WPROWADZENIE

Rosnąca ilość i złożoność danych wymaga nowych i elastycznych podejść do ich wydobywania. Tradycyjna metoda ręcznej analizy danych stała się obecnie nieefektywna, a analiza komputerowa stała się konieczna. Metody statystyczne, systemy eksperckie, rozmyte sieci neuronowe i algorytmy uczenia maszynowego (ML) są szeroko badane i stosowane w eksploracji danych. Lekarstwa odgrywają bardzo istotną rolę w życiu każdego człowieka. Wyprodukowanie jakiegokolwiek nowego leku do leczenia choroby i rozpowszechnienie go w interesie publicznym przy użyciu konwencjonalnych metod jest bardzo długotrwałe/męczące i bardzo nieopłacalne. Kilka molekuł łączy się w nowy lek, dlatego konieczne jest znalezienie struktury molekuł, co można przeprowadzić poprzez eksplorację danych. Eksploracja danych oznacza znajdowanie nowych informacji z wielu danych biologicznych/strukturalnych, które zostały wcześniej przekazane do systemu, a uzyskane informacje są nowe i użyteczne. Lek jest środkiem chemicznym używanym głównie do rozpoznawania/identyfikacji lub leczenia choroby/choroby. Choroby są rozpoznawane na poziomie subatomowym zwanym cząsteczką docelową. Każda choroba jest leczona poprzez równoważenie oddziaływania cząsteczki docelowej z cząsteczką aktywnego leku. W przypadku różnych chorób istnieją różne cele, a także wiele aktywnych związków leczniczych. Tak więc bardzo konieczne staje się zaklasyfikowanie lub zidentyfikowanie aktywnych związków chemicznych z grupy dużych informacji/statystyk biochemicznych, która obejmuje związki aktywne, nieaktywne i niejednoznaczne. W praktyce projektowanie nowego leku obejmuje takie procesy, jak badanie przesiewowe związków chemicznych, a także kategoryzowanie tych związków, co jest bardzo nieekonomiczne pod względem kosztów (średni koszt to ok. 1 mld USD) i czasu (około 10–15 lat). Dlatego projektowanie leków to najważniejsza i najbardziej wymagająca praca w badaniach klinicznych. Eksploracja danych oznacza badanie istniejących wcześniej dużych baz danych w celu odkrycia/wydobycia czegoś, co jest dla nas nowe i przydatne. Podobnie jak eksploracja danych, eksploracja molekularna obejmuje odkrywanie nowych cząsteczek na podstawie dużych zbiorów danych biologicznych. Odkryte nowe cząsteczki mogą skuteczniej docierać do celu, a tym samym mają lepszą skuteczność terapeutyczną. Większość zadań eksploracji danych w bioinformatyce obejmuje skanowanie ogromnych grup cząsteczek w celu odkrycia pewnej jednorodności między cząsteczkami określonej klasy. Przykładem tego samego jest odkrycie leku, w którym naukowiec chce odkryć nowego kandydata na lek w oparciu o rozpoznawcze potwierdzenie aktywności przeciwko pewnej chorobie, zebrane przez skanowanie kilku tysięcy cząsteczek. Współczesny nacisk kładzie się na prognozy sukcesu syntezy chemicznej, której głównym celem jest znalezienie cech molekularnych, które utrudniają/utrudniają pożądaną reakcję.

DLACZEGO GÓRNICTWO MOLEKULARNE?

Wyprodukowanie jakiegokolwiek nowego leku do leczenia konkretnej choroby i wprowadzenie go do publicznego zainteresowania przy użyciu konwencjonalnych metod jest bardzo długotrwałe/męczące i nieopłacalne. Kilka molekuł łączy się ze sobą, tworząc nowy lek, dlatego konieczne jest znalezienie struktury molekuł, co można przeprowadzić poprzez eksplorację danych. Eksploracja danych oznacza znajdowanie nowych informacji z wielu danych biologicznych/strukturalnych, które zostały wcześniej przekazane do systemu, a uzyskane informacje są nowe i użyteczne.

