Uniwersytet Rzeszowski

Rzeszów, 24.01.2025 r.

**Recenzja pracy doktorskiej mgr. inż. Czesława Horynia pt.: „Identyfikacja anomalii w dziedzinowych zbiorach danych złożonych" przygotowanej pod kierunkiem promotora dr hab. Agnieszki Nowak-Brzezińskiej, prof. UŚ, w dziedzinie Nauk Ścisłych i Przyrodniczych, w dyscyplinie Informatyka.**

Recenzja została sporządzona w związku z powołaniem przez Radę Naukową Instytutu Informatyki Uniwersytetu Śląskiego w Katowicach w dniu 05.11.2024 roku do pełnienia funkcji recenzenta w postępowaniu o nadanie stopnia naukowego doktora Nauk Ścisłych i Przyrodniczych w dyscyplinie Informatyka panu mgr. inż. Czesławowi Horyniowi.

Niniejsza recenzja ma za zadanie zgodnie z Art. 13 ust. 1 ustawy z dnia 14 marca 2003 r. o stopniach i tytule naukowym oraz o stopniach i tytule naukowym w zakresie sztuki (t.j. Dz.U. 2017 poz. 1789) ocenić, czy rozprawa doktorska stanowi oryginalne rozwiązanie problemu naukowego oraz wkład w dyscyplinę, zgodnie z art. 175 ust. 1 Przepisów wprowadzających ustawę – Prawo o szkolnictwie wyższym i nauce z 3.7.2018 r. (Dz.U. 2018 r. poz. 1669).

W ramach przeprowadzonej recenzji zostaną ocenione następujące punkty:

1. **Tematyka pracy doktorskiej i jej wkład w dyscyplinę.**
2. **Ocena zgodności pracy z wymaganiami formalnymi.**
3. **Ocena merytoryczna.**
4. **Ocena języka i stylu**
5. **Uwagi krytyczne i sugestie.**
6. **Podsumowanie i ocena końcowa.**

**Ad. 1.** Temat rozprawy doktorskiej brzmi: „Identyfikacja anomalii w dziedzinowych zbiorach danych złożonych".

W pracy została sformułowana teza badawcza, brzmiąca następująco:

*Zastosowanie zaawansowanych technik, takich jak SOM, LOF z podziałem na bloki oraz autoenkodery w ramach zespołu algorytmów, umożliwia zwiększenie czułości i wydajności wykrywania anomalii w rzeczywistych, złożonych zbiorach danych, przy jednoczesnym zachowaniu wydajności procesu analizy, szczególnie dzięki optymalizacji LOF.*

Treść pracy odpowiada jej tytułowi a równocześnie z postawioną tezą zostały określone następujące cele badawcze:

- zdefiniowanie pojęcia anomalii oraz identyfikacja wyzwań związanych z ich wykrywaniem w złożonych, kategorycznych i wielowymiarowych danych,
- zastosowanie i ocena nowoczesnych technologii, takich jak uczenie maszynowe i głębokie uczenie, w kontekście identyfikacji anomalii, z naciskiem na zwiększenie czułości,

- rozwój i optymalizacja zaawansowanych technik zespołowych, w celu zwiększenia skuteczności i wydajności procesu wykrywania anomalii,
- implementacja i testowanie proponowanych metod na rzeczywistych zbiorach danych, aby ocenić ich efektywność, czułość oraz użyteczność w praktycznych zastosowaniach.

Problematyka poruszana w pracy dotyczy opracowania zespołowego podejścia do analizy i wykrywania anomalii w danych. W ramach prac wybrano trzy metody bazowe oceny anomalii jak Local Outlier Factor, Auto Encoder, Self-Organizing Map. Badania dotyczyły niejako dwóch obszarów: optymalizacji rozmiaru bloku w algorytmie LOF oraz zespołowego środowiska Trinity SALT do oceny skuteczności identyfikacji anomalii. Badania przeprowadzono z zastosowaniem zestawu 14 różnych zbiorów danych.

Na podstawie wyników przedstawionych w recenzowanej pracy, uważam, że teza została spełniona, a niniejsza rozprawa spełnia wszystkie standardy obowiązujące w przypadku prac doktorskich oraz w znacznym stopniu przyczynia się do rozwoju dyscypliny naukowej jaką jest Informatyka.

