the convolutional network.

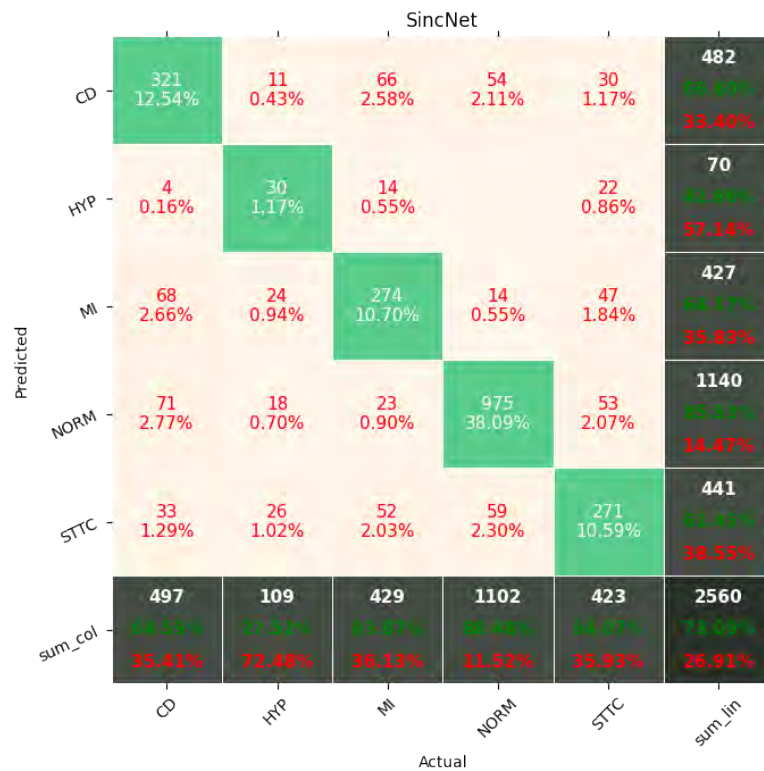


Figure 11. Confusion matrix of results for 5 classes for SincNet.

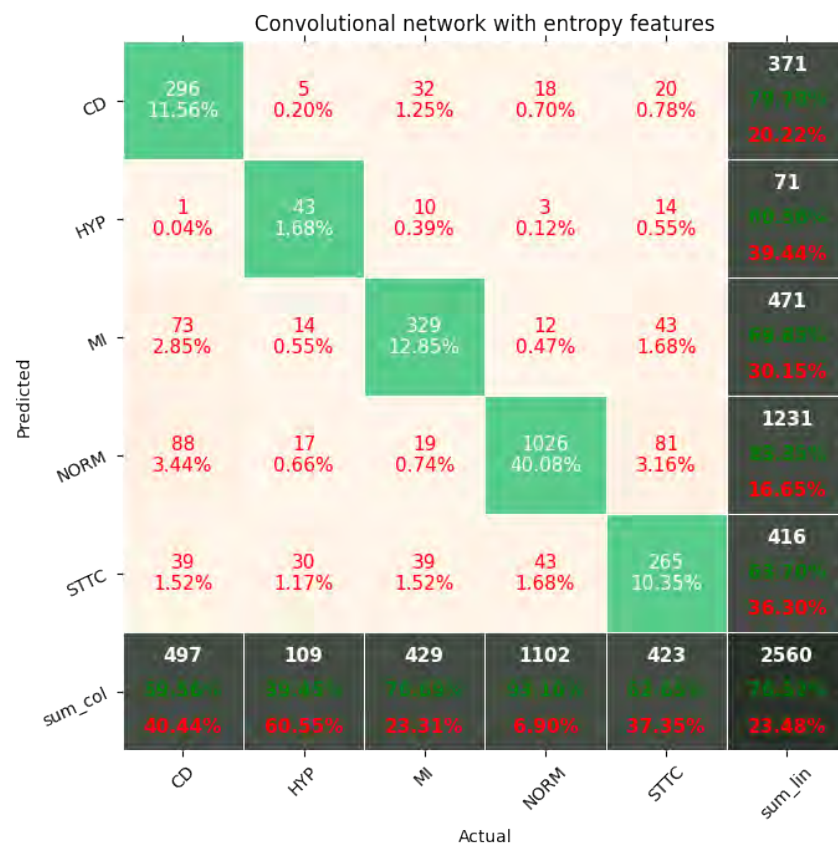


Figure 12. Confusion matrix of results for 5 classes for the convolutional network with entropy features.

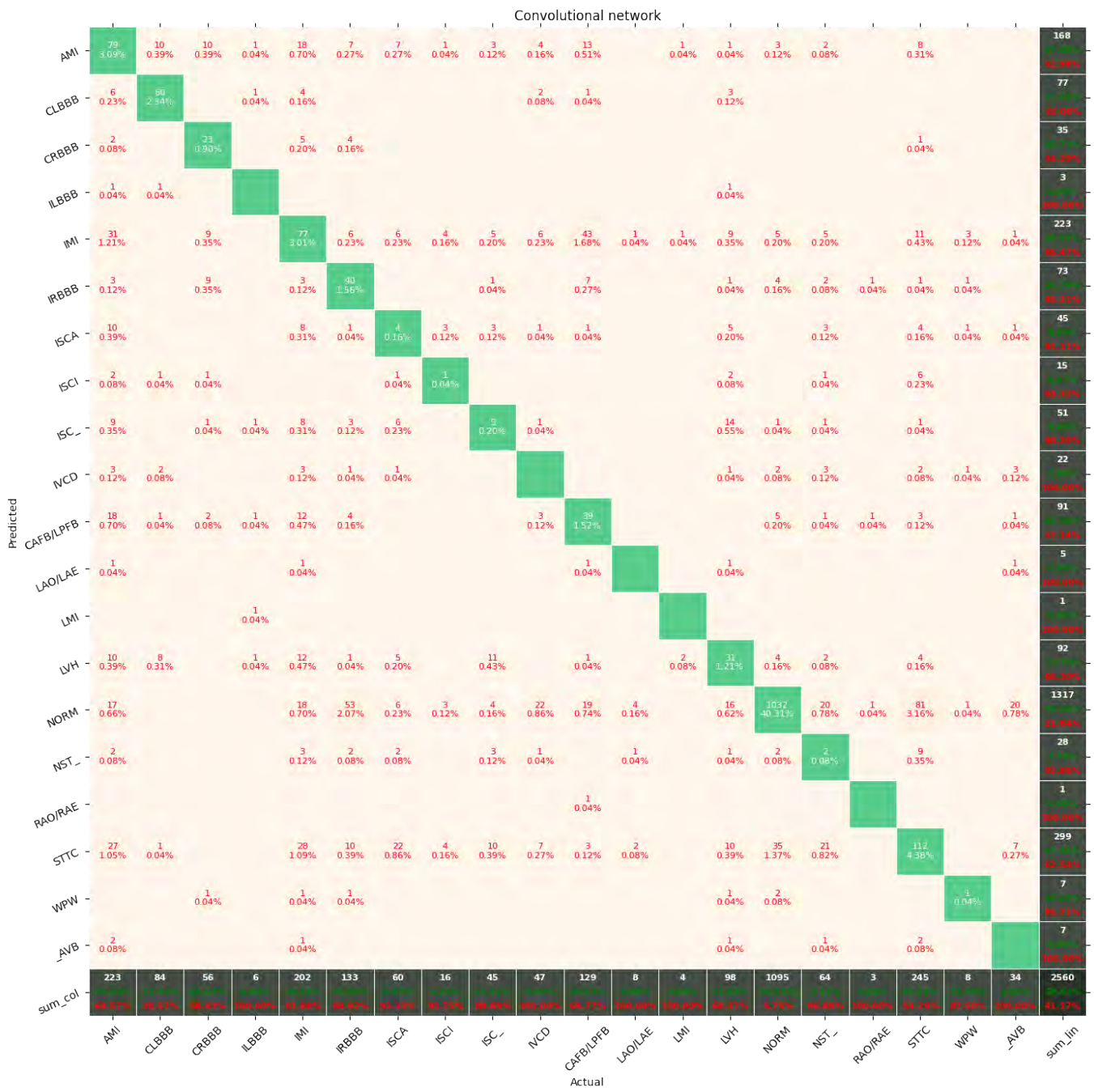


Figure 13. Confusion matrix of results for 5 classes for the convolutional network.

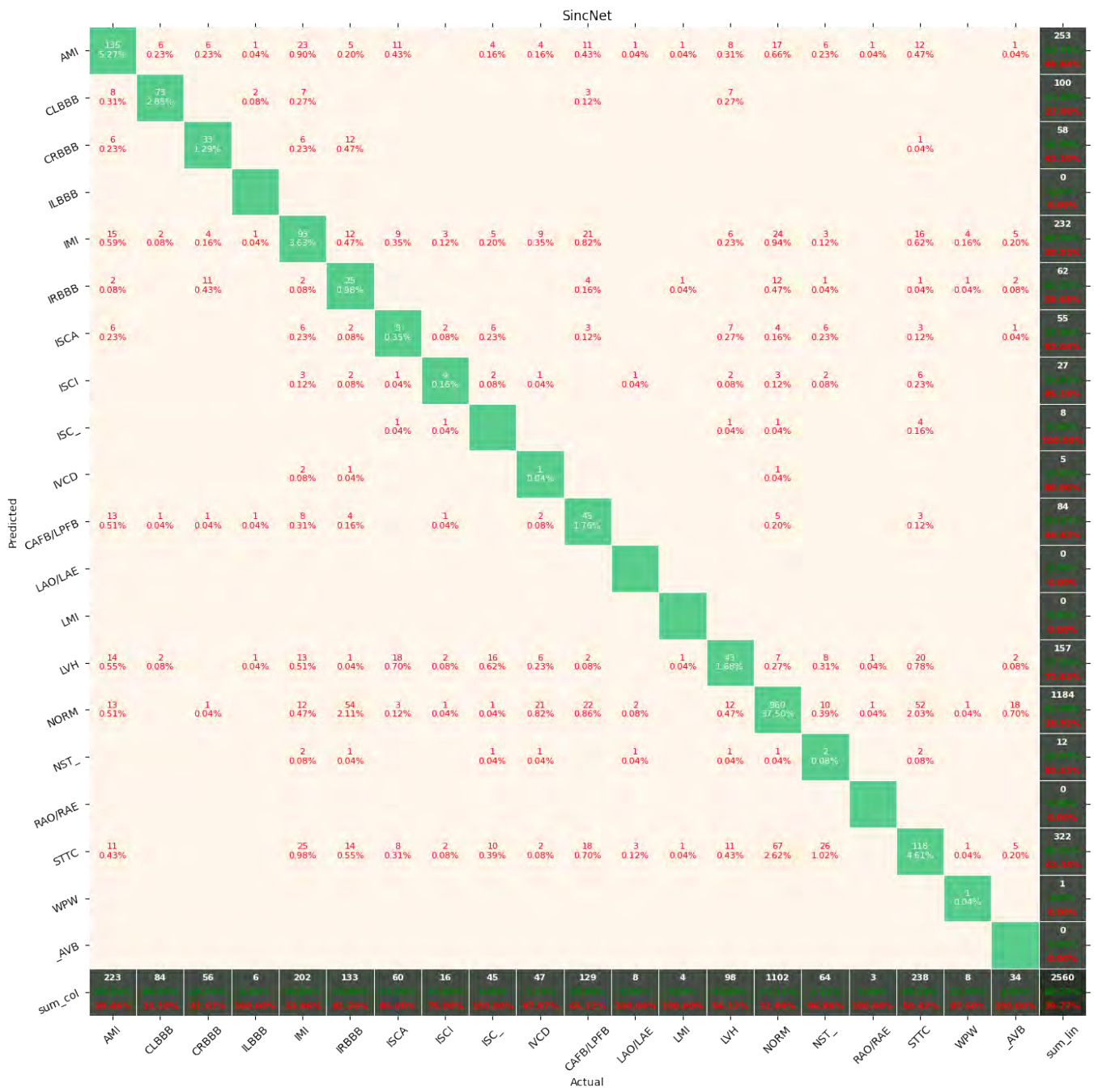


Figure 14. Confusion matrix of results for 5 classes for SincNet.

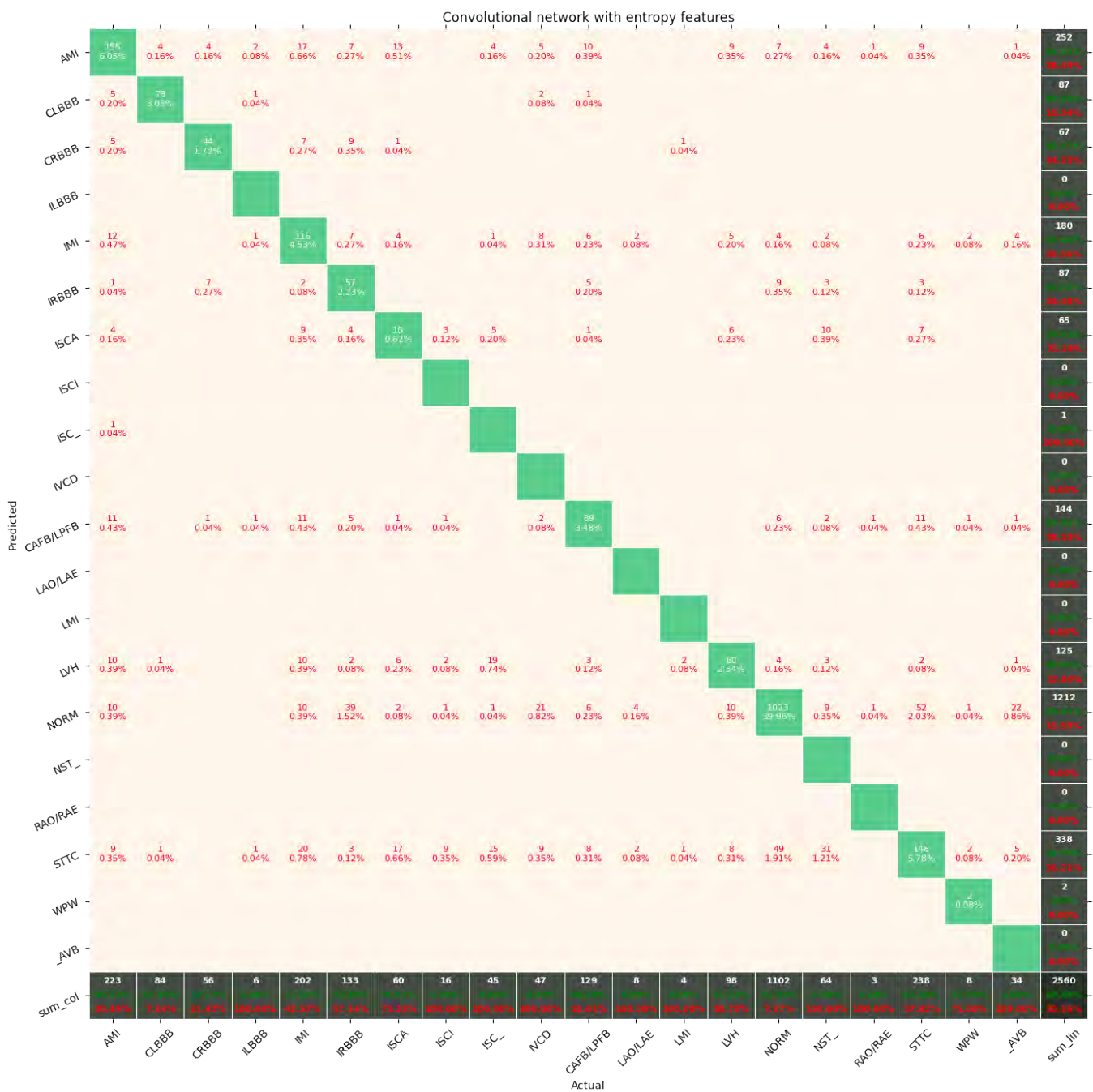


Figure 15. Confusion matrix of results for 5 classes for the convolutional network with entropy features.

In all cases of the evaluated networks, the NORM class obtained the highest value, which resulted from the large number of ECG recordings in this class.

4. Discussion

This paper presented a new model of convolutional neural networks, optimized to limit the computational and memory complexity for ECG recognition and classification of cardiovascular diseases. The research was carried out using a CNN network based on the convolutional network, which is relatively light and yields good results. The advantage of this approach is the possibility of using it on mobile and embedded devices, such as a Raspberry Pi or smartphone graphics cards.

The application of additional entropy-based features significantly improved the results. Such a solution also increased the weight of the network several times, however. As a

result, in applications where a very light network is needed, a compromise between weight and accuracy should be sought.

SincNet is a promising solution, but due to being designed to work with the human voice, it does not cope well with ECG signals in its original format. This results from the use of a set of initialization frequencies used in the computation of the Mel-frequency cepstral coefficients that are adapted to the spectral characteristics of the human voice. In the future, it would be worth considering the possibility of adapting SincNet to work with ECG.

The authors were unable to obtain better results due to the issue of overfitting on the training dataset. It was presumed that the addition of customized features may further boost the performance. The authors plan to investigate this claim in their next work.

Sampling determines the amount of measurements used to describe the signal. By changing the sampling, the signal is described by either more or fewer samples, whereas a stack of convolutional layers processes a fixed number of measurements in one context window. As a result, through a modification of the signal sampling, the network may either come to focus on more global features by reducing the amount of samples describing the signal or increase its attention to the details by increasing the measurements per signal.

Interpreting signals with different samplings may prove beneficial. In this work, we used only signals encoding 10 s of experiment on 1000 samples. It may well be the case that a network simultaneously interpreting a signal sampled with frequencies of 500 samples per second, 100 samples per second, and 50 samples per second will return better results. This is because signals sampled at lower frequencies can have entire ECG waves interpreted by one convolutional block, while signals sampled more frequently provide more detailed series for the extraction of features encoded by a small part of an ECG wave.

The proposed network based on a convolutional network is relatively uncomplicated. It is likely that better results could be obtained with the use of Inception models. This model uses heterogeneous subnets to improve the result. It is comparable to the case of wavelet transform, which may prove to be more advantageous than the use of fast Fourier transform. According to the authors, the proposed solution could be used in small devices for continuous monitoring of ECG signals, for example to alert about anomalies and make an initial diagnosis or support a doctor in this.

The authors assumed that a network's performance may be improved with a manageable cost increase by expanding its architecture with Inception-style heterogeneous subnetworks with varying kernels and poolings. The authors intend to investigate this assumption in their future work.

The authors further assumed that the integration of SincNet layers for low-level feature extraction in the first step of signal processing with the successful implementation of the first network based on convolutional layers may prove a benefit. The authors intend to investigate this assumption in their future work.

5. Conclusions

This study presented the capability of convolutional neural networks in the classification of heart diseases by the examination of ECG signals. The network proposed by the authors is both accurate and efficient as it is lightweight, allowing it to be computed on nonspecialized devices. The application of entropy-based features proved beneficial due to the improvements in the accuracy of heart disease classification. Entropy-based features are promising additions to data preprocessing that may prove beneficial in other signal-processing-related tasks.

Author Contributions: Conceptualization, S.Š., K.P. and D.L.; methodology, S.Š., K.P. and D.L.; software, S.Š., K.P., and D.L.; validation, S.Š., K.P. and D.L.; formal analysis, S.Š., K.P., and D.L.; investigation, S.Š., K.P. and D.L.; resources, S.Š., K.P., and D.L.; data curation, S.Š., K.P. and D.L.; writing—original draft preparation, S.Š., K.P., and D.L.; writing—review and editing, S.Š., K.P. and D.L.; visualization, S.Š., K.P. and D.L. All authors read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Benjamin, E.J.; Virani, S.S.; Callaway, C.W.; Chamberlain, A.M.; Chang, A.R.; Cheng, S.; Chiuve, S.E.; Cushman, M.; Delling, F.N.; Deo, R.; et al. Heart disease and stroke statistics—2018 update: A report from the American Heart Association. *Circulation* **2018**, *137*, e67–e492. [[CrossRef](#)] [[PubMed](#)]
2. Gupta, D.; Bajpai, B.; Dhiman, G.; Soni, M.; Gomathi, S.; Mane, D. Review of ECG arrhythmia classification using deep neural network. *Mater. Today Proc.* **2021**, In Press. [[CrossRef](#)]
3. World Health Organization. *Global Status Report on Noncommunicable Diseases*; WHO: Geneva, Switzerland, 2014.
4. Bogun, F.; Anh, D.; Kalahasty, G.; Wissner, E.; Serhal, C.B.; Bazzi, R.; Weaver, W.D.; Schuger, C. Misdiagnosis of atrial fibrillation and its clinical consequences. *Am. J. Med.* **2004**, *117*, 636–642. [[CrossRef](#)] [[PubMed](#)]
5. Schläpfer, J.; Wellens, H.J. Computer-interpreted electrocardiograms: Benefits and limitations. *J. Am. Coll. Cardiol.* **2017**, *70*, 1183–1192. [[CrossRef](#)]
6. Houssein, E.H.; Kilany, M.; Hassanien, A.E. ECG signals classification: A review. *Int. J. Intell. Eng. Informatics* **2017**, *5*, 376–396. [[CrossRef](#)]
7. Jambukia, S.H.; Vipul, K.D.; Harshadkumar, B.P. Classification of ECG signals using machine learning techniques: A survey. In Proceedings of the 2015 International Conference on Advances in Computer Engineering and Applications, Ghaziabad, India, 19–20 March 2015.
8. Macfarlane, P.W.; Devine, B.; Clark, E. The university of Glasgow (Uni-G) ECG analysis program. In Proceedings of the Computers in Cardiology, Lyon, France, 25–28 September 2005.
9. Wang, J.; Qiao, X.; Liu, C.; Wang, X.; Liu, Y.; Yao, L.; Zhang, H. Automated ECG classification using a non-local convolutional block attention module. *Comput. Methods Programs Biomed.* **2021**, *203*, 106006. [[CrossRef](#)] [[PubMed](#)]
10. Goldberger, A.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E.; et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220. [[CrossRef](#)]
11. Wagner, P.; Strodthoff, N.; Boussejot, R.; Samek, W.; Schaeffter, T. PTB-XL, a large publicly available electrocardiography dataset (version 1.0.1). *Sci. Data* **2020**, *7*, 1–5. [[CrossRef](#)]
12. Jia, W.; Xu, X.; Xu, X.; Sun, Y.; Liu, X. Automatic Detection and Classification of 12-lead ECGs Using a Deep Neural Network. In Proceedings of the Computing in Cardiology, Rimini, Italy, 13–16 September 2020; pp. 1–4.
13. Zhu, Z.; Lan, X.; Zhao, T.; Guo, Y.; Kojodjojo, P.; Xu, Z.; Liu, Z.; Liu, S.; Wang, H.; Sun, X.; et al. Identification of 27 abnormalities from multi-lead ECG signals: An ensemble SE_ResNet framework with sign loss function. *Physiol. Meas.* **2021**, *42*, 065008. [[CrossRef](#)]
14. Strodthoff, N.; Wagner, P.; Schaeffter, T.; Samek, W. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *arXiv* **2020**, arXiv:2004.13701.
15. Smisek, R.; Nemcova, A.; Marsanova, L.; Smital, L.; Vitek, M.; Kozumplik, J. Cardiac Pathologies Detection and Classification in 12-lead ECG. In Proceedings of the Computing in Cardiology, Rimini, Italy, 13–16 September 2020; pp. 1–4.
16. Zhang, D.; Yang, S.; Yuan, X.; Zhang, P. Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram. *Iscience* **2021**, *4*, 102373. [[CrossRef](#)]
17. Warrick, P.A.; Lostonlen, V.; Eickenberg, M.; Andén, J.; Homsí, M.N. Arrhythmia Classification of 12-lead Electrocardiograms by Hybrid Scattering-LSTM Networks. In Proceedings of the Computing in Cardiology, Rimini, Italy, 13–16 September 2020; pp. 1–4.
18. Acharya, U.R.; Fujita, H.; Adam, M.; Lih, O.S.; Hong, T.J.; Sudarshan, V.K.; Koh, J.E. Automated characterization of arrhythmias using nonlinear features from tachycardia ECG beats. In Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, 9–12 October 2016.
19. Jo, Y.Y.; Cho, Y.; Lee, S.Y.; Kwon, J.M.; Kim, K.H.; Jeon, K.H.; Cho, S.; Park, J.; Oh, B.H. Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram. *Int. J. Cardiol.* **2021**, *328*, 104–110. [[CrossRef](#)] [[PubMed](#)]
20. Lepek, M.; Pater, A.; Muter, K.; Wiszniewski, P.; Kokosińska, D.; Salamon, J.; Puzio, Z. 12-lead ECG Arrhythmia Classification Using Convolutional Neural Network for Mutually Non-Exclusive Classes. In Proceedings of the Computing in Cardiology, Rimini, Italy, 13–16 September 2020; pp. 1–4.
21. Ramaraj, E.; Virgeniya, S.C. A Novel Deep Learning based Gated Recurrent Unit with Extreme Learning Machine for Electrocardiogram (ECG) Signal Recognition. *Biomed. Signal Process. Control* **2021**, *68*, 102779.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

23. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
24. Caruana, R.; Lawrence, S.; Giles, L. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In Proceedings of the 14th Annual Neural Information Processing Systems Conference, Denver, CO, USA, 27 November–2 December 2020, pp. 402–408.
25. Ravanelli, M.; Yoshua, B. Speaker recognition from raw waveform with sincnet. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018.
26. Molau, S.; Pitz, M.; Schluter, R.; Ney, H. Computing Mel-frequency cepstral coefficients on the power spectrum. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001.
27. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [[CrossRef](#)]
28. Richman, J.S.; Moorman, J.R. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol.-Heart Circ. Physiol.* **2000**, *278*, H2039–H2049. [[CrossRef](#)]
29. Bandt, C.H.; Bernd, P. Permutation entropy: A natural complexity measure for time series. *Phys. Rev. Lett.* **2002**, *88*, 174102. [[CrossRef](#)]
30. Inouye, T.; Shinosaki, K.; Sakamoto, H.; Toi, S.; Ukai, S.; Iyama, A.; Katsuda, Y.; Hirano, M. Quantification of EEG irregularity by use of the entropy of the power spectrum. *Electroencephalogr. Clin. Neurophysiol.* **1991**, *79*, 204–210. [[CrossRef](#)]
31. Renyi, A. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; University of California Press: Oakland, CA, USA, 1961; pp. 547–561.
32. Bezerianos, A.; Tong, S.; Thakor, N. Time dependent entropy of EEG rhythm changes following brain ischemia. *Ann. Biomed. Eng.* **2003**, *31*, 221–232. [[CrossRef](#)]
33. Lad, F.; Sanfilippo, G.; Agrò, G. Extropy: A complementary dual of entropy. *arXiv* **2011**, arXiv:1109.6440.
34. Granero-Belinchón, C.; Roux, S.G.; Garnier, N.B. Information Theory for Non-Stationary Processes with Stationary Increments. *Entropy* **2019**, *21*, 1223. [[CrossRef](#)]

Article

ECG Signal Classification Using Deep Learning Techniques Based on the PTB-XL Dataset

Sandra Śmigiel ^{1,*}, Krzysztof Pałczyński ² and Damian Ledziński ²

¹ Faculty of Mechanical Engineering, UTP University of Science and Technology in Bydgoszcz, 85-796 Bydgoszcz, Poland

² Faculty of Telecommunications, Computer Science and Electrical Engineering, UTP University of Science and Technology in Bydgoszcz, 85-796 Bydgoszcz, Poland; krzysztof@palczynski.com.pl (K.P.); damian.ledzinski@utp.edu.pl (D.L.)

* Correspondence: sandra.smigiel@utp.edu.pl; Tel.: +48-52-340-8346

Abstract: The analysis and processing of ECG signals are a key approach in the diagnosis of cardiovascular diseases. The main field of work in this area is classification, which is increasingly supported by machine learning-based algorithms. In this work, a deep neural network was developed for the automatic classification of primary ECG signals. The research was carried out on the data contained in a PTB-XL database. Three neural network architectures were proposed: the first based on the convolutional network, the second on SincNet, and the third on the convolutional network, but with additional entropy-based features. The dataset was divided into training, validation, and test sets in proportions of 70%, 15%, and 15%, respectively. The studies were conducted for 2, 5, and 20 classes of disease entities. The convolutional network with entropy features obtained the best classification result. The convolutional network without entropy-based features obtained a slightly less successful result, but had the highest computational efficiency, due to the significantly lower number of neurons.

Keywords: ECG signal; classification; PTB-XL; deep learning



Citation: Śmigiel, S.; Pałczyński, K.; Ledziński, D. ECG Signal Classification Using Deep Learning Techniques Based on the PTB-XL Dataset. *Entropy* **2021**, *23*, 1121. <https://doi.org/10.3390/e23091121>

Academic Editor: Ernestina Menasalvas

Received: 5 July 2021

Accepted: 25 August 2021

Published: 28 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

According to publicly available reports, cardiovascular disease remains the leading cause of mortality worldwide [1]. One of the main causes of cardiovascular diseases is cardiac arrhythmia, in which the heartbeat deviates from typical beating patterns [2]. However, there are many types of irregular heartbeat. Accurate classification of heart disease types can aid in diagnosis and treatment [3].

An electrocardiogram (ECG) is a widely used, reliable, noninvasive approach to diagnosing cardiovascular disease. The standard ECG consists of 12 leads [4]. Traditionally, ECG results are manually interpreted by cardiologists based on a set of diagnosis criteria and experience. However, manual interpretation is time consuming and requires skill. Incorrectly interpreted ECG results may give rise to incorrect clinical decisions and lead to a threat to human life and health. With the rapid development of ECG and, at the same time, an insufficient number of cardiologists, the accurate and automatic diagnosis of ECG signals has become an interesting research topic for many scientists.

Over the past decade, numerous attempts have been made to identify a 12-lead clinical ECG, largely on the basis of the availability of large, public, open-source ECG data collections. Previous literature on ECG databases has shown a methodological division: signal processing and machine learning [5,6]. On the one hand, digital signal processing methods mainly include low- or high-pass filters, fast Fourier transform, and wavelet transform [7]. In this area, many algorithms are based on three processes: feature extraction, feature selection, and classification [8]. On the other hand, an alternative method is the application of machine learning methods. Such an application would primarily focus on

the automatic recognition of patterns that classify various disease entities, a method that is gaining greater importance in medical practice.

Algorithms known as deep neural networks have become particularly important in the last five years. Deep learning models have proven to be useful in increasing the effectiveness of diagnoses of cardiovascular diseases using ECG signals. By using the cascade of heterogeneous layers of neural networks to gradually extract increasingly high-level features, they lead to ever-improving neural networks built on their basis. Deep neural networks are reaching their zenith in various areas where artificial intelligence algorithms are applied.

In recent years, machine learning models have given rise to huge innovations in many areas, including image processing, natural language processing, computer games, and medical applications [9]. To date, however, the lack of adequate databases, well-defined assessment procedures, and unambiguous labels identifying signals has limited the possibilities for creating an automatic interpretation algorithm for the ECG signal. Known databases provided by PhysioNet, such as the MIT-BIH Arrhythmia Database and the PTB Diagnostic ECG Database, were deemed insufficient [10,11]. Data from single, small, or relatively homogeneous datasets, further limited by a small number of patients and rhythm episodes, prevented the creation of algorithms in machine learning models.

The work of the PhysioNet/Computing in Cardiology Challenge 2020 project to develop an automated ECG classifier provided an opportunity to address this problem by adding data from a wide variety of sources. Among these, there are numerous works, including the development of a comprehensive deep neural network model for the classification of up to 27 clinical diagnoses from the electrocardiogram. The authors of one of these achieved results, using the ResNet model, at the level of AUC = 0.967 and ACC = 0.43 in their study [12]. A similar approach was proposed [13], using the SE_ResNet model to improve the efficiency of the classification of various ECG abnormalities. Others, focusing on the comparative analysis of the recently published PTB-XL dataset, assessed the possibility of using convolutional neural networks, in particular those based on the ResNet and Inception architectures [14]. A different approach in the classification of cardiovascular diseases was demonstrated by the authors of a work [15] related to the detection of QRS complexes and T & P waves, together with the detection of their boundaries. The ECG classification algorithm was based on 19 classes. Features were extracted from the averaged QRS and from the intervals between the detected points.

The 12-lead ECG deep learning model found its reference mainly to ECG diagnosis in the automatic classification of cardiac arrhythmias. A deep learning model trained on a large ECG dataset was used with a deep neural network [16] based on 1D CNN for automatic multilabel arrhythmia classification with a score of ACC = 0.94 – 0.97. The authors of this study also conducted experiments on single-lead ECG with an analysis of the operation of every single lead. The subject of arrhythmia classification is also of interest to other authors [17], where, with the use of long-short term memory (LSTM), a model with an LSTM score of 0.6 was proposed. The choice of ECG for arrhythmia detection was undertaken by the authors of the paper [18], where they designed a computer-aided diagnosis system for the automatic diagnosis of four types of serious arrhythmias. In this approach, the ECG was analyzed using thirteen nonlinear features, known as entropy. The features extracted in this way were classified using ANOVA and subjected to automated classification using the K-nearest neighbor and decision tree classifiers. The obtained results were for KNN – ACC = 93.3% and DT – ACC = 96.3%. Various deep learning models for the examination of the ECG signal have also been proposed for atrial fibrillation, obtaining the result of ACC = 0.992 [19]. It is worth noting that the presented model successfully detected atrial fibrillation, and the tests were carried out with the use of various ECG signals. Attempts to investigate cardiac arrhythmias and cardiovascular diseases were also carried out in a new convolutional neural network [9] with a nonlocal convolutional block attention module (NCBAM), which focused on representative features along space, time, and channels. For the classification problem of ECG arrhythmia detection, the authors

obtained AUC = 0.93. The approach to convolutional neural networks, the possibilities and usability of tools, and the analysis of biomedical signals were also proposed by the authors of other papers [20]. The research included the implementation of a multilabel classification algorithm with the use of machine learning methods based on a CNN. The work described the details of the algorithm necessary for reconstruction and presented limitations and suggestions for improvement. A different approach to the ECG signal was presented by the authors of [21], where the focus was instead placed on processing the ECG signal, data sampling, feature extraction, and classification. They used a deep learning class model with gated recursive complex (GRU) and extreme learning machine (ELM) to recognize the ECG signal.

The aim of the study was to check the effectiveness of multiclass classification of ECG signals with the use of various neural network architectures. An additional aim was to test the effectiveness of very light nets for classification. A novelty in the article is the combination of a neural network with entropy-based features.

2. Materials and Methods

2.1. PTB-XL Dataset

In this article, data from the PTB-XL ECG database were used [11]. The PTB-XL database is a clinical ECG dataset of unprecedented size, with changes applied to evaluate machine learning algorithms. The PTB-XL ECG dataset contains 21,837 clinical 12-lead ECGs from 18,885 patients of 10 s in length, sampled at 500 Hz and 100 Hz with 16 bit resolution. Figure 1 shows examples of rhythms, consistent with the data contained in Table 1, which were used in the work. Among them there are examples of the following ECG signals: NORM—normal ECG, CD—myocardial infarction, STTC—ST/T change, MI—conduction disturbance, HYP—hypertrophy.

Table 1. The numbers of individual classes.

| Number of Records | Class | Description |
|-------------------|-------|------------------------|
| 7185 | NORM | Normal ECG |
| 3232 | CD | Myocardial Infarction |
| 3064 | STTC | ST/T Change |
| 2936 | MI | Conduction Disturbance |
| 815 | HYP | Hypertrophy |

The PTB-XL database is gender balanced. The data included were derived from 52% males and 48% females, ranging in age from 2 to 95 years (median 62). The data were enriched with additional information about the patient (age, sex, height, weight). Each ECG by the authors of the dataset was classified into one or more of 23 diagnostic subclasses in 5 diagnostic classes, or into classes that are not diagnostic classes. Each class was assigned a probability. Classes are marked according to the standard with the codes SCP_ECG.

The research methodology included classification studies carried out in 3 categories of binary classifications, where the classes were NORM (healthy patient) and all other classes (sick patient), where 5 diagnostic classes were used and where 20 diagnostic subclasses were used.

The research methodology was as follows (Figure 2): Data from the PTB-XL database were filtered and then divided into training, validation, and test groups. These data were then normalized and used as inputs for the neural networks that were examined. The network performed a classification. The signal class was obtained as an output, and this was then subjected to evaluation.

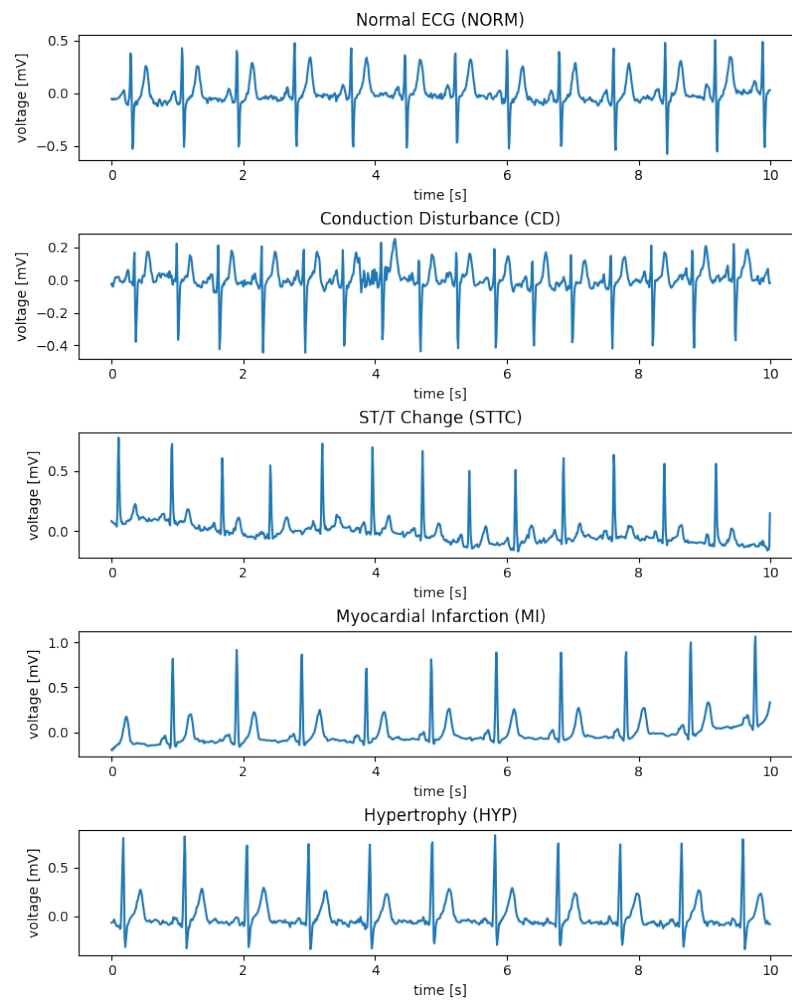


Figure 1. Examples of rhythm ECG signals.



Figure 2. General overview diagram of the method.

During the filtering stage, a set of 21,837 ECG records from the PTB-XL database was included in the simulation. ECGs not classified into diagnostic classes were filtered from the dataset. Subsequently, the ECGs in which the probability of classification was less than 100% were filtered out. In the next stage, ECGs were filtered out of those subclasses whose presence in the dataset was less than 20. A sampling frequency of 100 Hz was selected for the study, with 10 s as the length.

The dataset was divided into training, validation, and test sets in proportions of 70%, 15%, and 15%, respectively. The training set was used to train the network; the validation set was used to select the model; the test set was used to test the network's effectiveness.

As a result of the above activities, a total of 17,232 ECG records were used for the experimental analysis (Figure 3).

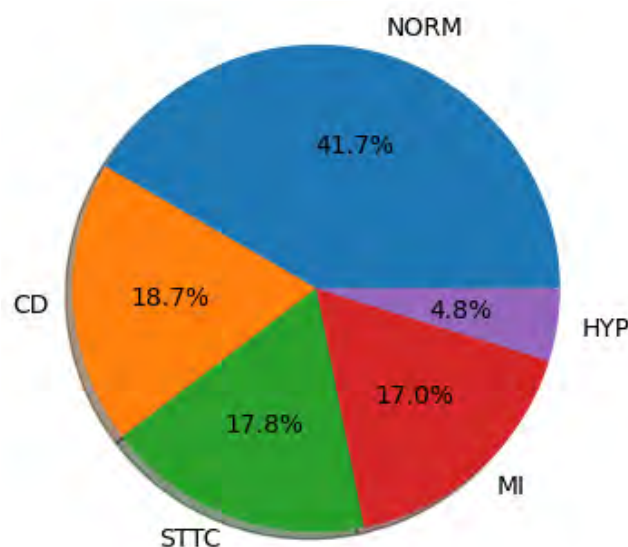


Figure 3. Diagnostic classes used in the study.

A detailed summary of the size of the individual classes used in the study and resulting from the above-described activities on the basis of PTB-XL is presented in Tables 1 and 2. The tables show the number of individual records used in the study, assigned to the appropriate diagnostic classes and subclasses defining cardiovascular diseases sorted by number of records.

Table 2. Numbers of individual subclasses.

| Number of Records | Subclass | Class | Description |
|-------------------|-----------|-------|---|
| 7185 | NORM | NORM | Normal ECG |
| 1713 | STTC | STTC | Non-diagnostic T abnormalities, suggests digitalis effect, long QT interval, ST-T changes compatible with ventricular aneurysm, compatible with electrolyte abnormalities |
| 1636 | AMI | MI | Anterior myocardial infarction, anterolateral myocardial infarction, in anteroseptal leads, in anterolateral leads, in lateral leads |
| 1272 | IMI | MI | Inferior myocardial infarction, inferolateral myocardial infarction, inferoposterolateral myocardial infarction, inferoposterior myocardial infarction, in inferior leads, in inferolateral leads |
| 881 | LAFB/LPFB | CD | Left anterior fascicular block, left posterior fascicular block |
| 798 | IRBBB | CD | Incomplete right bundle branch block |
| 733 | LVH | HYP | Left ventricular hypertrophy |
| 527 | CLBBB | CD | (Complete) left bundle branch block |
| 478 | NST_ | STTC | Nonspecific ST changes |
| 429 | ISCA | STTC | In anterolateral leads, in anteroseptal leads, in lateral leads, in anterior leads |
| 385 | CRBBB | CD | (Complete) right bundle branch block |
| 326 | IVCD | CD | Nonspecific intraventricular conduction disturbance |
| 297 | ISC_ | STTC | Ischemic ST-T changes |
| 204 | _AVB | CD | First-degree AV block, second-degree AV block, third-degree AV block |
| 147 | ISCI | STTC | In inferior leads, in inferolateral leads |
| 67 | WPW | CD | Wolff–Parkinson–White syndrome |
| 49 | LAO/LAE | HYP | Left atrial overload/enlargement |
| 44 | ILBBB | CD | Incomplete left bundle branch block |
| 33 | RAO/RAE | HYP | Right atrial overload/enlargement |
| 28 | LMI | MI | Lateral myocardial infarction |

2.2. Designed Network Architectures

This research compared three neural networks (convolutional network, SincNet, convolutional network with entropy features) in terms of the correct classification of the ECG signal. The research consisted of the implementation and testing of the proposed models of the neural networks. Cross-entropy loss as a loss function was applied to all networks.

The artificial neural networks proposed in this article were based on layers performing one-dimensional convolutions. This is a state-of-the-art solution in signal processing using deep learning due to its ability to extract features based on changes in consecutive samples, while simultaneously being faster and easier to train than recurrent layers such as LSTMs. The convolutional networks described in this article also contain residual connections between convolutional layers as described in [22]. These shortcut connections eliminate the so-called vanishing gradient problem and increase the capacity of models for better representation learning.

The networks were trained using the Adam optimizer as described in [23]. The optimizer trained the neural network using mini-batches of 128 examples in one pass. The learning rate was set at 0.001 at the beginning of the training and was later adjusted to 0.0001 to perform final corrections before ending the training. To prevent overfitting, early stopping was employed as described in [24]. The training of the neural network was stopped as soon as the network was unable to obtain better results on the validation dataset. This was to prevent overfitting. Following testing, the neural network was trained on the test dataset.

The tests were carried out using hardware configurations on a dual-Intel Xeon Silver 4210R, 192 GB RAM, and Nvidia Tesla A100 GPU. In this research, PyTorch and Jupyter Lab programming solutions were used for the implementation of the neural networks.

2.2.1. Convolutional Network

The first network examined is presented in Figure 4. It consists of five layers of one-dimensional convolutions with LeakyReLU activation functions and one fully connected layer with a softmax activation function. The network accepts ECG signals consisting of 12 channels containing 1000 samples each as inputs and outputs a class distribution vector normalized by application of the softmax function. The network determines the class to which an input signal belongs by determining the index of the vector maximum value. The class represented by this index is considered as a class of the input signal.

LeakyReLU was used instead of basic ReLU to preserve gradient loss in neurons outputting negative values. The coefficient describing a negative slope was set to 0.01; thus, the activation function can be described by the equation below:

$$f(x) = \begin{cases} 0.01x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases} \quad (1)$$

This configuration was used in every network proposed in this article.

This architecture was tested on both the normalized signal taken from the dataset without any other transformations and a spectrogram, and the results obtained from the former were better than from the latter. The network computing the spectrogram interpreted each spectrogram as a multichannel one-dimensional signal. Each of the twelve signals' spectrograms was processed by five one-dimensional convolutional blocks with the LeakyReLU activation function. The results of the convolutions were aggregated by performing adaptive average pooling. Afterwards, the results of pooling were flattened to the format of a one-dimensional vector and processed by a fully connected layer with a softmax activation function, and the output was used as a vector describing the probability distribution of the input signals belonging to each of the defined classes.

This is a simplified architecture designed to achieve both better computation time and memory storage efficiency. This network design has only 6 layers and, depending on the number of classes in classification, has just 8882 weights for binary classification and

11,957 weights for detecting 5 different classes of signal. The last segment of the network is a fully connected layer, which has a number of neurons equal to the quantity of possible classes to which the signal may belong. As a result, the more granular the classification process is, the more neurons are required, which increases the number of total weights in the network. The addition of residual connections did not increase the performance of the network significantly, but enlarged the quantity of parameters and computational steps required to process the signal.

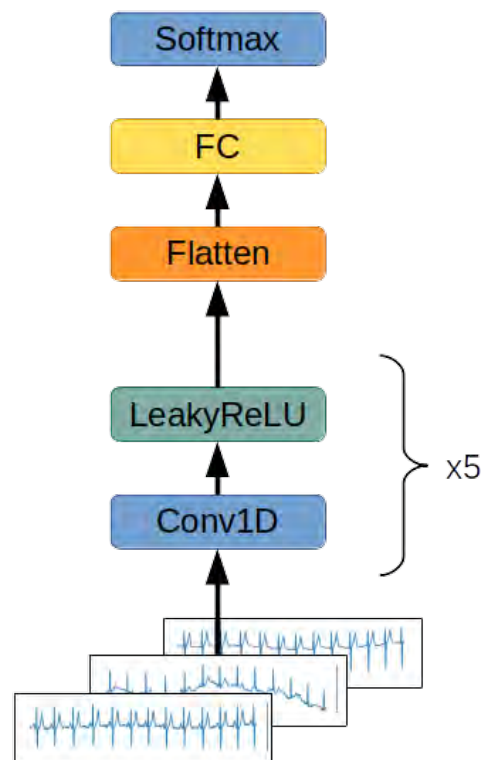


Figure 4. Convolutional network architecture. A twelve-channel ECG signal is passed through five subsequent one-dimensional convolutional layers with the LeakyReLU activation function. The results of the computation are flattened to the format of a one-dimensional vector. The results of the calculation are processed by a fully connected layer with a softmax activation function. The output value is a one-dimensional vector describing the probability distribution of the input signal belonging to each of the defined classes.

