dr hab. inż. Jakub Nalepa, prof. PŚ
Katedra Algorytmiki i Oprogramowania
Politechnika Śląska
Jakub.Nalepa@polsl.pl

Gliwice, 25 września 2024 r.

Recenzja rozprawy doktorskiej

Tytuł rozprawy: A Multi-Level Perspective on the Deep Learning Models and Human-Oriented Explanations with Applications to Medical Images

Autor: Mgr inż. Weronika Hryniewska-Guzik

Promotor: Prof. dr hab. inż. Przemysław Biecek

1 Tematyka rozprawy

Recenzowana rozprawa doktorska Pani mgr inż. Weroniki Hryniewskiej-Guzik dotyczy niezwykle istotnego problemu tworzenia wyjaśnialnych i interpretowalnych modeli sztucznej inteligencji, zwłaszcza w zastosowaniach medycznych. Opracowywanie algorytmów automatyzujących analizę i ułatwiających kliniczną interpretację surowych danych medycznych, np. badań obrazowych, jest szczególnie ważne ze względu na ogromne ilości pozyskiwanych danych, których ręczna analiza jest czasochłonna (oraz kosztowna), trudna do powtórzenia[1] i podatna na ludzkie błędy. Z drugiej strony, wykorzystanie algorytmów (głębokiego) uczenia maszynowego w medycynie, których działanie jest trudne lub niemożliwe do interpretacji, a sam proces ich tworzenia i oceny nie jest precyzyjnie udokumentowany może negatywnie wpływać na „zaufanie" użytkowników do predykcji wypracowywanych przez takie modele. W swojej rozprawie, Pani mgr inż. Weronika Hryniewska-Guzik dokładnie omawia praktyczne aspekty i problemy związane z tworzeniem, oceną i wykorzystaniem algorytmów sztucznej inteligencji w medycynie, jednocześnie uwypuklając istotność prowadzonych przez siebie badań.

W niniejszej rozprawie Doktorantka podjęła się opracowania metod, procedur i algorytmów, które umożliwiają odpowiedzialne tworzenie i rzetelną ocenę technik uczenia głębokiego do analizy badań obrazowych klatki piersiowej w diagnostyce COVID-19 (obrazów RTG klatki piersiowej oraz obrazów tomografii komputerowej klatki piersiowej). Prace podjęte przez Autorkę objęły stworzenie listy kontrolnej wspierającej opracowywanie i weryfikację odpowiedzialnych technik głębokiego uczenia maszynowego, zaprojektowanie i ocenę eksperymentalną głębokiego modelu wielozadaniowego do analizy obrazów medycznych, opracowanie „polireprezentacji" danych medycznych oraz zaproponowanie interaktywnej procedury wyjaśniania modeli uczenia głębokiego, a także metod łączenia różnych rodzajów wyjaśnień.

Uważam, że tematyka rozprawy doktorskiej Pani mgr inż. Weroniki Hryniewskiej-Guzik jest aktualna, zgodna z obecnym nurtem badań i dotyczy bardzo istotnych aspektów praktycznego wykorzystywania technik i modeli uczenia maszynowego w rzeczywistych problemach, w których istotną rolę odgrywa możliwość zrozumienia i zinterpretowania ich działania.

[1] Analiza tego samego obrazu medycznego, np. obrysowanie zmian nowotworowych w obrazie pozyskanym z wykorzystaniem rezonansu magnetycznego, przeprowadzona przez dwie osoby – lub nawet tę samą osobę w różnych punktach w czasie – często prowadzi do niespójnych wyników [1, 2].

2 Ocena strony merytorycznej rozprawy doktorskiej

Rozprawa jest napisana w języku angielskim i składa się z pięciu rozdziałów oraz bibliografii.