**Ad. 2.** Przedstawiona praca odpowiada wymaganiom stawianym rozprawom doktorskim. Stanowi oryginalne rozwiązanie dotyczące usprawnienia metod identyfikacji anomalii w danych. Autor samodzielnie nakreślił problematykę, przedstawił tło oraz motywację dotyczące zagadnienia analizy anomalii i przypadków odstających, oraz samodzielnie określił zakres badań i eksperymentów przewidzianych do realizacji.

Całość przedstawionej rozprawy doktorskiej składa się z 393 stron (wliczając poszczególne rozdziały, spis rysunków, spis tabel, spis algorytmów, wykaz badań dodatkowych, załączniki (dodatki) - A, B i C, wykaz cytowanej literatury). Rozprawa składa się z 9 rozdziałów (wliczając Wprowadzenie oraz Podsumowanie), 83 rysunków, 90 tabel (w tym 56 w załącznikach), 8 algorytmów oraz aż 436 pozycji literaturowych. W pierwszych rozdziałach autor nakreślił tło zagadnienia związanego z identyfikacja anomalii i przypadków odstających, czyli jego definicję, rodzaje i obszary występowania, a także wskaźniki skuteczności identyfikacji ich występowania. Następnie przedstawił koncepcję proponowanego rozwiązania, eksperymenty obliczeniowe i uzyskane wyniki. Praca zawiera również wymagane spisy rysunków, tabel, algorytmów, a także dodatki zawierające bardziej szczegółowe wyniki eksperymentów i pozostałe informacje.

**Ad. 3.** Praca wnosi nowe, istotne wartości do dyscypliny naukowej Informatyka. Jej najważniejszymi nowatorskimi osiągnięciami są:

- zastosowanie zespołowej oceny w procesie identyfikacji anomalii w danych, z opracowaniem nowatorskiej miary agregacji oceny modeli bazowych,

- optymalizacja technik zespołowych w calu zwiększenia czułości wykrywania anomalii,

- opracowanie autorskiego środowiska Trinity SALT implementującego opracowane podejścia i udostępnionego publicznie w sieci Internet,

- przeprowadzenie szeregu eksperymentów obliczeniowych, których wyniki potwierdzają postawioną tezę badawczą.

Cele badawcze zostały sformułowane jasno. Podjęty temat identyfikacji anomalii w danych jest tematem ważnym i wciąż aktualnym w dziedzinie analizy i drążenia danych oraz uczenia maszynowego. Autor w pierwszej części pracy realizuje pierwszy z celów jakim jest zdefiniowanie anomalii i wyzwań z tym związanych, w kontekście występowania w różnorodnych danych. Następnie próbuje scharakteryzować istniejące technologie i podejścia do identyfikacji anomalii a w konsekwencji zdefiniowanie własnego zespołowego podejścia do oceny i identyfikacji anomalii z lepszą skutecznością niż metody bazowe. W

końcowej części rozprawy opisuje implementację programową proponowanego podejścia i eksperymenty obliczeniowe z zastosowaniem szeregu różnych zbiorów danych rzeczywistych.

Zastosowane metody i narzędzia badawcze zostały dobrane poprawnie i adekwatnie do postawionych celów. Również środowisko Shiny w którym zaimplementowano rozwiązanie jest znane w dziedzinie produkcji aplikacji internetowych przeznaczonych dla analizy i wizualizacji danych w sieci Internet. Opracowane zostały własne implementacje algorytmów LOF, AE oraz SOM. Zaimplementowano autorską aplikację Trinity SALT i udostępniono ją publicznie do wykorzystania przez innych badaczy. Udostępniono również przygotowane do eksperymentów zbiory danych.

Wyniki badań zostały przedstawione w sposób przejrzysty i poddane zostały odpowiedniej interpretacji. Wnioski z przedstawionych wyników są poprawne jednak można mieć pewne zastrzeżenia przedstawione w sekcji Uwagi krytyczne i sugestie.

Doktorant wykazał wystarczająco szeroką znajomość literatury przedmiotu zagadnienia analizy i identyfikacji anomalii w danych o czym może świadczyć również ogromny zasób użytych źródeł literaturowych.

Wyniki uzyskane w pracy mogą mieć znaczenie praktyczne i mogą być wdrożone do zastosowań obliczeniowych. Ponadto również w istotny sposób wzbogacają wiedzę teoretyczną dotyczącą metod identyfikacji anomalii w danych.

**Ad. 4.** Poprawność językową pracy oceniam na wysokim poziomie, nie znaleziono większych błędów stylistycznych, ortograficznych czy gramatycznych. Struktura tekstu jest przejrzysta i logicznie ułożona. Można mieć nawet uwagi, że rozprawa jest zbyt obszerna, często fragmenty treści się powtarzają a powinny być raczej zwarte i konkretne, ale nie ujmuje to jej wartości.