2.2.2. SincNet

The second examined network uses the SincNet layers described in [25]. SincNet layers are designed for the extraction of low-level features from a raw signal's data samples. SincNet layers train "wavelets" for feature extraction by performing convolution on the input signal:

$$y[n] = x[n] \cdot g[n, \theta] \quad (2)$$

where n is the index of the probe and θ are the parameters of the wavelets determined during training. The wavelet function g is described with the equation:

$$g[n, f_1, f_2] = 2f_2 \text{sinc}(2\pi f_2 n) - 2f_1 \text{sinc}(2\pi f_1 n) \quad (3)$$

where *sinc* function is defined as:

$$\text{sinc}(x) = \frac{\sin(x)}{x} \quad (4)$$

f_1 and f_2 are the cutoff frequencies determined by the SincNet layer during the training phase and form a set of trainable parameters θ :

$$\theta = \{(f_{i,1}, f_{i,2}) | i \in C^+ \cap i \leq l\} \quad (5)$$

where l is the number of wavelets in the SincNet layer.

The pair of filters (f_1, f_2) are initialized using the frequencies used for calculation of Mel-frequency cepstral coefficients [26].

SincNet layers are designed to interpret only the signal's singular channel at once, so the second network's architecture consists of a subnetwork using a SincNet layer, which encodes each signal's channel separately. The features extracted by the subnetwork are concatenated into one feature vector, which is fed to a block of fully connected layers. The softmax layer serves the role of the output classification layer, while the SincNet subnetwork consists of the SincNet layer adjusting the wavelets to the raw signal, two convolutional layers with LeakyReLU activation functions and layer normalizations, and three fully connected layers with batch normalization and LeakyReLU activation functions (Figure 5).

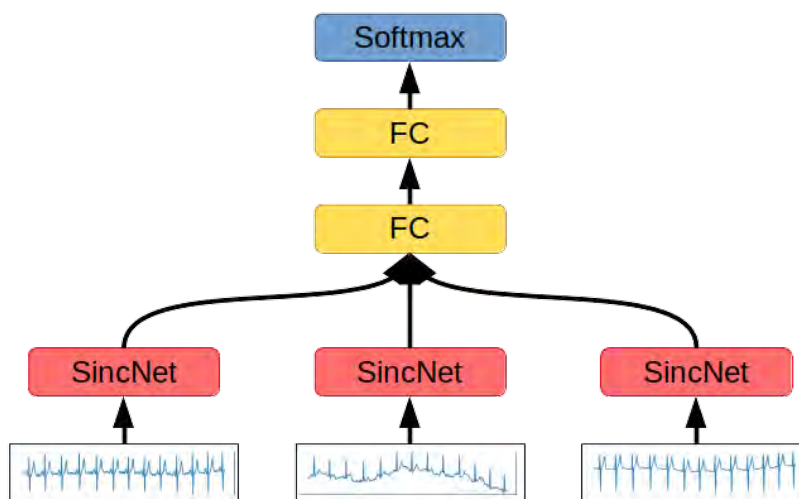


Figure 5. SincNet-based network architecture. Each channel of the 12-channel ECG signal is processed by a dedicated SincNet block. The results of each block are concatenated, flattened to the format of a one-dimensional vector, and used as an input for two subsequent fully connected layers, with LeakyReLU and softmax activation functions, respectively. The output value is a one-dimensional vector describing the probability distribution of the input signal belonging to each of the defined classes.

2.2.3. Convolutional Network with Entropy Features

The third network examined is presented in Figure 6. This network is an extended variant of the convolutional network. The network processes the ECG signal, and the values of the entropies are calculated for every channel of the signal. These entropies are:

- Shannon entropy—the summation of the informativeness of every possible state in the signal by measuring its probability. As a result, Shannon entropy is the measurement of the spread of the data [27];
- Approximate entropy—the measurement of series regularity. It provides information on how much the ECG fluctuates and its predictability [28];
- Sample entropy—an improvement on approximate entropy due to the lack of the signal length's impact on the entropy computations [28];
- Permutation entropy—the measurement of the order relations between ECG samples. This quantifies how regular and deterministic the signal is [29];

- Spectral entropy—the quantification of the energy spread uniformness across the frequency spectrum [30];
- SVD entropy—the measurement of how possible the dimensionality reduction of time series matrix is through factorization using the eigenvector approach;
- Rényi entropy—the generalization of the Shannon entropy by introducing the fractal order of the subsequent informativeness of each signal's state [31];
- Tsallis entropy—the generalization of the Boltzmann–Gibbs entropy, able to detect long-term memory effects on the signal [32];
- Extropy—the measurement of the amount of uncertainty represented by the distribution of the values in the observed ECG signal [33].

Granelo-Belinchon et al., in their article [34], stated that the tools of information theory can be straightforwardly applied to any nonstationary time process when considering small chunks of data spanning a short enough time range, allowing a slow evolution of higher-order moments to be neglected. The augmented Dickey–Fuller test has been conducted on ten-second-long training chunks of signals to determine the momentary stationarity of ECG signals. It turned out that 89.5% of tested signals were deemed stationarity in this small period of time, allowing the use of entropy methods for their interpretation.

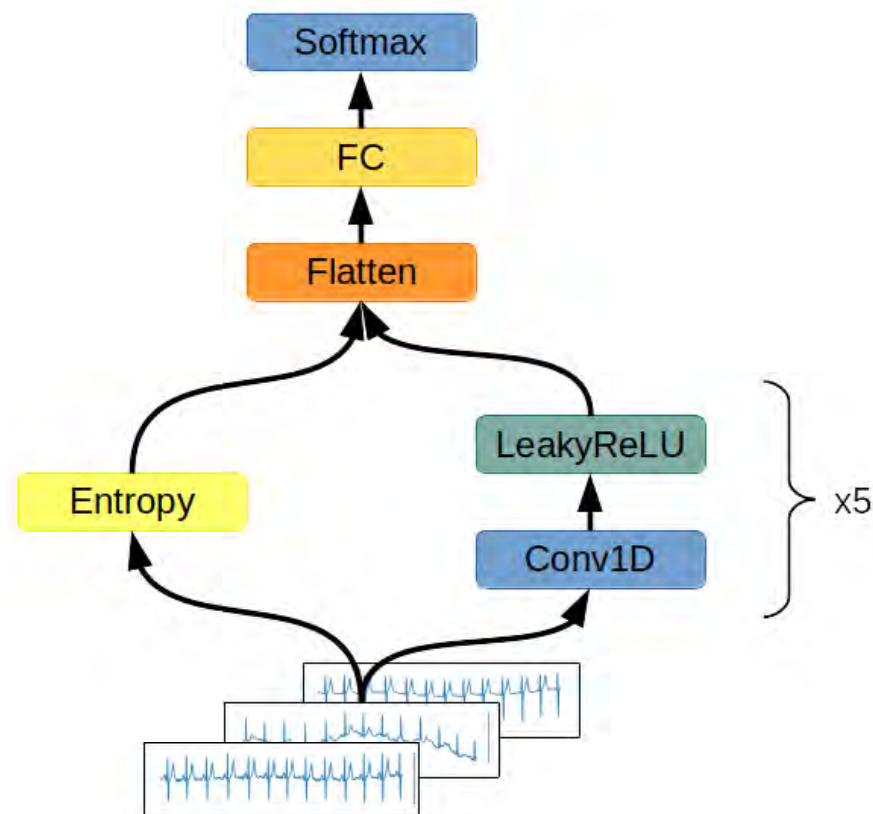


Figure 6. Convolutional network with the entropy features' block architecture. The computational graph of the network is made up of two branches. In the first branch, a twelve-channel ECG signal is passed through five subsequent one-dimensional convolutional layers with the LeakyReLU activation function. In the second branch, the input signal is used to compute the vector of entropies for every channel of the signal. The results of the computations from both branches are concatenated and flattened to the format of a one-dimensional vector. The results of the calculation are processed by a fully connected layer with softmax activation function. The output value is a one-dimensional vector describing the probability distribution of the input signal belonging to each of the defined classes.

The artificial neural network consists of two blocks: convolutional and fully connected. In the first step, a raw ECG signal is encoded by a convolutional block formed by five

one-dimensional convolutional layers with the LeakyReLU activation function. Each layer has a stride parameter equal to 2 to reduce the number of samples representing the time vector. Each layer also has a residual connection with the original, raw signal. Because of the signal's sample reduction due to the applied stride parameter, the ECG signal for each step of the residual connection is shrunk by average pooling with a window size of 2.

The encoded raw ECG signal is concatenated with the values of the entropies of every channel. Such a feature vector is fed to three fully connected layers with LeakyReLU activation functions in the first two and a softmax function in the last layer. The result of the softmax function is the output vector of the network and is used in order to classify the signal. For regularization purposes, there was a dropout with a chance of zeroing the input equal to 20% applied before each layer. The dropout was turned off during the network's evaluation.

2.3. Metrics

The neural networks were evaluated using the metrics described below. For the purpose of the simplicity of the equations, certain acronyms were created, as follows: TP—true positive, TN—true negative, FP—false positive, FN—false negative. The metrics used for the network evaluation are:

- Accuracy: $\text{Acc} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN})$;
- Precision = $\text{TP} / (\text{TP} + \text{FP})$;
- Recall = $\text{TP} / (\text{TP} + \text{FN})$;
- $\text{F1} = 2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$;
- AUC—area under the curve, ROC—area under the receiver operating characteristic curve. The ROC is a curve determined by calculating $\text{TFP} = \text{true positive rate} = \text{TP} / (\text{TP} + \text{FN})$ and $\text{FPR} = \text{false positive rate} = \text{FP} / (\text{TN} + \text{FP})$. The false positive rate describes the x-axis and the true positive rate the y-axis of a coordinate system. By changing the threshold value responsible for the classification of an example as belonging to either the positive or negative class, pairs of TFP-FPR are generated, resulting in the creation of the ROC curve. The AUC is a measurement of the area below the ROC curve;
- Total Params—number of neurons in the network. The smaller this number, the better, as less computation is required in order to perform classification.

3. Results

The results of the networks based on the convolutional network, SincNet, and the convolutional network with entropy features are summarized in Tables 3–5. With the recognition of two classes, the network based on the convolutional network achieved 88.2% ACC and with five classes 72.0% ACC. Similarly, the network based on SincNet achieved 85.8% ACC with the recognition of two classes and 73.0% with the recognition of five classes. The network based on the convolutional network with entropy features achieved 89.82% ACC with the recognition of two classes and 76.5% with the recognition of five classes. The network based on the convolutional network turned out to be slightly better than that based on SincNet. The situation changed with the recognition of 20 classes, where SincNet turned out to be slightly more effective. However, the network based on the convolutional network with entropy features turned out to be the best in all cases. It is worth noting that, depending on the number of recognized classes, the convolutional network had 200–600-times less weight than the SincNet-based network, which means it is much lighter. Adding entropy-based features to the convolutional network increases its weight two- to seven-fold. The convolutional neural network with entropy features achieved the highest accuracy in every classification task, scoring 89.2%, 76.5%, and 69.8% for 2, 5, and 20 classes, respectively. The basic convolutional network achieved better accuracy than SincNet during the classification of two classes (healthy/sick), but SincNet performed better on the classification of five and twenty classes. As described by Ravanelli et al. in [25], the neural network was designed to process the human voice without any

data preprocessing and did so successfully according to the authors. However, the results of its usage on ECG signals are far from ideal, as presented in Tables 3–5.

Table 3. The results of the convolutional network.

| Number of Classes | Acc | Avg Precision | Avg Recall | Avg F1 | Avg AUC | Total Params |
|-------------------|-------|---------------|------------|--------|---------|--------------|
| 2 | 0.882 | 0.879 | 0.882 | 0.88 | 0.953 | 8882 |
| 5 | 0.72 | 0.636 | 0.602 | 0.611 | 0.877 | 11,957 |
| 20 | 0.589 | 0.259 | 0.228 | 0.238 | 0.856 | 27,332 |

Table 4. The results of SincNet.

| Number of Classes | Acc | Avg Precision | Avg Recall | Avg F1 | Avg AUC | Total Params |
|-------------------|-------|---------------|------------|--------|---------|--------------|
| 2 | 0.858 | 0.855 | 0.854 | 0.855 | 0.93 | 6,109,922 |
| 5 | 0.73 | 0.666 | 0.589 | 0.6 | 0.884 | 6,109,922 |
| 20 | 0.593 | 0.287 | 0.269 | 0.262 | 0.807 | 6,269,204 |

Table 5. The results of the convolutional network with entropy features.

| Number of Classes | Acc | Avg Precision | Avg Recall | Avg F1 | Avg AUC | Total Params |
|-------------------|-------|---------------|------------|--------|---------|--------------|
| 2 | 0.892 | 0.889 | 0.893 | 0.891 | 0.96 | 58,178 |
| 5 | 0.765 | 0.714 | 0.662 | 0.68 | 0.910 | 58,259 |
| 20 | 0.698 | 0.355 | 0.339 | 0.332 | 0.815 | 58,664 |

Figures 7–15 show the confusion matrices of the results of the evaluated networks.

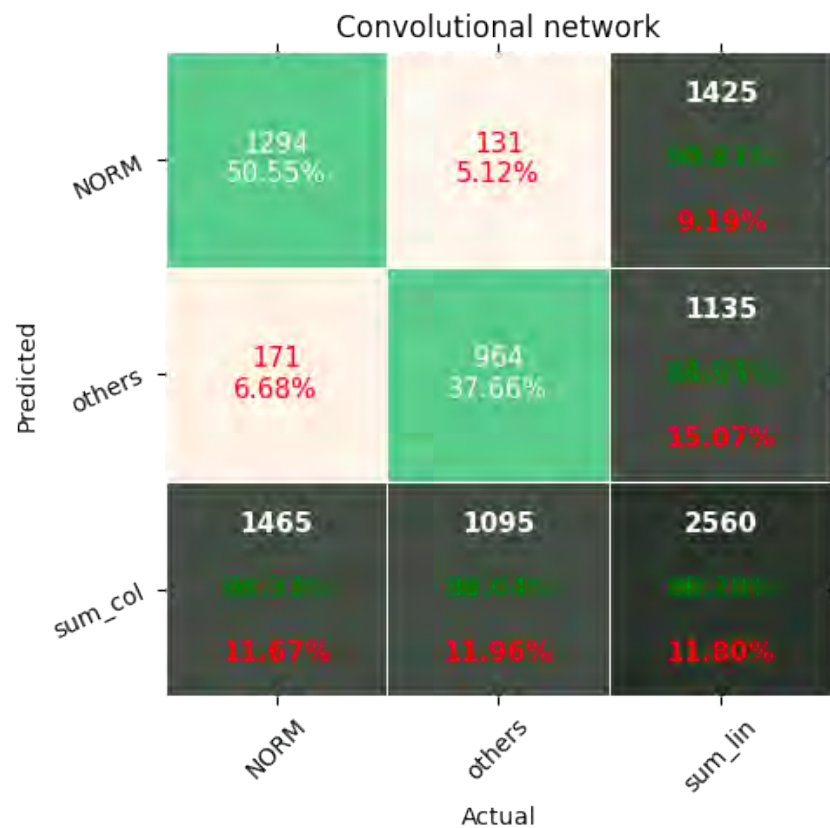


Figure 7. Confusion matrix of results for 2 classes for the convolutional network.

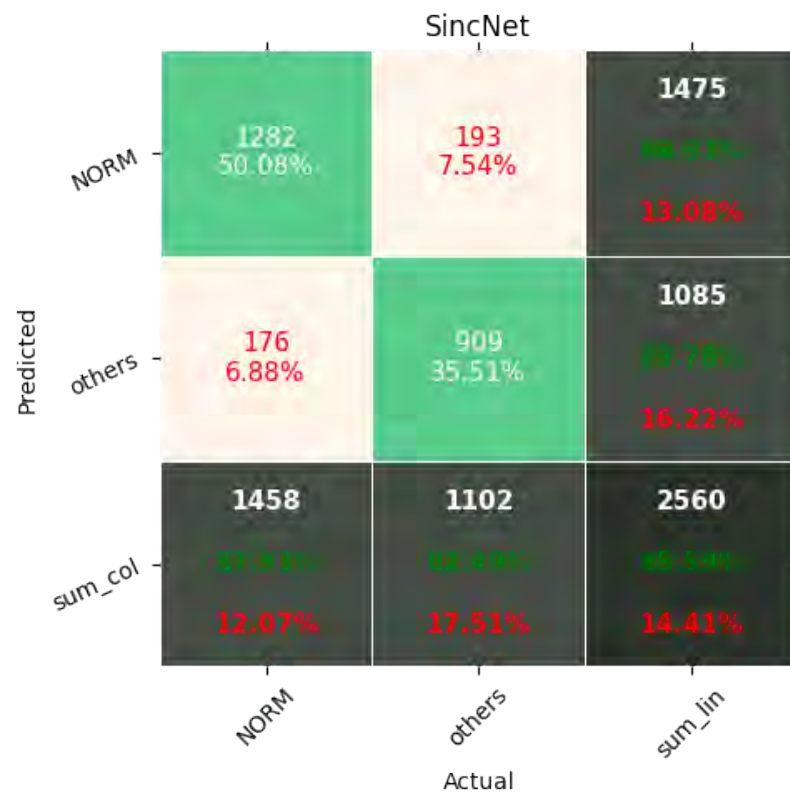


Figure 8. Confusion matrix of results for 2 classes for SincNet.

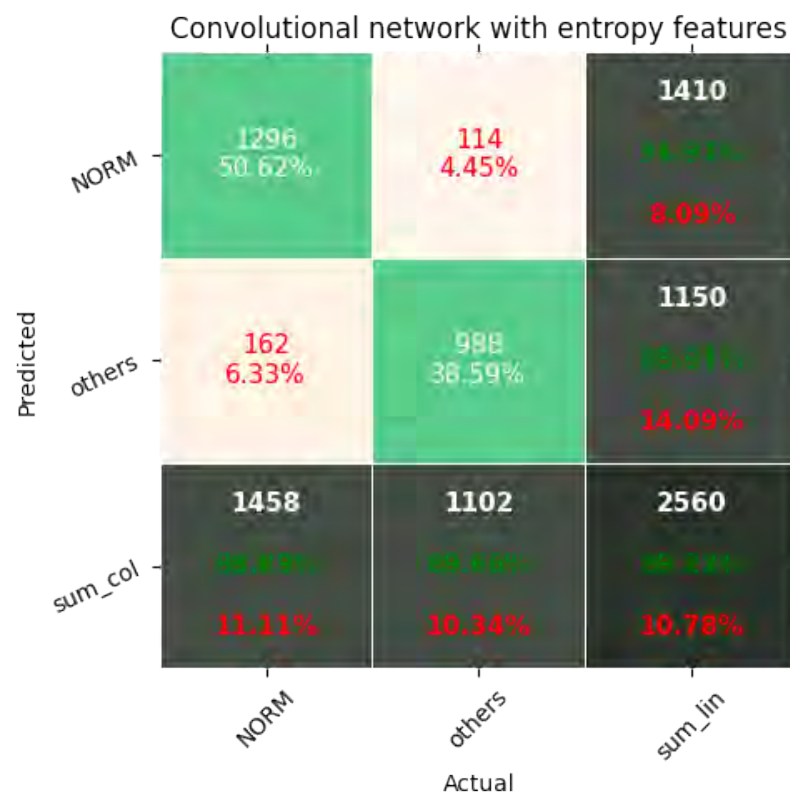


Figure 9. Confusion matrix of results for 2 classes for the convolutional network with entropy features.

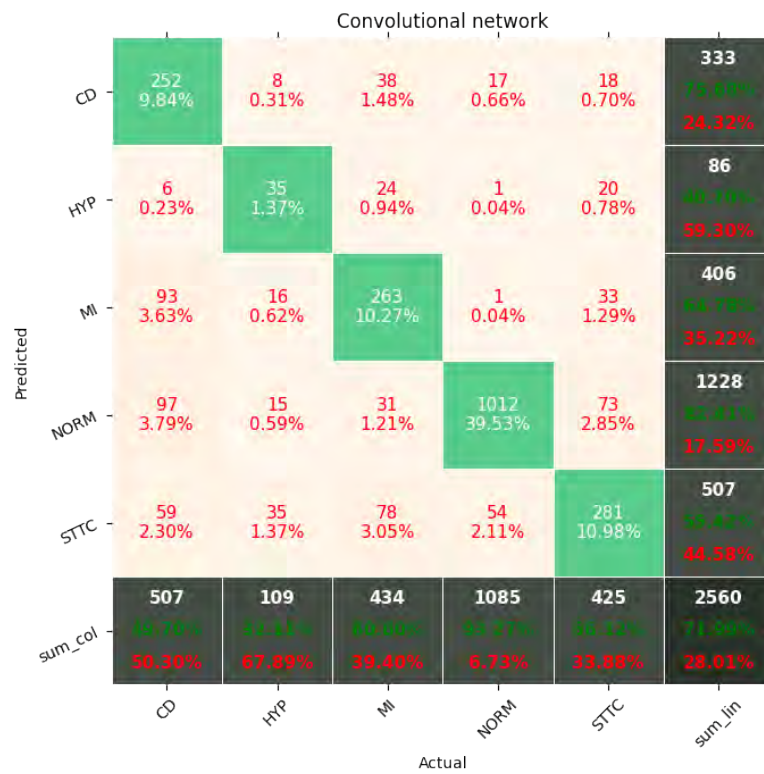


Figure 10. Confusion matrix of results for 5 classes for the convolutional network.

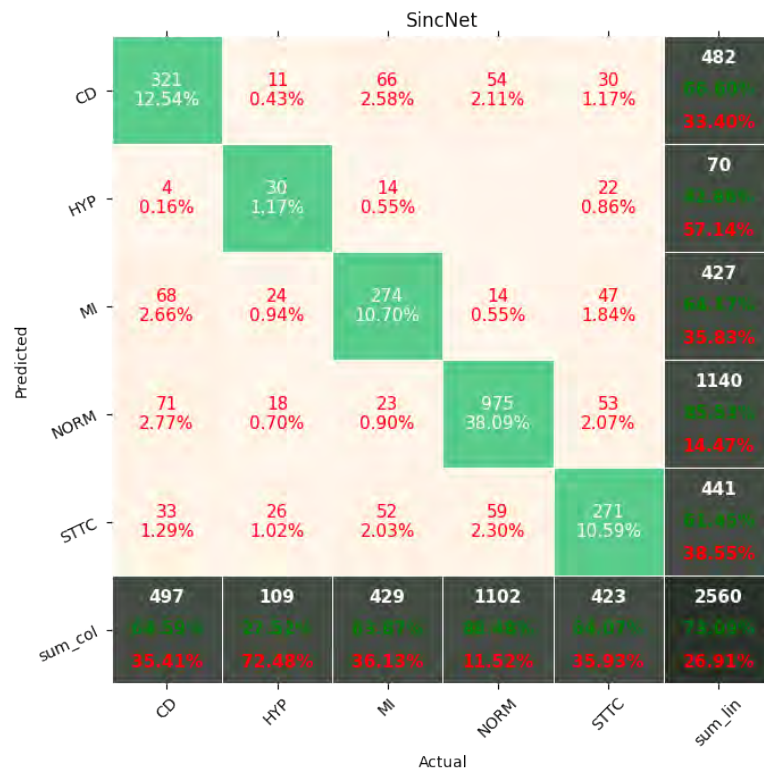


Figure 11. Confusion matrix of results for 5 classes for SincNet.

