Wydział Automatyki, Elektroniki i Informatyki
Katedra Inżynierii i Analizy Eksploracyjnej Danych

Prof. dr hab. inż.
Joanna Polańska
Kierownik Katedry

Recenzja rozprawy doktorskiej

Autor: Mgr Krzysztof Koras

Tytuł: Computational methods for anti-cancer drug sensitivity prediction

Promotor: dr hab. Ewa Szczurek

Dziedzina: Nauki ścisłe i przyrodnicze

Dyscyplina: Informatyka

Niniejsza recenzja została przygotowana na zlecenie Rady Naukowej Dyscyplin Matematyka i Informatyka Uniwersytetu Warszawskiego, zgodnie z pismem z dnia 28.10.2022 roku.

Ogólna charakterystyka rozprawy i ocena wyboru tematyki rozprawy

Rozprawa doktorska mgr. Krzysztofa Korasa jest napisana w języku angielskim. Tekst rozprawy liczy 108 stron. Składa się z siedmiu rozdziałów, suplementów oraz spisu literatury liczącego 219 pozycji.

Tematyką rozprawy jest badanie i rozwijanie metod bioinformatycznych oraz technik uczenia maszynowego do przewidywania *in silico* oddziaływania leków przeciwnowotworowych. W aspekcie bardzo intensywnego rozwoju prac naukowych i klinicznych nad terapiami antynowotworowymi tematyka recenzowanego doktoratu mieści się w bardzo aktywnym i szerokim polu badawczym. Jednym z wiodących obecnie problemów jest personalizacja terapii, to znaczy dobranie leku (cytostatyku, przeciwciała monoklonalnego lub innej substancji) do odpowiednio zmierzonej lub ocenionej charakterystyki populacji komórek nowotworowych u pacjenta. W literaturze ukazuje się bardzo wiele prac poświęconych personalizacji terapii antynowotworowych. Oczywiście uzyskiwane

Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki
ul. Akademicka 16, 44-100 Gliwice
+48 32 237 2144 / +48 601 891 245
joanna.polanska@polsl.pl

postępy są w bardzo dużym stopniu napędzane przez nowe techniki eksperymentalne biologii molekularnej, biochemii, chemii leków. Jednak w obliczu złożoności zjawisk i bardzo dużych wolumenów danych analitycznych istotną rolę we wspieraniu postępów w onkologii personalizowanej grają obliczenia komputerowe oraz nowe metody modelowania matematycznego. W tym dokładnie obszarze lokują się badania naukowe przeprowadzone przez Doktoranta i przedstawione w pracy. Doktorant zaproponował nowe metody badawcze i uzyskał oryginalne wyniki naukowe w zakresie obliczeniowej onkologii personalizowanej. Oryginalność jego podejścia polega na odpowiednim doborze i zastosowaniu metod uczenia maszynowego i sztucznej inteligencji do przewidywania oddziaływania leków na populacje komórek nowotworowych u pacjentów, budowania systemów rekomendacyjnych dla pewnych typów leków antynowotworowych, a także do rozszerzenia systemu rekomendacyjnego „proponowania" leków antynowotworowych o funkcje generatywne (generacyjne). Przeprowadzone badania naukowe są zrealizowane na bazie dostępnych internetowo danych eksperymentalnych oddziaływania związków chemicznych na linie komórek nowotworowych.

Biorąc pod uwagę powyższe, tematykę badań podjętą w przedstawionej rozprawie doktorskiej uważam za niezwykle istotną i bez wątpienia wpisującą się w obszar informatyki, ze szczególnym wskazaniem na jej zastosowania w obszarze biologii obliczeniowej.

Omówienie rozprawy

Poniżej pokrótce przedstawię podsumowanie poszczególnych części rozprawy doktorskiej. Praca jest bardzo ściśle związana z trzema oryginalnymi artykułami naukowymi, których pierwszym autorem jest Doktorant. Jednak kompozycja pracy, scharakteryzowana poniżej, jest szersza i ogólniejsza niż samo tylko przedstawienie wyników tych trzech artykułów naukowych.

