SZKOŁA GŁÓWNA
GOSPODARSTWA
WIEJSKIEGO

Dr hab. Bartosz Świderski, prof SGGW
Instytut Informatyki Technicznej
Katedra Sztucznej Inteligencji

Warszawa, 22.01.2026

Recenzja rozprawy doktorskiej mgr Michała Kassjańskiego „Zastosowanie algorytmów sztucznej inteligencji do analizy audiometrii tonalnej"

Problem naukowy poruszany w rozprawie dotyczy automatycznej, wiarygodnej (klinicznie użytecznej) klasyfikacji typu niedosłuchu na podstawie danych audiometrii tonalnej oraz przeniesienia tej funkcjonalności do praktycznego systemu wsparcia decyzji (w tym mobilnego). Problem został sformułowany jasno i konsekwentnie realizowany w całej rozprawie. Zostało to jasno ujęte wprost w hipotezach H1-H3 oraz celach G1-G4. Tematyka pracy koncentruje się zatem wokół zastosowania metod uczenia maszynowego / sztucznej inteligencji do klasyfikacji ubytków słuchu, gdzie wejściem jest audiogram w postaci cyfrowej (XML) lub obrazu (skanu).

Tematyka jest aktualna i istotna: dotyczy diagnostyki niedosłuchu (powszechny problem zdrowia publicznego) oraz deficytu specjalistów, a proponowane rozwiązania wpisują się w silny trend klinicznych systemów decision support i mHealth. W pracy zaproponowano i przebadano szereg podejść (architektur modeli), zaczynając od prostej binarnej klasyfikacji rozróżniając słuch prawidłowy od niedosłuchu, kończąc na zagadnieniu klasyfikacji do 4 klas (prawidłowy, przewodzeniowy, mieszany, odbiorczy). Niewątpliwym atutem pracy są rzeczywiste dane opracowywane we współpracy z pracownikami Kliniki Otolaryngologii Uniwersyteckiego Centrum Klinicznego w Gdańsku.

Praca jest napisana w języku angielskim, składa się z pięciu zasadniczych rozdziałów oraz załączonego cyklu 5 artykułów. Układ pracy jest klarowny, podział na części intuicyjny. Ponieważ główny trzon pracy stanowią artykuły, przechodzę do ustosunkowania się do nich.

1. Development of an AI-based audiogram classification method for patient referral

Artykuł dotyczy klasyfikacji wyników audiometrii tonalnej (dane dyskretne z progami słyszenia w funkcji częstotliwości) na dwie klasy: „prawidłowy słuch" vs „patologiczny ubytek słuchu".
Dane, na jakich oparto badanie, to 2400 serii wyników (XML) opisanych przez audiologów. Przetestowano kilka architektur sieci neuronowych (MLP, CNN, RNN), a następnie trzy warianty RNN (simple RNN, GRU, LSTM). Najlepsze wyniki uzyskano dla LSTM ACC ~0.98, AUC ~0.985, czyli osiągnięto wskaźniki świadczące o wysokiej jakości modelu. Należy przy tym podkreślić potencjalnie wysoką istotność kliniczną przedstawionych badań (niedobór specjalistów i potrzeba szybkiej wstępnej interpretacji). Ponadto charakter pracy jest zorientowany wdrożeniowo.

Uwagi krytyczne:
Wydaje się, że część informacji jest podana nieprecyzyjnie:
1. W Sekcji *Data* podana jest informacja, że zbiór zawiera 650 przypadków „normal hearing" i 1750 „pathological hearing loss". W części dotyczącej eksperymentów Autorzy podają, że stosują 5-fold cross-validation. Przyjmując za Autorami N jako klasę zdrowych pacjentów („TN - patients who have been properly classified with good hearing") otrzymujemy odwróconą proporcję w danych testujących (sięgając np. po macierz konfuzji z fig. 7 dla LSTM):

	ALL	spodziewany test	Dane z Fig.7 (Confusion matrix - LSTM)
normal hearing	650	130	(347+8) = 355
pathological hearing loss	1750	350	(1+124) = 125

total	2400	480	480

Taka odwrotna proporcja przemawia za hipotezą błędnego oznaczenia klas lub wymaga wytłumaczenia.