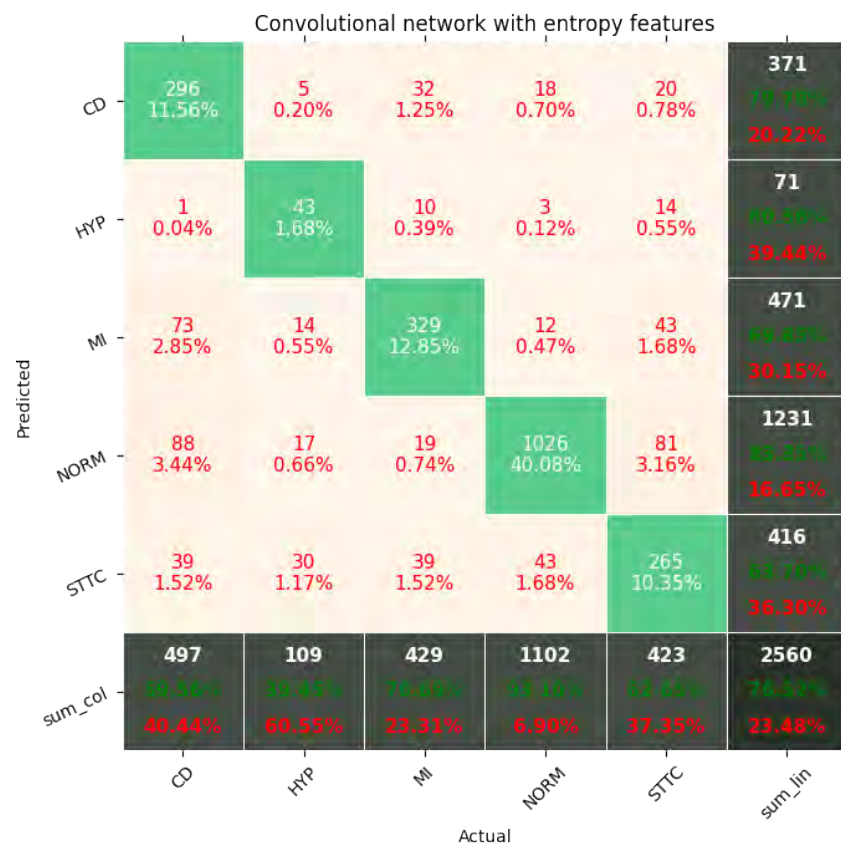


Figure 12. Confusion matrix of results for 5 classes for the convolutional network with entropy features.

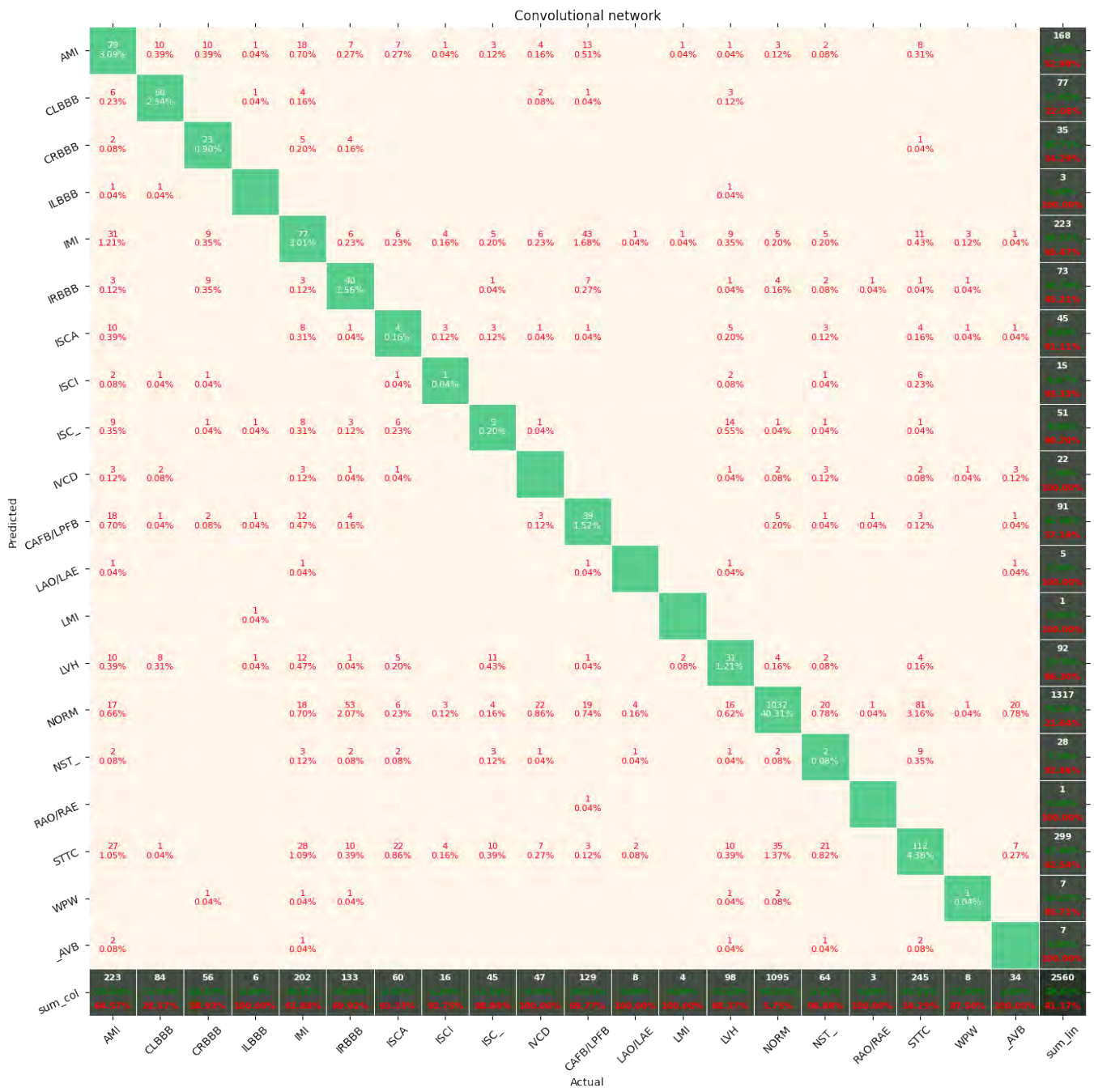


Figure 13. Confusion matrix of results for 5 classes for the convolutional network.

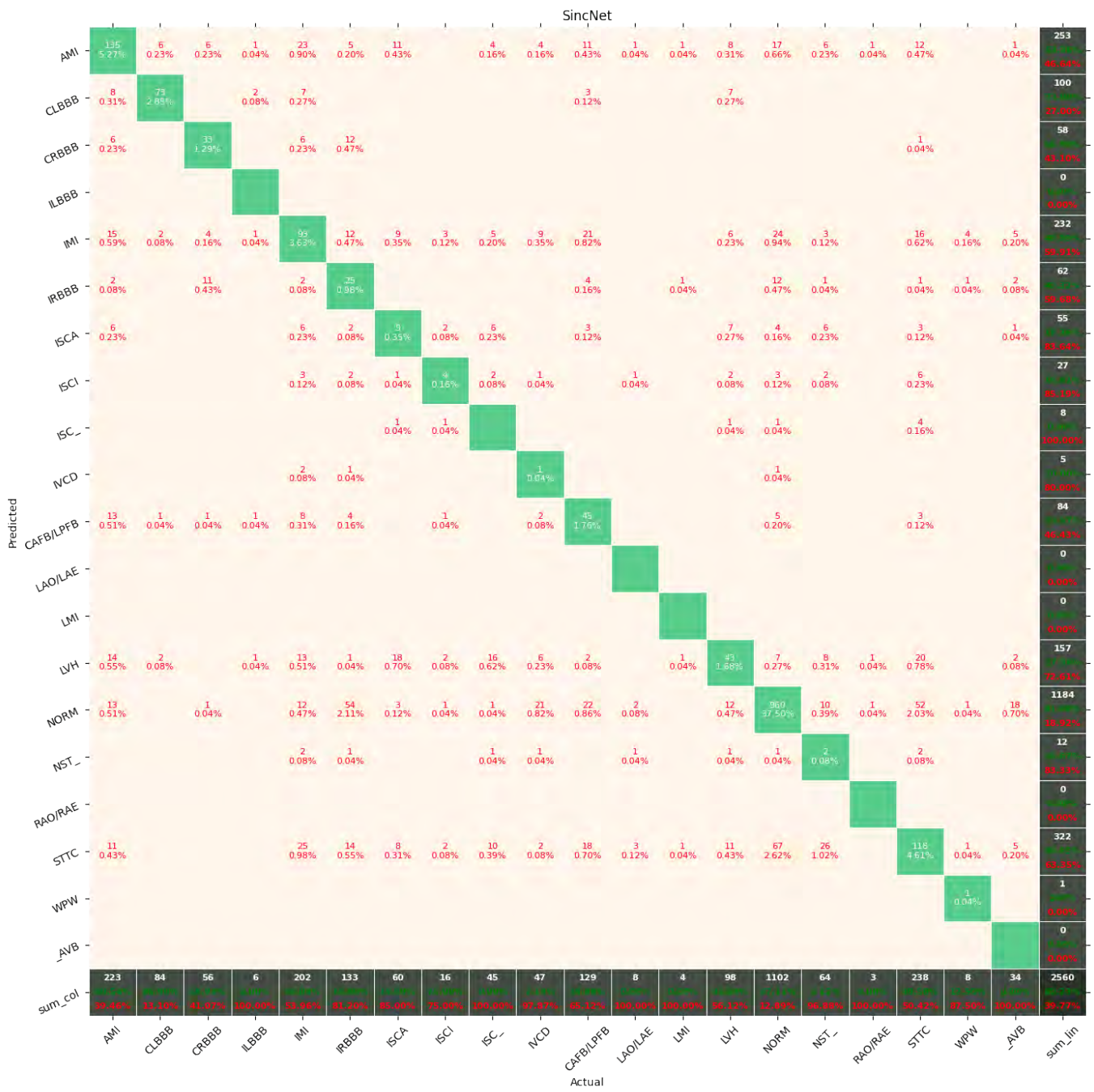


Figure 14. Confusion matrix of results for 5 classes for SincNet.

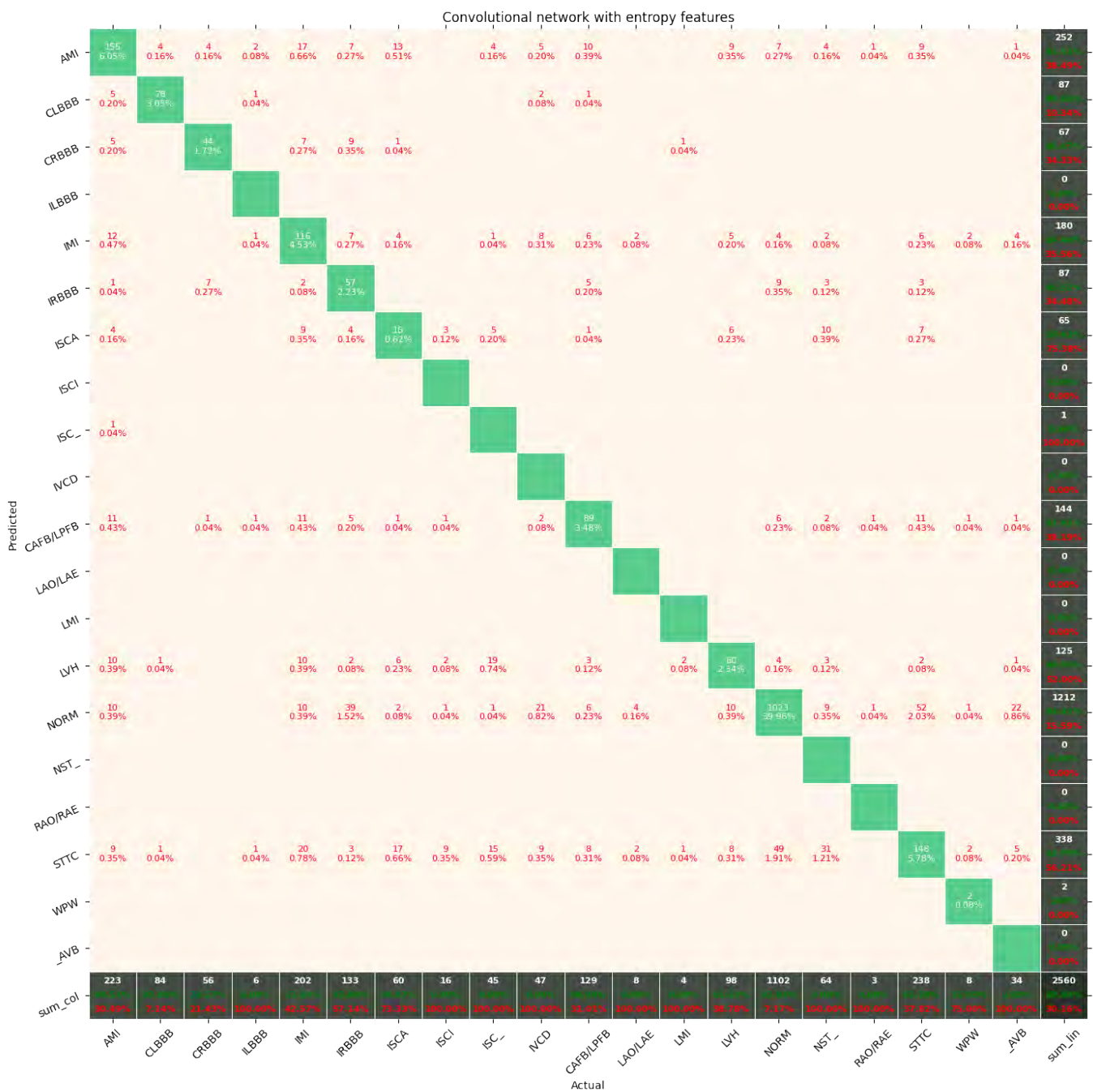


Figure 15. Confusion matrix of results for 5 classes for the convolutional network with entropy features.

In all cases of the evaluated networks, the NORM class obtained the highest value, which resulted from the large number of ECG recordings in this class.

4. Discussion

This paper presented a new model of convolutional neural networks, optimized to limit the computational and memory complexity for ECG recognition and classification of cardiovascular diseases. The research was carried out using a CNN network based on the convolutional network, which is relatively light and yields good results. The advantage of this approach is the possibility of using it on mobile and embedded devices, such as a Raspberry Pi or smartphone graphics cards.

The application of additional entropy-based features significantly improved the results. Such a solution also increased the weight of the network several times, however. As a

result, in applications where a very light network is needed, a compromise between weight and accuracy should be sought.

SincNet is a promising solution, but due to being designed to work with the human voice, it does not cope well with ECG signals in its original format. This results from the use of a set of initialization frequencies used in the computation of the Mel-frequency cepstral coefficients that are adapted to the spectral characteristics of the human voice. In the future, it would be worth considering the possibility of adapting SincNet to work with ECG.

The authors were unable to obtain better results due to the issue of overfitting on the training dataset. It was presumed that the addition of customized features may further boost the performance. The authors plan to investigate this claim in their next work.

Sampling determines the amount of measurements used to describe the signal. By changing the sampling, the signal is described by either more or fewer samples, whereas a stack of convolutional layers processes a fixed number of measurements in one context window. As a result, through a modification of the signal sampling, the network may either come to focus on more global features by reducing the amount of samples describing the signal or increase its attention to the details by increasing the measurements per signal.

Interpreting signals with different samplings may prove beneficial. In this work, we used only signals encoding 10 s of experiment on 1000 samples. It may well be the case that a network simultaneously interpreting a signal sampled with frequencies of 500 samples per second, 100 samples per second, and 50 samples per second will return better results. This is because signals sampled at lower frequencies can have entire ECG waves interpreted by one convolutional block, while signals sampled more frequently provide more detailed series for the extraction of features encoded by a small part of an ECG wave.

The proposed network based on a convolutional network is relatively uncomplicated. It is likely that better results could be obtained with the use of Inception models. This model uses heterogeneous subnets to improve the result. It is comparable to the case of wavelet transform, which may prove to be more advantageous than the use of fast Fourier transform. According to the authors, the proposed solution could be used in small devices for continuous monitoring of ECG signals, for example to alert about anomalies and make an initial diagnosis or support a doctor in this.

The authors assumed that a network's performance may be improved with a manageable cost increase by expanding its architecture with Inception-style heterogeneous subnetworks with varying kernels and poolings. The authors intend to investigate this assumption in their future work.

The authors further assumed that the integration of SincNet layers for low-level feature extraction in the first step of signal processing with the successful implementation of the first network based on convolutional layers may prove a benefit. The authors intend to investigate this assumption in their future work.

5. Conclusions

This study presented the capability of convolutional neural networks in the classification of heart diseases by the examination of ECG signals. The network proposed by the authors is both accurate and efficient as it is lightweight, allowing it to be computed on nonspecialized devices. The application of entropy-based features proved beneficial due to the improvements in the accuracy of heart disease classification. Entropy-based features are promising additions to data preprocessing that may prove beneficial in other signal-processing-related tasks.

Author Contributions: Conceptualization, S.Ś., K.P. and D.L.; methodology, S.Ś., K.P. and D.L.; software, S.Ś., K.P., and D.L.; validation, S.Ś., K.P. and D.L.; formal analysis, S.Ś., K.P., and D.L.; investigation, S.Ś., K.P. and D.L.; resources, S.Ś., K.P., and D.L.; data curation, S.Ś., K.P. and D.L.; writing—original draft preparation, S.Ś., K.P., and D.L.; writing—review and editing, S.Ś., K.P. and D.L.; visualization, S.Ś., K.P. and D.L. All authors read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Benjamin, E.J.; Virani, S.S.; Callaway, C.W.; Chamberlain, A.M.; Chang, A.R.; Cheng, S.; Chiuve, S.E.; Cushman, M.; Delling, F.N.; Deo, R.; et al. Heart disease and stroke statistics—2018 update: A report from the American Heart Association. *Circulation* **2018**, *137*, e67–e492. [[CrossRef](#)] [[PubMed](#)]
2. Gupta, D.; Bajpai, B.; Dhiman, G.; Soni, M.; Gomathi, S.; Mane, D. Review of ECG arrhythmia classification using deep neural network. *Mater. Today Proc.* **2021**, In Press. [[CrossRef](#)]
3. World Health Organization. *Global Status Report on Noncommunicable Diseases*; WHO: Geneva, Switzerland, 2014.
4. Bogun, F.; Anh, D.; Kalahasty, G.; Wissner, E.; Serhal, C.B.; Bazzi, R.; Weaver, W.D.; Schuger, C. Misdiagnosis of atrial fibrillation and its clinical consequences. *Am. J. Med.* **2004**, *117*, 636–642. [[CrossRef](#)] [[PubMed](#)]
5. Schläpfer, J.; Wellens, H.J. Computer-interpreted electrocardiograms: Benefits and limitations. *J. Am. Coll. Cardiol.* **2017**, *70*, 1183–1192. [[CrossRef](#)]
6. Houssein, E.H.; Kilany, M.; Hassanien, A.E. ECG signals classification: A review. *Int. J. Intell. Eng. Informatics* **2017**, *5*, 376–396. [[CrossRef](#)]
7. Jambukia, S.H.; Vipul, K.D.; Harshadkumar, B.P. Classification of ECG signals using machine learning techniques: A survey. In Proceedings of the 2015 International Conference on Advances in Computer Engineering and Applications, Ghaziabad, India, 19–20 March 2015.
8. Macfarlane, P.W.; Devine, B.; Clark, E. The university of Glasgow (Uni-G) ECG analysis program. In Proceedings of the Computers in Cardiology, Lyon, France, 25–28 September 2005.
9. Wang, J.; Qiao, X.; Liu, C.; Wang, X.; Liu, Y.; Yao, L.; Zhang, H. Automated ECG classification using a non-local convolutional block attention module. *Comput. Methods Programs Biomed.* **2021**, *203*, 106006. [[CrossRef](#)] [[PubMed](#)]
10. Goldberger, A.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E.; et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220. [[CrossRef](#)]
11. Wagner, P.; Strodthoff, N.; Boussejot, R.; Samek, W.; Schaeffter, T. PTB-XL, a large publicly available electrocardiography dataset (version 1.0.1). *Sci. Data* **2020**, *7*, 1–5. [[CrossRef](#)]
12. Jia, W.; Xu, X.; Xu, X.; Sun, Y.; Liu, X. Automatic Detection and Classification of 12-lead ECGs Using a Deep Neural Network. In Proceedings of the Computing in Cardiology, Rimini, Italy, 13–16 September 2020; pp. 1–4.
13. Zhu, Z.; Lan, X.; Zhao, T.; Guo, Y.; Kojodjojo, P.; Xu, Z.; Liu, Z.; Liu, S.; Wang, H.; Sun, X.; et al. Identification of 27 abnormalities from multi-lead ECG signals: An ensemble SE_ResNet framework with sign loss function. *Physiol. Meas.* **2021**, *42*, 065008. [[CrossRef](#)]
14. Strodthoff, N.; Wagner, P.; Schaeffter, T.; Samek, W. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *arXiv* **2020**, arXiv:2004.13701.
15. Smisek, R.; Nemcova, A.; Marsanova, L.; Smital, L.; Vitek, M.; Kozumplik, J. Cardiac Pathologies Detection and Classification in 12-lead ECG. In Proceedings of the Computing in Cardiology, Rimini, Italy, 13–16 September 2020; pp. 1–4.
16. Zhang, D.; Yang, S.; Yuan, X.; Zhang, P. Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram. *Iscience* **2021**, *4*, 102373. [[CrossRef](#)]
17. Warrick, P.A.; Lostonlen, V.; Eickenberg, M.; Andén, J.; Homsí, M.N. Arrhythmia Classification of 12-lead Electrocardiograms by Hybrid Scattering-LSTM Networks. In Proceedings of the Computing in Cardiology, Rimini, Italy, 13–16 September 2020; pp. 1–4.
18. Acharya, U.R.; Fujita, H.; Adam, M.; Lih, O.S.; Hong, T.J.; Sudarshan, V.K.; Koh, J.E. Automated characterization of arrhythmias using nonlinear features from tachycardia ECG beats. In Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Budapest, Hungary, 9–12 October 2016.
19. Jo, Y.Y.; Cho, Y.; Lee, S.Y.; Kwon, J.M.; Kim, K.H.; Jeon, K.H.; Cho, S.; Park, J.; Oh, B.H. Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram. *Int. J. Cardiol.* **2021**, *328*, 104–110. [[CrossRef](#)] [[PubMed](#)]
20. Lepek, M.; Pater, A.; Muter, K.; Wiszniewski, P.; Kokosińska, D.; Salamon, J.; Puzio, Z. 12-lead ECG Arrhythmia Classification Using Convolutional Neural Network for Mutually Non-Exclusive Classes. In Proceedings of the Computing in Cardiology, Rimini, Italy, 13–16 September 2020; pp. 1–4.
21. Ramaraj, E.; Virgeniya, S.C. A Novel Deep Learning based Gated Recurrent Unit with Extreme Learning Machine for Electrocardiogram (ECG) Signal Recognition. *Biomed. Signal Process. Control* **2021**, *68*, 102779.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

23. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
24. Caruana, R.; Lawrence, S.; Giles, L. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In Proceedings of the 14th Annual Neural Information Processing Systems Conference, Denver, CO, USA, 27 November–2 December 2020, pp. 402–408.
25. Ravanelli, M.; Yoshua, B. Speaker recognition from raw waveform with sincnet. In Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018.
26. Molau, S.; Pitz, M.; Schluter, R.; Ney, H. Computing Mel-frequency cepstral coefficients on the power spectrum. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001.
27. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [[CrossRef](#)]
28. Richman, J.S.; Moorman, J.R. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol.-Heart Circ. Physiol.* **2000**, *278*, H2039–H2049. [[CrossRef](#)]
29. Bandt, C.H.; Bernd, P. Permutation entropy: A natural complexity measure for time series. *Phys. Rev. Lett.* **2002**, *88*, 174102. [[CrossRef](#)]
30. Inouye, T.; Shinosaki, K.; Sakamoto, H.; Toi, S.; Ukai, S.; Iyama, A.; Katsuda, Y.; Hirano, M. Quantification of EEG irregularity by use of the entropy of the power spectrum. *Electroencephalogr. Clin. Neurophysiol.* **1991**, *79*, 204–210. [[CrossRef](#)]
31. Renyi, A. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; University of California Press: Oakland, CA, USA, 1961; pp. 547–561.
32. Bezerianos, A.; Tong, S.; Thakor, N. Time dependent entropy of EEG rhythm changes following brain ischemia. *Ann. Biomed. Eng.* **2003**, *31*, 221–232. [[CrossRef](#)]
33. Lad, F.; Sanfilippo, G.; Agrò, G. Extropy: A complementary dual of entropy. *arXiv* **2011**, arXiv:1109.6440.
34. Granero-Belinchón, C.; Roux, S.G.; Garnier, N.B. Information Theory for Non-Stationary Processes with Stationary Increments. *Entropy* **2019**, *21*, 1223. [[CrossRef](#)]

Article

Study of the Few-Shot Learning for ECG Classification Based on the PTB-XL Dataset

Krzysztof Pałczyński ¹, Sandra Śmigiel ^{2,*}, Damian Ledziński ¹ and Sławomir Bujnowski ¹

¹ Faculty of Telecommunications, Computer Science and Electrical Engineering, Bydgoszcz University of Science and Technology, 85-796 Bydgoszcz, Poland; krzysztof@palczynski.com.pl (K.P.); damian.ledzinski@pbs.edu.pl (D.L.); slawomir.bujnowski@pbs.edu.pl (S.B.)

