Prof. dr hab. inż. Krzysztof Krawiec
Instytut Informatyki
Politechnika Poznańska
ul. Piotrowo 2
60-965 Poznań

Poznań, 25.07.2022

Recenzja rozprawy doktorskiej mgr inż. Dominiki Sułot p.t. "Development of machine learning algorithms operating on small datasets to assist in the diagnosis of eye diseases"

1. Tematyka rozprawy

Tematem rozprawy doktorskiej mgr inż. Dominiki Sułot są algorytmy inteligentnej analizy danych na potrzeby wspomagania diagnostyki obrazowej w okulistyce. W szczególności, doktorantka skupia się na rozwijaniu algorytmów uczących się z danych, w oparciu o szeroko rozumiane techniki uczenia maszynowego, w tym zarówno techniki tradycyjne, jak i te wpisujące się w paradygmat uczenia głębokiego (deep learning). Innym wyróżnikiem rozprawy jest dążenie do rozwijania metod zdolnych do skutecznego uczenia się z niewielkiej liczby przykładów.

Obszar tematyczny rozprawy plasuje się zatem w analizie obrazów, uczeniu maszynowym, a zatem pośrednio także w obszarze sztucznej inteligencji. Praca ma charakter stosowany, tj. proponowane algorytmy i architektury weryfikowane są na rzeczywistych danych medycznych. Taka charakterystyka rozprawy pozwala mi zaklasyfikować ją do dyscypliny naukowej Inżynieria Biomedyczna.

2. Ocena treści rozprawy i wkładu oryginalnego

2.1 Treść rozprawy

Przedłożona rozprawa przygotowana została w języku angielskim, ma objętość około 100 stron, a jej zasadnicza część składa się ze wstępu, czterech rozdziałów oraz bibliografii. Według

deklaracji Autorki, praca oparta jest głównie na trzech opublikowanych przez nią wcześniej artykułach.

Rozdział pierwszy stanowi wprowadzenie do sztucznej inteligencji, w szczególności do metod uczenia maszynowego. Wprowadzając uczenie maszynowe, Doktorantka zwięźle zaprezentowała jego główne metodyki i paradygmaty, w szczególności podział na uczenie nadzorowane, nienadzorowane i inne. Dalej mgr Sułot krótko charakteryzuje standardowe podejścia uczenia maszynowego, w tym między innymi proste modele neuronowe, klasyfikatory Bayesowskie, drzewa decyzyjne, metody minimalnoodległościowe oraz metodę wektorów wspierających. Następnie znajdziemy w pracy prezentację podstaw modeli głębokiego uczenia, w tym sieci splotowych. Rozdział zawiera także krótką prezentację podstawowych metryk skuteczności systemów klasyfikacyjnych. Począwszy od sekcji 1.2, Autorka koncentruje się na zasygnalizowanym w tytule pracy aspekcie małych danych, a dokładniej scenariuszach użycia w których liczba przykładów uczących jest niewielka, w tym także niewielka względem zawartości informacyjnej poszczególnych przykładów, wyrażonej na przykład liczbą cech (tzw. klątwa wymiarowości). Autorka prezentuje wybrane techniki które adresują ten problem, w tym między innymi techniki augmentacji danych, w szczególności augmentacji obrazów. Rozdział kończy się krótkim przeglądem zastosowań metod uczenia maszynowego w okulistyce. Przegląd ten podzielony jest na trzy wątki, które odpowiadają scenariuszem użycia rozważanym w dalszych częściach pracy, tj. klasyfikacji pod kątem diagnostyki medycznej, segmentacji struktur anatomicznych w obrazach, oraz selekcji cech.

Rozdziały 2 i 3 rozprawy stanowią jej główną część i prezentują całość oryginalnych przyczynków. Pierwszy z nich koncentruje się na metodyce eksperymentalnej oraz eksperymentach wstępnych; rozdział 3 z kolei prezentuje finalne wyniki ilościowe i przeprowadza ich dyskusję. W obu rozdziałach poszczególne sekcje odnoszą się do wyżej wymienionych trzech scenariuszy użycia, dlatego pozwolę sobie dalej charakteryzować tę część rozprawy wspólnie dla rozdziałów 2 i 3.