2. Autorzy wykorzystują podejście 5-fold cross-validation, dla sieci LSTM raportują ACC:

iteration	1	2	3	4	5	average
ACC	97.96	98.33	97.96	97.91	98.22	98.076

Jeśli iteration oznacza fold, jako średnie ACC otrzymujemy: 98.076, powstaje wówczas pytanie, skąd wartość ACC 0.9812 w tabeli II? Liczbowo wydaje się zgadzać ten wynik z macierzą konfuzji (fig 7) z jednego folda (?).

3. Podobnie przy innych miarach (dotyczy także macierzy konfuzji) warto wyjaśnić co one reprezentują - średnią po k-foldach ?. Naturalne byłoby podanie np. średniej i odchylenia w agregacji po foldach.

4. Można odnieść wrażenie, że krzywa ROC nie jest policzona po wyjściach ciągłych z modelu (scorach), a po progowaniu, co skutkuje 3-punktową krzywą łamaną. Taki sposób prezentacji wyników nie daje pełnej informacji o *discriminatory power of the model* oraz *trade-off* pomiędzy czułością a specyficznością. Hipoteza ta wydaje się zgadzać z wyliczeniem:

$$AUC = \frac{1 + TPR - FPR}{2}$$

Jeśli tak - nie jest to wówczas krzywa ROC w standardowym sensie (tylko wizualizacja 1 punktu z krzywej ROC). Wynik ten nie wnosi informacji ponad macierz konfuzji dla danego thresholdu. Jeśli próg nie był typowym progiem 0.5 (argmax dla softmax, 0.5 dla sigmoidy) warto byłoby wspomnieć, jak był ustalany.

5. Nie zaszkodziłyby tutaj informacje o leave patient out (jeśli tak było), procedurze „foldowania" (wyjaśniając liczności foldów), oraz podanie podstawowych informacji dot. architektury.

6. W tabeli III Standaryzacja jest przed podziałem na train/test, co nie jest standardem i może wiązać się z *data leakage*.

7. W opisie standardów pojawia się powtórzenie „ISO 8253 and ISO 8253" – traktuję to jak literówkę.

8. Pracę uatrakcyjniłoby wspomnienie w jaki sposób AC i BC zostały uwzględnione architektonicznie.

2. Detecting type of hearing loss with different AI classification methods

Artykuł dotyczy klasyfikacji audiogramów na 3 typy ubytku słuchu: przewodzeniowy, odbiorczy i mieszany. Danymi są wartości liczbowe z plików XML. Zbiór obejmuje 4007 audiogramów z lat 2017-2021 (jeden ośrodek). Wejściem jest zestaw progów dla 5 częstotliwości (250, 500, 1000, 2000, 4000 Hz) osobno dla przewodnictwa powietrznego i kostnego (łącznie 10 wartości). Ocena dokonywana jest przez 5-fold cross-validation. Przebadano dość szeroką paletę modeli ML i DL. Uzyskano najlepszy wynik ACC dla sieci RNN ~94.46% (±0.91) co potencjalnie świadczy również o wysokiej jakości modeli (warto podkreślić relatywnie większy zbiór danych niż w w poprzednim artykule co wzmacnia istotność otrzymanych wyników).

Jednocześnie autorzy sami zaznaczają, że 94.46% może być zbyt niskie do zastosowań klinicznych, zwłaszcza przez „prohibitively large number of false negatives" i że potrzebne są większe, bardziej reprezentatywne dane. To ważne i uczciwe zastrzeżenie.

Uwagi krytyczne:

1. W artykule pada informacja, że pacjent mógł mieć maks. dwa wyniki (lewe/prawe ucho), to oznacza, że jeśli k-fold dzieli losowo – wyniki mogą być nieco bardziej optymistyczne (versus walidacja krzyżowa po ID pacjentów). Zasadny byłby tu komentarz doprezycujący tę kwestię.

