



**UNIWERSYTET
MIKOŁAJA KOPERNIKA
W TORUNIU**

Wydział Farmaceutyczny
Collegium Medicum w Bydgoszczy

**Zastosowanie algorytmów uczenia maszynowego
w analizie predykcyjnej aktywności przeciwdrobnoustrojowej
pochodnych imidazoliowych i wznowy raka płuca**

Rozprawa doktorska

mgr Andżelika Lorenc

Rozprawa napisana pod kierunkiem:

prof. dr. hab. n. farm. Adama Bucińskiego

dr hab. n. med. Anny Romaszko-Wojtowicz

*Niniejsza praca nie powstałaby, gdyby nie ludzie,
którzy towarzyszyli mi na mojej naukowej ścieżce.*

Dlatego składam serdeczne podziękowania:

Mojemu promotorowi i szefowi, prof. dr. hab. n. farm. Adamowi Bucińskiemu

*- za wspieranie mojego rozwoju, zaufanie i za to,
że dostrzegł we mnie potencjał, zanim ja sama zdążyłam go zauważyć.*

Mojej promotor, dr hab. n. med. Annie Romaszko-Wojtowicz, prof. WUM

*- za pokazanie, czym jest interdyscyplinarność w nauce, cenne wskazówki
i nieustanną gotowość do dzielenia się swoim doświadczeniem.*

Koleżankom z Katedry Biofarmacji - za życzliwość, wsparcie i przypominanie,

*że nawet doktorat da się przetrwać dzięki kawie,
poczuciu humoru i ludziom, na których można liczyć.*

Szczególnie jednak dziękuję Moim Rodzicom, Siostrze, Dziadkom i Przyjaciółom

*- za wiarę we mnie, nawet wtedy, gdy sami nie do końca rozumieli,
czym właściwie się zajmuję.*

Dziękuję, że dmuchaliście mi w żagle i nie pozwoliliście mi zwątpić, że dam radę.

Spis treści

Wykaz skrótów i akronimów	4
Streszczenie	6
Abstract.....	7
1. Wstęp.....	8
1.1. Typy uczenia maszynowego.....	10
1.2. Wybrane algorytmy uczenia maszynowego	11
1.2.1. Drzewa decyzyjne, CART	11
1.2.2. Analiza Składowych Głównych, PCA.....	13
1.2.3. Sztuczne sieci neuronowe, ANN	15
2. Cel pracy.....	26
3. Materiały i metody	28
3.1. Metody badawcze	30
4. Wyniki	35
4.1. Publikacja nr 1	35
4.2. Publikacja nr 2	37
4.3. Publikacja nr 3	41
5. Podsumowanie i wnioski.....	44
Literatura.....	47
Spis rycin	51
Artykuły stanowiące zbiór publikacji	52
Artykuł nr 1	52
Artykuł nr 1 - materiał dodatkowy.....	60
Artykuł nr 2	65
Artykuł nr 2 - materiał dodatkowy.....	74
Artykuł nr 3	74
Artykuł nr 3 – zgoda Komisji Bioetycznej	98
Oświadczenie – Komisja Bioetyczna	100
Opis merytoryczny udziału w publikacjach.....	101
Oświadczenia współautorów	104
Życiorys naukowy.....	112
Wykaz publikacji	116

Wykaz skrótów i akronimów

ACC - dokładność, *ang. Accuracy*

ACoS – Amerykańskie Kolegium Chirurgów, *ang. American College of Surgeons*

AI - sztuczna inteligencja, *ang. Artificial Intelligence*

ANN - sztuczna sieć neuronowa, *ang. Artificial Neural Network*

AUC - pole pod krzywą ROC, *ang. Area Under the Curve*

BFGS - algorytm Broydena-Fletcher-Goldfarba-Shanno, *ang. Broyden-Fletcher-Goldfarb-Shanno algorithm*

CART - drzewa klasyfikacyjne i regresyjne, *ang. Classification and Regression Trees*

CG - gradienty sprzężone, *ang. Conjugate Gradients*

DDAC - chlorek dialkilodimetyloamoniowy, *ang. Didecyldimethylammonium Chloride*

DFT - teoria funkcjonału gęstości, *ang. Density Functional Theory*

DL - głębokie uczenie, *ang. Deep Learning*

FA - analiza czynnikowa, *ang. Factor Analysis*

FDR - odsetek wyników fałszywie dodatnich wśród pozytywnych, *ang. False Discovery Rate*

FN - fałszywie ujemny, *ang. False Negative*

FP - fałszywie dodatni, *ang. False Positive*

FPR - odsetek wyników fałszywie dodatnich, *ang. False Positive Rate*

GenAI - generatywna sztuczna inteligencja, *ang. Generative Artificial Intelligence*

GLM - uogólniony model liniowy, *ang. Generalized Linear Model*

LLM - duży model językowy, *ang. Large Language Model*

lnMIC - logarytm naturalny minimalnego stężenia hamującego, *ang. natural logarithm of Minimal Inhibitory Concentration*

MAE - średni błąd bezwzględny, *ang. Mean Absolute Error*

MCC - współczynnik korelacji Matthews, *ang. Matthews Correlation Coefficient*

MIC - minimalne stężenie hamujące, *ang. Minimal Inhibitory Concentration*

ML - uczenie maszynowe, *ang. Machine Learning*

MLP - perceptron wielowarstwowy, *ang. Multilayer Perceptron*

MSE - średni błąd kwadratowy, *ang. Mean Squared Error*

NAACCR - Północnoamerykańskie Stowarzyszenie Centralnych Rejestrów Nowotworów, *ang. North American Association of Central Cancer Registries*

NCI - Narodowy Instytut Raka, *ang. National Cancer Institute*

PC - składowa główna, *ang. Principal Component*

PCA - analiza składowych głównych, *ang. Principal Component Analysis*
PCM - model kontinuum polaryzowalnego, *ang. Polarizable Continuum Model*
PPV - wartość predykcyjna dodatnia, *ang. Positive Predictive Value*
 R^2 - współczynnik determinacji, *ang. Coefficient of Determination*
RMSE - pierwiastek średniego błędu kwadratowego, *ang. Root Mean Squared Error*
ROC - krzywa charakterystyki operacyjnej odbiornika, *ang. Receiver Operating Characteristic*
SEER - Program Nadzoru, Epidemiologii i Wyników Leczenia, *ang. Surveillance, Epidemiology, and End Results Program*
SOS - suma kwadratów, *ang. Sum of Squares*
SPC - swoistość, *ang. Specificity*
tanh - tangens hiperboliczny (funkcja aktywacji), *ang. Hyperbolic Tangent*
TN - prawdziwie ujemny, *ang. True Negative*
TP - prawdziwie dodatni, *ang. True Positive*
TPR - czułość, *ang. True Positive Rate*
XAI – wyjaśnialna sztuczna inteligencja, *ang. Explainable Artificial Intelligence*

Streszczenie

Dynamiczny rozwój metod uczenia maszynowego stwarza nowe możliwości w zakresie modelowania złożonych zjawisk biologicznych i klinicznych. Celem niniejszej rozprawy doktorskiej była ocena przydatności sztucznych sieci neuronowych typu perceptron wielowarstwowy w modelowaniu problemów medycznych dwójakiej natury – aktywności przeciwdrobnoustrojowej związków imidazoliowych oraz w predykcji wznowy raka płuca na podstawie danych klinicznych. Rozprawa obejmuje trzy powiązane tematycznie publikacje, prezentujące rozwój metodologiczny od klasyfikacji binarnej, poprzez regresję wielowymiarową, aż do modelowania klinicznego.

W pierwszej z prezentowanych prac opracowano model perceptronu wielowarstwowego, wspartego metodą drzew decyzyjnych oraz analizą składowych głównych, celem klasyfikacji aktywności przeciwdrobnoustrojowej 140 nowo syntetyzowanych pochodnych imidazoliowych wobec *Klebsiella pneumoniae*. Na podstawie deskryptorów molekularnych opisujących cząsteczki zbudowano model umożliwiający skuteczne rozróżnienie związków aktywnych i nieaktywnych, co potwierdziły wysokie wartości miar klasyfikacyjnych.

Drugie badanie dotyczyło predykcji wartości minimalnego stężenia hamującego dla 126 pochodnych imidazoliowych wobec *Staphylococcus aureus* oraz *Candida albicans*. Zastosowano podejście hybrydowe obejmujące selekcję predyktorów z wykorzystaniem drzew decyzyjnych, redukcję wymiarowości metodą analizy głównych składowych oraz modelowanie przy użyciu perceptronu wielowarstwowego. Model umożliwił jednoczesną predykcję wartości lnMIC dla obu drobnoustrojów jednocześnie, potwierdzając skuteczność integracji algorytmów uczenia maszynowego również w złożonych problemach natury regresyjnej.

Trzecia praca obejmowała retrospektywną analizę danych klinicznych 2296 pacjentów z rakiem płuca hospitalizowanych w latach 2012-2017. Opracowany model perceptronu wielowarstwowego uczony był dwuetapowo z wykorzystaniem algorytmu wstecznej propagacji błędów oraz gradientów sprzężonych. Model osiągnął wysoką swoistość i pozwolił zidentyfikować najważniejsze czynniki predykcyjne wznowy, jednak jego czułość oceniono jako umiarkowaną.

Uzyskane w pracy wyniki potwierdzają, że sztuczne sieci neuronowe stanowią skuteczne narzędzie modelowania zarówno danych chemometrycznych, jak i danych klinicznych, mogąc wspierać projektowanie nowych związków biologicznie aktywnych oraz proces podejmowania decyzji medycznych.

Abstract

The dynamic development of machine learning methods creates new opportunities for modeling complex biological and clinical phenomena. The objective of this doctoral dissertation was to evaluate the utility of Multilayer Perceptron (MLP) artificial neural networks in modeling two distinct medical problems: the antimicrobial activity of imidazolium compounds and the prediction of lung cancer recurrence based on clinical data. The dissertation comprises three thematically related publications, demonstrating a methodological evolution from binary classification to multidimensional regression and, ultimately, clinical modeling.

In the first study, an MLP model was developed, supported by decision tree methods and Principal Component Analysis (PCA), to classify the antimicrobial activity of 140 newly synthesized imidazolium derivatives against *Klebsiella pneumoniae*. Based on molecular descriptors determined via 2D and 3D methods, a model was constructed that successfully distinguished between active and inactive compounds, as confirmed by high classification performance metrics.

The second study focused on predicting the Minimum Inhibitory Concentration (MIC) values for 126 imidazolium derivatives against *Staphylococcus aureus* and *Candida albicans*. A hybrid approach was employed, incorporating predictor selection via decision trees, dimensionality reduction using PCA, and modeling with a Multilayer Perceptron. The model enabled the simultaneous prediction of $\ln\text{MIC}$ values for both microorganisms, confirming the efficacy of integrating machine learning algorithms even in complex regression-based problems.

The third study involved a retrospective analysis of clinical data from 2,296 lung cancer patients hospitalized between 2012 and 2017. An MLP model was developed and trained in two stages using the backpropagation and conjugated gradient algorithms. While the model achieved high specificity and identified the most significant predictors of recurrence, its sensitivity was assessed as moderate.

The results obtained in this dissertation confirm that artificial neural networks are an effective tool for modeling both chemometric and clinical data. They can significantly support the design of new biologically active compounds and enhance the medical decision-making process.

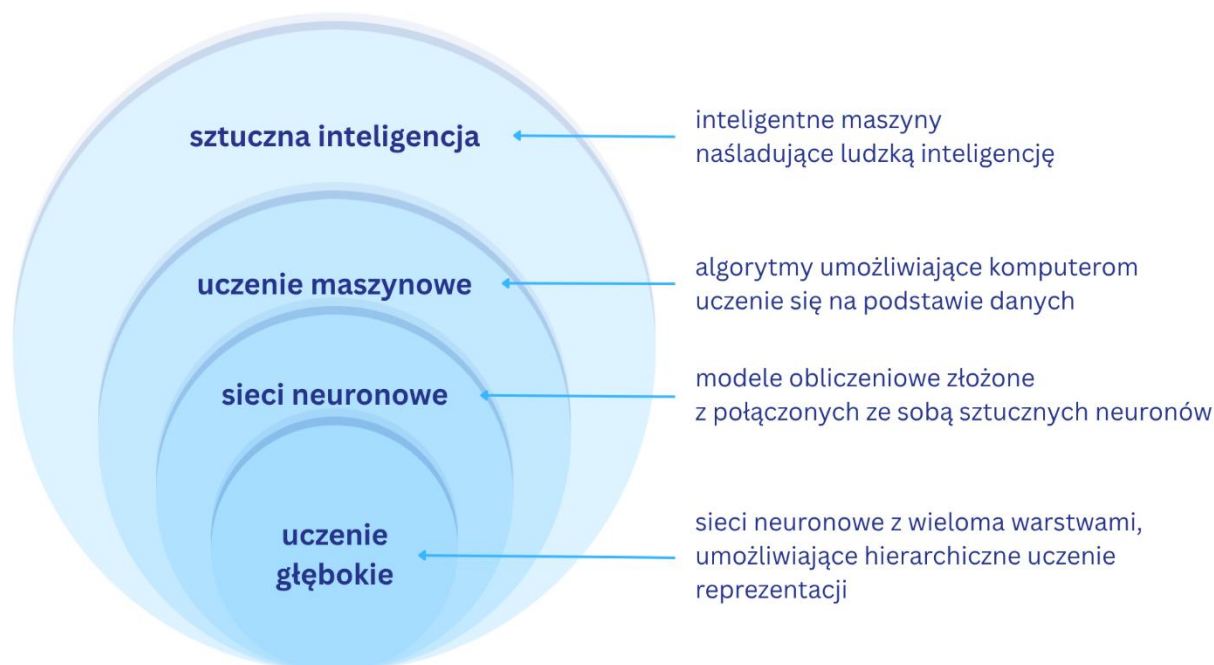
1. Wstęp

Pojęcia sztucznej inteligencji (ang. *Artificial Intelligence*, AI) oraz uczenia maszynowego (ang. *Machine Learning*, ML) jeszcze do niedawna funkcjonowały głównie w dwóch odrębnych przestrzeniach – w środowisku wysoko wyspecjalizowanych badań naukowych oraz w popkulturze, która nadawała im często uproszczone i nierzadko zniekształcone znaczenie. W ostatnich latach obserwuje się jednak wyraźne zbliżenie tych obszarów, a zagadnienia związane z AI stały się elementem szerokiego dyskursu społecznego i technologicznego, obejmującego zarówno debatę publiczną, jak i dyskusję nad rolą sztucznej inteligencji w różnych obszarach życia [1,2]. Należy jednak podkreślić, że za powszechnie używanym terminem „sztuczna inteligencja” kryją się złożone modele obliczeniowe, oparte na metodach statystycznych, rachunku prawdopodobieństwa oraz zaawansowanych operacjach matematycznych [3]. To właśnie te mechanizmy umożliwiają systemom komputerowym analizę danych, podejmowanie decyzji oraz formułowanie predykcji w oparciu o proces uczenia się na podstawie dostępnych informacji.

Historia przewidywania zjawisk sięga głęboko w czasy nowożytne, czego przykładem jest XVI-wieczna myśl Paracelsusa, który w swoich pracach postulował istnienie deterministycznych powiązań między stanem natury a losem człowieka, tym samym kładąc podwaliny pod wczesne prognozowanie zdarzeń medycznych [4]. Przejście do predykcji opartej na rygorystycznej, ilościowej analizie danych masowych przypisuje się Johnowi Grauntowi, który w XVII wieku sformalizował metody statystyczne w demografii [5]. To jego prace stanowią fundament współczesnych metod ML, znajdujących zastosowanie w wielu obszarach nauki i praktyki gospodarczej, w tym w medycynie, naukach przyrodniczych i społecznych, finansach, marketingu, zarządzaniu operacjami oraz analizie procesów biznesowych [6]. Choć metody oraz cele poszczególnych dziedzin potrafią się znacząco różnić, podstawowe mechanizmy leżące u podstaw modeli predykcyjnych pozostają w dużej mierze wspólne i opierają się na tych samych zasadach matematycznych oraz statystycznych [7]

Termin *machine learning* został wprowadzony w 1959 roku przez Arthura Samuela, jednego z pionierów sztucznej inteligencji [8]. Pomimo ponad sześćdziesięciu lat rozwoju tej dziedziny oraz upowszechnienia się pojęcia AI w przestrzeni publicznej, współcześnie jest ono często utożsamiane przede wszystkim z dużymi modelami językowymi (ang. *Large Language Models*, LLM), a tym samym z generatywną sztuczną inteligencją (ang. *Generative Artificial Intelligence*, GenAI). W rzeczywistości jednak AI stanowi znacznie szerszy obszar,

obejmujący różnorodne metody ML oraz uczenia głębokiego (ang. *deep learning*, DL) – od klasycznych modeli regresyjnych, klasyfikacyjnych i klasteryzacyjnych, poprzez drzewa decyzyjne, aż po złożone architektury głębokich sieci neuronowych (**Rycina 1**).



Rycina 1. Relacje hierarchiczne między sztuczną inteligencją (AI), uczeniem maszynowym (ML), sieciami neuronowymi oraz uczeniem głębokim (DL). Opracowanie własne na podstawie: Goodfellow i in. [9].

Pojęcie ML odnosi się do algorytmów przetwarzania danych, które na podstawie obserwowanych wzorców uczą się zależności między zmiennymi oraz dostosowują swoje parametry w celu minimalizacji błędu predykcji. Wysokiej jakości modele charakteryzują się również zdolnością do generalizacji, co oznacza, że są w stanie poprawnie przewidywać wyniki dla danych, które nie były wcześniej wykorzystywane w procesie uczenia [10]. Cecha ta jest szczególnie istotna, ponieważ warunkuje praktyczną użyteczność modelu – jedynie modele zachowujące wysoką jakość predykcji dla nowych danych mogą być adaptowane w rozwiązywaniu rzeczywistych problemów i wspomaganiu procesów decyzyjnych. Brak zdolności do generalizacji prowadzi do nadmiernego dopasowania do danych treningowych, co znacząco ogranicza wiarygodność oraz wartość translacyjną opracowanego rozwiązania [11].

Szczególnie interesującym, ze względu na złożoność oraz dynamiczny charakter procesów biologicznych zachodzących zarówno w stanie zdrowia, jak i choroby,

wyduje się wykorzystanie ML w kontekście nauk medycznych i farmaceutycznych. W obszarach tych dodatkowym wyzwaniem jest istotna zmienność międzyosobnicza – te same jednostki chorobowe mogą manifestować się odmiennym obrazem klinicznym u różnych pacjentów, co wynika z różnic genetycznych, molekularnych i środowiskowych [7]. Zdolność metod ML do identyfikowania złożonych, również nieliniowych zależności w danych, a w przypadku sztucznych sieci neuronowych (ang. Artificial Neural Networks, ANNs) – do aproksymacji funkcji o niemal dowolnie wysokim stopniu złożoności – sprawia, że stanowią one obiecujące narzędzie do modelowania tak wielowymiarowych układów biologicznych.

1.1. Typy uczenia maszynowego

Jak wspomniano wcześniej, ML to szeroki zbiór metod obliczeniowych o potencjale modelowania danych o wysokiej wymiarowości i zdolnych do generalizacji wiedzy na podstawie danych uczących. Aby jednak model ML był zdolny do generowania trafnych predykcji, konieczne jest przeprowadzenie iteracyjnego procesu uczenia. W zależności od tego w jaki sposób przebiega uczenie, algorytmy ML dzielone są na nadzorowane (z nauczycielem, ang. *supervised*) lub nienadzorowane (bez nauczyciela, ang. *unsupervised*).

Uczenie nadzorowane opiera się na wykorzystaniu zbioru danych zawierającego zarówno zmienne wejściowe (opisujące analizowane zjawisko), jak i odpowiadające im zmienne wyjściowe (stanowiące prawidłowe odpowiedzi). W trakcie procesu uczenia algorytm modyfikuje parametry w celu minimalizacji różnicy pomiędzy wartościami przewidywanymi a rzeczywistymi, zgodnie z przyjętą funkcją straty. Wśród metod nadzorowanych wyróżniane są klasyczne metody klasyfikacyjne oraz regresyjne, takie jak np. drzewa klasyfikacyjne i regresyjne (ang. *Classification and Regression Trees*, CART), jak również niektóre rodzaje sztucznych sieci neuronowych (ang. *Artificial Neural Networks*, ANNs) – np. sieci typu perceptron wielowarstwowy (ang. *Multilayer Perceptron*, MLP) [12].

W przypadku, gdy znajomość odpowiedzi na zadane pytanie nie jest konieczna do opracowania rozwiązania, stosuje się metody uczenia nienadzorowanego. Model samodzielnie odkrywa strukturę danych poprzez poszukiwanie ukrytych w nich wzorców i zależności. Do tej grupy należą między innymi metody redukcji wymiarowości, takie jak analiza głównych składowych (ang. *Principal Component Analysis*, PCA), klasyczne metody liniowe, takie jak analiza czynnikowa (ang. *Factor Analysis*, FA), a także podejścia oparte na sieciach neuronowych [13].

Niezależnie od przyjętego paradygmatu uczenia – nadzorowanego lub nienadzorowanego – konstrukcja modelu ML wymaga świadomego określenia celu analizy, wysokojakościowego zbioru danych uczących oraz doboru odpowiedniej metody modelowania [11,13].

1.2. Wybrane algorytmy uczenia maszynowego

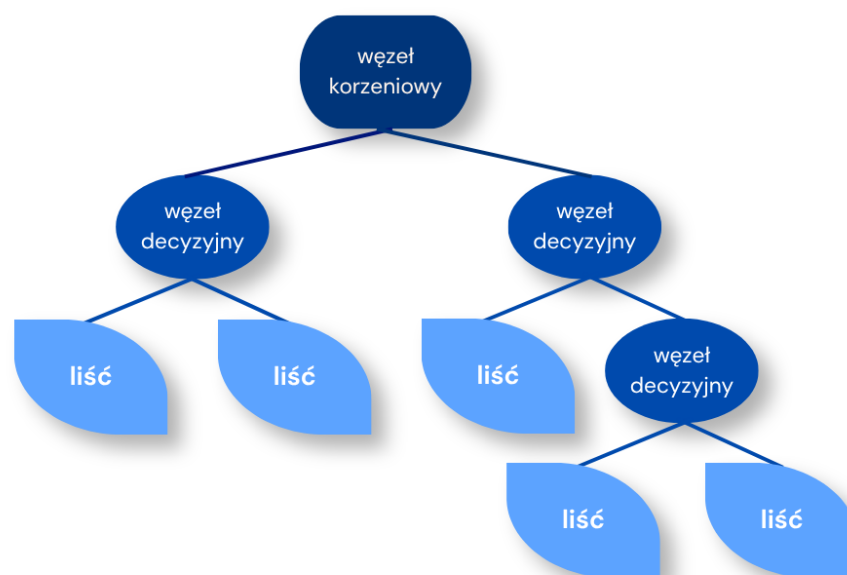
Po przedstawieniu ogólnych założeń metod uczenia maszynowego zasadne jest skoncentrowanie się na algorytmach wykorzystanych w niniejszej pracy. W dalszej części omówione zostaną metody reprezentujące różne podejścia do modelowania danych – drzewa decyzyjne CART, techniki redukcji wymiarowości w postaci PCA oraz sztuczne sieci neuronowe typu MLP. Wybór tych algorytmów wynika z ich komplementarnego charakteru oraz przydatności w analizie danych biomedycznych o wysokiej złożoności i wielowymiarowości.

W tym miejscu należy również zwrócić uwagę, że pojęcie algorytmu uczenia w metodach ML funkcjonuje na dwóch poziomach – wysokopoziomowo – jako rodzina modeli (np. CART, PCA, ANN) oraz niskopoziomowo – jako mechanizm określający, w jaki sposób parametry modelu są uczone (np. algorytm wstecznej propagacji błędów w ANN) [10].

1.2.1. Drzewa decyzyjne, CART

Jak wspomniano wcześniej, drzewa decyzyjne należą do nadzorowanych metod uczenia maszynowego. Algorytm ten został zaprezentowany przez Breimana i współpracowników w 1984 roku [14]. Jego nazwa wynika z charakterystycznej struktury modelu, przypominającej odwrócone drzewo – z węzłem korzeniowym (ang. *root node*) na szczycie, węzłami decyzyjnymi (ang. *decision nodes*) w części pośredniej oraz węzłami końcowymi, zwanymi liśćmi (ang. *terminal nodes*), na dole struktury (**Rycina 2**).

Zasada działania drzewa decyzyjnego polega na rekurencyjnym podziale zbioru danych rozpoczynającym się w węźle korzeniowym, który obejmuje całą analizowaną populację. Podział dokonywany jest na podstawie określonego kryterium optymalizacyjnego, a powstałe węzły potomne podlegają kolejnym podziałom według tej samej zasady. Proces ten kontynuowany jest do momentu spełnienia przyjętych kryteriów zatrzymania, prowadząc do utworzenia struktury reguł decyzyjnych typu „if-then”, umożliwiających klasyfikację lub predykcję analizowanych przypadków [15,16].



Rycina 2. Schemat struktury drzewa decyzyjnego przedstawiający węzeł korzeniowy, węzły decyzyjne oraz liście końcowe. Węzeł korzeniowy inicjuje proces podziału danych, węzły decyzyjne odpowiadają za kolejne rozgałęzienia na podstawie wartości zmiennych, natomiast liście reprezentują ostateczne decyzje klasyfikacyjne lub wartości predykcyjne.

Drzewa decyzyjne wykorzystywane są zarówno w problemach klasyfikacyjnych, jak i regresyjnych. W zadaniach klasyfikacyjnych ich celem jest skonstruowanie struktury podziałów umożliwiającej przypisanie obiektu do jednej z określonych kategorii na podstawie wartości zmiennych objaśniających. W przypadku problemów regresyjnych mechanizm podziału pozostaje analogiczny, jednak zmienna zależna ma charakter ciągły, a w liściach drzewa estymowana jest wartość numeryczna. Zastosowanie algorytmu CART w analizie regresyjnej umożliwia zatem identyfikację zmiennych stanowiących istotne predyktory badanej cechy oraz modelowanie zależności między nimi a zmienną odpowiedzi [17].

Istotną zaletą CART jest fakt, że jest to metoda nieparametryczna i nieliniowa, co oznacza, że jej architektura nie wymaga przyjmowania wstępnych założeń dotyczących liniowości, monotoniczności ani określonej postaci funkcji łączącej pomiędzy predyktorami a zmienną zależną, jak ma to miejsce w uogólnionych modelach liniowych (ang. *Generalized Linear Models*, GLM) [18]. W związku z powyższym algorytm CART jest szczególnie przydatny w analizie eksploracyjnej danych, zwłaszcza w sytuacji ograniczonej wiedzy *a priori* dotyczącej charakteru zależności między zmiennymi.

Ponadto, drzewa decyzyjne charakteryzują się wysokim poziomem interpretowalności zarówno uzyskanych wyników, jak i samej struktury modelu. Struktura drzewa, powstała w wyniku kolejnych podziałów danych, dostarcza informacji nie tylko o przypisanej klasie lub przewidywanej wartości zmiennej zależnej, lecz również – dzięki jednoznacznie zdefiniowanym warunkom podziału – pozwala prześledzić logiczną ścieżkę prowadzącą do danej decyzji [19]. Z praktycznego punktu widzenia drzewa decyzyjne, oparte na sekwencji reguł typu „if-then”, są bardziej przejrzyste niż modele opisane wyłącznie za pomocą miar klasyfikacyjnych, czy równań matematycznych [18].

W niniejszej pracy algorytm CART został wykorzystany jako narzędzie selekcji cech w procesie modelowania aktywności przeciwdrobnoustrojowej analizowanych homologów imidazoliowych. Dzięki swojej strukturze hierarchicznych podziałów umożliwił identyfikację zmiennych o największej mocy predykcyjnej oraz redukcję wysokowymiarowej przestrzeni deskryptorów molekularnych. Uzyskany w ten sposób zestaw najbardziej istotnych dla rozwiązywanego problemu predyktorów stanowił podstawę do dalszego etapu modelowania z wykorzystaniem metod redukcji wymiarowości oraz sztucznych sieci neuronowych.

1.2.2. Analiza Składowych Głównych, PCA

Analiza składowych głównych stanowi metodę uczenia nienadzorowanego umożliwiającą zarówno identyfikację struktury danych i zależności między zmiennymi, jak i redukcję ich wymiarowości, co ma szczególne znaczenie w analizie zbiorów wysokowymiarowych.

PCA jest metodą projekcyjną polegającą na ortogonalnym przekształceniu pierwotnego układu współrzędnych do nowej przestrzeni, w której kolejne składowe główne stanowią liniowe kombinacje zmiennych wejściowych [20]. Wartości współczynników określających udział poszczególnych zmiennych w danej składowej pozwalają zidentyfikować, które cechy w największym stopniu przyczyniają się do jej utworzenia. Nowe osie noszą nazwę składowych głównych (ang. *Principal Components*, PC) i są uporządkowane według malejącej wariancji, tak że pierwsza składowa wyjaśnia największą możliwą część zmienności danych, druga – największą część pozostałej zmienności i tak dalej. Wariancja poszczególnych PC odpowiada ich wartościom własnym, a suma wariancji wszystkich składowych jest równa całkowitej wariancji pierwotnego zbioru danych (w przypadku uwzględnienia wszystkich komponentów) [21].

W praktyce redukcja wymiarowości polega na zachowaniu jedynie tych składowych, które wyjaśniają istotną część całkowitej wariancji, co oznacza pozostawienie wyłącznie pierwszych kilku PC, które opisują największą część całkowitej zmienności zbioru. Oznacza to, że dane mogą zostać przedstawione w przestrzeni o mniejszej liczbie wymiarów przy zachowaniu istotnej części informacji zawartej w pierwotnym zbiorze [22]. Pomijanie kolejnych składowych wiąże się jednak z utratą części informacji, dlatego konieczne jest określenie akceptowalnego poziomu wyjaśnionej wariancji. W praktyce liczba zachowanych PC ustalana jest na podstawie łącznego procentu wyjaśnionej wariancji lub – w kontekście modelowania predykcyjnego – poprzez ocenę wpływu liczby składowych na jakość modelu w procedurze walidacyjnej [13]. Metoda PCA jest szczególnie efektywna w przypadku zmiennych pierwotnych wykazujących współzależności, ponieważ pozwala na koncentrację skorelowanej informacji w mniejszej liczbie nieskorelowanych składowych. W sytuacji, gdy zmienne wejściowe są w niewielkim stopniu skorelowane, potencjał redukcji wymiarowości jest ograniczony, gdyż każda z nich wnosi w dużej mierze niezależną informację. W takim przypadku pominięcie części składowych może prowadzić do utraty istotnych elementów zmienności danych [20].

Implikacje praktyczne zastosowania PCA są istotne zarówno z perspektywy obliczeniowej, jak i interpretacyjnej. Redukcja wysokowymiarowego zbioru danych do przestrzeni o mniejszej liczbie wymiarów umożliwia wykorzystanie składowych głównych jako syntetycznych reprezentacji zmiennych pierwotnych, przy zachowaniu istotnej części informacji zawartej w danych. Prowadzi to do zmniejszenia złożoności obliczeniowej modeli oraz ograniczenia zasobów niezbędnych do ich estymacji [23]. Z perspektywy interpretacyjnej analiza składowych głównych pozwala na identyfikację ukrytych struktur oraz współzależności między zmiennymi obserwowanymi, które w pierwotnej przestrzeni danych mogą być trudne do uchwycenia. Ma to szczególne znaczenie w naukach medycznych i farmaceutycznych, gdzie analizowane zbiory danych często charakteryzują się wysoką wymiarowością i znaczną liczbą potencjalnie współzależnych predyktorów [24,25]. W niniejszej pracy wybrane składowe główne, wyznaczone w procesie redukcji wymiarowości danych, zostały następnie wykorzystane jako dane wejściowe do budowy modelu predykcyjnego opartego na sztucznych sieciach neuronowych.

1.2.3. Sztuczne sieci neuronowe, ANN

Sztuczne sieci neuronowe stanowią algorytmy ML inspirowane strukturą biologicznych układów nerwowych. Pojęcie sztucznego neuronu, będącego podstawową jednostką budulcową i funkcjonalną ANN, po raz pierwszy pojawiło się w pracy McCulloch i Pitts *“A logical calculus of the ideas immanent in nervous activity”* z 1943 roku [26].

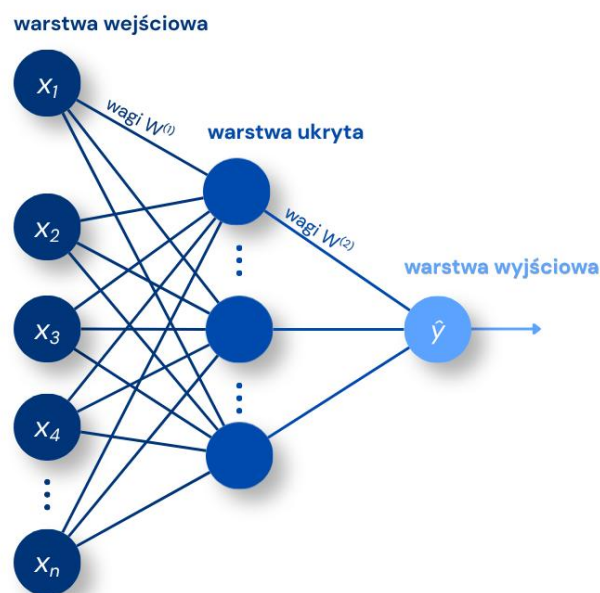
ANN składają się z warstw połączonych ze sobą sztucznych neuronów, w których przepływ informacji determinowany jest przez wagi połączeń oraz funkcje aktywacji, stanowiące część hiperparametrów sieci. Ze względu na swoją adaptacyjną strukturę ANN znajdują zastosowanie zarówno w problemach regresyjnych, jak i klasyfikacyjnych, szczególnie w sytuacjach, w których zależności między zmiennymi mają charakter nieliniowy i trudny do opisanego za pomocą modeli liniowych [27,28]. W związku z powyższym, ANN mają zdolność aproksymacji złożonych funkcji nieliniowych, co pozwala im modelować wielowymiarowe relacje między danymi wejściowymi a zmienną wyjściową [29]. Cecha ta wynika z iteracyjnego charakteru procesu uczenia sieci neuronowej – pełny zestaw danych treningowych prezentowany jest modelowi wielokrotnie, a przejście modelu przez wszystkie przypadki stanowiące zbiór treningowy nazywane jest epoką. Po zakończeniu każdej epoki uczenia model aktualizuje swoje hiperparametry w taki sposób, aby zminimalizować popełniany błąd predykcji. Odpowiednio dobrana architektura sieci, funkcje aktywacji oraz procedura optymalizacji parametrów umożliwiają minimalizację funkcji straty i dopasowanie modelu do obserwowanych danych na poziomie pozwalającym z powodzeniem implementować go do nowych danych.

Adaptowalność algorytmów ANN, będąca ich niewątpliwą zaletą, może prowadzić jednak do nadmiernego dopasowania modelu do danych. Warunkiem uzyskania modelu o wysokiej zdolności generalizacji jest dostępność wystarczająco licznej i reprezentatywnej zbioru uczącego [30]. Zbyt mała liczba obserwacji może prowadzić do nadmiernego dopasowania modelu do danych treningowych, określanego jako przeuczenie (ang. *overfitting*), co ogranicza jego użyteczność w predykcji nowych przypadków [31]. Warto również nadmienić, że obok wielkości zbioru danych stanowiącego podstawę treningową ANN, istotnym czynnikiem determinującym jakość modelu jest jakość danych wykorzystanych w procesie uczenia. W literaturze często przywoływana jest zasada „garbage in – garbage out”, podkreślająca, że model nie jest w stanie wygenerować wiarygodnych predykcji na podstawie danych obarczonych błędami, szumem lub systematycznymi zniekształceniami [32].

Jakość danych obejmuje nie tylko ich liczebność, lecz również reprezentatywność, kompletność, poprawność pomiarową oraz adekwatność względem modelowanego zjawiska. Nawet duży zbiór danych nie gwarantuje uzyskania modelu o wysokiej zdolności generalizacji, jeśli zawiera istotne błędy, liczne braki lub zmienne nieodzwierciedlające rzeczywistych mechanizmów rządzących danym zjawiskiem [33]. Z uwagi na fakt, że sieci neuronowe uczą się wyłącznie na podstawie dostarczonych im obserwacji, ich zdolność do poprawnego modelowania zależności jest bezpośrednio uzależniona od jakości i struktury danych wejściowych [29,34].

Jednym z typów ANN, zastosowanym w niniejszej pracy, jest jednokierunkowa, wielowarstwowa sieć neuronowa typu MLP [35], której architektura obejmuje warstwę wejściową, co najmniej jedną warstwę ukrytą oraz warstwę wyjściową (**Rycina 3**) [36]:

- Warstwa wejściowa (ang. *input layer*) składa się z neuronów odpowiadających liczbie zmiennych wejściowych wykorzystywanych w modelu.
- Warstwa ukryta (ang. *hidden layer*) stanowi zasadniczy element przetwarzania informacji – każdy neuron w warstwie ukrytej oblicza ważoną sumę sygnałów wejściowych, a następnie wynik ten poddawany jest nieliniowemu przekształceniu określone przez funkcję aktywacji. Liczba warstw oraz liczba neuronów w warstwach ukrytych stanowią hiperparametry modelu i dobierane są empirycznie w oparciu o jakość predykcji oraz procedury walidacyjne.
- Warstwa wyjściowa (ang. *output layer*) generuje ostateczną odpowiedź modelu, której charakter (np. wartość ciągła lub prawdopodobieństwo przynależności do klasy) zależy od typu rozwiązywanego problemu oraz zastosowanej funkcji aktywacji.



Rycina 3. Schemat architektury sztucznej sieci neuronowej typu MLP.

Siła sygnału przekazywanego do neuronu zależy od wartości sygnałów pochodzących z neuronów poprzedniej warstwy oraz od wag przypisanych do poszczególnych połączeń. Dla pojedynczego neuronu sygnał wejściowy przyjmuje postać liniowej kombinacji zmiennych wejściowych, w której każda zmienna mnożona jest przez odpowiadającą jej wagę, a następnie do uzyskanej sumy dodawany jest parametr przesunięcia (ang. *bias*). Wagi określają zarówno siłę, jak i kierunek wpływu danego sygnału wejściowego na aktywację neuronu w kolejnej warstwie [26]. Waga dodatnia wzmacnia wpływ sygnału, waga ujemna go osłabia, natomiast wartość bliska zeru oznacza marginalny udział danej zmiennej w procesie przetwarzania informacji [37]. Wagi stanowią podstawowe parametry modelu i są iteracyjnie aktualizowane w kolejnych epokach procesu uczenia w taki sposób, aby minimalizować funkcję straty, czyli różnicę pomiędzy wartościami przewidywanymi przez model a wartościami rzeczywistymi [9]. Proces prezentacji danych oraz aktualizacji wag powtarzany jest wielokrotnie, aż do osiągnięcia satysfakcjonującego poziomu błędu lub spełnienia ustalonego kryterium zatrzymania [38].

1.2.3.1. Parametry architektury i uczenia sieci neuronowej

Przed rozpoczęciem procesu uczenia sieci neuronowej definiuje się jej architekturę, obejmującą liczbę warstw oraz liczbę neuronów w każdej z nich. Określa się również funkcje aktywacji, funkcję straty, algorytm uczenia, np. wsteczną propagację błędu, algorytm optymalizacyjny oraz początkowe wartości hiperparametrów modelu.

➤ Funkcja aktywacji

Każdy neuron w sieci oblicza ważoną sumę sygnałów wejściowych, do której dodawany jest parametr przesunięcia - *bias*. Otrzymana wartość poddawana jest następnie przekształceniu przez funkcję aktywacji. To właśnie funkcja aktywacji wprowadza do modelu nieliniowość i umożliwia sieci neuronowej modelowanie złożonych zależności między zmiennymi. W praktyce w nowoczesnych sieciach neuronowych stosowane są przede wszystkim funkcje ciągłe i różniczkowalne, takie jak funkcja sigmoidalna, czy tangens hiperboliczny ($\tanh$). Ich zastosowanie pozwala na wykorzystanie metod gradientowych w procesie optymalizacji parametrów sieci, w szczególności algorytmu wstecznej propagacji błędów. Wybór funkcji aktywacji wpływa zarówno na zakres wartości wyjściowych neuronu, jak i na dynamikę procesu uczenia, w tym szybkość zbieżności oraz stabilność optymalizacji. Parametr *bias* pełni rolę przesunięcia funkcji aktywacji, regulując próg odpowiedzi neuronu i umożliwiając elastyczne dopasowanie modelu do danych. W przypadku braku funkcji aktywacji (lub zastosowania wyłącznie funkcji liniowej) sieć wielowarstwowa redukuje się do modelu liniowego [39,40].

➤ Funkcja straty

Funkcja straty określa sposób ilościowego pomiaru błędów modelu podczas uczenia sieci neuronowej i bezpośrednio wpływa na przebieg oraz efektywność procesu trenowania [9].

Standardową funkcją straty stosowaną w problemach regresyjnych jest suma kwadratów (ang. *sum of squares*, SOS), definiowana jako suma kwadratów różnic pomiędzy wartościami rzeczywistymi a wartościami przewidywanymi przez model. Funkcja ta pozwala określić wielkość odchylenia predykcji od wartości rzeczywistych, a zastosowanie kwadratu powoduje, że duże odchylenia mają nieproporcjonalnie duży wpływ na wartość funkcji straty, co zwiększa wrażliwość modelu na obserwacje odstające oraz powoduje, że proces uczenia koncentruje się przede wszystkim na minimalizacji największych błędów [22,38].

W problemach natury klasyfikacyjnej najczęściej stosowaną funkcją straty jest entropia krzyżowa (ang. *cross-entropy*), która odpowiada ujemnemu logarytmowi wiarygodności modelu (ang. *likelihood*), tj. prawdopodobieństwu, jakie model przypisuje obserwowanym danym. Na podstawie danych treningowych model estymuje prawdopodobieństwa przynależności do poszczególnych klas, a entropia krzyżowa ocenia, czy wysokie prawdopodobieństwo przypisywane jest klasie rzeczywistej – im niższe prawdopodobieństwo, tym większa wartość funkcji straty. Minimalizacja entropii krzyżowej jest równoważna

maksymalizacji wiarygodności modelu, czyli maksymalizacji prawdopodobieństwa danych obserwowanych przy zadanych parametrach modelu. [9,37]

1.2.3.2. **Uczenie i optymalizacja sieci neuronowych**

Głównym celem procesu uczenia jest minimalizacja funkcji straty, która określa różnicę pomiędzy wartościami przewidywanymi przez model a wartościami rzeczywistymi. W celu osiągnięcia tego minimum w procesie uczenia i optymalizacji stosowane są odpowiednie algorytmy umożliwiające iteracyjną aktualizację parametrów modelu, takich jak wagi połączeń oraz wyrazy wolne [36].