Dla **scenariusza użycia wspomagania diagnostyki jaskry** Autorka wykorzystuje obrazy pozyskane techniką scanning laser ophthalmology (SLO) i pracuje ze zbiorem 227 badań składającym się ze 105 przypadków pozytywnych i 122 przypadków kontrolnych. Doktorantka przedstawia najpierw próby wykorzystania tradycyjnych technik analizy obrazu na potrzeby tego problemu diagnostycznego, wykazując ich niewystarczającą skuteczność, co stanowi motywację dla wykorzystania technik uczenia maszynowego. Zestaw algorytmów uczenia maszynowego wykorzystanych w tym celu pokrywa się z tym prezentowanym w rozdziale pierwszym rozprawy. Wykorzystanie tych algorytmów pozwoliło na polepszenie skuteczności predykcyjnej w porównaniu z klasycznymi technikami analizy obrazu, jednak w stopniu nadal dalekim od satysfakcjonującego. W efekcie Autorka zaproponowała wykorzystanie splotowych sieci neuronowych w kilku wariantach, w szczególności architektur znanych już z poprzednich prac oraz pewnej architektury autorskiej. W celu dalszego polepszenia skuteczności predykcyjnej, mgr Sułot użyła także komitetu klasyfikatorów bazującego na podzbiorach otrzymanych w efekcie podziału kolekcji danych techniką walidacji krzyżowej (cross validation).

Wyniki ilościowe dla tego scenariusza użycia, zaprezentowane w rozdziale 3, ilustrują stopniową progresję jakości, wynikającą z wykorzystania różnych danych wejściowych (oryginalny obraz, grubość warstwy RNFL), typów klasyfikatorów i architektur sieci neuronowych. W ostatecznym rozrachunku najlepsze wyniki uzyskano głęboką siecią splotową o dedykowanej architekturze z jednoczesnym wykorzystaniem komitetu takich modeli, gdzie finalna decyzja podejmowana jest na podstawie głosowania.

W ramach **scenariusza użycia dotyczącego segmentacji obrazu**, Doktorantka skoncentrowała się na zadaniu segmentacji błony Brucha w obrazowaniu OCT, a dokładniej w pojedynczych tomogramach pochodzących z tego obrazowania. Celem było w szczególności wyznaczenie otwarcia w błonie Brucha (Bruch's Membrane Opening, BMO), którego wyznaczenie ma szczególnie przełożenie diagnostyczne. Wykorzystano zbiór badań częściowo pokrywający się z poprzednim scenariuszem użycia, obejmujący 325 pacjentów, dla których z wolumenu danych OCT wybrano po 16 tomogramów obejmujących obszar dysku. W efekcie powstała kolekcja 5200 obrazów, która została następnie ręcznie oetykietowania przez eksperta, który w każdym z tomogramów oznaczył dwa punkty charakteryzujące położenie końców BMO. Następnie tak otrzymane dane przełożone zostały na zadanie klasyfikacji, w którym wszystkie piksele w tomogramie znajdujące się w obrębie BMO stanowią klasę pozytywną, a pozostałe piksele klasę negatywną. Autorka zaproponowała trzy sposoby prezentacji tak przygotowanych danych systemom uczącym się: w postaci pojedynczych a-skanów, podokien, oraz całych obrazów. Dla każdej z tych trzech reprezentacji rozważano także wariant trójwymiarowy (3D), w którym poza bieżącym tomogramem na wejście systemu uczącego się podawane są także dwa tomogramy sąsiednie, tj. poprzedni i następny (w sensie osi Y).