2. Wydaje się, że niektóre metryki mogłyby być znów nieco precyzyjniej opisane:
Autorzy deklarują („Due to the aforementioned class imbalance, macro averaging was calculated") zastosowanie *macro averaging* z uwagi na niezbalansowanie klas, jednak w Tabelach I–II wartości *Recall* są takie same jak *Accuracy* (także odchylenia). W problemie single-label multiclass równość Accuracy = Recall zachodzi dla *weighted recall* (oraz dla *micro recall*), natomiast dla *macro recall* jest możliwa jedynie w szczególnych przypadkach. Ponieważ raportowany *F1* nie jest równy *Accuracy*, wariant *micro* wydaje się mało prawdopodobny. Wskazane byłoby tu także doprecyzowanie, jakiego uśredniania (macro/micro/weighted) użyto dla precision/recall/F1 oraz jak agregowano wyniki po foldach.

3. Krzywe ROC – cenne byłoby wyjaśnienie jak były liczone (w tym agregowane po foldach) no i najważniejsze – wydają się być liczone znów po progowaniu zamiast przed.

4. Artykuł wymienia różne modele z nazwy, ale nie podaje szczegółów, które są kluczowe do odtworzenia wyników jak i oceny ryzyka overfittingu (np dla modeli deep: liczby warstw, rozmiarów, regularizacji (dropout/L2), optymalizatora, learning rate, liczby epok, kryterium stopu, itp). Dla klasycznych ML: np. kernel i gamma w SVM, k w KNN, parametry RF/DT, etc. Samo zestawienie, typu „model SVM uzyskał ACC 83.38% podczas gdy Random Forest 81.26%" bez podania kluczowych hiperparametrów może prowadzić do zubożonego (a nawet mylącego) obrazu różnic między metodami.
5. W tekście podobnie jak przy poprzednim artykule jest drobna redundancja „ISO 8253, ISO 8253" przy opisie standardów.

3. Efficiency of artificial intelligence methods for hearing loss type classification

Autorzy znów porównują wiele metod ML/DL do klasyfikacji typu niedosłuchu (przewodzeniowy, odbiorczy, mieszany) na podstawie surowych danych z audiometrii tonalnej (XML) zebranych klinicznie. Jest on rozszerzeniem poprzedniego m.in. poprzez uwzględnienie:

a) modeli GRU i LSTM
b) analizie wpływu standaryzacji/skalowania danych dla sieci (Z-score, MinMax, MaxAbs), pokazano, że wybór skali może mieć wyraźny wpływ na wyniki.
c) Jednym z ciekawszych aspektów jest augmentacja danych tablicowych (podwojenie) przez CTAB-GAN (conditional GAN) i analiza jej wpływu na jakość modeli (osobno ML i DL)

Autorzy otrzymali najlepszy wynik ACC ~97.56% dla sieci LSTM przy Z-score + GAN.

Uwagi krytyczne:
1. W zasadzie większość uwag krytycznych z poprzedniego artykułu odnoszą się również do tego, czyli:

a) kwestia wyjaśnienia wyjścia z 10 foldów na jeden wynik (znów autorzy deklarują makro averaging) i znów wszędzie Accuracy = Recall (włącznie z odchyleniem) ale F1 ≠ Accuracy (co przemawia za weighted).
b) Zabezpieczenie się przed potencjalnym wyciekiem danych (pacjent mógł mieć maks. dwa wyniki (lewe/prawe ucho)) – czy grupy są rozłączne pacjentowo. Standaryzacja na diagramie – znów przed podziałem – co może potencjalnie wiązać się z data leakage.
c) Krzywe ROC – zakładam że liczone Micro-averaged One-vs-Rest (agregacja po klasach), ale czy także w zakresie foldów? I znów, nieco ważniejsza sprawa: wyglądają jak po progowaniu (wydaje się 3-punktowe).
d) Brak opisu szczegółów dotyczących poszczególnych modeli (parametry, uczenie), co utrudnia odtwarzalność.

Skąpość opisu dotyczy także CTAB-GAN (typu: liczba epok, batch size, learning rate, typ optymalizatora, etc.), zakładam że był uczony na zbiorze train.
2. W opisie K-fold pojawia się zdanie sugerujące proporcję trening:test 10% : 90% (str. 4), co wygląda na błąd (standardowo jest odwrotnie). Dalej jest mowa o wydzieleniu 10% dedykowanego testu (str. 4), co już brzmi racjonalnie.
3. W tekście występuje norma, sugerujaca „literówkę": „ISO 8253, ISO 8253" (powtórzenie).