W pierwszym rozdziale pracy, jako wstęp przedstawia się motywację do przeprowadzonych badań naukowych, wyzwania jakie stoją przed obliczeniową onkologią personalizowaną. W scenariuszu zastosowania metod modelowania matematycznego do przewidywania efektów leków antynowotworowych Doktorant słusznie umieszcza problem selekcji cech. Przez wielką liczbę

dostępnych danych pomiarowych przy znacznie mniejszej liczbie eksperymentów walidacyjnych problem selekcji cech (i wstępnej redukcji wymiarowości) jest dużym wyzwaniem, a wybór metody silnie zależy od charakteru danych i postawionego zadania i często decyduje o jakości opracowanego wyniku. Następnie Doktorant charakteryzuje zagadnienia związane z zastosowaniem sztucznej inteligencji w onkologii obliczeniowej, omawia problemy uczenia wielokryterialnego oraz przedstawia niezwykle istotny problem interpretowalności wyników. Opisuje też problem definiowania transformacji danych oraz omawia pokrótce zadanie modelowania generatywnego. Następnie stara się uzasadnić związek pomiędzy wynikami naukowymi przedstawionymi w artykułach naukowych powiązanych z pracą z opisywanymi problemami, których rozwiązanie, lub badania w kierunku rozwiązania, jest bardzo istotne dla postępu terapii nowotworowych.

W drugim rozdziale pracy Doktorant przedstawia pewne fakty związane z biologicznym tłem rozwoju nowotworów, a także ich leczeniem i planowaniem leczenia. Omawia znane tło biologiczne inicjacji i progresji procesu nowotworzenia. Przedstawia rolę mutacji kierunkowych w nowotworach, klonalność guzów nowotworowych oraz oddziaływanie populacji komórek nowotworowych z układem odpornościowym. Następnie przedstawia fakty związane z klasyfikacją leków i terapii antynowotworowych. Wyodrębnia i charakteryzuje rodzaje farmakoterapii, chemioterapię, terapię hormonalną, immunoterapię, terapię kierowaną i wymienia związane z nimi substancje farmakologiczne. Przedstawia także źródła danych o lekach antynowotworowych. Kolejną częścią drugiego rozdziału są opisy, odnośniki literaturowe i charakterystyka danych eksperymentalnych związanych z badaniem wrażliwości komórek nowotworowych na leki. Wymienia źródła danych fakrmakogenomicznych, w szczególności bazę danych GDSC (Genomics of Drug Sensitivity in Cancer) wykorzystywaną w swoich publikacjach związanych z pracą doktorską. Dalej charakteryzuje szerzej techniki eksperymentalne związane z farmakogenomiką. Przedstawia dane eksperymentalne oraz ich repozytoria internetowe, związane z genomiką nowotworów, pochodzące z eksperymentów sekwencjonowania komórek/tkanek nowotworowych. W kolejnym podpunkcie rozdziału opisywane są techniki transkryptomiczne, pomiarów poziomów ekspresji genów oraz ich repozytoria. Ostatnim elementem zawartym w rozdziale jest opis metod charakteryzowania związków chemicznych, które

mają zdolność oddziaływania na komórki/tkanki nowotworowe. Doktorant przedstawia dwa główne podejścia, przez charakterystykę celów oddziaływania leków oraz przez odwzorowania (zwykle uproszczone) ich struktury chemicznej. Doktorant wspomina o systemie SMILES kodowania związków chemicznych.

W rozdziale trzecim Doktorant przedstawia techniki informatyczne uczenia maszynowego związane z badaniami przeprowadzonymi w pracy. Przedstawia najpierw podstawowe techniki uczenia maszynowego, następnie zasady uczenia głębokiego. Omawia technikę i metody konstrukcji modeli generatywnych, to znaczy takich, które na bazie wcześniej przedstawionych w procesie uczenia przykładów posiadają zdolność do generowania nowych. Istotną cechą tych modeli jest istnienie warstwy ukrytej, której modyfikacje prowadzą do generowania nowych danych. W tym kontekście omawia ideę budowy autoenkoderów wariacyjnych. Przedstawia także ideę zmiennych ukrytych oraz jej ilustrację i zastosowanie do budowy modeli mieszanin rozkładów, w szczególności do budowy mieszanin rozkładów normalnych. W ostatnim podpunkcie rozdziału trzeciego Doktorant rozważa interesującą kwestię, formułując odpowiedź na pytanie: *jaką specyfikę posiada zastosowanie sztucznej inteligencji w problemach przewidywania oddziaływania leków antynowotworowych w stosunku do ogólnej metodologii rozwijania aplikacji związanych ze sztuczną inteligencją*. Przedstawia pewne koncepcje w tym zakresie, związane głównie z scenariuszami analiz danych charakteryzujących leki, danych charakteryzujących tkanki/komórki nowotworowe, w powiązaniu z pomiarami odpowiedzi nowotworów na terapie.