➤ **Wsteczna propagacja błędów**

Jednym z najlepiej poznanych i nadal najczęściej używanych algorytmów uczenia jest wsteczna propagacja błędów (ang. *backpropagation*). Jej opracowanie przypisuje się trzem niezależnym zespołom – Rumelhart et al. [41], Werbos [42] oraz Parker [43].

Algorytm ten obejmuje trzy główne etapy:

- (1) propagację sygnału w przód (ang. *forward pass*), w której dla zadanych danych wejściowych wyznaczane są kolejne aktywacje neuronów aż do warstwy wyjściowej,
- (2) obliczenie wartości funkcji straty na podstawie różnicy między predykcją modelu a wartością rzeczywistą,
- (3) propagację błędu wstecz przez kolejne warstwy sieci w celu wyznaczenia gradientów funkcji straty względem wag i parametrów przesunięcia.

Kluczowym elementem wstecznej propagacji błędów jest stopniowe przekazywanie informacji o błędzie z warstwy wyjściowej do wcześniejszych warstw sieci. Po obliczeniu różnicy między wartością przewidywaną a rzeczywistą, algorytm określa, w jakim stopniu poszczególne neurony i ich połączenia przyczyniły się do powstania tego błędu. W rezultacie dla każdej wagi wyznaczany jest gradient określający kierunek oraz wielkość zmiany parametru minimalizującego funkcję straty [10].

Algorytm wstecznej propagacji błędów stanowi standardową metodę obliczania gradientów w większości współczesnych architektur sieci neuronowych, w tym w perceptronach wielowarstwowych, sieciach konwolucyjnych oraz rekurencyjnych [44]. Jego popularność wynika z efektywności obliczeniowej oraz możliwości skalowania do modeli o dużej liczbie parametrów.

Obliczenie gradientów funkcji straty względem parametrów sieci umożliwia aktualizację wag i wyrazu wolnego w procesie optymalizacji. Proces ten odbywa się z zastosowaniem algorytmów optymalizacyjnych, a do najczęściej stosowanych metod optymalizacji należą m.in. gradienty sprzężone (ang *Conjugate Gradient*, CG) oraz algorytmy quasi-Newtonowskie, takie jak Broyden–Fletcher–Goldfarb–Shanno (BFGS).

➤ **Gradienty sprzężone, CG**

Metoda gradientów sprzężonych stanowi metodę optymalizacji wykorzystywaną do minimalizacji funkcji straty, szczególnie w przypadku problemów o dużej liczbie parametrów. W odróżnieniu od klasycznej metody gradientu prostego, która w każdej iteracji porusza się w kierunku najszybszego spadku funkcji (ujemnego gradientu), metoda gradientów sprzężonych uwzględnia również informacje z poprzednich kroków optymalizacji. Oznacza to, że kolejne kierunki optymalizacji są konstruowane tak, aby uwzględniały informacje z poprzednich iteracji i nie „cofały” wcześniej wykonanych kroków minimalizacji. W efekcie ogranicza się zjawisko oscylacji charakterystyczne dla prostego spadku gradientowego, co zwykle przekłada się na szybsze osiągnięcie minimum funkcji celu [45].

W kontekście sieci neuronowych metoda CG może być stosowana do aktualizacji wag po obliczeniu gradientów metodą wstecznej propagacji błędu. CG wykorzystują pełną informację o zbiorze uczącym w każdej iteracji, co może zwiększać stabilność procesu optymalizacji, szczególnie w przypadku zbiorów o umiarkowanej wielkości, jednak aktualizacja parametrów jest bardziej kosztowna obliczeniowo niż w przypadku metod stochastycznych, które wykorzystują tylko część zbioru uczącego [46–48].

➤ **Metody quasi-Newtonowskie - algorytm Broyden–Fletcher–Goldfarb–Shanno, BFGS**

Algorytm Broyden–Fletcher–Goldfarb–Shanno (BFGS) należy do metod optymalizacyjnych wykorzystywanych do minimalizacji funkcji straty. W odróżnieniu od prostych metod gradientowych, które w każdej iteracji aktualizują parametry wyłącznie na podstawie aktualnego kierunku spadku funkcji, BFGS dodatkowo uwzględnia informacje o tym, jak funkcja zachowywała się w poprzednich krokach optymalizacji [49]. W praktyce oznacza to, że algorytm nie tylko „schodzi w dół” w kierunku największego spadku, lecz również dostosowuje długość i kierunek kolejnych kroków na podstawie wcześniejszych zmian parametrów i gradientu. Pozwala to często na szybsze i bardziej stabilne osiągnięcie minimum funkcji straty w porównaniu z klasycznym spadkiem gradientowym.

Z drugiej strony, metoda BFGS wymaga większych zasobów obliczeniowych oraz pamięci niż klasyczne metody gradientowe, co ogranicza jej zastosowanie w bardzo dużych modelach lub przy analizie bardzo rozbudowanych zbiorów danych [47].

W kontekście sieci neuronowych algorytm BFGS może być wykorzystywany do aktualizacji wag po obliczeniu gradientów metodą wstecznej propagacji błędów, szczególnie w modelach o umiarkowanej liczbie parametrów, takich jak klasyczne perceptrony wielowarstwowe [50].

Efektywność procesu uczenia zależy jednak nie tylko od samego algorytmu obliczania gradientu, czy zastosowanej metody optymalizacji, lecz również od doboru hiperparametrów oraz jakości danych. Niewłaściwa konfiguracja tych elementów może prowadzić do wolnej zbieżności lub zatrzymania procesu uczenia w obszarach o małym gradiencie.

1.2.3.3. Proces uczenia i walidacji modelu

Proces uczenia sieci neuronowej obejmuje kolejno: propagację danych wejściowych w przód w celu wyznaczenia predykcji $\hat{y}$, obliczenie wartości funkcji straty, wyznaczenie gradientów metodą wstecznej propagacji błędów oraz aktualizację wag i parametrów przesunięcia przy użyciu wybranego algorytmu optymalizacyjnego. Jak wspomniano wcześniej, uczenie ma charakter iteracyjny – dane uczące prezentowane są sieci wielokrotnie aż do spełnienia przyjętego kryterium zatrzymania (np. osiągnięcia minimalnej wartości funkcji straty lub stabilizacji błędów walidacyjnego) [9].

Poza wymienionymi wcześniej parametrami sieci, które definiują działanie modelu, istnieją również takie, które regulują proces uczenia sieci, i które należy zdefiniować przed rozpoczęciem tego procesu – są nimi współczynnik uczenia (ang. *learning rate*) i momentum [42, 52].

- **Współczynnik uczenia** określa wielkość kroku, o jaki aktualizowane są wagi sieci w kierunku przeciwnym do gradientu funkcji straty. Zbyt wysoka jego wartość może prowadzić do niestabilności procesu optymalizacji i przeskakiwania minimum funkcji celu, natomiast zbyt niska – do bardzo wolnej zbieżności modelu.
- **Parametr momentum** wprowadza do procesu uczenia element „bezwładności”, uwzględniając wpływ wcześniejszych aktualizacji wag. W praktyce oznacza to, że kierunek zmiany parametrów zależy nie tylko od aktualnego gradientu, lecz również od

poprzednich kroków optymalizacji. Pozwala to ograniczyć oscylacje w wąskich dolinach funkcji straty oraz przyspieszyć zbieżność w obszarach o stałym kierunku spadku.

➤ Podział danych

W celu zapewnienia obiektywnej oceny jakości modelu dane z reguły dzielone są na trzy rozłączne zbiory:

- zbiór uczący (ang. *training set*) – wykorzystywany jest do estymacji parametrów modelu. To na jego podstawie sieć neuronowa uczy się zależności między zmiennymi wejściowymi a zmienną wyjściową poprzez minimalizację funkcji straty. W kolejnych epokach wagi są modyfikowane tak, aby zmniejszać błąd predykcji w tym zbiorze;
- zbiór walidacyjny (ang. *validation set*) – służy do monitorowania zdolności modelu do generalizacji oraz do kontroli zjawiska przeuczenia. Dane te nie są wykorzystywane do aktualizacji wag, lecz jedynie do oceny jakości modelu po zakończeniu danej epoki. Jeżeli błąd w zbiorze uczącym maleje, natomiast błąd w zbiorze walidacyjnym zaczyna rosnąć, oznacza to nadmierne dopasowanie modelu do danych treningowych. W takiej sytuacji proces uczenia może zostać zatrzymany (procedura *early stopping* opisana w dalszej części rozdziału), a jako końcowe przyjmowane są parametry modelu odpowiadające minimalnemu błędowi walidacyjnemu;
- zbiór testowy (ang. *test set*) – stanowi całkowicie niezależny zestaw danych, który nie bierze udziału ani w procesie uczenia ani w doborze hiperparametrów. Służy on do końcowej, obiektywnej oceny jakości modelu oraz jego zdolności predykcji dla nowych, wcześniej niewidzianych przypadków. Ocena modelu w zbiorze testowym wykonywana jest jednorazowo po zakończeniu procesu uczenia i walidacji [37].

➤ Generalizacja i przeuczenie

Zdolność generalizacji stanowi kluczową cechę modelu uczenia maszynowego i oznacza jego umiejętność poprawnego przewidywania wartości wyjściowych dla danych, które nie były wykorzystywane w procesie uczenia. Model o wysokiej zdolności generalizacji odtwarza ogólną strukturę zależności występujących w danych, zamiast zapamiętywać pojedyncze obserwacje ze zbioru uczącego [13].

Ocena zdolności generalizacji dokonywana jest poprzez analizę jakości predykcji w zbiorze walidacyjnym podczas procesu uczenia, a ostatecznie – w niezależnym zbiorze

testowym. W początkowej fazie treningu błędy w zbiorze uczącym oraz walidacyjnym zazwyczaj maleją równolegle. W miarę dalszego dopasowywania modelu może jednak dojść do sytuacji, w której błąd w zbiorze uczącym nadal spada, natomiast błąd walidacyjny zaczyna rosnąć. W efekcie błąd w zbiorze uczącym jest niski, natomiast jakość predykcji w zbiorze walidacyjnym lub testowym ulega pogorszeniu. Oznacza to, że model nadmiernie dopasowuje się do danych uczących, odwzorowując nie tylko istotne zależności, lecz również przypadkowy szum i specyficzne cechy zbioru treningowego. Taki model traci zdolność generalizacji i nie zapewnia wiarygodnych wyników dla nowych danych [51].

Z uwagi na możliwość nadmiernego dopasowania modelu do danych uczących, w praktyce stosuje się dodatkowe strategie regularyzacyjne i walidacyjne, których celem jest kontrola procesu uczenia oraz stabilne oszacowanie błędu generalizacji. W niniejszej pracy zastosowano w tym celu procedurę wczesnego zatrzymania uczenia, walidację krzyżową oraz podejście zespołowe.

1.2.3.4. Strategie poprawy jakości modelu

➤ Wczesne zatrzymanie uczenia

Jedną z metod ograniczania przeuczenia jest wczesne zatrzymanie uczenia (ang. *early stopping*). Metoda ta opiera się na monitorowaniu błędu w zbiorze walidacyjnym podczas kolejnych epok uczenia.

W procedurze *early stopping* jako ostateczne przyjmowane są parametry modelu odpowiadające minimalnej wartości błędu w zbiorze walidacyjnym. Proces uczenia może zostać przerwany:

- po osiągnięciu maksymalnej liczby epok,
- po przekroczeniu określonej liczby kolejnych epok bez poprawy błędu walidacyjnego,
- po osiągnięciu ustalonego progu wartości funkcji straty [9,22,52].

➤ Walidacja krzyżowa

W klasycznej procedurze zbiór danych dzielony jest jednokrotnie na część uczącą i walidacyjną (oraz ewentualnie testową). Choć metoda ta jest prosta i skuteczna, jej wyniki mogą zależeć od konkretnego podziału danych.

W przypadku zbiorów danych o ograniczonej liczbie obserwacji korzystnym rozwiązaniem jest zastosowanie walidacji krzyżowej k-krotnej (ang. *k-fold cross-validation*).

W tej procedurze zbiór danych dzielony jest na k równych części. W każdej iteracji jedna część pełni rolę zbioru walidacyjnego, natomiast pozostałe $k-1$ części tworzą zbiór uczący. Proces powtarzany jest tyle razy, by każda część zbioru pełniła funkcję walidacyjną dokładnie jeden raz.

Końcowa ocena jakości modelu uzyskiwana jest poprzez uśrednienie wyników z poszczególnych iteracji. Takie podejście zmniejsza zależność wyniku od pojedynczego podziału danych i pozwala uzyskać bardziej stabilne oszacowanie zdolności generalizacji modelu [22,38,53].

➤ Zespoły sieci

Inną strategią zwiększenia jakości predykcyjnej modeli sieci neuronowych jest podejście zespołowe. Metody zespołowe (ang. *ensemble methods*) stanowią strategię zwiększania jakości modeli predykcyjnych poprzez łączenie wyników wielu niezależnie wytrenowanych modeli bazowych. Zamiast opierać predykcję na pojedynczym estymatorze, agreguje się wyniki kilku modeli, co prowadzi do poprawy stabilności i zdolności generalizacji.

Podstawowym założeniem podejścia zespołowego jest redukcja wariancji estymatora. W przypadku modeli o wysokiej złożoności, takich jak sieci neuronowe typu MLP, proces uczenia jest wrażliwy na inicjalizację wag, podział danych treningowych oraz parametry optymalizacji. W rezultacie różne realizacje tego samego modelu mogą prowadzić do odmiennych rozwiązań lokalnych funkcji błędu. Agregacja predykcji wielu modeli pozwala na uśrednienie błędów indywidualnych estymatorów i zmniejszenie wpływu losowych fluktuacji w procesie uczenia [54].

W zadaniach regresyjnych agregacja predykcji najczęściej przyjmuje postać średniej arytmetycznej wyników modeli składowych, natomiast w problemach klasyfikacyjnych stosuje się głosowanie większościowe lub uśrednianie prawdopodobieństw. Takie podejście prowadzi do zmniejszenia niestabilności predykcji oraz ograniczenia ryzyka nadmiernego dopasowania pojedynczego modelu do danych uczących [55].

Zastosowanie metod zespołowych jest szczególnie uzasadnione w warunkach:

- ograniczonej liczności próby,
- wysokiej złożoności przestrzeni cech,
- nieliniowego charakteru zależności między zmiennymi,
- podatności modeli na lokalne minima funkcji straty.

W kontekście sieci neuronowych podejście zespołowe stanowi skuteczne narzędzie zwiększania jakości predykcyjnej bez konieczności modyfikowania architektury pojedynczego modelu [56].

1.2.3.5. Zastosowanie sieci neuronowych

➤ Sieci neuronowe w rozwiązywaniu problemów klasyfikacyjnych

Celem klasyfikacji jest przypisanie analizowanego przypadku do jednej z uprzednio zdefiniowanych kategorii. Problemy klasyfikacyjne można podzielić na dwuklasowe (binarnie decyzyjne) oraz wieloklasowe.

W zadaniach dwuklasowych sieć neuronowa najczęściej posiada pojedynczy neuron wyjściowy, którego wartość interpretowana jest jako prawdopodobieństwo przynależności do jednej z klas. W praktyce stosuje się funkcję aktywacji sigmoidalną, a wynik poddawany jest progowaniu w celu uzyskania decyzji klasyfikacyjnej [9].

W przypadku klasyfikacji wieloklasowej w warstwie wyjściowej występuje zazwyczaj tyle neuronów, ile klas w analizowanym problemie. Każdy neuron reprezentuje jedną klasę, a do normalizacji wyjść stosowana jest najczęściej funkcja softmax, pozwalająca interpretować wartości jako rozkład prawdopodobieństwa. Docelowa etykieta przyjmuje wówczas postać wektora kodowanego metodą „one-hot” [37].

➤ Sieci neuronowe w rozwiązywaniu problemów regresyjnych

W zadaniach regresyjnych celem modelu jest przewidywanie wartości zmiennej wyjściowej o charakterze ciągłym na podstawie zestawu zmiennych wejściowych. W takim przypadku warstwa wyjściowa sieci zazwyczaj zawiera jeden neuron z liniową funkcją aktywacji, umożliwiającą generowanie nieograniczonych wartości liczbowych [37].

Sieci neuronowe znajdują szerokie zastosowanie w problemach regresyjnych ze względu na zdolność modelowania złożonych, nieliniowych zależności pomiędzy zmiennymi. W literaturze określane są często jako uniwersalne aproksymatory, co oznacza, że przy odpowiednio dobranej architekturze i liczbie parametrów są w stanie modelować niemalże każdą funkcję, a w przeciwieństwie do klasycznych modeli regresyjnych nie wymagają one uprzedniego założenia konkretnej postaci funkcji opisującej zależność między zmiennymi [57].

2. Cel pracy

Celem niniejszej pracy było zbadanie możliwości wykorzystania wybranych metod uczenia maszynowego – drzew decyzyjnych (CART), analizy składowych głównych (PCA) oraz sieci neuronowych typu perceptron wielowarstwowy (MLP) – w modelowaniu i predykcji złożonych problemów biomedycznych związanych z właściwościami przeciwdrobnoustrojowymi oraz chorobami nowotworowymi.

Praca stanowi próbę oceny skuteczności wyżej wymienionych metod w analizie danych o wysokiej wymiarowości, zarówno na poziomie molekularnym (deskryptory molekularne), jak i klinicznym (dane pacjenckie), z uwzględnieniem ich potencjału aplikacyjnego w badaniach przedklinicznych oraz procesach wspomagania decyzji medycznych.

Cel szczegółowy 1

Opracowanie i walidacja modeli sztucznych sieci neuronowych typu MLP do klasyfikacji aktywności przeciwdrobnoustrojowej pochodnych imidazoliowych wobec *Klebsiella pneumoniae*, na podstawie deskryptorów molekularnych, z oceną wpływu selekcji cech metodą CART oraz redukcji wymiarowości metodą PCA na jakość klasyfikacji i zdolność generalizacji modelu.

Cel szczegółowy 2

Opracowanie i walidacja regresyjnego modelu sztucznej sieci neuronowej typu MLP do ilościowej predykcji wartości lnMIC pochodnych imidazoliowych wobec *Staphylococcus aureus* oraz *Candida albicans*, z wykorzystaniem selekcji predyktorów metodą CART, redukcji wymiarowości metodą PCA oraz oceny stabilności i jakości predykcyjnej modelu.

Cel szczegółowy 3

Opracowanie i ocena modelu MLP do predykcji wznowy raka płuca na podstawie retrospektywnych danych klinicznych, z uwzględnieniem:

- doboru zmiennych predykcyjnych o znaczeniu klinicznym,
- optymalizacji architektury sieci oraz algorytmów uczenia
- analizy jakości klasyfikacyjnej

Cel szczegółowy 4

Przeprowadzenie porównawczej analizy skuteczności zastosowanych metod uczenia maszynowego w modelowaniu danych o różnym charakterze (molekularnych i klinicznych), ze szczególnym uwzględnieniem:

- wpływu wysokowymiarowości danych na proces uczenia,
- roli selekcji zmiennych i redukcji wymiarowości w budowie stabilnych modeli,
- znaczenia procedur walidacyjnych w ocenie zdolności generalizacji.

3. Materiały i metody

Materiał do badań w niniejszej pracy stanowiły dwa rodzaje danych:

1. dane chemometryczne nowo syntetyzowanych związków imidazoliowych
2. retrospektywne dane kliniczne pochodzące od pacjentów

Ad 1. Część chemometryczna pracy obejmowała dane opisujące dwa niezależne zbiory nowo syntetyzowanych pochodnych imidazoliowych:

Zbiór I – 140 chlorków 3,3'-(α , ω -dioksaalkan)bis(1-alkiloimidazoliowych) wykorzystanych w analizie klasyfikacyjnej aktywności przeciwdrobnoustrojowej wobec *Klebsiella pneumoniae* (szczep ATC 27853).

Na podstawie wartości minimalnego stężenia hamującego (ang. *Minimal Inhibitory Concentration*, MIC) związki przypisano do klasy niskiej lub wysokiej aktywności, za punkt odcięcia przyjmując MIC chlorku didecyldimetyloamoniowego (ang. *didecyldimethylammonium chloride*, DDAC) (MIC=0.856 mol/L) - czwartorzędowej soli amoniowej używanej do dezynfekcji i zaakceptowanej przez Europejski Obszar Ekonomiczny jako biocyd.

Zbiór II – 126 chlorków α,ω -bis[(alkilo)imidazol-1-ylo-3-alkoksymetylowych] wykorzystanych do budowy modeli regresyjnych przewidujących wartości lnMIC wobec *Staphylococcus aureus* (szczep ATCC 25213) oraz *Candida albicans* (szczep ATC 90028).

Dla wszystkich 266 analizowanych cząsteczek oznaczono eksperymentalnie wartości MIC oraz opracowano modele 2D i 3D, które następnie poddano optymalizacji geometrycznej metodą teorii funkcjonału gęstości (ang. *Density Functional Theory*, DFT) z zastosowaniem modelu polaryzowalnego kontinuum (ang. *Polarizable Continuum Model*, PCM). Na podstawie zoptymalizowanych struktur wyznaczono 5270 deskryptorów molekularnych z 29 bloków logicznych przy użyciu programu Dragon 7.0 (Talete, Mediolan, Włochy) [58].

Zbiory te analizowano niezależnie, przy zastosowaniu analogicznego schematu metodologicznego obejmującego selekcję najlepszych predyktorów z wykorzystaniem CART, redukcję wymiarowości z użyciem PCA oraz modelowanie sztucznych sieci neuronowych typu MLP.

W analizie danych dotyczących wartości MIC wobec *Staphylococcus aureus* oraz *Candida albicans* zastosowano dodatkowy etap wstępnego przetwarzania danych w środowisku Python (Python 3.10.11, biblioteki: pandas 2.2.2, numpy 1.26.4,

scikit-learn 1.5.1). Celem tej procedury było oczyszczenie zbioru deskryptorów molekularnych oraz ograniczenie jego wymiarowości przed dalszym etapem modelowania.

Dane eksperymentalne dotyczące wartości MIC, struktury oraz deskryptorów molekularnych nowo syntetyzowanych pochodnych imidazoliowych zostały pozyskane w ramach współpracy z naukowcami Katedry Technologii Postaci Leku Wydziału Farmaceutycznego Uniwersytetu Mikołaja Kopernika, Katedry i Zakładu Biofarmacji i Farmakodynamiki Wydziału Farmaceutycznego Gdańskiego Uniwersytetu Medycznego oraz Wydziału Technologii Chemicznej Politechniki Poznańskiej. Zespoły te przeprowadziły modelowanie molekularne badanych cząsteczek oraz wyznaczyły wartości minimalnego stężenia hamującego (MIC). W ramach niniejszej pracy przeprowadzono analizę uzyskanych danych z wykorzystaniem metod uczenia maszynowego.

Ad 2. Trzeci zbiór danych obejmował retrospektywne dane medyczne 2296 pacjentów hospitalizowanych w latach 2012–2017 w Warmińsko-Mazurskim Centrum Chorób Płuc w Olsztynie. Kryterium włączenia stanowiło histopatologiczne potwierdzenie rozpoznania raka płuca. Dane dotyczące przeżycia pacjentów uzyskano z Głównego Urzędu Statystycznego.

Zmienna docelowa obejmowała wystąpienie wznowy choroby (ang. *recurrence*), zdefiniowanego zgodnie z wytycznymi Północnoamerykańskiego Stowarzyszenia Centralnych Rejestrów Nowotworów oraz Programu Nadzoru, Epidemiologii i Wyników Leczenia. Zbiór danych obejmował zmienne socjodemograficzne, dane dotyczące stylu życia, informacje kliniczne oraz dotyczące terapii zastosowanej u pacjentów.

Badanie zostało przeprowadzone zgodnie z Deklaracją Helsińską i uzyskało zgodę Komisji Bioetycznej Uniwersytetu Mikołaja Kopernika w Toruniu przy Collegium Medicum im. Ludwika Rydygiera w Bydgoszczy (zgoda nr KB 355/2020 z dnia 23.06.2020).

3.1. Metody badawcze

W analizie zbiorów danych zastosowano metody uczenia maszynowego opisane w części teoretycznej pracy:

- drzewa klasyfikacyjne i regresyjne – CART,
- analizę głównych składowych – PCA,
- sztuczne sieci neuronowe typu MLP.

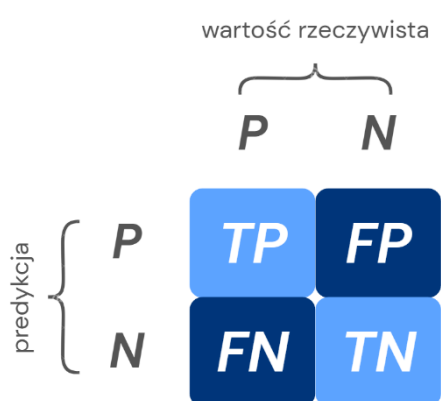
Zastosowanie metod ML miało charakter dwojaki:

- Redukcja wymiarowości i selekcja zmiennych – w przypadku analizy danych chemometrycznych ze względu na wysoką liczbę deskryptorów molekularnych, ograniczoną liczbę przypadków oraz potencjalne ryzyko przeuczenia modeli, wykorzystano algorytmy CART do selekcji najbardziej informatywnych zmiennych oraz PCA do redukcji wymiarowości danych.
- Modelowanie predykcyjne – właściwy etap modelowania realizowano przy użyciu sieci neuronowych typu MLP, które – dzięki swojej nieliniowej strukturze – pozwalają uchwycić złożone zależności między zmiennymi wejściowymi a zmienną docelową.

Wszystkie etapy doboru optymalnych predyktorów, redukcji wymiarowości danych oraz budowy i walidacji modeli sztucznych sieci neuronowych zostały przeprowadzone z wykorzystaniem oprogramowania Statistica 7.1 (Statsoft, Kraków, Polska) oraz TIBCO Statistica 13 (TIBCO Software Inc., Palo Alto, USA).

3.2. Ocena jakości modeli klasyfikacyjnych

W problemach klasyfikacyjnych jakość modelu oceniana jest na podstawie zgodności przewidywanych klas z klasami rzeczywistymi. W tym celu tworzona jest macierz pomyłek (**Rycina 4**), której licznosci w każdej grupie służą do obliczenia poszczególnych miar.



Rycina 4. Macierz pomyłek dla problemu klasyfikacji binarnej.

Interpretacja elementów:

- **TP** (ang. *True Positive*) – przypadki pozytywne poprawnie rozpoznane jako pozytywne
- **TN** (ang. *True Negative*) – przypadki negatywne poprawnie rozpoznane jako negatywne
- **FP** (ang. *False Positive*) – przypadki negatywne błędnie zaklasyfikowane jako pozytywne (błąd I rodzaju)
- **FN** (ang. *False Negative*) – przypadki pozytywne błędnie zaklasyfikowane jako negatywne (błąd II rodzaju)

Najprostszą miarą jest dokładność klasyfikacji (ang. *accuracy*, **ACC**), określająca odsetek poprawnie sklasyfikowanych przypadków w stosunku do wszystkich obserwacji:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

W praktyce jednak sama dokładność może być niewystarczająca, szczególnie w przypadku nie zrównoważonych zbiorów danych, dlatego w pracy zastosowano również inne miary:

czułość

(ang. *sensitivity*, *True Positive Rate*, **TPR**) – zdolność modelu do poprawnego wykrywania przypadków pozytywnych:

$$TPR = \frac{TP}{TP + FN}$$

swoistość

(ang. *specificity*, *True Negative Rate*, **SPC**) – zdolność do poprawnego rozpoznawania przypadków negatywnych:

$$SPC = \frac{TN}{TN + FP}$$

częstość wyników fałszywie dodatnich

(ang. *False Positive Rate*, **FPR**) – odsetek przypadków negatywnych błędnie zaklasyfikowanych jako pozytywne:

$$FPR = \frac{FP}{FP + TN} = 1 - SPC$$

precyzja

(ang. *precision*, *Positive Predictive Value*, **PPV**) – odsetek poprawnych klasyfikacji wśród wszystkich przewidzianych jako pozytywne:

$$PPV = \frac{TP}{TP + FP}$$

współczynnik fałszywych odkryć

(ang. *False Discovery Rate*, **FDR**) – odsetek błędnych klasyfikacji wśród wszystkich przewidzianych jako pozytywne:

$$FDR = \frac{FP}{TP + FP} = 1 - PPV$$

ujemna wartość predykcyjna

(ang. *Negative Predictive Value*, **NPV**) – odsetek prawdziwie negatywnych klasyfikacji wśród wszystkich przewidzianych jako negatywne:

$$NPV = \frac{TN}{TN + FN}$$

Ponadto, oprócz wyżej wymienionych klasycznych miar jakości klasyfikacji do oceny jakości modeli wykorzystano współczynnik korelacji Matthews (ang. *Matthews Correlation Coefficient*, MCC). Miara ta uwzględnia wszystkie elementy macierzy pomyłek (TP, TN, FP, FN) i stanowi zrównoważony wskaźnik jakości klasyfikacji, szczególnie zalecany w przypadku niezrównoważonych zbiorów danych.

Wartość MCC mieści się w przedziale od -1 do 1 , gdzie 1 oznacza klasyfikację idealną, 0 wynik losowy, a wartości ujemne wskazują na systematycznie błędną klasyfikację.

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

W celu kompleksowej oceny jakości klasyfikatora wyznaczono krzywe ROC (ang. *Receiver Operating Characteristic*), przedstawiające zdolność modelu do rozróżniania dwóch klas przy różnych wartościach progu decyzyjnego oraz wartości pól pod krzywą ROC (ang. *Area Under the Curve*, AUC).

Krzywa ROC powstaje poprzez stopniową zmianę progu klasyfikacji w zakresie od 0 do 1 oraz obliczanie odpowiadających mu wartości TPR i FPR. Pozwala to na ocenę kompromisu pomiędzy czułością a swoistością modelu niezależnie od konkretnego punktu odcięcia.

W warunkach istotnej nierównowagi klas, pole pod krzywą ROC (AUC) stanowi bardziej adekwatną miarę jakości klasyfikatora niż sama dokładność klasyfikacji (ACC), ponieważ uwzględnia zdolność modelu do rozróżniania klas w całym zakresie możliwych progów decyzyjnych [59].

Wartość AUC przyjmuje zakres od 0 do 1 i stanowi miarę globalnej zdolności modelu do rozróżniania klas:

- $AUC = 1$ – klasyfikator idealny,
- $AUC = 0.5$ – klasyfikator losowy,
- $AUC < 0.5$ – klasyfikator o zdolności dyskryminacyjnej gorszej niż losowa.

AUC można interpretować jako prawdopodobieństwo, że model przypisze wyższe prawdopodobieństwo przynależności do klasy pozytywnej losowo wybranemu obiektowi rzeczywiście należącemu do tej klasy niż losowo wybranemu obiektowi klasy negatywnej.

Zastosowane miary umożliwiły kompleksową ocenę zdolności modelu do poprawnej identyfikacji przypadków pozytywnych i negatywnych uwzględniając wpływ nierównowagi liczebności klas.

3.3. Ocena jakości modeli regresyjnych

W zadaniach regresyjnych jakość modelu określana jest na podstawie różnicy pomiędzy wartościami przewidywanymi a rzeczywistymi. Najczęściej stosowane miary obejmują:

średni błąd kwadratowy

(ang. *Mean Squared Error*, **MSE**) – miara średniej kwadratowej różnicy między wartościami rzeczywistymi a przewidywanymi:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

pierwiastek z MSE

(ang. *Root Mean Squared Error*, **RMSE**) – pierwiastek kwadratowy z MSE, wyrażony w tych samych jednostkach co zmienna zależna:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

średni błąd bezwzględny

(ang. *Mean Absolute Error*, **MAE**) – średnia wartość bezwzględnych różnic między wartościami rzeczywistymi a przewidywanymi:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

współczynnik determinacji (R^2) – miara dopasowania modelu, określająca, jaka część wariancji zmiennej zależnej jest wyjaśniona przez model:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Gdzie:

y_i – wartość rzeczywista dla i -tej obserwacji

n – liczba obserwacji

$\hat{y}_i$ – wartość przewidziana przez model

$\bar{y}$ – średnia wartość rzeczywista

Miary te pozwalają ocenić zarówno średnią wielkość błędu predykcji, jak i stopień wyjaśnienia zmienności danych przez model. Podobnie jak w przypadku klasyfikacji, ocena powinna być przeprowadzona na zbiorze niezależnym od procesu uczenia, aby uniknąć przeszacowania jakości modelu.

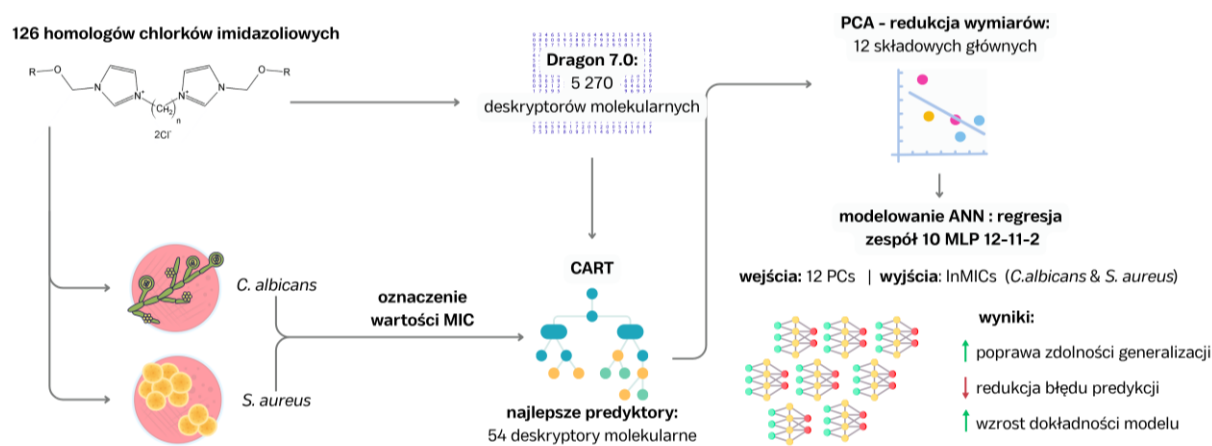
4. Wyniki

4.1. Publikacja nr 1

Lorenc A., Badura A., Karolak M., Pałkowski Ł., Kubik Ł., Bucíński A.:

Antimicrobial activity classification of imidazolium derivatives predicted by artificial neural networks,

Pharmaceutical Research, vol. 41, nr 5, 2024, s. 891-898, DOI:10.1007/s11095-024-03699-x, 70 punktów, IF(4,3)



Rycina 5. Abstrakt graficzny przedstawiający proces klasyfikacji aktywności przeciwdrobnoustrojowej 140 homologów chlorków imidazoliowych wobec *Klebsiella pneumoniae*.

Pierwsza z prac wchodzących w skład niniejszej rozprawy dotyczyła klasyfikacji 140 nowo syntetyzowanych pochodnych imidazoliowych wykazujących aktywność wobec *Klebsiella pneumoniae*. Projektowane związki modelowano w przestrzeni 2D i 3D, a następnie opisano za pomocą 5270 deskryptorów molekularnych. Po syntezie i potwierdzeniu struktury chemicznej wyznaczono wartości MIC, które stanowiły podstawę podziału związków na dwie klasy aktywności: niską (low) i wysoką (high) (**Rycina 5**).

Ze względu na bardzo dużą liczbę deskryptorów molekularnych ($N = 5270$) zastosowano wieloetapowy preprocessing danych. W pierwszym etapie wykorzystano metodę drzew klasyfikacyjnych CART w celu redukcji liczby zmiennych. Selekcja deskryptorów oparta była na wartości statystyki F oraz odpowiadającym jej poziomie istotności (p-value), które odzwierciedlały siłę związku pomiędzy danym deskryptorem a klasą aktywności przeciwdrobnoustrojowej. Na tej podstawie wybrano 200 deskryptorów, które następnie

stanowiły dane wejściowe do dalszej redukcji wymiarowości z wykorzystaniem PCA. Redukcja wymiarowości doprowadziła do uzyskania pięciu PC, wyjaśniających 99,57% całkowitej wariancji pierwotnych 200 deskryptorów.

Zbiór danych podzielono na zbiór uczący, walidacyjny i testowy, kolejno w proporcji 70:20:10. Na tak przygotowanych danych zbudowano model sieci neuronowej typu MLP, której architektura obejmowała:

- warstwę wejściową złożoną z 5 neuronów (odpowiadających pięciu PC),
- jedną warstwę ukrytą (12 neuronów),
- warstwę wyjściową z jednym neuronem odpowiadającym binarnej klasie aktywności (niska/wysoka).

W warstwie wyjściowej zastosowano funkcję aktywacji typu logistycznego (sigmoidalną), umożliwiającą interpretację wyjścia jako prawdopodobieństwa przynależności do klasy. Proces uczenia realizowano z wykorzystaniem algorytmu wstecznej propagacji błędów.

Ze względu na relatywnie niewielką liczebność zbioru zastosowano 12-krotną walidację krzyżową w celu zwiększenia wiarygodności oceny zdolności generalizacyjnej modelu. Opracowany model osiągnął procent poprawnych klasyfikacji na poziomie 88,61% dla zbioru uczącego, 90,48% dla zbioru walidacyjnego oraz 86,90% dla zbioru testowego. W analizie metryk klasyfikacyjnych uzyskano ACC = 0,907, SPC = 0,905 oraz PPV = 0,891, co wskazuje na zrównoważoną zdolność modelu do identyfikacji związków o wysokiej aktywności przeciwdrobnoustrojowej.

Uzyskane wyniki potwierdziły, że integracja metod selekcji cech (CART), redukcji wymiarowości (PCA) oraz modelowania nieliniowego (MLP) stanowi skuteczne podejście w zadaniach klasyfikacji aktywności biologicznej związków chemicznych charakteryzowanych wielowymiarowymi zbiorami danych. Przedstawione badania wyznaczają punkt wyjścia dla kolejnych prac wchodzących w skład niniejszej rozprawy, w których rozwijano podejście metodologiczne do predykcji właściwości biologicznych z wykorzystaniem sztucznych sieci neuronowych.

4.2. Publikacja nr 2

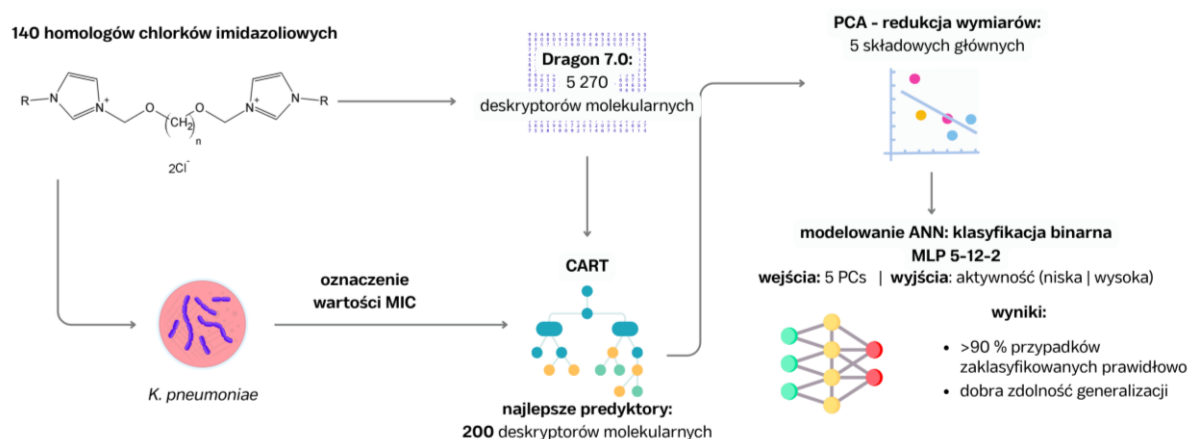
Lorenc A., Badura A., Karolak M., Pałkowski Ł., Kubik Ł., Skrzypczak A., Buciński A.:

Hybrid machine learning model for dual prediction of antimicrobial activity against *Staphylococcus aureus* and *Candida albicans*,

Journal of Pharmaceutical Innovation, vol. 21, nr 2, 2026, Numer artykułu: 174, s. 1-9,

DOI:10.1007/s12247-026-10412-5,

70 punktów, IF(2,7)



Rycina 6. Abstrakt graficzny przedstawiający hybrydowe zastosowanie metod ML do symultanicznej predykcji aktywności przeciwdrobnoustrojowej 126 homologów chlorków imidazoliowych wobec *Staphylococcus aureus* i *Candida albicans*.

Naturalną konsekwencją wyników uzyskanych w poprzedniej pracy, poświęconej klasyfikacji aktywności przeciwdrobnoustrojowej pochodnych imidazoliowych, było postawienie kolejnego pytania badawczego. O ile wcześniejsze badanie koncentrowało się na problemie klasyfikacyjnym (przynależność związku do określonej klasy aktywności), o tyle w niniejszej pracy podjęto próbę rozwiązania zadania istotnie bardziej złożonego – predykcji dokładnych wartości MIC. Dodatkowo rozszerzono zakres analizy na dwa drobnoustroje należące do odmiennych królestw biologicznych: bakterię *Staphylococcus aureus* oraz grzyb *Candida albicans*, co było uzasadnione ich udokumentowaną skłonnością do współwystępowania oraz wywoływania zakażeń mieszanych (**Rycina 6**). Tak sformułowany problem wymagał weryfikacji, czy zaproponowana wcześniej metodologia selekcji i modelowania deskryptorów molekularnych zachowa skuteczność również w warunkach regresji wielowyjściowej oraz w obliczu większej złożoności biologicznej modelowanego zjawiska.