Rozdział pierwszy jest syntetycznym wprowadzeniem do tematyki rozprawy, w którym Doktorantka omawia najistotniejsze wyzwania związane z wdrażaniem metod sztucznej inteligencji w dziedzinach, w których możliwość wyjaśnienia modeli uczenia maszynowego jest kluczowa. W kolejnych podrozdziałach Autorka omawia cztery precyzyjnie i jasno sformułowane pytania badawcze, na które będzie starała się odpowiedzieć w dalszych częściach rozprawy, a także przedstawia sześć publikacji, które zawierają najistotniejsze wyniki będące częścią rozprawy. W podrozdziale 1.2 (*Main scientific contribution from the dissertation publication)* Autorka przedstawia bardzo informatywny rysunek, który wizualizuje najważniejsze elementy rozprawy – warto zauważyć, że wszystkie rysunki i wizualizacje zawarte w rozprawie są przemyślane i łatwe w interpretacji. Chciałbym podkreślić, że umiejętność prezentowania i wizualizacji (nie tylko) otrzymywanych wyników jest niezwykle ważna w pracach naukowych, a rysunki umieszczone w rozprawie są dowodem na to, że Doktorantka przywiązuje istotną wagę do sposobu prezentowania swoich idei i wyników (a to z kolei jest jednym z dowodów dojrzałości naukowej Autorki). Dwie z sześciu prac, które zawierają wyniki badań omówionych w rozprawie zostały opublikowane w prestiżowych czasopismach (*Pattern Recognition* [3] i *Machine Learning* [4]), dwie zostały zaprezentowane na konferencjach, a dwie zostały opublikowane jako preprinty – we wszystkich artykułach Doktorantka jest „wiodącym" (pierwszym) autorem. Ponadto – w kolejnym podrozdziale – wymienione zostały trzy publikacje, które nie zostały „zawarte" w rozprawie. Należy zauważyć, że wymienione prace były prezentowane na workshopach w ramach prestiżowych konferencji międzynarodowych (takich jak CVPR czy AAAI, obie konferencje są notowane w rankingu CORE jako A*). W podrozdziale 1.3 (*Research aims*) przedstawione zostały cele badawcze, a rozdział został zwieńczony krótkim omówieniem struktury pracy.

W **rozdziale drugim**, Autorka wprowadza czytelnika w tematykę analizy obrazów medycznych i krótko omawia rozważane w pracy modalności obrazów medycznych (obrazy RTG i obrazy tomografii komputerowej) oraz przedstawia podstawowe operacje, które są zwykle wykorzystywane w ramach przetwarzania wstępnego obrazowych danych medycznych. Wstęp do obrazowania medycznego został zwieńczony omówieniem najistotniejszych wyzwań związanych z przetwarzaniem i analizą takich danych, zwłaszcza z wykorzystaniem technik uczenia głębokiego oraz z tworzeniem systemów wykorzystujących metody sztucznej inteligencji do analizy obrazów medycznych. W kolejnym podrozdziale 2.2 (*Deep Learning*), Doktorantka krótko omawia podstawowe architektury sieci neuronowych oraz przedstawia najważniejsze zadania związane z analizą obrazów medycznych, takich jak ich rekonstrukcja, segmentacja semantyczna, detekcja obiektów czy klasyfikacja, do których są wykorzystywane. W podrozdziale zawarto tabelę podsumowującą architektury sieci neuronowych, wraz z ich istotnymi cechami, np. liczbą warstw konwolucyjnych oraz liczbą parametrów trenowalnych, które są wykorzystywane w rozprawie. Autorka omawia również zagadnienia związane z uczeniem reprezentacji danych, przedstawia własności „dobrych" reprezentacji danych, a także krótko omawia metody uczenia architektur głębokich z przeniesieniem wiedzy, uczenia samonadzorowanego oraz wielozadaniowego. Kolejny podrozdział 2.3 (*Explainable Artificial Intelligence*) został poświęcony przedstawieniu technik analizy i wyjaśniania modeli uczenia maszynowego – Autorka przedstawiła ich taksonomię (wraz z przykładami wybranych narzędzi), która pozwala czytelnikowi lepiej zrozumieć, na których rodzajach narzędzi do wyjaśniania modeli Doktorantka skupi się w dalszych częściach rozprawy. Kolejne podrozdziały zawierają omówienie sposobów oceny wyjaśnień (metryki, które są powszechnie używane do tego celu mogłyby zostać tutaj dokładniej omówione) oraz przedstawienie istotności tworzenia narzędzi wygodnych dla użytkownika. W ostatnim podrozdziale Autorka przedstawia powiązanie swoich naistotniejszych wyników badań z (pod)rozdziałami podsumowującymi obecny stan wiedzy – uważam, że był to świetny pomysł, który pozwala umiejscowić wyniki Doktorantki w obecnym nurcie badań związanych z tworzeniem algorytmów uczenia głębokiego, uczenia reprezentacji oraz wyjaśnialnej sztucznej inteligencji.