W rozdziale czwartym przedstawione są wyniki zawarte w publikacji, „Koras K., Juraeva D., Kreis J., Mazur J., Staub E. & Szczurek E. *Feature selection strategies for drug sensitivity prediction*. Scientific Reports, 2020, 10(1):9377", której pierwszym autorem jest Doktorant. Modele obliczeniowe opracowane w tym artykule bazują na danych eksperymentalnych dostępnych w repozytorium GDSC (Genomics of Drug Sensitivity in Cancer) pochodzącym z publikacji zamieszczonej w czasopiśmie Nucleic Acid Research w 2012 roku. Praca ta udostępnia wyniki doświadczalne dotyczące 138 leków antynowotworowych, 700 linii komórkowych nowotworów związanych łącznie z około 75 000

eksperymentów. Publikacja, której współautorem jest Doktorant, stawia sobie za cel jak najszerszą ocenę strategii selekcji cech w aspekcie ich użyteczności w zadaniu nadzorowanej klasyfikacji związków chemicznych w ich oddziaływaniu na komórki nowotworowe. Jako miarę oddziaływania leku na komórki nowotworowe przyjęto wskaźnik AUC, pole pod krzywą „dawka – odpowiedź". Cały zbiór cech, obejmujący około 20 000 cech podzielono na kilka kategorii, cechy zdefiniowane przez cele terapeutyczne leków (typy tkanek, specyficzne ścieżki genowe), sygnatury ekspresji genów, sygnatury genomowe. Wybierane zbiory cech kombinowano z dwoma rodzajami klasyfikatorów nadzorowanych, klasyfikatorem lasów losowych oraz nieliniowym klasyfikatorem nazywanym siecią elastyczną. Dla różnych kategorii cech opisujących leki oraz tkanki nowotworowe porównuje się rzeczywiste oraz przewidywane oddziaływanie leku na komórki nowotworowe. Wprowadza się wskaźnik RelRMSE, względny spierwiastkowany wskaźnik sumy kwadratów, służący do skompensowania różnic wynikających ze stosowania cech należących do różnych klas. Obszerne eksperymenty numeryczne prowadzą do sformułowania interesujących wniosków dotyczących potencjału bardzo szerokiego spektrum cech farmakogenomicznych oraz biochemicznych w charakteryzowaniu i predykowaniu odpowiedzi populacji komórek nowotworowych na terapię środkami farmakologicznymi.

W rozdziale piątym przedstawiane są wyniki kolejnego artykułu, „Koras K., Kizling E., Juraeva D., Staub E. & Szczurek E. *Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines.* Scientific reports, 2021, 11(1):1-16", którego również pierwszym autorem jest Doktorant. Praca ta jest z jednej strony pogłębieniem wcześniejszych analiz, a z drugiej strony ogniskuje badania na węższej klasie środków farmaceutycznych, tzn. na inhibitorach kinaz. Jako cel tej pracy postawione jest zbudowanie systemu rekomendacyjnego dla inhibitorów kinaz w aspekcie ich oddziaływania na komórki nowotworowe. Praca bazuje na dwóch źródłach danych eksperymentalnych, na bazie danych GDSC wykorzystywanej także w poprzednim artykule oraz na bazie danych KINOMEscan obejmującej wyniki dotyczące dokowania ok. 440 kinaz do różnych czynników farmaceutycznych. Model matematyczny opracowany przez autorów publikacji obejmował powiązanie danych dotyczących oddziaływania leków związanych z aktywnościami farmakogenomicznycznymi z bezpośrednio mierzonymi profilami inhibicji leków przez kinazy