Pierwszym etapem analizy było wstępne oczyszczenie danych z użyciem języka Python i jego bibliotek (numpy, pandas, sklearn). W tym kroku ze zbioru danych wyeliminowano zmienne zawierające braki, a następnie charakteryzujące się niską wariancją (za próg odcięcia przyjmując wartość wariancji $< 0,05$). Następnie, celem uniknięcia problemu multikolinearności, wysokoskorelowane zmienne zostały odfiltrowane (za próg odcięcia przyjmując siłę korelacji $> 0,95$). Pełny skrypt oraz opis tego procesu dostępny jest w sekcji Artykuł nr 2 – materiał dodatkowy: 4. Preprocessing script.

Kolejnym etapem analizy była selekcja deskryptorów molekularnych z wykorzystaniem algorytmu drzew decyzyjnych CART. Procedurę przeprowadzono oddzielnie dla każdego z drobnoustrojów, identyfikując odrębny zestaw zmiennych istotnych dla przewidywania wartości MIC. Następnie dokonano porównania obu zbiorów i wyodrębniono predyktory wspólne dla obu mikroorganizmów. Pozwoliło to na uzyskanie 54 zmiennych potencjalnie niosących informację zarówno o aktywności przeciwbakteryjnej, jak i przeciwgrzybiczej. Przyjęta strategia miała na celu ograniczenie przestrzeni cech wyłącznie do parametrów wykazujących znaczenie predykcyjne w odniesieniu do obu typów aktywności biologicznej, co umożliwiło konstrukcję jednego modelu regresyjnego o szerszym zakresie zastosowania.

Liczba 54 zmiennych wejściowych okazała się jednak zbyt znaczna względem liczby dostępnych do analizy przypadków ($N = 126$). W związku z tym zastosowano PCA jako metodę redukcji wymiarowości. Transformacja pozwoliła na ograniczenie liczby wejść do 12 PC, zachowujących zdecydowaną większość (91,7%) wariancji pierwotnego zbioru danych. Zastosowanie PCA pełniło w tym podejściu podwójną funkcję – redukowało wymiar przestrzeni cech oraz eliminowało współliniowość deskryptorów molekularnych.

Ostatnim etapem łańcucha analitycznego było skonstruowanie sieci neuronowej typu MLP, której zadaniem była jednoczesna predykcja wartości MIC dla *S. aureus* oraz *C. albicans*. Zastosowanie pojedynczego modelu dla dwóch drobnoustrojów stanowiło istotny element nowości metodologicznej. W przeciwieństwie do budowy dwóch niezależnych modeli regresyjnych, zaproponowane rozwiązanie zakładało istnienie wspólnych, strukturalnych determinant aktywności przeciwdrobnoustrojowej, które mogą być reprezentowane przez wspólną przestrzeń cech. Podejście to można interpretować jako formę uczenia wielozadaniowego, w którym model uczy się wspólnej reprezentacji cech istotnych dla dwóch powiązanych, lecz biologicznie odrębnych problemów predykcyjnych.

W procesie modelowania analizowano warianty sieci neuronowych wykorzystujących od 3 do 12 PC jako zmienne wejściowe. Każdy wariant oceniano pod kątem jakości predykcyjnej oraz stabilności wyników pomiędzy zbiorem uczącym, walidacyjnym i testowym, zwracając szczególną uwagę na rozbieżności błędu pomiędzy zbiorami jako wskaźnik potencjalnego przeuczenia.

Przeprowadzone analizy wykazały, że modele wykorzystujące 12 głównych składowych charakteryzowały się najwyższą stabilnością predykcyjną oraz najbardziej wyrównanymi wynikami w poszczególnych zbiorach danych. Na tej podstawie przyjęto 12 PC jako optymalny wymiar przestrzeni wejściowej.

W kolejnym etapie skonstruowano 10 niezależnych modeli regresyjnych o architekturze 12-11-2 (12 neuronów wejściowych odpowiadających głównym składowym, 11 neuronów w warstwie ukrytej oraz 2 neurony wyjściowe odpowiadające wartościom $\ln\text{MIC}$ dla *S. aureus* i *C. albicans*). Modele różniły się inicjalizacją wag oraz przebiegiem procesu uczenia.

Ostatecznie zdecydowano o zastosowaniu podejścia zespołowego, polegającego na agregacji predykcji uzyskanych z 10 modeli. Takie rozwiązanie pozwoliło na ograniczenie wpływu losowej inicjalizacji parametrów na wynik końcowy oraz zwiększenie stabilności i wiarygodności predykcji.

Dodatkowo przeprowadzono analizę wrażliwości dla wszystkich sieci wchodzących w skład zespołu w celu oceny względnej istotności głównych składowych jako zmiennych wejściowych. Wyniki jednoznacznie wskazały, że pierwsza główna składowa (PC1) stanowiła najistotniejszy predyktor we wszystkich analizowanych modelach. Obserwacja ta jest zgodna z charakterem transformacji PCA, ponieważ PC1 wyjaśnia największą część wariancji pierwotnego zbioru deskryptorów molekularnych i tym samym niesie największy ładunek informacyjny.

W przypadku kolejnych głównych składowych nie zaobserwowano jednak stałej i jednoznacznej hierarchii istotności. Rangi PC2–PC12 różniły się pomiędzy poszczególnymi sieciami, co wskazuje na współdzielony i rozproszony charakter informacji predykcyjnej w przestrzeni zredukowanych cech. Oznacza to, że poza dominującą rolą PC1, model nie opierał się na pojedynczych, izolowanych komponentach, lecz na złożonej kombinacji współdziałających zmiennych. Taki rozkład istotności potwierdza zasadność zastosowania podejścia zespołowego, które ogranicza wpływ przypadkowych fluktuacji rang

poszczególnych komponentów w pojedynczych modelach i zwiększa stabilność końcowej predykcji.

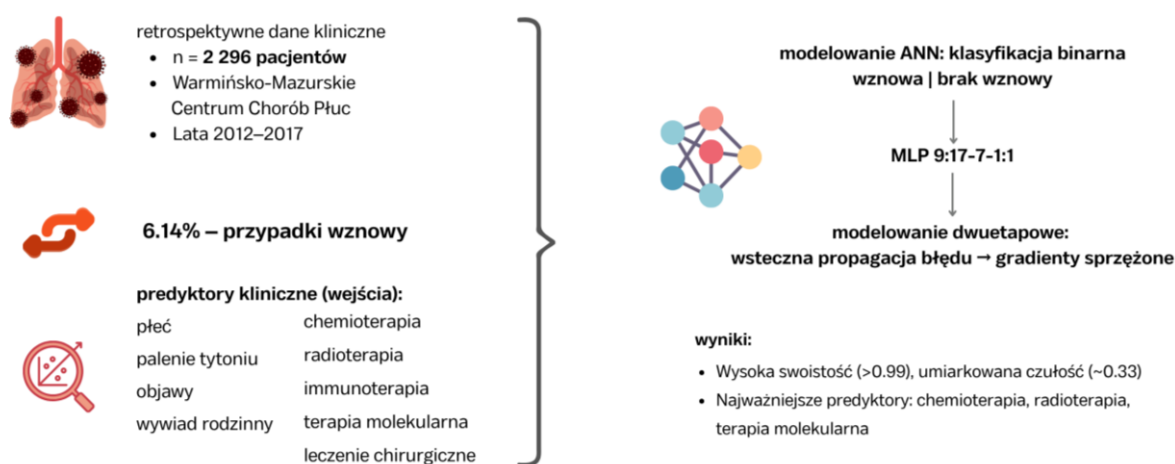
Zastosowany w niniejszej pracy, analogiczny do poprzedniego, proces analityczny pozwolił skutecznie wyodrębnić informację predykcyjną z wielowymiarowej przestrzeni deskryptorów molekularnych oraz odwzorować zależności pomiędzy strukturą chemiczną badanych związków a wartościami MIC równocześnie wobec dwóch mikroorganizmów. Uzyskane wyniki wskazują, że wspólna reprezentacja cech może być efektywnie wykorzystywana do predykcji aktywności przeciwdrobnoustrojowej wobec organizmów należących do różnych królestw biologicznych.

4.3. Publikacja nr 3

Lorenc A., Romaszko-Wojtowicz A., Jaśkiewicz Ł., Doboszyńska A., Buciński A.:

Exploring the efficacy of artificial neural networks in predicting lung cancer recurrence: a retrospective study based on patient records,

Translational Lung Cancer Research, Society for Translational Medicine (STM), vol. 12, nr 10, 2023, s. 2083-2097, DOI:10.21037/tlcr-23-350, 140 punktów, IF(4)



Rycina 7. Abstrakt graficzny przedstawiający zastosowanie sztucznej sieci neuronowej do predykcji wznowy raka płuca na podstawie retrospektywnych danych klinicznych 2 296 pacjentów hospitalizowanych w latach 2012-2017 w Warmińsko-Mazurskim Centrum Chorób Płuc.

Zgoła odmienny charakter miało trzecie z badań wchodzących w skład niniejszego cyklu – „*Exploring the efficacy of artificial neural networks in predicting lung cancer recurrence: a retrospective study based on patient records*”. W przeciwieństwie do wcześniejszych prac opartych na danych chemometrycznych związków imidazolowych, przedmiotem analizy stał się problem kliniczny – predykcja wznowy raka płuca na podstawie retrospektywnych danych medycznych pacjentów [Rycina 7].

Badanie przeprowadzono w oparciu o 2296 historii chorób pacjentów hospitalizowanych w latach 2012-2017 w Warmińsko-Mazurskim Centrum Chorób Płuc w Olsztynie. Kryterium włączenia stanowiło histopatologicznie potwierdzone rozpoznanie raka płuca. Wznowę definiowano zgodnie z wytycznymi NAACCR, ACoS oraz SEER/NCI. Dane zostały losowo podzielone na zbiór uczący (50%), walidacyjny (25%) oraz testowy (25%).

Istotnym utrudnieniem w analizie był zarówno kliniczny charakter danych, jak i ich typ. W odróżnieniu od wcześniejszych badań, w których każdy przypadek opisywany był przez liczne deskryptory chemiczne o charakterze zmiennych ciągłych, w niniejszym badaniu wszystkie zmienne wejściowe miały charakter jakościowy.

Zastosowane predyktory obejmowały:

- płeć,
- status palenia tytoniu,
- rodzaj chemioterapii,
- zastosowanie radioterapii, immunoterapii, terapii molekularnej,
- leczenie chirurgiczne,
- obecność objawów,
- wywiad rodzinny.

Zmienne te miały charakter dychotomiczny (0–1) lub typu „jeden z N” (one-of-N), które następnie zakodowano metodą one-hot encoding, uzyskując łącznie 17 neuronów w warstwie wejściowej. Ostatecznie architektura wybranego modelu przyjęła postać MLP 9:17-7-1:1 (9 zmiennych logicznych przekodowanych do 17 neuronów wejściowych, 7 neuronów w warstwie ukrytej, 1 neuron wyjściowy).

W przeciwieństwie do analiz chemometrycznych, gdzie selekcja zmiennych wejściowych była wynikiem działania algorytmów takich jak CART czy PCA, w tym badaniu dobór predyktorów opierał się przede wszystkim na aktualnej wiedzy medycznej oraz doświadczeniu zespołu klinicznego. Oznacza to, że o istotności zmiennych nie decydował algorytm eksploracyjny, lecz przesłanki kliniczne dotyczące czynników potencjalnie związanych z procesem wznowy.

Model uczono z wykorzystaniem dwuetapowej procedury optymalizacyjnej. W pierwszej fazie zastosowano algorytm wstecznej propagacji błędu przez 100 epok, co umożliwiło wstępne dostrojenie wag sieci. Następnie proces optymalizacji kontynuowano przez 20 epok z wykorzystaniem algorytmu CG. Takie podejście pozwoliło połączyć stabilność klasycznego algorytmu BP w początkowej fazie uczenia z szybszą konwergencją i efektywniejszym przeszukiwaniem przestrzeni rozwiązań charakterystycznym dla metod drugiego rzędu. W efekcie uzyskano model o wysokiej zdolności generalizacji, bez widocznych objawów przeuczenia.

Wybrany model MLP osiągnął wysoką jakość klasyfikacji:

- ACC: 0,875 (zbiór uczący); 0,891 (walidacyjny); 0,899 (testowy),
- AUC ROC: odpowiednio 0,937; 0,927; 0,920.

Ze względu na niezrównoważony charakter zbioru danych (wznowa występowała w ok. 6% przypadków), ocenę modelu rozszerzono o dodatkowe miary klasyfikacyjne, w tym TPR, SPC, PPV oraz współczynnik korelacji Matthews (MCC $\approx$ 0,5). Analiza ROC potwierdziła wysoką zdolność rozróżniania przypadków ze wznową i bez wznowy.

Przeprowadzona analiza wrażliwości wykazała, że największe znaczenie dla predykcji miały informacje dotyczące terapii – w szczególności chemioterapii, radioterapii oraz terapii molekularnej, co pozostaje spójne z aktualną wiedzą onkologiczną.

Uzyskane wyniki potwierdzają, że sieci neuronowe typu MLP mogą stanowić efektywne narzędzie wspomagające ocenę ryzyka wznowy raka płuca, nawet przy wykorzystaniu podstawowych danych klinicznych. Tym samym praca ta wpisuje się w szerszy nurt badań nad zastosowaniem metod sztucznej inteligencji jako narzędzi wspomagania decyzji klinicznych, podkreślając potencjał integracji uczenia maszynowego z praktyką medyczną.

5. Podsumowanie i wnioski

Cykl publikacji wchodzących w skład niniejszej rozprawy doktorskiej przedstawia systematyczną ewolucję zastosowań sztucznych sieci neuronowych typu MLP w predykcji zjawisk biologicznych i medycznych. Rozwój ten obejmuje modelowanie aktywności przeciwdrobnoustrojowej nowych związków chemicznych, ilościową predykcję wartości MIC wobec dwóch odmiennych patogenów, a następnie analizę retrospektywnych danych klinicznych w celu prognozowania wznowy choroby nowotworowej.

Zarówno w badaniach dotyczących danych eksperymentalnych pozyskiwanych metodami laboratoryjnymi i *in silico* („*Antimicrobial Activity Classification of Imidazolium Derivatives Predicted by Artificial Neural Networks*” oraz „*Hybrid machine learning model for dual prediction of antimicrobial activity against Staphylococcus aureus and Candida albicans*”), jak i w analizie danych klinicznych („*Exploring the Efficacy of Artificial Neural Networks in Predicting Lung Cancer Recurrence*”), zastosowano różne techniki uczenia maszynowego.

Mimo różnic tematycznych, wszystkie trzy publikacje łączy wspólna oś metodologiczna:

1. Zastosowanie sieci MLP jako głównego narzędzia modelowania;
2. Kontrola przeuczenia poprzez rygorystyczny podział danych na zbiory uczące, walidacyjne i testowe;
3. Optymalizacja architektury oraz algorytmów uczenia;
4. Analiza istotności predyktorów;
5. Ocena jakości modeli z wykorzystaniem rozbudowanych miar statystycznych.

Przeprowadzona analiza wykazała, że skuteczność modeli MLP w dużej mierze zależy od charakteru danych oraz sposobu ich przygotowania. Dane molekularne charakteryzowały się bardzo wysoką wymiarowością, co istotnie zwiększało ryzyko przeuczenia. Zastosowanie drzew decyzyjnych CART do selekcji najistotniejszych deskryptorów, a następnie analizy składowych głównych w celu dalszej redukcji wymiaru przestrzeni cech, pozwoliło na zachowanie około 99% informacji przy znacznym uproszczeniu struktury wejścia do sieci. Procedury te zwiększyły stabilność modeli oraz poprawiły ich zdolność do generalizacji w porównaniu z modelami opartymi na nieprzetworzonych deskryptorach. Ze względu na ograniczoną licznosc próby zastosowano walidację krzyżową lub podejście zespołowe, co bezpośrednio wpłynęło na zdolność predykcyjną modeli, ich stabilność oraz umożliwiło bardziej wiarygodną ocenę jakości w warunkach małych zbiorów uczących.

W przypadku danych klinicznych, cechujących się mniejszą liczbą zmiennych, lecz większą heterogenicznością populacji i nie zrównoważeniem klas, kluczowe znaczenie miała strategia uczenia modelu. Zastosowanie dwuetapowego procesu optymalizacji – obejmującego wsteczną propagację błędu, a następnie algorytm CG – zwiększyło stabilność treningu i umożliwiło skuteczniejsze minimalizowanie funkcji błędu. Strategia ta pozwoliła na początkowe uchwycenie ogólnych zależności w danych, a następnie ich precyzyjne dopasowanie, ograniczając ryzyko utknięcia w lokalnych minimach.

Badanie to stanowi metodologiczne rozszerzenie wcześniejszych prac zawartych w cyklu. O ile w analizach chemometrycznych głównym wyzwaniem była redukcja wymiarowości i selekcja spośród dużej liczby zmiennych ilościowych, o tyle w badaniu klinicznym kluczowe było modelowanie złożonych, nieliniowych zależności pomiędzy relatywnie niewielką liczbą zmiennych jakościowych o ograniczonej indywidualnej mocy informacyjnej.

Porównanie obu podejść wskazuje, że:

1. Wysokowymiarowość danych molekularnych wymaga formalnych procedur redukcji przestrzeni cech;
2. W danych klinicznych kluczowym wyzwaniem jest nie zrównoważenie klas oraz heterogeniczność populacji;
3. Niezależne zbiory walidacyjne i testowe są niezbędne dla rzetelnej oceny zdolności generalizacyjnej modeli.

Przeprowadzone badania potwierdzają, że sieci neuronowe typu MLP mogą stanowić skuteczne narzędzie modelowania w analizie danych chemometrycznych i klinicznych, pod warunkiem zachowania rygoru metodologicznego w zakresie selekcji zmiennych, redukcji wymiarowości, optymalizacji procesu uczenia oraz walidacji modeli. Uzyskane wyniki wskazują również, że kluczowym czynnikiem determinującym jakość predykcji nie jest wyłącznie wybór algorytmu, lecz świadome projektowanie całego procesu modelowania, uwzględniające charakter i strukturę analizowanych danych.

Wyniki przedstawione w cyklu publikacji potwierdzają również uniwersalność sztucznych sieci neuronowych jako narzędzi zdolnych do rozwiązywania problemów o odmiennym charakterze biologicznym i medycznym. Wykazano, że odpowiednio zaprojektowane modele MLP mogą być z powodzeniem stosowane zarówno do klasyfikacji aktywności biologicznej nowo projektowanych związków chemicznych, ilościowej predykcji parametrów

mikrobiologicznych, takich jak minimalne stężenie hamujące analizowanych związków, jak i do prognozowania wznowy nowotworu na podstawie danych pochodzących z dokumentacji medycznej pacjentów. Otrzymane rezultaty wskazują, że modele neuronowe zachowują wysoką skuteczność predykcyjną niezależnie od źródła danych, pod warunkiem dostosowania procedur przygotowania danych i walidacji modeli do specyfiki analizowanego problemu.

Istotnym osiągnięciem przedstawionych badań jest również wykazanie znaczenia metod wspomagających proces modelowania, takich jak drzewa decyzyjne CART, analiza składowych głównych czy analiza wrażliwości. Narzędzia te umożliwiły nie tylko poprawę jakości predykcyjnej modeli, ale również zwiększenie ich interpretowalności, co stanowi jeden z kluczowych warunków praktycznego wykorzystania metod sztucznej inteligencji w naukach biomedycznych. Uzyskane wyniki wpisują się w aktualny trend rozwoju metod *Explainable Artificial Intelligence* (XAI), którego celem jest zwiększenie przejrzystości modeli oraz lepsze zrozumienie zależności występujących w analizowanych danych.

Przeprowadzone badania wykazały, że metody uczenia maszynowego mogą stanowić wartościowe narzędzie wspierające zarówno proces odkrywania nowych substancji biologicznie aktywnych, jak i podejmowanie decyzji klinicznych. W perspektywie dalszych badań zasadne wydaje się rozszerzenie analiz na większe i bardziej zróżnicowane zbiory danych, wykorzystanie nowoczesnych architektur głębokiego uczenia oraz integracja danych pochodzących z wielu źródeł. Może to przyczynić się do dalszego zwiększenia dokładności predykcji oraz rozwoju spersonalizowanych rozwiązań wspomagających badania farmaceutyczne i medycynę precyzyjną.

Literatura

- [1] Aldasoro I, Gambacorta L, Pal R, Revoltella D, Weiss C, Wolski M. AI adoption, productivity and employment – Evidence from European firms. vol. 1325. Bank for International Settlements; 2026. <https://doi.org/doi/10.2867/1772538>.
- [2] Purificato E, Bili D, Jungnickel R, Ruiz-Serra V, Fabiani J, Abendroth-Dias K, et al. The role of artificial intelligence in scientific research – A science for policy, European perspective. Luxembourg: Publications Office of the European Union; 2025. <https://doi.org/doi/10.2760/7217497>.
- [3] Chowdhary KR. Fundamentals of artificial intelligence. 1st ed. Springer New Delhi; 2020. <https://doi.org/10.1007/978-81-322-3972-7>.
- [4] Pagel Walter. Paracelsus: An Introduction to Philosophical Medicine in the Era of the Renaissance. S.Karger AG; 1982. <https://doi.org/10.1159/ISBN.978-3-318-05013-4>.
- [5] Connor H. John Graunt F.R.S. (1620-74): The founding father of human demography, epidemiology and vital statistics. J Med Biogr 2022;32:57–69. <https://doi.org/10.1177/09677720221079826>.
- [6] Kelleher JD., Mac Namee B, D’Arcy A. Fundamentals of machine learning for predictive data analytics : algorithms, worked examples, and case studies. 1st ed. Cambridge, Massachusetts; London, England: The MIT Press; 2015. <https://doi.org/10.1007/978-81-322-3972-7>.
- [7] Ashley EA. Towards precision medicine. Nat Rev Genet 2016;17:507–22. <https://doi.org/10.1038/nrg.2016.86>.
- [8] Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. IBM J Res Dev 1959;3:210–29. <https://doi.org/10.1147/rd.33.0210>.
- [9] Goodfellow I, Bengio Y, Courville Aaron. Deep Learning. MIT Press; 2016.
- [10] Kubat Miroslav. An introduction to machine learning. 3rd ed. Springer Cham; 2021. <https://doi.org/10.1007/978-3-030-81935-4>.
- [11] Mitchell TM. Machine Learning. McGraw-Hill; 1997.
- [12] Nasteski V. An overview of the supervised machine learning methods. Horizons 2017;4:51–62. <https://doi.org/10.20544/HORIZONS.B.04.1.17.P05>.
- [13] James G, Witten D, Hastie T, Tibshirani R. Unsupervised Learning. An Introduction to Statistical Learning. Springer Texts in Statistics, New York: Springer; 2021, p. 497–552. https://doi.org/10.1007/978-1-0716-1418-1_12.
- [14] Breiman L, Friedman JH, Olsen RA, Stone CJ. Classification and Regression Trees. Chapter 7 1984:1–7.
- [15] Divakaran S. Data Science: Principles and Concepts in Modeling Decision Trees. Data Science in Agriculture and Natural Resource Management, 2021, p. 55–74. https://doi.org/10.1007/978-981-16-5847-1_3.

- [16] James G, Witten D, Hastie T, Tibshirani R. Tree-Based Methods. An Introduction to Statistical Learning, 2021, p. 327–65. https://doi.org/10.1007/978-1-0716-1418-1_8.
- [17] Loh W. Classification and regression trees. WIREs Data Mining and Knowledge Discovery 2011;1:14–23. <https://doi.org/10.1002/widm.8>.
- [18] Yohannes Y, Hoddinott JF. Classification and regression trees: an introduction. International Food Policy Research Institute; 1999. <https://doi.org/https://cgspace.cgiar.org/items/e6713c3d-0221-43c5-9caf-1ff7a78bb296>.
- [19] Molnar Christoph. Interpretable machine learning : a guide for making black box models explainable. Christoph Molnar; 2022.
- [20] Jolliffe IT, Cadima J. Principal component analysis: A review and recent developments. Philos Trans A Math Phys Eng Sci 2016;374. <https://doi.org/10.1098/rsta.2015.0202>.
- [21] Kherif F, Latypova A. Principal component analysis. Machine Learning: Methods and Applications to Brain Disorders, Elsevier; 2019, p. 209–25. <https://doi.org/10.1016/B978-0-12-815739-8.00012-2>.
- [22] James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning 2021. <https://doi.org/10.1007/978-1-0716-1418-1>.
- [23] Venkat N. The Curse of Dimensionality: Inside Out. 2018. <https://doi.org/10.13140/RG.2.2.29631.36006>.
- [24] Liton CP, Al Suman A, Sultan N. Methodological Analysis of Principal Component Analysis (PCA) Method. International Journal of Computational Engineering & Management 2013;16:32–8.
- [25] Hasan BMS, Abdulazeez AM. A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. Journal of Soft Computing and Data Mining 2021;2:20–30. <https://doi.org/10.30880/jscdm.2021.02.01.003>.
- [26] McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 1943;5:115–33. <https://doi.org/10.1007/BF02478259>.
- [27] Murtagh F. Multilayer perceptrons for classification and regression. Neurocomputing 1991;2:183–97. [https://doi.org/10.1016/0925-2312\(91\)90023-5](https://doi.org/10.1016/0925-2312(91)90023-5).
- [28] Tadeusiewicz R, Lula P. Wprowadzenie do sieci neuronowych. Kraków: StatSoft Polska Sp. z o.o. ; 2001.
- [29] Graupe D. Principles of Artificial Neural Networks. vol. 7. 3rd ed. Singapore: World Scientific Publishing Co. Pte. Ltd; 2013. <https://doi.org/10.1142/8868>.
- [30] Charilaou P, Battat R. Machine learning models and over-fitting considerations. World J Gastroenterol 2022;28:605. <https://doi.org/10.3748/WJG.V28.I5.605>.

- [31] Xu Y, Goodacre R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing* 2018 2:3 2018;2:249–62. <https://doi.org/10.1007/s41664-018-0068-2>.
- [32] Sambasivan N, Kapania S, Highfill H. Everyone wants to do the model work, not the data work: Data cascades in high-stakes ai. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* 2021. <https://doi.org/10.1145/3411764.3445518>.
- [33] Mohammed S, Budach L, Feuerpfeil M, Ihde N, Nathansen A, Noack N, et al. The effects of data quality on machine learning performance on tabular data. *Inf Syst* 2025;132. <https://doi.org/10.1016/J.IS.2025.102549>.
- [34] Popescu M-C, Balas VE, Perescu-Popescu L, Mastokaris N. Multilayer Perceptron and Neural Networks. *Wseas Transactions on Circuits and Systems* 2009;8:579–88.
- [35] Taud H, Mas JF. *Multilayer Perceptron (MLP). Geomatic Approaches for Modeling Land Change Scenarios*. 1st ed., Springer Cham; 2018, p. 451–5. https://doi.org/10.1007/978-3-319-60801-3_27.
- [36] Wilamowski BM. Neural network architectures and learning algorithms. *IEEE Ind Electron Mag* 2009;3:56–63. <https://doi.org/10.1109/MIE.2009.934790>.
- [37] Bishop CM. *Pattern recognition and machine learning*. 1st ed. Springer New York; 2009.
- [38] Hastie Trevor, Tibshirani Robert, Friedman JH. *The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations*. 2nd ed. Springer New York; 2009. <https://doi.org/https://doi.org/10.1007/978-0-387-84858-7>.
- [39] Lederer J. *Activation Functions in Artificial Neural Networks: A Systematic Overview* 2021. <https://doi.org/https://doi.org/10.48550/arXiv.2101.09957>.
- [40] Sharma S, Sharma S, Athaiya A. Activation function in neural networks. *International Journal of Engineering Applied Sciences and Technology* 2020;4:310–6. <https://doi.org/10.33564/IJEAST.2020.v04i12.054>.
- [41] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 1986;323:533–6. <https://doi.org/10.1038/323533a0>.
- [42] Werbos PJ. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. Harvard University, 1974.
- [43] Parker DB. *Learning-Logic: Casting the Cortex of the Human Brain in Silicon*. Cambridge: 1985.
- [44] Schmidhuber J. Deep learning in neural networks: An overview. *Neural Networks* 2015;61:85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>.

- [45] Shewchuk JR. An Introduction to the Conjugate Gradient Method Without the Agonizing Pain. Pittsburgh: 1994.
- [46] Nazareth JL. Conjugate gradient method. *Wiley Interdiscip Rev Comput Stat* 2009;1:348–53. <https://doi.org/10.1002/wics.13>.
- [47] Nocedal J, Wright SJ. Conjugate Gradient Methods. *Numerical Optimization. Springer Series in Operations Research and Financial Engineering.*, Springer New York; 2006. <https://doi.org/10.1007/978-0-387-40065-5>.
- [48] Møller MF. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks* 1993;6:525–33. [https://doi.org/10.1016/S0893-6080\(05\)80056-5](https://doi.org/10.1016/S0893-6080(05)80056-5).
- [49] Nocedal J, Wright SJ. Quasi-Newton Methods. *Numerical Optimization. Springer Series in Operations Research and Financial Engineering.*, Springer New York; 2006, p. 135–63. https://doi.org/10.1007/978-0-387-40065-5_6.
- [50] Kelley CT. The BFGS Method. *Iterative Methods for Optimization*, Society for Industrial and Applied Mathematics; 1999, p. 71–86. <https://doi.org/10.1137/1.9781611970920.ch4>.
- [51] Lawrence S, Tsoi AC, Giles CL. Lessons in Neural Network Training: Overfitting May be Harder than Expected. *Fourteenth National Conference on Artificial Intelligence*, California: AAAI Press; 1997, p. 540–5.
- [52] Prechelt L. Early Stopping — But When? *Neural Networks: Tricks of the Trade*. 2nd ed., Springer, Berlin, Heidelberg; 2012, p. 53–67. https://doi.org/10.1007/978-3-642-35289-8_5.
- [53] Bates S, Hastie T, Tibshirani R. Cross-Validation: What Does It Estimate and How Well Does It Do It? *J Am Stat Assoc* 2023. <https://doi.org/10.1080/01621459.2023.2197686>.
- [54] Zhou ZH. Ensemble methods: Foundations and algorithms. *Ensemble Methods: Foundations and Algorithms* 2012:1–218. <https://doi.org/10.1201/b12207>.
- [55] Opitz D, Maclin R. Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research* 1999;11:169–98. <https://doi.org/10.1613/jair.614>.
- [56] Hansen LK, Salamon P. Neural Network Ensembles. *IEEE Trans Pattern Anal Mach Intell* 1990;12:993–1001. <https://doi.org/10.1109/34.58871>.
- [57] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Networks* 1989;2:359–66. [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8).
- [58] Talete srl. Dragon (software for molecular descriptor calculation) version 7.0.6 2016.
- [59] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit* 1997;30:1145–59. [https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2).

Spis rycin

Rycina 1. Relacje hierarchiczne między sztuczną inteligencją (AI), uczeniem maszynowym (ML), sieciami neuronowymi oraz uczeniem głębokim (DL). Opracowanie własne na podstawie: Goodfellow i in. [9].....	9
Rycina 2. Schemat struktury drzewa decyzyjnego przedstawiający węzeł korzeniowy, węzły decyzyjne oraz liście końcowe. Węzeł korzeniowy inicjuje proces podziału danych, węzły decyzyjne odpowiadają za kolejne rozgałęzienia na podstawie wartości zmiennych, natomiast liście reprezentują ostateczne decyzje klasyfikacyjne lub wartości predykcyjne.	12
Rycina 3. Schemat architektury sztucznej sieci neuronowej typu MLP.	17
Rycina 4. Macierz pomyłek dla problemu klasyfikacji binarnej.	31
Rycina 5. Abstrakt graficzny przedstawiający proces klasyfikacji aktywności przeciwdrobnoustrojowej 140 homologów chlorków imidazoliowych wobec <i>Klebsiella pneumoniae</i>	35
Rycina 6. Abstrakt graficzny przedstawiający hybrydowe zastosowanie metod ML do symultanicznej predykcji aktywności przeciwdrobnoustrojowej 126 homologów chlorków imidazoliowych wobec <i>Staphylococcus aureus</i> i <i>Candida albicans</i>	37
Rycina 7. Abstrakt graficzny przedstawiający zastosowanie sztucznej sieci neuronowej do predykcji wznowy raka płuca na podstawie retrospektywnych danych klinicznych 2 296 pacjentów hospitalizowanych w latach 2012-2017 w Warmińsko-Mazurskim Centrum Chorób Płuc.	41



Antimicrobial Activity Classification of Imidazolium Derivatives Predicted by Artificial Neural Networks

Andżelika Lorenc¹ · Anna Badura¹ · Maciej Karolak² · Łukasz Pałkowski² · Łukasz Kubik³ · Adam Buciński¹

Received: 14 November 2023 / Accepted: 9 April 2024
© The Author(s) 2024

Abstract

Purpose This study assesses the Multilayer Perceptron (MLP) neural network, complemented by other Machine Learning techniques (CART, PCA), in predicting the antimicrobial activity of 140 newly designed imidazolium chlorides against *Klebsiella pneumoniae* before synthesis. Emphasis is on leveraging molecular properties for predictive analysis.

Methods Classification and regression decision trees (CART) identified the top 200 predictive molecular descriptors. Principal Component Analysis (PCA) reduced these descriptors to 5 components, retaining 99.57% of raw data information. Antimicrobial activity, categorized as high or low, was based on experimentally proven minimal inhibitory concentration (MIC), with a cut-point at MIC = 0.856 mol/L. A 12-fold cross-validation trained the MLP (architecture 5-12-2 with 5 Principal Components).

Results The MLP exhibited commendable performance, achieving almost 90% correct classifications across learning, validation, and test sets, outperforming models without PCA dimension reduction. Key metrics, including accuracy (0.907), sensitivity (0.905), specificity (0.909), and precision (0.891), were notably high. These results highlight the MLP model's efficacy with PCA as a high-quality classifier for determining antimicrobial activity.

Conclusions The study concludes that the MLP neural network, along with CART and PCA, is a robust tool for predicting the antimicrobial activity class of imidazolium chlorides against *Klebsiella pneumoniae*. CART and PCA, used in this study, allowed input variable reduction without significant information loss. High classification accuracy and associated metrics affirm the method's potential utility in pre-synthesis assessments, offering valuable insights for antimicrobial compound design.

Keywords artificial neural networks · classification · imidazolium derivatives · *klebsiella pneumoniae* · principal component analysis

✉ Andżelika Lorenc
andzelika.lorenc@cm.umk.pl

¹ Department of Biopharmacy, Faculty of Pharmacy, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, dr A. Jurasza 2, 85-089 Bydgoszcz, Poland

² Department of Pharmaceutical Technology, Faculty of Pharmacy, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, dr A. Jurasza 2, 85-089 Bydgoszcz, Poland

³ Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gen. J. Hallera 107, 80-416 Gdańsk, Poland

Published online: 17 April 2024

Introduction

Antibiotic resistance is a pressing concern as bacteria continue to develop resistance to antimicrobial substances employed in healthcare and industrial settings. This poses a significant threat to public health, as noted by the World Health Organization (WHO) and other reputable health organizations [1–4]. One of the most alarming microorganisms, with the ability to develop multi-resistance against antimicrobial agents, is *Klebsiella pneumoniae* [5]. This Gram-negative bacteria from the *Enterobacteriaceae* family was reported by the Centers for Disease Control and Prevention (CDC) as a high-risk pathogen due to its increasing rate of antibiotic resistance and potential to spread [6]. That trend of *K. pneumoniae* strains is especially dangerous in hospitals and other healthcare

institutions where sepsis and antiseptics have a huge impact on patients' outcomes and hospitalization time [7]. The problem increased due to large-scale disinfectant usage when the COVID-19 pandemic raised [8, 9]. The *K. pneumoniae*'s propensity to develop resistance against popular antibiotics and disinfection agents and the ability to create biofilm forces scientists to find new ways of searching for new antimicrobial compounds [10–12].

The identification of potential drugs can be a time-consuming and resource-intensive process. One of the problems in the area of searching for new active agents is the multitude of compounds that can be synthesized, not all of which will exhibit desired properties. However, this process can be streamlined and made more efficient, by leveraging computational chemistry and machine learning (ML) techniques in the initial research stage. This approach reduces the need for significant resources such as time, money, and chemicals. The ultimate goal is to identify the most promising compounds, which can then be further developed [13–16].

Traditional QSAR models may struggle to capture complex and non-linear relationships between chemical structures and antimicrobial activity while the non-linear nature of ML algorithms can represent these intricate connections. Those methods are able to extract patterns and insights directly from data without relying on predefined equations, making them more adaptive to the nuances of bis-imidazolium compounds' structure–activity relationships. It can also learn relevant features from the data, potentially uncovering subtle structural elements that contribute to antimicrobial activity. This is in contrast to QSAR, where feature selection is often a manual and hypothesis-driven process [17, 18].

Among various ML techniques used in drug research and development, the Artificial Neural Networks (ANNs) are one of the most promising. At the initial stages of molecular development, both 2D and 3D structures of compounds can be modeled, and subsequent to these models, molecular descriptors can be calculated. The obtained data can be used as predictors of antimicrobial activity by utilizing computational intelligence methods like ANNs [19–21]. This approach allows the preselection of potential antimicrobial compounds and the synthesis of only the most promising ones [22].

ANNs, widely employed in bioinformatics operate in a manner inspired by biological neural systems. These networks consist of interconnected artificial neurons organized into layers, and they become active under specific conditions. One prominent type of supervised ANN utilized in this field is the Multilayer Perceptron (MLP). The MLP is structured as a feed-forward network comprising a minimum of three layers of artificial neurons – input, (at least one) hidden, and output. Information flows from one layer to the next when it surpasses a predefined activation threshold

determined by the activation function within each layer's neurons [23, 24].

Since MLPs are supervised neural networks with known input–output pairs, the primary objective is to minimize the discrepancy between predicted and actual outcomes which is indicator of the predictive quality of the network. Achieving this with a single learning attempt is nearly impossible. Therefore, the training process of an MLP is based on epochs, which represent repeated cycles of presenting the entire dataset to the network in order to optimize its performance. During each epoch, the strengths of connections (weights) between neurons are recalibrated (starting from initial random values) based on the errors observed in previous epochs – the network learns on its own mistakes.. This iterative process continues until the network's error reaches its minimum, enhancing its predictive accuracy [25].

The problem which can occur in this process is overfitting, which means that the network overly adjusts to the data which deprives its ability to generalize the knowledge. In this case, the network will perform well on known data but the ability to correctly predict outcomes for the new data (generalization) is limited.

The reasons that the model overfits the data include sample size, input data dimensionality, or regularization techniques. A small number of cases taking part in the learning process can lead the model to treat the irrelevant information (noise) as important characteristics of presented data. As the prospects of to increase the number of cases are mostly limited, some kind of multiple sampling, for example, different types of cross-validation (CV), is suggested to avoid this problem and to optimize the model and enhance its performance by better hyperparameters tuning [26].

Another issue is the high dimensionality of the input data, known as the 'Curse of Dimensionality' [27]. This term, introduced by Bellman indicates that the rising number of features in the dataset should be followed by an exponentially rising number of samples to maintain the balance in the model. Similarly to the previous issue, as the number of samples cannot be enhanced, the way to avoid overfitting caused by excessive dimensionality is dimensional reduction. One of the methods applied as solution is Principal Component Analysis (PCA) which, based on correlations between features, calculates the Principal Components (PCs) describing as much information included in original features as possible [28].

An alternative method to mitigate overfitting involves the utilization of regularization techniques, with one such approach being early stopping. This strategy involves the introduction of a separate dataset, referred to as a validation set. The validation set containing new, never presented to the network cases predicts the solution based on the network architecture previously built using the learning set. The function of the validation set is to check the network's

generalization ability by comparing the errors made by ANN in each epoch. When the errors made in both sets are comparable, the learning process is continued, but if the error made in the learning set decreases and the validation set error increases, the network overfits the learning data and the process should be stopped [29].

The only set which doesn't take part in the learning process is the test set. As the cases presented for this set weren't revealed for it in the learning and test stages of network modeling, the test set role is to verify the ability of the developed model to generalize its knowledge for newly presented data [30].

In this paper, the authors would like to present a pre-synthesis approach for the classification of bis-imidazolium derivatives using Artificial Neural Networks in combination with other machine-learning techniques. The mechanism of action of studied compounds is closely related to their chemical structure and physical properties. Compounds interact electrostatically with the negatively charged cell surfaces of microbes and surface active compounds easily penetrate through the protein-lipid biological membranes, causing disturbances in their structural and functional coherence. That make them potentially active against considered in this research *Klebsiella pneumoniae* strains.

Materials and Methods

Structures, Synthesis, and Molecular Descriptors

The authors decided to investigate the 140 novel bis-imidazolium compounds (quaternary ammonium salts) as potential agents [31, 32].

The structures of 140 analyzed imidazolium homologs differed in the length of the linker chain (n value—from 2 to 12 CH_2 groups) and the substituent chain (Fig. 1, Table 1) (full list of homologs available in Table S1 in supplementary material). Designed compounds were modeled into 2D and 3D structures as neutral compounds and the quantum mechanical Density Functional Theory (DFT), B3LYP method using Pople's 6-31G as basis set was applied to optimization. Optimization was performed in the implicit solvent model SCRF (Self-Consistent Reaction Field) with dielectric constant set to that of water, with the use of the PCM (Polarizable Continuum Model).

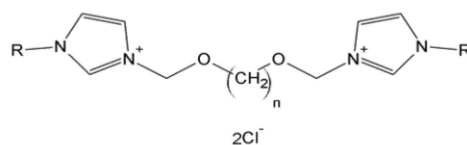


Fig. 1 General structure of analyzed imidazolium homologs.

Table 1 Informative System for 3,3'-(α, ω -Dioxaalkan)bis(1-Alkylimidazolium) Chlorides

Code	R
1	CH_3
2	C_2H_5
3	C_3H_7
4	C_4H_9
5	C_5H_{11}
6	C_6H_{13}
7	C_7H_{15}
8	C_8H_{17}
9	C_9H_{19}
10	$\text{C}_{10}\text{H}_{21}$
11	$\text{C}_{11}\text{H}_{23}$
12	$\text{C}_{12}\text{H}_{25}$
14	$\text{C}_{14}\text{H}_{29}$
16	$\text{C}_{16}\text{H}_{33}$

Calculations were performed employing Gaussian 09, Revision D.02 (Gaussian, Inc., Wallingford CT, USA), on a supercomputer cluster nodes with 12-core Intel® Xeon® E5 v3 2.3 GHz processors. Geometry optimization calculations were carried out at the Centre of Informatics—Tricity Academic Supercomputer & Network. In further analysis, the 5270 molecular descriptors from 29 logical blocks were determined using Dragon 7.0 software (Talete, Milan, Italy) [33].