4. Automated hearing loss type classification based on pure tone audiometry data

W pracy tej autorzy proponują model sieci neuronowej (wariant RNN, stacked LSTM z warstwą Bi-LSTM na wejściu) do automatycznej klasyfikacji typu niedosłuchu dla czterech klas: słuch prawidłowy, niedosłuch przewodzeniowy, mieszany i odbiorczy (normal, conductive, mixed, sensorineural).

Niewątpliwie zaletą jest duży zbiór danych - dane obejmują 15 046 wyników badań pochodzących od 9663 dorosłych pacjentów. Przy czym jedna obserwacja = jedno ucho (zatem maks. dwa wyniki na pacjenta: lewe i prawe ucho). W pracy osiągnięto bardzo wysokie ACC 99.33% (±0.23), co przemawia za dużą skutecznością stosowanej metody.

Uwagi krytyczne:
1. Autorzy deklarują, że trzy doświadczone osoby (audiolodzy) etykietowały typ niedosłuchu zgodnie z metodologią z Tabeli 1, co może (ale nie musi) przemawiać za hipotezą, że adnotacje mogły w dużej mierze

wynikać z formalnych progów. Artykuł nie doprecyzowuje, czy był brany dodatkowy kontekst kliniczny. Jeśli etykietowanie wyłącznie w oparciu o podane reguły- przewaga praktyczna budowanego modelu względem prostego (eksperckiego) może być ograniczona. Tutaj również komentarz autora byłby zasadny.
2. W artykule podano, że z jednego pacjenta pozyskano „maksymalnie dwa wyniki" – osobno dla lewego i prawego ucha. Stratyfikowana CV jest opisana na poziomie próbek, nie pacjentów (brak informacji, że lewe/prawe ucho zawsze trafia do tego samego folda). Nasuwa się uwaga podobna do wcześniejszej - korelacje między uszami tego samego pacjenta (ta sama etiologia, wiek, ekspozycja, konfiguracja audiometru) potencjalnie mogą zawyżać metryki. To potencjalnie może również sprzyjać wysokim wartościom AUC~1.00.
3. Brakuje szczegółowego opisu hiperparametrów modelu (np: liczba jednostek LSTM, optimizer, learning rate, batch size, liczba epok).
4. Zasadny byłby również opis stosowania testu McNemara czy był wykonany na jednym foldzie czy też jest po wszystkich foldach, jaką wersję testu zastosowano, jaka była statystyka (wzór i sposób liczenia), czy uwzględniono problem zależności obu uszu – ten sam pacjent. Podobnie sprawa wyjścia na jedną „finalną" macierz konfuzji (agregacja po foldach ?).
5. Normalizacja danych – przed podziałem na train/test może sprzyjać zawyżeniu wyników.
Krzywe ROC - jest podane wprawdzie micro-average ale brak wyjaśnienia jak przebiega agregacja po foldach. I znów wyglądają na 3 punktowe (po progowaniu).

5. Development and testing of an open source mobile application for audiometry test result analysis and diagnosis support

Artykuł przedstawia otwartoźródłową aplikację mobilną (Android) wspierającą analizę audiogramów. Na podstawie zdjęcia (także wykonanego telefonem) dokonywana jest digitalizacja audiogramu do postaci danych liczbowych (wartości progów na podstawie symboli, etykiet osi i linii siatki). Dane te służą następnie jako wejście do modelu (Bi-LSTM opisanego we wcześniejszej pracy) czyli klasyfikacji typu ubytku słuchu (prawidłowy, przewodzeniowy, mieszany, odbiorczy).

W zakresie korekcji perspektywy i automatycznego obrotu – wykorzystano ML Kit. Digitalizacja wykorzystuje trzy modele YOLOv5 (detekcja audiogramu, symboli i etykiet), OCR oraz detekcję linii metodą Probabilistic Hough Transform.
Wprowadzono także aproksymację niewykrytych linii na podstawie sąsiednich linii, co zwiększa odporność na słabe zdjęcia. Model Bi-LSTM skonwertowano do TensorFlow Lite i poddano kwantyzacji potreningowej, aby zmniejszyć rozmiar i przyspieszyć działanie na urządzeniu mobilnym.