NARZĘDZIA ZAANGAŻOWANE W DATA MINING

Eksploracja danych obejmuje trzy podstawowe narzędzia:

1. Statystyki

2. Sztuczna inteligencja (AI)

3. Uczenie maszynowe

Statystyka oznacza duży, istniejący wcześniej zbiór danych biologicznych o liczbie n struktur molekularnych, z których odkrywane są nowe i przydatne dane. Innymi słowy, możemy powiedzieć, że jest to rodzaj analizy matematycznej, która polega na wykorzystaniu modeli kwantyfikowanych dla dowolnego zestawu danych. Statystyka wykorzystuje metodologie do wydobywania, analizowania, rewizji i wyciągania wniosków z podanych danych. W informatyce sztuczna inteligencja jest powszechnie znana jako wgląd maszynowy, tj. wiedza prezentowana przez maszyny, a nie normalny wgląd pokazywany przez ludzi. Można go również zdefiniować jako zdolność zaawansowanego komputera PC lub robota sterowanego przez komputer do wykonywania zadań ogólnie związanych z inteligentnymi istotami. Uczenie maszynowe to techniczne uczenie się algorytmów i reprezentacji statystycznej, które komputer wykorzystuje do wykonania określonej pracy bez użycia poleceń bezpośrednich/wyraźnych, w zależności od modeli i wniosków. Jest postrzegany jako pododdział AI.

NAUKA O DANYCH

Zanim przejdziemy do szczegółów dotyczących sztucznej inteligencji, zapoznajmy się najpierw z nauką o danych. Data science to szerokie pojęcie, które obejmuje wszystkie aspekty przetwarzania danych i to nie tylko pod kątem analitycznym, ale również algorytmicznym. Obejmuje następujące aspekty:

1. Wizualizacja danych

Jest to wysiłek/środek umieszczenia danych w wizualizowanym formacie w celu lepszego zrozumienia.

2. Integracja danych

Oznacza to zbieranie danych z różnych źródeł, a następnie łączenie ich w jednym formacie. Integracja odbywa się za pomocą następujących procesów, takich jak czyszczenie, mapowanie i transformacja.

3. Architektura rozproszona

Obejmuje to modele, zasady, reguły i standardy, które pomagają w zarządzaniu rodzajem danych, które są grupowane, w jaki sposób są gromadzone, prezentowane, łączone i ustawiane do wykorzystania w systemach informatycznych i organizacjach.

4. Decyzje oparte na danych

Jest to metoda zarządzania biznesem, która ocenia wybory, które można wycofać za pomocą możliwych do udowodnienia statystyk.

5. Automatyzacja za pomocą ML

Jest to podstawowy sposób na zautomatyzowanie danych przez podejście ML.

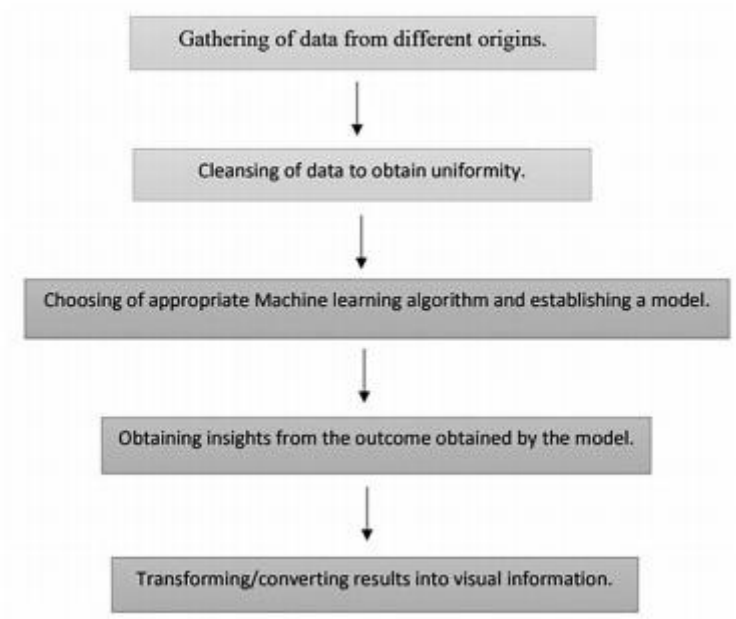
6. Inżynieria danych

Ten aspekt nauk o danych koncentruje się głównie na praktycznych zastosowaniach gromadzenia i analizy danych.