Rozdziały 3–4 stanowią kluczową część rozprawy doktorskiej, w której Doktorantka omawia swoje

najważniejsze osiągnięcia. **Rozdział trzeci** (*Insights into Lung Representations*) rozpoczyna się od omówienia badań związanych z tworzeniem listy kontrolnej, która ma usystematyzować i w efekcie ułatwić proces odpowiedzialnego tworzenia i rzetelnej oceny algorytmów uczenia głębokiego w diagnostyce COVID-19, aczkolwiek przedstawione koncepcje można łatwo generalizować na inne zadania związane z analizą danych (nie tylko) medycznych. Uważam, że tworzenie takich procedur standaryzujących metody opracowywania i wdrażania modeli sztucznej inteligencji jest niezwykle ważne, zwłaszcza w kontekście systemów medycznych, które mogą zamienić się w produkty medyczne (podlegające certyfikacji) oraz w obliczu „kryzysu reprodukowalności" badań związanych z uczeniem maszynowym, z którym obecnie mierzymy się w wielu dziedzinach[2] [5]. Autorka przeprowadziła systematyczny przegląd literatury związanej z diagnostyką COVID-19 z wykorzystaniem algorytmów uczenia głębokiego oraz poprawnie zidentyfikowała istotne błędy i niedociągnięcia (związane ze zbieraniem i przygotowaniem danych, modeli i technik ich wyjaśniania) analizowanych prac, często opublikowanych w prestiżowych czasopismach. Na podstawie dogłębnej analizy wybranych artykułów, przedstawiono „listę kontrolną", która następnie została wykorzystana do ich powtórnej oceny. Istotnym elementem opracowanej listy kontrolnej jest wskazanie, jacy „aktorzy" powinni brać udział w tworzeniu „odpowiedzialnych" systemów sztucznej inteligencji – Autorka podkreśla istotność wykorzystania wiedzy domenowej (w przypadku diagnostyki COVID-19 z obrazów medycznych – wiedzy radiologa) w procesie tworzenia systemów sztucznej inteligencji. W kolejnej części rozdziału trzeciego Autorka przechodzi do omówienia wielozadaniowej architektury głębokiej, którą Doktorantka zaprojektowała do wieloaspektowej analizy obrazów tomografii komputerowej klatki piersiowej (tj. do segmentacji, detekcji, rekonstrukcji i klasyfikacji). Architektura sieci i badania eksperymentalne zostały poprawnie zaprojektowane i pozwoliły na ilościową ocenę opracowanych technik. Ostatni podrozdział 3.3 (*X-ray transferable polyrepresentation learning*) został poświęcony pracom Doktorantki związanym z wprowadzeniem polireprezentacji danych medycznych, łączącej różnorodne reprezentacje obrazowych danych medycznych (np. wykorzystujące ich cechy radiomiczne oraz głębokie, wypracowane przez sieci syjamskie). W swoich badaniach eksperymentalnych Autorka skupiła się nie tylko na ocenie jakości polireprezentacji danych w problemie klasyfikacji obrazów medycznych, ale także na zrozumieniu, które składowe polireprezentacji są najistotniejsze.

