US Kancelaria
RPW/8977/2026 N
Data:2026-05-04

dr hab. Barbara Pękala, prof. UR

Rzeszów, 25 kwietnia 2026 r.

Wydział Nauk Ścisłych i Technicznych, Instytut Informatyki
Uniwersytet Rzeszowski
ul. Pigonia 1, 35-310 Rzeszów

RECENZJA ROZPRAWY DOKTORSKIEJ

Autor: Kwabena Frimpong Marfo
Tytuł rozprawy: *Classification based on dispersed data with deep learning issues*
Promotor: dr hab. Małgorzata Przybyła-Kasperek, prof. UŚ
Jednostka: Uniwersytet Śląski w Katowicach, Instytut Informatyki

1. Ocena tematyki i znaczenia pracy

Rozprawa doktorska dotyczy problematyki klasyfikacji danych rozproszonych, która stanowi jeden z kluczowych i aktualnych obszarów badań w informatyce, szczególnie w kontekście ograniczeń związanych z prywatnością danych oraz rosnącej heterogeniczności danych. Autor podejmuje istotny problem budowy globalnych modeli klasyfikacyjnych bez konieczności centralizacji danych, co ma duże znaczenie aplikacyjne m.in. w medycynie czy finansach.

Tematyka pracy wpisuje się w aktualne trendy badawcze związane z uczeniem federacyjnym oraz uczeniem rozproszonym, a jednocześnie wykracza poza klasyczne założenia tych paradygmatów, proponując nowe podejścia do radzenia sobie z heterogenicznością przestrzeni cech.

2. Cel pracy i teza badawcza

Główną tezą rozprawy jest stwierdzenie, że zastosowanie architektur opartych na sieciach neuronowych umożliwia poprawę jakości klasyfikacji w środowiskach danych rozproszonych i heterogenicznych .

Autor formułuje trzy główne pytania badawcze dotyczące: efektywności neuronowej fuzji predykcji lokalnych modeli, możliwości budowy globalnego modelu poprzez imputację brakujących cech, skuteczności podejścia opartego na ekstrakcji cech (PCA, SVD, UMAP).

Cele pracy są jasno określone i konsekwentnie realizowane.

3. Struktura i zawartość pracy

Rozprawa ma charakter cyklu publikacji i obejmuje 11 artykułów naukowych opublikowanych w czasopismach i materiałach konferencyjnych.

Rozprawa doktorska autorstwa Kwabeny Frimponga Marfo pt. *„Classification based on dispersed data with deep learning issues"* ma charakter pracy cyklu publikacji i poświęcona jest problematyce klasyfikacji danych rozproszonych i heterogenicznych z wykorzystaniem metod uczenia maszynowego, w szczególności sieci neuronowych.

Praca składa się z czterech głównych rozdziałów, poprzedzonych wprowadzeniem oraz uzupełnionych wykazem publikacji stanowiących jej podstawę. W rozdziale pierwszym Autor przedstawia motywację podjęcia badań, wynikającą z ograniczeń klasycznych metod uczenia maszynowego w środowiskach, gdzie dane są rozproszone, niejednorodne oraz objęte restrykcjami prywatności. Sformułowana zostaje główna hipoteza badawcza, zgodnie z którą zastosowanie architektur opartych na sieciach neuronowych pozwala na poprawę jakości klasyfikacji w warunkach danych rozproszonych, oraz określone zostają szczegółowe pytania badawcze.

Rozdział drugi ma charakter przeglądowy i teoretyczny. Autor definiuje pojęcie danych rozproszonych, przedstawia ich formalny opis oraz omawia podstawowe założenia funkcjonowania systemów rozproszonych. Następnie dokonuje analizy istniejących podejść do uczenia na danych rozproszonych, ze szczególnym uwzględnieniem uczenia federacyjnego oraz uczenia rozproszonego, wskazując ich zalety, ograniczenia oraz wyzwania, takie jak heterogeniczność danych, problemy komunikacyjne czy kwestie prywatności.