The synthesis, molecules determination using nuclear magnetic resonance spectroscopy (NMR Spectra), purity examination with thin-layer chromatography (TLC), and elemental analysis of studied imidazolium compounds is described in article by Pałkowski *et al.* [34].

Antimicrobial Activity

The minimal inhibitory concentration (MIC) value of each homolog was determined for the *K. pneumoniae* ATCC 27853 strain based on references of Clinical and Laboratory Institute (CLSI) which was also conducted and reported by mentioned above Pałkowski *et al.* With reference to the standard, which was didecylmethylammonium chloride (DDAC) (a quaternary ammonium salt used for disinfection and approved by the European Economic Area as biocide [35]) with $\text{MIC} = 0.856 \text{ mol/L}$, imidazolium compounds were categorized to high activity or low activity group. Imidazolium compounds with MICs below the MIC of the DDAC were classified in the high activity category, and those above—low activity category. This division resulted in the separation of 64 compounds classified as high activity class (45.71%) and 76 compounds classified as low activity class (54.29%) (Table S1 in supplementary material).

Data Preprocessing

In the preprocessing stage, from the 5270 descriptors obtained in previous steps with Dragon 7.0 software the ones that were constant for all the cases and/or incomplete were deleted. That left 2711 variables for further analysis. In the next step, the data was standardized to minimize the influence of variables scaling and to make them comparable.

All the calculations in this research were conducted using STATISTICA 13, provided by StatSoft Inc., Tulsa, USA.

Descriptors Selection

As each of the 140 examined molecules was characterized by a set of 2711 molecular descriptors, a volume of data deemed excessive for meaningful analysis within the framework of Artificial Neural Networks (ANN), so the authors opted to employ different ML techniques as a strategic approach to mitigate the challenges posed by highly probable overfitting issue.

1. Authors decided to use Classification and Regression Trees (CART) to select the best activity descriptive variables—descriptors. Based on chi-square statistics and p -value ($p < 0.001$) the trees have selected 200 best-fitted variables. Selected for this research 200 molecular descriptors belonged to 19 molecular blocs, and most of the descriptors were from the 2D matrix-based descriptors block (92 descriptors, 46%), 3D autocorrelations block (25 descriptors, 12.5%) and 2D autocorrelations (23 descriptors, 11.5%). (full list of selected descriptors available in Table 2 in supplementary material)
2. To limit the number of neurons and avoid data overload in the input layer, the principal component analysis (PCA) was applied. PCA is a linear method of dimension reduction, allowing to reduce the number of variables included in the network by searching the relationships between them. The correlated variables are transformed to save most of the variance of the reduced variables and create the new variable called the principal component (PC) [36] This way, 5 principal components describing 99.57% of the variability of the data contained in the 200 selected variables were extracted.

Artificial Neural Networks calculations

Obtained in previous steps 5 PCs were used as input variables to build ANN classification models and predict, based on previously presented conditions, whether the compound presents high activity or low activity against *K. pneumoniae*. Cases were randomly divided into 3 sets: learning—98 cases (70%), validation—28 cases (20%), and test—14 cases (10%). Using the Broyden-Fletcher-Goldfarb-Shanno

(BFGS) [37] learning algorithm, the software automatically modeled 500 artificial neural networks from which the one with the most optimal architecture, the best predictive ability, and the lowest error (the number of correctly classified cases) was chosen as principle ANN hyperparameters setting for further analysis.

Results

Model selection, metrics and evaluation

The chosen MLP neural network was built of 5 neurons in the input layer, 12 neurons in the hidden layer, and 2 neurons in the output layer. The graph of the modeled MLP with inputs, outputs, and activation functions for hidden and input layers is shown in Fig. 2.

As the sample size was relatively small, the 12-fold cross-validation (CV) sampling technique has been applied. The percentage of correctly classified cases in every CV fold is shown in Fig. 3.

The quality of the learning, validation, and test set was considered as an average of models obtained in the cross-validation step and established at 88.61%, 90.48%, and 86.90% respectively, which means that the model correctly classified almost 90% of cases in each set. To evaluate the quality of modeled ANN the classification metrics and statistics were calculated (Table II).

To compare the performance of MLPs with and without PCA dimension reduction, the authors constructed various MLP models with different input configurations. The minimum number of inputs for MLPs without PCA was determined by the number of inputs used in MLPs with dimension reduction. The maximum number of inputs was constrained to not exceed 10% of the total cases (given the dataset size of 140 cases, the maximum number of inputs was set at 14).

The inputs were selected from the top of the list of the 200 most significant variables (according to p -value) derived initially with the CART method used for PCA dimension reduction. These variables were not subjected to further reduction and were directly used as the inputs for MLPs. The resulting models are presented in Table III.

Discussion

Current reports show that effectiveness in the search for new antimicrobial agents is estimated at around 5% [38]. Released in 2015 in 'Review of Antimicrobial Resistance' [39] pointed out that the number of drugs under development showing notable antimicrobial activity, especially against Gram-negative bacteria with broaden resistance is

Fig. 2 MLP 5-12-2 graph

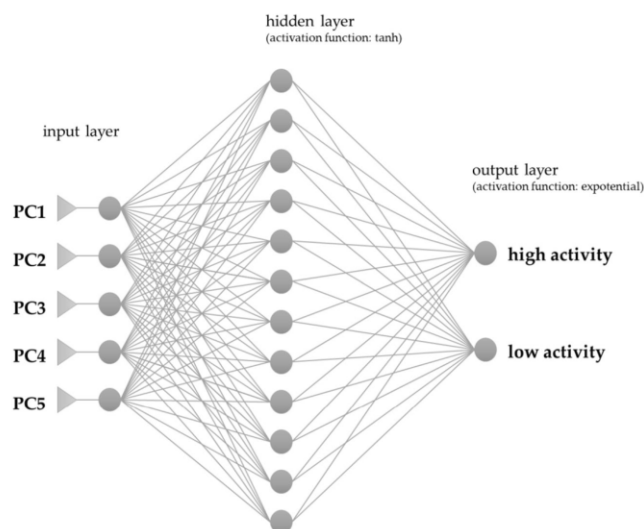
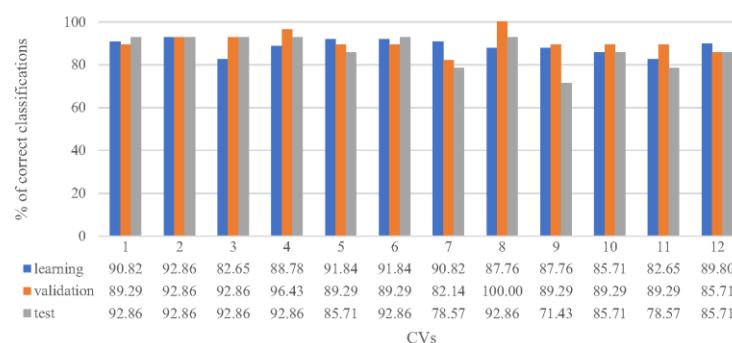


Fig. 3 Histogram and table of classification correctness (%) of learning, validation, and test sets



desperately low and it is estimated that among the vast of drugs under development only a few show potential against most resistant pathogens. In the mentioned review authors also pointed out that lowering the costs of antibiotics development is one of the major issues to focus on.

The majority of research is limited by financials, hence the most important problem seems to be the disproportion between the costs of the development of new drugs and the resulting benefits [40]. Implementing *in silico* methods in the preliminary selection of modeled drugs allows excluding the compounds with low antibacterial properties, which improves the whole process by minimizing the number of synthesized compounds, reducing costs, and lowering chemical waste as the methods of searching for new antimicrobial agents are hindered by vast of financial

and ecological restrictions nowadays. The combination of molecular modeling and ML methods seems to have a high potential for usage in the preclinical stage of drug research [41] limiting the needed resources. That also has been investigated on a large-scale dataset by Rahman *et al.* where they showed that the use of ML methods in the initial part of drug research can significantly increase the hit rate in searching for antimicrobial agents [42]. Research presented by Badura *et al.* shows that ANNs created to categorize 140 imidazolium compounds by their antimicrobial abilities have achieved over 90% accuracy for two bacterial strains – *E. coli* [43] and *S. aureus* [44] as well as for *C. albicans* fungus strain [45] using 20 molecular descriptors for each, but in contrast to the work presented in this paper, they worked on raw data.

Table II ANN Classification Model Metrics and Statistics

Model metrics			
Layer	Input	Hidden	Output
No. of neurons	5	12	2
Activation function		Tanh	Exponential
Model quality			
	Learning set	Validation set	Test set
Number of cases (%)	98 (70%)	28 (20%)	14 (10%)
Correctness (mean $\pm$ SD)	88.61 $\pm$ 3.45	90.48 $\pm$ 4.65	86.90 $\pm$ 7.36
Classification correctness			
	High activity	Low activity	All
General	64	76	140
Classified correctly	57 (89.06%)	70 (92.11%)	127 (90.71%)
Classified incorrectly	7 (10.94%)	6 (7.89%)	13 (9.29%)
Classification metrics			
	TPR	0.905	
	SPC	0.909	
	FPR	0.091	
	FDR	0.109	
	PPV	0.891	
	NPV	0.921	
	ACC	0.907	
	MCC	0.813	

The predictive ability of the selected in this research ANN allows classifying the compound to the group of high activity or low activity with almost 90% certainty. Used in the research MLP 5–12–2 neural network based on five Principal Components obtained good classification metrics with 0.905 sensitivity, 0.909 specificity, 0.891 precision, and 0.907 accuracy. The MCC which is considered as one of the most balanced classification metrics was established at 0.813. Also, FPR at the level of 0.091 shows a low probability of classifying compounds with low antimicrobial activity into the group of highly active homologs.

It should be kept in mind that appropriate feature selection is of great importance for developing accurate predictive models. The number of inputs should be adjusted to the analyzed dataset – with an increasing number of input features the risk of overfitting rises, but ANN with too small number of inputs could not gain the ability to generalize knowledge. Taking that into account, the authors used the PCA method for dimension reduction which allowed lowering the number of input variables to 5, maintaining over 99% of information carried by the 200 molecular descriptors selected with CART. This way of data condensation is widely used in computational intelligence problem-solving to avoid overfitting and increase algorithm performance speed. The concept of using PCA dimension reduction in combination with ML models is well known in medical sciences [46–49]. The paper by Chippalakatti *et al.*, comparing different classification models with and without PCA dimension reduction shows that those with PC as inputs reached better performance than models based on raw data which was also shown in this paper [50]. The results depicted in Table III reveal that none of the MLP models trained on the original descriptors data outperformed the predictive quality of the MLP model trained with PCA inputs. Remarkably, the MLP model constructed with only 5 principal components encompassing the data from the 200 molecular descriptors exhibited better activity prediction than larger models based on the raw data. That result reinforces the decision to use PCA as a dimension-reduction method.

Conclusions

The supervised MLP NN used in this research reached high values in predicting the activity class of presented imidazole compounds. The combination of different ML techniques, including dimension reduction with PCA for multivariable datasets elevated the model performance in comparison to

Table III Predictive Qualities Comparison of Different MLP Models

	Set quality			Error function	Activation function	
	Learning	Validation	Test		Hidden layer	Output layer
PCs MLP 5-6-2	88.61	90.48	86.90	SOS	Logistic	Tanh
MLP 5-5-2	83.25	84.23	80.95	Entropy	Exponential	Softmax
MLP 6-5-2	83.59	88.99	83.93	Entropy	Tanh	Softmax
MLP 7-16-2	82.65	89.29	83.93	Entropy	Tanh	Softmax
MLP 8-13-2	84.10	88.99	80.36	SOS	Logistic	Linear
MLP 9-13-2	82.91	89.29	82.14	SOS	Exponential	Logistic
MLP 10-16-2	84.18	90.48	79.17	Entropy	Tanh	Softmax
MLP 11-10-2	85.80	90.77	83.33	Entropy	Logistic	Softmax
MLP 12-3-2	84.27	88.69	86.31	Entropy	Tanh	Softmax
MLP 13-15-2	84.01	86.61	85.71	SOS	Tanh	Logistic
MLP 14-2-2	81.89	87.80	80.36	SOS	Logistic	Logistic

the models based on raw data.. The application of ML methods gives scientists the opportunity to study the antimicrobial properties of compounds even in the earliest design and development stages.

Furthermore, use of several different methods such as PCA for dimension reduction to limit the number of ANN inputs, the learning process early stopping method supported by a validation set and use of additional test set to evaluate the generalization ability of the network allowed to avoid overfitting problem maintaining as best as possible performance.

We are aware that artificial neural network models are designed to predict, in this case, antimicrobial properties of certain homogenous groups of molecules, and for another group, there will be a need to build a different model. To optimize the accuracy of our models and to improve their predictive abilities, it is essential to conduct a comprehensive analysis of a wider range of compounds. The research, however, shows that the NN models demonstrate high potential as a preliminary approach to selecting designed with computational methods molecules. In our opinion, the use of ANN as a tool for the selection of compounds with antibacterial potential can have a significant impact on the performance of the initial phase of drug development.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11095-024-03699-x>.

Author Contributions A.L. designed the project, performed the calculations, and wrote the manuscript. M.K., L.P. and E.K. provided data to analyze, supported and supervised the chemoinformatics part of the research. A.Badura and A.Buciński. supported and supervised the ML part of the research. All the authors reviewed the manuscript.

Funding Funding based on Elsevier and Nicolaus Copernicus University in Toruń (Elsevier SIS: 2055) agreement to provide Open Access publishing.

Data Availability The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of Interest The authors declare no competing financial interest.

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/>.

References

1. WHO Regional Office for Europe/European Centre for Disease Prevention and Control. Antimicrobial resistance surveillance in Europe 2022–2020 data. Copenhagen: WHO Regional Office for Europe; 2022.
2. Global antimicrobial resistance and use surveillance system (GLASS) report 2021. Geneva: World Health Organization; 2021. Licence: CC BY-NC-SA 3.0 IGO.
3. TalebiBezminAbadi A, Rizvanov AA, Haertlé T, Blatt NL. World Health Organization Report: Current Crisis of Antibiotic Resistance. Vol. 9, BioNanoScience. Springer New York LLC; 2019. p. 778–88.
4. Murray CJ, Ikuta KS, Sharara F, Swetschinski L, Robles Aguilar G, Gray A, *et al.* Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet*. 2022;399(10325):629–55.
5. Wang G, Zhao G, Chao X, Xie L, Wang H. The characteristic of virulence, biofilm and antibiotic resistance of *klebsiella pneumoniae*. *Int J Environ Res Public Health*. 2020;17:1–17.
6. CDC. Antibiotic Resistance Threats in the United States, 2019. Atlanta, GA: U.S. Department of Health and Human Services, CDC; 2019.
7. Naylor NR, Atun R, Zhu N, Kulasabanathan K, Silva S, Chatterjee A, *et al.* Estimating the burden of antimicrobial resistance: a systematic literature review. *Antimicrob Resist Infect Control*. 2018;7(58).
8. Getahun H, Smith I, Trivedi K, Paulin S, Balkhy HH. Tackling antimicrobial resistance in the COVID-19 pandemic. Vol. 98, Bulletin of the World Health Organization. World Health Organization; 2020.
9. Clancy CJ, Buehrle DJ, Nguyen MH. PRO: the COVID-19 pandemic will result in increased antimicrobial resistance rates. *JAC Antimicrobial Resistance*. 2020;2(3).
10. Navon-Venezia S, Kondratyeva K, Carattoli A. *Klebsiella pneumoniae*: a major worldwide source and shuttle for antibiotic resistance. *FEMS Microbiol Rev*. 2017;41(3):252–75.
11. Adler A, Katz DE, Marchaim D. The continuing Plague of extended-spectrum β -lactamase-producing *Enterobacteriaceae* Infections. *Infect Dis Clin N Am*. 2016;30(2):347–75.
12. Antoniadou A, Kontopidou F, Poulakou G, Koratzanis E, Galani I, Papadomichelakis E, *et al.* Colistin-resistant isolates of *Klebsiella pneumoniae* emerging in intensive care unit patients: first report of a multiclonal cluster. *J Antimicrob Chemother*. 2007;59(4):786–90.
13. Kore PP, Mutha MM, Antre RV, Oswal RJ, Kshirsagar SS. Computer-aided drug design: an innovative tool for modeling. *Open J Med Chem*. 2012;02(04):139–48.
14. Liota E, Spyrou G, Vassilatis D, Cournia Z. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr Top Med Chem*. 2014;14(16):1923–38.
15. Liu G, Stokes JM. A brief guide to machine learning for antibiotic discovery. *Curr Opin Microbiol*. 2022;1(69): 102190.
16. Macalino SJY, Gosu V, Hong S, Choi S. Role of computer-aided drug design in modern drug discovery. *Arch Pharm Res*. 2015;38(9):1686–701.
17. Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ. Machine learning in drug discovery: a review. *Artif Intell Rev*. 2022;55(3):1947–99.
18. Tsou LK, Yeh SH, Ueng SH, Chang CP, Song JS, Wu MH, *et al.* Comparative study between deep learning and QSAR classifications for TNBC inhibitors and novel GPCR agonist discovery. *Sci Rep*. 2020;10:1–11.
19. Grisoni F, Consonni V, Todeschini R. Impact of molecular descriptors on computational models. *Methods Mol Biol*. 2018;1825:171–209.

20. Mauri A, Consonni V, Todeschini R. Molecular descriptors. *Handbook of Computational Chemistry*. 2017;2065–93.
21. Tropsha A. Predictive quantitative structure–activity relationship modeling. *Compr Med Chem II*. 2007;4:149–65.
22. Torres MDT, de la Fuente-Nunez C. Toward computer-made artificial antibiotics. *Curr Opin Microbiol*. 2019;1(51):30–8.
23. Murtagh F. Multilayer perceptrons for classification and regression. *Neurocomputing*. 1991;2(5–6):183–97.
24. Popescu MC, Balas VE, Perescu-Popescu L, Mastokaris N. Multilayer perceptron and neural networks. *Wseas Transactions on Circuits and Systems*. 2009;8(7):579–88.
25. Camacho Olmedo MT et al. (eds). Geomatic approaches for modeling land change scenarios. *Lect Notes Geoinf Cartogr*. https://doi.org/10.1007/978-3-319-60801-3_27.
26. Bates S, Hastie T, Tibshirani R. Cross-validation: what does it estimate and how well does it do it? *J Am Stat Assoc*. 2023.
27. Bellman R. *Dynamic Programming*. Dover Publications; 1957.
28. Hasan BMS, Abdulazeez AM. A review of principal component analysis algorithm for dimensionality reduction. *J Soft Comput Data Min*. 2021;2(1):20–30.
29. Lawrence S, Tsoi AC, Giles CL. Lessons in neural network training: overfitting may be harder than expected. In: Fourteenth National Conference on Artificial Intelligence. AAAI Press; 1997. pp. 540–5.
30. Tadeusiewicz R, Lula P. *Wprowadzenie do sieci neuronowych*. Kraków: StatSoft Polska Sp. z o.o.; 2001.
31. Walsh SE, Maillard JY, Russell AD, Catrenich CE, Charbonneau DL, Bartolo RG. Activity and mechanisms of action of selected biocidal agents on Gram-positive and -negative bacteria. *J Appl Microbiol*. 2003;94(2):240–7.
32. Bharate SB, Thompson CM. Antimicrobial, antimalarial, and antileishmanial activities of mono- and bis-quaternary pyridinium compounds. *Chem Biol Drug Des*. 2010;76(6):546–51.
33. Pałkowski Ł, Karolak M, Błaszczyszński J, Krysiński J, Słowiński R. Structure-activity relationships of the imidazolium compounds as antibacterials of staphylococcus aureus and pseudomonas aeruginosa. *Int J Mol Sci*. 2021;22(15):7997.
34. Pałkowski Ł, Błaszczyszński J, Skrzypczak A, Błaszczak J, Kozakowska K, Wróblewska J, et al. Antimicrobial Activity and SAR Study of New Gemini Imidazolium-Based Chlorides. *Chem Biol Drug Res*. 2013;(83):278–88.
35. European Chemicals Agency. Imidazole - Substance Infocard. <https://echa.europa.eu/substance-information/-/substanceinfo/100.005.473>. Accessed 30 Sep 2023.
36. Kherif F, Latypova A. Chapter 12 - Principal component analysis. In: Mechelli A, Vieira S, editors. *Machine Learning*. Academic Press; 2020. p. 209–25.
37. Kelley CT. The BFGS Method. In: *Iterative methods for optimization*. Society for Industrial and Applied Mathematics; 1999. p. 71–86.
38. Payne DJ, Gwynn MN, Holmes DJ, Pompliano DL. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat Rev Drug Discov*. 2007;6:29–40.
39. The Review on Antimicrobial Resistance. Securing new drugs for future generations: the pipeline of antibiotics. https://amr-review.org/sites/default/files/SECURING%20NEW%20DRUGS%20FOR%20FUTURE%20GENERATIONS%20FINAL%20WEB_0.pdf. Accessed 2 Feb 2023.
40. Årdal C, Balasegaram M, Laxminarayan R, McAdams D, Outtersson K, Rex JH, et al. Antibiotic development — economic, regulatory and societal challenges. *Nat Rev Microbiol*. 2020;18(5):267–74.
41. Souza Leite M, Alles de Jesus A, Leite Araujo PJ, Ferreira dos Santos B. Chapter 4 - Advances in drug development with the application of artificial intelligence. In: Török M, editor. *Contemporary Chemical Approaches for Green and Sustainable Drugs*. Elsevier; 2022. p. 69–88. (Advances in Green and Sustainable Chemistry).
42. Zisanur Rahman ASM, Liu C, Sturm H, Hogan AM, Davis R, Hu P, et al. A machine learning model trained on a high-throughput antibacterial screen increases the hit rate of drug discovery. *PLoS Comput Biol*. 2022;18(10):e1010613.
43. Badura A, Krysiński J, Nowaczyk A, Buciński A. Application of artificial neural networks to prediction of new substances with antimicrobial activity against *Escherichia coli*. *J Appl Microbiol*. 2020;130:40–9.
44. Badura A, Krysiński J, Nowaczyk A, Buciński A. Prediction of the antimicrobial activity of quaternary ammonium salts against *Staphylococcus aureus* using artificial neural networks. *Arab J Chem*. 2021;14(7):103233.
45. Badura A, Krysiński J, Nowaczyk A, Buciński A. Application of artificial neural networks to the prediction of antifungal activity of imidazole derivatives against *Candida albicans*. *Chemom Intell Lab Syst*. 2022;15:222.
46. Sudharsan M, Thailambal G. Alzheimer's disease prediction using machine learning techniques and principal component analysis (PCA). *Mater Today Proc*. 2023;81:182–90.
47. Reddy KVV, Elamvazuthi I, Aziz AA, Paramasivam S, Chua HN. Heart disease risk prediction using machine learning with principal component analysis. 2020 8th International Conference on Intelligent and Advanced Systems (ICIAS), Kuching, Malaysia, 2021, pp. 1–6. <https://doi.org/10.1109/ICIAS49414.2021.9642676>.
48. Shimpi P, Shah S, Shroff M, Godbole A. A machine learning approach for the classification of cardiac arrhythmia. 2017 International Conference on Computing Methodologies and Communication (ICCMC), Erode, India; 2017. pp. 603–607. <https://doi.org/10.1109/ICCMC.2017.8282537>.
49. Jaques LE, Depoian AC, Xie D, Bailey CP, Guturu P. A machine learning approach to medical data identification through principal component analysis. *Big Data III: Learning, Analytics, and Applications*. SPIE. 2021;11730:7–15.
50. Chippalakatti S, Renumadhavi CH, Pallavi A. Comparative Review on the Machine Learning Algorithms for Medical Data. 2022 6th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bangalore, India; 2022. pp. 1–6. <https://doi.org/10.1109/CSITSS57437.2022.10026396>.

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

Artykuł nr 1 - materiał dodatkowy

Antimicrobial Activity Classification of Imidazolium Derivatives Predicted by Artificial Neural Networks

Andżelika Lorenc ^{a,*}, Anna Badura ^a, Maciej Karolak ^b, Łukasz Pałkowski ^b, Łukasz Kubik ^a, Adam Buciński ^a

Supplementary material

Table S1. Dragon 7.0 molecular descriptors selected with CART for PCA

No. Block	Name	Description
1 2D autocorrelations	ATS4v	Broto-Moreau autocorrelation of lag 4 (log function) weighted by van der Waals volume
2 2D autocorrelations	ATS6e	Broto-Moreau autocorrelation of lag 6 (log function) weighted by Sanderson electronegativity
3 2D autocorrelations	ATS6m	Broto-Moreau autocorrelation of lag 6 (log function) weighted by mass
4 2D autocorrelations	ATS7m	Broto-Moreau autocorrelation of lag 7 (log function) weighted by mass
5 2D autocorrelations	ATS7s	Broto-Moreau autocorrelation of lag 7 (log function) weighted by I-state
6 2D autocorrelations	ATS8e	Broto-Moreau autocorrelation of lag 8 (log function) weighted by Sanderson electronegativity
7 2D autocorrelations	ATS8v	Broto-Moreau autocorrelation of lag 8 (log function) weighted by van der Waals volume
8 2D autocorrelations	ATSC2e	Centred Broto-Moreau autocorrelation of lag 2 weighted by Sanderson electronegativity
9 2D autocorrelations	GATS1i	Geary autocorrelation of lag 1 weighted by ionization potential
10 2D autocorrelations	GATS1p	Geary autocorrelation of lag 1 weighted by polarizability
11 2D autocorrelations	JGI1	mean topological charge index of order 1
12 2D autocorrelations	JGI2	mean topological charge index of order 2
13 2D autocorrelations	JGI3	mean topological charge index of order 3
14 2D autocorrelations	MATS1i	Moran autocorrelation of lag 1 weighted by ionization potential
15 2D autocorrelations	MATS2i	Moran autocorrelation of lag 2 weighted by ionization potential
16 2D autocorrelations	MATS2v	Moran autocorrelation of lag 2 weighted by van der Waals volume
17 2D matrix-based descriptors	AVS_B(s)	average vertex sum from Burden matrix weighted by I-State
18 2D matrix-based descriptors	AVS_D	average vertex sum from topological distance matrix
19 2D matrix-based descriptors	AVS_Dz(e)	average vertex sum from Barysz matrix weighted by Sanderson electronegativity
20 2D matrix-based descriptors	AVS_Dz(i)	average vertex sum from Barysz matrix weighted by ionization potential
21 2D matrix-based descriptors	AVS_Dz(m)	average vertex sum from Barysz matrix weighted by mass
22 2D matrix-based descriptors	AVS_Dz(p)	average vertex sum from Barysz matrix weighted by polarizability
23 2D matrix-based descriptors	AVS_Dz(Z)	average vertex sum from Barysz matrix weighted by atomic number
24 2D matrix-based descriptors	Chi_Dz(p)	Randic-like index from Barysz matrix weighted by polarizability
25 2D matrix-based descriptors	Chi_Dz(v)	Randic-like index from Barysz matrix weighted by van der Waals volume
26 2D matrix-based descriptors	H_D/Dt	Harary-like index from distance/detour matrix
27 2D matrix-based descriptors	SM6_H2	spectral moment of order 6 from reciprocal squared distance matrix
28 2D matrix-based descriptors	SpAbs_Dz(e)	graph energy from Barysz matrix weighted by Sanderson electronegativity
29 2D matrix-based descriptors	SpAbs_Dz(i)	graph energy from Barysz matrix weighted by ionization potential
30 2D matrix-based descriptors	SpAbs_Dz(m)	graph energy from Barysz matrix weighted by mass
31 2D matrix-based descriptors	SpAbs_Dz(p)	graph energy from Barysz matrix weighted by polarizability
32 2D matrix-based descriptors	SpAbs_Dz(Z)	graph energy from Barysz matrix weighted by atomic number
33 2D matrix-based descriptors	SpAD_D	spectral absolute deviation from topological distance matrix
34 2D matrix-based descriptors	SpAD_Dz(e)	spectral absolute deviation from Barysz matrix weighted by Sanderson electronegativity
35 2D matrix-based descriptors	SpAD_Dz(i)	spectral absolute deviation from Barysz matrix weighted by ionization potential
36 2D matrix-based descriptors	SpAD_Dz(m)	spectral absolute deviation from Barysz matrix weighted by mass
37 2D matrix-based descriptors	SpAD_Dz(p)	spectral absolute deviation from Barysz matrix weighted by polarizability
38 2D matrix-based descriptors	SpAD_Dz(Z)	spectral absolute deviation from Barysz matrix weighted by atomic number
39 2D matrix-based descriptors	SpDiam_D	spectral diameter from topological distance matrix
40 2D matrix-based descriptors	SpDiam_Dz(e)	spectral diameter from Barysz matrix weighted by Sanderson electronegativity
41 2D matrix-based descriptors	SpDiam_Dz(i)	spectral diameter from Barysz matrix weighted by ionization potential
42 2D matrix-based descriptors	SpDiam_Dz(m)	spectral diameter from Barysz matrix weighted by mass
43 2D matrix-based descriptors	SpDiam_Dz(p)	spectral diameter from Barysz matrix weighted by polarizability
44 2D matrix-based descriptors	SpDiam_Dz(Z)	spectral diameter from Barysz matrix weighted by atomic number
45 2D matrix-based descriptors	SpMAD_B(i)	spectral mean absolute deviation from Burden matrix weighted by ionization potential
46 2D matrix-based descriptors	SpMAD_B(p)	spectral mean absolute deviation from Burden matrix weighted by polarizability
47 2D matrix-based descriptors	SpMAD_B(s)	spectral mean absolute deviation from Burden matrix weighted by I-State
48 2D matrix-based descriptors	SpMAD_B(v)	spectral mean absolute deviation from Burden matrix weighted by van der Waals volume
49 2D matrix-based descriptors	SpMAD_L	spectral mean absolute deviation from Laplace matrix
50 2D matrix-based descriptors	SpMax_D	leading eigenvalue from topological distance matrix

51	2D matrix-based descriptors	SpMax_Dz(e)	leading eigenvalue from Barysz matrix weighted by Sanderson electronegativity
52	2D matrix-based descriptors	SpMax_Dz(i)	leading eigenvalue from Barysz matrix weighted by ionization potential
53	2D matrix-based descriptors	SpMax_Dz(m)	leading eigenvalue from Barysz matrix weighted by mass
54	2D matrix-based descriptors	SpMax_Dz(p)	leading eigenvalue from Barysz matrix weighted by polarizability
55	2D matrix-based descriptors	SpMax_Dz(Z)	leading eigenvalue from Barysz matrix weighted by atomic number
56	2D matrix-based descriptors	SpMaxA_A	normalized leading eigenvalue from adjacency matrix
57	2D matrix-based descriptors	SpMaxA_B(e)	normalized leading eigenvalue from Burden matrix weighted by Sanderson electronegativity
58	2D matrix-based descriptors	SpMaxA_B(i)	normalized leading eigenvalue from Burden matrix weighted by ionization potential
59	2D matrix-based descriptors	SpMaxA_B(m)	normalized leading eigenvalue from Burden matrix weighted by mass
60	2D matrix-based descriptors	SpMaxA_B(p)	normalized leading eigenvalue from Burden matrix weighted by polarizability
61	2D matrix-based descriptors	SpMaxA_B(s)	normalized leading eigenvalue from Burden matrix weighted by I-State
62	2D matrix-based descriptors	SpMaxA_B(v)	normalized leading eigenvalue from Burden matrix weighted by van der Waals volume
63	2D matrix-based descriptors	SpMaxA_L	normalized leading eigenvalue from Laplace matrix
64	2D matrix-based descriptors	SpMaxA_X	normalized leading eigenvalue from chi matrix
65	2D matrix-based descriptors	SpPos_D	spectral positive sum from topological distance matrix
66	2D matrix-based descriptors	SpPos_Dz(e)	spectral positive sum from Barysz matrix weighted by Sanderson electronegativity
67	2D matrix-based descriptors	SpPos_Dz(i)	spectral positive sum from Barysz matrix weighted by ionization potential
68	2D matrix-based descriptors	SpPos_Dz(m)	spectral positive sum from Barysz matrix weighted by mass
69	2D matrix-based descriptors	SpPos_Dz(p)	spectral positive sum from Barysz matrix weighted by polarizability
70	2D matrix-based descriptors	SpPos_Dz(Z)	spectral positive sum from Barysz matrix weighted by atomic number
71	2D matrix-based descriptors	SpPosA_B(s)	normalized spectral positive sum from Burden matrix weighted by I-State
72	2D matrix-based descriptors	SpPosA_L	normalized spectral positive sum from Laplace matrix
73	2D matrix-based descriptors	TI1_L	first Mohar index from Laplace matrix
74	2D matrix-based descriptors	VE1_B(e)	coefficient sum of the last eigenvector from Burden matrix weighted by Sanderson electronegativity
75	2D matrix-based descriptors	VE1_B(i)	coefficient sum of the last eigenvector from Burden matrix weighted by ionization potential
76	2D matrix-based descriptors	VE1_B(m)	coefficient sum of the last eigenvector from Burden matrix weighted by mass
77	2D matrix-based descriptors	VE2_B(p)	average coefficient of the last eigenvector from Burden matrix weighted by polarizability
78	2D matrix-based descriptors	VE2_B(v)	average coefficient of the last eigenvector from Burden matrix weighted by van der Waals volume
79	2D matrix-based descriptors	VE2_Dz(e)	average coefficient of the last eigenvector from Barysz matrix weighted by Sanderson electronegativity
80	2D matrix-based descriptors	VE2_Dz(i)	average coefficient of the last eigenvector from Barysz matrix weighted by ionization potential
81	2D matrix-based descriptors	VE2_Dz(m)	average coefficient of the last eigenvector from Barysz matrix weighted by mass
82	2D matrix-based descriptors	VE2_Dz(Z)	average coefficient of the last eigenvector from Barysz matrix weighted by atomic number
83	2D matrix-based descriptors	VR2_A	normalized Randic-like eigenvector-based index from adjacency matrix
84	2D matrix-based descriptors	VR2_B(e)	normalized Randic-like eigenvector-based index from Burden matrix weighted by Sanderson electronegativity
85	2D matrix-based descriptors	VR2_B(i)	normalized Randic-like eigenvector-based index from Burden matrix weighted by ionization potential
86	2D matrix-based descriptors	VR2_B(m)	normalized Randic-like eigenvector-based index from Burden matrix weighted by mass
87	2D matrix-based descriptors	VR2_B(p)	normalized Randic-like eigenvector-based index from Burden matrix weighted by polarizability
88	2D matrix-based descriptors	VR2_B(s)	normalized Randic-like eigenvector-based index from Burden matrix weighted by I-State
89	2D matrix-based descriptors	VR2_B(v)	normalized Randic-like eigenvector-based index from Burden matrix weighted by van der Waals volume
90	2D matrix-based descriptors	VR2_D/Dt	normalized Randic-like eigenvector-based index from distance/detour matrix
91	2D matrix-based descriptors	VR2_H2	normalized Randic-like eigenvector-based index from reciprocal squared distance matrix
92	2D matrix-based descriptors	VR2_L	normalized Randic-like eigenvector-based index from Laplace matrix
93	2D matrix-based descriptors	Wi_Dz(v)	Wiener-like index from Barysz matrix weighted by van der Waals volume
94	2D matrix-based descriptors	WiA_B(e)	average Wiener-like index from Burden matrix weighted by Sanderson electronegativity
95	2D matrix-based descriptors	WiA_B(i)	average Wiener-like index from Burden matrix weighted by ionization potential
96	2D matrix-based descriptors	WiA_B(m)	average Wiener-like index from Burden matrix weighted by mass
97	2D matrix-based descriptors	WiA_B(p)	average Wiener-like index from Burden matrix weighted by polarizability
98	2D matrix-based descriptors	WiA_B(s)	average Wiener-like index from Burden matrix weighted by I-State
99	2D matrix-based descriptors	WiA_B(v)	average Wiener-like index from Burden matrix weighted by van der Waals volume
100	2D matrix-based descriptors	WiA_H2	average Wiener-like index from reciprocal squared distance matrix
101	3D autocorrelations	TDB01m	3D Topological distance based descriptors - lag 1 weighted by mass
102	3D autocorrelations	TDB01p	3D Topological distance based descriptors - lag 1 weighted by polarizability
103	3D autocorrelations	TDB01s	3D Topological distance based descriptors - lag 1 weighted by I-state
104	3D autocorrelations	TDB01v	3D Topological distance based descriptors - lag 1 weighted by van der Waals volume
105	3D autocorrelations	TDB02p	3D Topological distance based descriptors - lag 2 weighted by polarizability
106	3D autocorrelations	TDB02s	3D Topological distance based descriptors - lag 2 weighted by I-state
107	3D autocorrelations	TDB02u	3D Topological distance based descriptors - lag 2 unweighted
108	3D autocorrelations	TDB03e	3D Topological distance based descriptors - lag 3 weighted by Sanderson electronegativity
109	3D autocorrelations	TDB03i	3D Topological distance based descriptors - lag 3 weighted by ionization potential
110	3D autocorrelations	TDB03m	3D Topological distance based descriptors - lag 3 weighted by mass
111	3D autocorrelations	TDB03p	3D Topological distance based descriptors - lag 3 weighted by polarizability
112	3D autocorrelations	TDB03r	3D Topological distance based descriptors - lag 3 weighted by covalent radius

113	3D autocorrelations	TDB03u	3D Topological distance based descriptors - lag 3 unweighted
114	3D autocorrelations	TDB03v	3D Topological distance based descriptors - lag 3 weighted by van der Waals volume
115	3D autocorrelations	TDB04e	3D Topological distance based descriptors - lag 4 weighted by Sanderson electronegativity
116	3D autocorrelations	TDB04i	3D Topological distance based descriptors - lag 4 weighted by ionization potential
117	3D autocorrelations	TDB04m	3D Topological distance based descriptors - lag 4 weighted by mass
118	3D autocorrelations	TDB04p	3D Topological distance based descriptors - lag 4 weighted by polarizability
119	3D autocorrelations	TDB04r	3D Topological distance based descriptors - lag 4 weighted by covalent radius
120	3D autocorrelations	TDB04u	3D Topological distance based descriptors - lag 4 unweighted
121	3D autocorrelations	TDB04v	3D Topological distance based descriptors - lag 4 weighted by van der Waals volume
122	3D autocorrelations	TDB05e	3D Topological distance based descriptors - lag 5 weighted by Sanderson electronegativity
123	3D autocorrelations	TDB05s	3D Topological distance based descriptors - lag 5 weighted by I-state
124	3D autocorrelations	TDB05u	3D Topological distance based descriptors - lag 5 unweighted
125	3D autocorrelations	TDB06s	3D Topological distance based descriptors - lag 6 weighted by I-state
126	3D matrix-based descriptors	Chi_RG	Randic-like index from reciprocal squared geometrical matrix
127	3D matrix-based descriptors	H_G	Harary-like index from geometrical matrix
128	3D matrix-based descriptors	H_G/D	Harary-like index from distance/distance matrix
129	3D matrix-based descriptors	SM2_G/D	spectral moment of order 2 from distance/distance matrix
130	3D matrix-based descriptors	VR2_RG	normalized Randic-like eigenvector-based index from reciprocal squared geometrical matrix
131	3D matrix-based descriptors	WiA_RG	average Wiener-like index from reciprocal squared geometrical matrix
132	3D-MorSE descriptors	Mor01u	signal 01 / unweighted
133	Burden eigenvalues	SpMax1_Bh(m)	largest eigenvalue n. 1 of Burden matrix weighted by mass
134	CATS 2D	CATS2D_02_LL	CATS2D Lipophilic-Lipophilic at lag 02
135	CATS 2D	CATS2D_03_LL	CATS2D Lipophilic-Lipophilic at lag 03
136	CATS 3D	CATS3D_02_LL	CATS3D Lipophilic-Lipophilic BEN 02 (2.000 – 3.000 Å)
137	Charge descriptors	LDI	local dipole index
138	Connectivity indices	X0A	average connectivity index of order 0
139	Connectivity indices	X1A	average connectivity index of order 1
140	Constitutional indices	AMW	average molecular weight
141	Constitutional indices	GD	graph density
142	Constitutional indices	Me	mean atomic Sanderson electronegativity (scaled on Carbon atom)
143	Constitutional indices	Mv	mean atomic van der Waals volume (scaled on Carbon atom)
144	Constitutional indices	N%	percentage of N atoms
145	Constitutional indices	O%	percentage of O atoms
146	Edge adjacency indices	SM10_AEA(ed)	spectral moment of order 10 from augmented edge adjacency mat. weighted by edge degree
147	Edge adjacency indices	SM12_EA(bo)	spectral moment of order 12 from edge adjacency mat. weighted by bond order
148	Edge adjacency indices	SpMAD_EA(bo)	spectral mean absolute deviation from edge adjacency mat. weighted by bond order
149	Edge adjacency indices	SpMAD_EA(dm)	spectral mean absolute deviation from edge adjacency mat. weighted by dipole moment
150	Edge adjacency indices	SpMAD_EA(ed)	spectral mean absolute deviation from edge adjacency mat. weighted by edge degree
151	Edge adjacency indices	SpMaxA_AEA(bo)	normalized leading eigenvalue from augmented edge adjacency mat. weighted by bond order
152	Edge adjacency indices	SpMaxA_AEA(dm)	normalized leading eigenvalue from augmented edge adjacency mat. weighted by dipole moment
153	Edge adjacency indices	SpMaxA_AEA(ed)	normalized leading eigenvalue from augmented edge adjacency mat. weighted by edge degree
154	Edge adjacency indices	SpMaxA_AEA(ri)	normalized leading eigenvalue from augmented edge adjacency mat. weighted by resonance integral
155	Edge adjacency indices	SpMaxA_EA	normalized leading eigenvalue from edge adjacency mat.
156	Edge adjacency indices	SpMaxA_EA(bo)	normalized leading eigenvalue from edge adjacency mat. weighted by bond order
157	Edge adjacency indices	SpMaxA_EA(dm)	normalized leading eigenvalue from edge adjacency mat. weighted by dipole moment
158	Edge adjacency indices	SpMaxA_EA(ed)	normalized leading eigenvalue from edge adjacency mat. weighted by edge degree
159	Edge adjacency indices	SpMaxA_EA(ri)	normalized leading eigenvalue from edge adjacency mat. weighted by resonance integral
160	ETA indices	Eta_B_A	eta average branching index
161	ETA indices	Eta_beta_A	eta average VEM count
162	ETA indices	Eta_betaP_A	eta pi and lone pair average VEM count
163	ETA indices	Eta_betaS_A	eta sigma average VEM count
164	ETA indices	Eta_epsi_A	eta average electronegativity measure
165	ETA indices	Eta_FL_A	eta average local functionality index
166	ETA indices	Eta_sh_p	eta p shape index
167	ETA indices	Eta_sh_y	eta y shape index
168	GETAWAY descriptors	HATS5s	leverage-weighted autocorrelation of lag 5 / weighted by I-state
169	Information indices	AAC	mean information index on atomic composition
170	Information indices	BICO	Bond Information Content index (neighborhood symmetry of 0-order)
171	Information indices	Ges	Number of symmetry classes (based on electrotopological state)
172	Information indices	HDcpx	graph distance complexity index (log function)
173	Information indices	ICO	Information Content index (neighborhood symmetry of 0-order)
174	Information indices	IDET	total information content on the distance equality