W **rozdziale 4.**, Doktorantka opisuje wyniki badań związanych z metodami interaktywnych wyjaśnień oraz z tworzeniem algorytmów grupowania wyjaśnień wypracowywanych przez różnorodne techniki wyjaśniania algorytmów uczenia maszynowego. Autorka precyzyjnie przedstawia najistotniejsze wady metody LIME, takie jak brak możliwości ręcznego wyboru analizowanych spójnych obszarów (superpikseli), trudność w ocenie cech istotnych z punktu widzenia działania modelu czy brak odporności na szum w danych, które stanowiły motywację do stworzenia interaktywnego algorytmu LIMEcraft. Ma on ułatwić użytkownikowi (nawet niedoświadczonemu w wykorzystywaniu technik wyjaśniania modeli uczenia maszynowego) zrozumienie i analizę zadanego modelu sieci głębokiej. Aby zweryfikować, czy algorytm LIMEcraft w istocie jest wolny od wad metody LIME zaprojektowano szereg eksperymentów, zwieńczonych eksperymentem, w którym udział wzięło 20 użytkowników systemu, których opinie (skwantyfikowane w ramach opracowanej ankiety) pozwoliły na wykazanie, że możliwość ręcznego wyboru spójnego obszaru w obrazie okazała się kluczowa dla użytkowników. W kolejnym podrozdziale 4.2, Autorka przedstawia prostą oraz oszczędną obliczeniowo (i bardzo „elegancką") metodę grupowania wyjaśnień (NormEnsembleXAI) wypracowywanych przez szereg narzędzi wyjaśniania modeli uczenia maszynowego. W podrozdziale zostały omówione zalety i wady innych technik agregacji wyjaśnień, a algorytm NormEnsembleXAI został porównany z innymi dostępnymi technikami (z wykorzystaniem szeregu metryk oceny wyjaśnień znanych z literatury). W ostatnim podrozdziale rozdziału czwartego Autorka skupia się na omówieniu systemu agregującego wyjaśnienia (za pomocą sieci konwolucyjnych), w którym wykorzystano nowe metryki oceny zbiorów, reprezentacji danych i samych wyjaśnień. Opracowane podejście wykorzystuje bardzo ciekawą koncepcję użycia wyjaśnień do oceny różnych (kluczowych) aspektów systemów sztucznej inteligencji i zostało wszechstronnie zweryfikowane w ramach badań eksperymentalnych.

[2]Autorka zwraca uwagę, że tylko w sześciu (z 25) analizowanych pracach autorzy zawarli informację o ich implementacji i upublicznili ich kod.

W ostatnim **rozdziale 5.**, Autorka podsumowuje rozprawę, krótko omawiając najistotniejsze elementy pracy oraz możliwe kierunki dalszych badań. Doktorantka kończy pracę stwierdzeniem, że *„the era of creating classifiers that are only good in performance metrics is slowly coming to an end. Now, the time for safe and trustworthy artificial intelligence begins."*, z którym zdecydowanie się zgadzam.

Rozprawa doktorska mgr inż. Weroniki Hryniewskiej-Guzik prezentuje zarówno wiedzę teoretyczną jak i praktyczną Doktorantki w dyscyplinie *Informatyka Techniczna i Telekomunikacja*, a jej stronę merytoryczną oceniam jednoznacznie pozytywnie. Na szczególną uwagę zasługuje fakt, że Autorka udostępniła implementacje swoich metod. W ramach badań eksperymentalnych Doktorantka zweryfikowała wszystkie hipotezy badawcze i osiągnęła cele badawcze, które dotyczą istotnych problemów bezpośrednio związanych z opracowywaniem modeli głębokiego uczenia maszynowego oraz metod wyjaśniania i analizy takich algorytmów. Problemy te są niezwykle ważne w wielu dziedzinach – Autorka, w ramach swoich badań, skupiła się na analizie danych medycznych, które są znakomitym przykładem, w którym zaproponowane metody odpowiedzialnego tworzenia wyjaśnialnych algorytmów sztucznej inteligencji są szczególnie istotne.

Z pełnym przekonaniem stwierdzam, że Doktorantka jest świetnie przygotowana do tego, żeby prowadzić dalsze badania związane z projektowaniem, implementowaniem i wszechstronną weryfikacją systemów wykorzystujących metody (wyjaśnialnego) uczenia maszynowego.

2.1 Pytania i uwagi

W trakcie lektury rozprawy nasunęły mi się następujące pytania i uwagi:

1. W pracy zabrakło mi bardziej „formalnego" zdefiniowania kluczowych metryk (*localization, faithfulness, robustness, complexity, randomization*) – ich bardziej szczegółowy opis pojawia się dopiero na str. 162.