Kluczową część pracy stanowi rozdział trzeci, w którym zaprezentowano autorskie podejścia do problemu klasyfikacji danych rozproszonych. Rozdział ten podzielony jest na trzy zasadnicze nurty badawcze. W pierwszym z nich Autor proponuje metody fuzji lokalnych predykcji z wykorzystaniem sieci neuronowych (MLP oraz RBF), analizując ich skuteczność oraz odporność na szum i stopień rozproszenia danych. W drugim nurcie przedstawiono podejście polegające na agregacji lokalnych modeli neuronowych poprzez uzupełnianie brakujących atrybutów z wykorzystaniem sztucznie generowanych obiektów, co umożliwia budowę globalnego modelu przy zachowaniu prywatności danych. Trzeci nurt badań dotyczy zastosowania metod ekstrakcji cech (m.in. PCA, SVD, UMAP) w celu odwzorowania danych lokalnych do wspólnej przestrzeni cech, co pozwala na integrację modeli bez konieczności ich bezpośredniej synchronizacji. Każde z podejść zostało poddane szerokiej weryfikacji eksperymentalnej z wykorzystaniem danych rzeczywistych i syntetycznych.

Rozdział czwarty stanowi podsumowanie przeprowadzonych badań, w którym Autor syntetyzuje uzyskane wyniki, odnosi się do postawionej hipotezy badawczej oraz wskazuje

kierunki dalszych prac. Integralną częścią rozprawy jest także zestaw publikacji naukowych (11 pozycji), stanowiących podstawę przedstawionych badań i dokumentujących ich rozwój.

Struktura pracy jest logiczna i spójna, od przedstawienia problemu i przeglądu literatury, poprzez szczegółową prezentację metod, aż do ich empirycznej weryfikacji i syntetycznego podsumowania. Układ rozprawy umożliwia czytelnikowi stopniowe przechodzenie od zagadnień ogólnych do szczegółowych rozwiązań zaproponowanych przez Autora.

4. Oryginalność i wkład naukowy

Rozprawa doktorska wnosi istotny i wieloaspektowy wkład do rozwoju metod uczenia maszynowego w środowiskach danych rozproszonych i heterogenicznych. Oryginalność pracy przejawia się nie tylko w zaproponowanych rozwiązaniach algorytmicznych, ale również w sposobie ujęcia problemu. Autor odchodzi od klasycznych założeń jednorodności danych oraz możliwości ich centralizacji, co stanowi jedno z głównych ograniczeń istniejących metod, takich jak federated learning czy distributed learning.

Kluczową wartością pracy jest zaproponowanie metod, które:

- **nie wymagają wspólnej przestrzeni cech**,
- **nie zakładają dostępu do danych surowych**,
- **umożliwiają integrację modeli uczonych na strukturalnie różnych danych**.

Jest to istotne rozszerzenie klasycznych paradygmatów uczenia federacyjnego, które zazwyczaj wymagają zgodności przestrzeni cech lub synchronizacji modeli.

Do najważniejszych osiągnięć autora należą:

Fuzja predykcji lokalnych modeli

Zastosowanie sieci neuronowych (MLP, RBF) do agregacji wyników klasyfikatorów lokalnych (k-NN), co pozwala na zachowanie prywatności danych i poprawę jakości klasyfikacji względem klasycznych metod ensemble.

Agregacja modeli poprzez sztuczne obiekty

Nowatorskie podejście polegające na ujednoliceniu przestrzeni cech oraz budowie globalnego modelu poprzez agregację wag sieci neuronowych.

Podejście oparte na ekstrakcji cech

Zastosowanie transformacji (PCA, SVD, UMAP) do odwzorowania danych do wspólnej przestrzeni i eliminacji potrzeby imputacji danych.

Autor nie ogranicza się do zaproponowania metod, ale przeprowadza szeroką analizę wpływu: liczby atrybutów, liczby klas decyzyjnych, poziomu szumu, stopnia rozproszenia danych, a także stopnia nakładania się obiektów między zbiorami.

Szczególnie wartościowe jest wykazanie, że:

- liczba atrybutów ma największy wpływ na jakość klasyfikacji,
- modele są odporne na umiarkowany poziom szumu,
- stopień rozproszenia ma nieliniowy wpływ na wyniki.

Proponowane metody mają duży potencjał aplikacyjny w obszarach takich jak: medycyna (analiza danych pacjentów rozproszonych między szpitalami), systemy IoT, analiza danych finansowych, systemy o wysokich wymaganiach prywatności.

W szczególności możliwość pracy na danych heterogenicznych bez ich udostępniania stanowi istotną przewagę nad istniejącymi rozwiązaniami.

