POLITECHNIKA GDAŃSKA

Imię i nazwisko autora rozprawy: Michał Kassjański
Dyscyplina naukowa: informatyka techniczna i telekomunikacja

ROZPRAWA DOKTORSKA

Tytuł rozprawy w języku polskim: Zastosowanie algorytmów sztucznej inteligencji do analizy audiometrii tonalnej

Tytuł rozprawy w języku angielskim: Application of artificial intelligence algorithms for analysis of pure tone audiometry

Promotor *podpis*	Promotor pomocniczy *podpis*
dr hab. inż. Marcin Kulawiak	
Promotor pomocniczy *podpis*	Kopromotor *podpis*

Gdańsk, rok 2025

Acknowledgments

I would like to express my deep gratitude to Professor Marcin Kulawiak for his extraordinary commitment as my supervisor, especially for his immense patience and continuous encouragement to pursue my research. Thanks to the assistance obtained, it was possible to accomplish the entire plan concerning research and publications.

I also wish to thank the doctors from the Department and Clinic of Otolaryngology at the Medical University of Gdańsk for their hard work in gathering and categorizing data. Moreover, I would like to extend my special thanks to Professor Tomasz Przewoźny for his guidance regarding the medical aspects of this doctoral thesis.

I am also grateful to Professor Marcin Ciecholewski for initiating collaboration with the doctors and for his support during the initial phase of this PhD project.

Additionally, I would like to express my appreciation to my wonderful friends, whose support I could always rely on during times of uncertainty.

Ultimately, I want to extend my heartfelt thanks to my parents for their role in fostering a robust sense of responsibility in me, especially my mother, whose unwavering support has always been a reliable source of strength.

Abstract

The presented study investigated the classification of hearing loss types based on tonal audiometry test results. A comprehensive multi-stage study was designed and executed, employing various neural network architectures. This work was conducted in collaboration with the Department of Otolaryngology at the University Clinical Centre in Gdańsk, which provided the audiometric dataset for the study. In the initial phase, a deep neural network architecture was proposed for binary classification, differentiating between normal hearing and hearing loss. The subsequent phase focused on classifying different types of hearing loss. To achieve this, various classification algorithms have been tested on the collected dataset, including machine learning methods such as random forest, logistic regression, support vector machine, stochastic gradient descent, and decision trees, as well as neural network architectures such as multi-layer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN), long short-term memory network (LSTM) and gated recurrent unit (GRU). The knowledge gained from these experiments was applied to develop a complete classification model based on the Bi-LSTM architecture (which considered both normal hearing and specific types of hearing loss). The developed classifier achieved a 99.33% accuracy result, which is state-of-the-art in classification of hearing loss type based on audiometric data at the time of writing. The final phase of the research involved the development of a mobile application that allows medical staff to identify the type of hearing loss from a photograph of the test results taken with a smartphone. This required optimizing the classifier for mobile devices and developing a method to digitize the audiogram using OCR techniques, Hough transformation, and object detection with the YOLO architecture. The source code for the developed application has been released under an open-source license to facilitate future enhancements with additional features aimed at supporting the medical community.

Streszczenie

Tematem badań była klasyfikacja typu ubytku niedosłuchu na podstawie wyników badań audiometrii tonalnej. W tym celu zaplanowano i zrealizowano wieloetapowe badania z wykorzystaniem zróżnicowanych architektur sieci neuronowych. Prace te zostały przeprowadzone w ścisłej współpracy z pracownikami Kliniki Otolaryngologii Uniwersyteckiego Centrum Klinicznego w Gdańsku, którzy dostarczyli wykorzystywany w badaniach zbiór danych audiometrycznych. W pierwszym etapie badań zaproponowano architekturę głębokich sieci neuronowych do klasyfikacji binarnej, rozróżniając słuch prawidłowy od niedosłuchu. W kolejnym etapie skupiono się na rozwiązaniu umożliwiającym klasyfikację różnych typów niedosłuchu. W tym celu na zgromadzonym zbiorze przeprowadzono testy algorytmów klasyfikacyjnych wykorzystujących metody uczenia maszynowego, takich jak losowy las decyzyjny, regresja logistyczna, maszyna wektorów nośnych, metoda stochastycznego spadku wzdłuż gradientu i drzewa decyzyjne, jak również architektury sieci neuronowych, takie jak jednokierunkowa sieć neuronowa (MLP), konwolucyjna sieć neuronowa (CNN), rekurencyjna sieć neuronowa (RNN), długa pamięć krótkotrwała (LSTM) oraz bramkowane jednostki rekurencyjne (GRU). Zgromadzone w ten sposób doświadczenia zostały wykorzystane w kolejnym etapie badań do opracowania opartego o architekturę Bi-LSTM modelu klasyfikacji pełnej (uwzględniającego słuch normalny jak również poszczególne typy niedosłuchu). Opracowany klasyfikator w przeprowadzonych badaniach osiągnął wynik 99.33% dokładności, osiągając najlepszy rezultat klasyfikacji typu niedosłuchu na podstawie danych audiometrycznych według bieżącego stanu wiedzy. Finalnym etapem badań było stworzenie aplikacji mobilnej umożliwiającej personelowi medycznemu identyfikację typu ubytku słuchu na podstawie zdjęcia wyników badań wykonanego za pomocą smartfona. W tym celu dokonano optymalizacji opracowanego klasyfikatora pod kątem wykorzystania na urządzeniu mobilnym oraz opracowano metodę digitalizacji audiogramu opartą na metodach OCR, transformacji Hougha i detekcji obiektów z wykorzystaniem architektury YOLO. Kod źródłowy opracowanej aplikacji został udostępniony na licencji otwartej w celu ułatwienia jej przyszłej rozbudowy o nowe funkcjonalności wspomagające pracę środowiska medycznego.

Contents

1. Introduction

1.1. Dissertation outline

This dissertation is based on a series of five publications, of which three are peer-reviewed journal papers published in Springer Nature Scientific Reports and Journal of Automation, Mobile Robotics and Intelligent Systems, while the two remaining papers are peer-reviewed conference papers published as conference materials, indexed by renowned science databases such as Scopus, Web of Science and DBLP. The five papers included in the series were published in 2022—2025. All five papers, which are included in this dissertation, together comprise a consistent set on the topic of application of artificial intelligence algorithms for analysis of pure tone audiometry.

The outline of this dissertation is as follows: the Introduction Chapter 1 Section 1.2 describes the motivation for the research leading to this dissertation, in Section 1.3 the research hypotheses are formulated and described, Section 1.4 outlines the scope of this thesis and Section 1.5 presents the series of publications along with their scientific metrics.

In Chapter 2, a description of medical terminology is provided, along with a characterization of the data and of the metrics that are utilized in the evaluation of models. Next, in Chapter 3, the research from the four papers (P1 - P4) is summarized with detailed comment about author's contribution to the state of the art in terms of classification of hearing loss type. Chapter 4, based on paper P5, presents the original mobile app proposed by the author. Chapter 5 summarizes all the presented and published research material and outlines the future research areas. Chapter 6 outlines the computational resources that were utilized, while Chapter 7 presents references. The final chapter of the dissertation, Chapter P, includes all the papers that are part of the publication series for this doctoral dissertation, accompanied by statements of contribution.

1.2. Motivation

Auditory perception represents a crucial sensory function that is essential to the survival of humans and animals alike. Any impairment in auditory abilities can significantly hinder communication skills, negatively influence interpersonal relationships and jeopardize an individual's capacity to navigate and understand their surroundings. Untreated hearing loss is recognized as the third most common cause of long-term disability worldwide [1]. This condition crosses demographic lines, affecting individuals from a diverse range of age groups, and leads to substantial consequences not only for those directly impacted and their families but also for entire economic systems. The global economy encounters an estimated annual loss of around 1 trillion US dollars due to deficiencies in the diagnosis and management of hearing loss [1]. The urgency of tackling this public health issue is further emphasized by forecasts suggesting a significant rise in the incidence of hearing impairment in the forthcoming decades. Currently, it is estimated that more than 1.5 billion people suffer from varying levels of hearing loss, a number anticipated to increase to 2.5 billion by 2050, as reported by the World Health Organization (WHO) [1]. Addressing this looming crisis requires an immediate and unified effort to raise public awareness, improve access to hearing healthcare services, and implement effective intervention strategies that produce tangible results.

The timely identification and effective management of hearing impairment, especially in children, are crucial for minimizing the adverse effects associated with auditory deficiencies. Research has shown that early detection of hearing loss can significantly reduce the prevalence of auditory impairments in the pediatric population, leading to improved developmental outcomes [1]. Medical and surgical interventions for ear conditions have proven effective in restoring hearing function, frequently allowing patients to regain their original auditory capabilities. However, the successful diagnosis and management of hearing loss fundamentally depend on the availability of adequate and sustainable resources for hearing healthcare. A major challenge to the effectiveness of hearing health systems is the lack of trained professionals who can deliver essential audiological services [1]. This issue is particularly acute in low-income countries, where the ratio of ear, nose, and throat (ENT) specialists is fewer than one per million individuals. The scarcity of audiologists further complicates efforts to address the hearing health needs of these populations [1].

The complexity of the issue is further intensified by the fact that, though skilled hearing healthcare professionals can manually detect and manage certain forms of hearing loss, many conditions can only be precisely diagnosed through the use of pure tone audiometry - a technique widely regarded as the gold standard for evaluating auditory function. This assessment method measures audiometric threshold shifts,

thus enabling the categorization of hearing loss into distinct types: conductive, sensorineural, or mixed. The degree of hearing loss can range from mild to profound, significantly affecting an individual's quality of life. The use of pure tone audiometry is essential not only for individual diagnostic needs but also for enhancing epidemiological studies and creating effective rehabilitation approaches [1]. The results of pure tone audiometry are generally illustrated in an audiogram, which acts as a visual representation of the lowest sound intensity, expressed in decibels, that a person can detect across various frequencies. This information offers detailed insights into an individual's auditory abilities and serves as a vital tool for professionals in developing tailored intervention plans for those experiencing hearing difficulties.

Artificial intelligence (AI) has the potential to mitigate the disparity between the availability of hearing professionals and the growing demand for their services. AI employs algorithms that allow computers to recognize specific patterns within data analysis and derive meaningful conclusions. This capability has facilitated the formulation of research hypotheses, which are elaborated upon in section 1.3.

1.3. Research hypotheses

The research hypotheses, which have been the foundation of the presented dissertation, have been formulated in 2021. Basing on the review of the state of the art in the area of automated hearing loss type classification (more in Section 2.1), and the initial implementation of deep learning models in mobile devices (details in Section 3.1), the following statements have been formulated:

H1. ***The application of modern neural network architectures to classification of hearing loss types based on audiometric data can push the state of the art and deliver performance and accuracy viable for introduction in clinical practice.***

H2. ***Modern neural network architectures dedicated for processing raster and time-series data are capable of accurate classification of raw tonal audiometry test results.***

H3. ***It is possible to optimize modern neural network architectures to efficiently operate on smartphones which cost less than 100 USD, thus providing healthcare professionals around the world with a mobile application for classification of hearing loss types based on images of hearing test results captured with a smartphone camera.***

1.4. Scope and contribution

On the basis of the research hypotheses, formulated in Section 1.3, the following goals of this dissertation have been defined:

G1. Review of existing classification models of pure tone audiometry data and their viability for application in medical settings.

G2. Testing different neural network architectures on raw audiometry data to develop a model for hearing loss type classification.

G3. Development of a deep learning model for hearing loss type classification which would be accurate enough for implementation in clinical settings.

G4. Creation of a mobile application allowing the use of the previously developed model to classify the type of hearing loss from a photograph of audiometric test results.

1.5. Series of publications

This section describes the series of five publications that comprise a consistent set on the topic formulated as the title of this dissertation. The series consists of three journal articles and two peer reviewed conference papers.

The first article [2], referred to as (P1), is a conference material prepared for the Workshop on Artificial Intelligence for Next-Generation Diagnostic Imaging which was part of the 17th Conference on Computer Science and Intelligence Systems (FedCSIS), hosted in Sofia in 2022. In the paper several different artificial neural network models, including MLP, CNN and RNN, have been developed and tested for classification of audiograms into two classes - normal and pathological represented hearing loss.

The second paper [3], referred to as (P2), is a conference material prepared for the Doctoral Symposium - Recent Advances in Information Technology which was part of the 18th Conference on Computer Science and Intelligence Systems (FedCSIS), hosted in Warsaw in 2023. In the paper several AI-based models were used to classify audiograms into three types of hearing loss: mixed, conductive, and sensorineural.

Both paper [2] and [3] are indexed in renowned databases, including Web of Science, SCOPUS and DBLP. The FedCSIS conference rank in the Computing Research and Education Association of Australasia (CORE) ranking was assigned as B until November 2022, when the FedCSIS has been classified as multiconference and not ranked. Moreover, the Computer Science conferences ranking [4] prepared for 2012-2016 based on Google Scholar Metrics for 2000 conferences places FedCSIS on position 216, which is in the first quartile (Q1).

The third paper [5], referred as (P3), is a journal article published in 2024 in the Journal of Automation, Mobile Robotics and Intelligent Systems – JAMRIS. The paper is an extended version of conference paper P2 [3], which investigates the application of a wider range of AI based algorithms and neural network architectures to the problem of classification of tree types of hearing loss. The paper also presents the influence of training dataset augmentation with the use of a Conditional Generative Adversarial Network on the results produced by different classification methods.

The fourth paper [6], referred as (P4), is a journal article published in 2024 in Scientific Reports. The paper proposed a neural network model based on the Bidirectional Long Short-Term Memory architecture, which has been devised and evaluated for classifying audiometry results into four classes, representing normal hearing, conductive hearing loss, mixed hearing loss and sensorineural hearing loss.

The last paper [7], referred as (P5), is a journal article published in 2025 in Scientific Reports. The paper presents a novel Open Source mobile application for the Android operating system that allows users to scan and analyse audiograms using a smartphone camera and subsequently classify the type of hearing loss.

The details of each publications, including scientific metrics and Ministry of Education (MEiN) rank points [8] are presented in Table 1.

Paper ID	Title	Authors	Published in	Scientific metrics	Author's contribution
P1	Development of an AI-based audiogram classification method for patient referral [2]	Michał Kassjański, Marcin Kulawiak, Tomasz Przewoźny	Annals of Computer Science and Information Systems, IEEE (2022)	MEiN points: 70 in 2022	70%
P2	Detecting type of hearing loss with different AI classification methods: a performance review [3]	Michał Kassjański, Marcin Kulawiak, Tomasz Przewoźny,	Annals of Computer Science and Information Systems, IEEE (2023)	MEiN points: 70 in 2023	70%

Paper ID	Title	Authors	Published in	Scientific metrics	Author's contribution
P3	Efficiency of Artificial Intelligence Methods for Hearing Loss Type Classification: an Evaluation [5]	Dmitry Tretiakow, Jagoda Kuryłowicz, Andrzej Molisz, Krzysztof Koźmiński, Aleksandra Kwaśniewska, Paulina Mierzwińska-Dolny, Miłosz Grono	Journal of Automation, Mobile Robotics and Intelligent Systems – JAMRIS (2024)	CiteScore: 0.9 MEiN points: 70 (100 in 2023)	70%
P4	Automated hearing loss type classification based on pure tone audiometry data [6]		Scientific Reports (2024)	CiteScore: 6.7 Impact Factor: 3.9 MEiN points: 140	70%
P5	Development and testing of an open source mobile application for audiometry test result analysis and diagnosis support [7]	Michał Kassjański, Marcin Kulawiak, Tomasz Przewoźny, Dmitry Tretiakow, Andrzej Molisz	Scientific Reports (2025)	CiteScore: 6.7 Impact Factor: 3.9 MEiN points: 140	70%

Table 1. The information regarding each publication that is part of the series, together with its scientific metrics.

2. The problem of hearing loss type classification

2.1. Pure-tone audiometry

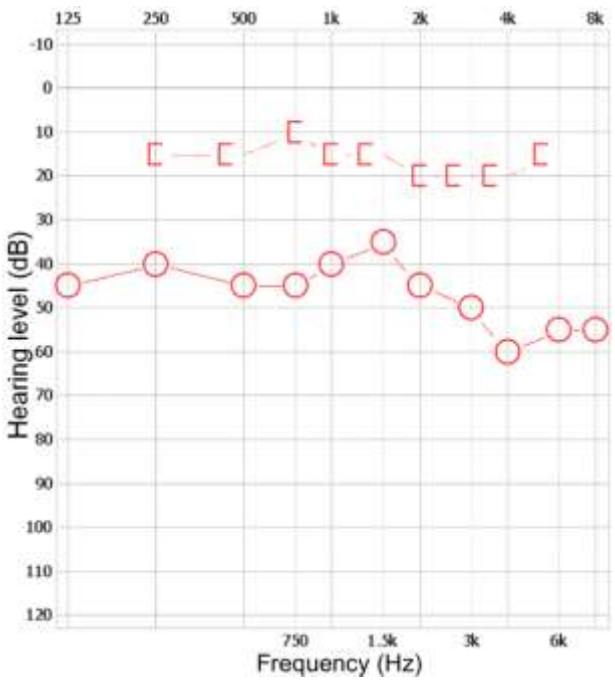
Hearing impairment is primarily evaluated through the use of pure-tone audiometry, which is traditionally performed in a soundproof setting while the individual is seated. This technique involves the presentation of a series of pure tones that gradually increase in loudness, delivered at predetermined threshold levels, usually via headphones. The aim is to determine the auditory threshold for both air and bone conduction. Air conduction assesses the functionality of the entire auditory system, which includes the auricle and extends to the auditory centers situated in the temporal lobe. Any level of impairment within this system leads to a decrease in the air conduction curve. On the other hand, bone conduction evaluates the auditory mechanism from the standpoint of the bony structure of the cochlea, bypassing the transmission of sound through the outer and middle ear. Although it offers an alternative pathway for sound transmission, its importance is generally regarded as lesser than that of air conduction. By employing pure-tone audiometry, which assesses both air and bone conduction, it becomes possible to identify the characteristics of the hearing deficit. Conductive hearing loss is usually linked to conditions affecting the external auditory canal and/or the middle ear. In contrast, sensorineural hearing loss results from damage to the sensory cells and/or the nerve fibers within the inner ear [9]. Mixed hearing loss signifies a combination of both sensorineural and conductive hearing impairments [10]. Hearing loss can present as unilateral or bilateral, may occur suddenly or develop gradually, and ranges in severity from mild to profound. Hearing impairment is widespread, especially among individuals with auditory disorders and the elderly population [11].

2.2. Data

The results obtained from pure tone audiometry are usually represented through an audiogram, which serves as a graphical representation that displays the minimum sound intensity, measured in decibels, that a person is able to hear at various frequencies. This data offers a detailed understanding of an individual's hearing capabilities and is an important resource for professionals in designing personalized interventions for those with hearing issues [12].

The datasets utilized in all articles included in this series were sourced from adult patients who were tested between 2010 and 2022 at the Otolaryngology Clinic of the University Clinical Centre in Gdansk, Poland. Tonal audiometry evaluations were carried out in soundproof booths (ISO 8253, ISO 8253). The signals were generated using calibrated Itera II and Midimate 622 clinical audiometers, produced by Madsen Electronics (Otometrics, Copenhagen, Denmark) (PN-EN 60645-1, ISO 389, ISO 8789, ISO 7566, ISO 8798). The equipment was designed to accommodate corrections for ANSI S 3.6-1989 and 2004 standard hearing levels. The assessment of participants' hearing through tonal audiometry followed the guidelines established by the American Speech-Language-Hearing Association (ASHA) [13]. During air conduction tests, the signal from the audiometer was connected to TDH-39P headphones. For bone conduction tests, the audiometer was linked to a B-71 bone vibrator (New Eagle, PA). Each patient provided a maximum of two test results, one for each ear, which ensured that there was no duplication of data from the same patient and promoted a rich variety of data [6].

In addition to the audiograms, the provided datasets also encompass XML files generated by audiology software, which contain comprehensive information about the tonal points present in the audiogram. In the P1-P4 papers, XML files were utilized to analyze the raw audiometry data. A sample audiogram along with a fragment of the corresponding XML file that includes the coordinates of the consecutive tonal points, is displayed in Fig 1.



<MeasCond><TonePoint

Freq1="125" Intensity 1="450"

></TonePoint>

<TonePoint

Freq1="250" Intensity1="400"

></TonePoint>

<TonePoint>

Freq1="500" Intensity1="450"

></TonePoint>

[…]

Fig. 1. Two methods of representing tonal audiometry test results: audiogram (left) and XML (right) [6].

As shown in Fig 1, the horizontal axis of the audiogram represents frequency, which is quantified in Hertz (Hz) and typically ranges from 125 Hz to 8000 Hz, encompassing the human hearing spectrum. The vertical axis indicates the hearing level, measured in decibels (dB), usually spanning from -10 dB (indicating very good hearing) to 120 dB (indicating profound hearing loss). A higher value on this axis means that the sound must be louder for the individual to detect it.

Additionally, audiograms utilize specific symbols to denote the results of hearing tests for each ear. For the right ear, the symbol "O" signifies air conduction thresholds, indicating the softest sounds the individual can hear through air conduction, while the symbol ">" indicates bone conduction thresholds for the right ear, showing the faintest sounds heard through vibrations applied to the skull. For the left ear, the symbol "X" represents air conduction thresholds, and the symbol "<" indicates bone conduction thresholds. Furthermore, symbols such as "□" and "△" may be used to indicate masked air conduction thresholds, which are employed when testing one ear while the other is masked to prevent cross-hearing. The symbols "[" and "]" can denote masked bone conduction thresholds [13].

2.3. Target metrics

2.3.1. Classification metrics

Classification metrics provide a way to quantitatively evaluate how well a classification model performs. They offer a deeper insight into the model's effectiveness by taking into account various aspects, ranging from the basic measure of accuracy to more complex metrics that differentiate between different types of errors. This distinction can be crucial for a comprehensive assessment of the model's performance.

The metric that is most intuitive, which is accuracy, can be mathematically articulated as the ratio of correct predictions to the overall predictions made for a certain dataset.

$$Accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions\ made}$$

In many scenarios, evaluating accuracy alone is inadequate, particularly when working with an unbalanced dataset. Therefore, to improve the assessment of the model, a confusion matrix is applied. The confusion

matrix is a table which indicates the number of correct and incorrect predictions made by the model against the actual classifications found in the test set, in addition to the nature of the errors that were made. The results from the confusion matrix can be divided into four categories:

- **True Positives (TP):** when positive predicted was true;
- **True Negatives (TN):** when negative predicted was true;
- **False Positives (FP)**: when positive predicted was false;
- **False Negatives (FN):** when negative predicted was false.

From these four parameters (TP, FN, FP, and TN), one can compute precision, recall and the F1 score. Precision is defined as the classification model's ability to accurately identify only the relevant data points, which is calculated as the ratio of all samples the model has classified as positive to the actual number of positive samples. Recall, in contrast, is the classification model's ability to identify all relevant data points; it measures the number of positive class predictions made from all instances of the positive class. Finally, the F1 score, is a single metric that combines both precision and recall, representing their harmonic mean.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Moreover, the performance of classification models is often depicted graphically in the form of the ROC curve (Receiver Operating Characteristic curve) and the AUC score (area under the ROC curve). The ROC curve demonstrates the balance between recall, which is also known as True Positive Rate (TPR), and the False Positive Rate (FPR) at various decision thresholds. The FPR shows the share of objects falsely assigned a positive class out of all objects of the negative class. In more precise terms, it pertains to the percentage of negative data samples that are mistakenly classified as positive (FP) among all negative data samples (TN + FP).

$$False\ Positive\ Rate = \frac{FP}{TN + FP}$$

2.3.2. Object detection metrics

The problem of object detection employs metrics similar to those used in classification. However, this task is more complex as it involves both localization (bounding box) and classification simultaneously. The accuracy metric that assesses the overlap between the predicted bounding box of a detected feature and the ground truth bounding box is known as Intersection over Union (IoU). Additional calculations are also derived from the confusion matrix, however, the TP, TN, FP and FN metrics need to be adjusted within the context of object detection:

- **True Positive (TP):** This refers to a precise identification where the object detection model successfully recognizes and locates objects, IoU score between the predicted bounding box and the actual ground truth bounding box meeting or surpassing a set threshold.
- **True Negative (TN):** This term is not applicable in object detection as it primarily aims to accurately confirm the absence of objects. The primary objective is to detect and identify objects rather than to validate their nonexistence.
- **False Positive (FP):** This denotes an erroneous detection, occurring when the model incorrectly identifies an object that is absent in the ground truth or when the predicted bounding box has an IoU score that falls below the established threshold.
- **False Negative (FN):** This represents a failure to detect ground truth, occurring when the model fails to recognize an object that is present in the ground truth, effectively indicating that it has overlooked these objects.

By employing the formulas defined in the preceding section it is possible to determine both precision and recall. Precision is concerned with the accurate identification of relevant objects, whereas recall highlights the model's ability to detect all ground truth bounding boxes. Collectively, precision and recall assess the equilibrium between the quality and quantity of predictions.

3. Automated classification of pure tone audiometry data

This chapter summarizes research on automated classification of hearing loss type, originally published in four peer-reviewed papers (P1-P4). The research was divided into several stages. The first stage involved creating a binary classifier to distinguish normal hearing from hearing loss (P1). This was followed by the development of a three-class classifier distinguishing between the three types of hearing loss (P2, P3). Finally, by integrating the experience and expertise from prior research, a complete classification model, consisting of 4 classes (normal hearing, sensorineural hearing loss, conductive hearing loss and mixed hearing loss), was proposed (P4).

Section 3.1 discusses the state-of-the-art. Section 3.2 describes the work related to creation of the binary classifier. Section 3.3 outlines the work related to development of the classifier distinguishing between three types of hearing loss. Section 3.4 presents the full automated classifier of hearing loss type.

3.1. State-of-the-art in audiometry data classification

In the realm of medical practice, the identification of hearing impairment types is based on pure-tone audiometry test results, which are analyzed according to their configuration, severity, lesion location (type of hearing loss) and symmetry [14]. The lesion's site is determined by the air and bone conduction thresholds on the audiogram, while the configuration is characterized by its shape. The severity is assessed by the degree of hearing loss.

The area of automatic audiometry data classification has been explored for a considerable duration overtime. In the last decade, multiple attempts have been made to establish an automated classification method that is accurate enough to be applied in practice. This work can be categorized into two main thematic areas: the classification of audiogram shapes to determine the initial configurations of hearing aids and the diagnosis of hearing loss types. In the first category, there are numerous documented attempts found in the literature, beginning with Chelz Belitz et al [15], who integrated unsupervised and supervised machine learning techniques to correlate audiograms with a limited number of hearing aid configurations. More recently, Abeer Elkhouly et al [16] proposed a machine learning solution to classify audiograms into hearing aid configurations based on their shapes using unsupervised spectral clustering. The topic of automatic hearing aid configuration is a popular one [17,18,19], yet it is quite distinct from the focus of the current PhD thesis. These publications concentrate on the shape of the audiogram, which seeks to predefine the configuration of hearing aids from a specific selection of settings through clustering methods. The popularity of this issue is not by chance; it is due to the direct applicability of these methods in the commercial market. In contrast, the classification of hearing loss types, which is directly related to a medical diagnostic problem, has attracted significantly less attention than automated hearing aid configuration.

In this context, Elbaşı and Obali [20] provided a comparison of several methodologies for assessing hearing loss, including the Decision Tree C4.5 (DT-J48), Naive Bayes, and the Neural Network Multilayer Perceptron (NN) model. The study was performed on a dataset consisting of 200 samples, categorized into four distinct groups: normal hearing, conductive hearing loss, sensorineural hearing loss, and mixed hearing loss. The input data was organized as a series of numeric values representing Decibels at constant frequency levels (750 Hz, 1 kHz, 1.5 kHz, 2 kHz, 3 kHz, 4 kHz, 6 kHz, 8 kHz). The classification algorithms were executed using Weka software, resulting in an accuracy of 95.5% for the Decision Tree, 86.5% for Naive Bayes, and 93.5% for the NN model.

In a recent investigation, Crowson et al. [21] utilized ResNet models to systematically categorize audiogram images into three categories of hearing loss: sensorineural, conductive and mixed, along with a classification for normal hearing. The study made use of a dataset that included 1007 audiograms, which were pre-processed into static plots with a resolution of 500 × 500 pixels. Instead of executing a complete training from scratch for the classifier, the authors strategically utilized transfer learning techniques, leveraging well-established raster classification models. While all assessed architectures were based on convolutional neural network (CNN) frameworks, the ResNet-101 model particularly excelled, achieving a remarkable classification accuracy of 97.5%.

Paper	Audiogram classification problem	Data size	Data type	Accuracy (%)
Ersin Elbaşı and Murat Obali [20]	Hearing loss types: normal, conductive, mixed and sensorineural	200	Raw audiometry data	95.5
Crowson et al. [21]		1007	Audiograms (raster data)	97.5

Table 2. Existing approaches to the classification of hearing loss types.

In summary, the subject of AI-based audiometry data classification has not been thoroughly explored. The existing solutions have been developed and tested on relatively small datasets, and thus their applicability tin general medical practice is limited (Table 2). Clinical guidelines indicate that the acceptable margin of error should ideally be below 5%, aiming for a target closer to 3% [22][23]. Among the classifiers reviewed, only one satisfies these criteria. Crowson et al. [21] created the most efficient audiogram classifier to date, employing transfer learning to modify an existing image classification network for the analysis of audiogram images. Although this approach achieved an impressive classification accuracy of 97%, it possesses significant limitations. As an image classifier, it cannot be directly utilized on the original data series produced by tonal audiometry. Consequently, the data must be transformed into audiogram images, which may result in the loss of critical information. Furthermore, while audiograms typically share a similar structure, those generated by different hardware and software can exhibit considerable variation. These discrepancies may encompass differences in background and line colors, as well as the volume of information displayed (for example, whether the data pertains to one ear or both). Therefore, a universal classification approach for tonal audiometry cannot depend solely on an image classifier. Moreover, since the existing studies have been performed on relatively small datasets, this limited sample size may have resulted in an overly optimistic and potentially unreliable performance evaluations. The small size of the training dataset also complicates the identification of significant patterns within specific classes, which could lead to biased validation results when applied to the test dataset.

3.2. Binary classification

This section is a summary of conference paper (P1) entitled „Development of an AI-based audiogram classification method for patient referral”.

The main objective of the research (P1) was to develop an AI-driven system designed to classify audiometry data, aiming to enhance patient referrals within the realm of hearing healthcare. Audiometry tests play a crucial role in the diagnosis of hearing impairments, but their interpretation necessitates the expertise of trained audiologists. The number of available audiologists rarely follows the growth dynamics of the patient population, which results in delays in obtaining diagnoses. This research aimed to tackle this issue by creating an AI tool capable of automatically categorizing tonal audiometry test results into two distinct groups: normal hearing and pathological hearing loss (which indicates the existence of hearing impairment). By implementing this system, the AI could support general practitioners (GPs) and primary care providers in swiftly and accurately identifying patients who require further assessment by specialists, thereby expediting the referral process and enhancing the overall delivery of healthcare.

The study used a dataset consisting of 2,400 data series contained numerical information about tonal points, defined as loudness (dB) for a given frequency (Hz), in XML format.. The dataset included the following range of frequencies: 125Hz, 250Hz, 375Hz, 500Hz, 750Hz, 1000Hz, 1500Hz, 2000Hz, 3000Hz, 4000Hz, 6000Hz, 8000Hz. These data were collected from various clinical settings and labeled by experienced audiologists, who classified each audiogram as either indicating normal hearing or hearing loss. The outline of the research described in the paper P1 is presented in Figure 2.