udostępnianymi przez repozytorium KINOMEscan. Powiązanie to okazało się być bardzo skuteczne. Model matematyczny opracowany przez autorów nosi nazwę DEERS (Drug Eficacy Estimation Recommender System). Jego schemat jest przedstawiony na rysunku 1 artykułu. Zawiera głęboką sieć neuronową powiązaną z dwoma autoenkoderami, jeden z nich jest związany z oddziaływaniem leku a drugi z charakterem linii komórkowej. Metoda zaproponowana w pracy wykazuje wyższą skuteczność niż referencyjna metoda bazująca na macierzowych modelach faktoryzacyjnych.

W rozdziale szóstym, bazującym na publikacji, „Koras K., Możejko M., Szymczak P., Izdebski A., Staub E. & Szczurek, E. *A generative recommender system with GMM prior for cancer drug generation and sensitivity prediction.* In Proceedings of the 17th Machine Learning in Computational Biology meeting, PMLR 200:61-73, 2022", przedstawione jest dalsze rozwinięcie koncepcji z poprzednich dwóch prac. Także w tej pracy Doktorant jest pierwszym autorem. Praca rozwija poprzednie kierunki badań o postawienie i rozwiązanie problemu skonstruowania systemu rekomendacyjnego z funkcjonalnością generatywną dla oddziaływania środków farmaceutycznych na kinazy. Schemat zaproponowanego „generatywnego" autoenkodera przedstawiony jest na rysunku 1 artykułu. Przetwarzanie jest dość złożone i należy uznać za ciekawe osiągnięcie samo uzyskanie stabilności takiej struktury. Moduły wejściowe potoku przetwarzania danych zawierają formaty SMILES dla związków chemicznych, cechy transkryptomiczne i genomowe linii komórkowych, dane dotyczące profili inhibicji kinaz oraz dane eksperymentalne dawka – odpowiedź. Moduł wyjściowy z zaprojektowanej sieci autoenkodera zawiera predykcje odpowiedzi linii komórkowych na działanie podawanych leków. Interesującą i nowatorską konstrukcją jest zaprojektowanie warstwy ukrytej tej sieci w postaci rozkładu opisanego mieszaniną gęstości gaussowskich. Autorzy przedstawiają oryginalną metodę estymacji parametrów tych mieszanin. W pracy przeprowadzone są badania numeryczne w dwóch głównych kierunkach. Pierwszym jest zdolność zbudowanej struktury do przewidywania oddziaływania leków na komórki nowotworowe, a drugim jest funkcjonalność generatywna stworzonej sieci to znaczy jej zdolność do generowania nowych leków. Autorzy argumentują, że zastosowana mieszanina rozkładów normalnych jako warstwa ukryta zwiększa

precyzję przewidywania w działaniu sieci. W toku zrealizowanych prac obliczeniowych wykazują skuteczność i efektywność zaproponowanej metodologii.

Rozdział siódmy stanowi podsumowanie pracy. Ponadto dwa dodatki zawierają pewne szczegóły implementacyjne wykorzystywanych w artykułach procedur i algorytmów.

W trakcie lektury dysertacji nasunęło mi się kilka pytań, które przedstawiam poniżej:

1. Czym kierował się Doktorant wybierając wykorzystane w badaniach algorytmy uczenia maszynowego oraz selekcji cech?
2. Istnieją liczne bazy danych dotyczące oddziaływania substancji chemicznych na biomolekuły, np. BidingDB, https://www.bindingdb.org/rwd/bind/index.jsp. Czy i w jaki sposób Doktorant dokonywał przeglądu możliwych źródeł danych, które miałyby potencjał do wykorzystania w budowanych przez Doktoranta sieciach?
3. Czy Doktorant uważa, że możliwe jest zastosowanie metod sztucznej inteligencji do badania i planowania immunoterapii? Czy doktorant zna jakieś wyniki z tego zakresu?
4. Jak Doktorant wyobraża sobie dalszy rozwój badań w zakresie zastosowania sztucznej inteligencji w onkologii obliczeniowej?
5. Czy Doktorant uważa, że rozwijane metody znajdą w przyszłości zastosowania kliniczne?