**Ad. 5.** Analizując rozprawę oraz uzyskane wyniki eksperymentów można wskazać pewne słabości, niedociągnięcia czy też braki a na ich podstawie postawić pytania. Syntetyczne zestawienie najważniejszych krytycznych uwag i sugestii zamieszczono poniżej.

- Uzupełnienie poszczególnych zagadnień w części teoretycznej pracy o praktyczne przykłady i implementacje, pozwoliłoby na uzyskanie pełnej wersji książkowej.
- Można mieć pewne uwagi do stylistyki treści pracy w kontekście tego, że często treści są niejako dublowane i powtarzane wielokrotnie w sposób zbędny. W przypadku ścisłych i technicznych dyscyplin można spodziewać się raczej konkretów. Często sformułowania są niejasne i „przegadane".
- Nie jest jasne na jakiej podstawie została określona wartość bonusów $b_1$, $b_2$, $b_3$, które mają wpływ na ocenę rankingu przypadków odstających.
- Nie jest jasne na jakiej podstawie na rysunku 8.4, czy w tabeli 8.2 są zaznaczone przypadki odstające, które nie zostały wykryte. Ale skąd wiemy, że to są tego rodzaju przypadki, czy też jak autor twierdzi anomalie.
- Stylistyka i kolorystyka rysunków powinna być zbliżona i podobna a nie jest, np. rysunki 8.1, 8.2, 8.3. Wyglądają one jakby pochodziły z różnych źródeł.
- Autor często używa nieprecyzyjnie pojęcia czułości, ponieważ nie wiadomo czy odnosi się ono do oceny jakości klasyfikacji czy też do zdarzenia rozpoznania przypadku odstającego, jest to dość niejasne i mylące.
- Brakuje w pracy jasnego przykładu rankingów oceny przypadków odstających z wykorzystaniem systemu SALT. Brakuje również pokazania przykładu syntetycznego przypadku odstającego, który jest dodawany do rzeczywistych danych. Poza tym niejasne jest w jaki sposób generowano syntetyczne przykłady dla rzeczywistych danych. Jeśli losowo to z jakiej przestrzeni wartości, rozkładu, czy ewentualnie jest to jakaś permutacja istniejących wartości atrybutów. Stąd pojawia

się pytanie czy na pewno generowany przypadek jest odstającym, jaka jest pewność że utworzyliśmy przypadek odstający. Nie jest też jasne czy autor wziął pod uwagę możliwość, że w rzeczywistych przykładach też może znajdować się przypadek odstający, a mimo to otrzymał z założenia etykietę non-outlier.