UCZENIE MASZYNOWE

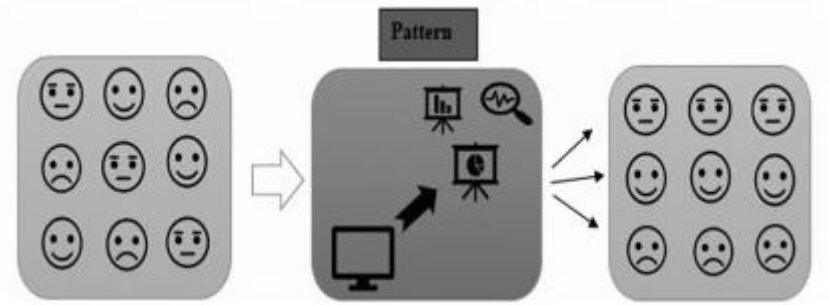
Uczenie maszynowe to część sztucznej inteligencji, która kształci/szkoli program komputerowy lub oblicza zdolność do naturalnego zdobywania biegłości za pomocą zadań i konsekwentnego rozwoju. W większości skupia się na ulepszaniu programów komputerowych, które mogą uczyć się i uzyskiwać informacje dla siebie. Deweloperzy muszą kodować i sprawdzać ostrożnie, zgodnie z potrzebą, w celu,

aby platforma mogła autonomicznie wykonywać ulepszenia wizyt. Proces ML przedstawiono na rysunku 1.

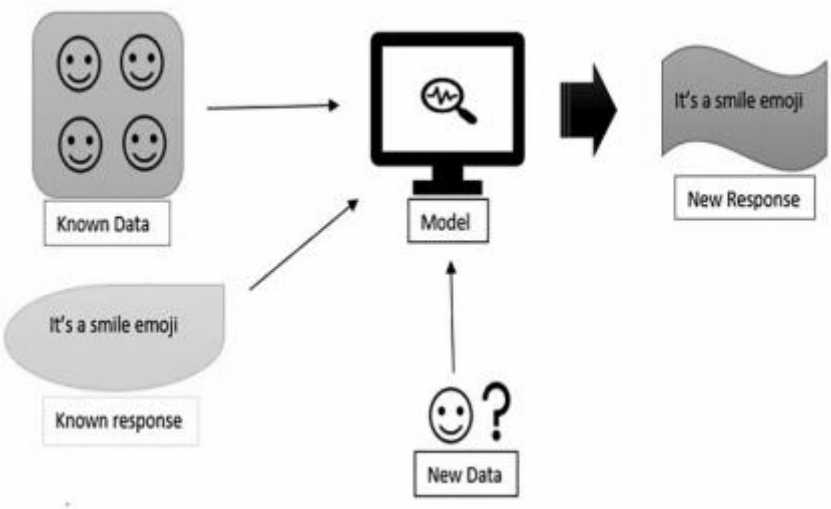


Istnieją trzy rodzaje ML:

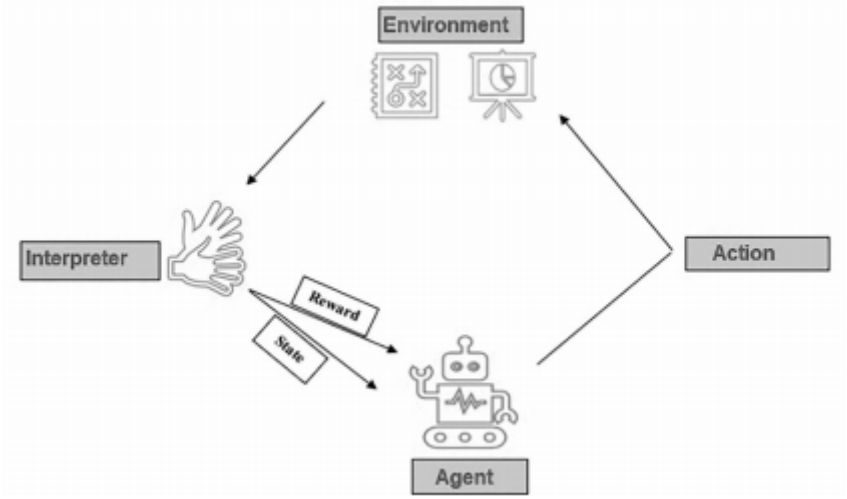
• Nauka nienadzorowana (rysunek 2)



• Uczenie nadzorowane (rysunek 3)



• Uczenie się przez wzmacnianie (rysunek 4)