W tym scenariuszu użycia wykorzystano wyłącznie modele uczenia głębokiego, których architektury były dostosowane do trzech wymienionych wyżej reprezentacji danych wejściowych: architekturę gęstą dla pojedynczych a-skanów, architekturę splotowo-gęstą dla klasyfikacji podokien (patches, windows), oraz architekturę typu U-net dla wariantu reprezentacji w którym wykorzystywano całe tomogramy. W tym ostatnim przypadku wartości zwracane przez model dla poszczególnych pikseli musiały być poddane jeszcze dodatkowemu przetwarzaniu, którego celem była agregacja decyzji na poziomie całego a-skanu. Decyzje wskazywane przez modele wymagają następnie dalszego przetworzenia, autorskim algorytmem 1 (sekcja 2.2.2), celem jednoznacznego wyznaczenia położeń BMO. Rozważano także inne warianty architektury U-net. Proces uczenia i testowania modeli powtarzano trzykrotnie celem uzyskania istotności statystycznej wyników. Najlepsze wyniki w kategoriach trafności klasyfikacji osiągnął model U-net. Poza tym Doktorantka przeprowadziła także ocenę jakości modeli w kategoriach rozbieżności wskazywanej lokalizacji otwarcia BMO względem lokalizacji rzeczywistej. Także i te wyniki potwierdziły przewagę modelu U-net. Dodatkowe eksperymenty z wariantami architektury U-net nie doprowadziły już do dalszych popraw.

Scenariusz użycia selekcji cech dotyczył danych klinicznych (tzw. biomarkerów). Doktorantka proponuje metodę selekcji wykorzystującą komitet klasyfikatorów, w którym każdy z klasyfikatorów bazowych używa potencjalnie innego podzbioru cech, który losowany jest nie

jednostajnie, lecz w sposób ukierunkowany rozkładem prawdopodobieństwa skonstruowanym na bazie wartości statystyki F testu ANOVA (co różni proponowane podejście od znanej metody Random Subspace). W części eksperymentalnej, mgr Sułot aplikuje zaproponowaną metodę do zbioru 211 pacjentów opisanych 48 biomarkerami, gdzie zadaniem jest przydzielenie pacjenta do jednej z 3 klas decyzyjnych (jaskra, podejrzenie jaskry, pacjent zdrowy). Ocena ilościowa wskazuje na przewagę proponowanej metody nad Random Subspace oraz podejściem referencyjnym (brak selekcji cech).

Dodatkowym wątkiem, prezentowanym w sekcjach 2.4 i 3.4, jest wykorzystanie **ortogonalnej sieci splotowej**, w której algorytm uczenia/optymalizacji wykorzystuje funkcję straty wzbogaconą o dodatkowy człon penalizujący nieortogonalność wektorów parametrów (reprezentujących poszczególne filtry w danej warstwie splotowej). Celem tego członu jest promowanie dywersyfikacji wykształcanych filtrów. Doktorantka zaaplikowała tę architekturę do zbioru danych CIFAR-10 (zadanie klasyfikacji) oraz zbioru 7340 tomogramów OCT (zadanie segmentacji warstw anatomicznych), testując skuteczność procesu uczenia w funkcji zmieniającego się rozmiaru zbioru uczącego. Eksperymenty obliczeniowe podsumowane w sekcji 3.4 wykazały skuteczność członu ortogonalizacyjnego w obu problemach, tj. modele wykorzystujące ten człon szybciej osiągały wyższą trafność klasyfikowania i utrzymywały ją także po zakończeniu uczenia; efekt ten był szczególnie widoczny dla małych zbiorów uczących.

Rozprawę zamyka rozdział 4, podsumowujący wypracowane wyniki i krótko dyskutujący możliwe kierunki dalszych prac.