Oryginalność rozprawy polega na:

- zaproponowaniu nowych metod integracji modeli w środowiskach rozproszonych,
- odejściu od klasycznych założeń jednorodności danych,
- połączeniu podejść neuronowych, imputacyjnych i transformacyjnych w spójny framework.

Wkład naukowy pracy należy ocenić jako **znaczący, aktualny i dobrze udokumentowany eksperymentalnie**.

5. Ocena metodologii i poprawności badań

Metodologia zastosowana w rozprawie doktorskiej jest kompleksowa, wieloetapowa i dobrze dostosowana do specyfiki analizowanego problemu, jakim jest klasyfikacja danych rozproszonych i heterogenicznych. Autor wykazuje się dużą świadomością zarówno ograniczeń istniejących metod, jak i konieczności ich modyfikacji w warunkach braku centralizacji danych.

Autor przyjął podejście polegające na:

- analizie istniejących metod (federated learning, distributed learning),
- identyfikacji ich ograniczeń (m.in. założenie jednorodności danych),
- zaproponowaniu alternatywnych rozwiązań,
- ich systematycznej walidacji eksperymentalnej.

Takie podejście jest metodologicznie poprawne i zgodne z dobrymi praktykami badań w informatyce.

Na szczególne podkreślenie zasługuje fakt, że Autor nie ogranicza się do jednej metody, lecz buduje **spójny ciąg rozwiązań**, gdzie kolejne podejścia eliminują słabości poprzednich, a także badania mają charakter **iteracyjny i ewolucyjny**, co zwiększa ich wiarygodność.

Eksperymenty zostały zaprojektowane w sposób przemyślany i systematyczny. W szczególności:

- zastosowano zarówno **dane rzeczywiste**, jak i **syntetyczne**,
- uwzględniono różne scenariusze rozproszenia danych (od 3 do 11 lokalnych zbiorów),
- analizowano wpływ wielu parametrów:
  - liczby atrybutów,
  - liczby klas,
  - liczby obiektów,
  - poziomu szumu,
  - stopnia heterogeniczności.

Szczególnie wartościowe jest wykorzystanie **danych syntetycznych generowanych kontrolowanie**, co pozwala na izolowanie wpływu poszczególnych czynników.

Autor stosuje odpowiednie narzędzia statystyczne do oceny wyników, w tym: test Wilcoxona, test Friedmana, analizę istotności statystycznej różnic między metodami.

Takie podejście zwiększa wiarygodność wyników, pozwala na formułowanie wniosków o charakterze ogólnym, eliminuje przypadkowość obserwowanych efektów. Jest to istotny element podnoszący poziom metodologiczny pracy.

Ponadto, autor analizuje szerokie spektrum modeli: MLP, RBF, GRU, LSTM, czy sieci rekurencyjne.

Natomiast, porównanie różnych architektur pozwala na ocenę ich przydatności w środowiskach rozproszonych, zwiększa kompletność badań oraz umożliwia wskazanie najlepszego rozwiązania (np. przewaga RBF w fuzji predykcji).

Na uwagę zasługuje również analiza wpływu liczby warstw i neuronów, a także badanie wpływu parametrów takich jak liczba sztucznych obiektów czy liczba komponentów PCA.

Metodologicznie bardzo interesujące są trzy ww. podejścia zaproponowane przez Autora:

1. **Fuzja predykcji** – podejście pośrednie, unikające integracji danych,
2. **Imputacja poprzez sztuczne obiekty** – umożliwiająca agregację modeli,

3. **Transformacja do wspólnej przestrzeni cech** – eliminująca problem niezgodności danych.

Każde z podejść jest dobrze uzasadnione, posiada własną procedurę eksperymentalną i zostało poddane niezależnej ocenie. Takie ujęcie świadczy o wysokiej dojrzałości metodologicznej pracy.

Zatem, metodologia zastosowana w rozprawie jest poprawna, dobrze uzasadniona, kompleksowa, adekwatna do postawionych celów badawczych i należy ją ocenić jako **wysokiej jakości i spełniającą wymagania stawiane rozprawom doktorskim.**

6. Ocena dorobku publikacyjnego oraz analiza źródeł

Cykl publikacji obejmuje artykuły o dobrej jakości: czasopisma (m.in. *Entropy*, *PLOS ONE*) oraz konferencje międzynarodowe. Dorobek jest spójny tematycznie i stanowi solidną podstawę rozprawy doktorskiej.