Skuteczność detektorów YOLO oceniono metrykami mAP w 5-krotnej walidacji krzyżowej osiągając bardzo wysokie wyniki mAP50 ~0.98–0.995 (aczkolwiek metryka mAP50–95 już ~0.646). Aplikację przetestowano na kilku smartfonach i w dwóch warunkach oświetlenia. Wartym podkreślenia jest utrzymywanie poprawności działania również w trudniejszych warunkach dzięki interpolacji linii.

Całość przetwarzania odbywa się lokalnie na urządzeniu, co ma praktyczny walor. Kod udostępniono publicznie (choć bez danych, co w tym przypadku jest zrozumiałe). Dla porządku wypada dodać, że system nie był projektowany pod audiogramy odręczne.

Uwagi krytyczne:
Brak metryki „end-to-end" na materiale fotograficznym (np. odsetek poprawnych klasyfikacji typu niedosłuchu liczony od zdjęcia). Artykuł m.in. raportuje metryki detekcji i liczbę interpolowanych linii, a nie końcową skuteczność rozpoznania klasy po całym łańcuchu.

W ramach uwag krytycznych do samej pracy (nie artykułów) sformułowałbym następującą uwagę:
Porównywanie performance modelu po samym accuracy (dokładności) niekoniecznie daje obiektywny obraz jakości (gradacyjny) modelu. Sam wynik może zależeć (oprócz specyfiki danych) ale także od rozkładu klas.

Pomimo uwag krytycznych (które w zasadzie są nieodłącznym elementem każdej recenzji) praca (w tym załączone prace) świadczy o bardzo dobrej znajomości zagadnień z zakresu sztucznej inteligencji, przetwarzania danych medycznych oraz audiologii, na poziomie zaawansowanym i odpowiadającym obszarowi prowadzonych badań. Autor rozwiązał postawiony problem, stosując adekwatne i nowoczesne metody z zakresu uczenia maszynowego i głębokiego, w tym sieci rekurencyjne (LSTM, Bi-LSTM). Przeprowadzone badania dowodzą dobrej znajomości metodyki badań naukowych.

Podsumowując, oryginalny dorobek obejmuje m.in.:
1. Opracowanie i walidację modelu Bi-LSTM do klasyfikacji surowych danych audiometrii na dużym zbiorze klinicznym (z bardzo wysoką skutecznością),
2. Porównanie metod (w tym także innych algorytmów niż Bi-LSTM) i wpływu przygotowania danych.
3. Praktyczne wdrożenie w postaci aplikacji mobilnej analizującej zdjęcia audiogramów (pipeline: detekcja – digitalizacja - klasyfikacja). Wartym podkreślenia jest również znaczenie praktyczne: potencjalne wsparcie lekarzy POZ i odciążenie audiologów / otolaryngologów oraz redukcja ryzyka błędów interpretacyjnych.
Dorobek ten ma też wyraz w postaci pięciu publikacji o punktacji odpowiednio 70, 70, 70, 140, 140 punktów ministerialnych (liczonym na moment publikacji) co przemawia za dorobkiem znaczącym.

W podsumowaniu chciałbym wyrazić opinię, że przedłożona praca Michała Kassjańskiego stanowi wartościowe opracowanie zarówno pod względem naukowym, jak i praktycznym, stanowiąc przy tym istotny wkład w zagadnienie automatycznej diagnostyki niedosłuchu na podstawie danych audiometrii tonalnej.

Wobec powyższego stwierdzam, że przedstawioną do oceny rozprawę doktorską mgra Michała Kassjańskiego pt. „Zastosowanie algorytmów sztucznej inteligencji do analizy audiometrii tonalnej" oceniam pozytywnie. Rozprawa zadowalająco spełnia wymagania stawiane rozprawie doktorskiej określone w art. 187 ustawy z dnia 20 lipca 2018 r. „Prawo o szkolnictwie wyższym i nauce" (tekst jednolity: Dz.U. z 2024 r. poz. 1571 z późn. zm.). W związku z powyższym wnoszę o dopuszczenie Doktoranta do dalszych etapów postępowania o nadanie stopnia doktora.

Bartosz Świderski

Dr hab. Bartosz Świderski, prof SGGW