² Faculty of Mechanical Engineering, Bydgoszcz University of Science and Technology, 85-796 Bydgoszcz, Poland

* Correspondence: sandra.smigiel@pbs.edu.pl; Tel.: +48-52-340-8346

Abstract: The electrocardiogram (ECG) is considered a fundamental of cardiology. The ECG consists of P, QRS, and T waves. Information provided from the signal based on the intervals and amplitudes of these waves is associated with various heart diseases. The first step in isolating the features of an ECG begins with the accurate detection of the R-peaks in the QRS complex. The database was based on the PTB-XL database, and the signals from Lead I–XII were analyzed. This research focuses on determining the Few-Shot Learning (FSL) applicability for ECG signal proximity-based classification. The study was conducted by training Deep Convolutional Neural Networks to recognize 2, 5, and 20 different heart disease classes. The results of the FSL network were compared with the evaluation score of the neural network performing softmax-based classification. The neural network proposed for this task interprets a set of QRS complexes extracted from ECG signals. The FSL network proved to have higher accuracy in classifying healthy/sick patients ranging from 93.2% to 89.2% than the softmax-based classification network, which achieved 90.5–89.2% accuracy. The proposed network also achieved better results in classifying five different disease classes than softmax-based counterparts with an accuracy of 80.2–77.9% as opposed to 77.1% to 75.1%. In addition, the method of R-peaks labeling and QRS complexes extraction has been implemented. This procedure converts a 12-lead signal into a set of R waves by using the detection algorithms and the k-mean algorithm.

Keywords: ECG signal processing; few-shot learning; R wave detection; distance-based classification; PTB-XL dataset; deep learning



Citation: Pałczyński, K.; Śmigiel, S.; Ledziński, D.; Bujnowski, S. Study of the Few-Shot Learning for ECG Classification Based on the PTB-XL Dataset. *Sensors* **2022**, *22*, 904. <https://doi.org/10.3390/s22030904>

Academic Editor: Christoph Hintermüller

Received: 28 October 2021

Accepted: 21 January 2022

Published: 25 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Machine learning, especially Deep Learning (DL) approaches, has been of interest in academia and industry. This has resulted in numerous changes in the approach to automatic detection or classification processes. However, the reliability of such studies has not always been high and differs depending on the methods used.

Since recently, it has been proved that Artificial Intelligence (AI) and machine learning has numerous applications in all engineering fields. Among them are the areas of electrical engineering [1], civil engineering [2], and petroleum engineering [3]. In addition, classification using DL methods [4] have several practical applications in various areas of medicine, such as the diagnosis of diseases based on physiological parameters [5], the classification of cardiac arrhythmias based on ECG signals [6,7], and the recognition of human activity [8]. Various ECG classification schemes based on DL were used to detect heart diseases [9–12], for example, using Long Short-Term Memory networks [13] and one-dimensional Convolution Neural Networks [14–16]. In addition, DL methods have been used to classify pathological conditions of the heart, such as arrhythmia, atrial fibrillation, ventricular fibrillation, and others.

Cardiovascular disease is a general term for a series of cardiovascular abnormalities that are the world's leading cause of death [17]. Each of them is identified and interpreted using an electrocardiogram (ECG). The ECG is an important non-invasive diagnostic method for the interpretation and identification of various types of heart disease. Figure 1 shows an illustrative waveform of the ECG signal. Every day, approximately 3 million ECGs are produced worldwide [18]. ECG data contain rich information about the rate and rhythm of the heartbeat. Clinically, the ECG is analyzed over a short period using a graph of several consecutive cardiac cycles. The process begins with R-peak detection. It is usually the most visible part of the ECG that can be easily identified. The ECG reflects the depolarization of the main mass of the ventricles and refers to the maximum amplitude in the QRS complex. QRS complexes are the starting point for the analysis of the ECG signal. They serve as rhythm items and provide information about intraventricular rhythm and conduction [19,20].

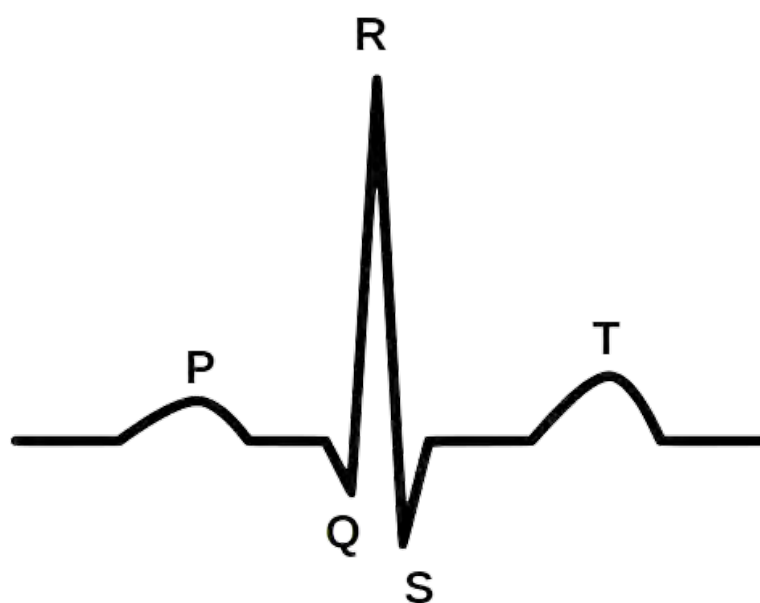


Figure 1. The illustrative waveform of the ECG signal.

Several methods and techniques have been used to locate the R-peak in the ECG signal, based on standard techniques such as digital filtering, wavelet transform, Fourier Transform, signal decomposition, and Hilbert Transform. However, only a few proposed works use DL methods in the literature to detect QRS complexes. One of the works in [21] is where a 300-point Convolutional Neural Network (CNN) and clustering on the neural output are used to detect QRS complexes on the pre-processed input signal. Another method using CNN has been proposed [22], demonstrating the reliable detection of the fetal QRS complex. The authors of the work [23] proposed a 1-D CNN and Multi-Layer Perceptron (MLP) classifier that determines the QRS positions. Another approach was the work [19] in which two DL models based on multi-dilated convolutional blocks were used: CNN and CRNN. Finally, this group of works includes [24], where a stacked autoencoder deep neural network is proposed to extract the QRS complex.

Regardless of the DL methods chosen, problems are identified, including classification efficiency, the detection of undesirable results, dependence on computing power, and the high sample count. In response to these problems, a few newly published articles propose using Few-Shot Learning (FSL) to identify new concepts in medicine and fill the gap between the efficiency and the size of the training samples. FSL mimics humans' ability to acquire knowledge from a few samples. This technique involves training a neural network to encode input data into small-sized vectors, which distances to other vectors encoding objects of the same class are smaller than to vectors representing objects from different

classes. The distance between vectors is usually computed by measuring the Euclidean distance between two vectors. In addition, FSL can encode information regarding the object's belonging to a particular class in the output vector. Because of that, the layer of neurons representing defined classes is not required, which allows the FSL network to distinguish between classes that were not seen during training, thereby enabling learning from limited samples and rapidly generalizing to new tasks, giving a different perspective on DL.

There are many areas of application of FSL methods. In the medical field, the use of FSL methods occurs in conjunction with medical images and medical signals. One of the application directions is to use the network-based FSL method to classify rare human peripheral blood leukocyte images. The proposed Siamese network by the authors of [25] contains two identical Convolutional Neural Networks and a logistic regression network. In justifying their research, the authors point to the relationship between the number of leukocytes and various diseases, including cancer. The obtained results show that the Siamese network can overcome the scarcity and imbalance of datasets used in this research. The results are promising and give hope for addressing the issue of rare leukocyte images recognition in medicine.

Another view is the use of Few-Shot Deep Learning in medical imaging, for example, COVID-19-infected areas in Computed Tomography (CT) images. Recent studies indicate that detecting radiographic patterns on chest CT scans can provide high sensitivity and specificity in identifying COVID-19. One of the works [26] was undertaken to investigate the efficacy of FSL in U-Net architectures, allowing for a dynamic fine-tuning of the network weights as new samples are fed into the U-Net. The obtained results confirmed the improvement of the segmentation accuracy improvement in the identification of COVID-19-infected regions. A similar approach was proposed by the authors of another study [27], pointing to the use of FSL for the computerized diagnosis of emergencies due to coronavirus-infected pneumonia on CT images. A similar application of FSL was demonstrated by the authors of the study [28], who undertook the classification of COVID-19 infected areas on X-rays. As part of the research, the method was tested to classify images showing unknown symptoms of COVID-19 in an environment designed to learn several samples, with prior meta-learning only on images of other diseases.

Diagnostics of disease states based on medical images using DL methods have also been applied in dermatology. The authors of the work [29] demonstrated the possibility of using FSL for Dermatological Disease Diagnosis. Skin diseases are increasingly becoming one of the most common human diseases, contributing to dangerous cancerous changes or affecting motor disability. The proposed method is scalable to new classes and can effectively capture intra-class variability. A similar approach was used by the authors of [30], who proposed a Few-Shot segmentation network for skin lesion segmentation, which requires only a few pixel-level annotations. The authors emphasize that the proposed method is a promising framework for Few-Shot segmentation of skin lesions. The conducted experiments show that removing the background region of the query image both accelerates the speed of network convergence and significantly improves the segmentation efficiency.

The works of other authors in medicine with the use of FSL indicate the possibility of application in creating predictive models of drug reactions based on screens of cell lines. For example, the authors' work in [31] applied Few-Shot machine learning to train a versatile neural network model in cell lines that can be tuned to new contexts using a few additional samples. The model quickly adapted to switching between different tissue types and shifting from cell line models to clinical contexts.

In biomedical signals, an interesting approach is to use the FSL method of Electroencephalography (EEG)-based Motor Imagery (MI) Classification. The authors of the work [32] drew attention to an essential aspect of research on the brain-computer interface using EEG signals. In their justification, they indicated the potential of EEG in designing key technologies in both healthcare and other industries. The research proposed a two-

way Few Shot network that can efficiently learn representative features of unseen subject categories and classify them with limited MI EEG data.

In the area of the ECG signal, the authors in [33] proposed a meta-transfer-based FSL method to handle arrhythmia classification with the ECG signal in wearable devices. The results obtained by the authors indicate that the proposed method exceeds the accuracy of other comparative methods when performing various Few Shot tasks within the same training samples.

The study aimed to determine the usefulness of the FSL for ECG signal proximity-based classification. The research was conducted by training Deep Convolutional Neural Networks to recognize 2, 5, and 20 different heart disease classes. For this task, two neural networks were trained. The first one was optimized by performing FSL to classify input samples based on Euclidean distance to the defined classes' vectors. The second one was trained to perform softmax-based classification. It serves as a basis for comparison due to its well-known effectiveness in recognizing classes established during training. This work also examines classification strategies in FSL by comparing the results obtained from proximity-based classification to training machine learning algorithms on top of optimized FSL neural networks. The tested machine learning algorithms are XGBoost, Random Forest, Decision Tree, K-Nearest Neighbors, and SVMs. The neural network proposed for this task interprets a set of QRS complexes extracted from ECG signals. The method of R-peaks labeling and QRS complexes extraction has been implemented. This procedure converts a 12-lead signal into a set of R waves by using the detection algorithms and the k-mean algorithm. The novelty of this work involves using the FSL learning style for training on known, fixed classes; its comparison with more typical, softmax-based classifications; and the evaluation of classification strategies to be employed on top of the trained FSL network.

This paper is organized as follows: Section 2 closely describes the methods, the architectures of the artificial intelligence system, and the previously carried out data filtering, R Wave detection, and QRS extraction. Then, Section 3 presents the result of the research. Then, the discussion is given in Section 4. Finally, Section 5 concludes the paper and provides a look at further studies on this topic.

2. Materials and Methods

The methodology used in the paper was as follows (Figure 2): The PTB-XL dataset containing the labeled 10-second raw-signal ECG was used for the research. First, the records in the database have been filtered. Then, the R waves were labeled in the records in the next step. On this basis, QRS segments were separated. Finally, the dataset has been split into training, test, and validation data (respectively 70%, 15%, 15%). These data were used to train two neural networks, based on softmax and a Few Shot, as classifiers of 2, 5, and 20 classes of heart diseases. In the last stage, the network performance was evaluated.

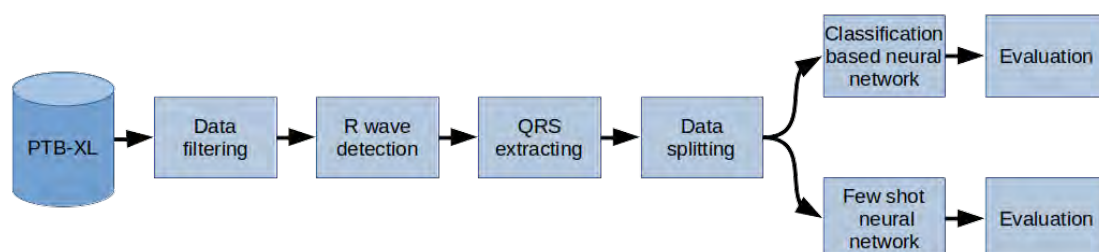


Figure 2. General overview diagram of the method.

2.1. PTB-XL Dataset

In this study, all the ECG data used come from the PTB-XL dataset [34,35]. PTB-XL is the publicly available and most extensive set of clinical ECG data. It provides a rich set of ECG annotations and additional metadata, which together constitute an ideal source for

training and evaluating machine learning algorithms. The PTB-XL dataset contains 12-lead 10 s ECGs from 18,885 different patients for a total of 21,837 records. ECG files come in two other options with 500 Hz and 100 Hz sampling rates with 16-bit resolution. The research used ECGs with 500 Hz sampling rates. The database contains 71 types of heart diseases with 5 significant classes: normal ECG (NORM), myocardial infarction (CD), ST/T change (STTC), conduction disturbance (MI), and hypertrophy (HYP).

2.2. Data Filtering

Initially, the PTB-XL had 21,837 records. However, not all records have labels (assigned classes), and not all assigned classes were 100% sure. For this reason, both cases were filtered out of the original dataset. Each record has a given class and a subclass for specific heart disease. Records with the number of subclasses less than 20 were also filtered from the original dataset. In this way, 17,232 records were obtained, each belonging to 1 of the 5 classes and 1 of the 20 subclasses. Figure 3 shows a detailed distribution of classes and subclasses. Descriptions of the classes of diseases are included in the in Appendixes A and B.

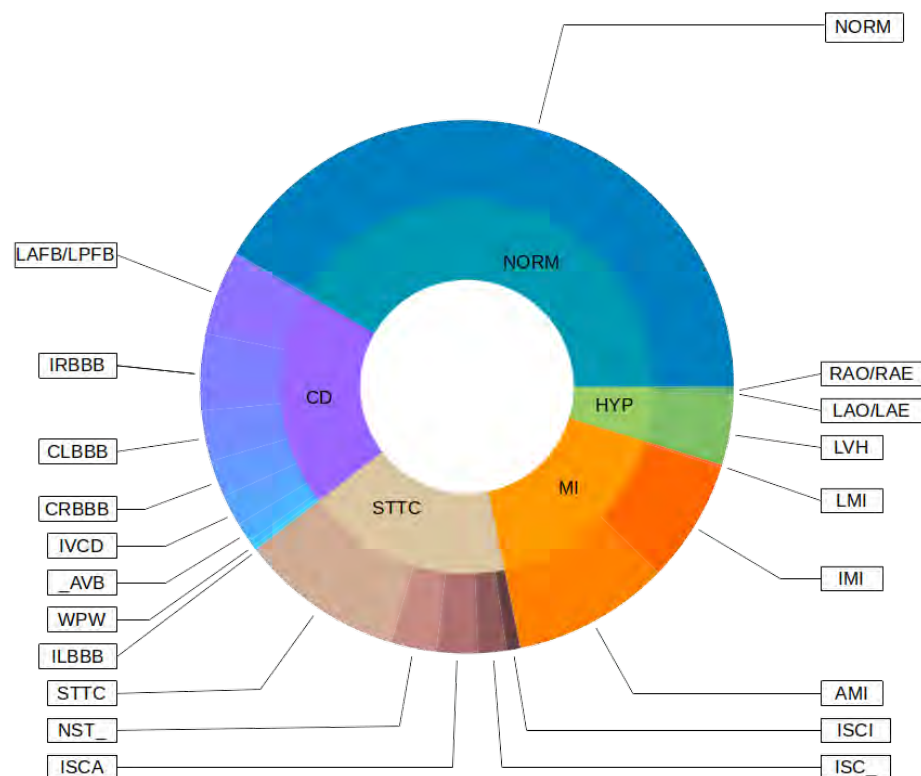


Figure 3. Classes and subclasses of used records.

2.3. R Wave Detection and QRS Extraction

None of the known R-peak detection methods tested by the authors were 100% effective. In addition, these methods use only a 1-lead signal. For this reason, the authors decided to propose their own method, using several known methods (Hamilton detector [36], Two Average detector [37], Stationary Wavelet Transform detector [38], Christov detector [39], Pan–Tompkins detector [40], and Engzee detector [41] with modification [42]) for all 12-leads and obtaining a consensus from them using k-mean algorithm. The designated R-peaks were used to cut the 10-s records into segments referred to further in the work as QRS complexes. The cuts were determined in the middle of the distance between the designated R-peaks (Figure 4). The first and last segments were removed. The following segments were resampled to 100 samples. In this way, for each record, a set of

QRS complexes and metadata as BPM (Beat Per Minute) and resampling ratio for each QRS complex were obtained.

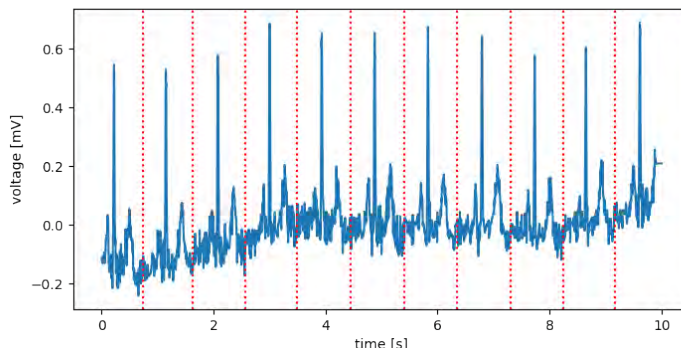


Figure 4. Sample record of NORM class for I lead, with places for section cuts (Red).

2.4. Designed Network Architectures

This chapter describes the architecture of the Deep Neural Networks used in this research (Figure 5) and the methodology of processing QRS complexes, applied loss functions, and training procedure.

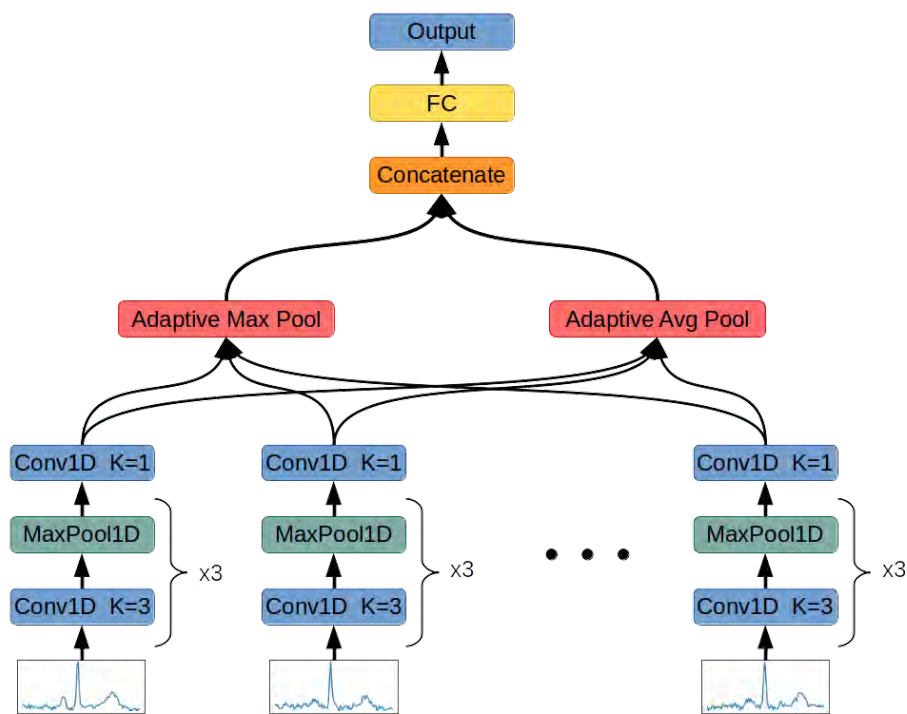


Figure 5. Designed Neural Network architecture.

The system receives the collection of QRS complexes stored in the input signal:

$$X_i = \{Q_1, \dots, Q_n\}, n \in N^+ \tag{1}$$

where:

X —set of input signals after QRS extraction performed;

i —index of signal being processed by the system;

Q_n — n -th extracted QRS complex containing 100 12-dimensional samples:

$$|Q_j| = 1200, j \in N^+ \cap j \leq n \tag{2}$$

Then, a set of QRS complexes is processed by the function designed to transform each wave into a 24-dimensional vector containing abstract features allowing for similarity calculation between vectors representing classes defined in the PTB-XL dataset:

$$f : R^{12 \times 100} \rightarrow R^{24} \quad (3)$$

The function has been approximated by the deep convolutional neural function described in Table 1. The process of learning this neural network has been presented in the Section 2.5.

Table 1. The architecture of Deep Convolutional Neural Network encoding one QRS complex.

| Layer | Channels In | Channels Out | Kernel Size | Padding | Stride |
|-----------|-------------|--------------|-------------|---------|--------|
| Conv1d | 12 | 24 | 3 | 1 | 1 |
| MaxPool1d | 24 | 24 | 2 | 0 | 2 |
| Conv1d | 24 | 48 | 3 | 0 | 1 |
| MaxPool1d | 48 | 48 | 2 | 0 | 2 |
| Conv1d | 48 | 96 | 3 | 0 | 1 |
| MaxPool1d | 96 | 96 | 2 | 0 | 2 |
| Conv1d | 96 | 2 | 1 | 0 | 1 |

Each convolutional layer's output is subjected to the LeakyReLU activation function with parameter equal to 0.01. The last convolutional layer operates using a kernel of size 1. This computation has been inspired by GoogLeNet architecture [43], and its task is to perform dimensionality reduction. This layer requires only 192 weights to reduce the activation map size 48 times.

The function approximated by Convolutional Neural Network is used to encode each QRS in the input data:

$$Z_i = \{f(X_{i,j}) | j \in N^+ \cap j < |X_i|\} \quad (4)$$

As a result, Z_i is a set of 24-dimensional vectors with varying cardinality. This set is now processed by Adaptive Maximum Pooling and Adaptive Average Pooling functions.