Doktorant wykazał się dobrą znajomością dorobku literaturowego dotyczącego zagadnień, którym rozprawa jest poświęcona. Doktorant uwzględnił prace odnoszące się zarówno do najnowszych badań w tematyce rozprawy, jak i prace dokumentujące rozwój metod klasyfikacji danych na przestrzeni ostatnich kilkudziesięciu lat. Taki dobór literatury potwierdza dostateczną wiedzę Doktoranta w obszarze badawczym rozprawy doktorskiej i należy ocenić go pozytywnie.

7. Słabe strony rozprawy, uwagi dyskusyjne

Pomimo wysokiego poziomu merytorycznego rozprawy oraz jej istotnego wkładu naukowego, należy wskazać kilka aspektów wymagających doprecyzowania lub dalszego rozwinięcia:

7.1. Liczba rzeczywistych zastosowań

Chociaż Autor wykorzystuje zarówno dane rzeczywiste, jak i syntetyczne, to jednak:

- dominują **zbiory benchmarkowe (UCI, dane syntetyczne)**,
- brakuje szerzej zakrojonych eksperymentów w rzeczywistych środowiskach aplikacyjnych (np. medycyna, przemysł).

W kontekście deklarowanego zastosowania w systemach wrażliwych (np. ochrona zdrowia), wskazane byłoby przeprowadzenie eksperymentów na danych rzeczywistych o większej skali, oraz analiza wpływu specyfiki domeny na działanie modeli.

7.2. Odniesienie do najnowszych metod uczenia federacyjnego

Przegląd literatury jest szeroki, aczkolwiek warto rozszerzyć porównania eksperymentalne z nowoczesnymi metodami FL (np. FedAvg, FedProx, FedOpt).

7.3. Analizy wrażliwości metod na dobór hiperparametrów

Proponowane rozwiązania wymagają doboru wielu parametrów, takich jak:

- liczba neuronów,
- liczba warstw,
- liczba sztucznych obiektów,
- liczba komponentów PCA/SVD/UMAP,
- parametry sieci RBF (centra, szerokość funkcji).

Interesująca byłaby szersza analiza stabilności wyników względem zmian parametrów.

7.4. Ograniczona interpretowalność modeli

Zastosowanie głębokich modeli neuronowych powoduje, że:

- modele mają charakter **„czarnej skrzynki"**,
- trudna jest interpretacja decyzji klasyfikacyjnych.

Jest to szczególnie istotne w kontekście:

- zastosowań medycznych,
- systemów decyzyjnych wymagających uzasadnienia.

Brakuje głębszej analizy interpretowalności i wskazania potencjalnego wykorzystania technik explainable AI.

7.5. Założenia dotyczące danych rozproszonych

Proponowane podejścia zakładają m.in.:

- brak możliwości współdzielenia danych,
- częściową zgodność klas decyzyjnych,
- określoną strukturę danych lokalnych.

Nie wszystkie te założenia są jednoznacznie uzasadnione czy zweryfikowane eksperymentalnie w różnych scenariuszach.

Czy były brane pod uwagę przez Autora scenariusze dynamiczne (zmieniające się dane, klienci)?

7.6. Analiza odporności na ataki i aspektów bezpieczeństwa

W kontekście danych rozproszonych i prywatności warto szerzej rozważyć podatności metod na ataki na model, manipulację lokalnymi danymi, czy inferencję z modeli. Jest to istotne szczególnie w porównaniu z literaturą FL, gdzie bezpieczeństwo stanowi kluczowy element badań.

Wskazane uwagi mają charakter **uzupełniający i rozwijający**, nie podważają zasadniczych osiągnięć i wysokiego poziomu naukowego pracy.

8. Wnioski końcowe

W podsumowaniu stwierdzam, że rozprawa doktorska mgr Kwabeny Frimponga Marfo, pomimo pewnych uwag krytycznych, spełnia wymagania stawiane kandydatom do stopnia naukowego doktora, określone w art. 187 ust. 1 i 2 Ustawy z dnia 20 lipca 2018r. Prawo o szkolnictwie wyższym i nauce (Dz.U. z 2020r. poz. 85, z późn. zm.). Autor wykazał się bardzo dobrą znajomością literatury, zaproponował oryginalne rozwiązania, przeprowadził rzetelną analizę eksperymentalną. Wnoszę zatem o dopuszczenie jej do publicznej obrony.

dr hab. Barbara Pękala, prof. UR

Instytut Informatyki
Uniwersytet Rzeszowski