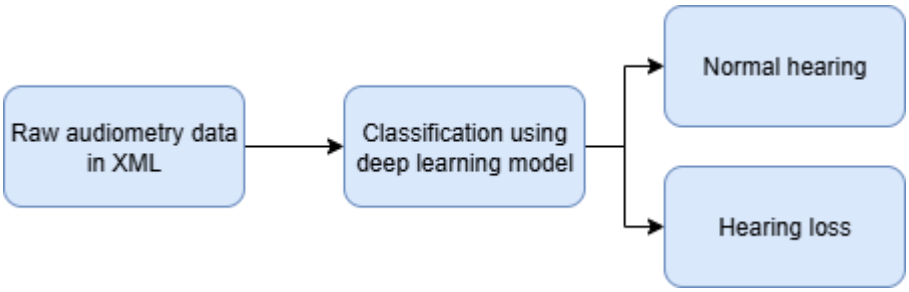


Fig. 2. An overview of the aim of paper P1.

In order to identify the most effective method for classifying the audiometry data, the paper explored various deep learning architectures. Each model has been assessed using k-fold cross-validation [24], which consists of dividing the data into k subsets and training the model k-times with k-1 subsets, with a different subset being used for testing in every iteration. The presented research used k = 5, which resulted in train to test dataset proportions of 80% to 20%, respectively. The general workflow of the study is shown in Fig. 3.

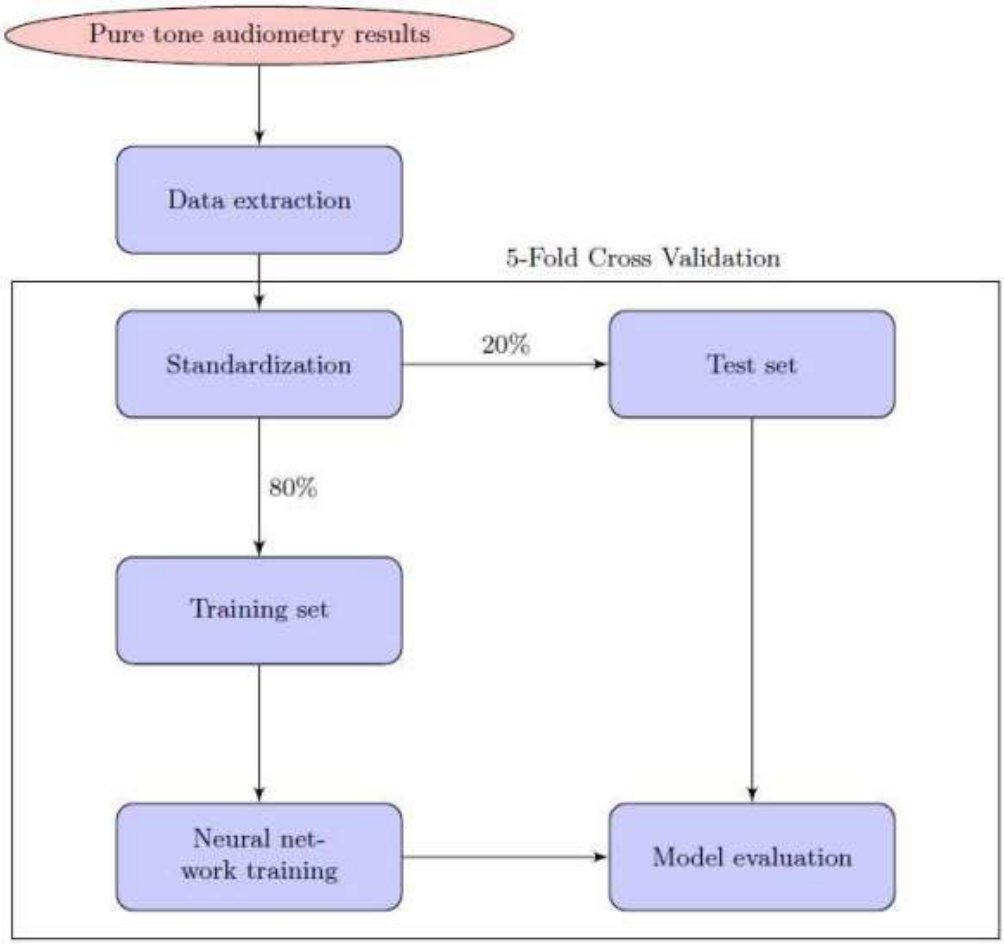


Fig. 3. The systematic approach to processes resulting in model evaluation [2].

In the discussed paper, the following artificial neural network architectures were used:

(a) Multilayer Perceptron (MLP) [25]:

The multilayer perceptron is the most prominent and commonly employed neural network architecture. It can be used to construct standalone networks as well as segments of considerably more intricate models, which will be discussed in detail later. The structure of the MLP is delineated by its design,

which comprises an input layer, one or more hidden layers, and an output layer. The network is completely connected, indicating that each unit obtains connections from all units in the prior layer. This implies that each unit has its own bias, and there is a weight associated with every pair of units in two consecutive layers. Therefore, the calculations for the l'th hidden layers of the network can be articulated as:

$$h_i^{(1)} = \theta^{(1)}\left(\sum_j w_{ij}^{(1)} x_j + b_i^{(1)}\right),$$

$$h_i^{(2)} = \theta^{(2)}\left(\sum_j w_{ij}^{(2)} h_j^{(1)} + b_i^{(2)}\right),$$

...

$$y_i = \theta^{(l)}\left(\sum_j w_{ij}^{(l)} h_j^{(l-1)} + b_i^{(l)}\right),$$

where the x_j are the inputs to the unit and the $w_{ij}^{(l)}$are the weights, the $b_i^{(l)}$ is the bias and $\theta^{(l)}$ is the nonlinear activation function respected to layer l. Moreover, the y_i is the activation of the output unit. It is important to recognize that each hidden layer may utilize distinct activation functions; nevertheless, the Rectified Linear Unit (ReLU) is currently the most prevalent activation function employed in hidden layers:

$$f(x) = ReLU(x) = max(0, x).$$

Concerning the output layer (y_i), pertaining to the k-class classification problem, the softmax function determines the output probabilities:

$$y_i = \frac{e^{h_i}}{\sum_{j=1}^{k} e^{h_j}} \; for \; i = 1,2, \dots, k,$$

where k is the number of classes and h_i is the output from the last layer before applying softmax. When the probability y has been established, L the difference between the predicted output and the expected value is evaluated by the loss function L.. In classification tasks, the latter is realized via cross-entropy loss:

$$L_{cross-entropy} = -\sum_{i=1}^{k} y_i log\big((\hat{y}_i)\big),$$

where $\hat{y}_i$ is the predicted probability for class i. For the purpose of optimizing the loss function (L), the gradient descent optimization algorithm is utilized to determine the ΔL:

$$\Delta w_i = -\alpha \frac{\partial L}{\partial w_i}$$

where α is the learning rate. Using the generated loss value, the chain rule is applied to updating individual weights:

$$w_{new} = w_{old} - \alpha \frac{\partial L}{\partial w},$$

where $\frac{\partial L}{\partial w}$ is derivative of loss with respect to weight.

(b) Convolutional Neural Network (CNN) [26]:

A Convolutional Neural Network (CNN) represents a category of artificial neural networks that excels in processing structured grid data, such as images. A widely adopted form of CNN, resembling the MLP, is characterized by several convolutional layers that are succeeded by sub-sampling (pooling) layers, culminating in fully connected layers at the end.

The input X for each convolutional layer is represented as a 3D tensor, encompassing values for height, width, and depth. The depth, often referred to as the channel number, is three in the case of an RGB image, and one for a grayscale image. A convolutional layer contains a set of K learnable filters (kernels) Kwhich process the input image to create feature maps. The label 'feature map' refers to the representation of the occurrence of certain features in an image, for instance, straight lines, edges, or distinct objects. The output of a convolutional layer can be articulated as follows:

$$Z = X * K + b,$$

where X is the input image, K is the filter (kernel) which performs the convolution operation and b is the bias term. In addition, after the convolutional operation an activation function is used to introduce non-linearity (ReLU). Afterward, pooling layers are typically employed, aiming to downsample the feature maps created by the convolutional layer into a smaller quantity of parameters, consequently lowering computational complexity and improving management of overfitting. The most widely used pooling operation is referred to as max pooling, which is defined as:

$$P_{i,j} = max_{m,n} Z_{i+m,j+n},$$

where m and n define the pooling window size. After the image undergoes the feature-learning procedure utilizing convolutional and pooling layers, the result from the last pooling layer is converted into a vector and then directed through one or more fully connected layers (MLP). In classification tasks, the final output probabilities are obtained by utilizing the softmax function. Analogous to MLP, the backpropagation process is utilized to enhance the loss function, which is mainly cross-entropy in classification scenarios. It is essential to underscore that the optimization process relates to both the weights of the fully connected layers and the filters employed in the convolutional layers, including the biases in those types of layers.

(c) Recurrent Neural Network (RNN) [27]:

Recurrent Neural Networks (RNN) are a type of neural network architecture mainly used for detecting patterns in sequential data, including handwriting, genomes, text, or numerical time series that are commonly generated in industrial environments. Unlike MLP, which transmit information in a unidirectional manner without cycles, RNNs incorporate cycles that allow them to relay information back into their own structure. This capability enhances the performance of Feedforward Networks by integrating previous inputs.

The essential part of an RNN is the recurrent layer, which, unlike feedforward networks that process all inputs at once, handles one input at a time for each time step. This sequential processing allows the network to maintain a dynamic that changes over time. At every time step t, the RNN takes an input x_t and updates its hidden state h_t, relying on the previous hidden state h_{t-1} and the current input. Mathematically, this update can be defined as:

$$h_t = f(W_h h_{t-1} + W_x x_t + b),$$

where W_h is the weight matrix for the hidden state, W_x is the weight matrix for the input, b is the bias vector and f is an activation function (e.g. ReLU). The output at each time step can be computed as:

$$y_t = W_y h_t + b_y,$$

where W_y is the weight matrix for the output layer and b_y is the output bias. Similarly to MLP and CNN, the RNN loss function quantifies the disparity between the predicted and expected outputs, however cross-entropy loss also accounts for the total number of time steps T:

$$L_{RNN} = -\sum_{t=1}^{T}\sum_{i=1}^{k} y_{t,i} log\ (\hat{y}_{t,i}),$$

where k is the number of classes, $y_{(t,i)}$ is the true label, and $\hat{y}_{(t,i)}$ is the predicted probability for class i at time step t. The architecture of RNNs incorporates a specialized backpropagation method termed Backpropagation Through Time (BPTT) [34], which is utilized to optimize the loss function and subsequently update the weights and biases. This approach involves unrolling the RNN temporally, thus treating it as a feedforward network for the duration of the sequence. Each timestep of the unrolled recurrent neural network can be seen as an extra layer due to the order dependency of the issue, with the internal state from the preceding timestep serving as input for the succeeding timestep.

(d) Gated Recurrent Units (GRU) [28]:

Gated Recurrent Units, as proposed by Cho et al. [28], are a form of RNNs that incorporates gating mechanisms to facilitate better information flow management and to mitigate the challenges of vanishing and exploding gradients when learning long-term dependencies. The structure of a GRU merges the hidden state and cell state into a single state and includes two gates: the update gate and the reset gate. Thus, the functionality of a GRU realized via the following operations:

- Update Gate: $z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$;
- Reset Gate: $r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$;
- Candidate Activation: $\tilde{h}_t = tanh\ (W_h x_t + U_h(r_t \odot h_{t-1}) + b_h)$;
- Hidden State Update: $h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$;

where σ is the sigmoid activation function, $\odot$ denotes element-wise multiplication, W is the weight matrix for the input states, U is weight matrix for hidden states and b are bias vectors.

Gating mechanisms play a crucial role in preserving significant information across lengthy sequences with enhanced efficiency. To begin with, by selectively permitting pertinent information to pass through the gates, GRUs mitigate the risk of gradients vanishing completely. Furthermore, owing to the effective gating mechanisms, GRUs can frequently be trained more rapidly than conventional RNNs, thereby enhancing effectiveness and decreasing the number of necessary training iterations.

(e) Long Short-Term Memory (LSTM) [27]:

Long Short-Term Memory (LSTM) introduced by Sepp Hochreiter and Jurgen Schmidhuber [27] is an advanced form of RNN, similar to GRU, designed specifically to resolve the vanishing gradient problem. However, LSTM is more widely used than GRU because of its superior performance in tasks that require long-term memory, particularly in the area of natural language processing. LSTM can be differentiated from GRU by its more elaborate architecture, which consists of three gates: an input gate, a forget gate, and an output gate, compared to the two gates present in GRU (update and reset gates). The hidden state h_t and cell state c_t of LSTM are updated using the following operations:

- Forget Gate: $f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$;
- Input Gate: $i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$, $\tilde{c}_t = tanh\ (W_c x_t + U_c h_{t-1} + b_c)$;
- Cell State Update: $c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t$;

- Output Gate: $o_t = \sigma(W_o x_t + U_o h_{t-1} + b_0), h_t = o_t \odot tanh\,(c_t)$;

where σ represents the sigmoid activation function, $\odot$ signifies element-wise multiplication, W denotes the weight matrix associated with the input states, U refers to the weight matrix for hidden states and b indicates the bias vectors.

The design of all models was analogous in terms of their layers, starting with the input layer, followed by a specialized network layer (MLP, CNN, RNN, GRU, or LSTM) that included 12 neurons and employed a ReLU activation function, and subsequently applying Dropout at a rate of 10%. The following layer was another specialized network, this time made up of 6 neurons and also utilizing a ReLU activation function. Again, a Dropout layer was applied at a rate of 10% to help mitigate the risk of overfitting. The network architectures concluded with a dense layer that contained two neurons and a softmax activation function. The shape of the input layer was consistent across MLP, RNN, GRU, and LSTM, which was (2,12) – covering information from both conditions (air on bone) across a range of 12 frequencies. Additionally, the CNN input layer required an extra dimension to accommodate the number of colors, resulting in a shape of (2,12,1).

Initial investigations evaluated Multilayer Perceptron (MLP), Convolutional (CNN), and Recurrent (RNN) neural networks, yielding accuracy rates of 94.58%, 95.63%, and 96.04%, respectively. The RNN architecture demonstrated the highest classification performance, prompting further exploration of RNN-based architectures, including Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM). Both models exhibited comparable accuracy, achieving 97.71% for GRU and 98.12% for LSTM. Furthermore, the confusion matrices and ROC curve with the AUC parameter were analyzed for each model, revealing that the LSTM model attained the best scores across all metrics.

In conclusion, this study proposed an AI-driven method for classifying audiograms as either normal or pathological, aimed at supporting referral decisions in primary care. Using a dataset of 2,400 expert-labeled pure audiometry data, the authors trained and compared MLP, CNN, and RNN models. The LSTM-based RNN achieved the highest accuracy (98.12%), which meets the predetermined margin of error standards and surpasses the 97.5% classification accuracy of the leading algorithm for audiogram data classification, as proposed by Crowson et al. [21]. It is important to highlight that this study focused solely on binary classification, whereas Crowson et al. [21] provided a methodology for the complete four-class classification.

3.2.1. Author's contribution to the state of the art

This section summarizes paper (P1) in the context of author's contribution to the state of the art in the area of automated classification of pure tone audiometry data. The paper contributes in the following subjects:

- ✓ While previous studies have explored aspects of audiogram interpretation, this work delivers a complete end-to-end binary classification system (normal vs pathological) using machine learning.
- ✓ This study is among the first to apply RNN architectures (LSTM, GRU) specifically to audiogram data, treating hearing thresholds across frequencies as sequential patterns.
- ✓ The paper demonstrates that temporal dependencies in audiometric patterns can be effectively captured using RNNs, outperforming traditional MLPs.
- ✓ The research achieves a 98% accuracy rate with LSTM, setting a high benchmark for audiometry data classification compered to state-of-the-art.
- ✓ The paper positions the developed AI model as a referral support tool in the clinical workflow, aiming to support GPs in low-resource settings and accelerate diagnosis.

3.3. Classification of three types of hearing loss

This section provides a summary of the papers (P2, P3) concerning the classification of three types of hearing loss. Section 3.2.1 outlines the findings from conference paper P2, titled "Detecting Types of Hearing Loss Using Various AI Classification Methods: A Performance Review," whereas section 3.2.2 presents the results from the extended conference paper published as a journal article entitled "Efficiency of Artificial Intelligence Methods for Hearing Loss Type Classification: an Evaluation". The research workflow of the papers P2 and P3 is presented in Figure 4.

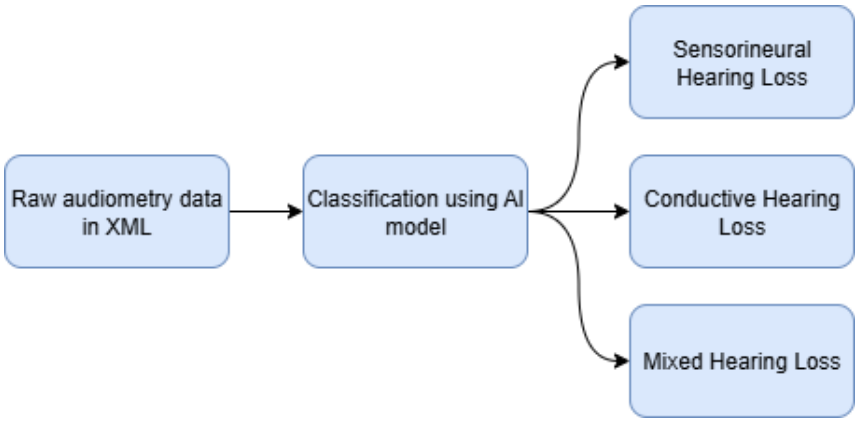


Fig. 4. Research workflow of papers P2 and P3.

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

The paper P2 presents a comprehensive examination of the ways in which artificial intelligence (AI), especially machine learning and deep learning methodologies, can be employed to categorize various types of hearing loss (mixed hearing loss, conductive hearing loss and sensorineural hearing loss) using raw pure tone audiometry data. The impetus for this study arises from the increasing need for rapid, precise, and economical diagnostic instruments that can assist audiologists in more effectively recognizing hearing deficiencies.

In the initial phase, the listed below machine learning classification algorithms were evaluated.

(a) Gaussian Naive Bayes [29]:

The foundation of the Gaussian Naive Bayes is Bayes' theorem, which can be expressed as

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)},$$

Where $P(C|X)$ is the probability of class C given the feature vector X, $P(X|C)$ is the probability of observing features X given class C, $P(C)$ is the prior probability of class C, $P(X)$ is the total probability of observing features X.

In the Naive Bayes classifier, it is presumed that the features are independent.

$$P(X|C) = P(x_1, x_2, \dots, x_n|C) = P(x_1|C) \cdot P(x_2|C) \cdots P(x_n|C),$$

where $x_1, x_2, \dots, x_n$ are the individual features of the vector X.

Also, we operate under the assumption that the features conform to a normal distribution. The likelihood of feature x_i belonging to class C can be represented by the probability density function of the normal distribution:

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}},$$

where $\mu$is the mean of the feature in class C and σ^2 is the variance of the feature in class C.

To classify a new feature vector X, the probability for each class C_k is computed.

$$P(C_k|X) \propto P(C_k) \cdot P(X|C_k),$$

$$P(X|C_k) = \prod_{i=1}^{n} P(x_i|C_k).$$

In the end, the class C_k with the highest likelihood is selected as the final candidate:

$$\hat{C} = argmax_{C_k} P(C_k|X).$$

(b) K-Nearest Neighbors (KNN) [30]:

The k-NN algorithm categorizes a new instance by considering the predominant class among its $k$closest neighbors within the feature space. Steps of the k-NN Algorithm:

- Determine the number of nearest neighbors k, that should be considered for the purpose of classification.
- To determine the distance for a particular test instance x, assess the distance between x and all training instances x_i, employing techniques like Euclidean Distance ($d(x, x_i)$).

$$d(x, x_i) = \sqrt{\sum_{j=1}^{m} \left(x_j - x_{ij}\right)^2},$$

where *m* is the number of features, and x_{ij} is the j-th feature of the i-th training instance.

- Sort the distances $d(x, x_i)$ and select the k training instances with the smallest distances donated as $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$.
- Determine the class labels of the k nearest neighbors. The predicted class label *C*(*x*) for the test instance *x* is given by:

$$\hat{C}(x) = argmax_c \left(\sum_{i=1}^{k} I\left(C\left(x_{(i)}\right) = C\right) \right),$$

where I is the indicator function that equals 1 if the condition is true and 0 otherwise.

(c) Logistic Regression [31]:

Logistic regression delineates the relationship between a binary dependent variable and one or more independent variables by means of the logistic function. The model's output is a probability value that ranges from 0 to 1, which can be understood as the probability of the input belonging to a certain class.

The primary step in logistic regression is to determine a linear combination of the input features. For a given input vector $x = [x_1, x_2, \ldots, x_n]$, the linear combination can be expressed as:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \beta^T x,$$

where β_0 is the intercept (bias term), $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients (weights) for each feature and β is the vector of coefficients.

Subsequently, the result of the linear combination z is transmitted through the logistic (sigmoid) function to generate a probability:

$$P(Y = 1|x) = \sigma(x) = \frac{1}{1 + e^{-z}},$$

where $P(Y = 1 \vee x)$ is the probability that the output Y is 1 given the input x $\sigma(z)$ is the logistic function.

In order to reach a classification decision, a threshold T (typically set at 0.5) is utilized on the predicted probability.

$$\hat{C}(x) = \begin{cases} 1, & P(Y = 1|x) \geq T \\ 0, & otherwise \end{cases}$$

To effectively train the logistic regression model, optimizing the coefficients β is necessary. The cost function utilized is the log loss (cross-entropy loss), which assesses the difference between the predicted probabilities and the true class labels:

$$J(\beta) = \frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} log\ (P(Y = 1|x^{(i)})) + (1 - y^{(i)}) log\ (1 - P(Y = 1|x^{(i)})) \right],$$

where *m* is the number of training examples, $y^{(i)}$ is the actual label for the *i*-th example and $x^{(i)}$ is the i-th training example.

(d) Support Vector Machines (SVMs) [32]:

The essential idea of Support Vector Machines (SVMs) is to discover a hyperplane that most effectively separates the data points of various classes in the feature space. The hyperplane is selected to ensure that the margin is as large as possible. The margin is defined as the space between the hyperplane and the nearest data points from either class, which are referred to as support vectors.

When addressing a binary classification problem, consider a dataset that contains n training examples, with each example depicted as a feature vector x_i and a corresponding label y_i (where $y_i \in \{-1,1\}$). The aim is to determine a hyperplane defined by the equation:

$$w^T x + b = 0,$$

where w is the weight vector (normal to the hyperplane) and $b$is the bias term. The decision function for classifying a new instance x is given by:

$$\hat{C}(x) = sgn(w^T x + b),$$

$$sgn = \begin{cases} -1, x < 0 \\ 0, x = 0 \\ 1, x > 0 \end{cases}$$

In order to optimize the margin, it is essential to accurately classify the data points and maximize the distance from the hyperplane to the closest points, known as support vectors. The margin γ can be defined as:

$$\gamma = \frac{2}{\|w\|}.$$

In order to identify the optimal hyperplane, it is essential to minimize the norm of the weight vector $\|w\|$while ensuring that all training examples are accurately classified:

$$\bigvee_i y_i\ (w^T\ x_i + b) \geq 1.$$

To address the aforementioned quadratic programming issue involving inequality constraints, the method of Lagrange multipliers can be employed. Consequently, the Lagrange function is defined as follows:

$$L(w, b, \alpha) = \frac{1}{2}w^2 + \sum_{i=1}^{n} \alpha_i (y_i(w^T x_i + b) - 1),$$

where a_i are the Lagrange multipliers.

(e) Stochastic Gradient Descent (SGD) [33]:

The objective of the SGD classifier is to reduce a loss function $L(w)$, which evaluates the model's effectiveness in classifying the data. For a given dataset $\{(x_i, y_i)\}_{i=1}^{n}$, where x_i is the feature vector and $y_i$is the class label, the loss function can be defined as:

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} l\big(y_i, f(x_i; w)\big),$$

where *l* is the loss function and $f(x_i; w)$ is the model function with parameters w.

In the next step the gradient of the loss function *L*(*w*) with respect to the parameters *w* is computed:

$$\nabla L(w) = \frac{\partial L(w)}{\partial w}.$$

Within the framework of the SGD algorithm, the modification of the parameters w occurs based on an individual example as demonstrated here:

$$w \leftarrow w - \eta \nabla l\big(y_i, f(x_i; w)\big),$$

where *η* is the learning rate and $\nabla l(y_i, f(x_i; w))$ is the gradient of the loss function.

The algorithm carries out the aforementioned procedures for all examples in the dataset over a series of epochs. In each epoch, the examples are shuffled in a random manner, which introduces randomness into the update process and helps to prevent local minima.

(f) Decision Tree [34]:

The Decision Tree classification algorithm creates a model that resembles a tree structure, where every internal node signifies a decision based on a specific feature, each branch illustrates the outcome of that decision, and each leaf node corresponds to a class label. For a given dataset

$$D = \{(x_i, y_i)\}_{i=1}^{n},$$

where each instance is represented by a feature vector *x* and a corresponding class label *y*. The main aim of the Decision Tree algorithm is to divide the dataset D into subsets that are as uniform as possible in relation to the target class. The frequently used criteria for evaluating the impurity of a node is entropy. The entropy of a dataset *D* is given by:

$$Entropy(D) = -\sum_{i=1}^{C} p_i log_2(p_i).$$

Subsequently, the Information Gain is calculated, which quantifies the decrease in entropy or impurity following the division of a dataset based on a feature. This metric aids in identifying the optimal feature for splitting at each node. The formula for Information Gain is:

$$IG(D,A) = Entropy(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} Entropy(D_v),$$

where $IG(D,A)$ is the information gain of dataset *D* when splitting on attribute A, $Values(A)$ are the possible values of attribute A, D_v is the subset of D for which attribute A has value *v*, D V is the total number of instances in dataset *D* and D_v V is the number of instances in subset D_v.

The Decision Tree algorithm chooses the feature that provides the highest information gain for splitting the dataset. This recursive process is carried out for each subset until a stopping criterion is reached (for example maximum depth of the tree).

(g) Random Forest [35]:

The Random Forest classifier can be regarded as an advanced or ensemble variant of the Decision Tree algorithm. Whereas a solitary Decision Tree generates predictions through a sequence of data splits, the Random Forest constructs numerous Decision Trees and amalgamates their results to enhance overall efficacy. This technique is called bootstrapping, which produce various subsets of the training data. For each tree t in the forest, a bootstrap sample D_t is generated by randomly drawing n instances from D with replacement:

$$D_t = \{(x_{i_1}, y_{i_1}), (x_{i_2}, y_{i_2}), \dots, (x_{i_n}, y_{i_n})\},$$

where i_j are randomly selected indices from the original dataset D. In constructing each decision tree, a random subset of features is chosen for each split. Let m denote the total number of features, and a subset of m′ features is selected, with m' being less than m. The number of features considered at each split is a hyperparameter indicated as m′. For each individual tree t, the algorithm develops a decision tree T_t through the use of the bootstrap sample D_t and the selected features. The criterion for splitting can be founded on entropy. Once all trees T_t are built, predictions for a new instance x are made by aggregating the predictions from all trees:

$$\hat{C} = mode(T_1(x), T_2(x), \dots, T_T(x)),$$

where *T* is the total number of trees in the forest. The mode is calculated by determining which class label appears most frequently among the predictions from all the trees.

The second phase of the research focused on assessing the following Artificial Neural Network (ANN) architectures: Feedforward Neural Network (FNN/MLP), Convolutional Neural Network (CNN) Recurrent Neural Network (RNN). Furthermore, an analysis was conducted to evaluate the performance of the Graph Neural Network (GNN) [36] on audiometry data. GNNs represent a category of neural networks specifically engineered to function on graph data, necessitating the transformation of input data into a graph $G = (V, E)$, where V is the set of nodes (vertices) and E is the set of edges connecting the node. Each node $v_i \in V$ can have an associated feature vector$x_i \in R^d$,where *d* is the dimensionality of the feature space. The essential function in GNNs is the message passing mechanism, consisting of two key phases: message aggregation (nodes gather and integrate messages from adjacent nodes) and node update (nodes modify their representations utilizing the aggregated messages). Typically, the message passing procedure is reiterated T times according to the following methodology:

Firstly, for each node v_i, messages are relayed from its neighbors $N(i)$. The message from neighbor v_j directed to v_i can be articulated as follows:

$$m_{ij}^{(t)} = Message\left(x_i^{(t)}, x_j^{(t)}, e_{ij}\right),$$

where $m_{ij}^{(t)}$ denotes the message transmitted from node v_j to node v_i during iteration *t,* and $Message$ is a function (e.g. element-wise sum, mean or max) that processes the message based on the characteristics of the nodes and the edge. Thus, for node v_i the message is computed as:

$$m_i^{(t)} = \sum_{j \in N(i)} m_{ij}^{(t)}.$$

On the basis of these aggregated messages, the GNN layer updates the features of source node i. At the conclusion of this update process, the node ought to be aware of both its own characteristics and those of its neighbouring nodes. This is achieved by integrating the feature vector of node i with the aggregated messages. Thus, the update function U can be defined as:

$$x_i^{(t+1)} = U\left(x_i^{(t)}, m_i^{(t)}\right),$$

The outputs can serve multiple downstream purposes, such as classifying nodes or graphs and predicting edges.

The models were developed utilizing a dataset of 4007 rows of audiometry data, which was categorized by experienced audiologists. The input data series comprised of vertical information regarding tonal points of both air and bone conduction, represented as volume (dB) for specific frequencies (Hz), sourced from XML files. The frequency spectrum of the dataset encompassed 250Hz, 500Hz, 1000Hz, 2000Hz, and 4000Hz.

Given that GNN necessitates graph input, the audiometry data was converted into a directed graph comprising 10 nodes and 18 edges. Frequency and loudness values were allocated to the nodes and the classification of hearing loss types in GNN was done in graph-level. Figure 5 illustrates a visual representation of the graph.

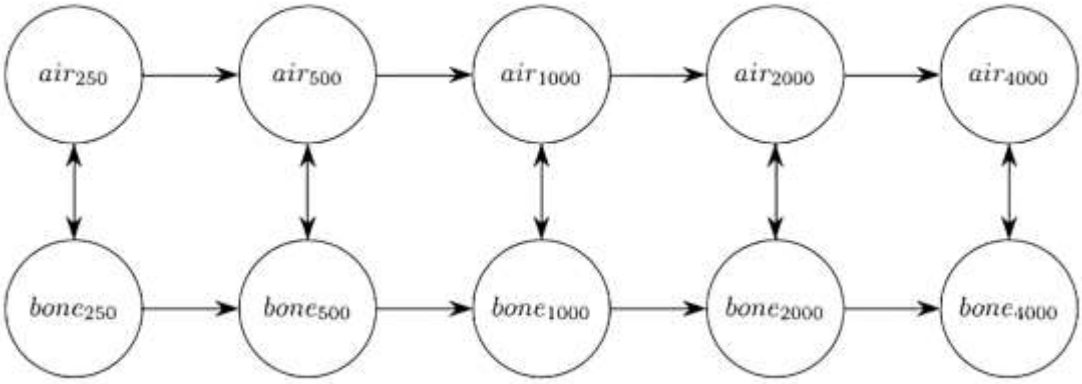


Figure 5. The GNN architecture's input graph structure [3].

The architectures of the developed artificial neural networks followed the general structure outlined in paper P1 (section 3.2), with the main difference being the shape of the input layer. The latter was structured as (5, 2), since this research examined the five primary frequencies (250, 500, 1000, 2000 and 4000 Hz) for both conductions.