The Adaptive Maximum Pooling function selects maximum value from each dimension of the vectors in the set:

$$Zmax_i = [\max(\{Z_{i,j,1} | j \in N^+ \cap j < |X_i|\}), \dots, \max(\{Z_{i,j,24} | j \in N^+ \cap j < |X_i|\})] \quad (5)$$

The Adaptive Average Pooling function averages values of every dimension from vectors in the set:

$$Zavg_i = [\frac{1}{|Z_i|} \sum_{j=1}^{|Z_i|} Z_{i,j,1}, \dots, \frac{1}{|Z_i|} \sum_{j=1}^{|Z_i|} Z_{i,j,24}] \quad (6)$$

The results of both Adaptive Average Pooling and Adaptive Maximum Pooling are combined into 1 48-dimensional vector:

$$A = [Zmax_i, Zavg_i] \quad (7)$$

In the last step, the result is inputted to a fully connected layer with 20 neurons turning the 48-dimensional vector of concatenated pooling results into a 20-dimensional final vector:

$$F_i = f(A); f : R^{48} \rightarrow R^{20} \quad (8)$$

Vector F_i describes the input signal using 20 abstract features. It is used in both classification neural networks to determine the signal's class by subjecting it to softmax function for class probability distribution computation or in FSL for signal's class determination by measuring Euclidean distance to the center of the class represented by vector made of averaging feature vectors obtained from signals on the training dataset. In the case of

standard classification, there is also one more fully connected layer added to adjust the size of the abstract features vector to the number of classes in the classification task.

2.5. Training

The neural networks' parameters have been adjusted using Adam [44] optimizer. In addition, the dataset has been split into training, validation, and test sets five times to reduce the impact of fortunate weights randomization on the network's performance. The split was performed by dividing the dataset by 70%, 15%, and 15%.

The training dataset was used to determine the values of the network's weights. In addition, the network was evaluated on the validation dataset during the training process to perform early stopping [45] for overfitting reduction purposes. The final network's evaluation has been performed on a test dataset using the last saved set of weights, which scored the best result on the validation dataset. Each time the network scored the best result on the validation dataset, its weights have been saved. The training lasted until 10,000 epochs elapsed or early stopping was performed.

For the purpose of this research, two neural networks have been trained, one for FSL and one for standard classification serving as a basis for a benchmark. Both networks are structurally almost identical and differ only in adding one fully connected layer in standard classification tasks and the interpretation of output vector and employed loss function.

2.5.1. Few-Shot Learning

Few-Shot Learning network was trained using the triplet margin loss function [46]. The task of this loss function is to decrease the distance between vectors belonging to the same class and increase it for vectors from different classes. This process can be described by the formula:

$$L(a, p, n) = \max(d(a, p) - d(a, n) + m, 0) \quad (9)$$

where:

a —"anchor" vector. This vector is compared with the other two vectors;

p —"positive" vector. This vector belongs to the same class as the "anchor" vector;

n —"negative" vector. This vector belongs to the different class as the "anchor" vector;

m —margin. Quantity describing desired separation of vectors from the same class with vectors from different classes. In this research, m was equal to 1;

d —distance function, $d : (R^{20}, R^{20}) \rightarrow R^1$.

For this research purpose, the Euclidean distance has been used as a distance function:

$$d(x, x') = \sum_{j=0}^{|x|} (x_j - x'_j)^2 \quad (10)$$

The purpose of the triplet margin loss function is to ensure that the distance between vectors from two different classes is higher than a distance between vectors of the same class in addition to constant margin m . The neural network is not penalized for its performance only if:

$$d(a, p) - d(a, n) + m \leq 0 \quad (11)$$

$$d(a, p) \leq d(a, n) - m \quad (12)$$

Minimizing this function ensures the separation of inter-class distances from distances to vectors of other classes by the margin of m .

During training, triplets of vectors, two from the same class and one from different classes, were randomly selected and fed to the network. At each step, classes were picked from the distribution created from the computing frequency of occurrence in the dataset. This approach was motivated by the a priori assumption that reciprocating class observation frequency from dataset to training process results in better network convergence. However, for more balanced training, a different approach may be undertaken, in which classes are picked from either a weighted frequency-based distribution or a univariate one.

Due to the PyTorch limitation of forming only homogenous-sized tensors, the process of forming batches requires one more restriction on the triplet sampling function. Every sample in the batch must have the exact number of QRS complexes. The batch-sampling function first randomly selects the number of QRS complexes required in this batch to obtain such tensors. Then, it randomly selects triplets from signals in the dataset that contain the same amount of QRS complexes as the value selected. Finally, the amount of QRS complexes in the batch is sampled from the distribution weighted by the frequency of each wave in the dataset. The evaluation process of the neural network consists of these steps:

1. Split evaluation dataset randomly into two sets while ensuring that QRS complexes for each class have the same cardinality. From now on, the first set is referenced as a “database” set and the second one as a “query” dataset.
2. Use an Artificial Intelligence system to convert each set of QRS complexes from both “database” and “query” datasets into 20-dimensional vectors.
3. For each class, take all vectors belonging to it from the “database” set and compute the average 20-dimensional vector. It results in average vectors being later referenced as “class center vectors”.
4. For each vector in the “query” dataset, compute its distance to every “class center vector”. The class, whose “center vector” has been the closest to the vector from the “query” dataset is the class associated with the entry in the “query” dataset.
5. Calculate evaluation metrics by comparing true labels of vectors in the “query” dataset with labels computed in the previous step.

This process emulates the behavior of the real-life working environment. The “database” set resembles the structure that stores previously measured and processed ECG signals labeled by professionals. This database is used to label incoming queries. In this research, entries in the database were aggregated by computing the average for each class. This solution involves the least amount of computational cost. It is because “class center vectors” are computed once. Then, the incoming query must be compared with only one vector per class instead of numerous database entries, as required in other strategies.

The other method of classification involved training machine learning models on top of network-encoded small-sized vectors. The machine learning models evaluated in this work are XGBoost, Random Forest, Decision Tree, K-Nearest Neighbors, and SVMs with linear, polynomial, radial basis function, and sigmoid kernels. In this approach, the FSL neural network generates small-size vectors encoding crucial features of the input signals. Then, the aforementioned machine learning algorithms are trained to classify these vectors.

2.5.2. Softmax-Based Classification

Softmax-based classification is a well-known process of training a neural network using the operation mentioned above as an activation function for converting the neural network’s output into a class probability distribution. The equation of the softmax function is given below:

$$\sigma(Z)_i = \frac{e^{Z_i}}{\sum_{j=1}^{|Z|} e^{Z_j}} \quad (13)$$

where:

Z —output vector computed by neural network;

$\sigma(Z)_i$ —value of class probability distribution function for i -th class.

The output of the softmax activation function is then compared with the desired results using cross-entropy loss function computed with the formula below:

$$L(p, y) = - \sum_{c=1}^M y_{o,c} \ln(p_{o,c}) \quad (14)$$

where:

p —probability that observation o belongs to the class c computed by application of softmax function on the output of neural network; y —binary value that is equal to 1 if observation o belongs to the class c and 0 if not.

The loss function forces the neural network to output the vector as close as possible to a one-hot encoded vector with the maximum value contained under the index of the class the signal belongs to. This is a well-established solution tested both by scientists and engineers and in this research, it serves as a basis for comparison between FSL network results and softmax-based one.

2.6. Metrics

Neural networks were evaluated using the metrics described below [16]. For simplicity of equations, specific acronyms have been created, as follows: TP —True Positive, TN —True Negative, FP —False Positive, FN —False Negative. The metrics used for network evaluation are:

- Accuracy: $Acc = (TP + TN) / (TP + FP + TN + FN)$;
- Precision = $TP / (TP + FP)$;
- Recall = $TP / (TP + FN)$;
- $F1 = 2 \cdot Precision \cdot Recall / (Precision + Recall)$;
- AUC—Area Under ROC. ROC (Receiver operating characteristic) is a curve determined by calculating the True Positive Rate = $TFP = TP / (TP + FN)$ and the False Positive Rate = $FPR = FP / (TN + FP)$. The False Positive Rate describes the x-axis and the True Positive Rate the y-axis of a coordinate system. By changing the threshold value responsible for the classification of an example as belonging to either the positive or negative class, pairs of TFP – FPR are generated, resulting in the creation of the ROC curve. AUC is a measurement of the area below the ROC curve.

3. Results

The networks have been evaluated using the k-fold cross-validation technique for $k = 5$. Each network has been trained five times from scratch on the randomly selected train, validation, and test datasets. The evaluation results on the test dataset are presented in Tables 2–7 for tasks involving the classification of 2, 5, and 20 classes, respectively. Tables show the averaged, minimal, and maximal accuracy values and the F1, AUC, and specificity and sensitivity scores with standard deviation. Additionally, the average accuracy and the F1 score achieved by the evaluated models have been presented in Figures 6 and 7.

Table 2. Results for two-class classification, part I.

| Technique | Acc | Acc Avg Std | F1 | F1 Avg Std | AUC | AUC Avg Std |
|----------------------------------|------------|---------------|-----------|--------------|-----------|---------------|
| FSL proximity-based | 89.5–91.1% | 90.4% 0.5% | 89.1–90.8 | 90.6 0.6 | 92.5–94.4 | 93.7 0.8 |
| Softmax-based classification | 89.2–90.5% | 89.7% 0.4% | 89.0–90.2 | 89.4 0.4 | 94.8–95.9 | 95.5 0.4 |
| FSL + XGBoost | 87.9–89.7% | 88.9% 0.7% | 86.5–88.5 | 87.7 0.8 | 95.1–97.2 | 96.1 0.7 |
| FSL + Random Forest | 87.8–91.2% | 89.4% 1.1% | 86.2–90.1 | 88.1 1.3 | 95.5–97.1 | 96.3 0.5 |
| FSL + Decision Tree | 84.9–88.9% | 86.4% 1.4% | 82.8–87.5 | 85.0 1.8 | 82.8–87.5 | 85.0 1.8 |
| FSL + KNN – 5 neighbors | 88.7–92.0% | 89.9% 1.2% | 87.1–91.2 | 88.8 1.4 | 93.9–96.4 | 94.6 0.9 |
| FSL + KNN – 20 neighbors | 88.1–93.3% | 90.9% 1.9% | 86.6–92.6 | 89.8 2.2 | 96.0–97.8 | 96.6 0.7 |
| FSL + SVM with linear kernel | 88.6–93.3% | 91.2% 1.6% | 87.4–92.9 | 90.3 1.8 | 96.1–97.6 | 96.9 0.5 |
| FSL + SVM with polynomial kernel | 87.2–93.0% | 89.6% 1.9% | 85.5–92.3 | 88.2 2.3 | 94.9–97.6 | 96.0 1.0 |
| FSL + SVM with RBF kernel | 89.2–93.3% | 91.3% 1.4% | 88.1–92.8 | 90.5 1.6 | 92.2–95.6 | 93.8 1.3 |
| FSL + SVM with Sigmoid kernel | 68.6–92.9% | 86.6% 9.1% | 66.2–92.2 | 85.3 9.6 | 83.6–95.3 | 88.2 4.0 |

Table 3. Results for two-class classification, part II.

| Technique | Specificity | Specificity Avg Std | Sensitivity | Sensitivity Avg Std |
|----------------------------------|-------------|-----------------------|-------------|-----------------------|
| FSL proximity-based | 89.6–91.0% | 90.4% 0.5% | 88.9–90.7% | 89.9% 0.6% |
| Softmax-based classification | 89.0–90.2% | 89.4% 0.5% | 88.9–90.7% | 89.9% 0.6% |
| FSL + XGBoost | 89.1–90.7% | 89.8% 0.6% | 86.5–88.5% | 87.7% 0.8% |
| FSL + Random Forest | 89.4–91.9% | 90.3% 1.0% | 86.3–90.1% | 88.2% 1.3% |
| FSL + Decision Tree | 86.4–89.7% | 87.6% 1.2% | 82.8–87.5% | 85.0% 1.8% |
| FSL + KNN – 5 neighbors | 89.6–92.7% | 90.8% 1.1% | 87.1–91.2% | 88.8% 1.4% |
| FSL + KNN – 20 neighbors | 89.4–93.9% | 91.6% 1.7% | 86.6–92.6% | 89.8% 2.2% |
| FSL + SVM with linear kernel | 89.5–93.5% | 91.6% 1.5% | 87.4–92.9% | 90.3% 1.8% |
| FSL + SVM with polynomial kernel | 89.2–93.7% | 91.0% 1.6% | 85.5–92.3% | 88.2% 2.3% |
| FSL + SVM with RBF kernel | 90.0–93.5% | 91.7% 1.3% | 88.1–92.8% | 90.5% 1.6% |
| FSL + SVM with Sigmoid kernel | 68.2–93.4% | 87.2% 9.6% | 66.2–92.2% | 85.3% 9.6% |

Table 4. Results for five-class classification, part I.

| Technique | Acc | Acc Avg Std | F1 | F1 Avg Std | AUC | AUC Avg Std |
|----------------------------------|------------|---------------|-----------|--------------|-----------|---------------|
| FSL proximity-based | 69.8–74.2% | 71.8% 1.7% | 60.6–66.9 | 63.7 2.4 | 85.6–88.9 | 87.6 1.2 |
| Softmax-based classification | 75.1–77.1% | 75.8% 0.8% | 66.8–69.6 | 67.9 1.0 | 87.5–90.9 | 89.6 1.3 |
| FSL + XGBoost | 74.8–76.1% | 75.2% 0.5% | 66.9–70.8 | 68.4 1.6 | 90.9–92.3 | 91.8 0.5 |
| FSL + Random Forest | 75.2–77.7% | 76.3% 0.8% | 66.8–69.9 | 68.4 1.1 | 92.0–93.0 | 92.5 0.4 |
| FSL + Decision Tree | 67.0–68.5% | 68.0% 0.5% | 58.7–63.3 | 61.3 1.4 | 75.2–77.8 | 76.7 0.8 |
| FSL + KNN – 5 neighbors | 74.4–76.7% | 75.8% 0.8% | 65.9–71.1 | 68.4 2.1 | 87.5–89.8 | 88.8 0.8 |
| FSL + KNN – 20 neighbors | 77.3–79.5% | 78.4% 0.9% | 68.2–71.6 | 69.6 1.2 | 92.0–93.2 | 92.5 0.5 |
| FSL + SVM with linear kernel | 77.0–79.8% | 78.8% 1.0% | 69.9–73.1 | 71.7 1.1 | 93.4–94.3 | 93.8 0.3 |
| FSL + SVM with polynomial kernel | 74.5–76.9% | 76.0% 0.9% | 64.7–69.2 | 66.6 1.6 | 92.4–93.2 | 92.9 0.3 |
| FSL + SVM with RBF kernel | 77.9–80.2% | 79.0% 0.9% | 69.0–71.8 | 70.6 1.0 | 93.3–93.8 | 93.6 0.3 |
| FSL + SVM with Sigmoid kernel | 64.4–76.6% | 72.9% 4.4% | 53.1–65.0 | 62.1 4.5 | 89.5–92.8 | 90.8 1.1 |

Table 5. Results for five-class classification, part II.

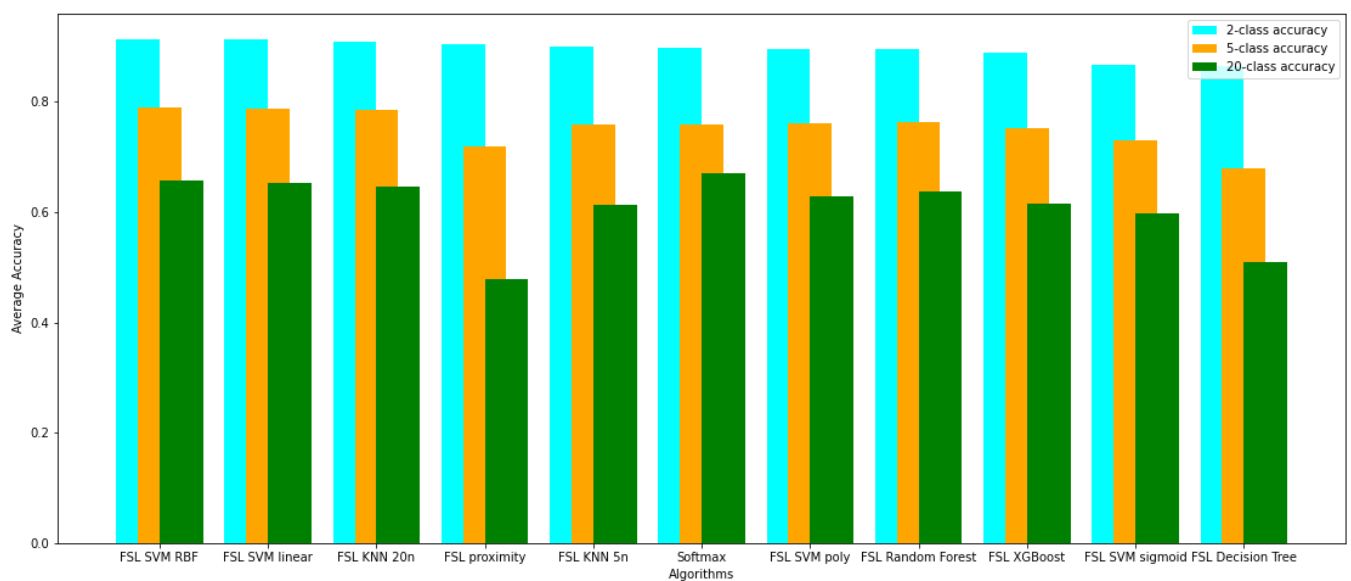
| Technique | Specificity | Specificity Avg Std | Sensitivity | Sensitivity Avg Std |
|----------------------------------|-------------|-----------------------|-------------|-----------------------|
| FSL proximity-based | 60.2–66.4% | 63.2% 2.1% | 62.7–68.1% | 65.9% 2.0% |
| Softmax-based classification | 68.3–70.6% | 69.5% 0.8% | 65.9–69.1% | 67.1% 1.1% |
| FSL + XGBoost | 65.7–68.0% | 66.9% 0.9% | 66.9–70.8% | 68.4% 1.7% |
| FSL + Random Forest | 66.8–68.3% | 67.8% 0.6% | 66.8–69.9% | 68.4% 1.2% |
| FSL + Decision Tree | 57.1–59.9% | 59.0% 1.0% | 58.8–63.4% | 61.3% 1.5% |
| FSL + KNN – 5 neighbors | 64.9–70.0% | 67.6% 1.8% | 65.9–71.1% | 68.4% 2.1% |
| FSL + KNN – 20 neighbors | 70.2–72.9% | 71.9% 1.0% | 68.2–71.6% | 69.6% 1.2% |
| FSL + SVM with linear kernel | 69.8–74.5% | 72.4% 1.6% | 69.9–73.1% | 71.7% 1.1% |
| FSL + SVM with polynomial kernel | 68.3–75.5% | 72.4% 2.5% | 64.7–69.2% | 66.6% 1.6% |
| FSL + SVM with RBF kernel | 70.8–75.5% | 73.5% 1.8% | 69.0–71.8% | 70.6% 1.0% |
| FSL + SVM with Sigmoid kernel | 56.1–70.8% | 65.4% 5.1% | 53.1–65.0% | 62.1% 4.5% |

Table 6. Results for 20-class classification, part I.

| Technique | Acc | Acc Avg Std | F1 | F1 Avg Std | AUC | AUC Avg Std |
|----------------------------------|------------|---------------|-----------|--------------|-----------|---------------|
| FSL proximity-based | 44.3–50.1% | 47.8% 2.1% | 23.8–26.0 | 24.7 0.8 | 78.8–84.4 | 80.8 2.5 |
| Softmax-based classification | 66.2–68.2% | 67.1% 0.8% | 31.9–33.0 | 32.4 0.4 | 82.4–86.3 | 84.4 1.5 |
| FSL + XGBoost | 58.7–66.2% | 61.6% 0.5% | 25.3–34.5 | 29.7 3.4 | 74.2–86.3 | 79.3 3.5 |
| FSL + Random Forest | 61.5–66.6% | 63.6% 2.5% | 27.1–36.7 | 30.9 3.5 | 73.6–80.3 | 77.1 2.3 |
| FSL + Decision Tree | 45.8–58.8% | 51.0% 4.5% | 19.8–30.0 | 25.0 4.1 | 58.4–60.3 | 59.5 0.6 |
| FSL +KNN – 5 neighbors | 58.4–67.2% | 61.3% 3.2% | 25.3–36.1 | 29.6 4.1 | 66.1–70.2 | 68.2 1.6 |
| FSL + KNN – 20 neighbors | 61.4–69.9% | 64.6% 2.9% | 26.7–36.5 | 30.9 4.0 | 70.1–76.6 | 74.1 2.3 |
| FSL + SVM with linear kernel | 62.9–70.0% | 65.3% 2.5% | 28.2–36.7 | 32.0 3.2 | 77.7–85.8 | 82.7 3.0 |
| FSL + SVM with polynomial kernel | 59.7–67.2% | 62.9% 2.8% | 23.4–34.0 | 28.7 4.3 | 74.7–84.9 | 80.1 3.7 |
| FSL + SVM with RBF kernel | 63.3–70.6% | 65.8% 2.5% | 27.8–37.2 | 31.4 3.9 | 77.1–82.5 | 80.5 2.2 |
| FSL + SVM with Sigmoid kernel | 56.7–65.4% | 59.8% 3.2% | 16.6–28.2 | 23.7 4.6 | 77.0–82.5 | 80.9 2.0 |

Table 7. Results for 20-class classification, part II.

| Technique | Specificity | Specificity Avg Std | Sensitivity | Sensitivity Avg Std |
|----------------------------------|-------------|-----------------------|-------------|-----------------------|
| FSL proximity-based | 24.9–26.8% | 25.6% 0.6% | 27.7–29.7% | 28.5% 0.7% |
| Softmax-based classification | 36.3–39.7% | 37.6% 1.2% | 32.2–33.1% | 32.6% 0.3% |
| FSL + XGBoost | 23.8–31.6% | 27.7% 2.6% | 25.3–34.5% | 29.7% 3.4% |
| FSL + Random Forest | 26.6–32.3% | 28.4% 2.2% | 27.1–36.7% | 31.0% 3.6% |
| FSL + Decision Tree | 19.4–28.5% | 24.8% 3.7% | 19.9–30.0% | 25.1% 4.2% |
| FSL +KNN – 5 neighbors | 24.0–33.7% | 28.3% 3.3% | 25.3–36.1% | 29.6% 4.1% |
| FSL + KNN – 20 neighbors | 24.9–31.1% | 28.2% 2.0% | 26.7–36.5% | 30.9% 4.0% |
| FSL + SVM with linear kernel | 25.6–34.1% | 28.8% 3.1% | 28.2–36.7% | 32.0% 3.2% |
| FSL + SVM with polynomial kernel | 26.0–33.0% | 29.1% 2.4% | 23.4–34.0% | 28.7% 4.3% |
| FSL + SVM with RBF kernel | 25.5–31.1% | 28.9% 2.7% | 27.8–37.2% | 31.4% 3.9% |
| FSL + SVM with Sigmoid kernel | 15.5–25.4% | 20.8% 3.4% | 16.6–28.2% | 23.7% 4.6% |

**Figure 6.** Comparison of average accuracy of evaluated models on 2, 5, and 20 classes detection.

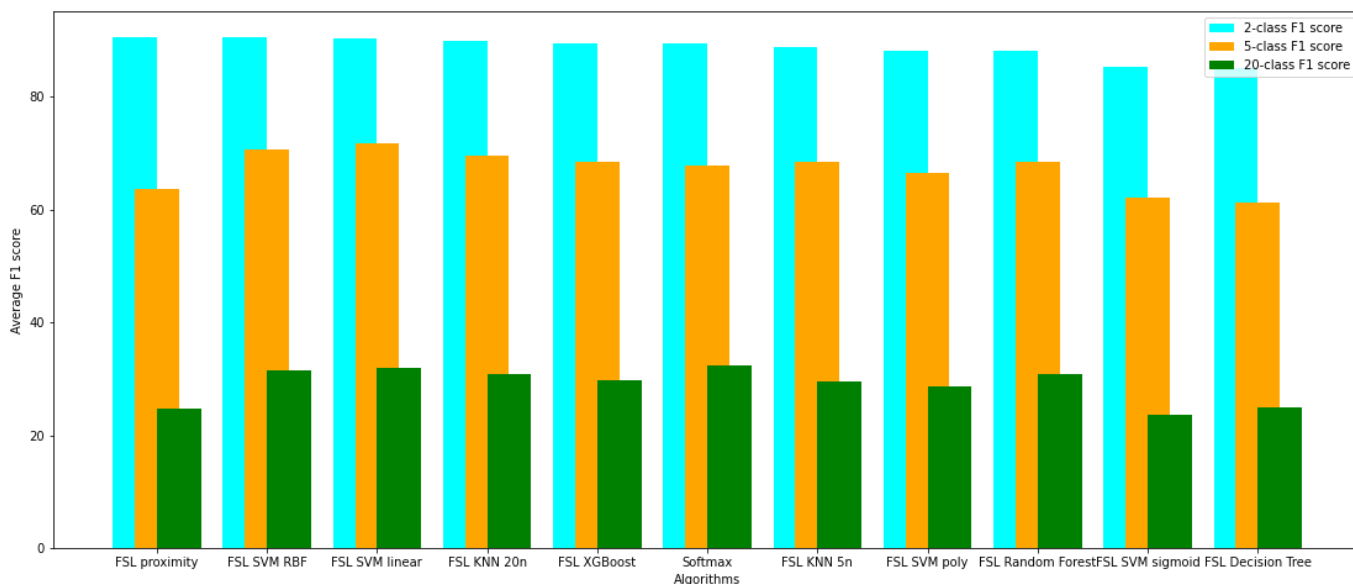


Figure 7. Comparison of average F1 score of evaluated models on 2, 5, and 20 classes detection.