2. W Tabeli 3.1, która podsumowuje dostępne zbiory danych, zabrakło mi informacji dotyczącej sugerowanego podziału zbiorów na podzbiory treningowe i testowe dla większości omówionych zbiorów (o takim podziale Autorka wspomina w przypadku np. zbioru nr 6). Czy takie podziały nie zostały opracowane dla pozostałych zbiorów danych (przez ich autorów)? Na str. 78 czytamy, że najczęstszym podziałem, który jest stosowany w analizowanych pracach jest podział 80/20 – warto byłoby również wspomnieć o możliwej stratyfikacji. Czy podzbiory treningowe i testowe, które były wykorzystywane w analizowanych pracach były w jakiś sposób stratyfikowane? Jeśli tak, to w jaki?

3. Autorka, na str. 78, zauważa, że w pracy [6] wymieniono zestaw sugerowanych metryk ilościowych, które powinny być wyznaczane do oceny jakości działania algorytmów klasyfikacji binarnej, wieloklasowej i wieloetykietowej. Korzystnie byłoby podsumować, które konkretnie metryki były wyznaczane w analizowanych pracach – czy w którejkolwiek pracy wyznaczany był współczynnik korelacji Matthewsa?

4. Niektóre rekomendacje, które zostały zawarte w opracowanej liście kontrolnej są nieprecyzyjne. W sekcji *Model performance* czytamy, że *Are at least a few metrics of those proposed in Albahri et al. [6] used?* – co w tym przypadku oznacza *„a few"*? Czy nie warto byłoby „wymusić" wyznaczanie wszystkich rekomendowanych metryk (być może rozszerzonych o współczynnik korelacji Matthewsa) w trakcie tworzenia i oceny modeli uczenia maszynowego? Jeśli dodatkowo algorytmy uczenia maszynowego byłyby walidowane na tych samych danych testowych (np. rekomendowanych zbiorach danych testowych, lub na rekomendowanych podzbiorach dostępnych zbiorów), to moglibyśmy łatwo i rzetelnie porównywać istniejące i nowe algorytmy diagnostyki COVID-19 z danych obrazowych. Taka strategia została wykorzystana np. w ramach konkursu BraTS (*Brain Tumor Segmentation Challenge*), skupiającego się na automatycznym obrysowywaniu zmian nowotworowych w obrazowaniu rezonansem magnetycznym głowy [6].

5. Pewien niedosyt pozostawia brak przeprowadzenia testów statystycznych, które mogłyby pozwolić na zweryfikowanie, czy różnice w jakości działania opracowanych technik, np. architektur wielozadaniowych, są statystycznie istotne (np. Tabele 3.10–3.11). Jakie testy statystyczne można byłoby zastosować?

6. Co Doktorantka ma na myśli pisząc, że „*In Figure 3.15, the cross-validation results show that in terms of accuracy...*"? W jaki sposób została przeprowadzona walidacja krzyżowa? Czy zbiór nie został podzielony raz na podzbiory treningowe i testowe („*The datasets were divided into training and validation sets in proportion 4:1.*", str. 103)? Umieszczenie tabeli podsumowującej podziały na podzbiory treningowe i testowe dla każdego eksperymentu mogłoby ułatwić zrozumienie procedury walidacyjnej w niektórych eksperymentach.

7. W jaki sposób zostały dobrane wagi 0.2 i 0.8 (Rys. 3.16) i jaki wpływ na ten wybór na jakość działania tej architektury?

8. W niektórych eksperymentach wykorzystano wczesny warunek stopu treningu sieci (np. na str. 108) – warto byłoby podsumować hiperparametry procedury uczenia sieci, np. w formie tabeli, dla każdego eksperymentu.

9. Czy w Tabeli 3.13 przedstawiono wyniki dla najlepszego modelu (tj. modelu po 28 epokach treningu)?

10. Na str. 112 czytamy, że „*The data from different sources were preprocessed by imputing missing values using...*" – kiedy pojawiały się takie braki w danych?

11. Na str. 122 czytamy, że „*LIMEcraft suggests how many superpixels should be inside and outside the selected areas to maintain the same size of the superpixels.*" – dlaczego, według Doktorantki, zachowanie podobnej wielkości superpikseli może być korzystne? Czy można byłoby zaproponować inną metrykę, która mogłaby pozwolić na automatyczne wyznaczenie docelowej liczby superpikseli w analizowanym obszarze?

12. Czy wykorzystanie bezparametrycznego algorytmu grupowania (zamiast k-średnich) w metodzie LIMEcraft mogłoby być korzystne?