- Brak eksperymentów przeprowadzonych na czysto syntetycznych zbiorach danych, dla których możemy generować różne proporcje anomalii i wartości odstających w zależności od potrzeb.
- Niejasne jest również przygotowanie danych do eksperymentów w rozdziale 8. W pierwszym eksperymencie jest 7 zbiorów, w drugim 9 a w trzecim 14. Dlaczego nie jest to ten sam zestaw danych dla każdego z nich? Ponadto:
  - W rozdziale 8.1 w podpunkcie „Różnorodność i specyfika badanych zbiorów" autor pisze: „*W przypadku zbioru „mushroom", zamiast generować syntetyczne anomalie jak w poprzednich eksperymentach, przyjęto, że trujące grzyby stanowią rzeczywiste anomalie. W związku z tym, wybrano 0,988% trujących grzybów z całego zbioru jako anomalie, a resztę usunięto. W efekcie powstał mniejszy, ale bardziej realistyczny zbiór anomalii, nazwany „mushroom real"*." Powstaje pytanie na jakiej podstawie przyjęto takie założenia, co to znaczy bardziej realistyczny zbiór anomalii?
  - Podobnie jest z innymi zbiorami, np. zbiór p53 Mutant:, „*Aby zoptymalizować zbiór danych do analizy, obliczono macierz korelacji, eliminując cechy o wysokiej korelacji (powyżej 0,5), co zredukowało liczbę kolumn z 5 408 do 444.*". Powstaje pytanie jaki był cel tej optymalizacji i dlaczego akurat ten zbiór tak przetwarzano, dlaczego usunięto prawie 5000 cech, czy ma to sens? Autor pisze, że zbiór „*był trudny do przetworzenia z powodu swojej ogromnej wymiarowości oraz silnej korelacji między cechami*". Nie sądzę by był to argument za takim przetwarzaniem danych.
  - Zbiór Labeled Vehicle Claims również został „*przefiltrowany, by skupić się na istotnych anomaliach. Z 268 255 obiektów wybrano 24 968 z anomaliami w kolumnach Label i Category_anomaly, a następnie ograniczono do 1 488 kluczowych przypadków, pozostawiając 212 994 obiekty do analizy. Zbiór danych jest mocno niezbalansowany, z anomaliami stanowiącymi tylko 0,699% wszystkich obiektów. Mniej istotne kolumny zostały usunięte, pozostawiając 19 kluczowych dla analizy.*" Powstaje pytanie dlaczego akurat tak postąpiono z tymi danymi. Dlaczego w bazie p53 Mutant usunięto cechy o wysokiej korelacji, a w bazie Labeled Vehicle Claims usunięto mniej istotne cechy? Co to są kluczowe przypadki?
  - Autor nie udziela wyczerpujących odpowiedzi na powyższe pytania, a opis zbiorów jest niejednoznaczny.
- W pracy technicznej z zakresu informatyki algorytmy powinny być jednak przedstawione w formie zwartego pseudokodu z użyciem symboli matematycznych.
- Dziwna forma tabel, np. 24, 25, 27, itd. Tabele z jednym wierszem nie powinny być tabelami lecz raczej wypunktowaną listą.
- Liczba 436 odnośników literaturowych wydaje się być zbyt duża, aczkolwiek może również stanowić pewien zasób wiedzy na temat zagadnienia analizy przypadków odstających.
- Wydaje się, że autor zbyt mało uwagi poświęcił wyjaśnieniu różnicy pomiędzy anomalią a przypadkiem odstającym. Anomalia i przypadek odstający (outlier) to dwa terminy używane w analizie danych, które czasem są stosowane zamiennie, ale różnią się od siebie w zależności od kontekstu i zastosowań.

**Ad. 6.** Uważam, że Doktorant z powodzeniem zrealizował wszystkie cele oraz udowodnił postawioną tezę badawczą, aczkolwiek praca zawiera kilka niejasno określonych elementów, które mogą mieć wpływ na wyniki badawcze. Głównym minusem jest brak eksperymentów przeprowadzonych na spreparowanych danych syntetycznych, gdzie możemy z całkowitą pewnością definiować i identyfikować obiekty które są anomalią lub przypadkiem odstającym.

Pomimo tych niedociągnięć pozytywnie oceniam pracę, a moje pytania, postawione wyżej mają raczej dociekliwy charakter. Należy podkreślić, iż autor zdefiniował także aż cztery cele badawcze, które w większym lub mniejszym stopniu zostały w pracy zrealizowane i pozwoliły udowodnić postawioną tezę badawczą.

Na szczególną uwagę zasługuje również ogólna działalność naukowa doktoranta, który jest współautorem 9 publikacji naukowych (wg Google Scholar), w tym 2 posiadających współczynnik Impact Factor (IF). Jego H-index wynosi: 4 (Google Scholar), 4 (Scopus), 2 (Web of Science), a prace były cytowane następującą ilość razy: 65 (Google Scholar), 43 (Scopus) oraz 9 (Web of Science), co świadczy o dobrej (na tym etapie kariery) rozpoznawalności w środowisku naukowym.

Moja ocena pracy **mgra inż. Czesława Horynia** jest **pozytywna**. Moim zdaniem niniejsza praca prezentuje cenne wyniki badań i jest znaczącym osiągnięciem naukowym w dyscyplinie naukowej **Informatyka**. **Spełnia ona również w mojej ocenie wszystkie wymogi zawarte w aktualnie obowiązującej Ustawie z dnia 20 lipca 2018 roku "Prawo o szkolnictwie wyższym i nauce" w sprawie warunków i trybu przeprowadzania przewodów doktorskich i może być przedmiotem publicznej obrony.**

**Wnioskuję do Rady Naukowej Dyscypliny Informatyka o dopuszczenie pana mgra inż. Czesława Horynia do dalszych etapów przewodu doktorskiego.**

Paja Wiesław

.....

Dr hab. inż. Wiesław Paja, prof. UR

Instytut Informatyki, Kolegium Nauk Przyrodniczych

Uniwersytet Rzeszowski

ul. St. Pigonia 1, 35-310 Rzeszów

wpaja@ur.edu.pl