All tested methods have been implemented in Python 3.10 in Jupyter Notebook environment. The implementations of machine learning classification algorithms have been imported from the scikit-learn

module [37]. The neural networks have been implemented in Keras/Tensorflow [38]. Due to the vast number of algorithms and neural network architectures examined, computational resources from the Centre of Informatics - Tricity Academic Supercomputer & Network (CI TASK) were employed to train the models.

In the realm of machine learning algorithms, the Support Vector Machine classifier has demonstrated the most impressive results, achieving an accuracy of 83.38%. This algorithm also excelled in metrics such as precision, recall, F1 score and AUC. Following closely behind, the Logistic Regression and Random Forest models also surpassed the 80% accuracy threshold. Among the artificial neural network models evaluated, the RNN emerged as the top performer, attaining an accuracy of 94.46% and a F1 score of 94.45%, excelling in precision, recall, and AUC as well. The CNN model ranked second, with an accuracy of approximately 93.46%, which may come as a surprise since CNNs are typically utilized for image analysis. This phenomenon was attributed to the fact that CNNs excel at extracting data and patterns from matrices, and a single audiometry test result could be interpreted as a small (5x2) matrix. The FFN model generally secured third place with an accuracy of 89.67%, while the GNN model recorded the lowest scores at 83.15%.

In conclusion, the study presented in paper P2 sought to evaluate various AI-driven algorithms for the classification of discrete tonal audiometry data series into three categories of hearing loss: sensorineural, conductive, and mixed. The Recurrent Neural Network achieved the highest classification accuracy, reaching 94.46%. Although multiple AI models demonstrated encouraging outcomes, no single approach consistently surpassed the others across all situations. Consequently, additional efforts were required to focus on enlarging the dataset and enhancing RNN models with respect to accuracy.

3.3.2. Efficiency of Artificial Intelligence Methods for Hearing Loss Type Classification

The article P3 serves as an extension of the conference paper P2. This study has been broadened to incorporate several new AI models and to deliver a more comprehensive evaluation of the employed deep learning algorithms, which includes an analysis of the influence of different data preprocessing techniques on the classification of hearing loss types. Additionally, the extended paper addresses the implications of augmenting the training dataset through the application of a generative adversarial network (GAN) [51]. The GAN consists of two neural networks, a generator G and a discriminator D, that are trained simultaneously through adversarial training. The GAN could be formally defined as:

Let X be the data distribution from which samples are to be generated. The goal of the GAN is to learn a mapping from a latent space Z to the data space X.

Generator *G* is a function that maps a random noise vector $z \in Z$ to a data sample $x \in X$:

$$x = G(z; \theta_G),$$

where θ_G are the parameters of the generator.

Discriminator D is a function that takes a data sample x and outputs a probability $D(x; \theta_D)$ that indicates whether the sample is real (from the data distribution) or fake (generated by *G*):

$$D(x; \theta_D) = P(D = 1|x),$$

where θ_D are the parameters of the discriminator.

The training process of GANs can be conceptualized as a two-player minimax game, in which the generator seeks to reduce the likelihood of the discriminator accurately identifying generated samples, whereas the discriminator strives to enhance its classification precision. The objective function can be articulated as:

$$min_G max_D V(D, G) = E_{x p_{data}}[log D(x)] + E_{z p_z}\left[log\left(1 - D\big(G(z)\big)\right)\right],$$

where p_{data} is the real data distribution, p_z is the distribution of the latent variable z, $E_{xp_{data}}[logD(x)]$ is expected value of the logarithm of the probability that real samples x are classified as real by the discriminator *D* and $E_{zp_z}\left[log\left(1-D\left(G(z)\right)\right)\right]$ is expected value of the logarithm of the probability that samples generated by *G* are classified as fake by the discriminator *D*.

The presented research utilized a conditional generative adversarial network (CGAN) [39], which represents a modification of the GAN framework that includes labels as extra data during the training stage. Consequently, this method resulted in an increase of the size of the dataset by a factor of two.

In a manner akin to P2, artificial intelligence and machine learning models have been applied to automatic classification of hearing loss types—conductive, sensorineural, or mixed—using pure-tone audiometry data based on 4,007 audiometry samples, each labeled by professional audiologists. Furthermore, a more comprehensive assessment was conducted focusing on advanced RNN models, specifically the Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM). Additionally, the research examined the effects of standard data preprocessing methods, including the normalization and scaling of audiometric features, on the ultimate accuracy value. The evaluation encompassed standardization techniques such as Z-Score (1), MinMax (2), and MaxAbs Scaler (3):

$$z_{score} = \frac{x-\mu}{\sigma}(1),$$

$$z_{minmax} = \frac{x-min}{max-min}(2),$$

$$z_{maxabs} = \frac{x}{|max|}(3),$$

where x is the raw score, μ is the mean, σ is the standard deviation, min is the minimum value of the feature and max is the maximum value of the feature. Furthermore, additional tests have been conducted after expanding the training dataset by means of a conditional generative adversarial network.

The efficacy of all evaluated models was measured using 10-fold cross-validation. Additionally, during the model evaluation process, the standard 10-fold set was reduced to 90%, leaving 10% to constitute a representative test dataset. Moreover, the separation into training and testing sets was conducted in a way that preserved the class proportions present in the complete dataset. This adjustment was made to facilitate a meaningful comparison of the performance between models trained with and without data generated by the CGAN.

In terms of machine learning methods, the application of CGAN yielded positive outcomes for only 5 out of the 7 algorithms that were examined in the dedicated test dataset. The generation of additional training data resulted in increasing the classification accuracy level in SVMs and logistic regression by approximately 5%. The largest increase, amounting to 8%, is shown in the SGD results as compared to those without CGAN. Table 3 provides a comprehensive overview of the results from this comparison.

The paper also discusses the influence of normalization strategies on the performance of classification tasks utilizing deep learning models. Clearly, the Z-Score normalization technique (1) exhibited exceptional performance across all evaluated architectures. The classification accuracy achieved with this method is, on average, 35% superior to that obtained with MinMaxScaler (2) and approximately 120% greater than the results yielded by MaxAbsScaler (3).

Algorithms	Default training (acc)	Training with CGAN (acc)
Gaussian Naive Bayes	63.09%	63.59% ⇧
K-Nearest Neighbors	80. 29%	79. 05% ⇩
Logistic Regression	89.77%	92.51% ⇧
Support Vector Machines	90.27%	93.04% ⇧
Stochastic Gradient Descent	79.55%	85.53% ⇧
Decision Trees	84. 53%	83.29% ⇩
Random Forest	87.78%	88.02% ⇧

Table 3. An analysis of the performance of machine learning models, both with and without the application of CGAN, evaluated on a specific test dataset [5].

Models	Default training (acc)	Training with CGAN (acc)
FNN	95.48%	91.66% ⇩
CNN	92.01%	88.19% ⇩
RNN	93.40%	94.44% ⇧
LSTM	94.79%	97.56% ⇧
GRU	92.70%	92.70% ⇔

Table 4. The performance comparison of deep learning models, trained with CGAN versus those without, as analyzed on a designated test dataset [5].

In terms of deep learning architectures, training on the expanded dataset has significantly increased the performance of certain deep learning models while impacting the performance of others. In particular, the classification accuracy of recurrent networks has increased by nearly 1% in the case of RNN, around 1.5% for GRU and nearly 3% for LSTM. On the other hand, the classification effectiveness of FNN and CNN has reduced by nearly 3%. Table 4 presents a detailed summary of the findings from the comparison of deep learning models, both with and without the use of CGAN. In conclusion, the most favorable outcomes were achieved using the Long Short-Term Memory model, which reached a peak classification accuracy of 97.56% through Z-Score normalization and CGAN data augmentation, which is similar results to state-of-the art of 97% [21]. Overall, all deep learning models demonstrated significantly superior classification performance compared to traditional machine learning algorithms.

3.3.3. Author's contribution to the state of the art

The papers (P2-P3) contribute in the following subjects:

- ✓ A systematic comparison of both traditional machine learning (e.g., Random Forest, SVM) and deep learning methods (CNN, RNN, LSTM, etc.) for the specific task of hearing loss type classification has been performed.
- ✓ Application of a Generative Adversarial Network (GAN) for augmentation in audiology data classification. This has shown to mitigate the common problem of small dataset sizes in medical AI, improving the generalization and robustness of deep learning models.
- ✓ The importance of selecting an appropriate method of data standardization has been investigated, revealing that Z-score standardization provides best results for audiometric data.
- ✓ The proposed LSTM model demonstrated a classification accuracy of 97.56%, aligning closely with the current state-of-the-art performance of 97%.
- ✓ The study demonstrates that AI models, particularly LSTM, can reliably assist or even automate the classification of hearing loss types, offering time-saving and accuracy benefits in clinical environments.

3.4. Full classification of hearing loss type

This section is a summary of journal article (P4) entitled „Automated hearing loss type classification based on pure tone audiometry data".

The article P4 details a deep learning approach that incorporates a Bi-LSTM model to classify hearing loss types - normal, conductive, sensorineural, and mixed - automatically, based on raw pure-tone audiometry data, with the purpose of aiding clinicians and general practitioners in diagnosis. The main workflow of the paper (P4) is presented in Figure 6.

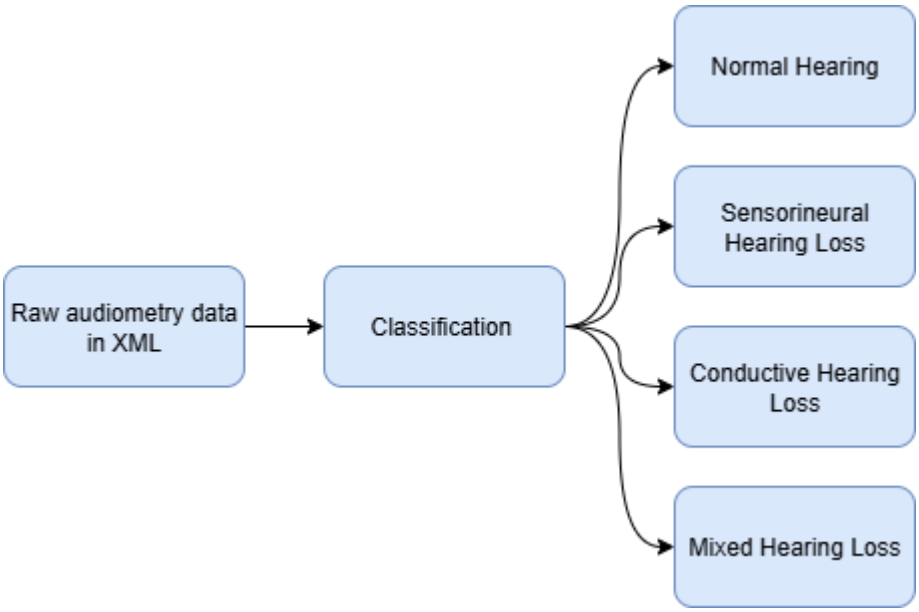


Fig 6. An overview of the workflow of paper P4.

The paper proposes a new data classifier model based on the Bi-LSTM architecture, which is a variant of Bi-RNN that utilizes two basic LSTMs to analyze input time series in both forward and backward orientations. The input layer, which has a shape of (7,2) – 7 timesteps of frequencies in both conductions - is followed by a Bi-LSTM layer with 7 neurons and a dropout layer, which helps prevent overfitting. The dropout layer is followed by a single LSTM layer with 4 neurons and another dropout layer. The final layer is a Dense one, which converts the input 492 parameters to one of four classification categories using the softmax function. An overview of the proposed architecture is shown in Figure 7.

The model has been trained on a total of 15,046 audiometry test results from 9,663 adult patients. The data for each individual measurement (one ear of one patient) comprised seven lists that represented air and bone conduction, with hearing levels quantified in decibels across frequencies of 125 Hz, 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz, and 8000 Hz, respectively. Three qualified audiologists classified the morphologies of hearing loss based on the audiometry test results, categorizing the data into four distinct classes: normal hearing, conductive hearing loss, mixed hearing loss and sensorineural hearing loss, in accordance with the methodology outlined in the paper. This resulted in 2584 (17.17%) normal samples, 657 (4.37%) samples of conductive hearing loss, 4028 (26.71%) samples of mixed hearing loss, and 7777 (51.69%) samples of sensorineural hearing loss.

Based on previous results in regard of audiometry data (P3), Z-score normalization has been applied to the training data and a system of class weight has been introduced to prevent unintended outcomes from occurring when processing unbalanced data.

In order to address the aforementioned class imbalance, the study employs stratified 10-fold cross validation. This method is an enhancement of standard 10-fold cross validation, tailored specifically for classification challenges where the proportion of target classes remains consistent across each fold as it does throughout the entire dataset.

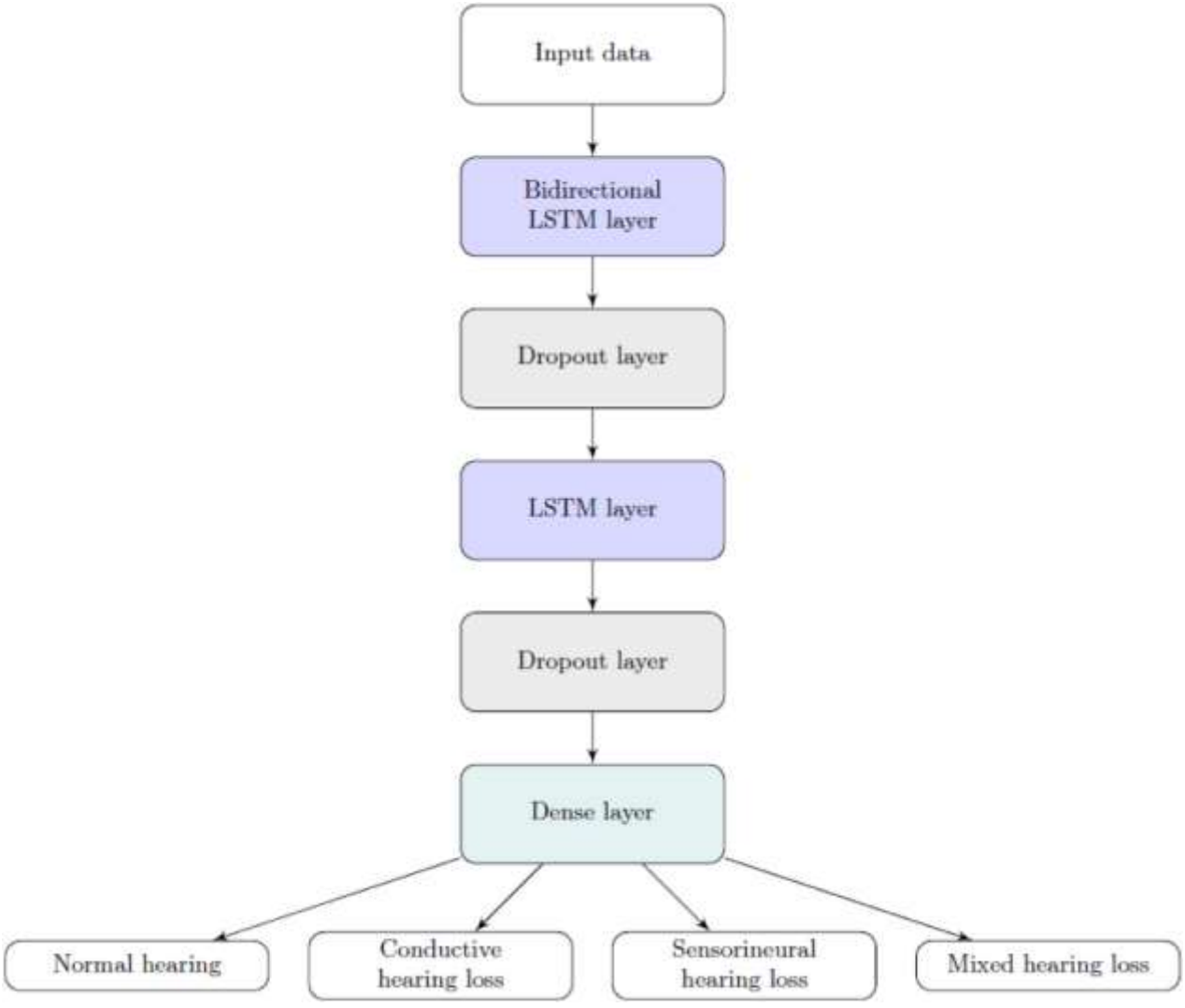


Fig 7. An overview of the proposed Bi-LSTM architecture [6].

The findings obtained from the stratified 10-fold cross-validation reveal that the proposed Bi-LSTM model successfully classified normal hearing, sensorineural hearing loss, conductive hearing loss and mixed hearing loss, achieving an average accuracy of 99.33%. The accuracy varied between 99.00 and 99.73 with

a standard deviation of 0.23%, demonstrating stability. Metrics such as precision, recall, and F1 score exhibited comparable trends in relation to accuracy. The proposed Bi-LSTM model significantly surpassed the current leading method in raw audiometry data classification, which is the C4.5 Decision Tree (DT-J48) method introduced by Elbaşı and Obali [20]. The application of the C4.5 classifier to the presented dataset yielded an accuracy level consistent with that reported in the original study, with a mean accuracy of 95.64% and a standard deviation of 0.69%. In the broader context of the audiogram classification problem, the overall accuracy of the proposed model (99.33%) exceeds that of the most effective existing method for hearing loss classification (97.5%), as presented by Crowson et al. [21] for raster data. While the numerical difference may not be substantial (99.33% compared to 97.5%), the results were derived from a considerably more representative dataset (15,046 in the paper versus 1007 samples in Crowson et al [21]).

In conclusion, this paper introduces a Bi-LSTM-based model designed to classify raw audiometry data into categories of normal hearing and three distinct types of hearing loss. This innovative solution enhances the classification of hearing loss types, surpassing the existing state-of-the-art methodologies. The findings indicate that the proposed neural network-based classifier for audiometry data holds potential for application in clinical settings, serving either as a classification tool for general practitioners or as a support system for professional audiologists.

3.4.1. Author's contribution to the state of the art

- ✓ The proposed model has achieved a classification accuracy of 99.33%, which surpasses the current state-of-the-art in raw audiometry data classification, as reported by Elbaşı and Obali [20], who achieved an accuracy of 95.5%.
- ✓ The proposed solution has also demonstrated superior performance compared to the existing state-of-the-art in raster audiogram classification, as presented by Crowson et al. [21], which attained an accuracy of 97.5%.
- ✓ This study was conducted on the largest and most varied tonal audiometry dataset to date, thus ensuring that the obtained classification results are representative of real-world performance..
- ✓ In contrast to previous methods that depend on audiogram images (Crowson et al. [21]), the proposed model utilizes raw air and bone conduction thresholds, enhancing interpretability, eliminating variability from different chart formats, and facilitating direct integration into audiometry equipment or hospital systems.
- ✓ The model introduces a bidirectional LSTM specifically designed to process the frequency-ordered characteristics of audiometric data, effectively capturing both local and long-range threshold patterns.
- ✓ The proposed approach grants professional audiologists the ability to utilize an AI decision support system, which may help decrease their workload, improve diagnostic accuracy, and lower the likelihood of human error.

3.5. Summary of pure-tone audiometry classification models

This section provides a summary of all AI-driven models utilized for the classification of audiometric data as detailed in articles P1-P4. In total, 15 distinct machine learning and deep learning models have undergone testing. Ultimately, these tests culminated in the creation of the 4-class Bi-LSTM model outlined in P4, which is designed for the classification of hearing loss types, including normal hearing.

K-fold cross-validation served as the primary method for assessing classification models. The evaluation metrics taken into account included: accuracy, precision, recall, F1 score, ROC curve with AUC score and confusion matrices.

The nature of the error derived from the error matrix was also considered in the assessment of the models. Given that the objective of the study was to implement the findings in a medical context, the main emphasis was on eradicating the error that could lead to a patient receiving unsuitable medical treatment due to an incorrect classifier outcome. In the context of the audiometric test evaluation issue, this pertains to a situation where a patient with any form of hearing impairment is misclassified by the model as having normal hearing.

Table 5 presents the details of all proposed models in P1-P4 regarding classification of pure tone audiometry data in the context of the state-of-the-art.

Model	Authors	Data type	Dataset size	Classification problem	Accuracy	Results Published
C4.5 (decision tree)	Elbaşı and Obali	Raw audiometry data	200	Hearing types: normal, conductive, mixed and sensorineural	95.5%	[20]
ResNet-101 (CNN)	Crowson et al.	Audiograms (raster data)	1007	Hearing types: normal, conductive, mixed and sensorineural	97.5%	[21]
LSTM	Kassjański et al.	Raw audiometry data	2400	Normal and hearing loss	98%	P1
RNN	Kassjański et al.	Raw audiometry data	4007	Hearing loss types: conductive, mixed and sensorineural	94.46%	P2
LSTM	Kassjański et al.	Raw audiometry data	4007	Hearing loss types: conductive, mixed and sensorineural	97.56%	P3
Bi-LSTM	**Kassjański et al.**	**Raw audiometry data**	**15046**	**Hearing types: normal, conductive, mixed and sensorineural**	**99.33%**	**P4**

Table 5: Summary of all proposed models considered in the (P1) — (P4) papers with state of the art comparison.

4. Processing of tonal audiometry data on mobile devices

This chapter is a summary of journal article (P5) entitled „Development and testing of an open source mobile application for audiometry test result analysis and diagnosis support".

The article presents a novel open-source Android application designed to aid clinicians in the analysis of audiograms and the diagnosis of hearing loss. Pure-tone audiometry, recognized as the clinical benchmark for assessing hearing, is represented through audiograms that necessitate expert interpretation to determine the type and severity of hearing loss. To enhance this process, an application that enables users to capture and analyse an image of a printed audiogram using a smartphone camera has been created. The workflow of the study (P5) is illustrated in Figure 8.

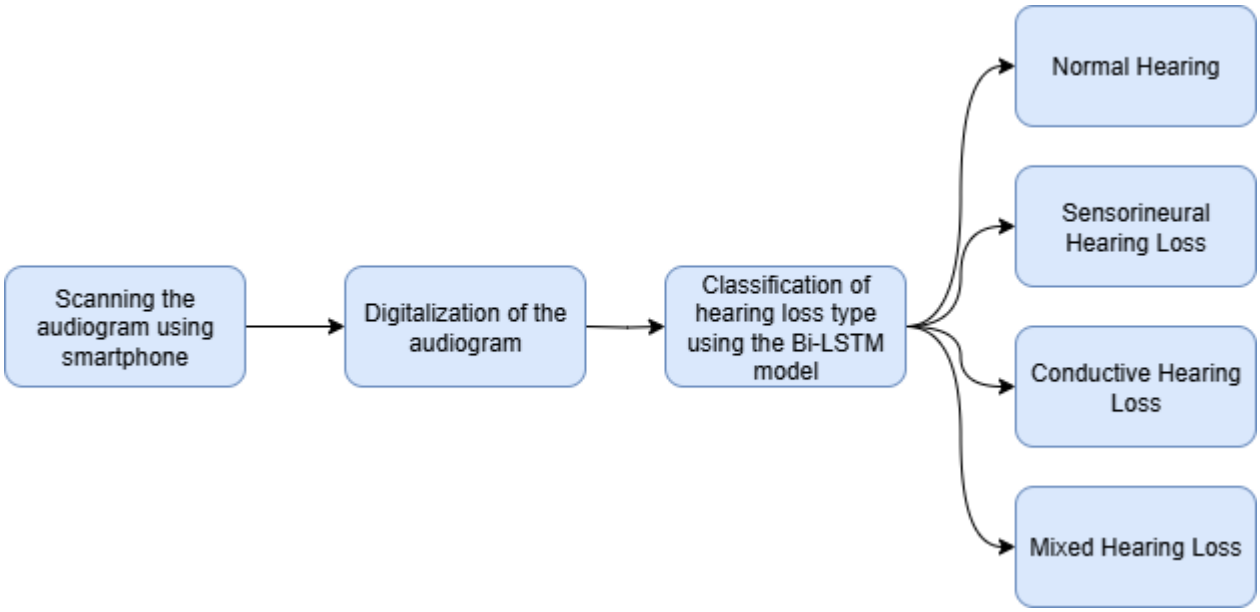


Fig 8. A summary of the research workflow of paper P5.

The processing and classification of pure-tone audiometry test results on mobile devices required their prior digitization and transformation from audiogram form. In this context, section 4.1 describes the state-of-the-art in audiogram data digitalization, while section 4.2 presents the process of developing the mobile application.

4.1. State-of-the-art in audiogram digitalization

The evaluation of tonal audiometry test outcomes is most precise when performed on raw audiometry data, thereby circumventing issues associated with the analysis of audiograms produced by various software. Furthermore, it mitigates potential errors that may occur during the generation and printing processes. Nevertheless, some clinical settings (such as a general practitioner's office) are limited to the printed results of the test, which limits the usability and efficiency of contemporary automated audiogram analysis models. In such cases, it is necessary to transition printed audiograms into a digital format. Currently, the literature that discusses this specific issue is confined to the publications listed below.

The initial research was performed by Li et al. [40], who designed various convolutional neural networks to extract audiograms, symbols, and axis labels from audiogram images. The synthesis of results from all models results in a digital representation of the audiogram. The system showed 98% accuracy on scanned images, while it reached 84% accuracy on images captured with a camera.

Following this, Chairh and Green [41] introduced a novel digitalization tool that employs YOLOv5 for the recognition of symbols and Tesseract for the identification of labels. The dataset included 3,200 reports, in

comparison to the 420 reports analyzed by Li et al. [40]. This study took into consideration all audiological symbols, including those obscured from air and bone conduction. The audiogram, axis label, and symbol models achieved mAP@0.5 scores of 84%, 34%, and 39%, respectively.

The most recent work was performed by Yang et al. [42], who presented a system similar to that of Chairh and Green [41], which includes a multi-stage integration of YOLOv5 models paired with an optical character recognition (OCR) model. The analysis focused on both pure tone audiometry and sound field testing. The accuracy rate at each stage was about 98%, based on 2,535 samples for audiogram detection and 2214 records for symbol detection.

In summary, the audiogram digitalization process can be outlined through two principal methodologies: the utilization of convolutional neural networks and the incorporation of YOLO together with OCR models. Recent innovations in this sector, as highlighted by Yang et al. [42], leverage the latter technique, achieving an accuracy rate close to 98%. The heightened accuracy illustrates the effectiveness of integrating YOLO and OCR technologies to enhance digitalization efforts, particularly in applications that require meticulous object detection and text recognition. This being said, none of these solutions were crafted with mobile device implementation in mind. It is vital to understand that mobile devices generally have reduced processing power compared to desktop or server environments. This difference can lead to longer inference times and may limit the complexity of the models, particularly sophisticated CNN models as indicated by Li et al. [40]. Moreover, mobile devices are limited by their RAM and storage capacity, which is a problem particularly in terms of Large CNN-based models which demand considerable amounts of memory. To deploy CNNs on mobile devices effectively, it is necessary to optimize models through techniques such as quantization, pruning, or the use of lightweight architectures such as MobileNet [43]. This optimization can be complex and may require specialized knowledge. As a result, the models proposed by Chairh and Green [41] and Yang et al. [42] are not suitable for direct implementation on mobile devices, as they utilize demanding YOLO and OCR architectures.

4.2. Mobile application for audiometry test result analysis

The methodology employed for the processing and classification of pure-tone audiometry test results on mobile devices was delineated into three distinct stages: scanning, digitization and classification of audiograms.

The scanning process was realized using the ML Document Scanner from Google's ML Kit, enabling the user to position their smartphone camera over the document for automated capture with perspective correction. Afterwards, the YOLOv5 object detection model has been applied to identify and extract the audiogram region from hearing test results report.

The procedure for digitizing an audiogram consists of three fundamental stages: line detection, symbol detection, and label detection.

The process of detecting lines on the audiogram was carried out through the Probabilistic Hough Transform approach [44], which is a modification of the classic Hough Transform. Before the application of the Hough method, the Canny Edge detection method [45] was implemented to derive an edge map from the images. The Canny edge detection algorithm includes five essential steps:

1. Noise Reduction. The input image is smoothed using a Gaussian filter to reduce noise and unwanted details.

Let $I(x, y)$be the input image, and $G(x, y)$ be the Gaussian filter. The smoothed image is obtained by convolving $I(x, y)$ with $G(x, y)$*:*

$$I'(x, y) = I(x, y) * G(x, y),$$

where Gaussian filter $G(x, y)$ is defined by the following formula:

$$G(x, y) = \frac{1}{2\pi\sigma^2} \cdot e^{\frac{-x^2-y^2}{2\sigma^2}},$$

where σ is the standard deviation of the Gaussian distribution.

2. Gradient Calculation. The gradient magnitude ($M(x,y)$) and direction ($\theta(x,y)$) are computed using the Sobel operator [46]:

$$G_x = \frac{\partial I'}{\partial x}, G_y = \frac{\partial I'}{\partial y},$$

$$M(x,y) = \sqrt{G_x^2 + G_y^2},$$

$$\theta(x,y) = tan^{-1}\left(\frac{G_x}{G_Y}\right).$$

3. Non-Maximum Suppression. The gradient magnitude is subjected to thresholding to suppress non-maximum values (T_1), which results in a binary image featuring edge candidates.

$$M'(x,y) = \begin{cases} M(x,y), & M(x,y) \geq T_1 \\ 0, & otherwise \end{cases}$$

4. Double Thresholding. Two thresholds (T_1 and T_2) are utilized on the edge candidates to classify them as strong (1) or weak edges (0).

$$E(x,y) = \begin{cases} 1, & M'(x,y) \geq T_2 \\ 0, & M'(x,y) \leq T_1 \end{cases}$$

5. Edge connection. The contour of the image's edge is associated with the strong edge as a reference. Upon connecting to the image edge's endpoint, a search is conducted for the edge point that can be continued in the weak edge, thereby obtaining the full edge information of the image.

Based on obtained edge points from Canny method, the Hough Transform method was used for line extraction. In the Hough Transform, a line in the Cartesian coordinate system can be represented in polar coordinates as:

$$r = xcos(\theta) + ysin(\theta),$$

where r denotes the perpendicular distance from the origin to the line, and θ represents the angle formed between the x-axis and the line that is perpendicular to the line being depicted. For every edge point (x_i, y_i) in the binary image derived from the Canny method, the associated values of r for a spectrum of angles θ are calculated:

$$r_i = x_i cos(\theta) + y_i sin(\theta).$$

This indicates that for every edge point, a sinusoidal curve is produced in the Hough space, with each θ representing a distinct line that may intersect the point (x_i, y_i).

Utilizing an accumulator array $A(r,\theta)$, the Hough Transform counts the number of points that correspond to each (r,θ) pair. The accumulator is initialized to zero:

$$\bigvee_{r,\theta} A(r,\theta) = 0.$$

For every edge point (x_i, y_i), the value r_i is calculated for a range of angles θ, and for each calculated pair (r_i, θ), the accumulator is increased by 1:

$$A(r_i, \theta) \leftarrow A(r_i, \theta) + 1.$$

After all edge points have been processed, the accumulator array will display peaks at sites where multiple points in the image relate to the same line in Hough space. To ascertain these peaks, a threshold T is applied:

$$If\ A(r, \theta) > T, \qquad then(r, \theta)\ is\ a\ detected\ line.$$

The parameters (r,θ) associated with the identified lines can be transformed back into the Cartesian coordinates corresponding to the lines in the initial image. The line can be represented in the slope-intercept format as:

$$y = \frac{-sin(\theta)}{cos(\theta)} x + \frac{r}{sin(\theta)}.$$

To enhance computational efficiency and maintain stable performance, an advanced variant of the traditional Hough Transform known as the Probabilistic Hough Transform has been applied. This method, rather than utilizing all edge points (x_i, y_i), randomly selects a subset of these points. Furthermore, the algorithm guarantees that the chosen points are adequately dispersed to accurately depict the overall edge structure. The Probabilistic Hough Transform is especially advantageous for real-time applications and minimizes computational complexity, which are a critical factor for mobile devices.