2.2 Wkład oryginalny

Do oryginalnych przyczynków rozprawy należą moim zdaniem przede wszystkim:

1. Dedykowana architektura głębokiej sieci neuronowej dla scenariusza diagnostycznego i jej połączenie ze schematem komitetu głosujących klasyfikatorów (sekcja 2.1).
2. Metoda selekcji cech w połączeniu z komitetem klasyfikatorów, wykorzystująca rozkład prawdopodobieństwa bazujący na analizie ANOVA, celem selekcji skorelowanych cech (sekcja 2.3.2).
3. Zaaplikowanie znanych metod uczenia maszynowego oraz modeli sieci neuronowych (oraz ich rozszerzeń, w szczególności członu promującego ortogonalizację wag filtrów splotowych) do specyficznych problemów diagnostyki medycznej oraz przetwarzania/interpretacji obrazów medycznych, w tym wykazanie szczególnej przydatności ortogonalizacji dla problemów z ograniczoną liczbą przykładów.

2.3 Ocena zawartości pracy i uwagi polemiczne

Rozprawa zredagowana jest przejrzyście i napisana przystępnym językiem angielskim. Organizacja tekstu jest klarowna, a lekturę ułatwiają adekwatne odwołania krzyżowe oraz lista skrótów. W całej treści pracy dopatrzyłem się jedynie kilku literówek.

Doktorantka zdefiniowała w części końcowej wstępu cztery hipotezy badawcze, które w mojej ocenie zostały zweryfikowane w ramach prac przedstawionych w rozprawie. Pozwolę sobie jednak nadmienić że hipotezy 2 i 3 mają bardzo ogólnikowy charakter, i osobiście uważam że ich prawdziwość była w znacznym stopniu przesądzona już w momencie ich formułowania.

Podczas lektury rozprawy dopatrzyłem się pewnych wątpliwości i uchybień, które krótko opisuję poniżej.

1. Pierwsza uwaga polemiczna dotyczy braku dobrze zdefiniowanego zbioru testującego, który stanowiłby jedynie finalny probierz skuteczności poszczególnych modeli. Autorka przeprowadza szereg eksperymentów wykorzystujących coraz bardziej wyrafinowane modele uczenia maszynowego, oceniając ich zdolność predykcyjną schematem walidacji krzyżowej. Sam schemat walidacji krzyżowej ma oczywiście na celu redukcję ryzyka przeuczenia. jednak realizując kolejne eksperymenty na bazie tego samego podziału tego samego zbioru, otrzymywana zdolność predykcyjna stanowi podstawę decyzji odnośnie wyboru i hiperparametryzacji kolejnych modeli. W szczególności, jeżeli podział kolekcji danych na podzbiory walidacji krzyżowej pozostaje niezmienny w kolejnych eksperymentach, istnieje ryzyko że otrzymane wyniki obciążone są tym konkretnym podziałem, a w konsekwencji otrzymane modele mógłby uzyskiwać inne (gorsze) wyniki gdyby testować je na innym podziale. Metodycznie bardziej poprawne byłoby wyodrębnienie właściwego docelowego zbioru testującego, na którym modele testowane byłyby dopiero post-factum, tj. po zakończeniu wszystkich eksperymentów.
2. Wykorzystanie testu Wilcoxona dla prezentacji wyników w tabelach 3.1 i 3.2 jest ryzykowne. W obecności wielu (więcej niż dwóch) konfiguracji, wykorzystywanie testów statystycznych dla par prób zwiększa ryzyko popełnienia błędu fałszywie dodatniego. W takich przypadkach stosować można na przykład test Friedmana z analizami post-hoc. Podobna uwaga dotyczy testu t-Studenta stosowanego w sekcji 3.3 (Tabela 3.8).
3. W scenariuszu 2, stosowanie trzech osobnych podsieci neuronowych dla bieżącego poprzedniego i kolejnego tomografu w podejściu klasyfikacji całych a-skanów w trybie 3D wydaje się dyskusyjne. Rozważane tu tomogramy mają takie same charakterystyki, a zatem bardziej naturalne byłoby wykorzystanie tej samej podsieci dla wszystkich trzech tomografów (tak jak np. w tzw. sieciach syjamskich, czyli alternatywnie tzw. współdzielenie wag).
4. Równoległa prezentacja współczynników Dice i Jaccard (DSC i JSC; tabela 3.4 i dalsze) jest redundantna, ponieważ wartość jednego z nich w pełni determinuje wartość drugiego, a zależność pomiędzy nimi jest monotoniczna.
5. Agregacja wyników z Tabeli 3.7 w Tabeli 3.8 przeprowadzona jest z wykorzystaniem średnich, co jest dyskusyjne, ponieważ agregowane zmienne charakteryzują się różnymi rozkładami. Lepszym podejściem byłoby rangowanie podejść (klasyfikatorów bazowych)

a następnie uśrednianie rang. Podobna uwaga dotyczy agregacji prezentowanej w Tabeli 3.9.