TECHNIKI ML

Dostępne są różne narzędzia ML do klasyfikacji różnych zestawów danych. Narzędzia ML, które znajdują rolę w projektowaniu leków farmaceutycznych, to sieć neuronowa, k-NN (k-najbliższy sąsiad), maszyna wektorów nośnych (SVM), Naive Bayes (NB) i drzewo decyzyjne (DT). Najlepsze narzędzie ML, które można dalej wykorzystać w projektowaniu leków, można znaleźć, stosując różne systemy macierzy do tych narzędzi ML, a wyniki można interpretować i porównywać. Sztuczna sieć neuronowa (ANN), NB, SVM, k-NN, DT, regresja logistyczna (LR), liniowa analiza dyskryminacyjna (LDA) i Random Forest (RF) mają różne algorytmy, na których działają. Każde pojedyncze narzędzie ML lub jednocześnie dwa narzędzia mogą być trenowane poprzez podanie zestawu danych. W naszym wyszkolonym programie stosowane są różne systemy macierzy, takie jak zamieszanie, dokładność, przypominanie, wynik F-1 i precyzja, aby uzyskać pożądane wyniki. Porównując wyniki dwóch równoczesnych algorytmów, można łatwo zidentyfikować, które narzędzie ML jest lepszym klasyfikatorem, a który najlepiej pasuje do tego badania.

PODEJŚCIA UCZENIA MASZYNOWEGO DO WYDOBYWANIA CZĄSTECZEK

Geerestein użył DT do rozróżnienia między potencjalnymi lekami a nielekami. Do uczenia modelu wykorzystano związki pobrane z dostępnych indeksów chemicznych i światowych baz danych indeksów leków. Stwierdzono, że miara niedokładności w niezależnym zestawie danych walidacyjnych wyniosła 17,4%. Przewidywanie modelu można również wykorzystać do eskortowania pozyskiwania lub ekstrakcji związków do skanowania biologicznego lub tworzenia bibliotek kombinatorycznych. Christian Borgelt wykorzystał algorytm do odkrywania fragmentów w grupie cząsteczek, które pomagają w rozróżnieniu między odmiennymi klasami, na przykład aktywności, która dodatkowo pomogłaby w dziedzinie odkrywania leków. Autor wykorzystał zestaw danych, który jest publicznie dostępny w National Cancer Institute, zestaw danych DTP AIDS Antiviral Screen. W tym skanowaniu autor wykorzystał analizę formazanu do oszacowania zachowania ludzkich komórek CEM z HIV-1. Związki mające 50% konserwację/ochronę przed komórkami CEM przebadano ponownie i zarejestrowano jako umiarkowanie aktywne (CM). Związki zapewniające 100% konserwację/ochronę zostały zarejestrowane jako potwierdzone jako aktywne (CA). A związki, które nie spełniają żadnego z tych kryteriów, zostały zarejestrowane jako potwierdzone jako nieaktywne (CI). W sumie przetestowano 37 171 związków i stwierdzono, że 325 związków należy do CA, 877 związków należy