Moreover, the paper proposes a method that calculates the position of any undetected lines by leveraging the spatial coordinates of the two closest detected lines. In the simplest case, when there exist two parallel lines $y_1 = mx + b_1$ and $y_2 = mx + b_2$ an interpolated line y_p that is exactly between these two lines can be expressed as:

$$y_p = mx + \frac{b_1 + b_2}{2}.$$

In more complex scenarios, when it is necessary to determine the equation of a line derived from two lines that are spaced further apart, the sole distinction will involve calculating the y-intercept (b) while considering the number of steps (assuming the lines are to maintain equal distance from one another).

In the realm of symbol detection, the architecture of YOLOv5s was employed to accurately identify symbols on audiograms. A total of 8 distinct classes were established, each corresponding to various audiological symbols, including those from air and bone conduction from both ears, along with a masked version of the symbols. For label detection, Optical Character Recognition (OCR) technology was utilized, particularly the Machine Learning Kit Text Recognition v2 API developed by Google, alongside the fine-tuned YOLOv5s model. Fine-tuning refers to the act of altering a pre-trained model to make it suitable for a new, related task, which in this instance involved the detection of labels in audiograms. For this purpose, the first 10 layers of the original YOLOv5 model (trained on the COCO dataset) were frozen, while the rest have been retrained on 987 instances of the audiogram label data for 1000 epochs. The fine-tuning process employs transfer learning, leveraging the features that the model has already acquired from large datasets, leading to expedited training times and often improved outcomes. Fine-tuning is particularly effective when working with a limited amount of labeled data. Since the model has been trained on a large dataset previously, it can successfully leverage this knowledge for the new task, needing fewer examples to achieve a satisfactory level of performance.

In the domain of audiogram classification, the Bi-LSTM model, as specified in P4, has been implemented. The original Bi-LSTM model, which was developed using Keras, has been adapted into TensorFlow Lite format through the use of post-training quantization techniques. Overall, quantization is the method of using lower-bit representations instead of higher-bit representations for a specific real-valued number. For example, a continuous real number, which is usually represented as a 32-bit floating-point number, can be

approximated with a discrete representation like an 8-bit integer. In deep learning, model parameters, such as weights and biases, are initially stored as 32-bit floating-point numbers to enable high-precision calculations during the training phase. After the training process is completed, these parameters can be reduced to 16-bit floating-point or 8-bit integer representations. This reduction in precision leads to a decrease in the overall size of the model, thus improving its efficiency for deployment on mobile devices [47]. Moreover, all YOLOv5s models utilized during the digitalization phase also required optimization for mobile functionality. Optimization minimizes the computational burden of the model, thereby reducing latency during inference, which enables models to operate more swiftly and efficiently—an essential requirement for real-time applications. Furthermore, optimized models utilize less power, a factor that is especially critical for mobile devices dependent on battery longevity.

In the process of evolving mobile-optimized AI models, the 5-fold cross-validation method was utilized. The performance assessment of the YOLOv5 model was performed using the mean average precision (mAP) metric, a recognized benchmark for evaluating the success of object detection. The overall average precision (AP) is calculated by averaging the AP values derived at each IoU threshold outlined below:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i,$$

where AP_i is the average precision of each class and N is the total number of classes. In the paper the IoU threshold was set at 0.5.

The paper also presents a manual evaluation of the application's performance across devices with varying budget options. The complete system underwent testing using a collection of twelve audiograms, each differing in complexity, under two distinct lighting conditions (50 lux and 500 lux) across three separate devices. The app's performance on various smartphone cameras was evaluated by quantifying the number of audiogram lines that were not accurately detected by the application due to inadequate image quality. Furthermore, any detection errors noted during the testing process were utilized to assess the efficacy of the line interpolation techniques incorporated within the application.

The essential technical requirements for using the developed mobile application consist of a minimum Android version of 9, at least 6 GB of RAM, a minimum of 400 MB of storage capacity, and a camera sensor with a resolution of at least 12 MP. At the time of writing, new devices that meet these specifications can be obtained for under 100 USD, while used devices can be had for less than 50 USD. This makes the application viable for use even in low-income areas.

The proposed model for audiogram detection based on YOLOv5, reached an 99% mAP50. Relative to the findings of Chairh and Green [41], this model demonstrated a significant enhancement in mAP50, with an improvement of 15 percentage points. In contrast, Yang et al. [42] reported an exceptional 100% accuracy; however, their results did not address mAP50, making it difficult to perform direct comparisons.

In analyzing the symbol detection model, the performance metrics stand out, with a mAP50 of 98%. When these results are compared to those from alternative symbol detection models, the study by Chairh and Green [41] shows a significantly lower score of 39% mAP@50. Additionally, when evaluating the outcomes presented by Yang et al. [42], the accuracy achieved is closely aligned with that of the model in question(98.11%), although it should be noted that in this case the authors also failed to provide a clear statement regarding the mAP50 value.

In terms of the label detection, the proposed model achieved a mAP@50 of 99%. The comparative analysis shows that the performance metrics of the presented model greatly exceed those of Chairh and Green [41], whose model achieved 34% mAP@50. Furthermore, the study conducted by Yang et al. [42] takes a different approach from training a YOLO-based model, focusing entirely on an OCR system that reaches an accuracy close to 99%, with aligns with the 99% mAP@50 of the model presented. This being said, in the proposed application the label detection system integrates OCR results with those from YOLO for a more thorough label detection. The merging of these two models produces superior results under conditions of noisy images. This is particularly noticeable when the OCR does not recognize blurred labels, while the YOLO model manages to detect them.

The manual evaluation of the application revealed it to be fully functional across all tested devices. In more intricate scenarios, the line interpolation feature became progressively essential, yet the system continued to exhibit proficiency in accurate audiogram classification. Higher-end devices displayed superior performance in the Hough line detection system, however even devices from lower-cost options exhibited commendable results.

In conclusion, the research introduces a mobile application aimed at the thorough classification of hearing loss types by utilizing audiograms obtained via a smartphone camera operating on the Android platform. The application employs state-of-the-art techniques for the scanning, digitization, and classification of audiograms. The digitized audiograms are classified using the Bi-LSTM model from a prior study (P4). Moreover, the application showcased has the capacity to function as an accessible and detailed diagnostic support tool for physicians in clinical settings.

4.3. Author's contribution to the state of the art

- ✓ The pioneering app equipped with a fully on-device AI-enabled tool for the interpretation of pure-tone audiograms utilizing a smartphone. This strategy delivers real-time and fully offline diagnostic support, representing a crucial innovation for limited-resource areas.
- ✓ The application has been released under an open-source license to promote transparency, reproducibility, and worldwide collaboration. It allows for customization and adaptation to different clinical settings.
- ✓ A three-stage processing architecture - audiogram detection, audiogram digitalization and hearing loss classification - has been integrated into a seamless automated diagnostic workflow.
- ✓ An innovative technical pipeline has been developed for label detection, integrating the results of OCR and YOLOv5.
- ✓ The developed app was tested across different smartphones, including low-, mid-, and high-tier options, in several scenarios to validate performance consistency.
- ✓ The software responds to the rising global challenge of hearing loss, especially in locations that do not have access to specialized audiological services, facilitating the app to be utilized on devices available for under 50 USD.

4.4. Summary of audiogram classification in mobile app

This section is a summary of the mobile app allowing to classify audiograms described in (P5).

Outside of tonal audiometry laboratories, hearing evaluations are usually conducted through analysis of audiogram images. Thus, application of the state-of-the-art classification models referenced in publications P1-P4 e.g. in a general practitioners office necessitated the development of a tool extract audiometric data from a printed audiogram. Furthermore, the objective was to design a user-friendly tool that could be integrated into a medical setting, while ensuring patient confidentiality and eliminating the need for costly equipment to support the classification model. Consequently, in light of contemporary medical trends [48], the decision was made to create an application capable of swiftly scanning the test result report, extracting audiometry data and immediate interpretation of the findings. An additional factor considered was to perform all calculations on the mobile device itself, thereby avoiding the transmission of data to an external server and mitigating the risk of storing sensitive patient information.

An Android application meeting the requirements described above has been successfully developed under an open-source license, thus allowing further development. All code is available on GitHub [49].

The application was developed by introducing a three-step process: scanning, digitizing, and classifying the audiogram. All detection models were evaluated using standard metrics, and the application was subjected to thorough manual testing to assess its overall functionality. The findings unequivocally demonstrate that the application is appropriate for deployment even in low-income medical settings.

5. Summary and conclusions

This chapter concludes the dissertation based on a consistent series of five publications, which covered the title research on application of artificial intelligence algorithms for analysis of pure tone audiometry.

Section 5.1 summarizes the results of the presented research in terms of meeting the set research goals and verifying the research hypotheses formulated in Section 1.3. Additionally, the obtained results are briefly commented in the context of other results from related literature. The final Section 5.2 outlines potential areas for the further research in application of AI in the field of audiology.

5.1. Summary of research goals and conclusion

The research goal (G1), aiming for review of existing classification models of pure tone audiometry data and their possibility to be applied in medical settings, has been achieved and is covered primarily by publication (P1) and (P4). Given that only two publications exist concerning the classification of hearing loss types, a standalone review paper on this topic was not feasible. In summary, the first of the existing solutions, proposed by Elbaşı and Obali [20], classified raw audiometry data with an accuracy of only 95.5% and was evaluated on a small dataset comprising merely 200 samples. The second solution, presented by Crowson et al. [21], attained a satisfactory accuracy level of 97.5% using the ResNet-101 model. However, this was accomplished on a specific set of audiogram images. While the structure of audiograms is generally consistent, there can be notable variations between audiograms produced by different hardware and software configurations. In addition to differences in background and line colors, audiograms may also vary in the volume of information presented (e.g. they may provide data for one ear or both). Consequently, a universal approach to classifying tonal audiometry results cannot rely solely on an image classifier.

The goal (G2), which aimed to test different neural network architectures on raw audiometry data to develop a model for hearing loss type classification has also been successfully achieved. The related research has been published in papers (P1) — (P3), which describe results of testing 15 distinct machine learning and deep learning models. In summary, the best results have been obtained by models based on CNN and RNN architectures.

To goal (G3), which aimed to develop a deep learning model for hearing loss type classification accurate enough to allow its implementation in clinical settings was successfully achieved in (P4). In conclusion, the Bidirectional Long Short-Term Memory architecture has been developed and assessed for the purpose of classifying audiometry test results into four distinct categories: normal hearing, conductive hearing loss, mixed hearing loss, and sensorineural hearing loss. The network has been trained on 15,046 hearing test results that were analyzed and categorized by professional audiologists. The proposed model attains a classification accuracy of 99.33% on external datasets, meeting the accuracy requirements and showing an improvement over the 97.5% accuracy reported by Crowson et al. [21].

Finally, to the goal (G4), which aimed to create of a mobile application allowing for the use of the previously developed to classify the type of hearing loss from a photograph of audiometric test results has been successfully achieved in (P5). In summary, the application facilitates the scanning of hearing reports, automatically detects and separates audiograms, digitizes them utilizing YOLO, OCR, and image processing techniques. Subsequently, it employs the model introduced in (P4), which is optimized for mobile devices, to classify the scanned audiograms as either normal hearing or one of the three types of hearing loss.

As a result, it can be concluded that all the research goals of this dissertation have been successfully accomplished. Meeting all the research targets derived from the research hypotheses also permits the evaluation of the research hypotheses themselves.

Hypothesis H1 stated that “The application of modern neural network architectures to classification of hearing loss types based on audiometric data can push the state of the art and deliver performance and accuracy viable for introduction in clinical practice”. Validating this hypothesis is directly associated with the accomplishment of objective G3 and the publication (P4), where the proposed Bi-LSTM model realized an accuracy of 99.33%, in contrast to the 95.5% accuracy of the solution offered by Elbaşı and Obali [20] on pure audiometric data. Moreover, the proposed Bi-LSTM model exhibited a greater accuracy than that reported by Crowson et al. [21], which was 97.5% on raster data. In both cases, the results were derived

from datasets that were significantly smaller and likely less representative than that on which the proposed Bi-LSTM model was trained. To summarize, the deployment of advanced neural networks, notably those founded on RNN principles, **can enhance the present state of the art in terms of hearing loss types**, while the achieved accuracy makes the developed model **viable for introduction in clinical practice**

Hypothesis H2 declared that "Modern neural network architectures dedicated for processing raster and time-series data are capable of accurate classification of raw tonal audiometry test results". Establishing this hypothesis is intrinsically connected to objective G2 and the publications (P1 – P3). The novel methodology for the interpretation of pure audiometric data as a time series yielded surprisingly positive results. This was already demonstrated in publication P1, where the RNN model achieved a notably better performance (96% achieved by simple RNN) on the binary classification challenge than the more commonly employed feedforward network model for this data type (94%). Furthermore, the CNN, which converted the tonal audiometry results into a matrix with pixel values corresponding to the individual tonal points, showed a slightly inferior performance compared to the RNN at 95%. Similar relationships were observed in the more intricate evaluations of the algorithms and architectures discussed in (P2) and (P3). None of the machine learning models crossed the 86% accuracy threshold when the trained on the original tabular data structure. In contrast, altering the data structure yielded results of 95.63% for the LSTM network (when converted to time series) and 93.76% for the CNN network (when transformed to raster). In conclusion, advanced neural network architectures that are specifically designed to handle raster and time series data **can effectively classify raw tonal audiometry test results**.

Finally, hypothesis H3 stated that "It is possible to optimize modern neural network architectures to efficiently operate on smartphones which cost less than 100 USD, thus providing healthcare professionals around the world with a mobile application for classification of hearing loss types based on images of hearing test results captured with a smartphone camera". The process of validating this hypothesis is directly related to the fulfillment of objective G4 and the publication (P5). Given the substantial need for this type of application in developing countries, the app was purposefully tested on devices that can be purchased for less than 100 USD. The results of the tests clearly indicated that the app is entirely capable of classifying the type of hearing loss based on a photograph taken with even a less powerful camera (with the minimum sensor resolution being set at 12 MP). Moreover, the application can operate independent of a network connection, with all calculations being performed locally by AI models optimized for mobile devices, ensuring enhanced security of patient data. The overall findings, presented in (P5), demonstrate that the application is appropriate for deployment in low-income medical settings. In conclusion, **it is possible to optimize modern neural network architectures to efficiently operate on smartphones which cost less than 100 USD**, thus **providing healthcare professionals around the world with a mobile application for classification of hearing loss types based on images of hearing test results captured with a smartphone camera**.

Additionally, this dissertation, based on the concise series of five published articles, proves the following contributions of the author to the state of the art in application of artificial intelligence algorithms for analysis of pure tone audiometry:

C1. The author illustrates that recurrent neural networks (RNNs) can effectively capture temporal dependencies in audiometric patterns, surpassing the performance of conventional multi-layer perceptron (MLPs).

C2. The Bi-LSTM model proposed by the author has reached a classification accuracy of 99.33%, exceeding the current state-of-the-art in the classification of raw audiometry data, as noted by Elbaşı and Obali [20], who reported an accuracy of 95.5%. Furthermore, the performance is also superior to the existing state-of-the-art in raster audiogram classification, as demonstrated by Crowson et al. [21], which achieved an accuracy of 97.5%.

C3. The author introduced an application that provides professional audiologists with the capability to employ an AI decision support system for tonal audiometry test result interpretation, potentially reducing their workload, enhancing diagnostic precision, and minimizing the chances of human error.

C4. The author has developed an innovative open-source mobile application that features a fully on-device AI-enabled tool designed for interpreting pure-tone audiograms through a smartphone. This approach significantly advances the field by providing diagnostic support that is both real-time and designed for offline use to professionals as well as general practitioners.

Table 6 presents the relationship between hypotheses (H1) – (H3), the corresponding research goals (G1) — (G4) and author's contributions to the state of the art (C1) — (C4).

Hypotheses	Research goal	Publication	Dissertation chapter	Author's contribution
H1	G3	P4	3.3	C2, C3
H2	G1, G2	P1, P2, P3	3.1, 3.2	C1
H3	G4	P5	4	C3, C4

Table 6. The relationship between hypotheses (H1) – (H3), the corresponding research goals (G1) — (G4) and author's contributions to the state of the art (C1) — (C4).

5.2. Closing remarks and areas for future research

This thesis presents a comprehensive study that leads to a proposed author's solution for the implementation of artificial intelligence algorithms in the analysis of tonal audiometry. The proposed model for classifying types of hearing loss (including normal hearing) has demonstrated enhanced accuracy and has been trained on a considerably larger dataset than those available in the existing literature. Furthermore, a robust mobile application has been created under an open-source license, facilitating easy access for medical professionals to the classification model on their smartphones. However, the research in that field might be continued, in particular in the new potential areas of AI application in audiology, which have been outlined by the author in conclusions of (P4) and (P5) articles.

The potential areas for further research include, but are not limited to:

A1. The development of a model intended to support a more accurate classification of test results, which would factor in the probability of certain hearing disorders (e.g. otitis media, otosclerosis, noise-induced hearing loss, Ménière's disease, acoustic schwannoma, etc.).

A2. Creation of a similar mobile application for audiogram classification, specifically designed for iOS platforms.

A3. Design of an audiogram digitalization system that is effective for both printed and hand-drawn audiograms.

A4. The continuous enhancement of the mobile app is intended to enable a precise hearing evaluation to be carried out at home, incorporating automatic classification.

A promising outcome of this dissertation is the cooperation between Gdansk University of Technology and Medical University of Gdansk, represented by doctors from the Department of Otolaryngology. The integration of knowledge from two entirely distinct domains facilitates the development of solutions that catalyze advancements in both information technology and healthcare.

6. Computing resources

Computational resources for network training were provided by the Centre of Informatics - Tricity Academic Supercomputer & Network (CI TASK) under grant No. PT01167.

7. References

[1] World Health Organization. World Report on Hearing. WHO. Available at: https://www.who.int/publications/i/item/9789240020481 (2021).

[2] Michał Kassjański, Kulawiak, M. & Tomasz Przewoźny. Development of an AI-based audiogram classification method for patient referral. Proceedings of the 17th Conference on Computer Science and Intelligence Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, D. Ślęzak (eds). ACSIS, Vol. 30, pages 163–168 https://doi.org/10.15439/2022f66 (2022).

[3] Michał Kassjański et al. Detecting type of hearing loss with different AI classification methods: a performance review. Proceedings of the 18th Conference on Computer Science and Intelligence Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, D. Ślęzak (eds). ACSIS, Vol. 35, pages 1017–1022 https://doi.org/10.15439/2023f3083 (2023).

[4] Delgado López-Cózar, Enrique Orduña-Malea, Proceedings Scholar Metrics 2017: H Index of proceedings on Computer Science, Electrical & Electronic Engi neering, and Communications according to Google Scholar Metrics (2012-2016). EC3 Reports, 22. Granada, 27th of December, https://doi.org/10.13140/RG.2.2.11387.41768 (2017).

[5] Michał Kassjański et al. Efficiency of Artificial Intelligence Methods for Hearing Loss Type Classification: an Evaluation. Journal of Automation Mobile Robotics & Intelligent Systems 28–38 https://doi.org/10.14313/jamris/3-2024/19 (2024).

[6] Michał Kassjański et al. Automated hearing loss type classification based on pure tone audiometry data. Scientific Reports 14, https://doi.org/10.1038/s41598-024-64310-2 (2024).

[7] Kassjański, M., Kulawiak, M., Przewoźny, T., Tretiakow, D. & Molisz, A. Development and testing of an open source mobile application for audiometry test result analysis and diagnosis support. Scientific Reports 15, https://doi.org/10.1038/s41598-025-99338-5 (2025).

[8] Komunikat Ministra Nauki z dnia 05 stycznia 2024 r. w sprawie wykazu czasopism naukowych i recenzowanych materiałów z konferencji międzynarodowych - Ministerstwo Nauki i Szkolnictwa Wyższego - Portal Gov.pl. Ministerstwo Nauki i Szkolnictwa Wyższego https://www.gov.pl/web/nauka/komunikat-ministra-nauki-z-dnia-05-stycznia-2024-r-w-sprawie-wykazu-czasopism-naukowych-i-recenzowanych-materialow-z-konferencji-miedzynarodowych (2024).

[9] Glossary. National Institute of Deafness and Other Communication Disorders. Available at: https://www.nidcd.nih.gov/glossary.

[10] Ballantyne, J. C., Graham, J. M. & Baguley, D. Ballantyne’s Deafness (Wiley, 2009).

[11] Nagaraj, Naveen K., Hearing Loss and Cognitive Decline in the Aging Population: Emerging Perspectives in Audiology. Audiology Research, vol. 14, no. 3, 23 , pp. 479–492, https://doi.org/10.3390/audiolres14030040 (2024).

[12] Le, Colleen G, et al. The Audiogram: Detection of Pure-Tone Stimuli in Ototoxicity Monitoring and Assessments of Investigational Medicines for the Inner Ear. Journal of the Acoustical Society of America, vol. 152, no. 1, pp. 470–490, https://doi.org/10.1121/10.0011739 (2022).

[13] Guidelines for Manual Pure-tone Threshold Audiometry, Vol. 20, 297–301 (ASHA, 1978).

[14] Margolis, R. H. & Saly, G. L. Toward a standard description of hearing loss. Int. J. Audiol. 46(12), 746–758. https://doi.org/10.1080/14992020701572652 (2007).

[15] Belitz, C., et al. A machine learning based clustering protocol for determining hearing aid initial configurations from pure-tone audiograms, in Interspeech. https://doi.org/10.21437/interspeech.2019-3091 (2019).

[16] Abeer, E. et al. Data-driven audiogram classifier using data normalization and multi-stage feature selection. Sci. Rep. https://doi.org/10.1038/s41598-022-25411-y (2023).

[17] Pasta, A., Petersen, M. K., Jensen, K. J., and Larsen, J. Rethinking hearing aids as recommender systems, in CEUR Workshop Proceedings, Vol. 2439, 11–17 (2019).

[18] Guo, R., Liang, R., Wang, Q. & Zou, C. Hearing loss classification algorithm based on the insertion gain of hearing aid. Multimed. Tools Appl. https://doi.org/10.1007/s11042-023-14886-0 (2023).

[19] Charih, F., Bromwich, M., Mark, A. E., Lefrançois, R. & Green, J. R. Data-driven audiogram classification for mobile audiometry. Sci. Rep. https://doi.org/10.1038/s41598-020-60898-3 (2020).

[20] Elbaşı, E., Obali, M. Classification of hearing losses determined through the use of audiometry using data mining, in Conference: 9th International Conference on Electronics, Computer and Computation (2012).

[21] Crowson, M. G. et al. AutoAudio: Deep learning for automatic audiogram interpretation. J. Med. Syst. https://doi.org/10.1007/s10916-020-01627-1 (2020).

[22] Aziz, B., Riaz, N., Rehman, A., Ur Malik, M. I. & Malik, K. I. Colligation of hearing loss and chronic otitis media. Pak. J. Med. Health Sci. 15(8), 1817–1819. https://doi.org/10.53350/pjmhs211581817 (2021).

[23] Raghavan, A., Patnaik, U. & Bhaudaria, A. S. An observational study to compare prevalence and demography of sensorineural hearing loss among military personnel and civilian population. Indian J. Otolaryngol. Head Neck Surg. 74(S1), 410–415. https://doi.org/10.1007/s12070-020-02180-6 (2020).

[24] Berrar, D. Cross-validation. In Encyclopedia of Bioinformatics and Computational Biology 542–545, https://doi.org/10.1016/b978-0-12-809633-8.20349-x (2019).

[25] Popescu, M.-C., Balas, V. E., Perescu-Popescu, L. & Mastorakis, N. Multilayer perceptron and neural networks. WSEAS Trans. Circuits Syst. 8, 579–588 (2009).

[26] Keiron O'Shea & Nash, R. R. An Introduction to Convolutional Neural Networks. arXiv (Cornell University) https://doi.org/10.48550/arxiv.1511.08458 (2015).

[27] Sherstinsky, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D: Nonlinear Phenomena 404, 132306 (2020).

[28] Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv.org http://arxiv.org/abs/1412.3555, https://doi.org/10.48550/arXiv.1412.3555 (2014).

[29] Hand, D. J. & Yu, K. Idiot's Bayes?Not So Stupid After All? International Statistical Review 69, 385–398 (2001).

[30] Cunningham, P. & Delany, S. J. k-Nearest Neighbour Classifiers - A Tutorial. ACM Computing Surveys 54, 1–25 (2021).

[31] Cramer, J. S. The Origins of Logistic Regression. SSRN Electronic Journal https://doi.org/10.2139/ssrn.360300 (2003).

[32] Evgeniou, T. & Pontil, M. Support Vector Machines: Theory and Applications. Machine Learning and Its Applications 249–257 https://doi.org/10.1007/3-540-44673-7_12 (2001).

[33] Ruder, S. An overview of gradient descent optimization algorithms *. https://arxiv.org/pdf/1609.04747 (2017).

[34] Rokach, L. & Maimon, O. Decision Trees. in Data Mining and Knowledge Discovery Handbook 165–192 https://doi.org/10.1007/0-387-25465-x_9 (2005).

[35] Cutler, A., Cutler, D. R. & Stevens, J. R. Random Forests. *Ensemble Machine Learning* 45, 157–175 (2012).

[36] Scarselli, F., Gori, M., Ah Chung Tsoi, Hagenbuchner, M. & Monfardini, G. The Graph Neural Network Model. *IEEE Transactions on Neural Networks* 20, 61–80 (2009).

[37] Scikit-learn. “Scikit-Learn: Machine Learning in Python.” Scikit-Learn.org, scikit-learn.org/stable/ (2024).

[38] TensorFlow. “TensorFlow.” TensorFlow, Google, www.tensorflow.org/ (2019).

[39] Z. Zhao, A. Kunar, Van, R. Birke, and L. Y. Chen,“CTAB-GAN: Effective Table Data Synthesizing,”arXiv (Cornell University (2021).

[40] Li, S. et al. Interpreting Audiograms with Multi-stage Neural Networks. https://arxiv.org/abs/2112.09357 (2021).

[41] Charih, F. & Green, J. R. Audiogram digitization tool for Audiological Reports. IEEE Access 10(110761), 110761–110769 (2022).

[42] Yang, T.-W. et al. A novel method for audiogram digitization in audiological reports. IEEE Access 12, 37862–37872 (2024).

[43] Howard, Andrew G, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. ArXiv.org, arxiv.org/abs/1704.04861 (2017).

[44] Kiryati, N., Eldar, Y. & Bruckstein, A. M. A probabilistic Hough transform. Pattern Recognition 24, 303–316 (1991).

[45] J. Canny, A Computational Approach to Edge Detection, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-8, no. 6, pp. 679-698, https://doi.org/10.1109/TPAMI.1986.4767851 (1986)

[46] I. Sobel, An isotropic 3x3 image gradient operator, Presentation at Stanford A.I. Project (2014).

[47] Wu, Jiaxiang, et al. Quantized Convolutional Neural Networks for Mobile Devices, https://doi.org/10.1109/cvpr.2016.521 (2016).

[48] Carter, J., Sandall, J., Shennan, A. H. & Tribe, R. M. Mobile phone apps for clinical decision support in pregnancy: A scoping review. BMC Med. Inform. Decis. Mak. 19, 1–13. https://doi.org/10.1186/s12911-019-0954-1 (2019).

[49] michal-kass. GitHub - michal-kass/AudiogramScan. GitHub https://github.com/michal-kass/AudiogramScan (2025).

P. Publications included in the series

P1. Publication P1

Author Contribution Statement

I declare that in the publication:

M. Kassjański, M. Kulawiak and T. Przewoźny, "Development of an AI-based audiogram classification method for patient referral," 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS), Sofia, Bulgaria, 2022, pp. 163-168, https://doi.org/10.15439/2022F66. (2022)

my contribution, in accordance with CRediT (Contributor Role Taxonomy) was as follows: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization. Furthermore, my contribution percentage to the development of the publication was 70%.

01.07.2025 Michał Kassjański

Date Michał Kassjański

I, the undersigned, hereby certify that the information given by Michał Kassjański is correct.

01.07.2025 Marcin Kulawiak

Date Marcin Kulawiak

16.07.2025 Przewoźny Tomasz

Date Tomasz Przewoźny

Proceedings of the of the 17th Conference on Computer Science and Intelligence Systems pp. 163–168
DOI: 10.15439/2022F66
ISSN 2300-5963 ACSIS, Vol. 30

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

Michał Kassjański, Marcin Kulawiak
Department of Geoinformatics,
Faculty of Electronics, Telecommunications and Informatics,
Gdansk University of Technology,
Gdansk, Poland
Email: {michal.kassjanski, markulaw}@pg.edu.pl

Tomasz Przewoźny
Department of Otolaryngology,
Medical University of Gdansk,
Smoluchowskiego Str. 17,
80-214 Gdansk, Poland
Email: tomasz.przewozny@gumed.edu.pl

***Abstract*—Hearing loss is one of the most significant sensory disabilities. It can have various negative effects on a person's quality of life, ranging from impeded school and academic performance to total social isolation in severe cases. It is therefore vital that early symptoms of hearing loss are diagnosed quickly and accurately. Audiology tests are commonly performed with the use of tonal audiometry, which measures a patient's hearing threshold both in air and bone conduction at different frequencies. The graphic result of this test is represented on an audiogram, which is a diagram depicting the values of the patient's measured hearing thresholds. In the course of the presented work several different artificial neural network models, including MLP, CNN and RNN, have been developed and tested for classification of audiograms into two classes - normal and pathological represented hearing loss. The networks have been trained on a set of 2400 audiograms analysed and classified by professional audiologists. The best classification performance was achieved by the RNN architecture (represented by simple RNN, GRU and LSTM), with the highest out-of-training accuracy being 98% for LSTM. In clinical application, the developed classifier can significantly reduce the workload of audiology specialists by enabling the transfer of tasks related to analysis of hearing test results towards general practitioners. The proposed solution should also noticeably reduce the patient's average wait time between taking the hearing test and receiving a diagnosis. Further work will concentrate on automating the process of audiogram interpretation for the purpose of diagnosing different types of hearing loss.**

I. Introduction

HEARING IS one of the most important senses and is crucial for a human to maintain full connectivity to the world. Early on in life, hearing helps one to establish language skills which lays the groundwork for quick development during school years. In daily tasks, hearing is used in communicating with other people as well as for listening to music, television and radio, and going to the cinema or theatre.

According to World Health Organization (WHO), currently, around 430 million people globally require rehabilitation services for their hearing loss [1]. Estimations show that by 2050 nearly 2.5 billion people will be living with some degree of hearing loss, at least 700 million of whom will require rehabilitation services [1]. Overall, hearing impairment has devastating consequences for interpersonal communication, psychosocial well-being, quality of life and economic independence [2]. The consequences of hearing loss are frequently underestimated and ignoring the initial symptoms usually leads to further degradation. Once diagnosed, early intervention is the key to successful treatment. Medical and surgical treatment can cure most ear diseases, potentially reversing the associated hearing loss. Research has shown that, particularly in children, almost 60% of hearing loss is due to causes that can be prevented [1], [6], [7].

The standard hearing test is carried out using pure tone audiometry, which determines the hearing thresholds at different frequencies. As a rule, a frequency range of the hearing test varies within 125 – 8000 Hz. The sound level of pure tones is given in dBHL, and the subject is tested in both air and bone conduction. The test results in two data series containing discrete hearing thresholds in the function of frequency, separately for both conductions. This data series is usually presented in the form of an inverted graph called audiogram. An audiogram helps to determine the degree of hearing loss, but also the type of pathology: sensorineural, conductive or mixed [3], [4].