175 Information indices	SIC0	Structural Information Content index (neighborhood symmetry of 0-order)
176 Molecular properties	BLTA96	Verhaar Algae base-line toxicity from MLOGP (mmol/l)
177 Molecular properties	BLTD48	Verhaar Daphnia base-line toxicity from MLOGP (mmol/l)
178 Molecular properties	BLTF96	Verhaar Fish base-line toxicity from MLOGP (mmol/l)
179 Molecular properties	Hy	hydrophilic factor
180 Molecular properties	MLOGP	Moriguchi octanol-water partition coeff. (logP)
181 Molecular properties	MLOGP2	squared Moriguchi octanol-water partition coeff. (logP ²)
182 Molecular properties	PDI	packing density index
183 P_VSA-like descriptors	P_VSA_MR_1	P_VSA-like on Molar Refractivity, bin 1
184 RDF descriptors	RDF030i	Radial Distribution Function - 030 / weighted by ionization potential
185 RDF descriptors	RDF155s	Radial Distribution Function - 155 / weighted by I-state
186 Ring descriptors	MCD	molecular cyclized degree
187 Topological indices	CENT	centralization
188 Topological indices	Gnar	Narumi geometric topological index
189 Topological indices	Psi_e_A	electrotopological state pseudoconnectivity index - type S average
190 Topological indices	Psi_i_A	intrinsic state pseudoconnectivity index - type S average
191 Topological indices	TIE	E-state topological parameter
192 Topological indices	UNIP	unipolarity
193 Topological indices	VAR	variation
194 Topological indices	Wap	all-path Wiener index
195 Walk and path counts	MPC07	molecular path count of order 7
196 Walk and path counts	MPC09	molecular path count of order 9
197 Walk and path counts	MWC08	molecular walk count of order 8
198 Walk and path counts	piPC06	molecular multiple path count of order 6
199 Walk and path counts	piPC08	molecular multiple path count of order 8
200 Walk and path counts	TPC	total path count

Table S2. Minimal inhibitory concentration and activity of 140 3,3'-(α , ω -dioxalkane)bis(1-alkylimidazolium) chlorides

object	MIC		activity	object	MIC		activity	object	MIC		activity	object	MIC		activity	object	MIC		activity
	n	R			n	R			n	R			n	R			n	R	
1	2	1	33.875	low	29	4	1	28.468	low	57	6	1	3.295	low	85	8	1	0.919	low
2	2	2	28.468	low	30	4	2	26.363	low	58	6	2	26.360	low	86	8	2	1.697	low
3	2	3	13.181	low	31	4	3	26.360	low	59	6	3	22.966	low	87	8	3	0.788	high
4	2	4	6.590	low	32	4	4	22.966	low	60	6	4	21.576	low	88	8	4	2.966	low
5	2	5	5.741	low	33	4	5	5.394	low	61	6	5	10.172	low	89	8	5	0.178	high
6	2	6	5.394	low	34	4	6	1.261	low	62	6	6	0.597	high	90	8	6	0.168	high
7	2	7	2.543	low	35	4	7	0.308	high	63	6	7	0.292	high	91	8	7	0.079	high
8	2	8	0.597	high	36	4	8	0.146	high	64	6	8	0.278	high	92	8	8	0.037	high
9	2	9	0.566	high	37	4	9	0.069	high	65	6	9	0.133	high	93	8	9	0.071	high
10	2	10	0.538	high	38	4	10	0.133	high	66	6	10	0.253	high	94	8	10	0.068	high
11	2	11	0.513	high	39	4	11	0.063	high	67	6	11	0.030	high	95	8	11	0.259	high
12	2	12	0.981	low	40	4	12	0.940	low	68	6	12	0.233	high	96	8	12	0.960	low
13	2	13	1.817	low	41	4	13	0.433	high	69	6	13	0.417	high	97	8	13	0.883	low
14	2	14	1.680	low	42	4	14	0.052	high	70	6	14	3.125	low	98	8	14	0.409	high
15	3	1	29.652	low	43	5	1	27.375	low	71	7	1	16.937	low	99	9	1	7.117	low
16	3	2	27.375	low	44	5	2	25.423	low	72	7	2	7.117	low	100	9	2	6.591	low
17	3	3	25.423	low	45	5	3	23.730	low	73	7	3	6.591	low	101	9	3	6.590	low
18	3	4	5.933	low	46	5	4	2.781	low	74	7	4	6.590	low	102	9	4	2.871	low
19	3	5	11.125	low	47	5	5	20.942	low	75	7	5	1.424	low	103	9	5	0.345	high
20	3	6	5.235	low	48	5	6	4.945	low	76	7	6	0.669	high	104	9	6	0.326	high
21	3	7	0.613	high	49	5	7	18.74	low	77	7	7	0.041	high	105	9	7	0.154	high
22	3	8	0.581	high	50	5	8	0.285	high	78	7	8	0.038	high	106	9	8	0.073	high
23	3	9	0.552	high	51	5	9	2.120	low	79	7	9	0.146	high	107	9	9	0.278	high
24	3	10	0.271	high	52	5	10	0.259	high	80	7	10	0.017	high	108	9	10	0.265	high
25	3	11	0.502	high	53	5	11	0.062	high	81	7	11	0.513	high	109	9	11	0.981	low
26	3	12	0.480	high	54	5	12	0.119	high	82	7	12	0.491	high	110	9	12	3.789	low
27	3	13	7.123	low	55	5	13	0.425	high	83	7	13	0.451	high	111	9	13	3.492	low
28	3	14	6.596	low	56	5	14	1.590	low	84	7	14	0.833	high	112	9	14	6.476	low



Hybrid Machine Learning Model for Dual Prediction of Antimicrobial Activity against *Staphylococcus aureus* and *Candida albicans*

Andżelika Lorenc¹ · Anna Badura¹ · Maciej Karolak² · Łukasz Pałkowski² · Łukasz Kubik³ · Andrzej Skrzypczak⁴ · Adam Buciński¹

Received: 5 August 2025 / Accepted: 24 January 2026

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2026

Abstract

Antimicrobial resistance poses a major global health challenge, particularly in infections involving *Staphylococcus aureus* and *Candida albicans*, which frequently occur together and exhibit synergistic pathogenicity. Rapid and efficient discovery of novel antimicrobials is essential and must align with sustainable, green chemistry practices. Here, we present a hybrid machine learning (ML) workflow for the simultaneous prediction of the logarithm of the minimal inhibitory concentration (ln(MIC)) of 126 bis-imidazolium chlorides against both pathogens. From 5,270 molecular descriptors, a multistage reduction pipeline—including redundancy filtering, classification and regression tree analysis, and principal component analysis—yielded twelve principal components explaining 91.7% of the variance. These were used as inputs for an ensemble of multilayer perceptron artificial neural networks (12-11-2 architecture, tanh activation for both hidden and output layers), enhancing predictive stability and generalizability. The resulting models achieved correlation coefficients of $r=0.966$ for *S. aureus* and $r=0.954$ for *C. albicans*, with ensemble modeling particularly improving the fungal predictions. This study demonstrates that even with limited data, a carefully designed hybrid ML approach can provide accurate, dual-target antimicrobial predictions, offering a computationally efficient and interpretable tool to accelerate antimicrobial screening while supporting sustainable drug discovery.

Highlights

- Predicts ln(MIC) for *S. aureus* and *C. albicans* simultaneously using a single workflow.
- CART and PCA reduced 5,270 descriptors to twelve principal components, capturing 91.7% of the variance.
- An optimized MLP-ANN (12-11-2) achieved encouraging predictive performance for both pathogens.
- Ensemble MLPs improved robustness, particularly for *C. albicans* predictions.
- The workflow illustrates a resource-efficient strategy aligned with green chemistry principles.

Keywords Artificial neural networks · Principal component analysis · Classification and regression trees · Imidazolium derivatives · *Staphylococcus aureus* · *Candida albicans*

✉ Andżelika Lorenc
andzelika.lorenc@cm.umk.pl

Anna Badura
anna.badura@cm.umk.pl

Maciej Karolak
maciej.karolak@cm.umk.pl

Łukasz Pałkowski
lukaszpałkowski@cm.umk.pl

Łukasz Kubik
lukasz.kubik@gumed.edu.pl

Andrzej Skrzypczak
andrzej.skrzypczak@put.poznan.pl

Adam Buciński
adam.bucinski@cm.umk.pl

¹ Department of Biopharmacy, Faculty of Pharmacy, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University, Toruń, Poland

² Department of Pharmaceutical Technology, Faculty of Pharmacy, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University, Toruń, Poland

³ Department of Biopharmaceutics and Pharmacodynamics, Medical University of Gdańsk, Gdańsk, Poland

⁴ Faculty of Chemical Technology, Poznań University of Technology, Poznań, Poland

Published online: 05 February 2026

Springer

Introduction

Artificial intelligence (AI) and machine learning (ML) are increasingly permeating healthcare, offering powerful tools to accelerate drug discovery and support clinical decision-making. One of the most pressing challenges where such tools are urgently needed is the global rise of antimicrobial resistance. The effectiveness of well-established antimicrobial compounds has steadily declined due to the rapid adaptability and genetic evolution of microorganisms [1]. At the same time, concerns about chemical waste and sustainability highlight the importance of developing efficient approaches that align with the principles of green chemistry [2–4]. By reducing experimental burden, ML offers an attractive route to accelerate antimicrobial discovery while minimizing resource use [5, 6].

In 2024, the World Health Organization (WHO) classified methicillin-resistant *Staphylococcus aureus* (MRSA) as a high-priority threat because of its global impact on healthcare-associated and community-acquired infections [7]. In parallel, *Candida albicans*, a common opportunistic fungus, has exhibited increasing resistance to antifungals in recent years [8, 9]. Even more alarming are coinfections involving these two species, which are frequently encountered in intra-abdominal [10] and bloodstream infections [11, 12], and are increasingly recognized in intensive-care and post-surgical settings [13], where synergistic biofilm formation [14], and metabolic cooperation [15, 16] amplify virulence and drug tolerance, resulting in more complicated therapy. Moreover, mortality rates for mixed bacterial–fungal infections are higher than for mono-infections [17, 18].

Given the risk posed by such infections, effective disinfection strategies are crucial, particularly in high-risk environments such as hospitals. Poly-cationic quaternary ammonium compounds (QACs), and in particular bis-imidazolium salts, represent a promising class of disinfectants with both antibacterial and antifungal activity. They have demonstrated strong potency and a lower likelihood of resistance development compared with mono-cationic QACs [19, 20]. Still, their systematic evaluation against diverse microbial species remains experimentally demanding. Computational modeling approaches could reduce this burden, offering a sustainable strategy for prioritizing promising compounds.

Despite considerable progress in quantitative structure–activity relationship (QSAR) modeling, several gaps remain. First, the majority of studies predict antimicrobial activity against a single microorganism, whereas coinfections involving *S. aureus* and *C. albicans* are increasingly recognized as clinically severe due to their synergistic pathogenicity. Predictive models that address both pathogens simultaneously are scarce. Second, most machine

learning approaches in QSAR require large datasets, yet in many cases - including this study - only a limited number of compounds are available, which increases the risk of overfitting. Third, the high dimensionality of molecular descriptors presents a methodological challenge; while either feature selection or dimensionality reduction is commonly applied, integrated hybrid approaches are less frequently explored. Finally, although deep learning methods can capture complex relationships, they typically demand extensive data and computational resources and are often criticized as “black-box” approaches. Thus, there is a need for efficient, interpretable, and resource-conscious workflows consistent with the principles of green chemistry.

Materials and Methods

Bis-Imidazolium Homologs

The study material consisted of a set of 5,270 molecular descriptors calculated for 126 chlorides of α,ω -bis[(alkyl)imidazol-1-yl-3-alkoxymethyl], comprising 9 homologous series differing in the length of the linker chain (n value— from 4 to 12 CH₂ groups) and the substituent chain (R).

The structural details, synthesis, conditions of MIC determination and values, and computational details of the molecular descriptor calculations for the analyzed bis-imidazolium homologs were determined via standardized methodologies and are provided in the Supporting Information— sections: (1) The structural formula of the α,ω -bis[(alkyl)imidazol-1-yl-3-alkoxymethyl] chlorides, (2) Synthesis and MIC determination of bis-imidazolium homologs, and (3) Molecular descriptor calculations.

Data Preprocessing

The dataset used in this study comprised numerical information corresponding to approximately 5,270 molecular descriptors for 126 α,ω -bis[(alkyl)imidazol-1-yl-3-alkoxymethyl] chlorides. Data preprocessing was performed using Python 3.10.11 with the NumPy 1.26.0, Pandas 2.1.1, and Scikit-learn 1.1.3 libraries (preprocessing script provided in the Supporting Information – 4. Preprocessing Script). Preprocessing involved the removal of variables exhibiting zero variance or containing missing values (NaN). Following this step, the dataset was reduced to 2,667 variables, which were further refined to 437 variables by eliminating highly correlated descriptors (correlation threshold $r=0.95$). Subsequently, Classification and Regression Tree (CART) analysis was employed to identify the 150 most relevant descriptors based on F-statistics and associated p-values ($p<0.05$) for *S. aureus* and *C. albicans* individually. Both

Table 1 Summary of variance explained by the twelve principal components (PCs). The R2X column indicates the proportion of variance explained by each individual component, whereas the cumulative R2X column shows the cumulative proportion of variance retained across components

PC	R2X	Cumulated R2X
1	0.507	0.507
2	0.098	0.605
3	0.079	0.684
4	0.063	0.746
5	0.042	0.788
6	0.029	0.817
7	0.025	0.842
8	0.021	0.863
9	0.019	0.882
10	0.013	0.895
11	0.012	0.907
12	0.010	0.917

Table 2 Summary of the hyperparameters used for training the ANN model

Hyperparameter	Values
Neurons in Hidden Layer	3–20
Activation function (hidden, output)	linear, logistic, hyperbolic tangent (tanh), exponential, or sinusoidal
Regularization	Early stopping determined by: - Minimal error change: 10^{-8} - Epochs without change: 20

lists were then compared, and only the descriptors common to *S. aureus* and *C. albicans* ($N=54$) were retained, ensuring methodological consistency for simultaneous modeling. The complete list of descriptors used for Principal Component Analysis (PCA) is provided in Table S3 [Supporting Information].

Dimensionality Reduction with Principal Component Analysis

Owing to the high dimensionality of the preprocessed dataset, PCA was applied, reducing the 54 variables obtained in the previous step to 12 PCs, which explained 91.7% of the original variance (0.917 cumulative R2X, where R2X refers to the fraction of variance described by the PCs) (Table 1).

Multilayer Perceptron Model Optimization

Given the nonlinear and interdependent nature [Supplementary Information – 5. Correlation Analysis Between Principal Components and Antimicrobial Activity] of the resulting features, an artificial neural network (ANN) - specifically a

Multilayer Perceptron (MLP) - was selected as the predictive model.

Table 2 presents the key hyperparameters applied during the training of the ANN. The model was optimized using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [21]. The number of neurons in the input layer ranged from 3 to 12, corresponding to the number of PCs obtained through PCA. The hidden layer comprised between 3 and 20 neurons, while the output layer consisted of two neurons representing the predicted $\ln(\text{MIC})$ values for *S. aureus* and *C. albicans*. Various activation functions were tested for the hidden and output layers, including linear, logistic, hyperbolic tangent (tanh), exponential, and sinusoidal. Model training was performed for up to 500 epochs, with early stopping applied based on either changes in the error value or a predefined number of epochs without improvement. The performance metric used during hyperparameter optimization was r , the Pearson correlation coefficient, which served as the primary criterion for selecting the best-performing models.

The dataset was partitioned as follows:

- Training set ($N=89$; 70.64%): used to train the ANN, where model learning occurs.
- Validation set ($N=25$; 19.84%): employed to monitor the learning process, prevent overfitting, and implement early stopping, thereby assessing the network's generalization capability.
- Test set ($N=12$; 9.52%): reserved for evaluating the final performance of the ANN. These cases were not involved in training or validation and remained unknown to the network during model development.

All modeling procedures, including PCA, MLP training, and performance evaluation, were performed in the Statistica 13.0 (StatSoft, Tulsa, USA) software environment, which was used for the entirety of the analytical workflow.

Results

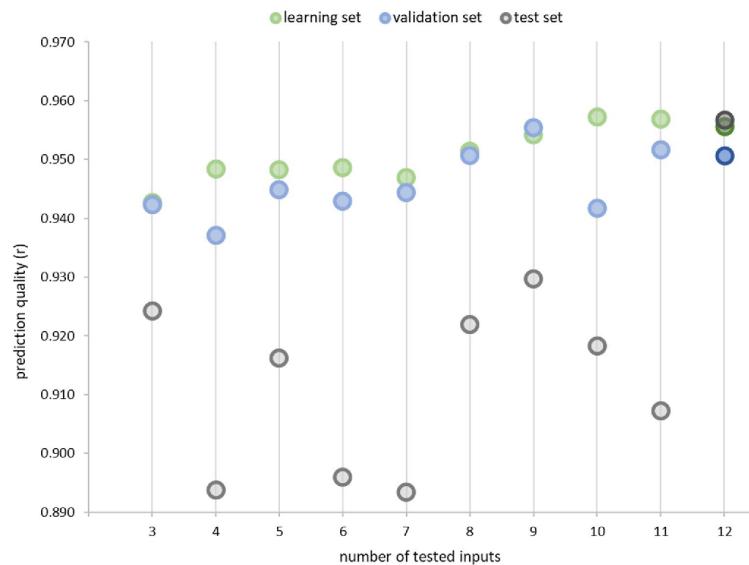
The grid search performed under the conditions described in the previous section identified the ten best-performing multilayer perceptron (MLP) architectures, along with their respective activation functions for the hidden and output layers. The models were evaluated based on prediction quality (r) for the training, validation, and test datasets. The top-performing networks demonstrated high generalization capabilities, with correlation coefficients (r) exceeding 0.93 for the test set. Most of the optimal architectures employed the hyperbolic tangent (tanh) activation function in the hidden layers, frequently paired with either tanh, linear, or sinusoidal functions in the output layer (Table 3).

Table 3 Qualities of prediction (r), learning time (epochs), and activation functions of grid search for 10 ANNs with 3–12 inputs

NN architecture	Training set quality (r)	Validation set quality (r)	Test set quality (r)	Activation (hidden layer)	Activation (output layer)
MLP 12-11-2	0.967	0.961	0.936	tanh	tanh
MLP 11-12-2	0.970	0.960	0.937	tanh	tanh
MLP 10-4-2	0.967	0.957	0.942	tanh	linear
MLP 9-11-2	0.966	0.962	0.951	logistic	sinus
MLP 8-9-2	0.962	0.958	0.937	tanh	tanh
MLP 7-9-2	0.962	0.953	0.956	logistic	tanh
MLP 6-20-2	0.956	0.954	0.952	tanh	sinus
MLP 5-8-2	0.967	0.959	0.962	tanh	tanh
MLP 4-16-2	0.965	0.955	0.954	tanh	tanh
MLP 3-7-2	0.960	0.951	0.961	tanh	linear

Table 4 Average prediction quality (r , SD) and learning time (epochs) of the best-performing ANN architectures obtained from 10-fold cross-validation

Model	Epochs	Average quality (r , SD)		
		Training set	Validation set	Test set
MLP 12-11-2	22-47	0.956; 0.010	0.951; 0.009	0.957; 0.018
MLP 11-12-2	23-63	0.957; 0.018	0.952; 0.015	0.907; 0.129
MLP 10-4-2	21-84	0.957; 0.014	0.942; 0.027	0.918; 0.081
MLP 9-11-2	20-56	0.954; 0.015	0.955; 0.017	0.930; 0.069
MLP 8-9-2	23-100	0.951; 0.018	0.951; 0.018	0.922; 0.070
MLP 7-9-2	21-70	0.947; 0.016	0.944; 0.025	0.893; 0.117
MLP 6-20-2	27-67	0.949; 0.019	0.943; 0.023	0.896; 0.130
MLP 5-8-2	31-101	0.948; 0.020	0.945; 0.022	0.916; 0.090
MLP 4-16-2	28-111	0.948; 0.013	0.937; 0.039	0.894; 0.107
MLP 3-7-2	26-249	0.943; 0.025	0.942; 0.020	0.924; 0.065

Fig. 1 Average prediction quality (r) for the learning, validation, and test sets as a function of the number of principal components (inputs) used in the ANN models

To identify the optimal number of principal components (PCs) that provided the best model performance, each configuration was evaluated using a 10-fold cross-validation procedure. For every network architecture, the mean correlation coefficient (r) and standard deviation (SD) were calculated across all folds for the training, validation, and test sets. The averaged performance metrics, along with the range of training epochs, are summarized in Table 4. Figure 1 illustrates how the prediction quality varied with the number of PCs used as inputs to the ANN model.

The best-performing and balanced model with a 12-11-2 architecture used a tanh activation function for both the hidden and output layers and was selected for further ensemble modeling. The ensemble approach, widely applied in machine learning, is used to enhance model accuracy and predictive performance. It is well established that ensemble models typically outperform the best individual model. Since identifying the optimal architecture required cross-validation, the authors selected the best-performing architecture based on the averaged performance metrics (learning, validation, and test sets across 10 folds, along with standard deviation and variance) and subsequently used the ANNs generated for each fold (Table 5) to construct the final ensemble model.

To determine the error made by the ensemble model, the following metrics were calculated: performance (r), coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and symmetric mean absolute percentage error

Table 5 Performance (r) of the MLP 12-11-2 ANN models across training, validation, and test sets for $\ln(\text{MIC}_{\text{SAU}})$ and $\ln(\text{MIC}_{\text{CAL}})$

Model	Epochs	$\ln(\text{MIC}_{\text{SAU}})$			$\ln(\text{MIC}_{\text{CAL}})$		
		Training (r)	Validation (r)	Test (r)	Training (r)	Validation (r)	Test (r)
1.	22	0.971	0.953	0.953	0.960	0.961	0.939
2.	24	0.971	0.965	0.905	0.958	0.965	0.955
3.	25	0.976	0.951	0.989	0.959	0.920	0.983
4.	29	0.958	0.945	0.959	0.932	0.961	0.971
5.	34	0.964	0.953	0.978	0.961	0.929	0.966
6.	37	0.959	0.953	0.973	0.940	0.952	0.945
7.	37	0.965	0.966	0.920	0.946	0.937	0.960
8.	40	0.957	0.964	0.966	0.937	0.955	0.917
9.	44	0.947	0.947	0.964	0.932	0.956	0.940
10.	47	0.963	0.952	0.987	0.959	0.929	0.965

Table 6 Errors analysis in ensemble ANN for *S. aureus* (SAU) and *C. albicans* (CAL)

	ensemble	
	SAU	CAL
r	0.966	0.954
R^2	0.933	0.910
MAE	0.489	0.477
MSE	0.359	0.413
RMSE	0.599	0.643
sMAPE	15.20%	10.62%

(sMAPE); the results are presented in Table 6. Additionally, parity plots comparing predicted and experimental $\ln(\text{MIC})$ values for both microorganisms were prepared (Fig. 2)

to visually assess prediction accuracy and the agreement between actual and modeled values.

Discussion

Artificial intelligence (AI) and machine learning (ML) are increasingly employed in medical and pharmacological research to enhance efficiency, accelerate compound development, and reduce the time to discovery, while also supporting the principles of green chemistry [22–24]. In the present study, these approaches enabled the prediction of $\ln(\text{MIC})$ values for bis-imidazolium homologs against *Staphylococcus aureus* and *Candida albicans*. Considering

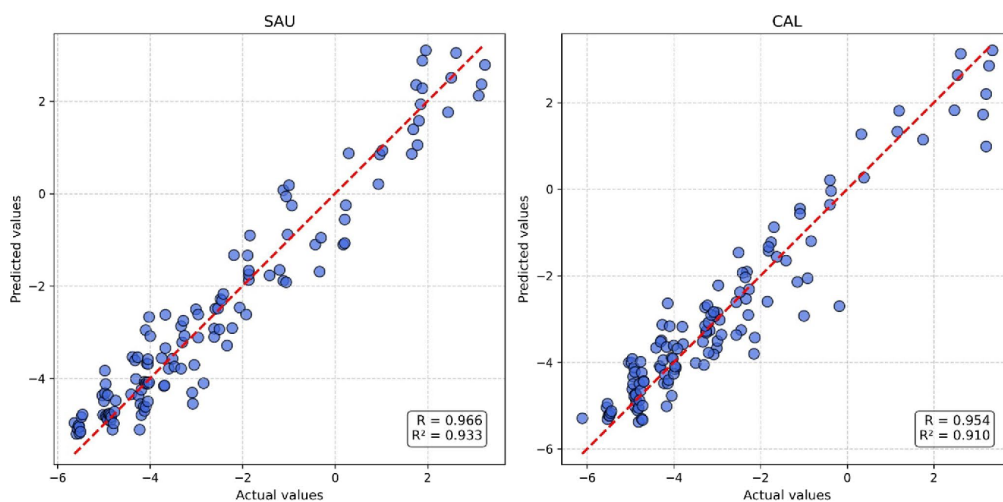


Fig. 2 Parity plots comparing predicted and experimental $\ln(\text{MIC})$ values for *Staphylococcus aureus* (SAU) and *Candida albicans* (CAL) obtained using the optimized ensemble ANN model. The dashed red line represents the ideal 1:1 correlation between predicted and actual

values. The strong linear relationship (SAU: $R=0.966$, $R^2 = 0.933$; CAL: $R=0.954$, $R^2 = 0.910$) indicates high predictive accuracy and model robustness for both microorganisms

that these microorganisms can co-occur in mixed bacterial–fungal infections and may interact synergistically, a tool capable of simultaneously estimating antimicrobial activity against both species was highly desirable. The integration of multiple ML techniques into a unified workflow proved effective, allowing the multi-methodological approach to achieve optimal predictive performance and demonstrating the potential of AI-driven strategies in guiding antimicrobial design.

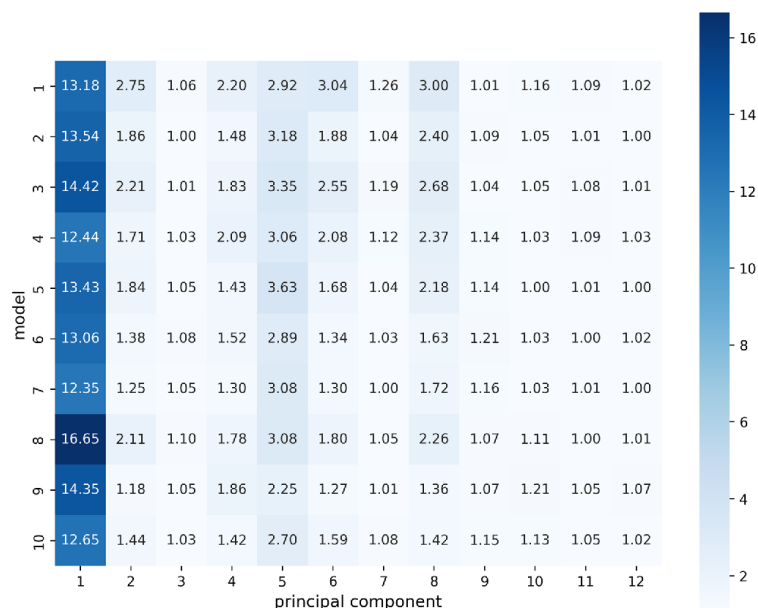
To predict the $\ln(\text{MIC})$ values of the analyzed compounds, molecular descriptors were calculated using Dragon 7.0 software, a widely used approach in QSAR studies [25, 26]. Initial feature selection involved the removal of variables with missing values, low variance, or high correlation. CART analysis was subsequently applied to identify the most relevant descriptors for both microorganism strains. Unlike black-box models, such as deep learning methods, decision trees can manage high-dimensional data while providing interpretable decision rules, making them well suited for understanding relationships between molecular features and antimicrobial activity. Although alternative methods, such as random forests or gradient boosting machines, could have been considered, CART was preferred due to its simplicity and ability to highlight the most significant predictors without requiring ensemble averaging. This approach also minimized computational resource requirements for subsequent ANN analyses and has been successfully applied in related studies [27]. Given that 54 input descriptors were available for only 126 cases, overfitting of the ANN was a potential concern. To address this, Principal Component Analysis (PCA) was employed as a dimensionality reduction technique, retaining the essential variance in the data while limiting the number of input variables [28, 29]. Moreover, PCA was preferred over alternatives such as autoencoders due to its computational efficiency and widespread acceptance in QSAR research [30, 31].

Initial analyses indicated a nonlinear relationship between the selected PCs and the target outputs [Supplementary Information – 5. Correlation Analysis Between Principal Components and Antimicrobial Activity]. Consequently, the MLP architecture was selected as the most suitable modeling approach given the statistical properties of the dataset and the nonlinear nature of the structure–activity relationships. Several of the molecular descriptors were intercorrelated, which limits the applicability of classical regression techniques that rely on variable independence and linearity assumptions. In contrast, MLPs are capable of learning from correlated inputs and capturing complex nonlinear dependencies through internal weight optimization. Moreover, in QSAR modeling, descriptor–activity relationships often exhibit interactive and nonmonotonic behavior that cannot be adequately represented by linear or additive

models. Moreover, in QSAR modeling, descriptor–activity relationships often exhibit interactive and nonmonotonic behavior that cannot be adequately represented by linear or additive models. To enhance robustness and reduce sensitivity to random initialization, an ensemble of independently trained MLPs was employed, with final predictions obtained as the mean of individual network outputs. This strategy improved prediction stability and generalization across both microbial targets.

Twelve PCs, accounting for 91.7% of the variance in the selected descriptors, were evaluated as inputs. Cross-validation analysis using the revised 10-fold procedure indicated that models incorporating a larger number of principal components achieved superior stability and generalization compared with those using fewer inputs. While reducing the number of inputs can help minimize model complexity, excessive dimensionality reduction resulted in a noticeable decline in predictive consistency (Table 4). The configuration employing 12 PCs demonstrated the most balanced performance across the learning, validation, and test sets, confirming it as the optimal input dimensionality for the present dataset (Fig. 1). These results suggest that, in this case, retaining a broader range of information improved the network's ability to capture nonlinear relationships within the data – an advantage that highlights the adaptability of MLP architectures compared with strictly linear methods. The parity plots presented in Fig. 2 demonstrate a strong agreement between the predicted and experimental $\ln(\text{MIC})$ values for both *S. aureus* and *C. albicans*. The data points cluster closely around the ideal 1:1 line, confirming the high predictive accuracy of the optimized ANN ensemble model. The obtained correlation coefficients ($r=0.966$ for *S. aureus* and $r=0.954$ for *C. albicans*) and determination coefficients ($R^2 = 0.933$ and $R^2 = 0.910$, respectively) are consistent with the low MAE and RMSE values reported, indicating excellent model generalization and reliability across both datasets. More complex deep learning architectures, such as convolutional or recurrent neural networks, were not employed due to the limited dataset size.

To further enhance prediction accuracy and stability, ensemble modeling was implemented. Error analysis [Supporting Information – 6. Performance of Individual Artificial Neural Networks (ANNs) Across Folds] indicated that individual network performance varied notably across folds and microorganisms. For *S. aureus*, average results across folds ($R^2 = 0.954$, $\text{RMSE}=0.633$, $\text{sMAPE}=15.8\%$) were slightly less stable than for *C. albicans* ($R^2 = 0.939$, $\text{RMSE}=0.639$, $\text{sMAPE}=11.8\%$), suggesting that the fungal model demonstrated greater internal consistency and generalization ability. However, considerable variation was observed among individual folds, with certain models (e.g., folds 3 and 6) achieving high predictive accuracy for *S.*

Fig. 3 Heatmap of Input Importance Across Principal Components for MLP Neural Networks

aureus ($R^2 > 0.97$) but showing weaker performance for *C. albicans* in the same runs. By contrast, the ensemble ANN (Table 6) effectively integrated predictions from all folds, yielding improved and more balanced performance across both microorganisms. The ensemble achieved $R^2 = 0.933$ and $RMSE = 0.599$ for *S. aureus*, closely matching the best single models, while for *C. albicans* it reached $R^2 = 0.910$ and $RMSE = 0.643$, outperforming the average of individual networks and reducing prediction variance. This confirms that ensemble aggregation enhances model robustness, minimizes sensitivity to fold-specific variability, and stabilizes accuracy across microbial targets.

Analysis of input importance across ANNs indicated that the first PC, which contained the greatest variance information, consistently had the highest influence (Fig. 3). Contributions of subsequent PCs varied among networks, underscoring the value of ensemble approaches, which can leverage different PC contributions across individual models to improve overall predictive performance.

Previous studies predicting antimicrobial activity using QSAR models have often relied on traditional statistical methods, such as multiple linear regression (MLR), partial least squares (PLS) [32, 33]. While these approaches can be effective in certain contexts, they are limited by the assumption of linear relationships between molecular descriptors and antimicrobial activity, which may fail to capture the complex interactions between chemical structure and

biological response. Additionally, conventional QSAR models often exhibit limited generalizability, particularly when applied to structurally diverse compounds [34]. In contrast, MLP-ANNs can model complex, nonlinear dependencies between molecular structure and antimicrobial efficacy, providing enhanced flexibility and predictive capability.

Moreover, unlike most QSAR studies that focus on a single microorganism, the present work successfully predicts $\ln(\text{MIC})$ values for both *S. aureus* and *C. albicans* within a single neural network. This dual-target approach substantially reduces experimental workload by enabling simultaneous predictions for bacterial and fungal pathogens, making it a practical and efficient tool for antimicrobial research. Recent studies exploring ML-based QSAR models, including deep learning approaches for molecular property prediction, often rely on large datasets to achieve accurate predictions [35, 36]. In contrast, the present study demonstrates that robust and meaningful predictions can be obtained even with a relatively small dataset, provided that careful feature selection and dimensionality reduction strategies are employed.

Compared with existing approaches, the hybrid ML pipeline implemented here offers several advantages. By integrating CART for feature selection, PCA for dimensionality reduction, and a single dual-target ensemble ANN for regression, this workflow effectively captures nonlinear relationships between descriptors and biological activity,

resulting in improved predictive accuracy. Furthermore, this approach is computationally efficient and aligns with green chemistry principles, reducing the resources required for antimicrobial screening and facilitating more sustainable drug discovery processes.

Conclusions

In this research, the authors decided to predict the $\ln(\text{MIC})$ values of two microorganism strains via an ANN. Given that the goal was to predict $\ln(\text{MIC})$ values with minimal computational overhead, the ensemble ANN combined with PCA and CART provided a balance between accuracy and efficiency. The ensemble model smooths out inconsistencies, offering strong and stable performance across both targets, which is especially important for *C. albicans*, where individual models tend to have high variance and underperform. The chosen methods align with the principles of green chemistry by optimizing computational resources. More complex ML approaches (e.g., ensemble methods such as XGBoost) often require extensive hyperparameter tuning, which increases the computational cost and training time. By integrating decision trees, PCA, and an MLP-ANN, this study builds upon previous QSAR research while addressing key limitations related to computational efficiency, generalizability, and multitarget prediction.

Limitations and Future Work

While the hybrid pipeline yielded encouraging predictive performance, several limitations should be acknowledged. First, the dataset comprised only 126 bis-imidazolium chlorides; the modest sample size restricts the network's capacity to generalize to chemically diverse scaffolds. Second, the model was validated with internal cross-validation and an ensemble strategy but no fully external dataset was available; independent validation would better quantify generalizability. Third, hyperparameters were tuned within the confines of a limited grid; broader searches or Bayesian optimization could potentially enhance performance. Finally, although we benchmarked a Random Forest regressor, further exploration of alternative learners (support vector regression, gradient boosting) and mechanistic descriptors could refine feature attribution. Future work should therefore focus on enlarging the training corpus, external validation, and systematic benchmarking to strengthen confidence in deployment for virtual screening.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s12247-026-10412-5>.

Acknowledgements The authors gratefully acknowledge the Centre of Informatics—Tricity Academic Supercomputer & Network (CI TASK) for providing the computational resources used for geometry optimization with Gaussian 09. We also acknowledge the use of commercial software accessed through institutional licenses: Gaussian 09, Revision D.02 (Gaussian, Inc., Wallingford, CT, USA) [41] for quantum chemical calculations, Dragon 7.0 (Talet, Milan, Italy) [42], for molecular descriptor generation, and Statistica 13.3 (StatSoft, Tulsa, USA) [43] for CART, PCA, and neural network modeling. This research was supported by the Faculty of Pharmacy at Nicolaus Copernicus University in Toruń.

Author Contributions A.L. designed the project, performed the calculations, and wrote the manuscript. L.S. synthesized the analyzed compounds M.K., L.P. and L.K. provided data to analyze, supported and supervised the chemoinformatics part of the research. A.Badura and A.Buciński. supported and supervised the ML part of the research. All authors have given approval to the final version of the manuscript.

Funding No funding was received to assist with the preparation of this manuscript.

Data Availability The datasets generated and analyzed during this study, as well as the custom scripts used for data preprocessing, are available at https://github.com/ikalore/imidazolium_CART_PCA_A_NN.

Declarations

Competing Interests The authors declare no competing interests.

References

- Aggarwal R, et al. Antibiotic resistance: a global crisis, problems and solutions. *Crit Rev Microbiol*. 2024;50(5):896–921. <https://doi.org/10.1080/1040841X.2024.2313024>.
- Naidu R, et al. Chemical pollution: a growing peril and potential catastrophic risk to humanity. *Environ Int*. 2021. <https://doi.org/10.1016/j.envint.2021.106616>.
- Castiello C, et al. GreenMedChem: the challenge in the next decade toward eco-friendly compounds and processes in drug design. *Green Chem*. 2023;25(6):2109–69. <https://doi.org/10.1039/D2GC03772F>.
- Kar S, Sanderson H, Roy K, Benfenati E, Leszczynski J. Green chemistry in the synthesis of pharmaceuticals. *Chem Rev*. 2022;122(3):3637–710. <https://doi.org/10.1021/ACS.CHEMREV.1C00631>.
- Kamkar M, et al. Artificial intelligence (AI) for sustainable resource management and chemical processes. *ACS Sustain Chem Eng*. 2024;12(8):2924–6. <https://doi.org/10.1021/ACSUSCHEMENG.4C01004>.
- Szymkuć S, et al. Computer-assisted synthetic planning: the end of the beginning. *Angew Chem Int Ed*. 2016;55(20):5904–37. <https://doi.org/10.1002/ANIE.201506101>.
- WHO. WHO bacterial priority pathogens list, 2024: Bacterial pathogens of public health importance to guide research, development and strategies to prevent and control antimicrobial resistance, 2024.
- Parambath S, et al. *Candida albicans*—a systematic review to inform the World Health Organization fungal priority pathogens list. *Med Mycol*. 2024. <https://doi.org/10.1093/MMY/MYAE045>.