Ocena rozprawy

Moja ogólna ocena rozprawy jest **pozytywna**. Praca bazuje na oryginalnych, współautorskich publikacjach naukowych. Doktorant jest pierwszym autorem wszystkich trzech publikacji. Ponadto, konstrukcja i kompozycja pracy, umieszczenie szeregu pomocniczych wyników i tłumaczeń, bardzo dobitnie wykazują bardzo dobrą orientację doktoranta w poruszanej problematyce oraz jego wiodącą rolę w badaniach. Należy podkreślić, że doktorat jest napisany ładnym językiem, jasnym i przejrzystym. Należy także docenić logikę konstrukcji wywodu w recenzowanej pracy. Pewien niedosyt budzi brak jawnie sformułowanych hipotez badawczych oraz celów poznawczych rozprawy. Widoczne jest

również bardzo duże podobieństwo tekstu dysertacji z fragmentami artykułów będących jej podstawą. Przyjmując, że jako pierwszy autor publikacji Doktorant miał wiodący udział w badaniach oraz opracowaniu tekstu, można uznać za akceptowalne.

Jak już wspomniano na początku recenzji, tematyka doktoratu lokuje się w bardzo aktywnym i szeroko rozpracowywanym obszarze badawczym. Zastosowania sztucznej inteligencji i uczenia maszynowego obejmują obecnie wszystkie kierunki badań naukowych w których udaje się zbudować odpowiednie formalizmy matematyczne. Onkologia obliczeniowa jest także bardzo intensywnie rozwijana przez wyniki z zakresu uczenia maszynowego. Jednak w tym intensywnie eksploatowanym obszarze badawczym Doktorantowi udaje się sformułować nowe i oryginalne podejścia. Tworzy konstrukcje głębokich sieci neuronowych, systemów rekomendacyjnych, funkcjonalności generatywnych, wyszukuje i opracowuje dla nich źródła danych treningowych i weryfikacyjnych. Przeprowadza wszechstronne badania obliczeniowe.

Bardzo wartościowa (rzadko spotykana w pracach bronionych na bazie publikacji) jest logiczna sekwencja zamieszczanych wyników, które zaczynają się od konstrukcji sieci, przechodzą do konstrukcji autoenkodera a następnie rozwijają funkcjonalność generacji nowych związków.

Doktorant wykazuje się kompetencjami i dojrzałością w swoim warsztacie naukowym w dyscyplinie informatyka. Posiada także zdolność badań i współpracy o charakterze interdyscyplinarnym.

Syntetyczna o cena rozprawy:

A. Czy rozprawa zawiera oryginalne rozwiązanie problemu naukowego? Zdecydowanie TAK

B. Czy kandydatka posiada ogólna wiedzę teoretyczną w dyscyplinie? Zdecydowanie TAK

C. Czy posiada umiejętność samodzielnego prowadzenia pracy naukowej? Zdecydowanie TAK

Konkluzja

Osiągnięcia i oryginalne elementy publikacji wchodzących w skład rozprawy, a także jakość całego tekstu rozprawy, są na pewno wystarczające do jej ogólnej pozytywnej oceny oraz spełniają zwyczajowe wymagania stawiane rozprawom doktorskim.

Stwierdzam zatem z pełnym przekonaniem, że opiniowana rozprawa Pana mgr Krzysztofa Korasa pt. „Computational methods for anti-cancer drug sensitivity prediction" zawiera samodzielne rozwiązanie ważnego i istotnego problemu naukowego, jednocześnie spełniając wszystkie wymagania przewidziane dla rozpraw doktorskich w aktualnie obowiązującej Ustawie o Tytule Naukowym i Stopniach Naukowych.

W związku z tym stawiam wniosek o dopuszczenie rozprawy doktorskiej do publicznej obrony.

Ponadto, biorąc pod uwagę liczne nowatorskie elementy w pracy, opublikowanie jej wyników w bardzo prestiżowych czasopismach naukowych, dojrzałość naukową Doktoranta i jego bardzo dobry warsztat badawczy, wnioskuję o wyróżnienie rozprawy.