According to projections, the demand for professional audiologists will burgeon in near future [1]. Nowadays, around 78% of low-income countries have less than one otorhinolaryngologist per million inhabitants and about 93% have less than one audiologist per million inhabitants [1], [5]. In this context, introduction of expert systems based on artificial intelligence for preliminary audiogram interpretation could significantly reduce the workload of specialists, while at the same time shortening the patient's wait for a diagnosis.

Over the last decade, a comparison of several approaches to hearing loss determination, including Decision Tree, Naive Bayes and Neural Network Multilayer Perceptron (NN) model, has been prepared by Eibaşı & Obali [10]. The tests have been carried out using a set of numerical values representing Decibels corresponding to fixed frequency levels (750Hz, 1kHz, 1,5kHz, 2kHz, 3kHz, 4kHz, 6kHz, 8kHz). The achieved accuracy was 95.5% in Decision Tree, 86.5 % in Naive Bayes and 93.5 % in NN.

A different approach was presented by Noma & Ghani [11], who developed a classification system based on the relationship between pure-tone audiometry thresholds and inner ear

disorders symptoms such as Tinnitus, Vertigo, Giddiness etc. The classifier, based on the multivariate Bernoulli model with feature transformation, has shown to provide 98% accuracy of predicting hearing loss symptoms based on audiometry results.

Recently, Charih et al. [12] presented their Data-Driven Annotation Engine, a decision tree based audiogram classifier which considers the configuration, severity, and symmetry of participant's hearing losses and compared it to AMCLASS [13], which fulfils the same purpose using a set of general rules. Both classifiers have achieved similar accuracy of around 90% across 270 different audiometric configurations by three licensed audiologists.

More recently, Crowson et al. [14] adopted the ResNet-101 model to classify audiogram images into three types of hearing loss (sensorineural, conductive or mixed) as well as normal hearing using a set of training and testing images consisting of 1007 audiograms. This approach resulted in 97.5% classification accuracy, however it is limited to processing images.

In summary, the combination of neural networks and increased computing resources of new hardware architectures has the potential to deliver faster overall tests results and more detailed assessments[15]. This being said, however, the currently proposed solutions deliver classification accuracy in the 90-95% range, which, although very high, still leaves considerable room for error. Clinical standards suggest that the margin of error should be kept under 5%[16] and optimally should be close to 3% [17]. These requirements are met only by two of the discussed classifiers. The method proposed by Noma & Ghani achieves 98% accuracy, however it has been designed to predict significant symptoms of inner ear disorder, and thus it cannot be used for general purposes such as early detection of hearing degradation. The best audiogram classifier to date has been presented by Crowson et al., who used transfer learning to adapt an established image classifier network to analysis of audiogram images. While this approach resulted in a 97% classification accuracy, it exhibits serious limitations. Because it is an image classifier, it cannot be used with the original data series produced by tonal audiometry. This means that the data series first need to be converted into audiogram images, which may result in data loss. Moreover, although the structure of audiograms generally is similar, there can still be significant differences between audiograms generated by different hardware and software configurations. Aside from differences such as background and line colours, audiograms can also differ in the amount of presented information (eg. they may contain data for a single ear or both). A sample comparison of significant differences between audiograms obtained from different sources is presented in Figures 1 and 2. In consequence, a universal solution for classifying results of tonal audiometry cannot be based on an image classifier.

This study presents the development of a neural network for classification of discrete tonal audiometry data series. In the course of this study, several different neural network architectures have been trained and tested with the use of 2400 audiogram data series analysed and classified by professional audiologists. The goal of the presented study was to achieve a high enough classification accuracy for the developed network to be applicable for use in a clinical environment.

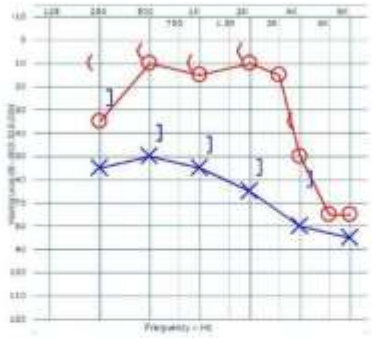


Fig. 1. A pure tone audiogram showing air and bone conduction thresholds for **both left and right ear** [8]. The "X" and "]" symbols are used to mark left-sided air and bone conduction, respectively. The "O" indicate air conduction, whereas the "<" denote bone conduction, both in the right ear.

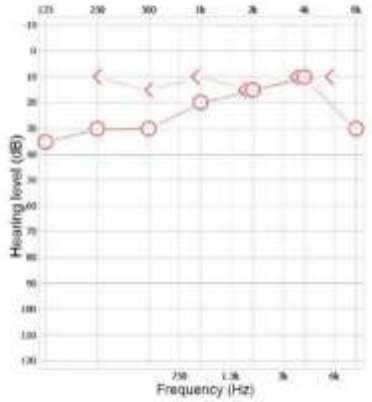


Fig. 2. A pure tone audiogram showing air and bone conduction thresholds **only for the right ear.** The "O" and "<" indicate left-sided air and bone conduction, respectively.

II. Materials & Methods

A. Data

The study has been conducted with the use of 2400 data series containing results of pure tone audiometry tests performed from 2020 to 2021 by clinicians working at the Otolaryngology Clinic of the University Clinical Centre in Gdansk, Poland. The data contains 650 examples of normal hearing and 1750 examples of pathological hearing loss. The tests had been performed in a soundproof booth, according to ISO 8253 and ISO 8253 standards. Air conduction tests employed TDH-39P headphones, while bone conduction testing involved a Radioear B-71 bone-conduction vibrator. The data series have been analysed and labelled by expert audiologists from the

Medical University of Gdansk Department of Otolaryngology according to established methodology [9]. In consequence, the dataset has been classified into two subsets: hearing pathology and normal hearing.

B. Preprocessing

The input data series contained numerical information about tonal points, defined as loudness (dB) for a given frequency (Hz), in XML format. The dataset included the following range of frequencies:

125Hz, 250Hz, 375Hz, 500Hz, 750Hz, 1000Hz, 1500Hz, 2000Hz, 3000Hz, 4000Hz, 6000Hz, 8000Hz.

Every tested frequency has been assigned a loudness level in the range from -10dB to 120dB. If certain frequencies had not been registered during the hearing test, they have not been included in the corresponding data series.

C. Testing methodology

Using the prepared dataset, three different neural network architectures have been trained to interpret tonal audiometry data and in order to differentiate normal hearing (N) from pathological hearing loss (P). The tested architectures included Multilayer Perceptron (MLP), Convolutional (CNN) and Recurrent (RNN) neural networks, all of which have been previously applied to data classification problems [18], [19], [20]. The general workflow of the presented study is shown in Fig. 3. Each model has been assessed using k-fold cross-validation, which consists of dividing the data into k subsets and training the model k-times with k-1 subsets, with a different subset being used for testing in every iteration. The presented research used $k = 5$, which resulted in train to test dataset proportions of 80% to 20%, respectively.

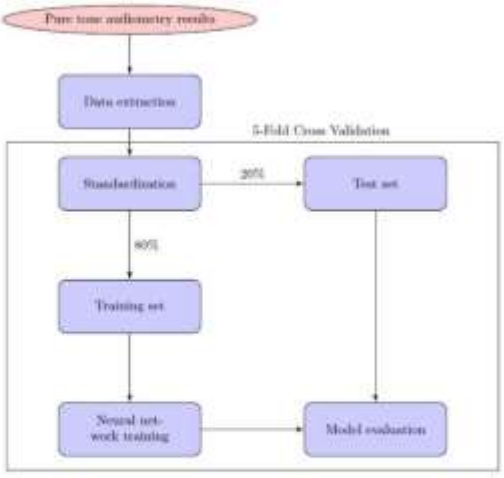


Fig. 3. Workflow of processes leading to model evaluation.

After revealing the best performing architecture, further tests and optimizations would be carried out in order to improve classification accuracy.

III. Results

The purpose of the initial tests was to reveal the best neural network architecture model for classification of pure tone audiometry data. The tested neural network architectures included MLP, CNN and RNN. The results of those tests are presented in Table I.

TABLE I
COMPARISON OF PERFORMANCE RESULTS OF PRELIMINARY MODELS.

Parameters	MLP	CNN	RNN
Accuracy	0.9458	0.9563	0.9604
Loss	0.6429	0.1185	0.1346
Precision	0.8255	0.8984	0.9062
Recall	1.0	0.9349	0.9430
F1	0.9044	0.9163	0.9243

As it can be seen, initial research revealed that the best classification performance has been produced by the RNN architecture model. Once the most promising neural network architecture has been identified, three of its variants have been trained and optimized in terms of hyper parameters, including number of nodes and hidden layers, dropout layers, learning and decay rate. The first model consisted of a simple RNN, second one was based on Gated Recurrent Units (GRU) [22] and the last one used Long Short-Term Memory (LSTM) [21]. The results of these tests are shown in Table II.

Receiver Operating Characteristics (ROC) curves with corresponding Area Under the Curve (AUC) parameters for these models are presented in Fig. 4.

TABLE II
COMPARISON OF PERFORMANCE RESULTS OF RNN MODELS.

Parameters	Simple RNN	GRU	LSTM
Accuracy	0.9646	0.9771	0.9812
Loss	0.0836	0.0530	0.0540
Precision	0.9030	0.9453	0.9394
Recall	0.9680	0.9680	0.9920
F1	0.9344	0.9565	0.9650

The cross validation scores for $k = 5$ with LSTM classifier are given in Table III. The average accuracy was 98.08% (+/- 0.17%).

TABLE III
K-FOLD VALIDATION SCORE OF LSTM MODEL ($k = 5$).

Iteration	1	2	3	4	5
Accuracy	97.96	98.33	97.96	97.91	98.22

A detailed analysis of classification performance achieved by the tested RNN models can be made using a confusion matrix, which visualizes the number of True Positives (TP

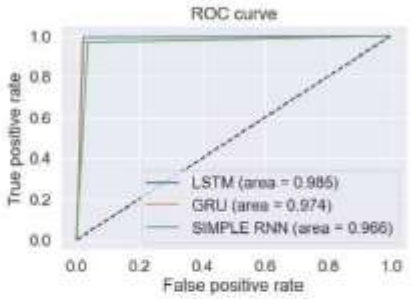


Fig. 4. ROC curve with AUC parameter of RNN models.

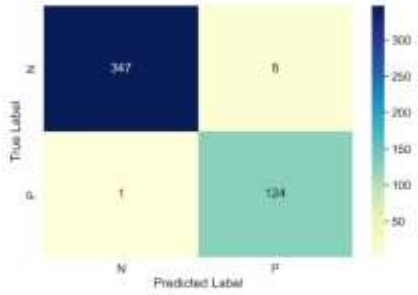


Fig. 7. Confusion matrix of LSTM.

- patients who have been properly classified with hearing loss), True Negatives (TN - patients who have been properly classified with good hearing), False Positives (FP - patients who have been improperly classified as hearing loss) and False Negatives (FN - patients who have been improperly classified with good hearing). The confusion matrix for the tested RNN models is presented in Figures 5, 6 and 7.

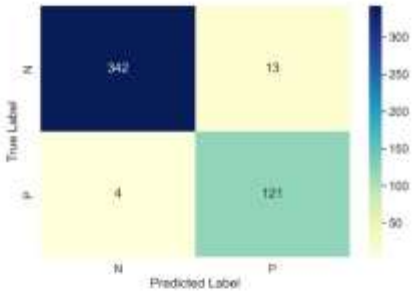


Fig. 5. Confusion matrix of simple RNN.

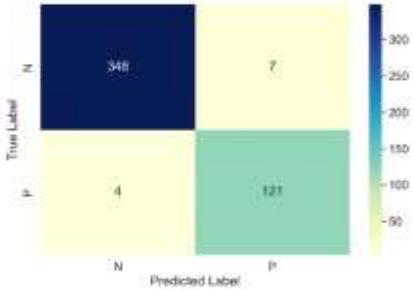


Fig. 6. Confusion matrix of GRU.

IV. Discussion

Initial tests have shown that the simple RNN architecture model delivers noticeably better pure tone audiometry classification results in comparison to MLP and CNN models, achieving accuracy of 96.04% versus 94.58% and 95.63% respectively (Tab. I). The chosen network architecture appears to have the largest impact on classification accuracy, as further tests and optimizations resulted in minor improvements. Optimization of parameters such as the number of nodes and hidden layers, dropout layers as well as learning and decay rate improved the accuracy of simple RNN from 96.04% to 96.46%. In comparison, applying the same optimization process to MLP and CNN models did not result in markedly improved evaluation parameters. A possible explanation for this could be the fact that RNN have been designed to process time series data, and structurally pure tone audiometry results could be interpreted as a special case of time series. This could be further explored by testing the effectiveness of more advanced RNN models such as GRU and LSTM. As it can be seen in Tab. II, both of these models obtained more than 97% accuracy, with the highest out-of-training set accuracy being achieved by LSTM at 98.12%. While these results, which have been cross-validated using the 5-fold method, would seem to indicate a general prevalence of the RNN architecture in processing audiometry data, establishing an effectiveness hierarchy of RNN models is a more complex matter. Although LSTM has shown the best classification accuracy, when analysed in terms of confusion matrix, the lowest number of False Positives (FP) was obtained by GRU (Figures 6 and 7), with LSTM taking second place. In comparison, the simple RNN produced over 62% more False Positives than LSTM and 85% more than GRU.

Overall, simple RNN and GRU performed equally well in terms of False Negatives (FN), producing them only in 0.8% of cases, whereas LSTM significantly outperformed the other models with only one case of error occurring. It can be argued that when classifying results of pure tone audiometry tests, the FN number is more important than FP because it shows that a patient does not have hearing loss when they actually do. In this case the patient may not receive treatment and

get worse because their disease was undetected. On the other hand, a false positive would only result in the patient being unnecessarily referred to an audiologist, who would properly interpret the test results and inform the patient that their level of hearing is normal.

Summing up, it can be said that the 98.12% classification accuracy achieved by LSTM fulfills the established margin of error criteria and is significantly better than the 97.5% classification accuracy offered by the best existing algorithm for audiogram data classification, proposed by Crowson et al. [14]. While some of the difference could be attributed to the rival method providing a larger set of classes, the presented method provides an additional advantage in the type of processed data: it works with original tonal audiometry data series instead of audiogram images and therefore is more universal. The only rival method also designed for processing tonal audiometry data series, presented by Elbaşı & Obali [10], provides an even lower 95.5% classification accuracy.

In terms of classifying pure tone audiometry data, the only existing solution with a similar classification accuracy level (98%, proposed by Noma & Ghani [11]), has been designed to predict significant symptoms of inner ear disorder and thus cannot be used for general classification of tonal audiometry test results.

V. Conclusions

The presented work aimed to develop a neural network for classification of discrete tonal audiometry data series with accuracy high enough for medical application. In the course of this study, several different neural network architectures, including MLP, CNN and RNN, have been trained and tested with the use of 2400 audiogram data series analysed and classified by professional audiologists. The highest classification accuracy was achieved with an optimized LSTM RNN at 98.12%. The high accuracy of the obtained neural network, particularly the low number of False Negatives (0.2%), allows for its application at the Otolaryngology Clinic of the University Clinical Centre in Gdansk, Poland. Results of pure tone audiometry tests, which thus far needed to be examined by professional audiologists, can now be classified with the developed neural network under the supervision of general practitioners. This change may result in a significant reduction of the workload of audiology specialists, as they will no longer need to deal with patients whose symptoms are not caused by hearing loss (which may amount to over 10% of all patients subjected to pure tone audiometry tests) [23], [24]. After it has been further tested in practice, the developed solution could be introduced directly in the audiometry laboratory, ensuring that the patient receives a first interpretation of the performed tests as soon as they have been completed. Further work will concentrate on expanding the classifier for the purpose of diagnosing different types of hearing loss.

Acknowledgement

The authors would like to thank M. Grono, K. Kożmiński, P. Mierzwińska and A. Romanowicz who helped to create the pure tone audiometry test dataset used in this study.

References

[1] World Health Organization. 2021. World report on hearing. https://www.who.int/publications/i/item/world-report-on-hearing.
[2] Olusanya, B. O., Neumann, K. J., Saunders, J. E. 2014. The global burden of disabling hearing impairment: a call to action. Bull World Health Organ. 92(5):367–373,http://dx.doi.org/92/5/13-128728
[3] Kapul AA, Zubova EI, Torgaev SN, Drobchik VV. 2017. Pure-tone audiometer. J Phys Conf Ser, http://dx.doi.org/10.1088/1742-6596/881/1/012010
[4] V. P. Aras. 2003. Audiometry techniques, circuits, and systems, M. Tech. Credit Seminar Report, Electronic Systems Group, EE Dept, IIT Bombay
[5] World Health Organization. 2013. Multi-country assessment of national capacity to provide hearing care
[6] Tukaj C, Kuczkowski J, Sakowicz-Burkiewicz M, Gulida G, Tretiakow D, Mionskowski T, Pawełczyk T. 2014. Morphological alterations in the tympanic membrane affected by tympanosclerosis: ultrastructural study. Ultrastruct Pathol. 38(2):69-73, http://dx.doi.org/10.3109/01913123.2013.833563
[7] Narozny W, Skorek A, Tretiakow D. 2021. Does Treatment of Sudden Sensorineural Hearing Loss in Patients With COVID-19 Require Anticoagulants? Otolaryngol Head Neck Surg. 165(1):236-237, http://dx.doi.org/10.1177/0194599820988511
[8] Prashanth Prabhu P, Jyothi Shivswamy. 2017. Audiological findings from an adult with thin cochlear nerves, Intractable & Rare Diseases Research, 6(1):72-75, http://dx.doi.org/10.5582/irdr.2016.01081
[9] Przewoźny T, Kuczkowski J. 2017. Hearing loss in patients with extracranial complications of chronic otitis media. Otolaryngol Pol. 71(3), pp. 31-41, http://dx.doi.org/10.5604/01.3001.0010.0130
[10] Ersin Elbaşı, Murat Obali. 2012. Classification of Hearing Losses Determined through the Use of Audiometry using Data Mining, Conference: 9th International Conference on Electronics,Computer and Computation
[11] Noma, N. G., & Ghani, M. K. A. 2013. Predicting Hearing Loss Symptoms from Audiometry Data Using Machine Learning Algorithms. In Proceedings of the Software Engineering Postgraduates Workshop (SEPoW), p. 86, Penang, Malaysia
[12] Charih F, Bromwich M, Mark AE, Lefrançois R, Green JR. 2020. Data-Driven Audiogram Classification for Mobile Audiometry. Sci Rep 10, 3962, http://dx.doi.org/10.1038/s41598-020-60898-3
[13] Margolis, R.H. and Saly, G.L. 2007. Toward a standard description of hearing loss. International journal of audiology, 46(12), pp.746-758, http://dx.doi.org/10.1080/14992020701572652
[14] Crowson MG, Lee JW, Hamour A, Mahmood R, Babier A, Lin V, Tucci DL, Chan TCY. 2020. AutoAudio: Deep Learning for Automatic Audiogram Interpretation. J Med Syst. 44(9):163, http://dx.doi.org/10.1007/s10916-020-01627-1
[15] Barbour, Dennis L. MD, PhD; Wasmann, Jan-Willem A. 2021. Performance and Potential of Machine Learning Audiometry, The Hearing Journal: Volume 74 - Issue 3 - p 40,43,44, http://dx.doi.org/10.1097/01.HJ.0000737592.24476.88
[16] Aziz, B., Riaz, N., Rehman, A.U., Malik, M.I., Malik, K.I. 2021. Colligation of Hearing Loss and Chronic Otitis Media. Pakistan Journal of Medical and Health Sciences Vol. 15, Issue 8, pp. 1817, http://dx.doi.org/10.53350/pjmhs211581817
[17] Raghavan, A., Patnaik, U. and Bhavdaria, A.S. 2020. An Observational Study to Compare Prevalence and Demography of Sensorineural Hearing Loss Among Military Personnel and Civilian Population. Indian Journal of Otolaryngology and Head & Neck Surgery, pp.1-6, http://dx.doi.org/10.1007/s12070-020-02180-6
[18] Zieliński, S. K., & Lee, H. 2018. Feature extraction of binaural recordings for acoustic scene classification. In 2018 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 585-588), http://dx.doi.org/10.15439/2018F182
[19] Agbehadji, I. E., Millham, R., Fong, S. J., & Yang, H. 2018. Kestrel-based Search Algorithm (KSA) for parameter tuning unto Long Short Term Memory (LSTM) Network for feature selection in classification of high-dimensional bioinformatics datasets. In 2018 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 15-20), http://dx.doi.org/10.15439/2018F52

[20] Lindén, J., Forsström, S., & Zhang, T. 2018. Evaluating combinations of classification algorithms and paragraph vectors for news article classification. In 2018 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 489–495), http://dx.doi.org/10.15439/2018F110

[21] Hochreiter, Sepp & Schmidhuber, Jurgen. 1997. Long Short-term Memory. Neural computation. 9. 1735-80 (1997), http://dx.doi.org/10.1162/neco.1997.9.8.1735

[22] Cho, Kyunghyun & Merrienboer, Bart & Gulcehre, Caglar & Bougares, Fethi & Schwenk, Holger & Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, http://dx.doi.org/10.3115/v1/D14-1179

[23] do Carmo LC, Médicis da Silveira JA, Marone SA, D'Ottaviano FG, Zagati LL, Dias von Söhsten Lins EM. 2018. Audiological study of an elderly Brazilian population. Braz J Otorhinolaryngol;74(3):342-9, http://dx.doi.org/10.1016/s1808-8694(15)30566-8

[24] Walker JJ, Cleveland LM, Davis JL, Seales JS. 2013. Audiometry screening and interpretation. Am Fam Physician.;87(1):41-7

P2. Publication P2

Author Contribution Statement

I declare that in the publication:

M. Kassjański, M. Kulawiak, T. Przewoźny, D. Tretiakow, , J. Kuryłowicz, A. Molisz, K. Koźmiński, A. Kwaśniewska, P. Mierzwińska-Dolny, M. Grono, "Detecting Type of Hearing Loss with Different AI Classification Methods: A Performance Review," 2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS), Warsaw, Poland, 2023, pp. 1017-1022, https://doi.org/10.15439/2023F3083. (2023)

my contribution, in accordance with CRediT (Contributor Role Taxonomy) was as follows: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization. Furthermore, my contribution percentage to the development of the publication was 70%.

01.07.2025	Michał Kassjański
Date	Michał Kassjański

I, the undersigned, hereby certify that the information given by Michał Kassjański is correct.

01.07.2025	Marcin Kulawiak	19.08.2025	Krzysztof Koźmiński
Date	Marcin Kulawiak	Date	Krzysztof Koźmiński
16.07.2025		20.07.2025	
Date	Tomasz Przewoźny	Date	Aleksandra Kwaśniewska
15.07.2025		18.08.2025	
Date	Dmitry Tretiakow	Date	Paulina Mierzwińska-Dolny
26.08.2025		16.07.2025	M. Grono
Date	Jagoda Kuryłowicz	Date	Miłosz Grono
16.07.2025			
Date	Andrzej Molisz		

Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 1017–1022
DOI: 10.15439/2023F3083
ISSN 2300-5963 ACSIS, Vol. 35

Detecting type of hearing loss with different AI classification methods: a performance review.

Michał Kassjański[1], Marcin Kulawiak[1], Tomasz Przewoźny[2], Dmitry Tretiakow[2], Jagoda Kuryłowicz[2], Andrzej Molisz[2], Krzysztof Kożmiński[4], Aleksandra Kwaśniewska[3], Paulina Mierzwińska-Dolny[4], Miłosz Grono[4]
[1] Department of Geoinformatics, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland
[2] Department of Otolaryngology, Medical University of Gdańsk, Poland
[3] Department of Otolaryngology, Laryngological Oncology and Maxillofacial Surgery, University Hospital No. 2, Bydgoszcz, Poland
[4] Student's Scientific Circle of Otolaryngology, Medical University of Gdańsk, Poland
Email: {michal.kassjanski, markulaw}@pg.edu.pl, {tprzew, d.tret}@gumed.edu.pl, jagoda.kurylowicz@gmail.com, {andrzej.molisz, krzyk}@gumed.edu.pl, kwasniewska.aleks@gmail.com, {paulinamierzwinska, milosz.grono}@gumed.edu.pl

***Abstract*—Hearing is one of the most crucial senses for all humans. It allows people to hear and connect with the environment, the people they can meet and the knowledge they need to live their lives to the fullest. Hearing loss can have a detrimental impact on a person's quality of life in a variety of ways, ranging from fewer educational and job opportunities due to impaired communication to social withdrawal in severe situations. Early diagnosis and treatment can prevent most hearing loss. Pure tone audiometry, which measures air and bone conduction hearing thresholds at various frequencies, is widely used to assess hearing loss. A shortage of audiologists might delay diagnosis since they must analyze an audiogram, a graphic representation of pure tone audiometry test results, to determine hearing loss type and treatment. In the presented work, several AI-based models were used to classify audiograms into three types of hearing loss: mixed, conductive, and sensorineural. These models included Logistic Regression, Support Vector Machines, Stochastic Gradient Descent, Decision Trees, RandomForest, Feedforward Neural Network (FNN), Convolutional Neural Network (CNN), Graph Neural Network (GNN), and Recurrent Neural Network (RNN). The models were trained using 4007 audiograms classified by experienced audiologists. The RNN architecture achieved the best classification performance, with an out-of-training accuracy of 94.46%. Further research will focus on increasing the dataset and enhancing the accuracy of RNN models.**

I. Introduction

HEARING is considered an essential sensory organ since it provides us with valuable information about the external environment. In addition, it enables us to interact with the outside world, communicate with others, remain safe, and derive enjoyment from a variety of auditory experiences. Hearing complements our other senses, such as sight and sensation, to provide a complete understanding of our surroundings.

According to the World Health Organization (WHO), more than 1.5 billion persons worldwide suffer from hearing loss, of which 430 million have moderate or severe hearing loss in their better hearing ear. According to the projections of the World Health Organization, by 2050 nearly 2.5 billion people will have hearing loss and at least 700 million will require rehabilitation services. Fortunately, many instances of hearing loss can be prevented through early detection and intervention [1].

Although the majority of ear diseases are curable, accurate diagnosis is a significant barrier to effective treatment. Audiologists, who are essential for the execution and interpretation of testing, are scarce worldwide. Approximately 93% of low-income countries have fewer than one audiologist per million people [1]. Given the disparity between the supply and demand for hearing specialists, artificial intelligence (AI) has the potential to resolve this problem. AI employs algorithms that enable computers to recognize particular data analysis patterns and make conclusions. The most prevalent AI application in tonal audiometry is hearing aid personalization, in which AI systems assist both the hearing-care expert and the patient in more precisely and efficiently adjusting hearing aids to the client's preferences [2, 3, 4].

Another possible application of expert systems in audiology is interpreting results of pure-tone audiometry, which is the standard method for diagnosing hearing loss. Typically, the examination is conducted while situated in an anechoic chamber. It entails conveying increasing-intensity pure tones through headphones and determining the threshold for air and bone conduction. In general, the results of the pure-tone audiometry test are presented as an inverted graph called an audiogram, which allows for identifying hearing impairment.

When describing hearing loss, three aspects are considered: the type of hearing loss, the degree of hearing loss, and the configuration of hearing loss. Three types of hearing loss are

distinguished: sensorineural, conductive, and mixed. The pattern of hearing loss across frequencies is determined by the configuration (shape) of the audiogram, whereas the severity is determined by the degree of hearing loss [5].

Classification of automated audiometry data has been investigated for a very long time. In the past ten years, there have been a number of initiatives to develop an automated classification system sufficiently accurate for clinical application. The most successful have been presented by Elbaşı and Obali [6], who compared Decision Tree, Naive Bayes, and Neural Network Multilayer Perceptron (NN) models for determining hearing loss. The research was conducted on a data set containing 200 samples divided into four categories: normal hearing, conductive hearing loss, sensorineural hearing loss, and mixed hearing loss. The accuracy of the classification algorithms was 95.5% for Decision Tree, 86.5% for Naive Bayes, and 93.5% for NN. While that work used raw audiometry test results, Crowson et al. [7] applied the ResNet models to classify rasterized results in the form of audiogram images into four categories of hearing (normal, sensorineural hearing loss, conductive hearing loss, mixed hearing loss) on a set of 1007 audiograms. Instead of completely training the classifier from scratch, the authors used transfer learning to train the classifier using widely recognized raster classification models. This method achieved a classification accuracy of 97.5%, but it is limited to image analysis.

In conclusion, the combination of machine learning and increased computational resources in innovative hardware architectures has the potential to generate faster overall test results and more exhaustive evaluations in audiology [8]. Despite the type of hearing loss, the classification accuracy of the currently proposed solutions ranges from 86 to 97%, which, while extremely high, still leaves a substantial margin of error. Moreover, while the best available audiogram classifier, presented by Crowson et al. [7], achieved 97.5% accuracy, it cannot be applied to the original data series produced by tonal audiometry due to being an image classifier. This means that before classification the datasets would need to be converted into a particular format of audiogram images (although the structure of audiograms is generally analogous, audiograms generated by different software can vary quite significantly). Additional problems would stem from the fact that some types of software generate two audiograms (one for each ear), while other software combines the information from both ears into a single audiogram, posing a great difficulty in universal analysis. Consequently, an image classifier cannot form the core of a versatile solution for classifying tonal audiometry results. Moreover, the abovementioned studies on determining the type of hearing loss were carried out with a relatively small data set, ranging from 200 test results in Elbaşı & Obali [6] to 1007 in Crowson et al. [7], which might have led to an optimistic and uncertain evaluation of model performance.

This study establishes the benchmark for machine learning and deep learning algorithms using a large set of discrete tonal audiometry data series. Throughout the course of this investigation, multiple AI models were trained and evaluated using 4007 audiogram data series analyzed and classified by professional audiologists. The purpose of this study was to investigate the performance of various AI solutions when applied to raw tonal audiometry data.

II. MATERIALS & METHODS

A. Data

The study was carried out on 4007 data series containing the results of pure tone audiometry tests performed between 2017 and 2021 by clinicians at the Otolaryngology Clinic of the University Clinical Centre in Gdansk, Poland. The data class proportion is presented in Fig. 1. Conductive hearing loss only has 674 examples, while mixed hearing loss has 1594 and sensorineural hearing loss has 1739. Each patient provided a maximum of two test results, one for the left ear and one for the right, resulting in no duplication of data from the same patient and ensuring adequate data variety.

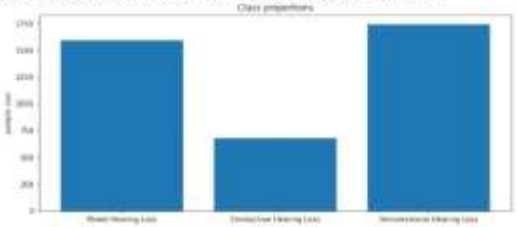


Figure 1. The three forms of hearing loss represented in the dataset, along with their respective proportions.

Tonal audiometry was used to evaluate patients' hearing according to the American Speech-Language-Hearing Association (ASHA) guidelines. All tests were conducted in soundproof chambers (ISO 8253, ISO 8253). The TDH39P headphones were utilized for air conduction testing, while the Radioear B-71 bone-conduction vibrator was used for bone conduction testing [9].

Experienced audiologists labeled the morphologies of hearing loss on the audiometry test results, dividing the set into three classes according to established methodology [5]: mixed hearing loss, conductive hearing loss and sensorineural hearing loss.

Typically, the results of pure-tone audiometry are depicted as an audiogram, which is a graphical representation of how loud sounds must be at various frequencies for them to be audible. In addition to a graphical representation, audiology software generates XML files that comprise all information regarding tonal points in the audiogram. This study processes raw audiometry data using XML files, analyzing five primary

frequencies (250, 500, 1000, 2000, 4000 Hz) from both air conduction and bone conduction.