9. Bohner F, Papp C, Gácsér A. The effect of antifungal resistance development on the virulence of *Candida* species. *FEMS Yeast Res.* 2022;22(1):1–10. <https://doi.org/10.1093/FEMSYR/FOAC019>.
10. Paul S, et al. A fungal metabolic regulator underlies infectious synergism during *Candida albicans*-*Staphylococcus aureus* intra-abdominal co-infection. *Nat Commun.* Dec. 2024;15(1):1–19. <https://doi.org/10.1038/S41467-024-50058-W>;TECHMETA.
11. Huang H-Y, et al. Clinical insights into mixed *Candida* and bacterial bloodstream infections: a retrospective cohort study. *Microbiol Spectr.* 2025. <https://doi.org/10.1128/SPECTRUM.01684-25>.
12. Wu YM, et al. Enhanced virulence of *Candida albicans* by *Staphylococcus aureus*: evidence in clinical bloodstream infections and infected zebrafish embryos. *J Fungi.* 2021. <https://doi.org/10.3390/JOF7121099>.
13. Hernandez-Cuellar E, et al. Characterization of *Candida albicans* and *Staphylococcus aureus* polymicrobial biofilm on different surfaces. *Rev Iberoam Micol.* 2022;39(2):36–43. <https://doi.org/10.1016/J.RIAM.2022.04.001>.
14. Harriott MM, Noverr MC. *Candida albicans* and *Staphylococcus aureus* form polymicrobial biofilms: effects on antimicrobial resistance. *Antimicrob Agents Chemother.* 2009;53(9):3914. <https://doi.org/10.1128/AAC.00657-09>.
15. Peters BM, et al. *Staphylococcus aureus* adherence to *Candida albicans* hyphae is mediated by the hyphal adhesin Als3p. *Microbiology.* 2012;158(Pt 12):2975. <https://doi.org/10.1099/MIC.0.062109-0>.
16. Paul S, et al. A fungal metabolic regulator underlies infectious synergism during *Candida albicans*-*Staphylococcus aureus* intra-abdominal co-infection. *Nat Commun.* 2024;15(1):1–19. <https://doi.org/10.1038/s41467-024-50058-w>.
17. Hu Y, et al. *Staphylococcus aureus* synergized with *Candida albicans* to increase the pathogenesis and drug resistance in cutaneous abscess and peritonitis murine models. *Pathogens.* Aug. 2021;10:1036. <https://doi.org/10.3390/PATHOGENS10081036/S1>.
18. Schlecht LM, et al. Systemic *Staphylococcus aureus* infection mediated by *Candida albicans* hyphal invasion of mucosal tissue. *Microbiology.* 2015;161(Pt 1):168. <https://doi.org/10.1099/MIC.0.083485-0>.
19. Habibi A, Bayat M, Omidi B, Ezabadi A, Mortazavi P. Synthesis of new imidazole-based ionic liquids with antifungal activity against *Candida albicans*. *Iran J Microbiol.* 2023;15(6):811. <https://doi.org/10.18502/IJM.V15I6.14162>.
20. Hodyna D, Kovalishyn V, Rogalsky S, Blagodatnyi V, Metelytsia L. Imidazolium ionic liquids as potential anti-*Candida* inhibitors: QSAR modeling and experimental studies. *Curr Drug Discov Technol.* 2016;13(2):109–19. <https://doi.org/10.2174/1570163813666160510122201>.
21. Kelley CT. The BFGS Method, in *Iterative Methods for Optimization*, Society for Industrial and Applied Mathematics, 1999, ch. 4, pp. 71–86. doi: <https://doi.org/10.1137/1.9781611970920.ch4>
22. Luo Y, Peng J, Ma J. Next Decade's AI-Based Drug Development Features Tight Integration of Data and Computation. *Health Data Sci.* 2022;2022:9816939. <https://doi.org/10.34133/2022/9816939>.
23. Ali KA, Mohin S, Mondal P, Goswami S, Ghosh S, Choudhuri S. Influence of artificial intelligence in modern pharmaceutical formulation and drug development. *Future J Pharm Sci.* 2024;10(1):1–15. <https://doi.org/10.1186/S43094-024-00625-1>.
24. Niazi SK, Mariam Z. Artificial intelligence in drug development: reshaping the therapeutic landscape. *Ther Adv Drug Saf.* Jan. 2025;16. https://doi.org/10.1177/20420986251321704/ASSET.1290B868C-A1FF-46C5-91CC-B0A0AA8E0AA9/ASSETS/IMA GES/LARGE/10.1177_20420986251321704-FIG3.JPG.
25. Ahmadi S, Ketabi S, Jebeli Javan M. Molecular descriptors in QSPR/QSAR modeling. *Challenges Adv Comput Chem Phys.* 2023;33:25–56. https://doi.org/10.1007/978-3-031-28401-4_2.
26. Karunakaran C, Setlur AS, Niranjana V. Chemoinformatics and molecular descriptors: Exploring the molecular space via chemoinformatics and comprehending the role of molecular description, *Molecular Modeling and Docking Techniques for Drug Discovery and Design*, pp. 71–89, Feb. 2025, <https://doi.org/10.4018/979-8-3693-5598-5.CH003>
27. Lorenc A, Badura A, Karolak M, Palkowski L, Kubik L, Bucinski A. Antimicrobial activity classification of imidazolium derivatives predicted by artificial neural networks. *Pharm Res.* 2024;41(5):891–8. <https://doi.org/10.1007/S11095-024-03699-X/TABLES/3>.
28. Howe TJ, Mahieu G, Marichal P, Tabruyn T, Vugts P. Data reduction and representation in drug discovery. *Drug Discov Today.* 2007;12(1–2):45–53. <https://doi.org/10.1016/J.DRUDIS.2006.10.014>.
29. Yoo S, Kim J, Choi GJ. Drug properties prediction based on deep learning. *Pharmaceutics.* 2022;14(2):467. <https://doi.org/10.3390/PHARMACEUTICS14020467>.
30. Li J, Luo D, Wen T, Liu Q, Mo Z. Representative feature selection of molecular descriptors in QSAR modeling. *J Mol Struct.* 2021;1244:131249. <https://doi.org/10.1016/J.MOLSTRUC.2021.131249>.
31. Das S, Amin SA, Gayen S, Jha T. 2D-QSAR Studies: Regression and Classification-Based QSAR Studies, *Modeling Inhibitors of Matrix Metalloproteinases*, pp. 3–31, Jan. 2023, <https://doi.org/10.1201/9781003303282-2/2D-QSAR-STUDIES-SANJIB-DAS-SK-ABDUL-AMIN-SHOVANLAL-GAYEN-TARUN-JHA> <https://doi.org/10.1201/9781003303282-2/2D-QSAR-STUDIES-SANJIB-DAS-SK-ABDUL-AMIN-SHOVANLAL-GAYEN-TARUN-JHA>
32. Alves CN, Pinheiro JC, Camargo AJ, Ferreira MMC, Romero RAF, Da Silva ABF. A multiple linear regression and partial least squares study of flavonoid compounds with anti-HIV activity. *Journal of Molecular Structure: THEOCHEM.* 2001;541(1–3):81–8. [https://doi.org/10.1016/S0166-1280\(00\)00755-7](https://doi.org/10.1016/S0166-1280(00)00755-7).
33. Yan Q, Wang F, Zhou B, Lin X. Hybrid 2D/3D-quantitative structure-activity relationship studies on the bioactivities and molecular mechanism of antibacterial peptides. *Amino Acids.* 2024;56(1):16. <https://doi.org/10.1007/S00726-024-03381-X>.
34. Vasilev B, Atanasova M. A (comprehensive) review of the application of quantitative structure-activity relationship (QSAR) in the prediction of new compounds with anti-breast cancer activity. *Appl Sci.* 2025;15(3):1206. <https://doi.org/10.3390/APP15031206>.
35. Rusic D, et al. Tackling the antimicrobial resistance ‘pandemic’ with machine learning tools: a summary of available evidence. *Microorganisms.* 2024;12(5):842. <https://doi.org/10.3390/MICROORGANISMS12050842>.
36. Cesaro A, Bagheri M, Torres M, Wan F, de la Fuente-Nunez C. Deep learning tools to accelerate antibiotic discovery. *Expert Opin Drug Discov.* 2023;18(11):1245. <https://doi.org/10.1080/17460441.2023.2250721>.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

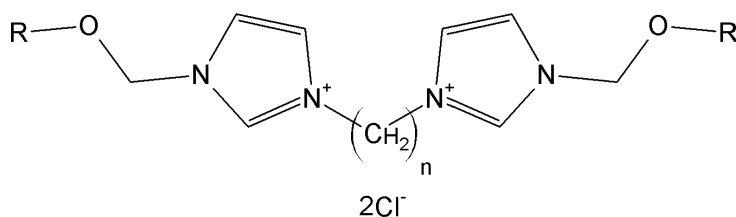
Artykuł nr 2 - materiał dodatkowy

Title: Hybrid Machine Learning Model for Dual Prediction of Antimicrobial Activity Against *Staphylococcus aureus* and *Candida albicans*

Supporting Information

1. The structural formula of the α,ω -bis[(alkyl)imidazol-1-yl-3-alkoxymethyl] chlorides

Figure S1. General structure of analysed α,ω -bis[(alkyl)imidazol-1-yl-3-alkoxymethyl] chlorides



2. Synthesis and MIC determination of bis-imidazolium homologs

Table S1 Coding of the linker chains (n) and substituent chain (R) in the structures of analysed α,ω -bis[(alkyl)imidazol-1-yl-3-alkoxymethyl] chlorides.

code	1	2	3	4	5	6	7	8	9	10	11	12	13	14
n				C ₄ H ₉	C ₅ H ₁₁	C ₆ H ₁₃	C ₇ H ₁₅	C ₈ H ₁₇	C ₉ H ₁₉	C ₁₀ H ₂₁	C ₁₁ H ₂₃	C ₁₂ H ₂₅		
R	CH ₃	C ₂ H ₅	C ₃ H ₇	C ₄ H ₉	C ₅ H ₁₁	C ₆ H ₁₃	C ₇ H ₁₅	C ₈ H ₁₇	C ₉ H ₁₉	C ₁₀ H ₂₁	C ₁₁ H ₂₃	C ₁₂ H ₂₅	C ₁₃ H ₂₇	C ₁₄ H ₂₉

The antimicrobial activity of the synthesized compounds was systematically evaluated against two clinically relevant reference strains: *Staphylococcus aureus* (SAU) ATCC 25213, a Gram-positive bacterium,

and *Candida albicans* (CAL) ATCC 90028, a pathogenic yeast. Both strains were obtained from Microbiologics (San Diego, CA, USA). To preserve viability and genetic stability prior to experimentation, all microbial cultures were cryopreserved at -80°C in cryovial vials (Simport, Canada). Before antimicrobial testing, cultures were revived and subcultured in Müller–Hinton broth to ensure robust and consistent microbial growth. To assess the antimicrobial efficacy of the synthesized compounds, the minimum inhibitory concentration (MIC) was determined for both analysed strains. MIC values were determined by preparing a series of agar plates with decreasing concentrations of the test compounds, ranging from 1.0% to 0.000125% (w/v).

To assess the antimicrobial potential of the compounds, minimum inhibitory concentration (MIC) values were determined using a standardized protocol. Microbial suspensions were prepared by adjusting the turbidity to a 0.5 McFarland standard, which corresponds to approximately 1.5×10^8 CFU/mL for bacteria and 1×10^6 CFU/mL for yeasts. This standardization was achieved by harvesting microbial colonies from blood trypticase soy agar (TSA, BBL) plates and suspending them in 2 mL of sterile saline to ensure uniform cell density across replicates.

The MIC determination was carried out using the agar dilution method, which allows for precise assessment of antimicrobial activity across a defined concentration gradient. Mueller–Hinton agar (bioMérieux) was used as the base medium. A series of agar plates were prepared with serial two-fold dilutions of each test compound, covering a concentration range from 1.0% to 0.000125% (w/v). These concentration values reflect standardized stepwise dilutions, consistent with Clinical and Laboratory Standards Institute (CLSI) guidelines for antimicrobial susceptibility testing. Standardized inocula of SAU and CAL were applied to the surfaces of the prepared agar plates using aseptic techniques. Plates inoculated with *S. aureus* were incubated at 37°C for 24 hours, while those inoculated with *C. albicans* were incubated for 48 hours at the same temperature to accommodate the slower growth kinetics of the fungal strain.

The MIC was defined as the lowest concentration of a compound at which no visible microbial growth occurred, with the exception of a faint haze or the presence of no more than a single colony. This definition aligns with accepted CLSI criteria and ensures a conservative and reproducible threshold for microbial inhibition. To validate the accuracy of the MIC determinations, two types of control plates were included in

each experimental batch. Growth control plates containing only the microbial inoculum without any test compounds were used to confirm the viability of the organisms and the suitability of the growth medium. In parallel, sterile control plates containing only the medium without microbial inoculation were included to verify the absence of contamination throughout the procedure. The determined MIC values of all 126 homologs for both bacterial strains are presented in **Table S2** in Supplementary Material.

All antimicrobial susceptibility assays were performed in full compliance with CLSI guidelines to guarantee methodological reliability, reproducibility, and data integrity. Additionally, MIC determinations were conducted with biological replicates, and observed variability between replicates did not exceed one dilution step.

Table S2 List of 126 α,ω -bis[(alkyl)imidazol-1-yl-3-alkoxymethyl] chlorides homologs with minimal inhibitory concentration for *Staphylococcus aureus* (MIC_SAU) and *Candida albicans* (MIC_CAL)

object					object					object					object				
	n	R	MIC_SAU	MIC_CAL		n	R	MIC_SAU	MIC_CAL		n	R	MIC_SAU	MIC_CAL		n	R	MIC_SAU	MIC_CAL
1	4	1	7.117	28.468	33	6	5	2.543	0.163	65	8	9	0.008	0.008	97	10	13	0.025	0.050
2	4	2	6.591	26.363	34	6	6	1.193	0.077	66	8	10	0.008	0.008	98	10	14	0.023	0.047
3	4	3	12.274	24.547	35	6	7	0.146	0.037	67	8	11	0.015	0.007	99	11	1	0.089	0.022
4	4	4	5.741	22.966	36	6	8	0.017	0.009	68	8	12	0.014	0.014	100	11	2	0.084	0.042
5	4	5	1.338	0.669	37	6	9	0.017	0.008	69	8	13	0.052	0.104	101	11	3	0.079	0.020
6	4	6	1.261	0.081	38	6	10	0.016	0.016	70	8	14	0.097	0.097	102	11	4	0.037	0.019
7	4	7	0.154	0.038	39	6	11	0.015	0.030	71	9	1	0.736	0.095	103	11	5	0.018	0.009
8	4	8	0.037	0.018	40	6	12	0.058	0.116	72	9	2	0.356	0.089	104	11	6	0.004	0.002
9	4	9	0.017	0.017	41	6	13	0.108	0.833	73	9	3	0.649	0.084	105	11	7	0.004	0.004

10	4	10	0.017	0.017	42	6	14	0.012	0.050	74	9	4	0.158	0.040	106	11	8	0.004	0.004
11	4	11	0.127	0.016	43	7	1	6.356	3.178	75	9	5	0.075	0.019	107	11	9	0.004	0.007
12	4	12	0.242	0.242	44	7	2	5.933	1.471	76	9	6	0.036	0.009	108	11	10	0.004	0.007
13	4	13	0.112	0.433	45	7	3	2.781	0.690	77	9	7	0.008	0.004	109	11	11	0.007	0.007
14	4	14	0.052	0.402	46	7	4	2.618	0.335	78	9	8	0.008	0.008	110	11	12	0.007	0.013
15	5	1	13.687	13.687	47	7	5	1.226	0.316	79	9	9	0.008	0.008	111	11	13	0.025	0.049
16	5	2	25.423	12.711	48	7	6	0.300	0.037	80	9	10	0.015	0.007	112	11	14	0.046	0.046
17	5	3	23.730	11.865	49	7	7	0.036	0.018	81	9	11	0.007	0.007	113	12	1	0.086	0.086
18	5	4	22.249	1.379	50	7	8	0.017	0.008	82	9	12	0.007	0.014	114	12	2	0.081	0.041
19	5	5	5.235	0.335	51	7	9	0.016	0.008	83	9	13	0.025	0.051	115	12	3	0.038	0.019
20	5	6	1.226	0.158	52	7	10	0.008	0.015	84	9	14	0.047	0.368	116	12	4	0.018	0.009
21	5	7	0.150	0.037	53	7	11	0.007	0.015	85	10	1	0.712	0.046	117	12	5	0.009	0.004
22	5	8	0.018	0.036	54	7	12	0.014	0.014	86	10	2	0.345	0.043	118	12	6	0.004	0.004
23	5	9	0.017	0.008	55	7	13	0.013	0.053	87	10	3	0.326	0.041	119	12	7	0.004	0.004
24	5	10	0.016	0.016	56	7	14	0.049	0.098	88	10	4	0.154	0.038	120	12	8	0.004	0.008
25	5	11	0.031	0.015	57	8	1	0.393	0.196	89	10	5	0.073	0.009	121	12	9	0.007	0.015
26	5	12	0.030	0.119	58	8	2	0.367	0.184	90	10	6	0.017	0.009	122	12	10	0.007	0.007
27	5	13	0.027	0.055	59	8	3	0.345	0.173	91	10	7	0.008	0.004	123	12	11	0.007	0.013
28	5	14	0.025	0.102	60	8	4	0.326	0.163	92	10	8	0.004	0.004	124	12	12	0.006	0.006
29	6	1	6.591	3.295	61	8	5	0.154	0.077	93	10	9	0.015	0.008	125	12	13	0.012	0.012
30	6	2	6.137	24.547	62	8	6	0.073	0.018	94	10	10	0.007	0.007	126	12	14	0.045	0.023
31	6	3	11.483	5.741	63	8	7	0.017	0.017	95	10	11	0.007	0.007					
32	6	4	5.394	0.669	64	8	8	0.017	0.008	96	10	12	0.013	0.013					

3. Molecular descriptors calculations

The analyzed compounds were modeled as 2D and 3D structures in their neutral forms, and their optimization was carried out using the quantum mechanical Density Functional Theory (DFT) with the B3LYP method and Pople's 6-31G basis set. The optimization process was conducted within the implicit solvent model SCRF (Self-Consistent Reaction Field), with the dielectric constant set to that of water, utilizing the PCM (Polarizable Continuum Model). Calculations were performed using Gaussian 09, Revision D.02 (Gaussian, Inc., Wallingford, CT, USA) [1] on a supercomputer cluster with nodes equipped with 12-core Intel® Xeon® E5 v3 2.3 GHz processors. Geometry optimization computations were carried out at the Centre of Informatics—Tricity Academic Supercomputer & Network. In subsequent analysis, 5270 molecular descriptors from 29 logical blocks were identified using Dragon 7.0 software (Talet, Milan, Italy) [2].

Table S3 Overview of selected molecular descriptors used in the PCA model development.

Name	Description	Molecular block
AVS_RG	average vertex sum from reciprocal squared geometrical matrix	3D matrix-based descriptors
CATS2D_03_AL	CATS2D Acceptor-Lipophilic at lag 03	CATS 2D
CATS2D_04_AL	CATS2D Acceptor-Lipophilic at lag 04	CATS 2D
CATS2D_05_AL	CATS2D Acceptor-Lipophilic at lag 05	CATS 2D
CATS2D_07_AL	CATS2D Acceptor-Lipophilic at lag 07	CATS 2D
CATS2D_09_AL	CATS2D Acceptor-Lipophilic at lag 09	CATS 2D
CATS3D_06_AA	CATS3D Acceptor-Acceptor BIN 06	CATS 3D
CATS3D_06_AL	CATS3D Acceptor-Lipophilic BIN 06	CATS 3D
CATS3D_07_AL	CATS3D Acceptor-Lipophilic BIN 07	CATS 3D
CATS3D_08_AL	CATS3D Acceptor-Lipophilic BIN 08	CATS 3D
CATS3D_09_AL	CATS3D Acceptor-Lipophilic BIN 09	CATS 3D
CATS3D_10_AL	CATS3D Acceptor-Lipophilic BIN 10	CATS 3D
DLS_02	modified drug-like score from Oprea et al. (6 rules)	Drug-like indices
EE_Dt	Estrada-like index (log function) from detour matrix	2D matrix-based descriptors
EE_Dz(v)	Estrada-like index (log function) from Barysz matrix weighted by van der Waals volume	2D matrix-based descriptors
Eig13_EA(dm)	eigenvalue n. 13 from edge adjacency mat. weighted by dipole moment	Edge adjacency indices
Eig15_EA(dm)	eigenvalue n. 15 from edge adjacency mat. weighted by dipole moment	Edge adjacency indices
F07[C-O]	Frequency of C - O at topological distance 7	2D Atom Pairs
GATS4e	Geary autocorrelation of lag 4 weighted by Sanderson electronegativity	2D autocorrelations
GATS4m	Geary autocorrelation of lag 4 weighted by mass	2D autocorrelations

GATS5e	Geary autocorrelation of lag 5 weighted by Sanderson electronegativity	2D autocorrelations
GATS5m	Geary autocorrelation of lag 5 weighted by mass	2D autocorrelations
HATS7u	leverage-weighted autocorrelation of lag 7 / unweighted	GETAWAY descriptors
MATS1s	Moran autocorrelation of lag 1 weighted by I-state	2D autocorrelations
MATS3e	Moran autocorrelation of lag 3 weighted by Sanderson electronegativity	2D autocorrelations
MATS3s	Moran autocorrelation of lag 3 weighted by I-state	2D autocorrelations
MATS4e	Moran autocorrelation of lag 4 weighted by Sanderson electronegativity	2D autocorrelations
MATS4m	Moran autocorrelation of lag 4 weighted by mass	2D autocorrelations
MATS4v	Moran autocorrelation of lag 4 weighted by van der Waals volume	2D autocorrelations
MATS6v	Moran autocorrelation of lag 6 weighted by van der Waals volume	2D autocorrelations
MATS8i	Moran autocorrelation of lag 8 weighted by ionization potential	2D autocorrelations
MAXDN	maximal electrotopological negative variation	Topological indices
Mor06m	signal 06 / weighted by mass	3D-MoRSE descriptors
Mor07v	signal 07 / weighted by van der Waals volume	3D-MoRSE descriptors
Mor08m	signal 08 / weighted by mass	3D-MoRSE descriptors
Mor08u	signal 08 / unweighted	3D-MoRSE descriptors
Mor24m	signal 24 / weighted by mass	3D-MoRSE descriptors
R7u+	R maximal autocorrelation of lag 7 / unweighted	GETAWAY descriptors
RDF105u	Radial Distribution Function - 105 / unweighted	RDF descriptors
RDF110u	Radial Distribution Function - 110 / unweighted	RDF descriptors
RDF150m	Radial Distribution Function - 150 / weighted by mass	RDF descriptors
RGyr	radius of gyration (mass weighted)	Geometrical descriptors
Ro5	Complementary Lipinski Alert index	Drug-like indices
SM11_AEA(dm)	spectral moment of order 11 from augmented edge adjacency mat. weighted by dipole moment	Edge adjacency indices
SpMax_EA(ed)	leading eigenvalue from edge adjacency mat. weighted by edge degree	Edge adjacency indices
SpMax3_Bh(s)	largest eigenvalue n. 3 of Burden matrix weighted by I-state	Burden eigenvalues
SpMin1_Bh(m)	smallest eigenvalue n. 1 of Burden matrix weighted by mass	Burden eigenvalues
SpPosA_B(p)	normalized spectral positive sum from Burden matrix weighted by polarizability	2D matrix-based descriptors
SpPosA_H2	normalized spectral positive sum from reciprocal squared distance matrix	2D matrix-based descriptors
VE1_A	coefficient sum of the last eigenvector from adjacency matrix	2D matrix-based descriptors
VE1_B(s)	coefficient sum of the last eigenvector from Burden matrix weighted by I-State	2D matrix-based descriptors
VE1_B(v)	coefficient sum of the last eigenvector from Burden matrix weighted by van der Waals volume	2D matrix-based descriptors
VE1_D/Dt	coefficient sum of the last eigenvector from distance/detour matrix	2D matrix-based descriptors
VE2_RG	average coefficient of the last eigenvector from reciprocal squared geometrical matrix	3D matrix-based descriptors

4. Preprocessing script

The script performs data preprocessing and feature selection on a microbiological dataset related to MIC (Minimum Inhibitory Concentration) values for *Staphylococcus aureus* (SAU) and *Candida albicans* (CAL). The main steps are:

- Data Loading: The script imports a CSV file (Lorenc_SAU_CAL_dataset_v1.csv) using a semicolon delimiter.
- Subset Selection: It selects only the columns from "MW" onward for further analysis.
- Data Cleaning: All occurrences of 'na' are converted to NaN, and non-numeric entries are coerced into missing values, which are then dropped column-wise.
- Low Variance Filtering: Features with variance below 0.05 are removed using VarianceThreshold.
- Correlation Filtering: Highly correlated features (correlation > 0.95) are eliminated to reduce multicollinearity.
- Final Dataset Assembly: The filtered features are combined with the MIC values (MIC_SAU, MIC_CAL) and sample identifier (No.), resulting in a cleaned and reduced dataset imidazole_selected.

```
import numpy as np
import pandas as pd
from sklearn.feature_selection import VarianceThreshold
from sklearn.preprocessing import StandardScaler

file_path = 'Lorenc_SAU_CAL_dataset_v1.csv'
data = pd.read_csv(file_path, delimiter=';')

# Start operations from the column named "MW"
start_column = 'MW'
start_index = data.columns.get_loc(start_column)
data_subset = data.iloc[:, start_index:]

# Replace 'na' with NaN and convert to numeric
data_subset = data_subset.replace('na', np.nan)
data_subset = data_subset.apply(pd.to_numeric, errors='coerce')

data_subset = data_subset.dropna(axis=1)

# Remove low variance features
selector = VarianceThreshold(threshold=0.05)
data_var = selector.fit_transform(data_subset)
```

```

selected_features = data_subset.columns[selector.get_support(indices=True)]
data_var_df = pd.DataFrame(data_var, columns=selected_features)

# Remove highly correlated features
corr_matrix = data_var_df.corr().abs()
upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(bool))

to_drop = [column for column in upper.columns if any(upper[column] > 0.95)]
data_corr = data_var_df.drop(columns=to_drop)

# Update selected features
selected_features = pd.DataFrame(data_corr.columns)
mic = pd.DataFrame(data, columns=['No.', 'MIC_SAU', 'MIC_CAL'])

imidazole_selected = mic.join(data_corr.copy())

```

Table S4. Predictive performance of MLP models for *S. aureus* (SAU) and *C. albicans* (CAL) MIC prediction across eight input feature sets. Performance metrics include Pearson correlation coefficient (r), coefficient of determination (R²), mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and symmetric mean absolute percentage error (sMAPE), reported separately for the training, validation, and test sets.

	1		2		3		4		5		6		7		8	
	SAU	CAL	SAU	CAL	SAU	CAL	SAU	CAL	SAU	CAL	SAU	CAL	SAU	CAL	SAU	CAL
training set																
r	0.954	0.934	0.968	0.951	0.975	0.973	0.946	0.918	0.962	0.973	0.969	0.967	0.964	0.935	0.965	0.948
R²	0.910	0.872	0.937	0.904	0.951	0.947	0.895	0.843	0.925	0.947	0.939	0.935	0.929	0.874	0.931	0.899
MAE	0.572	0.590	0.465	0.486	0.412	0.408	0.618	0.577	0.506	0.377	0.478	0.433	0.503	0.535	0.513	0.550
MSE	0.512	0.625	0.334	0.415	0.259	0.254	0.553	0.656	0.412	0.235	0.353	0.323	0.407	0.540	0.418	0.560
RMSE	0.715	0.790	0.578	0.644	0.509	0.504	0.744	0.810	0.642	0.484	0.594	0.569	0.638	0.735	0.647	0.748
sMAPE	15.70%	14.97%	15.27%	10.97%	12.87%	9.87%	18.63%	12.96%	13.66%	7.90%	15.24%	10.61%	14.94%	12.28%	14.68%	12.72%
validation set																
r	0.966	0.942	0.960	0.947	0.939	0.906	0.974	0.970	0.957	0.897	0.931	0.942	0.955	0.939	0.929	0.905
R²	0.933	0.887	0.922	0.897	0.882	0.821	0.949	0.941	0.916	0.805	0.867	0.887	0.912	0.882	0.863	0.819
MAE	0.556	0.699	0.533	0.656	0.654	0.705	0.507	0.525	0.647	0.799	0.710	0.624	0.639	0.775	0.723	0.558
MSE	0.423	0.696	0.427	0.721	0.762	0.837	0.392	0.499	0.653	1.341	0.834	0.580	0.572	0.949	0.687	0.509
RMSE	0.650	0.834	0.654	0.849	0.873	0.915	0.626	0.706	0.808	1.158	0.913	0.762	0.756	0.974	0.829	0.713
sMAPE	16.80%	13.85%	10.85%	16.55%	18.84%	14.44%	11.30%	9.36%	12.65%	15.73%	16.82%	14.96%	15.69%	19.19%	19.39%	12.42%
test set																
r	0.959	0.940	0.966	0.936	0.945	0.837	0.959	0.951	0.932	0.902	0.929	0.859	0.890	0.824	0.943	0.902
R²	0.920	0.884	0.933	0.876	0.893	0.701	0.920	0.904	0.869	0.814	0.863	0.738	0.792	0.679	0.889	0.814
MAE	0.584	0.716	0.606	0.646	0.709	0.932	0.630	0.581	0.787	0.820	0.618	0.889	0.906	0.875	0.592	0.796
MSE	0.435	0.609	0.600	0.737	0.896	2.125	0.624	0.591	0.860	0.850	0.593	1.367	1.208	1.509	0.482	0.965
RMSE	0.660	0.781	0.774	0.858	0.947	1.458	0.790	0.769	0.927	0.922	0.770	1.169	1.099	1.228	0.694	0.982
sMAPE	13.28%	10.86%	18.55%	15.12%	11.88%	16.57%	22.31%	14.15%	25.35%	18.33%	11.63%	17.92%	23.83%	12.69%	12.33%	12.59%

References:

- [1] M. J. Frish *et al.*, “Gaussian 09, Revision D.02,” 2013, *Gaussian, Inc, Wallingford CT*. Accessed: May 19, 2025. [Online]. Available: <https://gaussian.com/>
- [2] Kode srl, “Dragon (software for molecular descriptor calculation) version 7.0.6,” 2016. Accessed: May 19, 2025. [Online]. Available: <https://www.taletе.mi.it/>



Original Article

Exploring the efficacy of artificial neural networks in predicting lung cancer recurrence: a retrospective study based on patient records

Andżelika Lorenc^{1,8^}, Anna Romaszko-Wojtowicz^{2,3,8^}, Łukasz Jaśkiewicz^{4^}, Anna Doboszyńska^{2,3^}, Adam Buciński^{1^}

¹Department of Biopharmacy, Ludwik Rydygier Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, Bydgoszcz, Poland;

²Department of Pulmonology, School of Public Health, Collegium Medicum, University of Warmia and Mazury in Olsztyn, Olsztyn, Poland;

³The Center for Pulmonary Diseases, Olsztyn, Poland; ⁴Department of Human Physiology and Pathophysiology, School of Medicine, Collegium Medicum, University of Warmia and Mazury in Olsztyn, Olsztyn, Poland

Contributions: (I) Conception and design: A Lorenc, A Romaszko-Wojtowicz, A Buciński; (II) Administrative support: A Doboszyńska, A Buciński; (III) Provision of study materials or patients: A Romaszko-Wojtowicz, Ł Jaśkiewicz; (IV) Collection and assembly of data: All authors; (V) Data analysis and interpretation: A Lorenc, A Romaszko-Wojtowicz, A Buciński; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

⁸These authors contributed equally to this work.

Correspondence to: Andżelika Lorenc, MSc. Department of Biopharmacy, Ludwik Rydygier Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University in Toruń, Jurasza 2, 85-089 Bydgoszcz, Poland. Email: andzelika.lorenc@cm.umk.pl.

Background: Lung cancer remains a significant public health concern, accounting for a considerable number of cancer-related deaths worldwide. Neural networks have emerged as a promising tool that can aid in the diagnosis and treatment of various cancers. Consequently, there has been a growing interest in exploring the potential of artificial intelligence (AI) methods in medicine. The present study aimed to evaluate the effectiveness of a neural network in predicting lung cancer recurrence.

Methods: The study employed retrospective data from 2,296 medical records of patients diagnosed with lung cancer and admitted to the Warmińsko-Mazurskie Center for Lung Diseases in Olsztyn, Poland. The statistical software STATISTICA 7.1, equipped with the Neural Networks module (StatSoft Inc., Tulsa, USA), was utilized to analyze the data. The neural network model was trained using patient information regarding gender, treatment, smoking status, family history, and symptoms of cancer.

Results: The study employed a multilayer perceptron neural network with a two-phase learning process. The network demonstrated high predictive ability, as indicated by the percentage of correct classifications, which amounted to 87.5%, 89.1%, and 89.9% for the training, validation, and test sets, respectively.

Conclusions: The findings of this study support the potential usefulness of a neural network-based predictive model in assessing the risk of lung cancer recurrence. Further research is warranted to validate these findings and to explore AI's broader implications in cancer diagnosis and treatment.

Keywords: Artificial neural networks (ANNs); multilayer perceptron (MLP); lung cancer; recurrence prediction

Submitted May 31, 2023. Accepted for publication Sep 26, 2023. Published online Oct 26, 2023.

doi: 10.21037/tlcr-23-350

View this article at: <https://dx.doi.org/10.21037/tlcr-23-350>

[^] ORCID: Andżelika Lorenc, 0000-0002-1474-7864; Anna Romaszko-Wojtowicz, 0000-0003-2042-1382; Łukasz Jaśkiewicz, 0000-0002-0035-4314; Anna Doboszyńska, 0000-0001-9817-5287; Adam Buciński, 0000-0002-0558-9139.

Introduction

Lung cancer is the most frequent cause of death due to neoplastic diseases worldwide. A rapid growth in pulmonary cancer morbidity occurred in the early 20th century, most probably due to the increasing popularity of tobacco smoking (1,2).

Over the past few decades, there has been considerable progress in the diagnostics and treatment of lung cancer in the world. A breakthrough study, the National Lung Screening Trial (years 2002–2010), where a low-dose computer tomography (LDCT) of the chest was employed for early detection of lung cancer, demonstrated that an overall mortality rate due to lung cancer can be decreased by 6.7% owing to the acceleration of diagnosis (3). In 2018, this screening program was launched in Poland. Lung cancer diagnostics in Poland is carried out in pulmonology and thoracic surgery wards. Unfortunately, diagnostic procedures are unnecessarily prolonged because of the lack of optimal conditions [e.g., no access to endobronchial ultrasound (EBUS) bronchoscopy], the need to repeat biopsies, re-hospitalization, or difficulties arising from the way lung cancer diagnostics is financed from the National Health Fund in Poland (NFZ) (4). Another problem is the insufficient availability of advanced molecular diagnostics, which opens the door to a more modern treatment, that is targeted therapy.

Despite all advances in the diagnostics and treatment of lung cancer, the prognosis for patients remains poor. Five-year survival is 19% in the world, being slightly lower, 14.4%, in Poland (5). The main reason is the degree of

advancement of the illness at the moment of its diagnosis. This prompts scientists to search for alternative methods that allow for a faster diagnosis and timely initiation of treatment.

Artificial intelligence (AI) techniques have been rapidly expanding in various scientific fields, including medical science. Referring to the Dimensions research database over the last 5 years, a significant number of scientific papers—60,102 and 3,006 patents to be exact have been published on oncology and AI with a noticeable increase in the number of publications over the years (6). The first records of artificial neural network (ANN) in the PubMed database date back to 1986 (7,8). As early as 2004, networks were used to assess the topographical classification of electroencephalographic (EEG) patterns in people affected by Huntington's disease (HD) (9). Nevertheless, neural networks are also used in the broadly understood assessment of cancer, e.g., their recurrence, assessment of response to treatment, patient survival—prognosis (10–13). The prognosis can be assessed on the basis of some clinical data and not only the morphology of the tumor. A group of German scientists has shown that the condition of sentinel lymph nodes could be predicted with some probability based on digital melanoma preparations and data such as ulceration, tumor thickness, and patient's age (14). Progress in oncology fields has raised hopes among cancer patients for improved diagnosis and treatment, leading to a better quality of life and longer survival.

One particularly promising method is the use of ANNs, which have a high potential for medical applications due to their ability to process complex information. ANNs are a biologically inspired concept of machine learning based on structures formed by neurons in the brain. Their capability to learn and generalize knowledge made them considered useful and promising tools for solving intricate problems.

An example of an often-used ANN is a multilayer perceptron (MLP) which is a nonlinear data proceeding method characterized by a well-defined layered structure (15). MLP is built of at least three layers of artificial neurons—input and output layer and localized between them—one or more hidden layers. The outputs of one layer are connected to the input of every neuron in the next layer, creating a system of parallel data processing. The number of neurons in the input layer corresponds to the number of given input variables. The neurons of this layer are not processing the data but they transmit the signals to the neurons in the next layer—the hidden layer. The number of hidden layers in network architecture depends

Highlight box

Key findings

- Artificial neural network (ANN) can successfully be used to predict lung cancer recurrence based on retrospective data from patients.

What is known and what is new?

- ANNs are widely studied supporting tools in cancer diagnosis and treatment obtaining high predictive quality.
- A new approach to forecasting the risk of lung cancer recurrence is utilizing ANNs with basic clinical data as predictors.

What is the implication, and what should change now?

- By utilizing ANN, predicting the likelihood of lung cancer recurrence among patients can lead to improved care during remission and early detection of recurrence. This allows medical professionals to focus on patients who may be at higher risk, ultimately enhancing overall patient care.

on the problem's complexity. The neurons of the hidden layer process the signals and transmit them to the output layer neurons, which are furtherly processed and gathered to give the final solution (16,17).

The optimal choice of network architecture and learning algorithm is crucial to the model's predictive ability. The too-simple network could not identify the complex connections between the input and output data, which will effect on an inability to solve the given problem. On the other hand, architecture too complicated can lead to fast overlearning, which means that the network will precisely predict the outcome in the learning set, but the quality of prediction in validation and test sets will be poor. That phenomenon is called overfitting (18).

To evaluate the ANN performance the data is divided into three sets—learning set, validation set, and test set. The role of the learning set is to find connections between, at first sight, not related data. This set has the ability to prior see the given inputs and outputs of presented cases, which enables the learning algorithm to make necessary corrections of strength between the neurons (weights) in each epoch. This correction process takes place continuously until the error made by the network between real outputs and the predicted outputs is the smallest. The validation set role is to control the learning process and stop it at the moment when the error is the smallest to avoid overfitting. The test set is used for evaluation—the cases never presented before to the model are introduced to the designed ANN to check its ability to generalize knowledge and correctly predict the output.

The objective of this study was to evaluate the effectiveness of ANNs in lung oncology. The authors created an MLP neural network model using retrospective data, specifically focusing on lung cancer recurrence forecasting. They carefully developed, modeled, and validated their approach to assessing its usefulness in the field. We present this article in accordance with the TRIPOD reporting checklist (available at <https://tldr.amegroups.com/article/view/10.21037/tlcr-23-350/rc>).

Methods

Selection of the group and data collection

The retrospective study was conducted based on data from the medical histories of patients hospitalized in the Warmińsko-Mazurskie Center for Pulmonology Diseases in Olsztyn, Poland. The center acts as a reference hospital for

patients from northeastern Poland. The center comprises Pulmonological Wards and a Thoracic Surgery Ward, which provide patients with holistic care. The study included data gathered from the years 2012 to 2017 and extracted from medical histories of 2,296 patients with lung cancer diagnosed by histopathological tests.

The inclusion criterium was a histopathologically confirmed diagnosis of lung cancer. The exclusion criteria included: a lack of histopathological confirmation for various reasons (e.g., premature death of the patient) or a diagnosis of disseminated neoplastic disease with a point of origin outside the lung. Recurrence was defined as a local, regional or distant recurrence based on the North American Association of Central Cancer Registries (NAACCR), American College of Surgeon (AcoS)-Commission on Cancer and surveillance, epidemiology, and end results (SEER)/National Cancer Institute (NCI) in accordance with the Standards for Oncology Registry Entry (STORE) guidelines (19).

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Prior to conducting the study, approval was obtained from the management of the medical center and the Bioethics Committee of the Nicolaus Copernicus University in Toruń, Collegium Medicum in Bydgoszcz (No. KB 355/2020), on June 23rd, 2020. Patient survival data were obtained from the Main Statistical Office. Since the study is based on a retrospective analysis of patient medical records the individual consent for this retrospective analysis was waived.

Group characteristics

The proceeding above resulted in receiving the complete medical records of patients with lung cancer on which the study was conducted. Sixty-eight point five five percent of the patients (n=1,574) were men and 31.45% (n=722) were women. The average age of lung cancer diagnosis was 64.93 years [standard deviation (SD) =8.42], and the average age of death (n=2,072)—65.85 years (SD =8.41). Among the patients included in the study, patients that had never smoked were 5.44% (n=125), patients who quit 29.49% (n=677), and smokers—65.07% (n=1,494). The average packyears in the whole group (n=2,296) were established at 37.45 (SD =20.06), and the average body mass index (BMI) was 24.78 (n=2,296; SD =5.07). In the whole group, the symptoms occurred in 88.20% (n=2,025) cases.

The characteristics of the studied group with

Table 1 The descriptive statistics of studied group

Variables	Group	N	Me; Min; Max	SD
Age of diagnosis (years)		2,296	64.93; 29.99; 92.02	8.42
	M	1,574	65.32; 32.33; 92.02	8.35
	F	722	64.10; 29.99; 89.92	8.51
Age of death (years)		2,072	65.85; 32.57; 92.07	8.41
	M	1,428	66.28; 37.05; 92.07	8.38
	F	644	64.89; 32.57; 90.31	8.39
BMI (kg/m ²)		2,296	24.78; 13.10; 46.60	5.07
	M	1,574	24.62; 13.10; 46.60	4.82
	F	722	25.11; 13.38; 46.43	5.55
Packyears		2,294	37.45; 0.00; 200.00	20.06
	M	1,574	39.47; 0.00; 200.00	20.20
	F	720	33.03; 0.00; 120.00	19.00

Me, mean; Min, minimum; Max, maximum; SD, standard deviation; M, male; F, female; BMI, body mass index.

consideration of the type of therapy, smoking habits, family history, symptoms, and occurred recurrence are shown in *Table 1* and *Table 2*.

ANN analysis

In the initial stage of the study, the data describing 2,296 patients were randomly divided by the software into three sets: learning set, validation set, and test set, containing 1,148, 574, and 574 cases, respectively.

The preliminary models of MLP neural networks were built to find the optimal architecture of the network and the optimal set of input variables that allows predicting recurrence with the highest precision. In order for AI to be effectively applied in the medical field, it is essential for clinicians to be able to interpret the data and outcomes. Therefore, it is important to carefully and thoughtfully select the input variables. This is where explainable artificial intelligence (XAI) comes in, which places greater emphasis on human comprehension of AI processes rather than just strict calculations. This approach makes it easier to understand and apply the models to real-life problems (20,21). The selection of predictors for this research was made based on clinical importance for recurrence occurrence and the completeness of the data—the chosen variables contained complete data available for all of the cases. As a consequence of this approach, the models

Table 2 The descriptive statistics of input variables

Variables	Group	N (%)
Sex	M	1,574 (68.55)
	F	722 (31.45)
Chemotherapy	0	767 (33.41)
	1	107 (4.66)
	2	1,259 (54.83)
	3	100 (4.36)
	4	36 (1.57)
	5	20 (0.87)
Smoking	6	7 (0.30)
	Non-smoker	125 (5.44)
	Ex-smoker	677 (29.49)
Molecular therapy	Smoker	1,494 (65.07)
	N	2,258 (98.34)
Immunotherapy	Y	38 (1.66)
	N	2,274 (99.04)
Radiotherapy	Y	22 (0.96)
	N	1,877 (81.75)
Surgical therapy	Y	419 (18.25)
	N	2,257 (98.30)
Symptoms	Y	39 (1.70)
	N	2,025 (88.20)
Family history	N	271 (11.80)
	Y	2,074 (90.33)
Recurrence	Y	222 (9.67)
	N	2,155 (93.86)
	Y	2,155 (93.86)
	N	141 (6.14)

M, male; F, female.

were built based on nine input variables as predictors and recurrence as predicted variable (*Table 3*). The models differ in the number of neurons in the hidden layer and the activation function in the hidden and output layers, i.e., linear, hyperbolic, sigmoid, tangent, logistic, and exponential. In all the layers the aggregate function was linear and the error function was entropy. The models learned with four learning algorithms—quickprop, backpropagation, quasi-Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS), and conjugate gradients (22-24).