The influence of the dataset size on the FSL classification has been examined. During this evaluation, the Random Forest algorithm was used to classify few-shot encoded signals. The results are depicted in Figure 8, which shows the relationship between the size of the dataset used and the accuracy obtained during test evaluation. The sizes of the datasets evaluated are 1%, 5%, 10%, 50%, and 100% of the size of original test dataset.

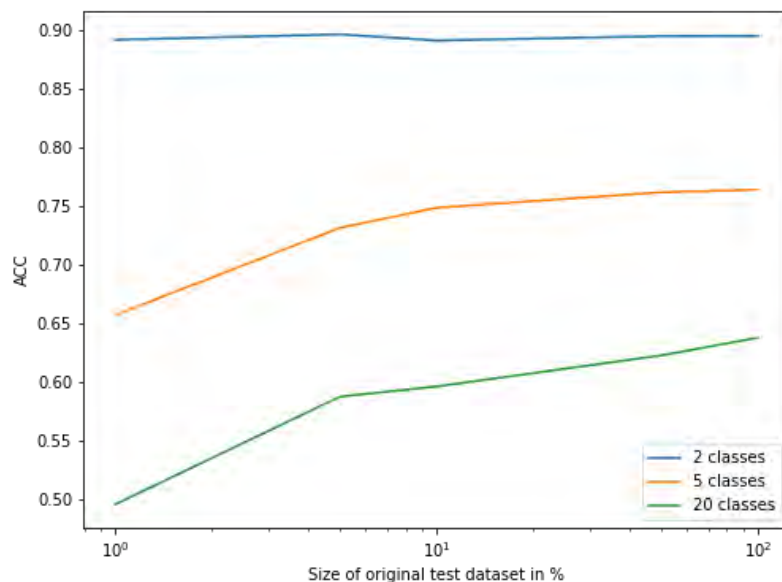


Figure 8. ACC as a function of the size of the original test dataset.

Figures 9–14 present the confusion matrices from the evaluation on one of the test datasets composed for k-fold cross validation conduction purposes. The Figures 15–20 depict the accuracy on the training and validation datasets during the training process.

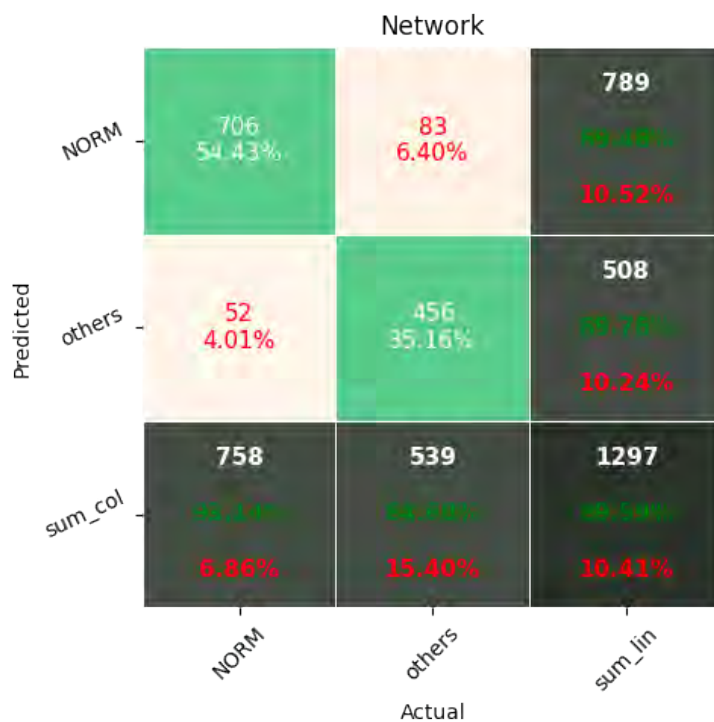


Figure 9. Confusion Matrix for Few-Shot (2 classes) with proximity-based classification.

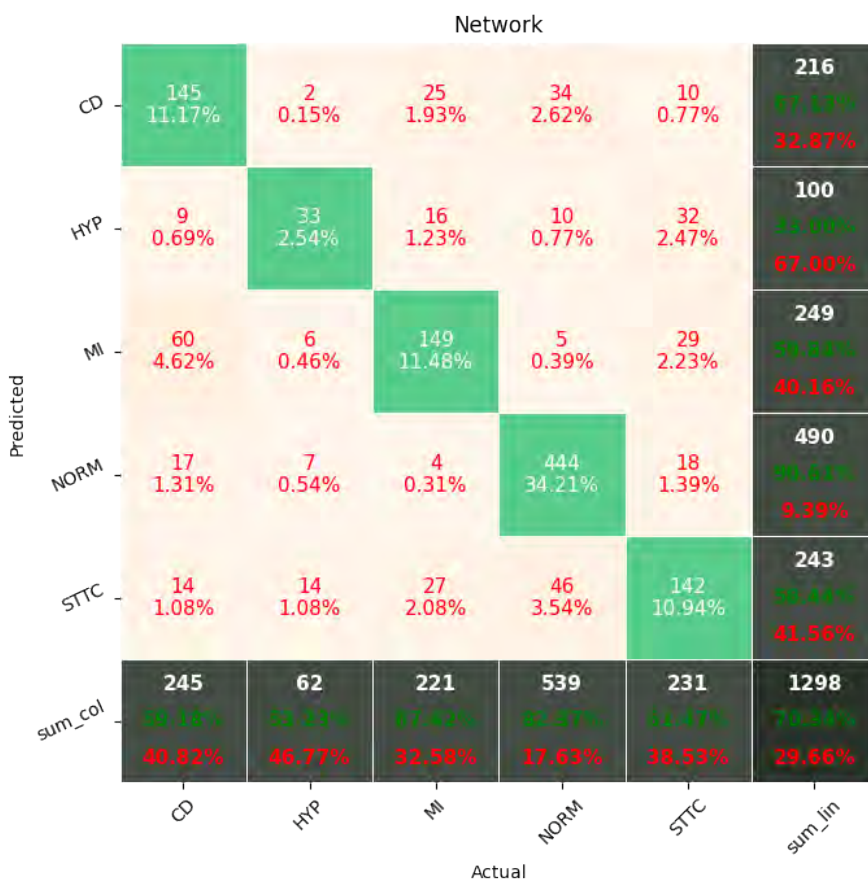


Figure 10. Confusion Matrix for Few-Shot (5 classes) with proximity-based classification.

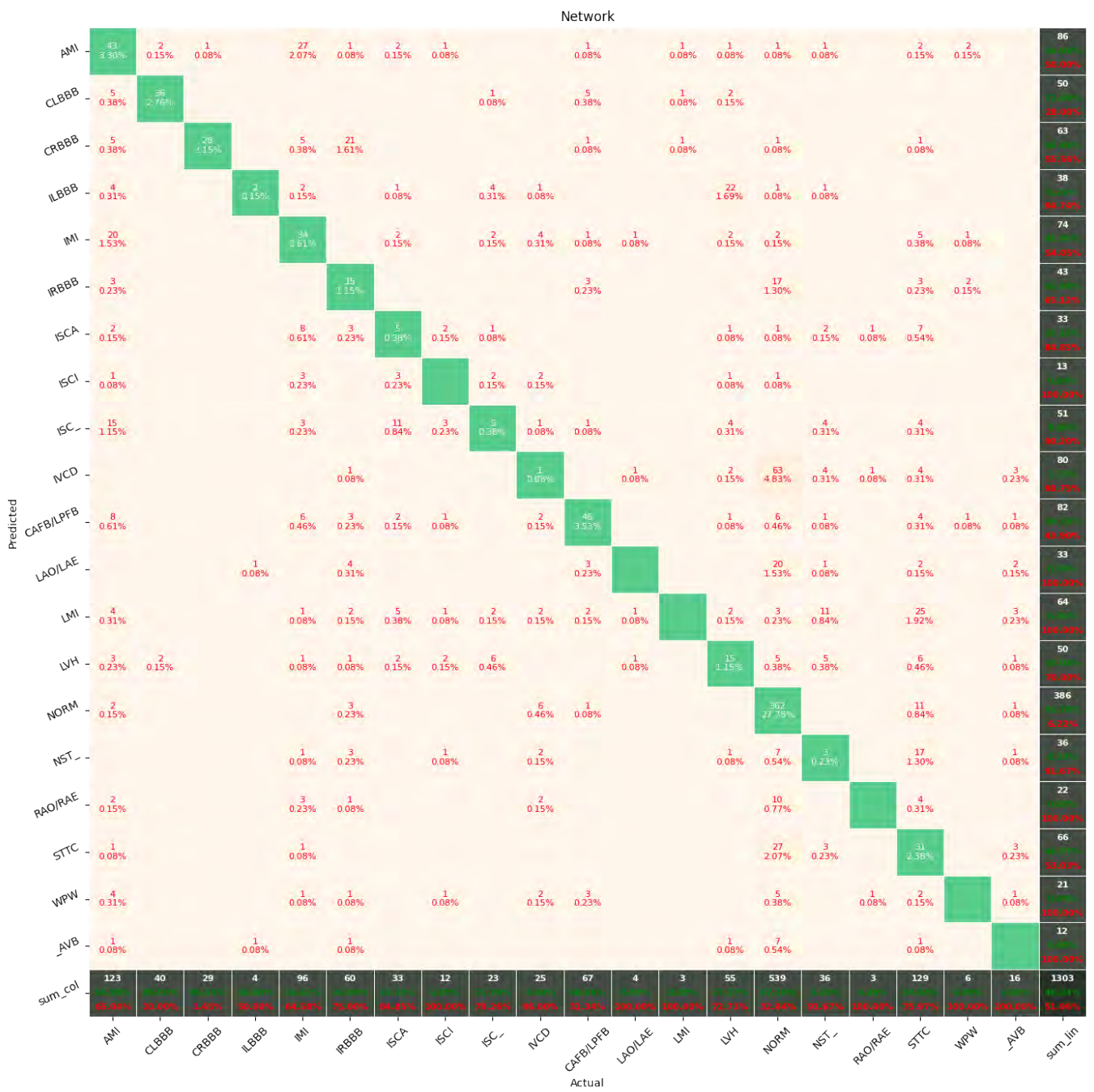


Figure 11. Confusion Matrix for Few-Shot (20 classes) with proximity-based classification.

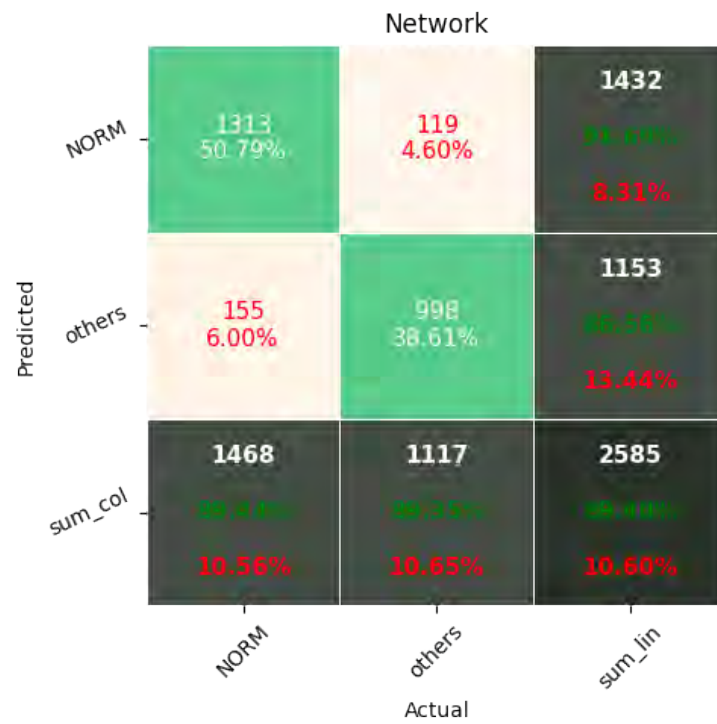


Figure 12. Confusion Matrix for softmax-based classification (2 classes).



Figure 13. Confusion Matrix for softmax-based classification (5 classes).

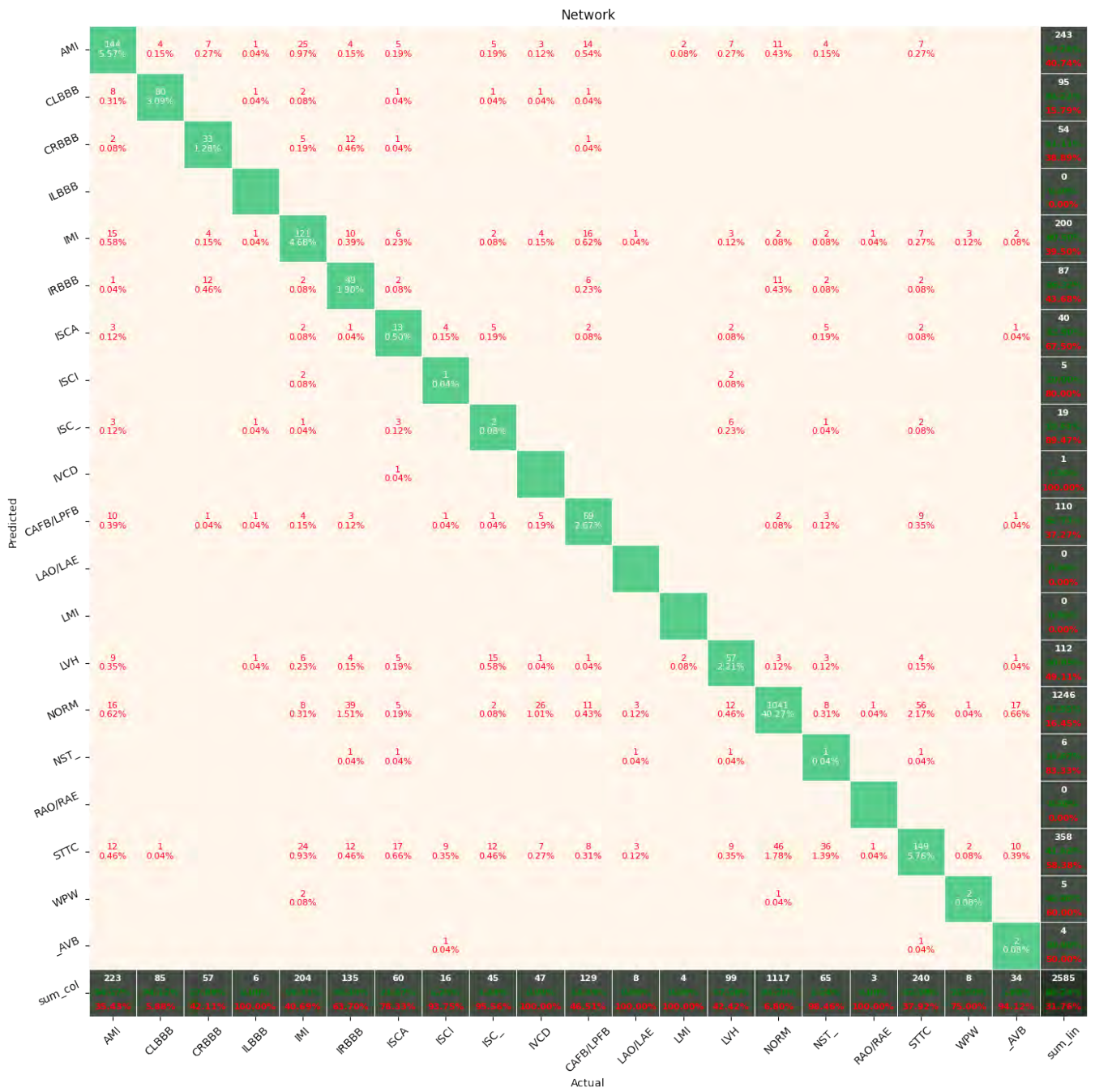


Figure 14. Confusion Matrix for softmax-based classification (20 classes).

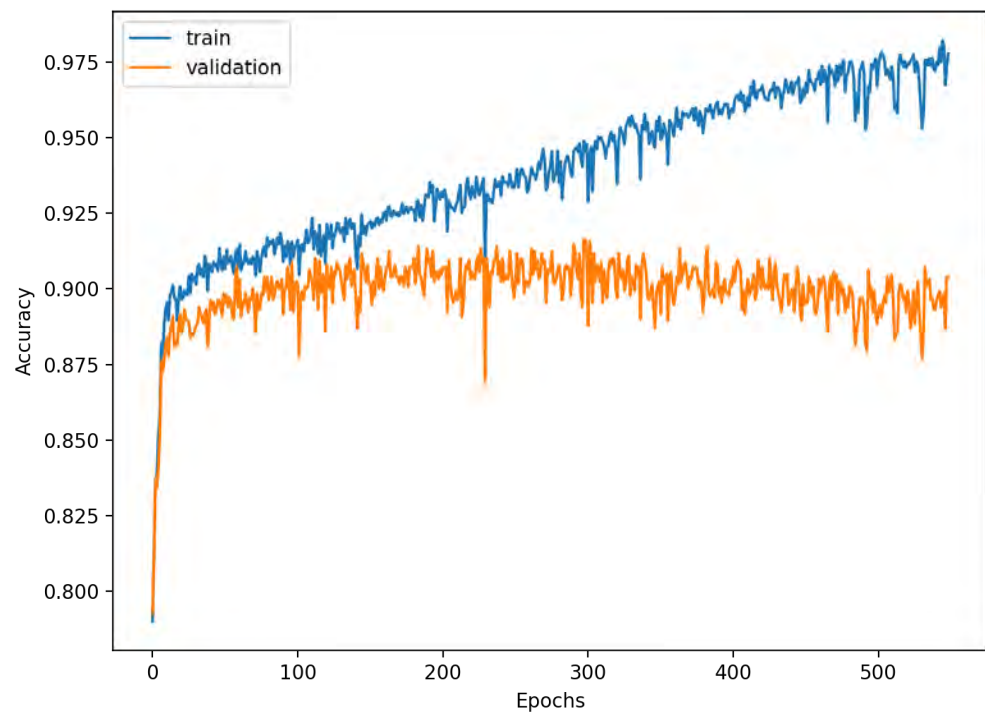


Figure 15. Learning process of the Neural Network for Few-Shot (2 classes) with proximity-based classification.

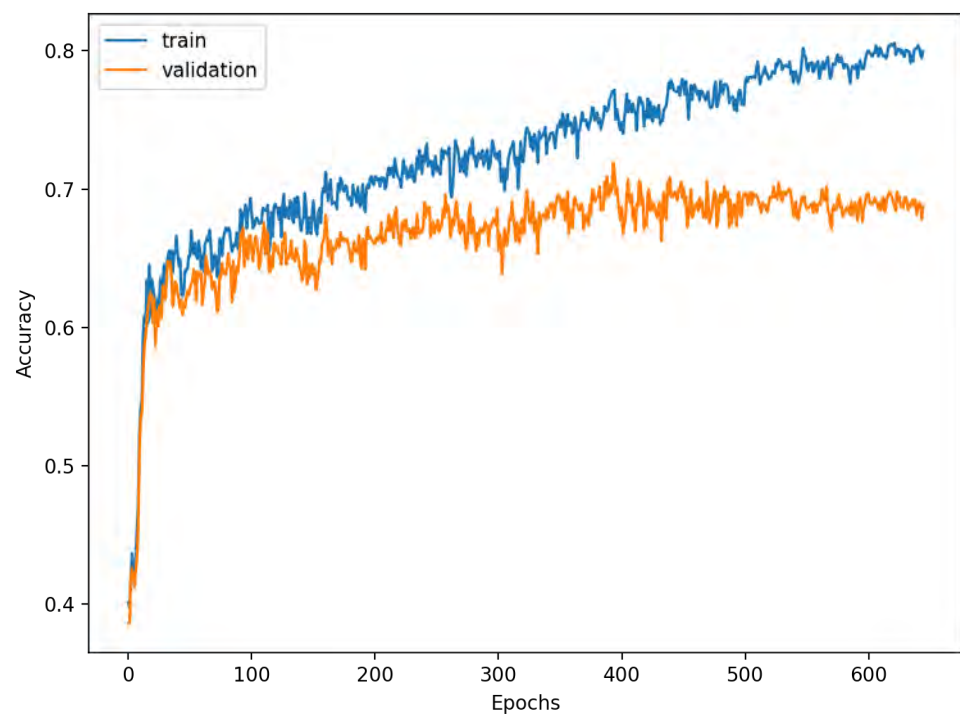


Figure 16. Learning process of the Neural Network for Few-Shot (5 classes) with proximity-based classification.

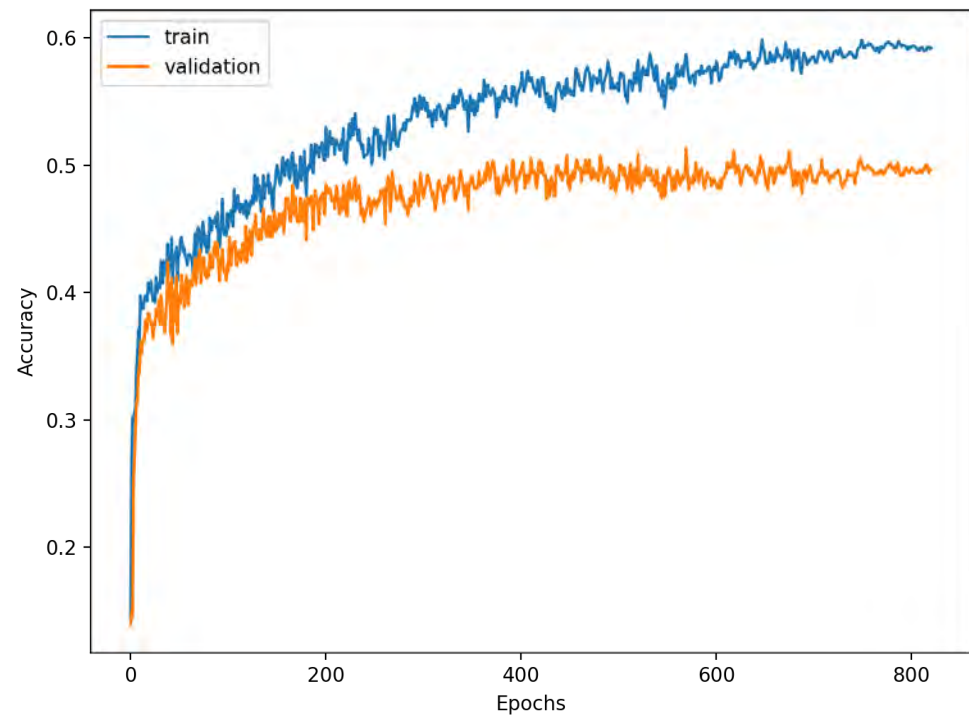


Figure 17. Learning process of the Neural Network for Few-Shot (20 classes) with proximity-based classification.