B. Methodology

The aim of the study was to test the performance of several different machine learning algorithms at the task of classifying tonal audiometry data. The goal of each method was to accurately categorize each dataset as mixed hearing loss (M), conductive hearing loss (C) or sensorineural hearing loss (S).

a) Machine learning algorithms

The initial phase of research involved testing the following machine learning classification algorithms: Gaussian Naive Bayes, K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machines (SVMs), Stochastic Gradient Descent (SGD), Decision Tree and Random Forest. The second phase of the study involved testing the following ANN architectures: Feedforward Neural Network (FNN), Convolutional Neural Network (CNN), Graph Neural Network (GNN), and Recurrent Neural Network (RNN). These techniques were previously applied to the classification problem of medical data [10, 11].

b) Data preprocessing

The input data series consisted of vertical information about tonal points of air and bone conduction, defined as volume (dB) for a given frequency (Hz), obtained from XML files. The frequency range of the dataset included 250Hz, 500Hz, 1000Hz, 2000Hz, and 4000Hz. Each frequency tested has been designated a loudness level between -10dB and 120dB. The dataset did not contain any empty values.

Since GNN requires graph input, the vector was turned into a directed graph with 10 nodes and 18 edges. Frequency and loudness values have been assigned to nodes. Figure 2 shows a graphical depiction of the graph.

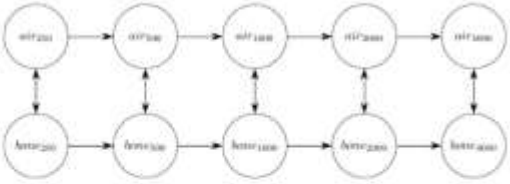


Figure 2. The GNN architecture's input graph structure.

c) Model evaluation

The performance of the tested models was evaluated using K-fold Cross-Validation, which is the process of splitting a dataset into K folds, using K-1 datasets for training and one for validation. The datasets are then rotated in consecutive tests, allowing for more accurate assessment of best, worst and average classification performance. Based on the magnitude of the dataset and the available computational resources, K was set to 5 in this study. Consequently, the ratio of train to test datasets is 80% to 20%, respectively.

III. RESULTS AND DISCUSSION

The initial stage of research tested the classification performance of a set of machine learning algorithms. The results have been expressed in terms of accuracy, precision, recall, and F1 score. Due to the aforementioned class imbalance, macro averaging was calculated. The outcome of those tests is presented in Table I.

Receiver Operating Characteristics (ROC) curves with corresponding Area Under the Curve (AUC) parameters, displaying the discrimination performance of the tested machine learning models in terms of true positives vs false positives are presented in Fig. 3. The ROC Curve and the ROC AUC score are essential tools for evaluating binary classification models, but they can also be applied to multi-classification problems. OvR method was selected, which stands for "One versus the Rest" and is a method for evaluating multiclass models that evaluates each class in comparison to the others simultaneously. In this scenario, one class is deemed the "positive" class, while the other classes are deemed the "negative" class. This reduces the multiclass classification output to a binary classification output, allowing the use of all known binary classification metrics to assess this scenario [12].

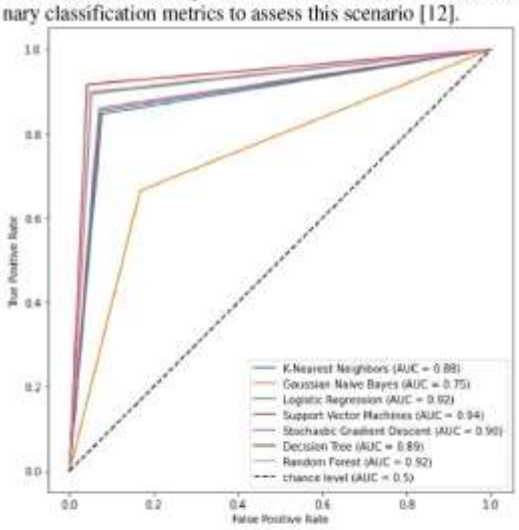


Figure 3. ROC curves with the AUC parameters for machine learning models.

As far as machine learning algorithms are concerned, the best results have been achieved by the Support Vector Machine classifier, which earned 83.38% accuracy. The algorithm also received best scores in precision, recall, F1, and AUC. The Logistic Regression and Random Forest models, which closely followed SVM, also scored above 80% accuracy.

TABLE I.
COMPARISON OF PERFORMANCE RESULTS OF MACHINE LEARNING MODELS. BEST RESULTS IN EACH CATEGORY HAVE BEEN HIGHLIGHTED IN GREEN

Algorithm	Gaussian Naive Bayes	K-Nearest Neighbors	Logistic Regression	Support Vector Machines	Stochastic Gradient Descent	Decision Trees	Random Forest
Accuracy	62.14% (+/- 8.43%)	74.40% (+/- 7.29%)	82.48% (+/- 7.21%)	83.38% (+/- 6.21%)	76.81% (+/- 7.78%)	79.49% (+/- 2.16%)	81.26% (+/- 4.46%)
Precision	87.68% (+/- 9.95%)	92.51% (+/- 5.92%)	94.74% (+/- 5.69%)	94.97% (+/- 4.08%)	90.96% (+/- 7.77%)	92.99% (+/- 5.68%)	94.27% (+/- 4.52%)
Recall	62.14% (+/- 8.43%)	74.40% (+/- 7.29%)	82.48% (+/- 7.21%)	83.38% (+/- 6.21%)	76.81% (+/- 7.78%)	79.49% (+/- 2.16%)	81.26% (+/- 4.46%)
F1	71.06% (+/- 5.32%)	81.12% (+/- 4.51%)	87.38% (+/- 5.62%)	88.05% (+/- 3.76%)	80.51% (+/- 9.62%)	85.16% (+/- 2.35%)	86.58% (+/- 2.70%)

Stochastic Gradient Descent and K-Nearest Neighbors achieved accuracy of 76.81% and 74.40%, respectively, which puts them well behind the three leading methods, but still a league above Gaussian Naive Bayes which scored only 62% accuracy.

It is worth noting that tree-based classifiers have shown the best accuracy stability in terms of 5-Fold validation, with approximately 2% standard deviation in Decision Tree and around 4.5% in Random Forest, whereas for all other models this parameter exceeds 6%. The problem of unbalanced data, which is definitely present in this study, is one of the elements that could have a negative impact on the scores of machine learning algorithms, which is particularly evident e.g. in the poor performance of Gaussian Naive Bayes.

The second phase of research involved deep learning architectures such as FNN, CNN, GNN, and RNN, which were examined using the same criteria as machine learning models. The results of these tests are shown in Table II. The ROC curves with AUC parameters are presented in Fig. 4.

Concerning the tested artificial neural network models, RNN performed best in terms of accuracy, precision, recall, F1 score and AUC, with 94.46% accuracy and 94.45% F1 score. This was to be expected, as the input datasets could be considered sequential data, which is a known strength of RNN [13]. These results also confirm the findings of a recent study [14], which evaluated different neural network designs in order to develop a binary classifier for normal and pathological hearing loss based on similar data, where the best results were also achieved by the RNN architecture. The second best model was CNN with roughly one percentage point less, which may be a little surprising given that CNNs are generally employed to evaluate images. This may be explained by the fact that CNNs perform best when processing data matrices, and the input datasets could be interpreted as small (5x2) matrices. FFN generally achieved third place, while GNN achieved the worst scores.

The overall performance differences between machine and deep learning models are largely in favor of artificial neural networks, with the exception of GNN, which remained at the level of machine learning techniques. The achieved results differ significantly from previous research (performed by Elbaşı and Obali [6]), which achieved 95.5 % accuracy in classifying raw audiometry data with Decision Tree. It should be noted, however, that the validity of those results may be questioned because they were obtained on only 200 samples, which is 20 times less than the dataset used in the current work. Furthermore, there is no information on the class proportion and the employed cross validation process.

TABLE II.
COMPARISON OF PERFORMANCE RESULTS OF DEEP LEARNING MODELS. BEST RESULTS IN EACH CATEGORY HAVE BEEN HIGHLIGHTED IN GREEN

Model	FFN	CNN	GNN	RNN
Accuracy	89.67% (+/-2.12%)	93.46% (+/- 0.83%)	83.15% (+/- 9.09%)	94.46% (+/-0.91%)
Precision	90.27% (+/-1.78%)	93.50% (+/- 0.83%)	86.04% (+/- 4.68%)	94.50% (+/- 0.91%)
Recall	89.67% (+/-2.12%)	93.46% (+/- 0.83%)	83.15% (+/- 9.09%)	94.46% (+/- 0.91%)
F1	89.71% (+/-2.09%)	93.46% (+/- 0.83%)	82.15% (+/- 11.02%)	94.45% (+/- 0.91%)

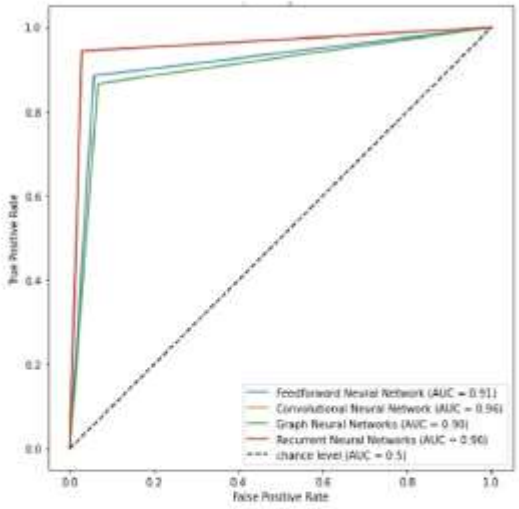


Figure 4. ROC curves with the AUC parameters for deep learning models.

In the above context, while best accuracy of 94,46%, achieved by RNN, is lower than the current state of the art in classification of audiometry test results (97.5%) held by Crowson et al. [7] for raster datasets, that score could be put in question as well. The most significant challenge with training deep learning models from scratch is that it must be done on a large dataset, or else it may miss important patterns. Reliable training of ANN classification models usually requires datasets consisting of at least 10000 samples. For raster datasets this may be alleviated somewhat by employing augmentation of much smaller datasets (which was the strategy applied by Crowson et al. [7]). Unfortunately, this method works best if the input dataset was sufficiently representative. In this case, various types of audiometry software can generate significantly different images, ranging from minor differences in plot color and measurement point indicator size to changes that can significantly impair the performance of an automated classifier, such as displaying test results from both ears on a single plot. As a result, unless an appropriately comprehensive audiogram database is constructed (which would require collection and classification of hundreds of thousands of audiograms produced by all types of audiometry software), image-trained classification models will only work with certain types of audiometry data. In comparison, a classifier which operates on raw audiometry data allows for more flexible and wider application in the clinical environment. This being said, the best classification accuracy of 94,46%, which was achieved in this test by RNN, could be considered too low for clinical application due to a prohibitively large number of false negatives. The latter would suggest that producing a reliably accurate raw audiometry data classifier will require constructing an appropriately large and representative training dataset.

IV. CONCLUSION

The presented work aimed to test several AI-based algorithms for classification of discrete tonal audiometry data series into three types of hearing loss: sensorineural, conductive, and mixed. In the course of this study, several different machine and deep learning models, including Gaussian Naive Bayes, K-Nearest Neighbors, Logistic Regression, Support Vector Machines, Stochastic Gradient Descent, Decision Trees, Random Forest, Feedforward Neural Network, Convolutional Neural Network, Graph Neural Network, and Recurrent Neural Network, have been trained and tested with the use of 4007 audiometry data series analyzed and classified by professional audiologists. The highest classification accuracy was achieved with Recurrent Neural Network at 94.46% (+/- 0.91%). The results of the study verified the general hierarchy of classification performance established by prior research, however they also suggest that the previously reported levels of classification accuracy (achieved for vastly inferior dataset sizes) might have been overly optimistic. In the above context, further work will concentrate on expanding the dataset and improving RNN models in terms of accuracy.

REFERENCES

[1] World Health Organization. 2021. World report on hearing. https://www. who.int/publications/i/item/world-report-on-hearing.
[2] Guo, R., Liang, R., Wang, Q. et al. 2023. Hearing loss classification algorithm based on the insertion gain of hearing aid. Multimed Tools Appl. http://dx.doi.org/10.1007/s11042-023-14886-0
[3] Belitz, C., Ali, H., Hansen, J. H. L. 2019. A Machine Learning Based Clustering Protocol for Determining Hearing Aid Initial Configurations from Pure-Tone Audiograms. Interspeech, 2325–2329, http://dx.doi.org/10.21437/interspeech.2019-3091
[4] Elkhouly, A., Andrew, A.M., Rahim, H.A. et al. 2023. Data-driven audiogram classifier using data normalization and multi-stage feature selection. Sci Rep 13, 1854, http://dx.doi.org/10.1038/s41598-022-25411-y
[5] Margolis, R. H., Saly, G. L. 2007. Toward a standard description of hearing loss. International journal of audiology, 46(12), 746–758, http://dx.doi.org/10.1080/14992020701572652
[6] Elbasi, E., Obali, M. 2012. Classification of Hearing Losses Determined through the Use of Audiometry using Data Mining, Conference: 9th International Conference on Electronics,Computer and Computation
[7] Crowson, M.G., Lee J.W., Hamour A., Mahmood, R., Babier, A., Lin, V., Tucci, D.L., Chan, T.C.Y. 2020. AutoAudio: Deep Learning for Automatic Audiogram Interpretation. J Med Syst. 44(9):163, http://dx.doi.org/10. 1007/s10916-020-01627-1
[8] Barbour, D. L., Wasmann, J. W. 2021. Performance and Potential of Machine Learning Audiometry, The Hearing Journal: Volume 74 - Issue 3 - p 40,43,44, http://dx.doi.org/10.1097/01.HJ.0000737592.24476.88
[9] Guidelines for manual pure-tone threshold audiometry. (1978). ASHA, 20(4), 297–301
[10] Ciszkiewicz A., Milewski G., Lorkowski J., 2018. Baker's Cyst Classification Using Random Forests, 2018 Federated Conference on Computer Science and Information Systems (FedCSIS), Poznan, Poland, 2018, pp. 97-100, http://dx.doi.org/10.15439/2018F89
[11] Kučera E., Haffner O., Stark E., 2017. A method for data classification in Slovak medical records, 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), Prague, Czech Republic, 2017, pp. 181-184, http://dx.doi.org/10.15439/2017F44.
[12] Landgrebe, T.C., Duin, R.P. 2006. A simplified extension of the Area under the ROC to the multiclass domain

[13] Al-Askar, H., Radi, N. MacDermott, A. 2016. Chapter 7 - Recurrent Neural Networks in Medical Data Analysis and Classifications, In Emerging Topics in Computer Science and Applied Computing, Applied Computing in Medicine and Health, Morgan Kaufmann,147-165, 9780128034682, http://dx.doi.org/10.1016/B978-0-12-803468-2.00007-2

[14] Kassjański, M., Kulawiak, M., Przewoźny, T. 2022. Development of an AI-based audiogram classification method for patient referral, 17th Conference on Computer Science and Intelligence Systems (FedCSIS), Sofia, Bulgaria, pp. 163-168, http://dx.doi.org/10.15439/2022F66.

P3. Publication P3

Author Contribution Statement

I declare that in the publication:

M. Kassjański, M. Kulawiak, T. Przewoźny, D. Tretiakow, , J. Kuryłowicz, A. Molisz, K. Koźmiński, A. Kwaśniewska, P. Mierzwińska-Dolny, M. Grono, "Efficiency of Artificial Intelligence Methods for Hearing Loss Type Classification: an Evaluation," Journal of Automation, Mobile Robotics and Intelligent Systems - JAMRIS, 28-38. https://doi.org/10.14313/jamris/3-2024/19. (2024)

my contribution, in accordance with CRediT (Contributor Role Taxonomy) was as follows: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization. Furthermore, my contribution percentage to the development of the publication was 70%.

01.07.2025	Michał Kassjański
Date	Michał Kassjański

I, the undersigned, hereby certify that the information given by Michał Kassjański is correct.

01.07.2025	Marcin Kulawiak	19.08.2025	Krzysztof Koźmiński
Date	Marcin Kulawiak	Date	Krzysztof Koźmiński
16.07.2025		20.07.2025	
Date	Tomasz Przewoźny	Date	Aleksandra Kwaśniewska
15.07.2025		18.08.2025	
Date	Dmitry Tretiakow	Date	Paulina Mierzwińska-Dolny
26.08.2025		16.07.2025	M. Grono
Date	Jagoda Kuryłowicz	Date	Miłosz Grono
16.07.2025			
Date	Andrzej Molisz		

sciendo

VOLUME 18, N° 3 2024
Journal of Automation, Mobile Robotics and Intelligent Systems

Efficiency of Artificial Intelligence Methods for Hearing Loss Type Classification: An Evaluation

Submitted: 9th December 2023; accepted: 26th March 2024

Michał Kassjański, Marcin Kulawiak, Tomasz Przewoźny, Dmitry Tretiakow, Jagoda Kuryłowicz, Andrzej Molisz, Krzysztof Koźmiński, Aleksandra Kwaśniewska, Paulina Mierzwińska-Dolny, Miłosz Grono

DOI: 10.14313/JAMRIS/3-2024/19

Abstract:
The evaluation of hearing loss is primarily conducted by pure tone audiometry testing, which is often regarded as the gold standard for assessing auditory function. This method enables the detection of hearing impairment, which may be further identified as conductive, sensorineural, or mixed. This study presents a comprehensive comparison of a variety of AI classification models, performed on 4007 pure tone audiometry samples that have been labeled by professional audiologists in order to develop an automatic classifier of hearing loss type. The tested models include random forest, support vector machines, logistic regression, stochastic gradient descent, decision trees, convolutional neural network (CNN), feedforward neural network (FNN), recurrent neural network (RNN), gated recurrent unit (GRU) and long short-term memory (LSTM). The presented work also investigates the influence of training dataset augmentation with the use of a conditional generative adversarial network on the performance of machine learning algorithms, and examines the impact of various standardization procedures on the effectiveness of deep learning architectures. Overall, the highest classification performance was achieved by LSTM, with an out-of-training accuracy of 97.56%.

Keywords: *classification, hearing loss types, pure-tone audiometry, RNN, LSTM, evaluation*

1. Introduction

Hearing is regarded as a vital sensory organ, as it furnishes us with crucial insights into our surroundings. It enhances our perception of the environment by complementing our visual and tactile senses, thereby facilitating an extensive comprehension of our environments. Furthermore, possessing adequate auditory perception allows us to engage in effective communication, maintain our safety, and receive gratification from a diverse range of audio activities, such as listening to music or watching theatrical performances.

In consequence, hearing loss has wide-ranging and significant consequences, which encompass, inter alia, the inability to engage in communication with others, as well as a delay in the acquisition of language skills in youngsters. This can result in social isolation, which in turn may lead to feelings of loneliness and frustration, especially in elderly individuals experiencing impaired hearing. According to data presented by the World Health Organization (WHO), the current global prevalence of hearing loss affects more than 1.5 billion people, of which 430 million suffer from moderate to severe hearing loss in their superior ear. As stated by the WHO, it is projected that by 2050, almost 2.5 billion individuals would experience varying levels of hearing impairment, and at least 700 million of them will need rehabilitation treatments [1]. At the same time, however, WHO also claims that almost half of all cases of hearing loss can be avoided by implementing public health interventions. Additional reductions in hearing impairment can be achieved by conducting screenings and implementing early interventions during childhood, such as utilizing assistive devices or considering surgical alternatives.

The evaluation of hearing loss is primarily conducted by pure tone audiometry testing, which has been considered as the most dependable approach for assessing auditory function. The procedure involves presenting pure tones at specific frequencies, either through headphones (air conduction) or by using a vibrator placed on the mastoid section of the temporal bone (bone conduction). The objective is to find the lowest level at which the individual can perceive the sound, known as the threshold, for each frequency [2]. The results of a hearing test are presented on an audiogram, which allows for the identification of the particular type and degree of hearing impairment.

In medical practice, the classification of hearing loss is determined by the configuration, severity, type (location of lesion), and symmetry found in the outcomes of pure-tone audiometry examinations.

The type of hearing loss may be categorized as conductive loss, which is caused by problems in the outer or middle ear, or sensorineural loss, which is a result of difficulties in the inner ear and auditory nerve. Alternatively, it could be a combination of both, known as mixed hearing loss. This classification must be performed by professional audiologists after each pure tone audiometry test. Particularly problematic on a global scale is the scarcity of specialized audiologists; in nearly 93% of low-income nations, there is fewer than one audiologist per million citizens [1].

Given the financial and social obstacles in reducing the large discrepancy between the demand and supply of hearing specialists, it is important to investigate the capability of artificial intelligence (AI) methods in resolving this issue. An automated decision support system could potentially offer a range of benefits, from minimizing human errors to entirely expediting the evaluation of pure-tone audiometry tests to general practitioners. The development of such a system could lead to a reduction in the workload required by specialists and a decrease in the waiting time for patients' diagnoses. Moreover, practical application of such a system would necessitate the establishment of clinical guidelines and best practices, ensuring that healthcare providers adhere to a uniform treatment process, improving patient diagnosis and decreasing treatment variability.

In the above context, the paper presents a comparison of machine learning and deep learning methods applied to the classification of 4007 tonal audiometry test results that were previously analyzed and labeled by expert audiologists. The objective of this study was to examine the efficacy of different artificial intelligence (AI) techniques when utilized with raw tone audiometry data. The latter is particularly significant because pre-classified pure tone audiometry data is relatively difficult to obtain in large quantities, which is why no prior works had the opportunity to perform an in-depth classification using state-of-the-art methods.

Furthermore, the presented work will serve as a basis for selecting an optimal model for classifying different types of hearing loss in clinical settings.

This article is an extension of the research presented in the 18th Conference on Computer Science and Intelligence Systems FedCSIS 2023 during the Doctoral Symposium—Recent Advances in Information Technology (DS-RAIT) [3]. The study was expanded to include several new AI models and provide a more thorough assessment of the applied deep learning algorithms, including an examination of the impact of various data preprocessing methods. Moreover, the extended paper also discusses the effects of expanding the training dataset with the use of a generative adversarial network (GAN).

2. Literature Review

Research on automatic audiometry data classification has been ongoing for an extended period of time. In past years, several endeavors have been made to develop an automatic classification system that is sufficiently accurate to justify its practical implementation. The papers can be categorized into two primary themes: one related to the determination of initial configurations of hearing aids, and the other focused on the classification of hearing loss types. In the literature there are numerous publications that discuss the former subject [4–6]; however, the subject of automatic classification of different forms of hearing loss is substantially less explored.

The first attempt at an automated classifier of hearing loss types was done by Elbaşı and Obali in 2012 [7] who carried a comparative analysis of various methods for identifying the type of hearing loss, including the implementation of multilayer perceptron (MLP) mode classifiers, Decision Tree C4.5, and Naive Bayes. The investigation was conducted on a dataset of 200 samples, which were classified in four distinct groups: normal hearing, sensorineural hearing loss, conductive hearing loss, and mixed hearing loss. The input data was formatted as a sequence of numerical values that represented decibels, which corresponded to constant frequency levels. The Decision Tree (C4.5) approach produced an accuracy of 95.5%, the Naive Bayes method achieved an accuracy of 86.5%, and the MLP algorithm obtained an accuracy of 93.5%.

A different method, which focused on raster images instead of tabular data, was presented several years later by Crowson et al. (2020) [8], who classified audiogram images using the ResNet model into three distinct hearing loss categories (conductive, sensorineural, or mixed) in addition to normal hearing. A dataset consisting of 1007 audiograms was utilized for both training and testing objectives. Instead of starting the classifier training process from the beginning, the scientists implemented transfer learning for training the classifier by utilizing well-established raster classification models. The classification accuracy of this approach reached 97.5%.

Overall, the integration of machine learning with enhanced computational resources in cutting-edge hardware architectures holds the promise of producing quicker overall test outcomes and more comprehensive assessments in the field of audiology [9]. Regarding the categorization of hearing loss types, the currently suggested methods exhibit classification accuracy ranging from 86% to 97%. Although this accuracy is remarkably high, it still allows for a significant margin of error. Furthermore, although the audiogram classifier developed by Crowson et al. [8] demonstrated the highest accuracy thus far, it is not suitable for analyzing the original tabular data generated by tonal audiometry, as it is designed only for image classification. Prior to classification, the datasets must be transformed into a specific format of audiogram images. Although audiograms generally have a similar structure, those produced by different tools can significantly differ in form and content. Some audiometry software generates individual audiograms for each ear, whereas others combine the data from both into just one audiogram. This poses a considerable difficulty when attempting to analyze all cases in a comprehensive manner. Hence, an image classifier is not suitable as the central component of a flexible system for categorizing pure tone audiometry results.

In addition, the aforementioned studies which attempted to create hearing loss classifiers were conducted using very small datasets. The sample sizes in the studies conducted by Elbaşı and Obali [7] and Crowson et al. [8] ranged from 200 to 1007 test results, respectively. With larger datasets, AI models can effectively capture a greater number of unique cases of hearing loss, resulting in more unbiased outcomes.

3. Methodology

The objective of this study was to evaluate the effectiveness of several artificial intelligence (AI) techniques in classification of pure tone audiometry data. The performance of different algorithms was evaluated by means of the accuracy with which each sample was classified as sensorineural hearing loss (S), mixed hearing loss (M), or conductive hearing loss (C) by each method.

3.1. Data

The study employed a dataset consisting of 4007 samples, which included the results of pure tone audiometry tests conducted by doctors at the Department of Otolaryngology of the University Clinical Centre in Gdansk between 2017 and 2021. Figure 1 illustrates the distribution of the data across different classes. There are 674 examples of conductive hearing loss, 1594 instances of mixed hearing loss, and 1739 samples of sensorineural hearing loss. The class imbalance arises from the patient treatment protocols implemented by medical institutions. Conductive hearing loss typically results from pathology affecting the ear canal, obstructing the passage of air. The diagnosis of this condition is usually made with an otoscope during the initial examination of the patient, thus eliminating the requirement for a pure-tone audiometry test.

Each patient contributed a maximum of two examination results, with one result assigned to the left ear and the other to the right ear, therefore eliminating any data redundancy for the same patient and assuring a sufficient diversity of data.

The hearing of the patients was assessed using pure tone audiometry in accordance with the guidelines set forth by the American Speech-Language-Hearing Association (ASHA) [10]. Every experiment was performed within soundproof enclosures (ISO 8253, ISO 8253). The TDH39P headphones were used for air conduction testing, while the Radioear B-71 bone-conduction vibrator was employed for bone conduction testing.

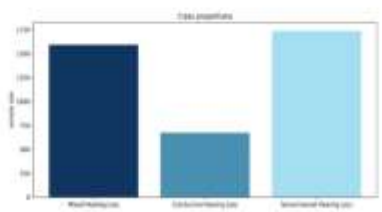


Figure 1. The class proportions in the input dataset

Alongside an audiogram, which is a standard visual representation of pure-tone audiometry test findings, audiology software produces XML files that contain comprehensive data on the tonal points in the audiogram. This study employs XML files containing raw audiometry data, concentrating on five fundamental frequencies (250 Hz, 500 Hz, 1000 Hz, 2000 Hz and 4000 Hz) acquired using both bone as well as air conduction.

3.2. Dataset Expansion

Because the size of the training dataset is rather small for machine learning standards, during the presented research this database was expanded through the application of a conditional generative adversarial network [11]. A generative adversarial network (GAN) is a deep learning network that has the ability to produce data that closely resembles the properties of the training data it was provided with. A conditional generative adversarial network (CGAN) is a variant of the GAN architecture that incorporates labels as additional information during the training phase. A CGAN comprises a pair of interconnected networks that undergo joint training:

1) Generator—this network takes a label and a random array as input and produces data that has the same structure as the training data samples associated with the given label.
2) Discriminator—this network aims to categorize observations as "real" or "generated" by using labeled batches of data that include observations from both the training data and the generated data.

In order to train a conditional GAN, it is necessary to concurrently train both networks with the objective of optimizing the performance of both. This involves training the generator to produce data that deceives the discriminator, while simultaneously training the discriminator to accurately differentiate between real and created data.

This research used CTAB-GAN [12] to augment the dataset by a factor of two. The CTAB-GAN is an expanded version of the initial research on CGAN for tabular data [13], enabling the handling of imbalanced data.

3.3. Preprocessing

In the first stage, feature scaling was utilized as a data preparation technique for standardizing the values of features in a dataset to uniform scale. As mentioned in the literature [14, 15], data standardization is advantageous in terms of enhancing efficiency throughout the training phase. This study used the widely used Z-Score (1) standardization approach:

$$Z_{score} = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the raw score, μ is the mean and σ is the standard deviation.

In addition, two more standardization formulas, MinMax (2) and MaxAbs Scaler (3), were tested on deep learning networks

$$Z_{\text{minmax}} = \frac{x - min}{max - min} \quad (2)$$

$$Z_{maxabs} = \frac{x}{|max|} \quad (3)$$

where x is the raw score, min is the minimum value of the feature and max is the maximum value of the feature.

3.4. Machine Learning Models

The research was initiated by evaluating the performance of various machine learning classification methods, including random forest (RF), Gaussian Naive Bayes, support vector machines (SVMs), logistic regression, stochastic gradient descent (SGD), K-nearest neighbors (KNN) and decision tree (DT). The tabular data format was used as the input for all the described algorithms.

All algorithms have been tested with different preprocessing methods, both on the initial as well as expanded dataset.

3.5. Machine Learning Models

The subsequent stage of the investigation entailed evaluating the following ANN architectures: convolutional neural network (CNN), recurrent neural network (RNN) and feedforward neural network (FNN). Furthermore, two of the most widely used RNN concepts, namely long short-term memory (LSTM) and gated recurrent unit (GRU), were evaluated. Both LSTM and GRU attempt to overcome the problem of vanishing gradients by introducing data flow control mechanisms [16].

Previously, these methods had been employed to classify relevant medical data [17, 18].

3.6. Evaluation Process

The performance of all tested models was assessed with the use of K-fold cross-validation. This process entailed partitioning the dataset into K subsets, referred to as folds, where K-1 subsets were allocated for training purposes and one subset was reserved for validation. Following this, the subsets have been sequentially rotated in subsequent tests, which enabled a more precise evaluation of the best, worst, and average performance of the classification. In the presented work the value of K was established at 10 in accordance with the literature standard and the scale of the dataset. Thus, the proportion of training to testing datasets is ten percent to ninety percent. During the evaluation of models, the default 10-fold set was decreased to 90%, with the remaining 10% forming a dedicated test dataset. This has been done to ensure that the performance of models trained with and without data generated with the use of CGAN can be effectively compared.

The general workflow of the presented study is shown in Figure 2.

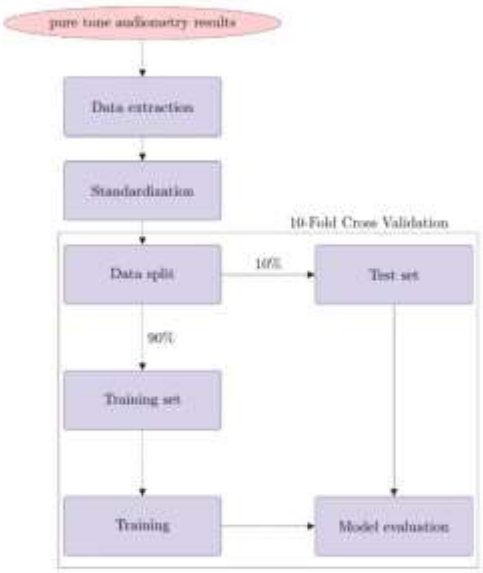


Figure 2. The workflow of the presented research into application of machine learning methods for the classification of hearing loss types based on pure-tone audiometry data

3.7. Evaluation Parameters

In addition to traditional measures such as accuracy, the presented research also employed precision-recall metrics derived from a confusion matrix [19] as well as receiver operating characteristics (ROC) curves which encompass the pertinent area-under-the-curve (AUC) data.