Table 3 Variables selected for the study

Data	Variables	Conversion	Variable value
Input	Sex	Binary	[1] Male
			[2] Female
	Smoking	One-of-N	[0] Non-smoker
			[1] Ex-smoker
			[2] Smoker
	Chemotherapy	One-of-N	The sum of chemotherapeutics used in the whole treatment
			[0] No chemotherapy
			[1] One chemotherapeutic
			[2] Two chemotherapeutics
			[3] Three chemotherapeutics
			[4] Four chemotherapeutics
			[5] Five chemotherapeutics
			[6] Six chemotherapeutic
	Molecular therapy	Binary	[0] No
			[1] Yes
	Immunotherapy	Binary	[0] No
			[1] Yes
	Radiotherapy	Binary	[0] No
			[1] Yes
	Surgical therapy	Binary	[0] No
			[1] Yes
Output	Recurrence	Binary	[0] No
			[1] Yes
			[0] No recurrence
			[1] Recurrence present

The random change in the sequence of cases introduced to the models was applied during the learning process.

For further analysis, the set of nine chosen input variables was used to build the MLP model. The predictor variables were binary (0–1) or multi-state (one-of-N) where all the states were equal. That resulted on obtaining seventeen neurons in the input layer with a linear activation

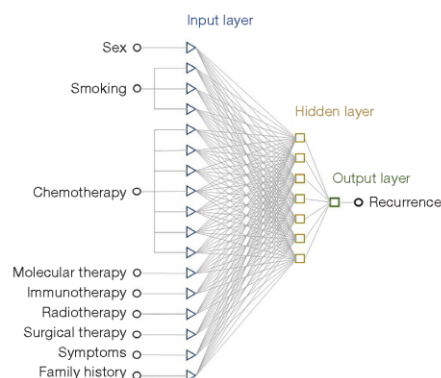


Figure 1 Graph of selected artificial neural network model—multilayer perceptron 9:17-7-1:1.

function. The hidden layer contained seven neurons in with a hyperbolic activation function, and output layer with one neuron and a logistic activation function. The learning rate η was established at 0.01, and the momentum μ was 0.3. The scheme of the chosen network is shown in Figure 1.

The best predictive results were reached using a two-stage learning process. In the first stage, the backpropagation algorithm was used for 100 epochs (meaning all of the cases from the learning set were presented to the network 100 times) and for the next 20 epochs, the learning was continued using the conjugate gradients algorithm.

Statistical analysis

The evaluation of the model is based on statistics for classification models: accuracy (ACC), true positive rate/sensitivity (TPR), specificity (SPC), false positive rate (FPR), false negative rate (FNR), false discovery rate (FDR), precision predictive value (PPV), negative predictive value (NPV), Matthews Correlation Coefficient (MCC) (25).

In the further evaluation of the ANN model the receiver operating characteristic (ROC) curves were drawn for all the sets. The ROC curve describes the SPC and sensitivity of the classifier. The closer to 1 the area under the curve (AUC) the higher quality of the classifier. The ideal classifier AUC is 1, the AUC =0.5 means it's random and useless (26).

In the next stage of the study, the importance of the

Table 4 Multilayer perceptron 9:17-7-1:1 model metrics

Variables	Recurrence, n (%)					
	Learning set (n=1,148, 50%)		Validation set (n=574, 25%)		Test set (n=574, 25%)	
	No	Yes	No	Yes	No	Yes
All	1,070	78	545	29	540	34
Classified correctly	938 (87.66)	67 (85.90)	485 (88.99)	27 (93.10)	487 (90.19)	29 (85.29)
Classified incorrectly	132 (12.34)	11 (14.10)	60 (11.01)	2 (6.90)	53 (9.81)	5 (14.71)

Table 5 Multilayer perceptron 9:17-7-1:1 classification statistics

Variables	Classification statistics		
	Learning set	Validation set	Test set
Quality (ACC)	0.875	0.891	0.899
Error	0.423	0.484	0.479
TPR	0.337	0.310	0.354
SPC	0.988	0.996	0.990
FPR	0.012	0.004	0.010
FNR	0.663	0.690	0.646
FDR	0.141	0.069	0.147
PPV	0.859	0.931	0.853
NPV	0.877	0.890	0.902
MCC	0.489	0.501	0.509

ACC, accuracy; TPR, true positive rate/sensitivity; SPC, specificity; FPR, false positive rate; FNR, false negative rate; FDR, false discovery rate; PPV, precision predictive value; NPV, negative predictive value; MCC, Matthews Correlation Coefficient.

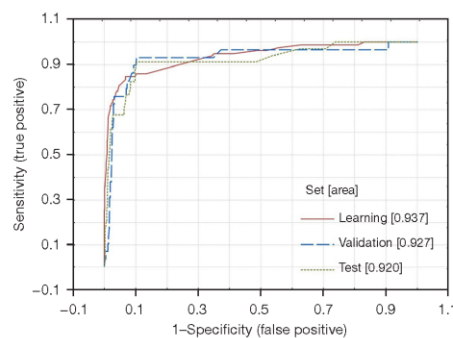


Figure 2 Receiver operating characteristics curves for multilayer perceptron 9:17-7-1:1 learning, test, and validation sets.

variables included in the study and used to build the ANN was evaluated using sensitivity analysis. It allowed to rank the input variables from the most to the least important for the network. Rank one means that without this variable network makes the most significant error, which draws information that this variable brings the most important data to the accurate prediction of ANN.

All the calculations were made with the STATISTICA 7.1 with Neural Networks module (StatSoft Inc., Tulsa, USA, RRID:SCR_014213) was used.

Results

The chosen ANN of architecture MLP 9:17-7-1:1 reached high recurrence predicting values in all three sets. The summary of the model quality and classification statistics of used ANN are shown in Table 4 and Table 5.

The ROC curves were drawn and the AUCs were calculated for learning, validation, and test set which is presented in Figure 2.

The sensitivity analysis carried out allowed us to rank the variables from the most to the least important for the network based on the error value. The results of the conducted analysis are shown in Table 6. In this study, the MLP model was used to predict recurrence occurrence based on retrospective data.

Discussion

AI systems are widely used in medical diagnostics (27-29). According to research, machine learning can effectively detect patterns and relationships among various variables in intricate datasets, such as those pertaining to oncological illnesses (30-32). The vast of described cases refers to predicting the occurrence of cancer, based on risk factors like smoking, family history, and work exposition (33,34). Research by Goryński *et al.* has demonstrated the

Table 6 Sensitive analysis results

Variables	Error	Rank
Chemotherapy	1.785	1
Radiotherapy	1.193	2
Molecular therapy	1.048	3
Smoking	1.039	4
Symptoms	1.023	5
Sex	1.015	6
Immunotherapy	1.008	7
Surgical therapy	0.992	8
Family history	0.982	9

potential of utilizing ANNs as an aid for physicians in diagnosing suspected cases of lung cancer (35). A study by Hsia *et al.* utilized ANN to analyze the longevity of patients with advanced lung cancer, taking into account both clinical data and genetic polymorphism of *p21* and *p53* genes (36). Meanwhile, Marchevsky *et al.* predicted the survival of patients with stage I and II non-small cell lung cancer (NSCLC) using clinical-pathological and immunohistochemical data (37) or based on computed tomography (CT) imaging features as in the study by Wang *et al.* (38). In 2021, Aonpong *et al.* published a study that used available CT images and gene expression data to assess the risk of disease recurrence (39). An article published by Yang *et al.* in the *Computational and Structural Biotechnology Journal* in 2022, demonstrated the efficacy of predictive models generated by machine learning in estimating the risk of recurrence and longevity in patients with NSCLC (40), in addition, machine learning methods were used to predict cancer recurrence after surgical treatment(41,42). Despite being powered by AI, the results of such a model still require human control and supervision to ensure their ACC and effectiveness.

The MLP model employed in our study proved to be highly effective in detecting lung cancer recurrence. Our chosen ANN model demonstrated exceptional predictive capability across all three sets: learning, validation, and test with quality at 0.875, 0.891, and 0.899 respectively.

To ensure an accurate evaluation of the model's performance, Table 5 displays a range of classification metrics. This is important because simply relying on ACC is not enough when dealing with imbalanced datasets. These metrics include TPR, SPC, FPR, FNR, FDR, PPV,

NPV, and MCC, which provide a more comprehensive assessment. The model's ability to accurately classify negative cases as negatives is highlighted by its high SPC or TNR values of 0.988, 0.996, and 0.990 in the learning, validation, and test sets, respectively. Additionally, the model's ability to correctly classify positive cases is underscored by its precision or PPV values of 0.859, 0.931, and 0.853 for the learning, validation, and test sets, respectively. It's worth noting that the FPR metric, which indicates the risk of negative cases being predicted as positive, is crucial given that positive cases only make up about 7% of the entire dataset. The FPR values are 0.012, 0.004, and 0.010 in the learning, validation, and test sets, respectively. However, it's important to pay attention to the TPR and FNR metrics, which reveal the number of correctly classified positive cases and the risk of true positives being classified as negatives. The TPR values are 0.337, 0.310, and 0.354 for the learning, validation, and test sets, respectively, while FNR values reach 0.663 in the learning set, 0.690 in the validation set, and 0.646 in the test set. Therefore, the model's performance needs to be improved in this area to decrease the number of negative misclassifications.

The reliability of ANN as a classifier to distinguish between patients with and without cancer recurrence was further reinforced by the consistently high AUC values of ROC curves, which were all above 0.92. Our sensitivity analysis indicated that the most crucial data pertained to the therapy, specifically chemotherapy, radiotherapy, and molecular therapy information.

The large diversity in lung cancer treatment is a consequence of both recommendations of oncological societies and the current clinical condition of a patient. The choice of specific therapeutic regimens always entails the need to balance benefits and risks for a patient. Cases of inoperable lung cancer may necessitate the treatment composed of sequential or simultaneous chemotherapy and radiotherapy or isolated radiotherapy (when there are contraindications to chemotherapy). The use of chemotherapy in patients in stage IV NSCLC improves the annual rate of overall survival (OS) from 10–20% to 30–50% (43). Combined therapy with the use of two cytostatic drugs is recommended. First-line chemotherapeutics are based on platinum (cisplatin, carboplatin) and administered in combination with taxanes (paclitaxel, docetaxel, or vinorelbine), antimetabolites (gemcitabine or pemetrexed) or alkaloids of the Madagascar periwinkle plant (vinblastine) of comparable activity.

Of significance is also the information on radiotherapy or molecular treatment. In the conducted sensitivity analysis, they were both assigned a high rank, two and three, respectively. For comparison, immunotherapy was assigned the rank of seven. It seems that the above is a consequence of the time period from which our data originated, namely from 2012 to 2017. Oncological therapy is a dynamically developing domain of medicine. On the other hand, radiotherapy is a well-grounded treatment method, which proves to be efficacious and is applied in all stages of a neoplastic disease. Over the past 20 years, we have witnessed a rapid development of radiotherapeutic technology, which can now localize a tumor very accurately, improve the therapeutic outcome and reduce undesirable side effects. For an entire population, in a 5-year-long observational study, the use of this therapeutic method can contribute to prolonging the survival time of patients by 4% (44). The combination of radiotherapy and immunotherapy has recently been the subject of numerous studies. Radiotherapy affects the immune system in several ways: it contributes to a change in the microenvironment of the tumor, to the release of cytokines and chemokines, infiltration of leukocytes, and increased vulnerability of neoplastic cells to cell apoptosis, which improves the effect of immunotherapy (45).

Targeted treatment can be given to nearly 30% of patients with lung cancer. It is currently possible in clinical practice to make an assessment of many mutations, e.g., *EGFR*, *ALK*, *KRAS*, *BRAF*, *ERBB2*, *PI3KC*, which can influence any potential modification in the treatment. The first drug used in practice was gefitinib, which contributed to a considerable, two-fold prolongation of progression-free survival (PFS) with the mutation of *EGFR* in NSCLC (46). The strategy of management and treatment of patients with lung cancer has undergone significant changes in recent years. Many years of clinical research, including preclinical and laboratory studies, have resulted in a modern approach to the oncological treatment. The optimal management strategy for NSCLC is radical surgical resection in patients with stage I and II tumors (47). Patients not qualified or refusing this method of treatment should be treated with radical radiotherapy. Chemoradiation is the preferred treatment for patients with stage II and III disease who are not eligible for surgery. The optimal treatment strategy for stage N2–3A/3B NSCLC remains controversial due to its heterogeneity. Karacz *et al.* in a paper published in 2020, characterized 1,549 cancer patients in terms of time of onset, location and recurrence in patients with lung cancer. The authors of this study showed that cases treated with

systemic therapy had a higher rate of recurrence (48).

For patients with stage IV disease, there are still many uncertainties as to whether local consolidation therapy with radiotherapy or surgery for residual disease can improve OS. In the presented study, we collected data on patients hospitalized in the years 2012–2017, so that the 5-year observation period was maintained. Therefore, the reports presented in the literature look ahead to the future in relation to the patients presented by us. Nevertheless, based on the data available in the literature, it can be concluded that specific monitoring of patients and combination treatment of chemotherapy with immunotherapy brings significant benefits in some patients. In Table 7, a brief review of the literature from the last 5 years on the management of patients with NSCLC is presented.

In our study, other high-ranking variables were symptoms and tobacco smoking. This study dealt with data from patients with lung cancer diagnosed in 2012–2017. At that time, no prophylactic tests with LDCT were carried out in Poland. For this reason, the diagnosis of a neoplastic illness was mostly made when patients had already presented symptoms (88.20%). These could be systemic symptoms (cachexia, paraneoplastic syndromes) or caused by intrabronchial hyperplasia (dyspnea, cough), infiltrations in the chest wall (pain), or distant metastases (brain, bones, adrenal glands, liver, etc.). Another problem is that these symptoms are initially non-specific and patients often postpone a visit to a doctor (62). Consequently, they initially receive symptomatic treatment, until a correct diagnosis is eventually arrived at (63). Most (about 70%) cases of lung cancer are diagnosed late, at the moment when the disease has already progressed to an advanced stage (64). Lung cancer in the IV stage of advancement is considered incurable and the aim of the therapy is to improve survival and mollify the symptoms. Five-year survival in this group of patients is 1–3% (65).

Tobacco smoking is considered to be the gravest risk factor for the development of lung cancer. About 70% of deaths among men and 50% among women due to lung cancer are linked to tobacco smoking (66). It has been proven that the survival of persons who have quit smoking is longer than those who continue the habit (67). In our study, smoking achieved the rank of four, which first and foremost was associated with the fact that we focused on the recurrence of lung cancer rather than on the development of a primary cancer.

In our study, family history and sex were less important factors in terms of the risk of lung cancer recurrence.

Table 7 Literature review

No.	Title	Stage	Group size	Conclusion	References data
1.	Timing, sites, and correlates of lung cancer recurrence	I to III	1,549	Cases treated with systemic therapy had a higher rate of recurrence	(48)
2.	Dynamic recurrence risk and adjuvant chemotherapy benefit prediction by ctDNA in resected non-small cell lung cancer (NSCLC)	II to III	116	Positive ctDNA results after both surgery and adjuvant chemotherapy are significantly associated with poorer recurrence-free survival	(49)
3.	Neoadjuvant nivolumab plus chemotherapy in resectable lung cancer	IB to IIIA	505	The addition nivolumab to chemotherapy resulted in significantly longer event-free survival and a higher percentage of patients with a pathological complete response than chemotherapy alone	(50)
4.	Postoperative intensity-modulated radiation therapy reduces local recurrence and improves overall survival in III-N2 non-small-cell lung cancer: a single-center, retrospective study	III-pN2	183	PORT significantly reduced the frequencies of local recurrence and improved OS in stage III-pN2 NSCLC, especially in the multiple-station pN2 group	(51)
5.	Osimertinib in resected <i>EGFR</i> -mutated non-small-cell lung cancer	IB to IIIA	682	<i>EGFR</i> mutation-positive NSCLC, DFS was significantly longer among those who received osimertinib	(52)
6.	Effect of postoperative radiotherapy for patients with pIIIA-N2 non-small cell lung cancer after complete resection and adjuvant chemotherapy: the phase 3 PORT-C randomized clinical trial	pIIIA-N2	394	Postoperative radiotherapy did not increase DSF	(53)
7.	Five year survival update from KEYNOTE-010: pembrolizumab versus docetaxel for previously treated, programmed death-ligand 1-positive advanced NSCLC	IIIB/IV	1033	The use of pembrolizumab has a positive effect on OS and PFS compared to docetaxel in patients with previously treated advanced NSCLC with PD-L1 expression	(54)
8.	Two-year survival comparing web-based symptom monitoring vs. routine surveillance following treatment for lung cancer	IIA to IV	133	Patients' monitoring during chemotherapy contributes to longer survival time and increasing the chances of detecting recurrence	(55)
9.	Randomized phase II study of pemetrexed or pemetrexed plus bevacizumab for elderly patients with previously untreated non-squamous non-small cell lung cancer: Results of the Lung Oncology Group in Kyushu (LOGIK1201)	IIIB to IV	51	The addition of Bevacizumab to Pemetrexed does not improve survival in elderly patients with NSqNSCLC	(56)
10.	Randomized phase II study of daily versus alternate-day administrations of S-1 for the elderly patients with completely resected pathological stage IA (tumor diameter >2 cm)-IIIA of non-small cell lung cancer: Setouchi Lung Cancer Group Study 1201	IA to IIIA	101	Postoperative adjuvant chemotherapy for NSCLC is associated with improved survival in the elderly population	(57)

Table 7 (continued)

Table 7 (continued)

No.	Title	Stage	Group size	Conclusion	References data
11.	International Tailored Chemotherapy Adjuvant (ITACA) trial, a phase III multicenter randomized trial comparing adjuvant pharmacogenomic-driven chemotherapy versus standard adjuvant chemotherapy in completely resected stage II-IIIa non-small-cell lung cancer	II-IIIa	773	There was no statistical difference in OS in the use of adjuvant chemotherapy in patients after complete resection of NSCLC	(58)
12.	Survival benefit of using pemetrexed for EGFR mutation-positive advanced non-small-cell lung cancer in a randomized phase III study comparing gefitinib to cisplatin plus docetaxel (WJTOG3405)		144	Sequential treatment including pemetrexed against <i>EGFR</i> -mutated NSCLC contributes to significantly longer OS	(59)
13.	Final overall survival results of WJTOG3405, a randomized phase III trial comparing gefitinib versus cisplatin with docetaxel as the first-line treatment for patients with stage IIIB/IV or postoperative recurrent EGFR mutation-positive non-small-cell lung cancer	IIIB/IV	177	The OS advantage of patients treated with gefitinib in a first line treatment compared to patients treated with cisplatin plus docetaxel has not been demonstrated. OS of patients with postoperative recurrence was better than that of stage IIIB/IV disease	(60)
14.	Randomized phase II study of 3 months or 2 years of adjuvant afatinib in patients with surgically resected stage I-III <i>EGFR</i> -Mutant non-small-cell lung cancer	IA to IIIB	46	Long-term (2-year) treatment with afatinib contributes to a reduction in the recurrence compared to shorter (3-month) treatment	(61)
15.	Machine learning application in personalised lung cancer recurrence and survivability prediction	IA to IV (with no subdivision stages)	998	Good predictiveness of CART models were demonstrated	(40)

ctDNA, circulating tumor DNA; PORT, postoperative radiotherapy; OS, overall survival; EGFR, endothelial growth factor receptor; DSF, disease-free survival; PFS, progression-free survival; PD-L1, programmed death ligand 1; NSqNSCLC, non-squamous non-small cell lung cancer; CART, correlation and regression trees.

Yu *et al.* proved the relationship between the history of lung cancer in the family and one's susceptibility to its development (68). Non-smokers who had cases of lung cancer in their families were at a higher risk of developing this type of neoplastic disease than non-smokers with no family members suffering from pulmonary diseases. The literature also provides evidence that there are some differences in patterns of lung cancer recurrence between the sexes (69). This might be associated with exposure to harmful factors, including the smoking habit, and with the clinical stage at which the neoplasm is diagnosed; women, for example, are more often diagnosed with lung adenocarcinoma. On the other hand, Watanabe *et al.* showed differences between men and women related to other histopathological types, e.g., squamous cell carcinoma (70). Such differences can arise from the influence of estrogens

and estrogen receptors on the pathogenesis of lung cancer. However, data related to this issue that is available in the literature seem contradictory (71,72).

In recent times, there has been a surge in the number of studies exploring the use of ANNs for predicting and diagnosing various types of cancer such as breast, ovarian, digestive tract, lung, and skin, among others (73-77). Several studies have been conducted to compare the effectiveness of AI-based techniques with traditional screening methods used in oncology. These methods include plasma biomarkers like prostate-specific antigen (PSA), large intestine endoscopy, and X-ray screening. The machine learning models evaluated the plasma PSA concentration and showed a high predicting ability (78). In a 2020 study, an ANN model was used to detect large intestine polyps and achieved a sensitivity level of 98.1% (79). The effectiveness

of AI algorithms in the evaluation of chest X-ray images was also proven by Chiu *et al.*, which supports early detection of lung cancer (80).

Recent literature suggests that ANN can be a crucial tool in developing decision support systems that can automate processes. In a study conducted by Buciński *et al.*, it was demonstrated that ANN can be used to predict the recurrence of breast cancer using retrospective data that includes basic information about the patient's disease and treatment. The results showed that ANN can achieve a predictive ability of up to 75% (81).

Through our research, we have successfully developed and trained the ANN that accurately predicts recurrence in lung cancer patients. One common issue faced by ANNs is overfitting, where the network memorizes rather than generalizes knowledge, leading to high error levels during testing (77). Despite previous concerns about overfitting in ANN, our study has demonstrated that the network consistently achieves the value of the ROC curves of 0.92 or higher. This indicates that ANN is an effective and reliable tool for identifying patients who are at high risk of lung cancer recurrence.

Our study has several limitations. According to the sensitivity analysis, chemotherapy is the most important parameter for classifying the occurrence of cancer recurrence. However, the multitude of chemotherapy regimens used and the complexity of the therapeutic process did not allow us to evaluate specific drugs used. In the future, it would be worth identifying a separate group of patients treated surgically. However, in our research, they account for only a small percentage. The data presented in this study are only illustrative. Based on this study, we aimed to draw attention to the possibility of using ANNs in assessing the risk of recurrence in lung cancer patients. Nevertheless, further research is required, particularly multicenter studies that would enable the collection of representative groups for each type of treatment used. Our study focused on patients diagnosed with lung cancer between 2012 and 2017. During this time, as mentioned earlier, no normal dose computer tomography (NDCT) studies were conducted in Poland to detect suspicious cancerous growths early. Furthermore, patients were treated according to the standards in force in a given year, published by the Polish Society of Oncology. Crucial for further development would be the creation of additional neural networks that take into account the changes occurring in the strategy of treating cancer patients.

Conclusions

ANNs represent a highly potent technology that demonstrates a remarkable ability to acquire knowledge through learning from examples and subsequently, to generalize such knowledge. Such networks have been effectively applied in the context of recurrence prediction for lung cancer patients. Therefore, it can be inferred that ANNs have the potential to improve both diagnostic and therapeutic processes, in addition to providing indispensable decision-making support for medical professionals.

Acknowledgments

Funding: The study was supported by Nicolaus Copernicus University in Toruń, Poland.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at <https://tlcr.amegroups.com/article/view/10.21037/tlcr-23-350/rc>

Data Sharing Statement: Available at <https://tlcr.amegroups.com/article/view/10.21037/tlcr-23-350/dss>

Peer Review File: Available at <https://tlcr.amegroups.com/article/view/10.21037/tlcr-23-350/prf>

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at <https://tlcr.amegroups.com/article/view/10.21037/tlcr-23-350/coif>). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). Prior to conducting the study, approval was obtained from the management of the medical center and the Bioethics Committee of the Nicolaus Copernicus University in Toruń, Collegium Medicum in Bydgoszcz (No. KB 355/2020), on June 23rd, 2020. Patient survival data were obtained from the Main Statistical Office. Since the study is based on a retrospective analysis of patient medical records the individual consent for this retrospective analysis was waived.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

- Rubin SA. Lung cancer: past, present, and future. *J Thorac Imaging* 1991;7:1-8.
- Pandi A, Mamo G, Getachew D, et al. A Brief Review on Lung Cancer. *Int J Pharm Res Health Sci* 2016;4:907-14.
- National Lung Screening Trial Research Team, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011;365:395-409.
- Adamek M, Biernat W, Chorostowska-Wynimko J, et al. Lung Cancer in Poland. *J Thorac Oncol* 2020;15:1271-6.
- American Cancer Society. Cancer Facts & Figures 2020. Atlanta: American Cancer Society; 2020.
- Dimensions AI. Digital Science & Research Solutions Inc. 2023 [cited 2023 May 17]. Machine Learning and Oncology Publications 2014-2023. Available online: https://app.dimensions.ai/discover/publication?search_mode=content&search_text=artificial%20intelligence%20AND%20oncology&search_type=kws&search_field=full_search&or_facet_year=2023&or_facet_year=2022&or_facet_year=2021&or_facet_year=2020&or_facet_year=2019
- Kleinfeld D. Sequential state generation by model neural networks. *Proc Natl Acad Sci U S A* 1986;83:9469-73.
- Thompson RF. The neurobiology of learning and memory. *Science* 1986;233:941-7.
- Bellotti R, De Carlo F, Massafra R, et al. Topographic classification of EEG patterns in Huntington's disease. *Neurol Clin Neurophysiol* 2004;2004:37.
- Comes MC, La Forgia D, Didonna V, et al. Early Prediction of Breast Cancer Recurrence for Patients Treated with Neoadjuvant Chemotherapy: A Transfer Learning Approach on DCE-MRIs. *Cancers (Basel)* 2021;13:2298.
- El Adoui M, Drisis S, Benjelloun M. Multi-input deep learning architecture for predicting breast tumor response to chemotherapy using quantitative MR images. *Int J Comput Assist Radiol Surg* 2020;15:1491-500.
- Bove S, Fanizzi A, Fadda F, et al. A CT-based transfer learning approach to predict NSCLC recurrence: The added-value of peritumoral region. *PLoS One* 2023;18:e0285188.
- Comes MC, Fucci L, Mele F, et al. A deep learning model based on whole slide images to predict disease-free survival in cutaneous melanoma patients. *Sci Rep* 2022;12:20366.
- Brinker TJ, Kiehl L, Schmitt M, et al. Deep learning approach to predict sentinel lymph node status directly from routine histology of primary melanoma tumours. *Eur J Cancer* 2021;154:227-34.
- Taud H, Mas J. Multilayer Perceptron (MLP). In: Camacho Olmedo M, Paegelow M, Mas JF, et al. editors. *Geomatic Approaches for Modeling Land Change Scenarios. Lecture Notes in Geoinformation and Cartography*. Springer, Cham; 2018:451-5.
- Graupe D. Principles of Artificial Neural Networks. In: Graupe D. *Advanced Series in Circuits and Systems*. 3rd ed. *Advanced Series in Circuits and Systems: Volume 7*. Singapore: World Scientific; 2013.
- Puri M, Solanki A, Padawer T, et al. Chapter 1 - Introduction to Artificial Neural Network (ANN) as a Predictive Tool for Drug Design, Discovery, Delivery, and Disposition: Basic Concepts and Modeling. In: Puri M, Pathak Y, Sutariya VK, et al. *Artificial Neural Network for Drug Design, Delivery and Disposition*. Elsevier Inc.; 2016:3-13.
- Wilamowski BM. Neural network architectures and learning algorithms. *IEEE Eng Med Biol Mag* 2009;3:56-63.
- American College of Surgeons Commission on Cancer. STandards for Oncology Registry Entry (STORE) v. 1.0 2018 [cited 2023 Jul 23]. Available online: https://www.facs.org/-/media/files/quality-programs/cancer/ncdb/store_manual_2018.ashx
- Amoroso N, Pomarico D, Fanizzi A, et al. A roadmap towards breast cancer therapies supported by explainable artificial intelligence. *Appl Sci* 2021;11:4881.
- Gunning D, Stefik M, Choi J, et al. XAI-Explainable artificial intelligence. *Sci Robot* 2019;4:eaay7120.
- Kelley CT. 4. The BFGS Method. Iterative Methods for Optimization. In: *Frontiers in Applied Mathematics* 1999:71-86. [cited 2023 Jul 23]. Available online: <https://doi.org/10.1137/1.9781611970920.ch4>
- Nazareth JL. Conjugate gradient method. *Wiley Interdiscip Rev Comput Stat* 2009;1:348-53.
- Manning T, Sleator RD, Walsh P. Biologically inspired

- intelligent decision making: a commentary on the use of artificial neural networks in bioinformatics. *Bioengineered* 2014;5:80-95.
25. Lever J, Krzywinski M, Altman N. Points of Significance: Classification evaluation. *Nature Methods* 2016;13:603-4.
 26. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol* 2010;5:1315-6.
 27. Wu Z, Wang L, Li C, et al. DeepLRHE: A Deep Convolutional Neural Network Framework to Evaluate the Risk of Lung Cancer Recurrence and Metastasis From Histopathology Images. *Front Genet* 2020;11:768.
 28. Le NQ, Ho QT, Ou YY. Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins. *J Comput Chem* 2017;38:2000-6.
 29. Araújo T, Aresta G, Castro E, et al. Classification of breast cancer histology images using Convolutional Neural Networks. *PLoS One* 2017;12:e0177544.
 30. Lai YH, Chen WN, Hsu TC, et al. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Sci Rep* 2020;10:4679.
 31. Sun Y, Goodison S, Li J, et al. Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics* 2007;23:30-7.
 32. Levitsky A, Pernemalm M, Bernhardson BM, et al. Early symptoms and sensations as predictors of lung cancer: a machine learning multivariate model. *Sci Rep* 2019;9:16504.
 33. Marcus MW, Chen Y, Raji OY, et al. LLPi: Liverpool Lung Project Risk Prediction Model for Lung Cancer Incidence. *Cancer Prev Res (Phila)* 2015;8:570-5.
 34. Park S, Nam BH, Yang HR, et al. Individualized risk prediction model for lung cancer in Korean men. *PLoS One* 2013;8:e54823.
 35. Goryński K, Safian I, Grądzki W, et al. Artificial neural networks approach to early lung cancer detection. *Cent Eur J Med* 2014;9:632-41.
 36. Hsia TC, Chiang HC, Chiang D, et al. Prediction of survival in surgical unresectable lung cancer by artificial neural networks including genetic polymorphisms and clinical parameters. *J Clin Lab Anal* 2003;17:229-34.
 37. Marchevsky AM, Patel S, Wiley KJ, et al. Artificial neural networks and logistic regression as tools for prediction of survival in patients with Stages I and II non-small cell lung cancer. *Mod Pathol* 1998;11:618-25.
 38. Wang X, Duan HH, Nie SD. Prognostic recurrence analysis method for non-small cell lung cancer based on CT imaging. *Proc SPIE* 11321, 2019 International Conference on Image and Video Processing, and Artificial Intelligence. 2019;113211T:411-7.
 39. Aonpong P, Iwamoto Y, Han XH, et al. Genotype-Guided Radiomics Signatures for Recurrence Prediction of Non-Small Cell Lung Cancer. *IEEE Access* 2021;9:90244-54.
 40. Yang Y, Xu L, Sun L, et al. Machine learning application in personalised lung cancer recurrence and survivability prediction. *Comput Struct Biotechnol J* 2022;20:1811-20.
 41. Hindocha S, Charlton TG, Linton-Reid K, et al. A comparison of machine learning methods for predicting recurrence and death after curative-intent radiotherapy for non-small cell lung cancer: Development and validation of multivariable clinical prediction models. *EBioMedicine* 2022;77:103911.
 42. Kim G, Moon S, Choi JH. Deep Learning with Multimodal Integration for Predicting Recurrence in Patients with Non-Small Cell Lung Cancer. *Sensors (Basel)* 2022;22:6594.
 43. da Cunha Santos G, Shepherd FA, Tsao MS. EGFR mutations and lung cancer. *Annu Rev Pathol* 2011;6:49-69.
 44. Shafiq J, Hanna TP, Vinod SK, et al. A Population-based Model of Local Control and Survival Benefit of Radiotherapy for Lung Cancer. *Clin Oncol (R Coll Radiol)* 2016;28:627-38.
 45. Vinod SK, Hau E. Radiotherapy treatment for lung cancer: Current status and future directions. *Respirology* 2020;25 Suppl 2:61-71.
 46. Gutiérrez L, Royuela A, Carcereny E, et al. Prognostic model of long-term advanced stage (IIIB-IV) EGFR mutated non-small cell lung cancer (NSCLC) survivors using real-life data. *BMC Cancer* 2021;21:977.
 47. Montagne F, Guisier F, Venissac N, et al. The Role of Surgery in Lung Cancer Treatment: Present Indications and Future Perspectives-State of the Art. *Cancers (Basel)* 2021;13:3711.
 48. Karacz CM, Yan J, Zhu H, et al. Timing, Sites, and Correlates of Lung Cancer Recurrence. *Clin Lung Cancer* 2020;21:127-135.e3.
 49. Qiu B, Guo W, Zhang F, et al. Dynamic recurrence risk and adjuvant chemotherapy benefit prediction by ctDNA in resected NSCLC. *Nat Commun* 2021;12:6770.
 50. Forde PM, Spicer J, Lu S, et al. Neoadjuvant Nivolumab plus Chemotherapy in Resectable Lung Cancer. *N Engl J Med* 2022;386:1973-85.
 51. Wei W, Zhou J, Zhang Q, et al. Postoperative intensity-modulated radiation therapy reduces local recurrence and improves overall survival in III-N2 non-small-cell lung

- cancer: A single-center, retrospective study. *Cancer Med* 2020;9(8):2820-32.
52. Wu YL, Tsuboi M, He J, et al. Osimertinib in Resected EGFR-Mutated Non-Small-Cell Lung Cancer. *N Engl J Med* 2020;383:1711-23.
 53. Hui Z, Men Y, Hu C, et al. Effect of Postoperative Radiotherapy for Patients With pIIIA-N2 Non-Small Cell Lung Cancer After Complete Resection and Adjuvant Chemotherapy: The Phase 3 PORT-C Randomized Clinical Trial. *JAMA Oncol* 2021;7:1178-85.
 54. Herbst RS, Garon EB, Kim DW, et al. Five Year Survival Update From KEYNOTE-010: Pembrolizumab Versus Docetaxel for Previously Treated, Programmed Death-Ligand 1-Positive Advanced NSCLC. *J Thorac Oncol* 2021;16:1718-32.
 55. Denis F, Basch E, Septans AL, et al. Two-Year Survival Comparing Web-Based Symptom Monitoring vs Routine Surveillance Following Treatment for Lung Cancer. *JAMA* 2019;321:306-7.
 56. Fukuda M, Kitazaki T, Ogawara D, et al. Randomized phase II study of pemetrexed or pemetrexed plus bevacizumab for elderly patients with previously untreated non-squamous non-small cell lung cancer: Results of the Lung Oncology Group in Kyushu (LOGIK1201). *Lung Cancer* 2019;132:1-8.
 57. Yamamoto H, Soh J, Okumura N, et al. Randomized phase II study of daily versus alternate-day administrations of S-1 for the elderly patients with completely resected pathological stage IA (tumor diameter > 2 cm)-IIIA of non-small cell lung cancer: Setouchi Lung Cancer Group Study 1201. *PLoS One* 2023;18:e0285273.
 58. Novello S, Torri V, Grohe C, et al. International Tailored Chemotherapy Adjuvant (ITACA) trial, a phase III multicenter randomized trial comparing adjuvant pharmacogenomic-driven chemotherapy versus standard adjuvant chemotherapy in completely resected stage II-IIIa non-small-cell lung cancer. *Ann Oncol* 2022;33:57-66.
 59. Haratake N, Shimokawa M, Seto T, et al. Survival benefit of using pemetrexed for EGFR mutation-positive advanced non-small-cell lung cancer in a randomized phase III study comparing gefitinib to cisplatin plus docetaxel (WJTOG3405). *Int J Clin Oncol* 2022;27:1404-12.
 60. Yoshioka H, Shimokawa M, Seto T, et al. Final overall survival results of WJTOG3405, a randomized phase III trial comparing gefitinib versus cisplatin with docetaxel as the first-line treatment for patients with stage IIIB/IV or postoperative recurrent EGFR mutation-positive non-small-cell lung cancer. *Ann Oncol* 2019;30:1978-84.
 61. Neal JW, Costa DB, Muzikansky A, et al. Randomized Phase II Study of 3 Months or 2 Years of Adjuvant Afatinib in Patients With Surgically Resected Stage I-III EGFR-Mutant Non-Small-Cell Lung Cancer. *JCO Precis Oncol* 2021;5:325-32.
 62. Swann R, McPhail S, Witt J, et al. Diagnosing cancer in primary care: results from the National Cancer Diagnosis Audit. *Br J Gen Pract* 2018;68:e63-72.
 63. Ellis PM, Vandermeer R. Delays in the diagnosis of lung cancer. *J Thorac Dis* 2011;3:183-8.
 64. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin* 2020;70:7-30.
 65. Bade BC, Dela Cruz CS. Lung Cancer 2020: Epidemiology, Etiology, and Prevention. *Clin Chest Med* 2020;41:1-24.
 66. O'Keeffe LM, Taylor G, Huxley RR, et al. Smoking as a risk factor for lung cancer in women and men: a systematic review and meta-analysis. *BMJ Open* 2018;8:e021611.
 67. Romaszko-Wojtowicz A, Lorenc A, Bucifski A, et al. Effects of Tobacco Smoking on the Survivability of Patients with Multiple Cancers and Single Lung Cancer. *Int J Environ Res Public Health* 2022;19:9179.
 68. Yu F, Xiao R, Li X, et al. Combined effects of lung disease history, environmental exposures, and family history of lung cancer to susceptibility of lung cancer in Chinese non-smokers. *Respir Res* 2021;22:210.
 69. Watanabe K, Tsuboi M, Sakamaki K, et al. Postoperative follow-up strategy based on recurrence dynamics for non-small-cell lung cancer. *Eur J Cardiothorac Surg* 2016;49:1624-31.
 70. Watanabe K, Sakamaki K, Nishii T, et al. Gender Differences in the Recurrence Timing of Patients Undergoing Resection for Non-Small Cell Lung Cancer. *Asian Pac J Cancer Prev* 2018;19:719-24.
 71. Stabile LP, Davis AL, Gubish CT, et al. Human non-small cell lung tumors and cells derived from normal lung express both estrogen receptor alpha and beta and show biological responses to estrogen. *Cancer Res* 2002;62:2141-50.
 72. Patel JD. Lung cancer in women. *J Clin Oncol* 2005;23:3212-8.
 73. Li Z, Koban KC, Schenck TL, et al. Artificial Intelligence in Dermatology Image Analysis: Current Developments and Future Trends. *J Clin Med* 2022;11:6826.
 74. Fan Z, Guo Y, Gu X, et al. Development and validation of an artificial neural network model for non-invasive gastric cancer screening and diagnosis. *Sci Rep* 2022;12:21795.
 75. Feng Y, Wang Z, Cui R, et al. Clinical analysis and

- artificial intelligence survival prediction of serous ovarian cancer based on preoperative circulating leukocytes. *J Ovarian Res* 2022;15:64.
76. Prisciandaro E, Sedda G, Cara A, et al. Artificial Neural Networks in Lung Cancer Research: A Narrative Review. *J Clin Med* 2023;12:880.
 77. Mao WB, Lyu JY, Vaishnani DK, et al. Application of artificial neural networks in detection and diagnosis of gastrointestinal and liver tumors. *World J Clin Cases* 2020;8:3971-7.
 78. Tătaru OS, Vartolomei MD, Rassweiler JJ, et al. Artificial Intelligence and Machine Learning in Prostate Cancer Patient Management-Current Trends and Future Perspectives. *Diagnostics (Basel)* 2021;11:354.
 79. Nadimi ES, Buijs MM, Herp J, et al. Application of deep learning for autonomous detection and localization of colorectal polyps in wireless colon capsule endoscopy. *Computers & Electrical Engineering* 2020;81:106531.
 80. Chiu HY, Peng RH, Lin YC, et al. Artificial Intelligence for Early Detection of Chest Nodules in X-ray Images. *Biomedicines* 2022;10:2839.
 81. Buciniński A, Baczek T, Krysiński J, et al. Clinical data analysis using artificial neural networks (ANN) and principal component analysis (PCA) of patients with breast cancer after mastectomy. *Reports of Practical Oncology & Radiotherapy* 2007;12:9-17.

Cite this article as: Lorenc A, Romaszko-Wojtowicz A, Jaśkiewicz Ł, Doboszyńska A, Buciniński A. Exploring the efficacy of artificial neural networks in predicting lung cancer recurrence: a retrospective study based on patient records. *Transl Lung Cancer Res* 2023;12(10):2083-2097. doi: 10.21037/tlcr-23-350

Artykuł nr 3 – zgoda Komisji Bioetycznej

Uniwersytet Mikołaja Kopernika w Toruniu
Collegium Medicum im L. Rydygiera w Bydgoszczy
KOMISJA BIOETYCZNA

Ul. M. Skłodowskiej-Curie 9, 85-094 Bydgoszcz, tel.(052) 585-35-63, fax.(052) 585-38-11

KB 355/2020

Bydgoszcz, 23.06.2020 r.

Działając na podstawie art.29 ustawy z dnia 5 grudnia 1996 roku o zawodzie lekarza (Dz.U. z 1997 r. Nr 28 poz. 152 (wraz z późniejszymi zmianami), rozporządzenia Ministra Zdrowia i Opieki Społecznej z dnia 11 maja 1999 r. w sprawie szczegółowych zasad powoływania i finansowania oraz trybu działania komisji bioetycznych (Dz.U. Nr 47 poz.480) oraz Zarządzenia Nr 21 Rektora UMK z dnia 4 marca 2009 r. z późn. zm. w sprawie powołania oraz zasad działania Komisji Bioetycznej Uniwersytetu Mikołaja Kopernika w Toruniu przy Collegium Medicum im Ludwika Rydygiera w Bydgoszczy oraz zgodnie z zasadami zawartymi w ICH – GCP

Komisja Bioetyczna przy UMK w Toruniu, Collegium Medicum w Bydgoszczy

(skład podano w załączeniu), na posiedzeniu w dniu **23.06.2020 r.** przeanalizowała wniosek, który złożyła kierownik badania:

mgr Andżelika Lorenc
Katedra Biofarmacji
Collegium Medicum w Bydgoszczy

z zespołem w składzie

- dr n. med. Anna Romaszko-Wojtowicz, mgr Andżelika Lorenc, dr hab. n. med. Anita Doboszyńska. prof. UWM, prof. dr hab. n. farm. Adam Buciński.

w sprawie badania:

„Wykorzystanie zaawansowanych technik inteligencji obliczeniowej w farmacji i medycynie.”