13. Wartości 0, 1 i -1 w podpisie Rys. 4.7 powinny zostać dokładniej wyjaśnione.

14. W czasie lektury pracy miałem wrażenie, że informacje zawarte w niektórych sekcjach są niepotrzebnie powtarzane (np. w sekcji 4.2.2 czy 4.3.2).

15. Czy, według Doktorantki, opracowane metody agregacji wyjaśnień mogłyby zostać wykorzystane w środowiskach uruchomieniowych z ograniczeniami sprzętowymi (np. na urządzeniach brzegowych z ograniczeniami pamięciowymi, obliczeniowymi i energetycznymi, takimi jak np. satelity obrazujące powierzchnię Ziemi)? Które metody wypracowywania wyjaśnień warto byłoby rozważyć w takim wypadku?

3 Ocena strony formalnej rozprawy doktorskiej

3.1 Ocena układu pracy

Układ rozprawy doktorskiej jest poprawny i logiczny, a kolejne rozdziały wprowadzają czytelnika w najistotniejsze aspekty prac Doktoranta.

3.2 Ocena zastosowanego piśmiennictwa

Bibliografia zawiera 295 pozycji. Dobór prac jest poprawny i jest jednym z dowodów na to, że Autorka świetnie orientuje się w aktualnym nurcie badań związanych z tematyką rozprawy. W trakcie analizy wykorzystanego piśmiennictwa zauważyłem niewielkie uchybienia:

- W niektórych wpisach w bibliografii możemy zauważyć niepoprawne wykorzystanie wielkich i małych liter, np. zamiast „COVID-19" i „X-ray" czytamy „covid-19" i „x-ray" (w [54]) oraz „HAZIZA" zamiast „Haziza" (w [168]).
- W bibliografii możemy znaleźć niekompletne wpisy, w których brakuje stron (np. [13], [88], [95], [109], [126], [175]).
- Zauważyłem, że w jednym wpisie nazwa konferencji pojawia się zapisana wyłącznie skrótem (ICCBR Workshops w [58]) – w pozostałych wpisach użyto pełnych nazw.

3.3 Uwagi redakcyjne

Rozprawa jest napisana bardzo starannie, a tekst jest poprawny pod względem językowym, stylistycznym i interpunkcyjnym. W pracy można dostrzec pewne drobne uchybienia, które w najmniejszy sposób nie wpływają na mój pozytywny odbiór rozprawy doktorskiej:

- Autorka używa dywizu zamiast (pół)pauzy (myślnika).
- W rozprawie zauważyłem nieliczne błędy typograficzne, np. „sieroty" (str. 7).
- Wszystkie skróty powinny być zdefiniowane przy ich pierwszym użyciu (np. DICOM i NIfTI na str. 27).
- Zamiast np. 7x7 lepiej byłoby zapisać 7×7.
- Na str. 43, str. 116 i na str. 168 zauważyłem brakującą spację („*As presented in Figure2.15...*", „*...of the neural network.The lack...*", „*testset*"), a na str. 60 brakującą kropkę w podpisie Rys. 3.3.
- Na str. 58 możemy znaleźć niepoprawne odwołania do sekcji II–III i sekcji IV.
- W nielicznych miejscach pracy Autorka niepoprawnie używa wielkich i małych liter, np. „*non-covid patients*" (zamiast „*non-COVID-19 patients*") na str. 93, „*dicom*" (zamiast „*DICOM*") na str. 103, „*auc*" (zamiast „*AUC*") na str. 156.
- Autorka naprzemiennie używa pisowni brytyjskiej (np. „*initialise*") i amerykańskiej (np. „*customization*") – powinno to zostać uspójnione.
- W podpisie Tabeli 3.12 warto byłoby wyjaśnić co oznacza pogrubienie wyników w tabeli (najlepsze wyniki dla zbioru B2000).
- Na Rys. 4.2 zauważyłem drobne literówki („*numer*"→„*number*", „*perturbated*"→„*perturbed*"). Literówka pojawia się też na str. 129 („*LIMEcraf*"→„*LIMEcraft*") oraz na str. 152 („*focuse*"→„*focus*").
- W Tabeli 4.1 Autorka użyła formy skróconej („*don't*"→„*do not*").
- Podpisy niektórych tabel są umieszczone pod nimi (np. Tabeli 4.1), a niektóre nad nimi (np. Tabeli 4.4).
- Formatowanie równania 4.1 jest niepoprawne (prawdopodobnie `scriptsize`).
- Po równaniu 4.6 nie powinno być wcięcia (str. 158).
- Separatorem dziesiętnym powinna być kropka (np. w Tabeli 4.6 separatorem jest przecinek).
- W Sekcji 4.2.3 i w sekcji 4.3.3 można było uspójnić symbole oznaczające wyjaśnienia dostępne dla *i-tej* instancji (E i $\mathcal{E}$).
- W rozdziale pierwszym występuje kolizja oznaczeń – artykuły, które zostały zawarte w rozprawie są oznaczone jako [P1]–[P4] (sekcja 1.2.1), a cele badawcze jako [P1]–[P6] (sekcja 1.3).
- Nie zauważyłem odnośników w tekście do Rys. 4.12 i 4.13.