These curves effectively demonstrate the discrimination performance of the evaluated models by comparing true positives and false positives. Furthermore, in addition to evaluating the efficacy of binary classification models, the receiver operating characteristic (ROC) curve and the area under the ROC curve (ROC AUC) score are valuable instruments for assessing multiple classification challenges. The chosen approach is OvR, an acronym for "one versus the rest," which assesses multiclass models by comparing each class to the others simultaneously. In this case, one class is designated as the "positive" class, while the remaining classes are designated as the "negative" class. This transforms the output of multiclass classification into binary classification, enabling the application of established binary classification metrics to evaluate this situation [20].

Table 1. Comparative analysis of performance outcomes of machine learning models without GAN

Algorithm	Gaussian Naive Bayes	K-Nearest Neighbors	Logistic Regression	Support Vector Machines	Stochastic Gradient Descent	Decision Trees	Random Forest
Accuracy	62.34% (± 12%)	77.02% (± 9%)	82.18% (± 9%)	85.15% (± 6%)	74.74% (± 9%)	80.09% (± 4%)	83.03% (± 4 %)
Precision	97.02% (± 4%)	97.34% (± 3 %)	97.92% (± 3%)	97.84% (± 3%)	97.91% (± 3%)	97.65% (± 3%)	97.62% (± 3 %)
Recall	62.34% (± 12 %)	77.02% (± 9 %)	82.18% (± 9%)	85.15% (± 6%)	74.74% (± 9%)	80.09% (± 4 %)	83.03% (± 4 %)
F1	74.68% (± 7%)	84.75% (± 8%)	88.36% (± 8%)	90.31% (± 5%)	83.76% (± 7%)	87.36% (± 4%)	89.12% (± 4%)

Table 2. Comparative analysis of performance outcomes of machine learning models with GAN

Algorithm	Gaussian Naive Bayes	K-Nearest Neighbors	Logistic Regression	Support Vector Machines	Stochastic Gradient Descent	Decision Trees	Random Forest
Accuracy	61.99% (± 10 %) ⇓	75.14% (± 7%) ⇓	86.67% (± 5 %) ⇑	89.52% (± 4 %) ⇑	80.68% (± 13%) ⇑	79.31% (± 2 %) ⇓	83.50% (± 4 %) ⇑
Precision	97.00% (± 4 %) ⇓	97.32% (± 4 %) ⇓	98.37% (± 2 %) ⇑	98.18% (± 2 %) ⇑	97.72% (± 3 %) ⇓	97.66% (± 3 %) ⇑	97.66% (± 3 %) ⇓
Recall	61.99% (± 10 %) ⇓	75.14% (± 7 %) ⇓	86.67% (± 5 %) ⇑	89.52% (± 4 %) ⇑	80.68% (± 13 %) ⇑	79.31% (± 2 %) ⇓	83.50% (± 4 %) ⇓
F1	74.56% (± 6 %) ⇓	83.86% (± 6 %) ⇓	91.75% (± 4 %) ⇑	93.22% (± 3 %) ⇑	87.01% (± 11 %) ⇑	86.91% (± 3 %) ⇓	89.45% (± 4 %) ⇑

4. Results and Discussion

The initial step of the presented study involved evaluation of the classification performance offered by a collection of machine learning algorithms. The outcomes have been evaluated in relation to accuracy, precision, recall, and F1 score. Macro averaging in 10-fold cross validation was used to offset the class imbalance in the training dataset. The test results are presented in Table 1.

The support vector machine classifier has achieved the highest level of success among machine learning algorithms, with an accuracy rate of 85.15%. The algorithm achieved the highest ratings in precision, recall, F1, and AUC. In close pursuit of SVM, the logistic regression and random forest models both exceeded 82% in terms of accuracy.

Stochastic gradient descent achieved an accuracy of 74.74%, while K-nearest neighbors obtained 77.02%, which puts both of them well below the top three algorithms, but still significantly higher than Gaussian Naive Bayes which only reached 62.34% accuracy.

Tree-based classifiers have demonstrated superior accuracy stability in 10-fold validation. The decision tree classifier exhibits a standard deviation of roughly 4%, while the random forest classifier has a standard deviation of around 4.65%. In contrast, all other models have a standard deviation over 6%. The issue of imbalanced data, which is certainly visible in this study, is one of the factors that might adversely affect the effectiveness of machine learning algorithms, as exemplified by the subpar results of Gaussian Naive Bayes.

The results in Table 2 depict the outcomes obtained by augmenting the training set using CTAB-GAN. The application of CGAN yielded positive outcomes for only 4 out of the 7 algorithms that were examined. Doubling the size of training data did not influence the accuracy of Naive Bayes and decision tree, which produced results differing by less than 1 percentage point. The KNN model exhibited a slight reduction in overall classification performance, losing less than 2 percentage points in accuracy and recall. On the other hand, the generation of additional training data resulted in increasing the classification accuracy level in SVMs and logistic regression by approximately 5%. The largest increase, amounting to an 8% increase, is shown in the SGD results as compared to those without CGAN.

This being said, the increase in accuracy, as well as improvements in other measures such as precision, recall, and F1 score shown by all three algorithms could be considered to be within their respective margins of error. In order to sidestep the issue of increased margins of error in the expanded datasets, the classification accuracy of selected methods was tested again on the dedicated test dataset, which had been extracted from the original data before training. Results of these tests are presented in the form of confusion matrices displayed in Figures 3, 4, 5 and Table 3. The matrix on the left depicts the outcomes obtained without the use of CGAN, while the matrix on the right illustrates the results following the implementation of CGAN. The S, M, and C indices represent sensorineural hearing loss, mixed hearing loss, and conductive hearing loss, respectively.

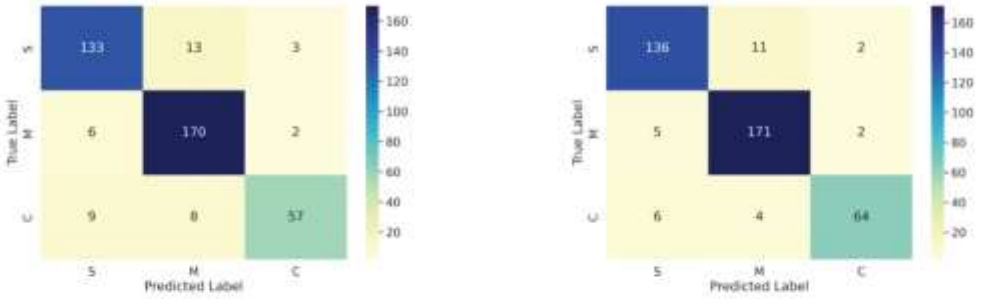


Figure 3. Confusion matrices of the logistic regression model trained without CGAN (left) and with CGAN (right)

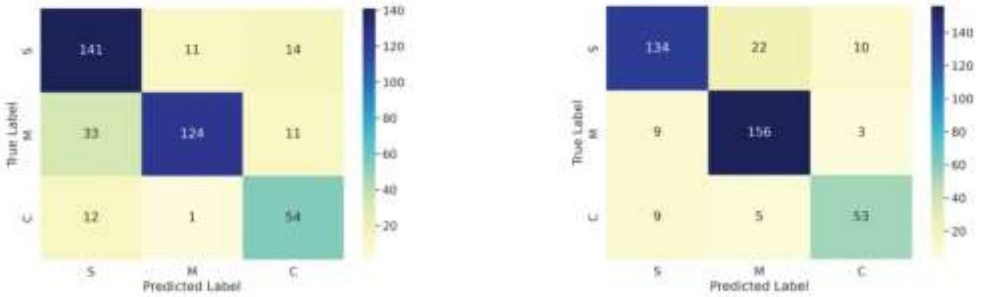


Figure 4. Confusion matrices of the stochastic gradient descent model trained without CGAN (left) and with CGAN (right)

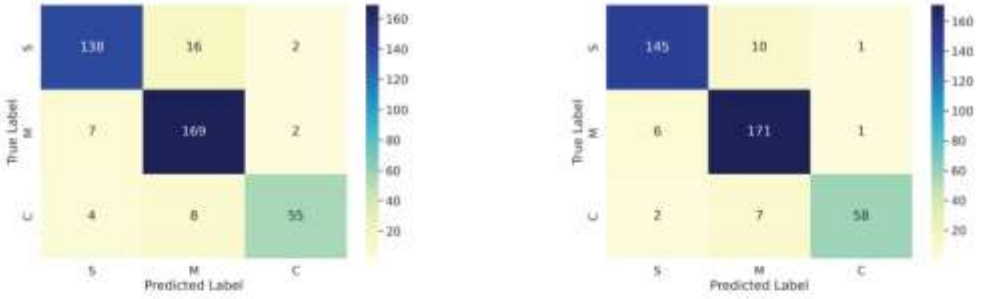


Figure 5. Confusion matrices of the support vector machines model trained without CGAN (left) and with CGAN (right)

Comparing the findings obtained from 10-fold cross validation to those obtained from a dedicated test, there is a similar improvement (Table 3). Logistic regression, support vector machines, and stochastic gradient descent exhibit considerable enhancements in accuracy, similar to the outcomes shown in 10-fold (Table 2). The results for Gaussian Naive Bayes and random forest show minimal variation, with a difference of less than one percentage point The most significant decline was observed in the performance of KNN and decision trees, with a difference of 1.24%, which is still comparable to the results obtained from the 10-fold analysis.

The improvements brought by artificially expanding the training dataset are best visible in the confusion matrices presented in Figures 3, 4, and 5.

In the case of the logistic regression model results depicted in Figure 3, it is noteworthy that, subsequent to the adoption of GAN, the number of conductive hearing loss cases (C) incorrectly labeled as sensorineural and mixed has demonstrated a drop of 30% and 50%, respectively. The improvements to classification of the remaining types are much smaller but persistent, with only the classification of mixed hearing loss as conductive showing no improvements. The performance of Stochastic Gradient Descent model has shown the largest improvements after training with GAN-derived data (Figure 4). The number of mixed hearing loss cases incorrectly classified as sensorineural decreased by 73% (from 33 to 9), while the number of conductive hearing loss cases labeled as sensorineural was reduced by 25% (12 to 9).

Table 3. Comparison of the accuracy of the tested machine learning models trained with and without the use of CGAN, analyzed on the dedicated test dataset

Algorithms	Default training (acc)	Training with CGAN (acc)
Gaussian Naive Bayes	63.09%	63.59% ⇑
K-Nearest Neighbors	80. 29%	79. 05% ⇓
Logistic Regression	89.77%	92.51% ⇑
Support Vector Machines	90.27%	93.04% ⇑
Stochastic Gradient Descent	79.55%	85.53% ⇑
Decision Trees	84. 53%	83.29% ⇓
Random Forest	87.78%	88.02% ⇑

At the same time, the number of sensorineural hearing loss cases improperly recognized as conductive decreased by 29% (from 14 to 10) and the number of mixed hearing loss datasets incorrectly labeled as conductive decreased by 73% (from 11 to 3). However, these gains are offset somewhat by a reduction in the accuracy of mixed hearing loss classification. After training on data generated by GAN, SGD has shown an increased tendency to label mixed hearing loss as either sensorineural (22 cases versus 11, a 100% increase) or conductive (5 cases versus 1, a 400% increase). This being said, the total number of properly recognized datasets still shows a considerable 8% increase (343 from 319).

Out of the three analyzed machine learning models, support vector machines (SVMs) is the only one which shows consistent improvements to all cases of classification inaccuracy after training with GAN-derived data. The number of sensorineural hearing loss cases improperly labeled as mixed and conductive is reduced by 38% (16 to 10) and 50% (2 to 1), respectively. The number of mixed hearing loss cases improperly labeled as sensorineural and conductive is reduced by 14% (7 to 6) and 50% (2 to 1), respectively. Finally, the number of conductive hearing loss cases incorrectly recognized as sensorineural and mixed is reduced by 50% (4 to 2) and 13% (8 to 7), respectively. These improvements increase the total number of correctly classified datasets from 362 to 375.

Given that in the current state of the art, deep learning models surpass the classification accuracy of all machine learning methods, the presented study also evaluated the performance of several deep learning architectures. These include feedforward neural networks (FNN), convolutional neural networks (CNN), and recurrent neural networks (RNN), which encompass gated recurrent units (GRU) and long short-term memory (LSTM). The evaluation was performed using a 10-fold cross-validation methodology, and involved assessment of the impact of implementing different data standardization methods. The results of these experiments are displayed in Tables 4–6.

Table 4. Classification performance of deep learning models using Z-Score normalization

	FNN	CNN	RNN	LSTM	GRU
Accuracy	93.06% (± 1%)	93.76% (± 1%)	94.07% (± 1%)	95.63% (± 1%)	93.83% (± 1%)
Precision	93.10% (± 1%)	93.82% (± 1%)	94.17% (± 1%)	95.68% (± 1%)	93.94% (± 1%)
Recall	93.06% (± 1%)	93.82% (± 1%)	94.07% (± 1%)	95.63% (± 1%)	93.83% (± 1%)
F1	93.0% (± 1%)	93.75% (± 1%)	94.04% (± 1%)	95.63% (± 1%)	93.83% (± 1%)

Table 5. Classification performance of deep learning models using MinMaxScaler normalization

	FNN	CNN	RNN	LSTM	GRU
Accuracy	66.44% (± 3%)	68.06% (± 2%)	68.23% (± 2%)	67.46% (± 1%)	68.95% (± 1%)
Precision	66.43% (± 3%)	57.30% (± 3%)	57.93% (± 3%)	57.69% (± 2%)	58.11% (± 2%)
Recall	66.43% (± 3%)	68.06% (± 2%)	68.23% (± 2%)	67.46% (± 1%)	68.95% (± 1%)
F1	60.09% (± 3%)	61.83% (± 2%)	68.23% (± 2%)	61.18% (± 2%)	62.64% (± 2%)

Table 6. Classification performance of deep learning models using MaxAbsScaler normalization

	FNN	CNN	RNN	LSTM	GRU
Accuracy	39.78% (± 1%)	39.78% (± 1%)	39.78% (± 1%)	39.78% (± 1%)	39.78% (± 1%)
Precision	15.84% (± 1%)	15.85% (± 1%)	15.85% (± 1%)	15.85% (± 1%)	15.85% (± 1%)
Recall	39.78% (± 1%)	39.78% (± 1%)	15.88% (± 1%)	39.78% (± 1%)	39.78% (± 1%)
F1	22.66% (± 1%)	22.66% (± 1%)	22.66% (± 1%)	22.66% (± 1%)	22.66% (± 1%)

As it can be seen in Tables 4–6, normalization strategy plays a fundamental part in obtaining good classification performance using deep learning models. Undoubtedly, the Z-Score normalization method delivered outstanding performance across all architectures (Table 4). These classification accuracy results are on average 35% better than in the case of MinMaxScaler (Table 5) and about 120% better than those produced by MaxAbsScaler (Table 6), which is clearly not suitable for audiometry data.

Concerning the results obtained by all networks with the Z-Score normalization method, LSTM exhibited the highest performance in terms of accuracy, recall, precision and F1 score. Specifically, it achieved an accuracy of 95.63% and an F1 score of 95.63%. It was predictable that the input datasets, being sequential data, would be well-suited for the RNN family of models, which is known for its strength in handling this type of data [18]. The results appear to validate the conclusions of a previous study [21] which assessed several neural network configurations to create a binary classifier for distinguishing between pathological hearing loss and normal hearing using similar data. Said investigation also concluded that the LSTM architecture yielded the most favorable results. The second-best results have been

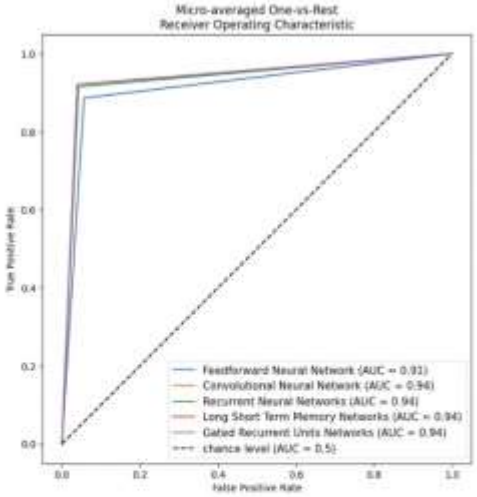


Figure 6. ROC curves with the AUC parameters for tested deep learning models during 10-Fold validation

achieved by the simple RNN model, with a difference of approximately 0.6%. While the difference is within the margin of error, this result is somewhat expected, considering that LSTM models typically offer superior performance over simple RNN models. The third place of the CNN model, which is prominently used for processing raster data, could be explained by the fact that each dataset in the current study is represented by a two-dimensional table which somewhat resembles a very small raster.

The classification performance of the presented deep learning models (Table 4) is visualized in Figure 6 in the form of ROC curves with corresponding AUC parameters. These illustrate the discriminatory capability of the evaluated deep learning models quantified by the ratio of true positives to false positives.

All CNN, RNN, LSTM, and GRU models have the same AUC parameter score of 0.94. With an AUC value of 0.91, the FNN model is conspicuously inferior to the others.

In general, the scaling technique has a substantial impact on the performance of classification models. Furthermore, this impact may vary depending on the specific types of models employed, such as monolithic and ensemble models [22].

Based on these results, all subsequent tests were performed with the use of Z-Score normalization, as it is the sole method that yields outcomes comparable to the state-of-the-art.

The final step of the presented research analyzed the performance of deep learning methods trained on the dataset augmented with the use of CGAN. The results are displayed in Table 7.

Table 7. Performance of deep learning models trained on data augmented with CGAN

	FNN	CNN	RNN	LSTM	GRU
Accuracy	90.64% (±1%) ⇓	90.71% (±1%) ⇓	94.92% (±0.5%)⇑	98.57% (±0.5%) ⇑	95.41% (±0.5%) ⇑
Precision	90.88% (±1%) ⇓	90.95% (±1%) ⇓	94.92% (±0.5%) ⇑	98.58% (±0.5%) ⇑	95.44% (±0.5%) ⇑
Recall	90.64% (±1%) ⇓	90.71% (±1%) ⇓	94.92% (±0.5%) ⇑	98.57% (±0.5%) ⇑	95.41% (±0.5%) ⇑
F1	90.60% (±1%) ⇓	90.74% (±1%) ⇓	94.92% (±0.5%) ⇑	98.57% (±0.3%) ⇑	95.41% (±0.5%) ⇑

Table 8. Comparison of the performance of deep learning models trained with and without the use of CGAN, analyzed on the dedicated test dataset

Models	Default training (acc)	Training with CGAN (acc)
FNN	95.48%	91.66% ⇓
CNN	92.01%	88.19% ⇓
RNN	93.40%	94.44% ⇑
LSTM	94.79%	**97.56%** ⇑
GRU	92.70%	92.70% ⇔
FNN	95.48%	91.66% ⇓

As it can be seen in Table 7, training on the expanded dataset has significantly increased the performance of certain deep learning models while impacting the performance of others, which mirrors the situation with machine learning algorithms. In particular, the classification accuracy of recurrent networks has increased by nearly 1% in the case of RNN, around 1.5% for GRU and nearly 3% for LSTM. On the other hand, the classification effectiveness of FNN and CNN has reduced by nearly 3%. This being said, considering the potential impact of testing the networks on CGAN-augmented data (which has been shown previously for machine learning methods), a subsequent analysis was conducted using the dedicated test set. The results of this test are presented in Table 8.

Similarly to the case of machine learning models, testing on the dedicated dataset yields similar overall results, however with somewhat different performance values. The performances of LSTM and RNN models have shown an increase, whereas those of FNN and CNN experienced a decline. An exception to this correlation is the GRU model, as its findings remain consistent regardless of the approach used. The LSTM model achieved the highest accuracy, reaching 97.56%. This result is lower by one percentage point compared to the figure reported in Table 7 for the 10-fold with GAN approach.

In general, artificial neural networks exhibit superior performance to deep learning models when comparing the two. However, the utilization of CGAN for training machine learning methods enables some of them to come closer to the accuracy delivered by the less performant deep learning methods. Still, the optimal outcomes are achieved by RNN-based models with Z-Score normalization and GAN augmentation, in particular simple RNN and LSTM models.

The achieved results significantly exceed those of prior investigations (conducted by Elbaşı and Obali [7]), which utilized a Decision Tree to classify raw audiometry data with an accuracy of 95.5%. Interestingly, when evaluated on the presented data, the same Decision Tree algorithm achieved an accuracy of approximately 83% on the dedicated test dataset. Yet, the validity of the cited findings may be questioned due to the limited sample size of just 200, which is significantly smaller than the dataset employed in the present study. Moreover, the results cannot be directly compared because the cited study was conducted on four classes (as opposes to three classes in the presented work), which included individuals with normal hearing, and there is no data regarding class distribution nor the method used for cross-validation.

At the same time, the greatest classification accuracy of 97.56% attained by LSTM on the dedicated test dataset is comparable to the present state of the art in classifying pure tone audiometry test results (97.5%) reported by Crowson et al. [8] for raster datasets. Similar to that work, training data augmentation has provided significantly better classification results (although the presented work augmented tabular data, whereas Crowson et al. augmented raster data). Again, these results cannot be directly compared due to the lower number of classes (three instead of four) used in the presented study. Moreover, Crowson et al. [8] classified raster audiograms instead of actual test results, and images produced by different types of audiometry software vary significantly. These variations can range from minor differences in the color of the plot and the size of the measurement point indicators to more significant changes that may adversely affect the performance of automated classifiers (e.g., presenting outcomes from both ears on a solitary plot). In order for image-trained classification models to be effective with all types of audiometry data, it is necessary to create a comprehensive audiogram database. This would include collecting and classifying thousands of audiograms created by different audiometry applications. By contrast, a classifier that utilizes unprocessed audiometry data offers greater versatility and broader potential for use in the clinical setting.

On the whole, despite attaining a relatively high classification accuracy of 97.56%, the presented LSTM-based classifier may not be adequate for clinical use due being trained on data augmented with CGAN. While this data has significantly improved the performance of certain classifiers, it has also decreased the performance of other methods, suggesting that not all of the generated datasets may properly reflect real-world audiometry data. Therefore, the creation of a reliable and precise classifier for raw audiometry data necessitates the establishment of a training dataset that is sufficiently large and representative, while also being closely controlled by medical experts.

5. Conclusion

The objective of the presented study was to assess the efficacy of different artificial intelligence algorithms in classifying discrete tonal audiometry data series into three specific types of hearing loss: conductive, sensorineural, and mixed. For this purpose, the study involved testing machine and deep learning models comprised of Gaussian Naive Bayes, support vector machines, random forest, K-nearest neighbors, logistic regression, stochastic gradient descent, decision trees, feedforward neural network, convolutional neural network and recurrent neural network (including long short-term memory and gated recurrent unit). The models indicated above have been trained and assessed using 4007 sets of tonal audiometry data, which had been analyzed and labeled by audiologists who are experts in the field.

Furthermore, the investigation also explored the impact of training dataset augmentation using a conditional generative adversarial network and examined how different standardization procedures affect the effectiveness of deep learning architectures.

The best overall results were obtained with the long short-term memory model, which attained the maximum classification accuracy of 97.56% with Z-Score normalization and CGAN data augmentation. On the whole, all deep learning models achieved substantially better classification results than machine learning algorithms when trained on the standard dataset, but training on the GAN-augmented dataset allowed support vector machines to achieve results similar to that of less performant deep learning models.

Thus, on the one hand the study's findings confirmed the overall ranking of classification performance that earlier research had established. On the other hand, the findings also suggest that the classification accuracy levels previously documented in literature, which were attained using considerably smaller datasets, might have been overly sanguine.

Finally, the results of the presented research indicate that using a GAN augmentation of training data may produce very positive results, however (as exemplified by the performance of the stochastic gradient descent model) unsupervised generation of input data may not always lead to optimal outcomes. In this context, future work could concentrate on enhancing the accuracy of the RNN-based classifier and increasing the size of training dataset as well as designing a GAN model which is more efficiently tuned for producing properly labeled tonal audiometry test data.

In general, the demonstrated outcomes indicate that the proposed AI-driven pure tone audiometry data classifier may have practical implications in clinical settings, functioning as either a classification system for general practitioners or a support system for professional audiologists. In both scenarios, the implementation of the classifier has the potential to minimize human error, enhance diagnostic accuracy, and reduce the waiting time for patients to receive their diagnosis.

AUTHORS

Michał Kassjański* – Department of Geoinformatics, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, 80-233, Gdansk, Poland, e-mail: michal.kassjanski@pg.edu.pl.

Marcin Kulawiak – Department of Geoinformatics, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, 80-233, Gdansk, Poland, e-mail: marcin.kulawiak@eti.pg.edu.pl.

Tomasz Przewoźny – Department of Otolaryngology, Medical University of Gdansk, Smoluchowskiego Str. 17, 80-214 Gdansk, Poland, e-mail: tomasz.przewozny@gumed.edu.pl.

Dmitry Tretiakow – Department of Otolaryngology, the Nicolaus Copernicus Hospital in Gdansk, Copernicus Healthcare Entity, Powstancow Warszawskich str. 1/2, 80-152, Gdansk, Poland, e-mail: d.tret@gumed.edu.pl.

Jagoda Kuryłowicz – Department of Otolaryngology, Medical University of Gdansk, 80-214, Gdansk, Poland, e-mail: jagoda.kurylowicz@gmail.com.

Andrzej Molisz – Department of Otolaryngology, Medical University of Gdansk, 80-214, Gdansk, Poland, e-mail: andrzej.molisz@gumed.edu.pl.

Krzysztof Koźmiński – Student's Scientific Circle of Otolaryngology, Medical University of Gdańsk, 80-214 Gdansk, Poland, e-mail: krzyk@gumed.edu.pl.

Aleksandra Kwaśniewska – Department of Otolaryngology, Laryngological Oncology and Maxillofacial Surgery, University Hospital No. 2, 85-168, Bydgoszcz, Poland, e-mail: kwasniewska.aleks@gmail.com.

Paulina Mierzwińska-Dolny – Student's Scientific Circle of Otolaryngology, Medical University of Gdańsk, 80-214 Gdansk, Poland, e-mail: paulinamierzwinska@gumed.edu.pl.

Miłosz Grono – Student's Scientific Circle of Otolaryngology, Medical University of Gdańsk, 80-214 Gdansk, Poland, e-mail: milosz.grono@gumed.edu.pl.

*Corresponding author

References

[1] World Health Organization, *World report on hearing*. Geneva: World Health Organization, 2021.

[2] R. W. Baloh and J. C. Jen, "Hearing and Equilibrium," Jan. 2012, doi: 10.1016/b978-1-4377-1604-7.00436-x.

[3] M. Kassjański et al., "Detecting type of hearing loss with different AI classification methods: a performance review," *Computer Science and Information Systems (FedCSIS), 2019 Federated Conference*, Sep. 2023, doi: 10.15439/2023f3083.

[4] C. Belitz, H. Ali, and J. Hansen, "A Machine Learning Based Clustering Protocol for Determining Hearing Aid Initial Configurations from Pure-Tone Audiograms," *PubMed Central*, Sep. 2019, doi: 10.21437/interspeech.2019-3091.

[5] F. Charih, M. Bromwich, A. E. Mark, R. Lefrançois, and J. R. Green, "Data-Driven Audiogram Classification for Mobile Audiometry," *Scientific Reports*, vol. 10, no. 1, Mar. 2020, doi: 10.1038/s41598-020-60898-3.

[6] A. Elkhouly et al., "Data-driven audiogram classifier using data normalization and multi-stage feature selection," *Scientific Reports*, vol. 13, no. 1, Feb. 2023, doi: 10.1038/s41598-022-25411-y.

[7] E. Elbaşı and M. Obali, "Classification of Hearing Losses Determined through the Use of Audiometry Using Data Mining," *Conference: 9th International Conference on Electronics,Computer and Computation*.

[8] M. G. Crowson et al., "AutoAudio: Deep Learning for Automatic Audiogram Interpretation," *Journal of Medical Systems*, vol. 44, no. 9, Aug. 2020, doi: 10.1007/s10916-020-01627-1.

[9] H. Shojaeemend and H. Ayatollahi, "Automated Audiometry: A Review of the Implementation and Evaluation Methods," *Healthcare Informatics Research*, vol. 24, no. 4, pp. 263–275, Oct. 2018, doi: 10.4258/hir.2018.24.4.263.

[10] Guidelines for Manual Pure-Tone Threshold Audiometry," *American Speech-Language-Hearing Association*. https://www.asha.org/policy/GL2005-00014/ (accessed Dec. 5, 2023).

[11] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," *arXiv.org*, 2014. https://arxiv.org/abs/1411.1784.

[12] Z. Zhao, A. Kunar, Van, R. Birke, and L. Y. Chen, "CTAB-GAN: Effective Table Data Synthesizing," *arXiv (Cornell University)*, Feb. 2021.

[13] L. Xu et al., "Modeling Tabular Data using Conditional GAN." Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf (accessed Dec 5, 2023).

[14] A. M. Annaswamy and Massoud Amin, *IEEE Vision for Smart Grid Controls: 2030 and Beyond*. Piscataway, Usa Ieee, 2013.

[15] M. Shanker, M. Y. Hu, and M. S. Hung, "Effect of data standardization on neural network training," *Omega*, vol. 24, no. 4, pp. 385–397, Aug. 1996, doi: 10.1016/0305-0483(96)00010-2.

[16] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[17] I. Banerjee et al., "Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification," *Artificial Intelligence in Medicine*, vol. 97, pp. 79–88, Jun. 2019, doi: 10.1016/j.artmed.2018.11.004.

[18] "Recurrent Neural Networks in Medical Data Analysis and Classifications," *Applied Computing in Medicine and Health*, pp. 147–165, Jan. 2016, doi: 10.1016/B978-0-12-803468-2.00007-2.

[19] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters*, vol. 30, no. 1, pp. 27–38, Jan. 2009, doi: 10.1016/j.patrec.2008.08.010.

[20] D. J. Hand and R. J. Till, "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems," *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001, doi: 10.1023/a:1010920819831.

[21] M. Kassjański, M. Kulawiak, and Tomasz Przewoźny, "Development of an AI-based audiogram classification method for patient referral," *Computer Science and Information Systems (FedCSIS), 2019 Federated Conference on*, Sep. 2022, doi: 10.15439/2022f66.

[22] L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, "The choice of scaling technique matters for classification performance," *Applied Soft Computing*, vol. 133, p. 109924, Jan. 2023, doi: 10.1016/j.asoc.2022.109924.

P4. Publication P4

Author Contribution Statement

I declare that in the publication:

M. Kassjański, M. Kulawiak, T. Przewoźny, D. Tretiakow, , J. Kuryłowicz, A. Molisz, K. Koźmiński, A. Kwaśniewska, P. Mierzwińska-Dolny, M. Grono, "Automated hearing loss type classification based on pure tone audiometry data," Scientific Reports, 14, 14203. https://doi.org/10.1038/s41598-024-64310-2. (2024)

my contribution, in accordance with CRediT (Contributor Role Taxonomy) was as follows: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization. Furthermore, my contribution percentage to the development of the publication was 70%.

01.07.2025	Michał Kassjański
Date	Michał Kassjański

I, the undersigned, hereby certify that the information given by Michał Kassjański is correct.