Po zapoznaniu się ze złożonym wnioskiem i w wyniku przeprowadzonej dyskusji oraz głosowania Komisja podjęła:

Uchwałę o pozytywnym zaopiniowaniu wniosku

w sprawie przeprowadzenia badań w zakresie określonym we wniosku pod warunkiem uzyskania zgody osób badanych na przetwarzanie danych osobowych w celach naukowych, a w przypadku braku takiej zgody, analizowania jedynie danych zanonimizowanych, pozbawionych danych personalnych (zgodnie z RODO). Zgoda obejmuje tylko dane z dokumentacji uczestników badania z okresu od 01.01.2012r. do 15.06.2020 r.

Zgoda obowiązuje od daty podjęcia uchwały (23.06.2020 r.) do końca 2023 r.

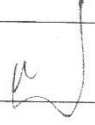
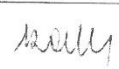
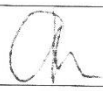
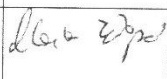
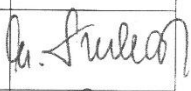
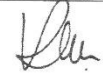
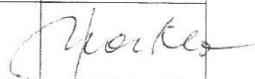
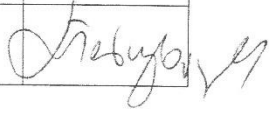
Wydana opinia dotyczy tylko rozpatrywanego wniosku z uwzględnieniem przedstawionego projektu; każda zmiana i modyfikacja wymaga uzyskania odrębnej opinii

Prof. dr hab. med. Karol Śliwka

Przewodniczący Komisji Bioetycznej

Otrzymuje:
mgr Andżelika Lorenc
Katedra Biofarmacji
Collegium Medicum w Bydgoszczy

Lista obecności
na posiedzeniu Komisji Bioetycznej
w dniu 23.06.2020 r.

Lp.	Imię i nazwisko	Funkcja/ Specjalizacja	Podpis
1.	Prof. dr hab. med. Karol Śliwka	Przewodniczący <i>medycyna sądowa</i>	
2.	Mgr prawa Joanna Połetek-Żygas	Z – ca przewodniczącego <i>prawnicza</i>	
3.	Prof. dr hab. med. Mieczysław Czerwionka-Szaflarska	<i>pediatra, alergologia i gastroenterologia dziecięca</i>	
4.	Prof. dr hab. med. Anna Balcar-Boroń	<i>pediatria, nefrologia</i>	
5.	Prof. dr hab. med. Marek Grabiec	<i>położnictwo, ginekologia onkologiczna</i>	
6.	Prof. dr hab. med. Zbigniew Włodarczyk	<i>chirurgia ogólna, transplantologia kliniczna</i>	
7.	Dr hab. n. med. Katarzyna Pawlak-Osińska, prof. UMK	<i>organizacja ochrony zdrowia, otolaryngologia</i>	
8.	Dr hab. n. med. Maria Kłopocka, prof. UMK	<i>choroby wewnętrzne, gastroenterologia</i>	
9.	Ks. dr hab. Wojciech Szukalski, prof. UAM	<i>duchowny</i>	
10.	Dr n. med. Radosława Staszak-Kowalska	<i>pediatria, choroby płuc</i>	
11.	Mgr prawa Patrycja Brzezicka	<i>prawnicza</i>	
12.	Mgr farm. Aleksandra Adamczyk	<i>farmaceutka</i>	
13.	Mgr Lidia Iwińska-Tarczykowska	<i>pielęgniarska</i>	

Oświadczenie – Komisja Bioetyczna

załącznik nr 1c do Uchwały RD 66/2024 z dnia 26.06.2024r

Oświadczenie

Oświadczam, że na prowadzenie badań będących przedmiotem mojej pracy doktorskiej pt. **Zastosowanie algorytmów uczenia maszynowego w analizie predykcyjnej aktywności przeciwdrobnoustrojowej pochodnych imidazoliowych i wznowy raka płuca**, stanowiącej cykl publikacji, zgoda odpowiedniej komisji bioetycznej:

a) nie jest wymagana dla

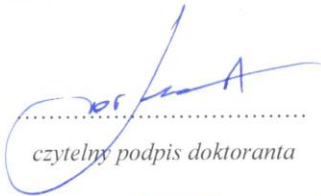
Publikacji nr 1. Lorenc A., Badura A., Karolak M., Pałkowski Ł., Kubik Ł., Buciński A.: Antimicrobial activity classification of imidazolium derivatives predicted by artificial neural networks, *Pharmaceutical Research*, vol. 41, nr 5, 2024, s. 891-898, DOI:10.1007/s11095-024-03699-x

Publikacji nr 2. Lorenc A., Badura A., Karolak M., Pałkowski Ł., Kubik Ł., Skrzypczak A., Buciński A.: Hybrid machine learning model for dual prediction of antimicrobial activity against *Staphylococcus aureus* and *Candida albicans*, *Journal of Pharmaceutical Innovation*, vol. 21, nr 2, 2026, s. 1-9, DOI:10.1007/s12247-026-10412-5,

b) została uzyskana dla

Publikacji nr 3. Lorenc A., Romaszko-Wojtowicz A., Jaśkiewicz Ł., Doboszyńska A., Buciński A.: Exploring the efficacy of artificial neural networks in predicting lung cancer recurrence: a retrospective study based on patient records, *Translational Lung Cancer Research*, Society for Translational Medicine (STM), vol. 12, nr 10, 2023, s. 2083-2097, DOI:10.21037/tlcr-23-350,

Zgoda nr **KB 355/2020** wydana przez Komisję Bioetyczną przy UMK w Toruniu, Collegium Medicum w Bydgoszczy w dniu 23.06.2020.


czytelny podpis doktoranta

Kierownik
Katedry Biofarmacji

prof. dr hab. Adam Buciński
podpis promotora

Opis merytoryczny udziału w publikacjach

załącznik nr 1b do Uchwały RD 66/2024 z dnia 26.06.2024r

Opis merytoryczny udziału w publikacji wieloautorskiej celem podjęcia właściwej uchwały dotyczącej potwierdzenia spełnienia wymagań, o których mowa w art. 186 ust. 1 pkt 3 lit. a i b Ustawy Prawo o Szkolnictwie Wyższym i nauce z dnia 20 lipca 2018, w przypadku publikacji wieloautorskich

Andżelika Lorenc
Imię i nazwisko kandydata

Bydgoszcz, 1.06.2026

Dotyczy publikacji:

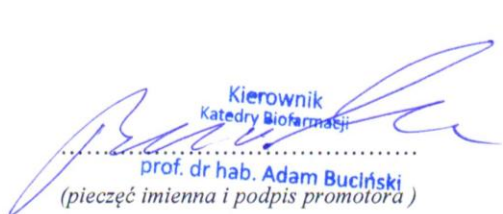
Dane publikacji: Lorenc A., Badura A., Karolak M., Pałkowski Ł., Kubik Ł., Buciński A.: Antimicrobial activity classification of imidazolium derivatives predicted by artificial neural networks, Pharmaceutical Research, vol. 41, nr 5, 2024, s. 891-898, DOI:10.1007/s11095-024-03699-x

Dane bibliometryczne IF: 4,3 , Liczba cytowań: 1 ; Punkty MNiSW: 70

Mój merytoryczny udział ww. publikacji jest następujący:

- opracowanie koncepcji zastosowania metod uczenia maszynowego w badaniu,
- przygotowanie i wstępne przetwarzanie danych do analizy,
- przeprowadzenie selekcji deskryptorów molekularnych z wykorzystaniem metody CART,
- redukcja wymiarowości wyselekcjonowanego zbioru deskryptorów metodą PCA,
- zaprojektowanie, wytrenowanie i walidacja modeli sztucznych sieci neuronowych typu MLP,
- analiza i interpretacja uzyskanych wyników,
- przygotowanie manuskryptu publikacji.

Mój udział procentowy wynosi: 65%


Kierownik
Katedry Biofarmacji
prof. dr hab. Adam Buciński
(pieczęć imienna i podpis promotora)


(kandydat)

Opis merytoryczny udziału w publikacji wieloautorskiej celem podjęcia właściwej uchwały dotyczącej potwierdzenia spełnienia wymagań, o których mowa w art. 186 ust. 1 pkt 3 lit. a i b Ustawy Prawo o Szkolnictwie Wyższym i nauce z dnia 20 lipca 2018, w przypadku publikacji wieloautorskich

Andżelika Lorenc
Imię i nazwisko kandydata

Bydgoszcz, 1.06.2026

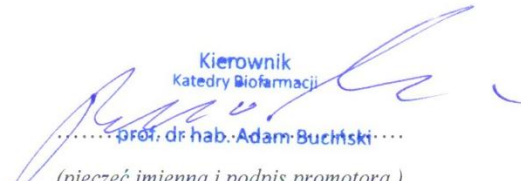
Dotyczy publikacji:

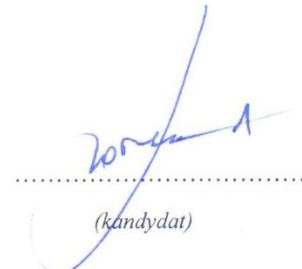
Dane publikacji: Lorenc A., Badura A., Karolak M., Pałkowski Ł., Kubik Ł., Skrzypczak A., Buciński A.: Hybrid machine learning model for dual prediction of antimicrobial activity against Staphylococcus aureus and Candida albicans, Journal of Pharmaceutical Innovation, vol. 21, nr 2, 2026, Numer artykułu: 174, s. 1-9, DOI:10.1007/s12247-026-10412-5, Dane bibliometryczne IF: 2,7 , Liczba cytowań: 0 ; Punkty MNiSW: 70

Mój merytoryczny udział ww. publikacji jest następujący:

- opracowanie koncepcji badania i strategii analitycznej,
- przygotowanie oraz przetwarzanie danych eksperymentalnych,
- przeprowadzenie selekcji zmiennych metodą CART oraz redukcji wymiarowości wyselekcjonowanych cech metodą PCA,
- projektowanie, optymalizacja i walidacja modeli MLP,
- opracowanie podejścia zespołowego do predykcji wartości $\ln(\text{MIC})$,
- analiza i interpretacja wyników badań,
- przygotowanie manuskryptu publikacji.

Mój udział procentowy wynosi: 60%


Kierownik
Katedry Biofarmacji
.....prof. dr hab. Adam Buchański.....
(pieczęć imienna i podpis promotora)


.....
(kandydat)

załącznik nr 1b do Uchwały RD 66/2024 z dnia 26.06.2024r

Opis merytoryczny udziału w publikacji wieloautorskiej celem podjęcia właściwej uchwały dotyczącej potwierdzenia spełnienia wymagań, o których mowa w art. 186 ust. 1 pkt 3 lit. a i b Ustawy Prawo o Szkolnictwie Wyższym i nauce z dnia 20 lipca 2018, w przypadku publikacji wieloautorskich

Andżelika Lorenc

Bydgoszcz, 1.06.2026

Imię i nazwisko kandydata

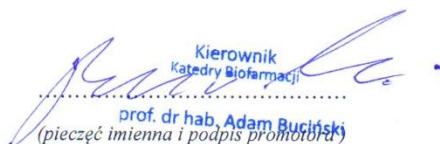
Dotyczy publikacji:

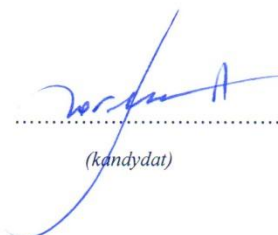
Dane publikacji: Lorenc A., Romaszko-Wojtowicz A., Jaśkiewicz Ł., Doboszyńska A., Bucinski A.: Exploring the efficacy of artificial neural networks in predicting lung cancer recurrence: a retrospective study based on patient records, Translational Lung Cancer Research, Society for Translational Medicine (STM), vol. 12, nr 10, 2023, s. 2083-2097, DOI:10.21037/tlcr-23-350, Dane bibliometryczne IF: 4 , Liczba cytowań: 10 ; Punkty MNiSW: 140

Mój merytoryczny udział ww. publikacji jest następujący:

- współudział w opracowaniu koncepcji badania,
- przygotowanie danych klinicznych do analizy,
- projektowanie, trenowanie i walidacja modeli sztucznych sieci neuronowych typu MLP,
- ocena jakości modeli z wykorzystaniem miar klasyfikacyjnych oraz analiz ROC,
- przeprowadzenie analizy wrażliwości predyktorów,
- analiza i interpretacja uzyskanych wyników,
- współudział w przygotowaniu manuskryptu publikacji.

Mój udział procentowy wynosi: 45%


Kierownik
Katedry Biofarmacji
prof. dr hab. Adam Bucinski
(pieczęć imienna i podpis promotora)


(kandydat)

Oświadczenia współautorów

Załącznik nr 5 do uchwały Nr 38 Senatu UMK z dnia 26 września 2023 r.
w sprawie postępowania o nadanie stopnia doktora
na Uniwersytecie Mikołaja Kopernika w Toruniu

Bydgoszcz, dnia 27.05.2026

dr n. med. i n. o zdr. Anna Maria Badura
Katedra Biofarmacji
Wydział Farmaceutyczny
Collegium Medicum im. Ludwika Rydygiera
Uniwersytet Mikołaja Kopernika w Toruniu

Rada Dyscypliny Nauki Farmaceutyczne
Uniwersytetu Mikołaja Kopernika w Toruniu

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracach:

- Lorenc A., Badura A., Karolak M., Pałkowski Ł., Kubik Ł., Skrzypczak A., Buciński A.:
Hybrid machine learning model for dual prediction of antimicrobial activity against *Staphylococcus aureus* and *Candida albicans*, Journal of Pharmaceutical Innovation, vol. 21, nr 2, 2026, Numer artykułu: 174, s. 1-9, DOI:10.1007/s12247-026-10412-5 mój udział polegał na **sprawowaniu merytorycznego nadzoru nad częścią badań dotyczącą metod uczenia maszynowego.**

Mój udział w powstaniu pracy wynosi **10%**.

- Lorenc A., Badura A., Karolak M., Pałkowski Ł., Kubik Ł., Buciński A.: **Antimicrobial activity classification of imidazolium derivatives predicted by artificial neural networks**, Pharmaceutical Research, vol. 41, nr 5, 2024, s. 891-898, DOI:10.1007/s11095-024-03699-x mój udział polegał na **sprawowaniu merytorycznego nadzoru nad częścią badań dotyczącą metod uczenia maszynowego.**

Mój udział w powstaniu pracy wynosi **10%**.

.....
A. Badura
(podpis)

Bydgoszcz, dnia 27.05.2026

dr Maciej Karolak
Katedra Technologii Postaci Leku
Wydział Farmaceutyczny
Collegium Medicum im. Ludwika Rydygiera
Uniwersytet Mikołaja Kopernika w Toruniu

Rada Dyscypliny Nauki Farmaceutyczne
Uniwersytetu Mikołaja Kopernika w Toruniu

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracach:

- Lorenc A., Badura A., Karolak M., Pałkowski Ł., Kubik Ł., Skrzypczak A., Buciński A.:
Hybrid machine learning model for dual prediction of antimicrobial activity against *Staphylococcus aureus* and *Candida albicans*, Journal of Pharmaceutical Innovation, vol. 21, nr 2, 2026, Numer artykułu: 174, s. 1-9, DOI:10.1007/s12247-026-10412-5 mój udział polegał na **pozyskaniu danych do analiz oraz sprawowaniu merytorycznego nadzoru nad częścią chemoinformatyczną badań.**

Mój udział w powstaniu pracy wynosi **10%**.

- Lorenc A., Badura A., Karolak M., Pałkowski Ł., Kubik Ł., Buciński A.: Antimicrobial activity classification of imidazolium derivatives predicted by artificial neural networks, Pharmaceutical Research, vol. 41, nr 5, 2024, s. 891-898, DOI:10.1007/s11095-024-03699-x mój udział polegał na **pozyskaniu danych do analiz oraz realizacji i nadzorowaniu części chemoinformatycznej badań.**

Mój udział w powstaniu pracy wynosi **10%**.

Maciej... Karolak
(podpis)

Łozanna, dnia 09.06.2026

dr Łukasz Kubik

Rada Dyscypliny Nauki Farmaceutyczne
Uniwersytetu Mikołaja Kopernika w Toruniu

Oświadczenie o współautorstwie

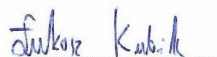
Niniejszym oświadczam, że w pracach:

- Lorenc A., Badura A., Karolak M., Pałkowski Ł., Kubik Ł., Skrzypczak A., Buciński A.: **Hybrid machine learning model for dual prediction of antimicrobial activity against *Staphylococcus aureus* and *Candida albicans***, Journal of Pharmaceutical Innovation, vol. 21, nr 2, 2026, Numer artykułu: 174, s. 1-9, DOI:10.1007/s12247-026-10412-5 mój udział polegał na **pozyskaniu deskryptorów molekularnych do analiz przy użyciu programu Dragon 7.0 (Talet, Mediolan, Włochy)**.

Mój udział w powstaniu pracy wynosi 5%.

- Lorenc A., Badura A., Karolak M., Pałkowski Ł., Kubik Ł., Buciński A.: **Antimicrobial activity classification of imidazolium derivatives predicted by artificial neural networks**, Pharmaceutical Research, vol. 41, nr 5, 2024, s. 891-898, DOI:10.1007/s11095-024-03699-x mój udział polegał na **pozyskaniu deskryptorów molekularnych do analiz przy użyciu programu Dragon 7.0 (Talet, Mediolan, Włochy)**.

Mój udział w powstaniu pracy wynosi 5%.


(podpis)

Bydgoszcz, dnia 06.05.2026

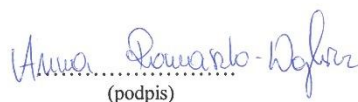
dr hab. n. med. Anna Romaszko-Wojtowicz, prof. UWM
Katedra Pulmonologii
Szkoła Zdrowia Publicznego
Collegium Medicum, Uniwersytet Warmińsko-Mazurski

Rada Dyscypliny Nauki Farmaceutyczne
Uniwersytetu Mikołaja Kopernika w Toruniu

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy Lorenc A., Romaszko-Wojtowicz A., Jaśkiewicz Ł., Doboszyńska A., Buciński A.: **Exploring the efficacy of artificial neural networks in predicting lung cancer recurrence: a retrospective study based on patient records**, *Translational Lung Cancer Research*, Society for Translational Medicine (STM), vol. 12, nr 10, 2023, s. 2083-2097, DOI:10.21037/tlcr-23-350) mój udział polegał na: współtworzeniu koncepcji badania, zbieraniu i kompletowaniu danych, oraz współudziale w przygotowaniu manuskryptu i jego ostatecznej akceptacji.

Mój udział w powstaniu pracy wynosi **35%**.


(podpis)

Bydgoszcz, dnia 06.05.2026

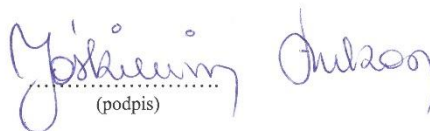
dr n. med. Łukasz Jaśkiewicz
Katedra Fizjologii i Patofizjologii Człowieka
Wydział Lekarski
Collegium Medicum, Uniwersytet Warmińsko-Mazurski

Rada Dyscypliny Nauki Farmaceutyczne
Uniwersytetu Mikołaja Kopernika w Toruniu

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy Lorenc A., Romaszko-Wojtowicz A., Jaśkiewicz Ł., Doboszyńska A., Bucinski A.: **Exploring the efficacy of artificial neural networks in predicting lung cancer recurrence: a retrospective study based on patient records**, *Translational Lung Cancer Research*, Society for Translational Medicine (STM), vol. 12, nr 10, 2023, s. 2083-2097, DOI:10.21037/tlcr-23-350) mój udział polegał na: przygotowaniu danych, ich interpretacji, przygotowaniu manuskryptu oraz akceptacji ostatecznej jego wersji.

Mój udział w powstaniu pracy wynosi 10%.


(podpis)

Bydgoszcz, dnia 06.05.2026

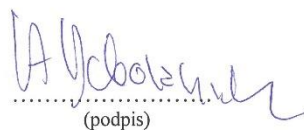
prof. dr hab. n. med. Anna Doboszyńska
Mazowieckie Centrum Leczenia Chorób Płuc i Gruzlicy

Rada Dyscypliny Nauki Farmaceutyczne
Uniwersytetu Mikołaja Kopernika w Toruniu

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracy Lorenc A., Romaszko-Wojtowicz A., Jaśkiewicz Ł., Doboszyńska A., Buciński A.: **Exploring the efficacy of artificial neural networks in predicting lung cancer recurrence: a retrospective study based on patient records**, *Translational Lung Cancer Research*, Society for Translational Medicine (STM), vol. 12, nr 10, 2023, s. 2083-2097, DOI:10.21037/tlcr-23-350) mój udział polegał na: współudziale w przygotowaniu manuskryptu oraz jego ostatecznej weryfikacji.

Mój udział w powstaniu pracy wynosi 5%.


(podpis)

Bydgoszcz, dnia 27.05.2026

prof. dr hab. n. farm. Adam Buciński
Katedra Biofarmacji
Wydział Farmaceutyczny
Collegium Medicum im. Ludwika Rydygiera
Uniwersytet Mikołaja Kopernika w Toruniu

Rada Dyscypliny Nauki Farmaceutyczne
Uniwersytetu Mikołaja Kopernika w Toruniu

Oświadczenie o współautorstwie

Niniejszym oświadczam, że w pracach:

- Lorenc A., Badura A., Karolak M., Pałkowski Ł., Kubik Ł., Skrzypczak A., Buciński A.: **Hybrid machine learning model for dual prediction of antimicrobial activity against *Staphylococcus aureus* and *Candida albicans***, Journal of Pharmaceutical Innovation, vol. 21, nr 2, 2026, Numer artykułu: 174, s. 1-9, DOI:10.1007/s12247-026-10412-5 mój udział polegał na **sprawowaniu merytorycznego nadzoru nad częścią badań dotyczącą metod uczenia maszynowego**.
Mój udział w powstaniu pracy wynosi 5%.
- Lorenc A., Badura A., Karolak M., Pałkowski Ł., Kubik Ł., Buciński A.: Antimicrobial activity classification of imidazolium derivatives predicted by artificial neural networks, Pharmaceutical Research, vol. 41, nr 5, 2024, s. 891-898, DOI:10.1007/s11095-024-03699-x mój udział polegał na **sprawowaniu merytorycznego nadzoru nad częścią badań dotyczącą metod uczenia maszynowego**.
Mój udział w powstaniu pracy wynosi 5%.
- Lorenc A., Romaszko-Wojtowicz A., Jaśkiewicz Ł., Doboszyńska A., Buciński A.: **Exploring the efficacy of artificial neural networks in predicting lung cancer recurrence: a retrospective study based on patient records**, Translational Lung Cancer Research, Society for Translational Medicine (STM), vol. 12, nr 10, 2023, s. 2083-2097, DOI:10.21037/tlcr-23-350, mój udział polegał na **sprawowaniu merytorycznego nadzoru nad częścią badań dotyczącą metod uczenia maszynowego**.
Mój udział w powstaniu pracy wynosi 5%.



.....
(podpis)

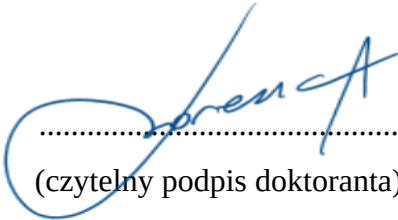
Bydgoszcz, 10.06.2026

mgr Andżelika Lorenc
Katedra Biofarmacji
Wydział Farmaceutyczny
Collegium Medicum im. Ludwika Rydygiera
Uniwersytet Mikołaja Kopernika w Toruniu

Rada Dyscypliny Nauki Farmaceutyczne
Uniwersytetu Mikołaja Kopernika w Toruniu

Oświadczenie

Informuję, że współautor Publikacji nr 2, wchodzącej w skład niniejszej rozprawy,
dr hab. Andrzej Skrzypczak, Instytut Technologii i Inżynierii Chemicznej, Wydział
Technologii Chemicznej Politechniki Poznańskiej nie może złożyć wymaganego oświadczenia
dotyczącego udziału w powstaniu publikacji z uwagi na jego śmierć.
Zakres jego wkładu określono na 5%, zgodnie z informacjami zawartymi w publikacji oraz
oświadczeniami pozostałych współautorów.



(czytelny podpis doktoranta)

Życiorys naukowy

Życiorys naukowy
mgr Andżelika Lorenc
ORCID: 0000-0002-1474-7864

I. Wykształcenie

2012-2017 – Uniwersytet Mikołaja Kopernika w Toruniu,
Collegium Medicum im. Ludwika Rydygiera w Bydgoszczy, Wydział Farmaceutyczny
studia magisterskie – analityka medyczna
uzyskanie tytułu zawodowego magistra na podstawie pracy magisterskiej
pt. Naczyniowo-śródbłonkowy czynnik wzrostu (VEGF) w raku sutka
promotor: dr hab. n. med. Barbara Ciastek

II. Zatrudnienie

2017 – obecnie - Uniwersytet Mikołaja Kopernika w Toruniu,
Collegium Medicum im. Ludwika Rydygiera w Bydgoszczy, Wydział Farmaceutyczny,
Katedra Biofarmacji
asystent – pracownik naukowo-dydaktyczny

III. Działalność naukowa

1. Artykuły naukowe i rozdziały z monografii:

Sumaryczny IF: **83.503**

Sumaryczna punktacja ministerialna: **1190.000**

Lorenc Andżelika, Badura Anna, Karolak Maciej, Pałkowski Łukasz, Kubik Łukasz, Skrzypczak Andrzej, Buciński Adam: Hybrid machine learning model for dual prediction of antimicrobial activity against Staphylococcus aureus and Candida albicans, Journal of Pharmaceutical Innovation, vol. 21, nr 2, 2026, Numer artykułu: 174, s. 1-9, DOI:10.1007/s12247-026-10412-5, **70 punktów, IF(2,7)**

Kokotkiewicz Adam, Król Agata, Al-Busairi Rafał, Godlewska Sylwia, Sparzak-Stefanowska Barbara, Badura Anna, **Lorenc Andżelika**, Marzec-Wróblewska Urszula, Siłuk Danuta, Buciński Adam: In-sample stability of phenolic abietane diterpenes : a case of Salvia apiana Jeps., Industrial Crops and Products, vol. 241, 2026, Numer artykułu: 122742, s. 1-12, DOI:10.1016/j.indcrop.2026.122742, łączna liczba autorów: 12, **200 punktów, IF(6,2)**

Król Agata, Kokotkiewicz Adam, Królicka Aleksandra, Hinc Krzysztof, Badura Anna, **Lorenc Andżelika**, Marzec-Wróblewska Urszula, Buciński Adam, Kuźma Łukasz, Łuczkiwicz Maria: Adventitious and hairy root cultures of Salvia apiana as a source of rosmarinic acid, International Journal of Molecular Sciences, MDPI, vol. 26, nr 7, 2025, Numer artykułu: 3138, s. 1-24, DOI:10.3390/ijms26073138, **140 punktów, IF(4,9)**

- Lorenc Andželika**, Badura Anna, Karolak Maciej, Pańkowski Łukasz, Kubik Łukasz, Buciński Adam: Antimicrobial activity classification of imidazolium derivatives predicted by artificial neural networks, *Pharmaceutical Research*, vol. 41, nr 5, 2024, s. 891-898, DOI:10.1007/s11095-024-03699-x, **70 punktów, IF(4,3)**
- Mendes Bárbara B., Zhang Zilu, Conniot João, Sousa Diana P., Ravasco João M. J. M., Onweller Lauren A., **Lorenc Andželika**, Rodrigues Tiago, Reker Daniel, Conde João: A large-scale machine learning analysis of inorganic nanoparticles in preclinical cancer research, *Nature Nanotechnology*, vol. 19, nr 6, 2024, s. 867-878, DOI:10.1038/s41565-024-01673-7, **200 punktów, IF(35,1)**
- Repczyńska Anna, Jułga Katarzyna, **Lorenc Andželika**, Skalska-Sadowska Jolanta, Wysocki Mariusz, Zaucha-Prażmo Agnieszka, Drabko Katarzyna, Bossowski Artur, Buciński Adam, Haus Olga: Cytogenetic findings in Polish patients with suspected Fanconi anemia, *Advances in Clinical and Experimental Medicine*, Uniwersytet Medyczny im. Piastów Śląskich we Wrocławiu, vol. 33, nr 4, 2024, s. 361-368, DOI:10.17219/acem/168825, łączna liczba autorów: 12, **70 punktów, IF(1,9)**
- Lorenc Andželika**, Romaszko-Wojtowicz Anna, Jaśkiewicz Łukasz, Doboszyńska Anna, Buciński Adam: Exploring the efficacy of artificial neural networks in predicting lung cancer recurrence: a retrospective study based on patient records, *Translational Lung Cancer Research*, Society for Translational Medicine (STM), vol. 12, nr 10, 2023, s. 2083-2097, DOI:10.21037/tlcr-23-350, **140 punktów, IF(4)**
- Romaszko-Wojtowicz Anna, **Lorenc Andželika**, Buciński Adam, Doboszyńska Anna: Effects of tobacco smoking on the survivability of patients with multiple cancers and single lung cancer, *International Journal of Environmental Research and Public Health*, vol. 19, nr 15, 2022, Numer artykułu: 9179, s. 1-11, DOI:10.3390/ijerph19159179, **140 punktów**
- Kokotkiewicz A., Badura Anna, Tabaczyńska Żaneta, **Lorenc Andželika**, Buciński Adam, Luczkiewicz M.: Optimization of distillation conditions for improved recovery of phthalides from celery (*Apium graveolens* L.) seeds, *Polish Journal of Food and Nutrition Sciences*, vol. 71, nr 2, 2021, s. 197-210, DOI:10.31883/pjfn/137612, **100 punktów, IF(2,736)**
- Lorenc Andželika, Mendes Bárbara B., Conniot João, Sousa Diana P., Conde João, Rodrigues Tiago: Machine learning for next-generation nanotechnology in healthcare, *Matter*, vol. 4, nr 10, 2021, s. 3078-3080, DOI:10.1016/j.matt.2021.09.014, **20 punktów, IF(19,967)**
- Badura Anna, Marzec-Wróblewska Urszula, Kamiński Piotr, Łakota P., Ludwikowski Grzegorz, Szymański Marek, Wasilow Karolina, Lorenc Andželika, Buciński Adam: Prediction of semen quality using artificial neural network, *Journal of Applied Biomedicine*, vol. 17, nr 3, 2019, s. 167-174, DOI:10.32725/jab.2019.015, **40 punktów, IF(1,7)**

2. Projekty badawcze:

- 2020 – obecnie** „Możliwości wykorzystania uczenia maszynowego, ze szczególnym uwzględnieniem sztucznych sieci neuronowych (ANN), w naukach farmaceutycznych i medycznych”, Katedra Biofarmacji, Wydział Farmaceutyczny, Collegium Medicum im. Ludwika Rydygiera w Bydgoszczy, Uniwersytet Mikołaja Kopernika w Toruniu - **członkini zespołu projektowego**
- 2023 – 2025** „Wybrane aspekty analizy fitochemicznej frakcji olejkowych oraz fenolowych w biomasach szalwii białej (*Salvia apiana*)” we współpracy z Katedrą i Zakładem Farmakognozji z Ogrodem Roślin Leczniczych Gdańskiego Uniwersytetu Medycznego. - **członkini zespołu projektowego**
- 2021 – 2024** “Machine learning for next-generation nanotechnology in healthcare” we współpracy z naukowcami Faculdade de Farmácia da Universidade de Lisboa, Portugalia - **członkini zespołu projektowego**
- 2020 – 2023** „Wykorzystanie zaawansowanych technik inteligencji obliczeniowej w farmacji i medycynie” we współpracy z Katedrą Pulmonologii Szkoły Zdrowia Publicznego Uniwersytetu Warmińsko-Mazurskiego w Olsztynie – **kierowniczka projektu**

IV. Udział w kursach i szkoleniach

- styczeń – czerwiec 2026** – kurs języka angielskiego WorkSmart English, Uniwersytet Otwarty UMK
- marzec – czerwiec 2025** - kurs języka angielskiego English for Startups, YUFE
- październik 2024** – Kurs Farmakokinetiki, Warsaw Health Foundation Hub, Agencja Badań Medycznych
- grudzień 2023 – marzec 2024** - Datamining dla niezaawansowanych, Ośrodek Analiz Statystycznych UMK
- październik 2023** - Statystyczna analiza danych, Politechnika Białostocka
- październik 2023** - Statystyka opisowa, Uniwersytet Ekonomiczny w Poznaniu
- wrzesień 2023** - Podstawy analizy danych w języku Python, Ośrodek Analiz Statystycznych UMK
- październik 2022 – czerwiec 2026** – kurs języka angielskiego B2 upper-intermediate, szkoła językowa English-Line
- maj – lipiec 2021** - Staż naukowy w ramach projektu “Machine learning for next generation nanotechnology in healthcare”, Faculdade de Farmácia da Universidade de Lisboa, Portugalia
- luty – czerwiec 2021** – kurs języka angielskiego B2 upper-intermediate, szkoła językowa English-Line

V. Działalność organizacyjna i dydaktyczna

Przygotowanie oraz prowadzenie zajęć dla studentów wizytujących w ramach programów SEP oraz ERASMUS+

Współprowadząca zajęcia podczas „V edycji Praktycznych Warsztatów dla Szkół Ponadpodstawowych” organizowanych w dniach 23-27 września 2024 roku przez Wydział Farmaceutyczny Collegium Medicum UMK

Przedstawicielka Samorządu Doktorantów UMK w Radzie Dyscypliny Nauki Farmaceutyczne (2019 – 2022)

Członkini Uczelnianej Komisji Stypendialnej (2019 – 2024)

Członkini Komisji ds. Studiów Doktoranckich Wydziału Farmaceutycznego (2019 – 2022)

Członkini Komisji Wyborczej Uniwersytetu (2019, 2022)

Członkini Samorządu Doktorantów UMK (2019 – 2022)

Członkini Wydziałowej Rady ds. Jakości Kształcenia (2019 – 2024)

Członkini Wydziałowej Komisji ds. Wyróżnień Studentów i Absolwentów (2023/2024)

VI. Nagrody i wyróżnienia

2025 – Nagroda zespołowa Rektora Uniwersytetu Mikołaja Kopernika za osiągnięcia uzyskane w dziedzinie naukowej w 2024r.

2024 – Nagroda Rektora Uniwersytetu Mikołaja Kopernika w Toruniu za publikacje w prestiżowym czasopiśmie naukowym

2024 – Nagroda Rektora Uniwersytetu Mikołaja Kopernika za osiągnięcia uzyskane w dziedzinie naukowej w 2023r.

2024 – Pierwsza dziesiątka najlepszych nauczycieli w Raporcie Oceny Zajęć Dydaktycznych na Wydziale Farmaceutycznym CM UMK w roku akademickim 2023/2024

2023 – Nagroda zespołowa Rektora Uniwersytetu Mikołaja Kopernika za osiągnięcia uzyskane w dziedzinie naukowej w 2022r.



mgr Andżelika Lorenc

Wykaz publikacji

UNIWERSYTET MIKOŁAJA KOPERNIKA W TORUNIU
Collegium Medicum im. Ludwika Rydygiera w Bydgoszczy
Biblioteka Medyczna
ul. M. Curie Skłodowskiej 9, 85-094 Bydgoszcz
tel. 52 585 3511, fax 52 585 3505
e-mail: biblio@cm.umk.pl

Bydgoszcz, 9.06.2026 r.

OBG-02-1.63.219.2026

WYKAZ PUBLIKACJI

mgr ANDŻELIKA LORENC

Łączna wartość punktacji KBN/MNiSzW: 1190.000

Wartość wskaźnika IF: 83.503

Cytowania (Web of Science): 120

Index H (Web of Science): 5

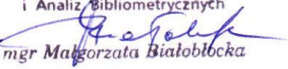
Cytowania (Scopus): 124

Index H (Scopus): 5

Lp.	Opis bibliograficzny	Punkty	IF
2026			
Artykuły z czasopism			
Artykuł naukowy			
1	Kokotkiewicz Adam, Król Agata, Al-Busairi Rafal, Godlewska Sylwia, Sparak-Stefanowska Barbara, Badura Anna, Lorenc Andżelika, Marzec-Wróblewska Urszula, Siluk Danuta, Buciński Adam: In-sample stability of phenolic abietane diterpenes : a case of <i>Salvia apiana</i> Jeps., Industrial Crops and Products, vol. 241, 2026, Numer artykułu: 122742, s. 1-12, DOI:10.1016/j.indcrop.2026.122742, łączna liczba autorów: 12	200	6,2
2	Lorenc Andżelika, Badura Anna, Karolak Maciej, Pałkowski Łukasz, Kubik Łukasz, Skrzypczak Andrzej, Buciński Adam: Hybrid machine learning model for dual prediction of antimicrobial activity against <i>Staphylococcus aureus</i> and <i>Candida albicans</i> , Journal of Pharmaceutical Innovation, vol. 21, nr 2, 2026, Numer artykułu: 174, s. 1-9, DOI:10.1007/s12247-026-10412-5	70	2,7
2025			
Artykuły z czasopism			
Artykuł naukowy			

1	Król Agata, Kokotkiewicz Adam, Królicka Aleksandra, Hinc Krzysztof, Badura Anna, Lorenc Andżelika, Marzec-Wróblewska Urszula, Buciniński Adam, Kuźma Łukasz, Łuczkiwicz Maria: Adventitious and hairy root cultures of <i>Salvia apiana</i> as a source of rosmarinic acid, International Journal of Molecular Sciences, MDPI, vol. 26, nr 7, 2025, Numer artykułu: 3138, s. 1-24, DOI:10.3390/ijms26073138	140	4,9
2024			
Artykuły z czasopism			
Artykuł naukowy			
1	Lorenc Andżelika, Badura Anna, Karolak Maciej, Pałkowski Łukasz, Kubik Łukasz, Buciniński Adam: Antimicrobial activity classification of imidazolium derivatives predicted by artificial neural networks, Pharmaceutical Research, vol. 41, nr 5, 2024, s. 891-898, DOI:10.1007/s11095-024-03699-x	70	4,3
2	Mendes Bárbara B., Zhang Zilu, Coniot João, Sousa Diana P., Ravasco João M. J. M., Onweller Lauren A., Lorenc Andżelika, Rodrigues Tiago, Reker Daniel, Conde João: A large-scale machine learning analysis of inorganic nanoparticles in preclinical cancer research, Nature Nanotechnology, vol. 19, nr 6, 2024, s. 867-878, DOI:10.1038/s41565-024-01673-7	200	35,1
3	Repczyńska Anna, Julga Katarzyna, Lorenc Andżelika, Skalska-Sadowska Jolanta, Wysocki Mariusz, Zaucha-Prażmo Agnieszka, Drabko Katarzyna, Bossowski Artur, Buciniński Adam, Haus Olga: Cytogenetic findings in Polish patients with suspected Fanconi anemia, Advances in Clinical and Experimental Medicine, Uniwersytet Medyczny im. Piastów Śląskich we Wrocławiu, vol. 33, nr 4, 2024, s. 361-368, DOI:10.17219/acem/168825, łączna liczba autorów: 12	70	1,9
2023			
Artykuły z czasopism			
Abstrakt (praca oryginalna opublikowana w streszczeniach)			
1	Król Agata, Kokotkiewicz Adam, Królicka Aleksandra, Badura Anna, Lorenc Andżelika, Marzec-Wróblewska Urszula, Buciniński Adam, Łuczkiwicz Maria: Hairy and excised roots of <i>Salvia apiana</i> as <i>in vitro</i> systems for rosmarinic acid production, European Journal of Translational and Clinical Medicine, Gdański Uniwersytet Medyczny, vol. 6, nr suppl. 4, 2023, s. 162-162	—	—
Artykuł naukowy			
1	Lorenc Andżelika, Romaszko-Wojtowicz Anna, Jaśkiewicz Łukasz, Doboszyńska Anna, Buciniński Adam: Exploring the efficacy of artificial neural networks in predicting lung cancer recurrence: a retrospective study based on patient records, Translational Lung Cancer Research, Society for Translational Medicine (STM), vol. 12, nr 10, 2023, s. 2083-2097, DOI:10.21037/tlcr-23-350	140	4
2022			
Artykuły z czasopism			
1	Romaszko-Wojtowicz Anna, Lorenc Andżelika, Buciniński Adam, Doboszyńska Anna: Effects of tobacco smoking on the survivability of patients with multiple cancers and single lung cancer, International Journal of Environmental Research and Public Health, vol. 19, nr 15, 2022, Numer artykułu: 9179, s. 1-11, DOI:10.3390/ijerph19159179	140	—
2021			
Artykuły z czasopism			

1	Kokotkiewicz A., Badura Anna, Tabaczyńska Żaneta, Lorenc Andżelika, Buciński Adam, Luczkiewicz M.: Optimization of distillation conditions for improved recovery of phthalides from celery (<i>Apium graveolens</i> L.) seeds, Polish Journal of Food and Nutrition Sciences, vol. 71, nr 2, 2021, s. 197-210, DOI:10.31883/pjfn/137612	100	2,736
2	Lorenc Andżelika, Mendes Bárbara B., Conniot João, Sousa Diana P., Conde João, Rodrigues Tiago: Machine learning for next-generation nanotechnology in healthcare, Matter, vol. 4, nr 10, 2021, s. 3078-3080, DOI:10.1016/j.matt.2021.09.014	20	19,967
2019			
Artykuły z czasopism			
1	Badura Anna, Marzec-Wróblewska Urszula, Kamiński Piotr, Łakota P., Ludwikowski Grzegorz, Szymański Marek, Wasilow Karolina, Lorenc Andżelika, Buciński Adam: Prediction of semen quality using artificial neural network, Journal of Applied Biomedicine, vol. 17, nr 3, 2019, s. 167-174, DOI:10.32725/jab.2019.015	40	1,7
poster			
1	Kokotkiewicz A., Marzec-Wróblewska Urszula, Badura Anna, Tabaczyńska Żaneta, Lorenc Andżelika, Buciński Adam, Luczkiewicz M.: Hydrodistillation parameters affect phthalide content in apium graveolens essential oil, 73 s., 2019, 16 th International Interdisciplinary Meeting on Bioanalysis, poster	—	—

Kierownik
Sekcji Bibliografii
i Analiz Bibliometrycznych

mgr Małgorzata Białobłocka