do CM, a 35 969 związków należy do CI. Serra i Thompson wygenerowali modele klasyfikacyjne do przewidywania wyniku cytogenetycznego in vitro dla grupy 383 związków organicznych. Wykorzystano dwie techniki k-NN i SVM. Wartości zaczerpnięto z testu przeprowadzonego na komórkach płuc chomika , która obejmuje zarówno ekspozycję 24-, jak i 48-godzinną. Do zakodowania topologicznych, elektronicznych, geometrycznych lub biegunowych cech powierzchni struktury użyto różnych deskryptorów. Ostateczny procent powodzenia kategoryzacji dla klasyfikatora k-NN złożonego tylko z 6 topologicznych stwierdzono, że deskryptory wynoszą 81,2% dla zestawu uczącego i 86,5% za zestaw testowy. Stwierdzono, że całkowity procent powodzenia kategoryzacji dla modelu SVM z trzema deskryptorami wyniósł 99,7% dla zestawu uczącego, 92,1% dla zestawu z walidacją krzyżową i 83,8% dla zestawu testowego. Evgeny Byvatov porównał systemy SVM i ANN do klasyfikacji leków/nieleków. Autor zastosował zarówno systemy do klasyfikowania zbioru danych o związkach na lekowe i nielekowe, jak i do filtrowania potencjalnie niechcianych cząsteczek jakobiblioteka złożona. W sumie wzięto pod uwagę 9208 cząsteczek (4998 leków i 4210 cząsteczek nielekowych). Wskaźniki wykonania obu klasyfikatorów porównano przy użyciu różnych deskryptorów - 120 standardowych deskryptorów fragmentów Ghose-Crippen oraz szerokiej gamy 180 różnych deskryptorów właściwości i fizykochemicznych z pakietu Molecular Operating Environment (MOE) oraz 225 topologicznych deskryptorów farmakoforów (CATS). Do walidacji krzyżowej uwzględniono łącznie 525 deskryptorów i stwierdzono, że wyniki uzyskane przez SVM były dokładniejsze (82% poprawnych przewidywań), a ANN wynosiło 80%. Andreas Bender (2004) wprowadził nowy sposób wyszukiwania podobieństw. W tej technice molekuły były nazywane środowiskiem atomowym, które następnie wprowadzano do systemu opartego na zdobywaniu informacji. Następnie do klasyfikacji związków stosuje się klasyfikator NB. Algorytm przewyższa wszystkie aktualne techniki odzyskiwania oceniane tutaj przy użyciu dwu- i trójwymiarowych deskryptorów. Ta technika może być również wykorzystana do rozpoznawania grup funkcyjnych w aktywnych cząsteczkach i jest skuteczna obliczeniowo. Gongde i Neagu (2005) zaproponowali energiczną technikę/procedurę, rozmyty model k-NN, do oceny toksyczności związków chemicznych. Metoda była zasadniczo zależna od metody nadzorowanego grupowania, znanej jako model k-NN, która działa na zasadzie rozmytej segregacji/klastrowania zamiast ostrej segregacji/klastrowania. Wyniki eksploracyjne rozmytego modelu k-NN nadzorowanego na 13 publicznych zestawach danych z repozytorium uczenia maszynowego UCI i siedmiu zestawach danych dotyczących toksyczności z rzeczywistych aplikacji skontrastowano z wynikami rozmytego klastrowania c-średnich, k-średnich klastrów, k- NN, rozmyte k-NN pod względem wykonania na podstawie klasyfikacji. Wyniki pokazały, że rozmyty model k-NN był najlepszą techniką przewidywania toksyczności związków chemicznych.

Shubhangi i Hiremath wykorzystali SVM z siecią neuronową do rozpoznawania znaków pisanych odręcznie. Technika rozpoznawania znaków to drukowany obraz/dokument lub rozpoznawanie znaków odręcznych. Wykazali współpracę łączenia klasyfikatorów SVM z siecią neuronową przy użyciu zestawu cech morfologicznych. Zaproponowano podejście, które można wykorzystać do wykrywania i rozpoznawania obu kategorii znaków (cyfrowych i pisanych odręcznie). Stwierdzono, że klasyfikatorzy poprawnie zaklasyfikowali aż do 97%. Mandal i Jana (2013) przeprowadzili względne badanie algorytmu NB i k-NN dla wieloklasowej kategoryzacji cząsteczek leków. Eksplorację algorytmów NB i k-NN przeprowadzono na biochemicznym zapisie danych pobranych z PubChem. Zestaw danych (tj. ocena qHTS w celu sprawdzenia autofluorescencji związku przy 460 nm (szary) w komórkach HEK293) został wzięty pod uwagę w analizie, która obejmuje łącznie 1280 rekordów, z których każdy jest powiązany z aktywnym, nieaktywnym lub niejednoznacznym Grupa. Ze zbioru danych 20% danych zostało zachowanych do celów badawczych, a reszta (80%) została wykorzystana do uczenia modelu. Zestaw danych testowych obejmował łącznie 256 związków, z których 14 było aktywnych, 54 były niejednoznaczne, a pozostałe związki były nieaktywne. Klasyfikator NB błędnie skategoryzował tylko