4 Konkluzja

Z pełnym przekonaniem stwierdzam, że recenzowana dysertacja Pani mgr inż. Weroniki Hryniewskiej-Guzik **spełnia** wymagania stawiane rozprawom doktorskim przez Ustawę – praca prezentuje ogólną wiedzę teoretyczną i praktyczną Doktorantki w dyscyplinie *Informatyka Techniczna i Telekomunikacja* oraz umiejętność samodzielnego prowadzenia pracy naukowej, a przedmiotem rozprawy doktorskiej jest oryginalne rozwiązanie precyzyjnie zdefiniowanego problemu naukowego.

W związku z powyższym, **wnioskuję o przyjęcie rozprawy doktorskiej oraz o dopuszczenie mgr inż. Weroniki Hryniewskiej-Guzik do publicznej obrony**.

Ze względu na bardzo wysoką jakość badań i wyników przedstawionych w rozprawie, bogaty dorobek publikacyjny Doktorantki oraz Jej dojrzałość badawczą, wnioskuję również o **wyróżnienie rozprawy doktorskiej**.

Jakub Nalepa

Literatura

[1] M. Benchoufi, E. Matzner-Lober, N. Molinari, A.-S. Jannot, P. Soyer, Interobserver agreement issues in radiology, Diagnostic and Interventional Imaging 101 (10) (2020) 639–641. doi:https://doi.org/10.1016/j.diii.2020.09.001.
URL https://www.sciencedirect.com/science/article/pii/S2211568420302175

[2] D. Gut, M. Trombini, I. Kucybała, K. Krupa, M. Rozynek, S. Dellepiane, Z. Tabor, W. Wojciechowski, Use of superpixels for improvement of inter-rater and intra-rater reliability during annotation of medical images, Medical Image Analysis 94 (2024) 103141. doi:https://doi.org/10.1016/j.media.2024.103141.
URL https://www.sciencedirect.com/science/article/pii/S1361841524000665

[3] W. Hryniewska, P. Bombiński, P. Szatkowski, P. Tomaszewska, A. Przelaskowski, P. Biecek, Checklist for responsible deep learning modeling of medical images based on COVID-19 detection studies, Pattern Recognition 118 (2021) 108035. doi:https://doi.org/10.1016/j.patcog.2021.108035.
URL https://www.sciencedirect.com/science/article/pii/S0031320321002223

[4] W. Hryniewska, A. Grudzień, P. Biecek, LIMEcraft: handcrafted superpixel selection and inspection for Visual eXplanations, Machine Learning 113 (5) (2024) 3143–3160. doi:10.1007/s10994-022-06204-w.
URL https://doi.org/10.1007/s10994-022-06204-w

[5] S. Kapoor, A. Narayanan, Leakage and the reproducibility crisis in machine-learning-based science, Patterns (Aug. 2023). doi:10.1016/j.patter.2023.100804.
URL https://www.cell.com/patterns/abstract/S2666-3899(23)00159-9

[6] S. Bakas, M. Reyes, A. Jakab, et al., Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge, CoRR abs/1811.02629 (2018). arXiv:1811.02629.
URL http://arxiv.org/abs/1811.02629