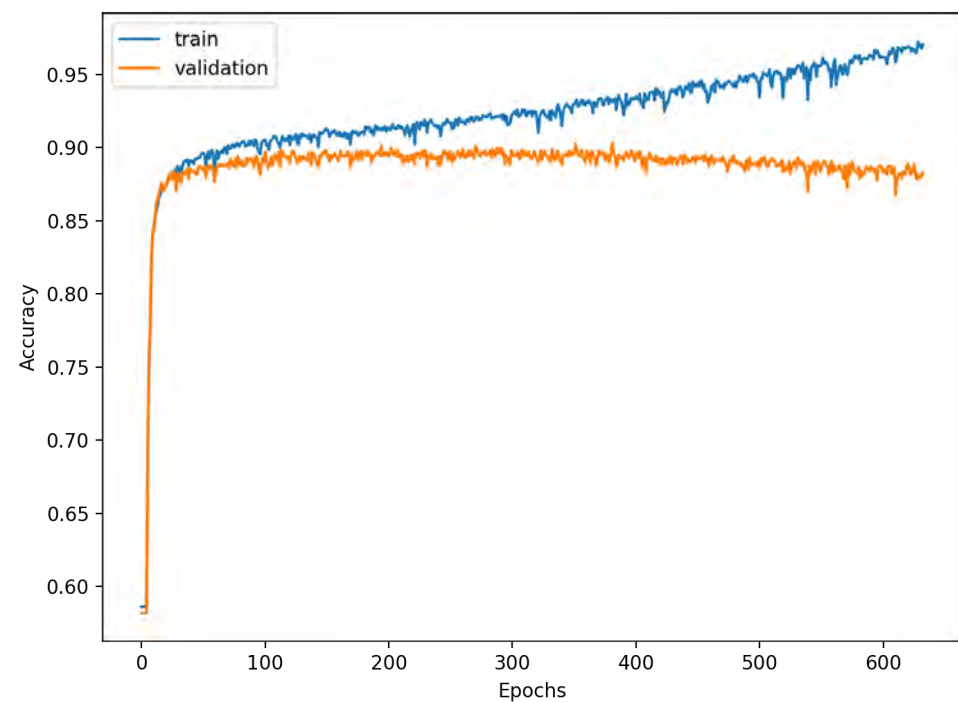


Figure 18. Learning process of the Neural Network for softmax-based classification (2 classes).

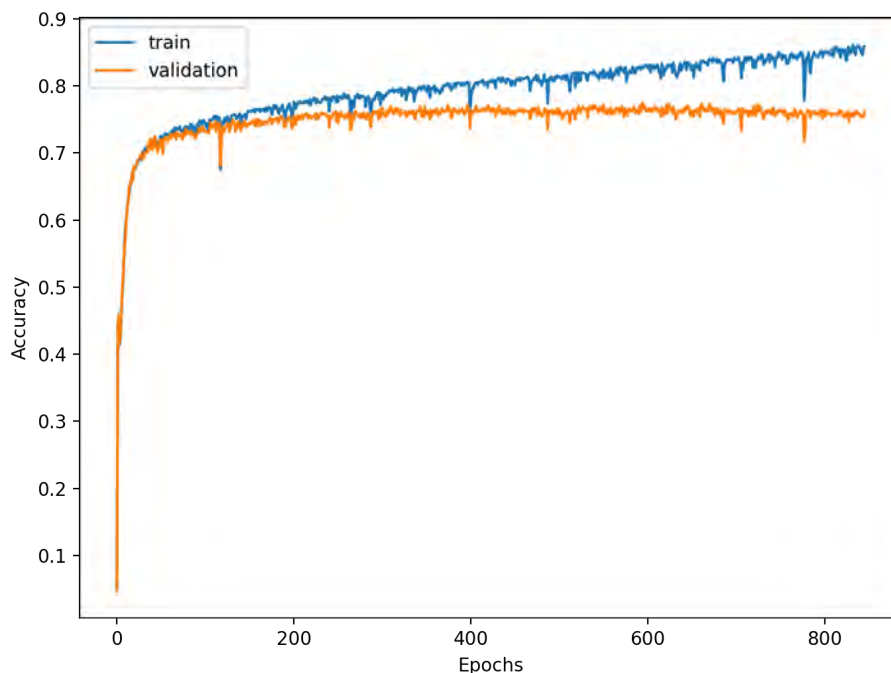


Figure 19. Learning process of the Neural Network for softmax-based classification (5 classes).

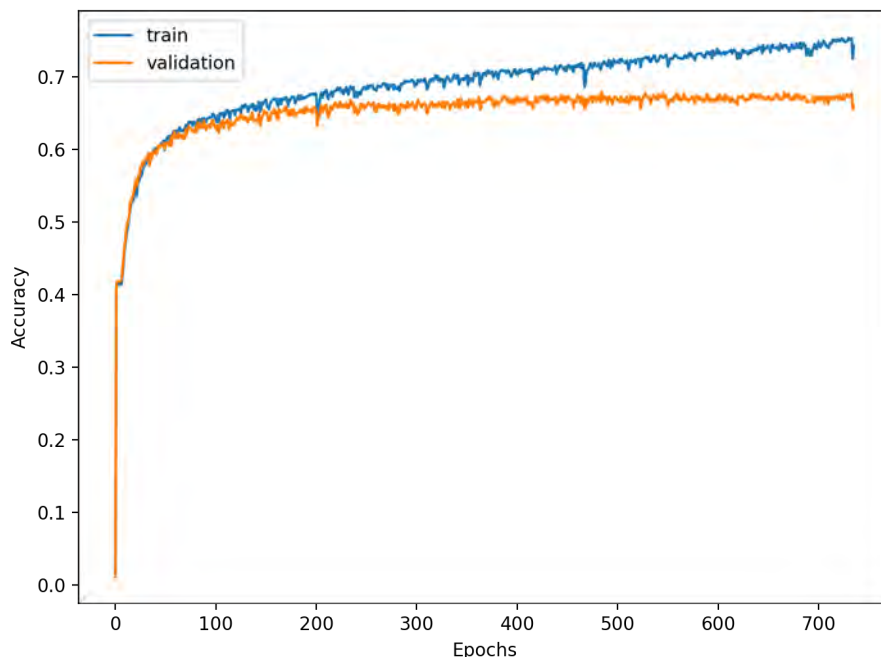


Figure 20. Learning process of the Neural Network for softmax-based classification (20 classes).

4. Discussion

The Deep Neural Network trained in a Few-Shot learning (FSL) fashion for proximity-based classification provides the benefit of improved accuracy through an embedded version of online learning, allowing for continuous classification augmentation without network weight adjustments. The network’s accuracy can be improved without the additional optimization of its weights through the expansion of the classified signals dataset. Such a set is used for referential class vector computation and is essential for the correct signal classification. Cardiological professionals can improve the network by labeling the signal and increasing the number of vectors used for class vector calculation, resulting in

better classification. Such a procedure does not require training of the network, which is cumbersome on production machines due to the higher computation complexity of the training network than using an already trained one. This augmentation procedure can be conducted on a CPU with low computation capabilities due to the simplicity of mean vector calculation.

The Few-Shot Learning neural network proved to be more accurate than the softmax-based network while classifying two classes. The FSL model had higher results in both averages, maximal and minimal accuracy. However, the network proved to be less accurate on tasks involving 5- and 20-class labeling. This phenomenon is most likely a result of insufficient representation of classes with low cardinality. For example: In Figure 11, the class "NORM" having the highest number of ECG records had the best precision and recall of all classes. The authors plan a further examination of the dataset size's influence on the quality of prediction.

This work classified the signals processed by an FSL neural network by computing the average vector representing each class and comparing the Euclidean distance between the classified sample and all class-representing vectors. The other methods evaluated in this work for classification use network-encoded signals in small-sized vectors to train models running algorithms such as XGBoost, Random Forest, Decision Tree, K-Nearest Neighbors, and SVMs with linear, polynomial, radial basis function, and sigmoid kernels. It turned out that the most promising classification algorithm for FSL in this particular task is SVM with a radial basis function kernel. This method proved to be the most effective among all the examined FSL classification strategies and achieved better results than softmax-based classification for both two and five classes. It achieved one of the highest scores in accuracy, specificity, sensitivity, F1, and AUC among all compared models. The outcomes are promising and suggest that the hybrid neural network systems based on proximity-differentiation classification with integrated machine learning models may provide better results than the typical softmax-based state-of-the-art classification. The authors plan on conducting further research to determine whether a combination of FSL with SVM with radial basis function kernel is beneficial in other tasks or merely the case in this particular example.

The accuracy of the FSL network during the training process varies significantly more than its softmax-based counterpart. This phenomenon is depicted in Figures 15–20. The softmax-based classification network reaches convergence faster and is less susceptible to the noise generated by the random selection of training data. This variance of the learning process is essential because of the commonness of early-stopping usage during network training. Typical early-stopping implemented in DL frameworks such as Keras stops the training if the evaluation score of the trained network on the validation dataset was not improved in a specific amount of time. This mechanism is important as it reduces the amount of wasted computation time and energy. However, due to the high variance of the FSL process, it is possible that controlling early-stopping based on local extremum may not be the best strategy. The results indicate that filtration of evaluation score's signal, such as averaging, may prove beneficial. The authors plan on further examination in future works.

In previous work [16], the best-obtained result in that research classifying sick/healthy patients (2 classes) is 89.2% accuracy. This value was increased in this research by the FSL neural network, the accuracy of which spans from 89.5% to 91.1%. As a result, even the worst performance of the studied network was better than the best in previous work. However, the results were not as promising during the classification of 5 and 20 classes. It is speculated that FSL can obtain better results for bigger datasets than Softmax-based classification, but the latter requires less training data than the former. The authors plan on conducting further research of this phenomenon.

The dataset size had almost no influence on the classification performance of the two classes. However, its impact was significant for the classification of 5 classes and even more important for the classification of 20 classes. It turns out that the more that classes are differentiated from each other, the more data are required.

5. Conclusions

The neural network trained for conducting Few-Shot Learning classification tasks proved to be more accurate than the softmax-based classification network when classifying signals using 2 and 5 labels but obtained worse results on 20 classes with fewer samples per class. In this experiment, the most efficient method for performing classification using the FSL network for signal encoding is the SVM model with an RBF kernel. Such networks can be successfully applied in systems that provide feedback from experts and data accumulation such as hospitals. The network can be improved without optimizing the network parameters in this environment, which requires high-end processing units such as GPUs. A proposed online learning strategy can be conducted on typical industrial CPUs. The FSL networks may prove beneficial as they allow for their performance to be improved after their rollout.

Author Contributions: Conceptualization, K.P. and S.Š.; methodology, K.P., S.Š. and D.L.; software, K.P., S.Š. and D.L.; validation, K.P., S.Š. and D.L.; formal analysis, K.P.; investigation, K.P. and S.Š.; resources, K.P., S.Š., D.L. and S.B.; data curation, S.B.; writing—original draft preparation, K.P. and S.Š.; writing—review and editing, K.P. and S.Š.; visualization, K.P., S.Š. and D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

| | |
|-----------------|---|
| ECG | Electrocardiogram |
| EEG | Electroencephalography |
| CT | Computed Tomography |
| QRS complex | Combination of three of the graphical deflections (Q wave, R wave, and S wave) seen on a typical ECG record. It represents an electrical impulse spreading through the ventricles of the heart and indicating their depolarization |
| Conv1d | Layer in Deep Neural Networks that performs a convolution on one-dimensional signal |
| MaxPool1d | Layer in Deep Neural Networks that performs pooling operation by selecting a maximum value from the moving window |
| Fully Connected | Layer in Deep Neural Networks that consists of neurons that process whole input data |
| Leaky ReLU | Activation function used in Deep Neural Networks |
| Padding | Parameter used in convolutional layers specifying the amount of zeroed samples added to the start and end of the processed signal. For example: Padding of 1 means that there is one sample of value zero artificially added at the beginning and the end of the signal. This operation is conducted to mitigate activation map shrinkage due to the application of convolution |
| Stride | Parameter used in convolutional layers specifying shift distance between subsequent windows of convolutions. For example: A stride of 1 means that the next convolution starts right after the beginning of the previous one, so the windows will overlap (provided that kernel size is bigger than 1) |
| RBF | Radial Basis Function |

Appendix A. Descriptions of the Classes of Diseases

| | |
|------|------------------------|
| NORM | Normal ECG |
| CD | Myocardial Infarction |
| STTC | ST/T Change |
| MI | Conduction Disturbance |
| HYP | Hypertrophy |

Appendix B. Descriptions of the Subclasses of Diseases

| | |
|-----------|---|
| NORM | normal ECG |
| STTC | non-diagnostic T abnormalities, suggests digitalis-effect, long QT-interval, ST-T changes compatible with ventricular aneurysm, compatible with electrolyte abnormalities |
| AMI | anterior myocardial infarction, anterolateral myocardial infarction, in anteroseptal leads, in anterolateral leads, in lateral leads |
| IMI | inferior myocardial infarction, inferolateral myocardial infarction, inferoposterolateral myocardial infarction, inferoposterior myocardial infarction, in inferior leads, in inferolateral leads |
| LAFB/LPFB | left anterior fascicular block, left posterior fascicular block |
| IRBBB | incomplete right bundle branch block |
| LVH | left ventricular hypertrophy |
| CLBBB | (complete) left bundle branch block |
| NST_ | non-specific ST changes |
| ISCA | in anterolateral leads, in anteroseptal leads, in lateral leads, in anterior leads |
| CRBBB | (complete) right bundle branch block |
| IVCD | non-specific intraventricular conduction disturbance |
| ISC_ | ischemic ST-T changes |
| _AVB | first degree AV block, second degree AV block, third degree AV block |
| ISCI | in inferior leads, in inferolateral leads |

References

- Roshani, S.; Jamshidi, M.B.; Mohebi, F.; Roshani, S. Design and Modeling of a Compact Power Divider with Squared Resonators Using Artificial Intelligence. *Wirel. Pers. Commun.* **2021**, *117*, 2085–2096. [[CrossRef](#)]
- Nazemi, B.; Rafiean, M. Forecasting house prices in Iran using GMDH. *Int. J. Hous. Mark. Anal.* **2020**, *14*, 555–568. [[CrossRef](#)]
- Roshani, M.; Sattari, M.A.; Ali, P.J.M.; Roshani, G.H.; Nazemi, B.; Corniani, E.; Nazemi, E. Application of GMDH neural network technique to improve measuring precision of a simplified photon attenuation based two-phase flowmeter. *Flow Meas. Instrum.* **2020**, *75*, 101804. [[CrossRef](#)]
- Narwariya, J.; Malhotra, P.; Vig, L.; Shroff, G.; Vishnu, T.V. Meta-learning for few-shot time series classification. In Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, Hyderabad, India, 5–7 January 2020.
- Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* **2018**, *8*, 1, 1–12. [[CrossRef](#)]
- Rajpurkar, P.; Hannun, A.Y.; Haghpanahi, M.; Bourn, C.; Ng, A.Y. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv* **2017**, arXiv:1707.01836.
- Mahajan, R.; Kamaleswaran, R.; Howe, J.A.; Akbilgic, O. Cardiac rhythm classification from a short single lead ECG recording via random forest. In Proceedings of the 2017 Computing in Cardiology (CinC), Rennes, France, 24–27 September 2017.
- Yang, J.; Nguyen, M.N.; San, P.P.; Li, X.L.; Krishnaswamy, S. Deep convolutional neural networks on multichannel time series for human activity recognition. In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015.
- Rizwan, A.; Zoha, A.; Mabrouk, I. B.; Sabbour, H. M.; Al-Sumaiti, A. S.; Alomainy, A.; Abbasi, Q. H. A review on the state of the art in atrial fibrillation detection enabled by machine learning. *IEEE Rev. Biomed. Eng.* **2020**, *14*, 219–239. [[CrossRef](#)]
- Bizopoulos, P.; Koutsouris, D. Deep learning in cardiology. *IEEE Rev. Biomed. Eng.* **2018**, *12*, 168–193. [[CrossRef](#)] [[PubMed](#)]
- Chandra, B.S.; Sastry, C.S.; Jana, S.; Patidar, S. Atrial fibrillation detection using convolutional neural networks. In Proceedings of the 2017 Computing in Cardiology (CinC), Rennes, France, 24–27 September 2017.

12. Rundo, F.; Conoci, S.; Ortis, A.; Battiato, S. An advanced bio-inspired photoplethysmography (PPG) and ECG pattern recognition system for medical assessment. *Sensors* **2018**, *18*, 405. [[CrossRef](#)]
13. Karim, F.; Majumdar, S.; Darabi, H.; Chen, S. LSTM fully convolutional networks for time series classification. *IEEE Access* **2017**, *6*, 1662–1669. [[CrossRef](#)]
14. Fawaz, H.I.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.A. Deep learning for time series classification: A review. *Data Min. Knowl. Discov.* **2019**, *33*, 917–963. [[CrossRef](#)]
15. Kashiparekh, K.; Narwariya, J.; Malhotra, P.; Vig, L.; Shroff, G. ConvTimeNet: A pre-trained deep convolutional neural network for time series classification. In Proceedings of the 2019 International Joint Conference on Neural Networks, Budapest, Hungary, 14–19 July 2019.
16. Śmigiel, S.; Pałczyński, K.; Ledziński, D. ECG Signal Classification Using Deep Learning Techniques Based on the PTB-XL Dataset. *Entropy* **2021**, *23*, 1121. [[CrossRef](#)] [[PubMed](#)]
17. Benjamin, E.J.; Blaha, M.J.; Chiuve, S.E.; Cushman, M.; Das, S.R.; Deo, R.; Muntner, P. Heart disease and stroke statistics—2017 update: A report from the American Heart Association. *Circulation* **2017**, *135*, e146–e603. [[CrossRef](#)] [[PubMed](#)]
18. Shenasa, M. Learning and teaching electrocardiography in the 21st century: A neglected art. *J. Electrocardiol.* **2018**, *51*, 357–562. [[CrossRef](#)] [[PubMed](#)]
19. Cai, W.; Hu, D. QRS complex detection using novel deep learning neural networks. *IEEE Access* **2020**, *8*, 97082–97089. [[CrossRef](#)]
20. Rashkovska, A.; Depolli, M.; Tomašić, I.; Avbelj, V.; Trobec, R. Medical-grade ECG sensor for long-term monitoring. *Sensors* **2020**, *20*, 6. [[CrossRef](#)]
21. Šarlija, M.; Jurišić, F.; Popović, S. A convolutional neural network based approach to QRS detection. In Proceedings of the 10th International Symposium on Image and Signal Processing and Analysis, Ljubljana, Slovenia, 18–20 September 2017.
22. Zhong, W.; Liao, L.; Guo, X.; Wang, G. A deep learning approach for fetal QRS complex detection. *Physiol. Meas.* **2018**, *39*, 045004. [[CrossRef](#)]
23. Xiang, Y.; Lin, Z.; Meng, J. Automatic QRS complex detection using two-level convolutional neural network. *Biomed. Eng. Online* **2018**, *17*, 1–17. [[CrossRef](#)]
24. Belkadi, M.A.; Daamouche, A.; Melgani, F. A deep neural network approach to QRS detection using autoencoders. *Expert Syst. Appl.* **2021**, *184*, 115528. [[CrossRef](#)]
25. Guo, Z.; Wang, Y.; Liu, L.; Sun, S.; Feng, B.; Zhao, X. Siamese Network-Based Few-Shot Learning for Classification of Human Peripheral Blood Leukocyte. In Proceedings of the 2021 IEEE 4th International Conference on Electronic Information and Communication Technology (ICEICT), Xi'an, China, 18–20 August 2021; pp. 818–822.
26. Voulodimos, A.; Protopapadakis, E.; Katsamenis, I.; Doulamis, A.; Doulamis, N. A Few-Shot U-Net Deep Learning Model for COVID-19 Infected Area Segmentation in CT Images. *Sensors* **2021**, *21*, 2215. [[CrossRef](#)]
27. Lai, Y.; Li, G.; Wu, D.; Lian, W.; Li, C.; Tian, J.; Jiang, G. 2019 Novel coronavirus-infected pneumonia on CT: A feasibility study of few-shot learning for computerized diagnosis of emergency diseases. *IEEE Access* **2020**, *8*, 194158–194165. [[CrossRef](#)]
28. Szűcs, G.; Németh, M. Double-View Matching Network for Few-Shot Learning to Classify Covid-19 in X-ray images. *Infocommun. J.* **2021**, *13*, 26–34. [[CrossRef](#)]
29. Prabhu, V.; Kannan, A.; Ravuri, M.; Chaplain, M.; Sontag, D.; Amatriain, X. Few-shot learning for dermatological disease diagnosis. In Proceedings of the Machine Learning for Healthcare Conference, Ann Arbor, MI, USA, 8–10 August 2019.
30. Xiao, J.; Xu, H.; Zhao, W.; Cheng, C.; Gao, H. A Prior-mask-guided Few-shot Learning for Skin Lesion Segmentation. *Computing* **2021**, 1–23. [[CrossRef](#)]
31. Ma, J.; Fong, S.H.; Luo, Y.; Bakkenist, C.J.; Shen, J.P.; Mourragui, S.; Ideker, T. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat. Cancer* **2021**, *2*, 233–244. [[CrossRef](#)] [[PubMed](#)]
32. An, S.; Kim, S.; Chikontwe, P.; Park, S.H. Few-shot relation learning with attention for EEG-based motor imagery classification. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October–24 January 2020.
33. Liu, T.; Yang, Y.; Fan, W.; Wu, C. Few-shot learning for cardiac arrhythmia detection based on electrocardiogram data from wearable devices. *Digit. Signal Process.* **2021**, *116*, 103094. [[CrossRef](#)]
34. Wagner, P.; Strodthoff, N.; Boussejot, R.; Samek, W.; Schaeffter, T. PTB-XL, a large publicly available electrocardiography dataset (version 1.0.1). *Sci. Data* **2020**, *7*, 1–5. [[CrossRef](#)] [[PubMed](#)]
35. Goldberger, A.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E.; et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220. [[CrossRef](#)]
36. Hamilton, P.S. *Open Source ECG Analysis Software Documentation*; E.P. Limited: Somerville, MA, USA, 2002.
37. Elgendi, M.; Jonkman, M.; De Boer, F. Frequency Bands Effects on QRS Detection. In Proceedings of the 3rd International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS2010), Valencia, Spain, 20–23 January 2010; pp. 428–431.
38. Kalidas, V.; Tami, L. Real-time QRS detector using Stationary Wavelet Transform for Automated ECG Analysis. In Proceedings of the 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), Washington, DC, USA, 23–25 October 2017.
39. Christov, I. Real time electrocardiogram QRS detection using combined adaptive threshold. *Biomed. Eng. Online* **2004**, *3*, 28. [[CrossRef](#)]

40. Pan, J.; Tompkins W. J. A Real-Time QRS Detection Algorithm. *IEEE Trans. Biomed. Eng.* **1985**, *BME-32*, 230–236. [[CrossRef](#)]
41. Zeelenberg, C. A single scan algorithm for QRS detection and feature extraction. *IEEE Comp. Cardiol.* **1979**, *6*, 37–42.
42. Lourenco, A.; Silva, H.; Leite, P.; Lourenco R.; Fred, A. Real Time Electrocardiogram Segmentation for Finger Based ECG Biometrics. *Biosignals* **2012**, 49–54.
43. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *arXiv* **2014**, arXiv:1409.4842
44. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
45. Caruana, R.; Lawrence, S.; Giles, L. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In Proceedings of the 14th Annual Neural Information Processing Systems Conference, Denver, CO, USA, 27 November–2 December 2000; pp. 402–408.
46. Ha, M.L.; Blanz, V. Deep Ranking with Adaptive Margin Triplet Loss. *arXiv* **2021**, arXiv:2107.06187.