cztery związki (jeden związek oczekiwano jako aktywny, ale w rzeczywistości był nieaktywny, a trzy związki oczekiwano jako aktywne, ale w rzeczywistości były one niejednoznaczne). Klasyfikator k-NN błędnie sklasyfikował tylko jeden związek (oczekiwany jako niejednoznaczny, ale w rzeczywistości był aktywny). Stwierdzono, że dokładność klasyfikatora NB wyniosła 93%, a klasyfikatora k-NN 99,6%. Ioannis wykorzystał ML i metody eksploracji danych w trwających badaniach nad cukrzycą w celu przekształcenia wszystkich inteligentnie dostępnych danych w cenną wiedzę. Gruntowne badania we wszystkich aspektach związanych/powiązanych z cukrzycą (takich jak diagnoza, etiopatofizjologia, leczenie itp.) doprowadziły do zebrania ogromnych ilości danych. Głównym celem tego badania była analiza wszystkich technik ML i eksploracji danych w zakresie eksperymentów z cukrzycą pod kątem wszystkich powyższych aspektów. Zastosowali szeroką gamę algorytmów ML. Głównie m.in. 85% charakteryzowało się uczeniem nadzorowanym, a pozostałe (15%) uczeniem się nienadzorowanym, czyli według zasad asocjacyjnych. Odkryli, że SVM był najbardziej udanym i powszechnie stosowanym algorytmem. Lei Zhang opracował wspomaganą komputerowo strategię planowania/przesiewania atomów dla struktury i badania cząsteczek zapachowych. Autor zastosował metodę przewidywania istoty cząstek przy użyciu techniki ML, a do przewidywania właściwości fizycznych, takich jak parametr rozpuszczalności, prężność pary i lepkość, zastosowano mechanizm udziału grupowego. Stworzono model MILP/MINLP do przesiewania/przedstawiania cząstek zapachowych. Wyniki pokazały, że nawet drobne zmiany w strukturze molekularnej mogą skutkować wyjątkowo różnymi aromatami. W związku z tym należy opracować obraz molekularny mający większe zestawy ram strukturalnych, aby ustanowić model, który może dokładniej przewidywać. Stwierdzono również, że cząsteczki o jednorodnym spektrum drgań mają jednorodne właściwości zapachowe. Konstantinos Vougas opisał nową procedurę skanowania in silico zależną zasadniczo od eksploracji reguł asocjacyjnych, rozpoznawania genów jako indywidualnych operatorów reakcji na leki i skontrastowania ich z odpowiednimi technikami eksploracji danych. Autorka opracowała komputerowy model przewidywania wyników biologicznych obejmujący głównie trzy etapy: pobranie zbioru danych, wybór odpowiedniego algorytmu i polecenie opracowania modelu przewidywania oraz przetestowanie go w nowych zbiorach danych.

PROCEDURA

Przygotowano model do porównania różnych narzędzi ML do klasyfikacji związków. Do analizy praktycznej opisy laptopów to Intel Core i5 7. generacji z 8 GB pamięci o dostępie swobodnym (RAM) z systemem operacyjnym Ubuntu (18.04.3 LTS). „Python” (wersja 3.7.3) był używany jako język programowania do uczenia i testowania różnych modeli. Głównym tematem tego modelu była klasyfikacja/kategoryzacja związków chemicznych na aktywne, nieaktywne i niejednoznaczne. Porównanie przeprowadzono między k-NN, SVM, DT, LR, LDA, RF i NB. Zastosowano wszystkie standardowe algorytmy z wyjątkiem NB (zastosowano NB Gaussa). Systemy metryk, które są tutaj używane do analizy wydajności, to precyzja, przypomnienie i wynik F1. Zestaw danych dotyczących związku chemicznego zaczerpnięto z PubChem (AID 720678). Ten zapis testu biologicznego (AID 720678) jest powiązany z łącznie 98 dodatkowymi zapisami testu biologicznego w PubChem, który obejmuje kilka projektów testów (https://pubchem.ncbi.nlm.nih.gov/bioassay/720678). W zbiorze danych znajdowało się łącznie 10486 związków. Spośród nich 10 374 było nieaktywnych, 40 aktywnych, a 72 niejednoznacznych. Zastosowano format dzielony (70/30); tj. 70% całego zestawu danych zostało wykorzystanych do uczenia modelu, a pozostałe 30% do celów testowych. Biblioteka ML sklearn została użyta w Pythonie.

WNIOSEK

Szkolenie i testowanie różnych technik ML przeprowadzono przy użyciu wspomnianego zestawu danych, a ich wyniki porównano przy użyciu trzech systemów metryk i stwierdzono, że DT jest

najskuteczniejszą techniką z 100% dokładnymi wynikami, a SVM i LR są najmniej skuteczne techniki z zaledwie 33% dokładnymi wynikami. Poza tymi technikami, LDA i RF zapewniają również obiecujące wyniki.