6. W większości eksperymentów przeprowadzonych w rozprawie brakuje bezpośredniej konfrontacji proponowanych metod z metodami znanymi z literatury (poza nielicznymi wyjątkami, np. porównanie z metodą Random Subset). Na przykład dyskusja finalnych wyników klasyfikacji diagnostyki jaskry nie odnosi się w żaden sposób do prac wcześniej wymienionych w przeglądzie literatury (sekcja 1.3). Dalej, w drugim scenariuszu użycia, w sekcji 3.2 na dole strony 78 Doktorantka sygnalizuje że otrzymane wyniki są lepsze niż te otrzymane we wcześniej publikowanych pracach, jednak nie cytuje tych prac.
7. W pracy nie przedstawiono wielu detali proponowanych podejść i architektur, co utrudnia, a nawet uniemożliwia replikację przeprowadzonych eksperymentów. Na przykład prezentacja architektury służącej do klasyfikacji a-skanów w sekcji 2.2 ogranicza się do rysunku 2.11; brak jest detali dotyczących liczby warstw w modelu, liczby unitów w warstwach, itd. Zakładam niemniej że przynajmniej część brakujących specyfikacji znaleźć można w artykułach Autorki na których oparta jest rozprawa.
8. Z drobnych niedociągnięć, komentarz do tabeli 3.2 na stronie 71 mówi o poprawie wyników o 30%, co wydaje się wartością zawyżoną gdy porównywać je z tabelą 3.1 – chyba że Doktorantka ma na myśli poprawę względem wcześniejszych wyników.

Z komentarzy dotyczących redakcji pracy, osobiście uważam że w dzisiejszej praktyce ogólne wstępy, jak na przykład tutaj wstęp do sztucznej inteligencji, są raczej nadmiarowe. Metody omawiane w sekcji 1.1 stanowią dziś elementarz uczenia maszynowego; uważam że lepiej było poświęcić więcej miejsca na bardziej szczegółowy przegląd zastosowań uczenia maszynowego w okulistyce, oraz na scharakteryzowanie problemu małych danych oraz technik adresowania tego problemu. Jestem też zdania że podział treści rozprawy na osobny rozdział Research Summary oraz rozdział Results and Partial Conclusions jest dość sztuczny (i prowadzi też do niepotrzebnych powtórzeń), zwłaszcza gdy badania prowadzone są w trzech równoległych wątkach; bardziej naturalne byłoby przetasowanie tej treści i rozbicie je na trzy-cztery osobne rozdziały dedykowane rozważanym scenariuszem użycia.

3. Konkluzja końcowa

Przedstawiona do oceny rozprawa doktorska mgr inż. Dominiki Sułot zawiera oryginalne i wartościowe osiągnięcia, zweryfikowane na wymagających i zróżnicowanych problemach rzeczywistych. Uważam że hipotezy postawione przez Autorkę pracy zostały zweryfikowane. Wymienione wyżej uwagi polemiczne mają stanowić przydatną informację zwrotną, i nie wpływają na moją ogólnie pozytywną ocenę przedłożonej rozprawy.

Wobec powyższego stwierdzam, że **rozprawa doktorska mgr inż. Dominiki Sułot spełnia warunki stawiane przez ustawę o tytule naukowym i stopniach naukowych w odniesieniu do rozpraw doktorskich, a zatem powinna być dopuszczona do publicznej obrony, o co wnoszę do Rady Dyscypliny Naukowej Inżynieria Biomedyczna Politechniki Wrocławskiej.**