01.07.2025	Marcin Kulawiak	19.08.2025	Krzysztof Koźmiński
Date	Marcin Kulawiak	Date	Krzysztof Koźmiński
16.07.2025	Tomasz Przewoźny	20.07.2025	Aleksandra Kwaśniewska
Date	Tomasz Przewoźny	Date	Aleksandra Kwaśniewska
15.07.2025	Dmitry Tretiakow	18.08.2025	Paulina Mierzwińska-Dolny
Date	Dmitry Tretiakow	Date	Paulina Mierzwińska-Dolny
26.08.2025	Jagoda Kuryłowicz	16.07.2025	M. Grono
Date	Jagoda Kuryłowicz	Date	Miłosz Grono
16.07.2025	Andrzej Molisz		
Date	Andrzej Molisz		

scientific reports

OPEN

Automated hearing loss type classification based on pure tone audiometry data

Michał Kassjański[1✉], Marcin Kulawiak[1], Tomasz Przewoźny[2], Dmitry Tretiakow[3], Jagoda Kuryłowicz[2], Andrzej Molisz[2], Krzysztof Koźmiński[2], Aleksandra Kwaśniewska[4], Paulina Mierzwińska-Dolny[2] & Miłosz Grono[2]

Hearing problems are commonly diagnosed with the use of tonal audiometry, which measures a patient's hearing threshold in both air and bone conduction at various frequencies. Results of audiometry tests, usually represented graphically in the form of an audiogram, need to be interpreted by a professional audiologist in order to determine the exact type of hearing loss and administer proper treatment. However, the small number of professionals in the field can severely delay proper diagnosis. The presented work proposes a neural network solution for classification of tonal audiometry data. The solution, based on the Bidirectional Long Short-Term Memory architecture, has been devised and evaluated for classifying audiometry results into four classes, representing normal hearing, conductive hearing loss, mixed hearing loss, and sensorineural hearing loss. The network was trained using 15,046 test results analysed and categorised by professional audiologists. The proposed model achieves 99.33% classification accuracy on datasets outside of training. In clinical application, the model allows general practitioners to independently classify tonal audiometry results for patient referral. In addition, the proposed solution provides audiologists and otolaryngologists with access to an AI decision support system that has the potential to reduce their burden, improve diagnostic accuracy, and minimise human error.

Keywords Classification, Bi-LSTM, Hearing loss, Tonal audiometry, Audiogram, AI decision support system

Hearing is a key sense in human daily existence, allowing for connectivity with the outside world in a manner that none of our other senses can accomplish. Aside from enabling efficient communication with others, good hearing is crucial for personal safety, e.g. when crossing the street on foot, operating a vehicle, or responding to a fire alarm, frequently enabling detection of a potential threat before it becomes visible. Other benefits that good hearing may bring to quality of life, such as listening to music, television and radio, also should not be overlooked. Extreme cases of communication difficulties, resulting in a decline in quality of life, may lead to psychiatric disorders such as depression[1].

According to the World Health Organization (WHO), hearing loss currently affects more than 1.5 billion people worldwide, of whom 430 million have moderate or higher levels of hearing loss in the better hearing ear. WHO predicts that by 2050, nearly 2.5 billion people will have some degree of hearing loss, with at least 700 million requiring rehabilitation services. Fortunately, early detection and efficient management can significantly mitigate numerous instances of hearing impairment, particularly those associated with childhood hearing loss. Medical and surgical methods can be effective in the treatment of ear diseases, in many cases leading to restoration of original hearing quality[2].

Hearing loss is predominantly determined with the use of pure-tone audiometry, typically performed while seated in a sound-proof chamber. It involves delivering a series of increasingly-intense pure tones at predetermined threshold levels, typically via headphones, and determining the auditory threshold for air and bone conduction. Air conduction determines the function of the complete auditory organ, from the auricle to the temporal lobe hearing centres. Any level of damage to this system decreases the air conduction curve. Bone conduction examines the organ of hearing from the level of the bony capsule of the cochlea, bypassing the conduction of

[1]Department of Geoinformatics, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, G. Narutowicza 11/12, 80-233 Gdańsk, Poland. [2]Department of Otolaryngology, Medical University of Gdańsk, Gdańsk, Poland. [3]Department of Otolaryngology, The Nicolaus Copernicus Hospital in Gdańsk, Copernicus Healthcare Entity, Gdańsk, Poland. [4]Department of Otolaryngology, Laryngological Oncology and Maxillofacial Surgery, University Hospital No. 2, Bydgoszcz, Poland. ✉email: michal.kassjanski@pg.edu.pl

sound through the outer and middle ear. It serves as an alternative pathway for sound conduction, although it is not as significant as air conduction. Using pure-tone audiometry, which assesses both air and bone conduction, it is possible to identify the type of hearing impairment. Conductive hearing loss is usually caused by diseases of the external auditory canal and/or middle ear. Sensorineural hearing loss occurs due to damage to the sensory cells and/or nerve fibers of the inner ear[3]. Mixed hearing loss is the result of both sensorineural and conductive hearing loss[4]. Hearing loss can be unilateral or bilateral, sudden or chronic, and can range in severity from mild to profound. Hearing impairment is common, particularly among patients with aural disease and the elderly.

The majority of hearing losses in clinical populations are sensorineural and mixed[5]. While the sensorineural components are rarely curable, accurate diagnosis is a major impediment in successful treatment. Audiologists, who are necessary for proper execution and interpretation of tests, are in short supply globally. Among low-income countries, in particular, approximately 93% have fewer than one audiologist per million. Even in nations with relatively high numbers of practitioners in the field of ear and hearing care, inequitable distribution and other factors can limit access to these specialists[5]. Artificial intelligence (AI) has the potential to address this issue, given the disparity between the supply and demand for hearing specialists. AI implements algorithms that allow computers to recognize specific data analysis patterns and draw conclusions. In the healthcare industry, this software analyses human cognition to establish links between various types of treatments and the subsequent medical outcomes. One of the most common uses of machine learning in medicine is the analysis of images such as computed tomography (CT) and magnetic resonance imaging (MRI) to detect various types of irregularities, including tumours, ulcers, fractures, internal bleeding etc., in order to provide crucial data for health care specialists and their patients. As a result, AI assists radiologists in automating daily administrative duties, improves diagnostic accuracy, eliminates human error risks, and allows researchers to concentrate on complex cases[6,7]. This is also true in tonal audiometry, where AI has been applied to the determination of edge frequency of a high-frequency dead zone in the cochlea as well as to assistance in fine-tuning hearing aids to the client's preferences more precisely and efficiently[8].

In the above context, this paper proposes a neural network model for classification of hearing loss types for discrete tonal audiometry data series. The primary objective was to obtain classification accuracy sufficient for clinical application of the developed network, allowing general practitioners to classify tonal audiometry results autonomously for further patient referral. This could result in lessening the burden on audiology specialists while still ensuring that the final decision on diagnosis is made by a physician. For audiologists, the system might eliminate simple cases, allowing them to concentrate on the more complex ones, as well as enhance diagnostic precision and prevent human error in daily practice. Furthermore, the aim is to exceed the current state-of-the-art in classification of raw audiometry data, which currently achieves an accuracy rate of 95.5% through the application of Decision Trees[9].

The paper is structured as follows: the second section describes the materials and techniques used to train and evaluate the classification model. Specifically, types of hearing loss are described in section "Hearing loss types", a literature review can be found in section "Automatic classification of audiometry data", the dataset is described in section "The tonal audiometry dataset", ethics declarations are included in section "Ethics declarations" and the study methodology is explained in section "Methodology". The third section provides detailed results. The fourth section discusses the obtained results and their comparison to the current state-of-the-art. Finally, the fifth section presents the conclusions.

Materials and methods

Hearing loss types

According to the WHO, hearing loss may be classified as conductive, sensorineural or mixed[2]. In conductive hearing loss, lesions develop in the conductive component (outer and middle ear). This type of hearing loss is characterised by good telephone speech comprehension, better hearing and speech discrimination in noise than in silence, improved speech comprehension after amplification, and preserved voice control. Conductive hearing loss may be diagnosed on the basis of audiometric tests and otoscopy. In tonal audiometry, it is characterised by a normal bone conduction, a lowering in the air conduction curve, and the presence of an air–bone gap, i.e. a 15–40 dB difference between the bone and air curves. Congenital deformities of the external and middle ear, otitis externa and otitis media, otosclerosis, injury of the external auditory canal, wax plug, obstruction of the auditory tube, and tumours of the temporal bone and nasopharynx are the most common causes of conductive hearing loss[10].

Sensorineural hearing loss is a hearing impairment induced by inner ear and auditory nerve disorders. It can be cochlear—caused by damage to the organ of Corti—or extracochlear—affected by damage to the auditory nerve up to the cochlear nuclei. The characteristics of sensorineural hearing loss are: better hearing through the air, impaired understanding of speech in noise, better hearing of low sounds, unpleasant perception of high sounds, different perception of sounds in both ears. On audiometric examination, bone and air conduction curves are at the same level, and there is no air–bone gap. The most common causes of sensorineural hearing loss with cochlear localization are: hearing loss caused by ageing, acute and chronic acoustic trauma, congenital defects, skull base fractures, pressure barotrauma, ototoxic drugs, chemotherapy, labyrinthitis, vascular disorders of the inner ear, Ménière's disease, cochlear otosclerosis, radiotherapy and metabolic disorders. Causes of sensorineural hearing loss with extracochlear and central location include presbycusis, multiple sclerosis, cranial trauma and fractures, meningitis, cerebello-pontine angle tumours, brain tumours and cerebrovascular diseases[11].

Mixed hearing loss is a combination of sensorineural and conductive hearing loss in a single ear. It may be the result of a single disease, such as otosclerosis or suppurative otitis media, or of the superimposition in one ear of two or more of the diseases listed above. It is characterised by a decreasing of the auditory threshold for bone conduction and air conduction with the existence of air–bone gap, impaired speech comprehension dependent

on the sensory-nervous component, and an audiometric curves demonstrating less decreasing in the low tones and more greater decreasing in the high tones.

Hearing impairment, particularly sensorineural hearing loss, is prevalent among the elderly and tends to aggravate with age. Conductive hearing loss is typical of adolescents and adults of working age; if it worsens, it does so very gradually, as in otosclerosis.

The described forms of hearing loss are treated differently. In the conductive type, surgical treatment predominates: paracentesis, ventilation tube placement, myringoplasty, and tympanoplasty. The majority of cases of sensorineural hearing loss are treated conservatively, as they result from sudden deafness, acute acoustic trauma and multiple sclerosis. In those instances, rehabilitation with the use of a hearing aid often proves effective. In specific cases, hearing rehabilitation must be combined with surgical treatment, as with cochlear implants. Mixed hearing loss is treated based on its aetiology. Stapedotomy is the surgical treatment for otosclerosis, whereas in non-surgical cases hearing devices are fitted[8,12].

Data selection criteria

The inclusion and exclusion criteria for tonal audiometry data in the dataset were determined according to rules given by Margolis and Saly[13]. During the initial stage of the classification process, every audiometry test result was evaluated to determine if it met the minimum standards for inclusion in this study. Initially, a thorough examination was performed to verify the existence of six octave frequencies, namely 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz and 8000 Hz in air conduction. In the context of bone condition, the auditory thresholds for four specific octave frequencies (500 Hz, 1000 Hz, 2000 Hz, 4000 Hz) were examined to determine their presence. If any of these values were absent, the data was rejected. Furthermore, any audiometry test result that satisfied any of the following criteria was also eliminated:

- Air conduction threshold is outside the range of − 10 to 110 dB HL (exceeding 250 Hz, when the limit is 90 dB HL);
- Air conduction threshold is beyond the 0–30 dB range of the next lower frequency.
- Bone conduction threshold is beyond the range of − 10 to 60 dB HL (exceeding 250 Hz, when the limit is 40 dB HL);
- The bone-conduction threshold should fall within the range of 50 to 10 dB relative to the air-conduction threshold at that frequency.

Classification of hearing loss types

Based on the latest (2021) WHO standards[2], normal hearing is defined as an average value for the air and bone conduction curve evaluated at 4 octave frequencies (0.5 kHz, 1 kHz, 2 kHz, 4 kHz) that is below 20 dB HL. These guidelines are also in line with the 1996 International Bureau for Audiophonology criteria[14]. The air-bone reserve, sometimes referred to as the air–bone gap, was calculated by subtracting the individual values of frequencies in the air-bone conduction threshold for in 500 Hz, 1000 Hz, 2000 Hz and 4000 Hz. The present air-bone reserve was determined to be 10 dB HL at three or more frequencies within the range of 500–4000 Hz or 15 dB HL for single frequency within this range[13]. The diagnosis of conductive hearing loss was made based on the existence of hearing loss in the air conduction curve, normal values in the bone conduction curve and the presence of air-bone reserve. The identification of sensorineural hearing loss was determined based on the observation of hearing loss in both the air and bone conduction curve as well as the absence of air-bone reserve. The medical diagnosis of mixed hearing loss has been established by noticing hearing loss in both the air and bone conduction curves, in addition to the presence of air-bone reserve. Table 1 provides a comprehensive list of the specific criteria that were utilized to classify the audiograms.

Automatic classification of audiometry data

In medical practice, the type of hearing impairment is determined from pure-tone audiometry test results according to their configuration, severity, location of lesion (hearing loss type) and symmetry[13,15]. The process is performed by audiology specialists on the graphical representation of an audiometry test result, known as an

	Air condition	Bone condition	Air–bone gap
Normal hearing	average value for frequencies of 500–4000 Hz < 20 dB	average value for frequencies of 500–4000 Hz < 20 dB	absence of an air–bone gap
Conductive hearing loss	average value for frequencies of 500–4000 Hz ≥ 20 dB	average value for frequencies of 500–4000 Hz < 20 dB	10 dB across minimum three frequencies (500–4000 Hz) or 15 dB across one frequency (500–4000 Hz)
Sensorineural hearing loss	average value for frequencies of 500–4000 Hz ≥ 20 dB and average value for frequencies of 4000–8000 Hz ≥ 20 dB	average value for frequencies of 500–4000 Hz ≥ 20 dB and average value for frequencies of 4000–6000 Hz ≥ 20 dB	absence of an air–bone gap
Mixed hearing loss	average value for frequencies of 500–4000 Hz ≥ 20 dB	average value for frequencies of 500–4000 Hz ≥ 20 dB	10 dB across minimum three frequencies (500–4000 Hz) or 15 dB across one frequency (500–4000 Hz)

Table 1. Classification criteria of hearing loss types.

audiogram. The site of lesion is determined by air and bone conduction thresholds of the audiogram, whereas the configuration is determined by shape. The severity is determined by the degree of hearing loss.

The subject of automatic audiometry data classification has been under investigation for a long time. Over the last decade, there have been a number of attempts at devising an automated method of classification that would be accurate enough to warrant practical application.

The first attempt in this regard was made by Cheng-Yung Lee et al.[16], who proposed a statistical classification system of audiogram shapes in an effort to enhance and integrate shape recognition across clinical settings. Based on 1633 audiograms, eleven audiometric shapes were classified using K-means cluster analysis. The authors anticipated that, in the future, the classification of audiogram shapes would result in a more efficient infrastructure for diagnosing hearing loss.

Further work may be divided into two thematic groups: classification of audiogram shapes for the purpose of determining the initial configurations of hearing aids[17,18] and diagnosing the type of hearing loss.

Chelzy Belitz et al.[19] combined unsupervised and supervised machine learning techniques for mapping audiograms to a limited number of hearing aid configurations. When mapping a single configuration to each audiogram, the best results were achieved with the Multi-layer Perceptron model at 64.19% accuracy. When mapping two configurations to each audiogram, the chance that at least one is correct increased to 92.70%.

Charih et al.[20] presented their Data-Driven Annotation Engine, a decision tree based audiogram multi-label classifier which considers the configuration, severity and symmetry of participant's hearing losses and compared it to AMCLASS[13], which fulfils the same purpose using a set of general rules. Dataset used in this study contained 270 distinct audiograms with seven tested frequencies at 500 Hz, 1,000 Hz, 2,000 Hz, 3,000 Hz, 4,000 Hz, 6,000 Hz and 8,000 Hz. However, bone conduction information is not included in the data set. Three licensed audiologists rated the method's accuracy at approximately 90 percent.

Abeer Elkhouly et al.[21] proposed a machine learning solution to classify audiograms for the purpose of configuring hearing aids based on their shapes using unsupervised spectral clustering, normalization, and multi-stage feature selection on a dataset of 28 244 audiograms. The authors normalized the data using 20 different normalization methods to increase the training data size in building a credible model, and then selected 10 normalized data sets to train the model. Firstly, the data was divided into 10 clusters, then classified using fine K nearest neighbour classifier with 95.4% accuracy.

In comparison to the subject of automated configuration of hearing aids, the problem of hearing loss type classification has been given considerably less attention.

In this regard, Elbaşı and Obali[9] presented a comparison of several approaches to hearing loss determination, including Decision Tree C4.5 (DT-J48), Naive Bayes and Neural Network Multilayer Perceptron (NN) model. The study was conducted on a dataset containing 200 samples divided into four categories, including normal hearing, conductive hearing loss, sensorineural hearing loss, and mixed hearing loss. Input data was formatted as a series of numeric values representing Decibels corresponding to constant frequency levels (750 Hz, 1 kHz, 1.5 kHz, 2 kHz, 3 kHz, 4 kHz, 6 kHz, 8 kHz). Classification algorithms have been implemented using Weka software, resulting in 95.5% accuracy in Decision Tree, 86.5% accuracy in Naive Bayes, and 93.5% accuracy in NN.

More recently, Crowson et al.[22] adopted the ResNet models to classify audiogram images into three types of hearing loss (sensorineural, conductive or mixed) as well as normal hearing using a set of training and testing images consisting of 1007 audiograms. The model was fed by 500 × 500 pixel images of static audiogram plots that had been pre-transformed. Instead of fully training the classifier, the authors applied transfer learning to well-established raster classification models. All tested architectures were based on convolutional neural network (CNN) architectures, but the ResNet-101 model achieved the highest classification accuracy at 97.5%.

In conclusion, the integration of neural networks with enhanced computational capabilities and more extensive training datasets should enable more comprehensive evaluations[23]. Despite this, the classification accuracy of the majority of the currently proposed solutions ranges between 90 and 95%, which, while very high, still leaves substantial room for error. According to clinical standards, the margin of error should be kept under 5%[24] and ideally should be close to 3%[25]. Only one of the discussed classifiers satisfies these requirements. Crowson et al.[22] presented the finest audiogram classifier to date, using transfer learning to adapt an established image classifier network to the analysis of audiogram images. Despite producing a classification accuracy of 97%, this method has significant limitations. Due to the fact that it is an image classifier, it cannot be applied to the original data series generated by tonal audiometry. This necessitates converting the data series into audiogram images, which may result in data loss. Moreover, although the structure of audiograms is generally similar, audiograms generated by distinct hardware and software configurations can still differ significantly. In addition to differences in background and line colours, audiograms can also differ in the amount of information conveyed (e.g. they may contain data for a single ear or both). Consequently, a universal classification solution for tonal audiometry results cannot rely on an image classifier[26].

In addition, each of the cited studies on determining hearing loss type was conducted with a relatively small data set, ranging from 200 test results in Elbaşı and Obali[9] to 1007 in Crowson et al.[22], which may have led to an optimistic and unreliable estimation of model performance. Moreover, the limited size of the training dataset poses a challenge in discerning relationship patterns within certain classes, potentially leading to a validation outcome that has bias when applied to the test dataset.

In the above context, a summary of the current state-of-the-art is presented in Table 2.

As can be seen in Table 2, thus far the issue of hearing loss type classification has been researched on relatively small data samples with undisclosed class ratios, and the best achieved classification accuracy has been produced for a raster dataset, resulting in limited application. Consequently, the presented work focused on the development of a classifier trained on a considerably larger and more representative data set. Moreover, the developed classifier has been designed for use with raw audiometry data, ensuring greater flexibility of application.

Paper	Audiogram classification problem	Data size	Data type	Accuracy (%)
Cheng-Yung Lee et al.[16]	Configuration—11 shapes	1633	Raster/raw data	–
Chelzy Belitz et al.[19]	Configuration—4 shapes	90 000	Raster/raw data	64
Charih etc.[20]	Configuration—8 shapes, severity, and symmetry	320	Raster data	90
Abeer Elkhouly et al.[21]	Configuration—10 shapes	28 244	Raster/raw data	95
Ersin Elbaşı and Murat Obali[9]	Hearing loss types: normal, conductive, mixed and sensorineural	200	Raw data	95.5
Crowson etc.[22]		1007	Raster data	97.5

Table 2. Overview of state-of-the-art methods and results.

The tonal audiometry dataset

The dataset includes 15,046 audiometry test results from 9663 adult patients tested between 2010 and 2022 in the Otolaryngology Clinic of the University Clinical Centre in Gdansk, Poland. Tonal audiometry tests were conducted in soundproof booths (ISO 8253, ISO 8253). Signals were generated by calibrated Itera II and Midimate 622 clinical audiometers, manufactured by Madsen Electronics (Otometrics, Copenhagen, Denmark) (PN-EN 60645-1, ISO 389, ISO 8789, ISO 7566, ISO 8798). The equipment had the ability to correct for ANSI S 3.6-1989 and 2004 standard hearing levels. The American Speech-Language-Hearing Association (ASHA) guidelines were used in the evaluation of participants' hearing by tonal audiometry[27]. Using air conduction tests, the signal generated by the audiometer was connected to TDH-39P headphones. For bone conduction tests, the audiometer was coupled to a B-71 bone vibrator (New Eagle, PA). Of the examined patients, aged between 18 and 98, 5591 were female (57.86%) and 4072 were male (42.14%). The patients' age distribution by sex has been illustrated in Fig. 1. A maximum of two test results were obtained from each patient, one for the left ear and one for the right, resulting in no replication of data from the same patient and ensuring good data variety.

Three experienced audiologists labelled the morphologies of hearing loss on the audiometry test results, dividing the set into four classes: normal hearing, conductive hearing loss, mixed hearing loss, and sensorineural hearing loss according to methodology presented in Table 1. The evaluation of every test results was conducted by three audiologists. In cases where the classification result was not unanimous, the final decision was made by majority vote in which the highest weight was given to the opinion of the senior audiologist (T.P.). Table 3 shows the quantity of samples for each class in the produced dataset.

The results of pure-tone audiometry are commonly presented in the form of an audiogram, which is a graphical representation of how loud sounds must be at various frequencies for them to be audible. In addition to a graphical representation, audiology software generates XML files containing all information regarding tonal points that appear in the audiogram. The presented research uses XML files to analyse raw audiometry data.

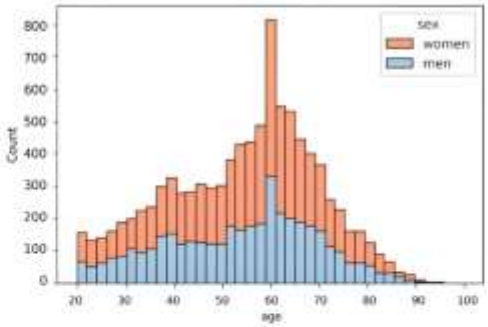


Figure 1. The distribution of patients' age and sex in the dataset.

Hearing type	Number of samples (%)
Normal	2584 (17.17%)
Conductive hearing loss	657 (4.37%)
Mixed hearing loss	4028 (26.71%)
Sensorineural hearing loss	7777 (51.69%)

Table 3. The four hearing types contained in the dataset and the number of samples in each group.

A sample audiogram, depicting masked right bone conduction ([) and non-masked right airconduction—(O) tresholds, with the corresponding XML file fragment containing the coordinates of consecutive tonal points is presented in Fig. 2.

The input data for a single measurement (one ear of one patient) consists of seven lists corresponding to air and bone conduction with hearing levels measured in decibels at frequencies of 125 Hz, 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz and 8000 Hz, respectively. It should be noted that two extreme frequencies (125 Hz and 8000 Hz) are not registered during bone conduction testing, which is why their values are set to null by default.

Ethics declarations

This study was approved by the Committee on Research Ethics at Medical University of Gdańsk (IRB KB/23/2024). Prior to participating in the hearing test and this study, the subjects provided informed consent. All methods were carried out according to the relevant guidelines and regulations.

Methodology

Data imbalance correction

An unbalanced training dataset, or in other words a dataset which does not evenly represent all possible classes, can significantly hinder the performance of machine learning models[28]. To prevent unintended outcomes from occurring when processing unbalanced data, a well-known system of class weight was employed[29]. This system permits the training procedure to account for the uneven distribution of classes by assigning different weights to the majority and minority classes. The objective is to penalise the model for the misclassification of the minority class by increasing its class weight and decreasing the class weight of the majority[30]. In the presented research, appropriate weight parameters were calculated and applied for each class during the training process.

Data normalisation

The process of data normalisation can aid in stabilising the gradient magnitude during training, particularly in the recurrent neural networks used in this study[31]. Experiments using several normalisation methods, such as linear normalisation, Robust Scaler and Max Abs Scaler, led to the selection of Z-score normalisation as the most effective[32]. Z-score normalisation refers to the process of normalising each value in a dataset so that the mean of all the values is 0 and the standard deviation is 1.

Network architecture

During a previous study[28], several neural network architectures were evaluated in order to construct a binary classifier for normal and pathological hearing loss. The tested architectures included Multilayer Perceptron (MLP), Convolutional (CNN) and Recurrent (RNN) neural networks. A multi-stage investigation revealed that the RNN architecture performs best with this type of medical data. Over the course of that study, Recurrent neural networks (RNN), Gated Recurrent Units (GRU), and Long Short-Term Memory (LSTM) have been tested on a subset of the presented dataset. The final accuracy performance of RNN, GRU and LSTM was revealed to be 96.46%, 97.71%, and 98.12%, respectively. In addition, the LSTM model achieved an exemplary False Negative

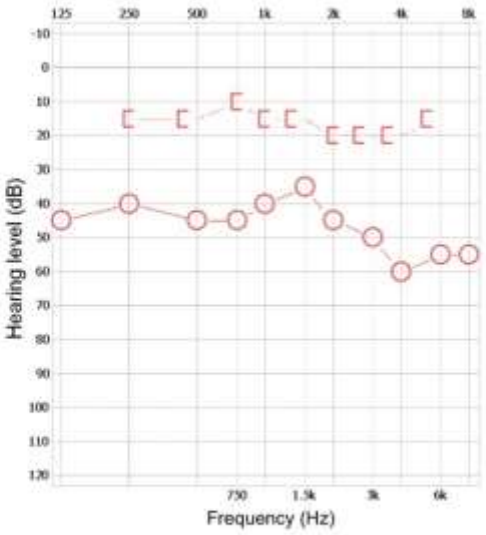


Figure 2. Two methods of representing tonal audiometry results: audiogram (left) and XML (right).

rate of 0.2%, which enabled its clinical application. Furthermore, these results have been corroborated by another study[33], which investigated different neural network architectures for categorizing three forms of hearing loss, with an RNN-based model demonstrating the best performance out of the 11 models evaluated. These results are in line with the general notion that RNN architectures perform well with sequential or time-series medical data[34–36], and LSTM in particular is well-known for successfully resolving problems with vanishing/exploding gradients[37].

Due to the success of RNN-LSTM networks in the above-mentioned classifications, they have been selected for use in the presented study. The initial proposed multi-class solution involves the processing of input data from a single ear of a patient's audiometry test in the first LSTM layer. In the input data, time steps correspond to the tested frequencies, while air and bone conduction represent features in each time step. Afterwards, the number of nodes is reduced by a dropout layer, which helps prevent overfitting. The next steps consist of a similar sequence of LSTM and dropout layers. The model is completed by a dense layer with softmax activation function. After additional investigation and optimization, the initial model was modified by replacing the LSTM with a Bidirectional LSTM (Bi-LSTM)[38] in the first layer. The Bi-LSTM is a variant of Bi-RNN that utilizes two basic LSTMs to analyse input time series in both forward and backward orientations. Utilizing data from both ways allows the model to detect patterns that could be overlooked when only using unidirectional LSTM. Thus, when considering pure tone audiometry data series, it can significantly improve classification accuracy. An overview of the proposed architecture is shown in Fig. 3.

Model evaluation

Due to the aforementioned class imbalance, the model has been trained with the use of stratified K-fold cross validation (SKCV)[39]. K-fold Cross-Validation is the process of dividing a dataset into K folds and evaluating the model's performance with new data. K represents the number of categories into which the data sample is divided. For instance, if the k-value is 10, it can be referred to as a tenfold cross-validation. At one point in the procedure, each fold serves as a test sample. SKCV is an extension of regular K-fold cross validation, which has been designed specifically for classification problems in which the ratio between the target classes is the same in each fold as it is in the entire dataset. In other words, the datasets are not distributed randomly into k-folds, but instead in a way that does not impact the sample distribution ratio across classes. Using stratified sampling

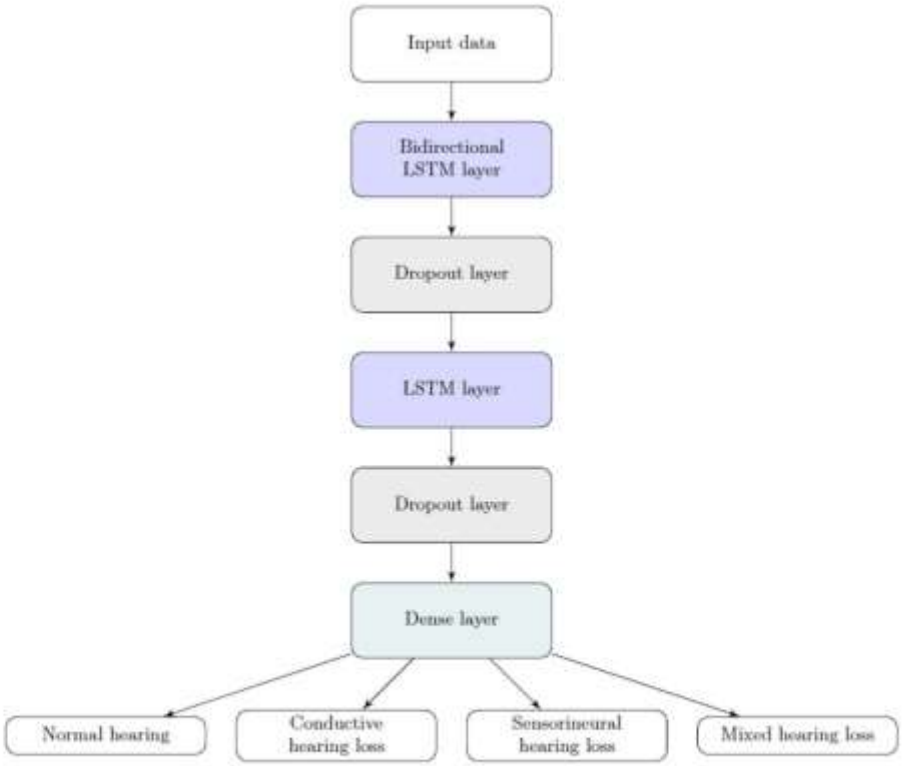


Figure 3. An overview of the proposed neural network architecture.

rather than random sampling ensures that relative class frequencies are effectively maintained across each train and test fold. The behaviour of SKCV is represented graphically in Fig. 4.

Thus, the general workflow of the presented research is depicted in Fig. 5.

Metrics and statistical test

In the classification context, the performance of a classifier is typically evaluated by computing functions on the resulting confusion matrix. In essence, the confusion matrix represents the proportion of class samples that have been misclassified as other classes. For every class, there are four types of distinguishable parameters:

- True positive (TP), when positive predicted was true;
- True negative (TN), when negative predicted was true;
- False positive (FP), when positive predicted was false;
- False negative (FN), when negative predicted was false.

On the basis of these parameters, classification accuracy (1) is a common classification metric that computes the proportion of correctly classified test data relative to the total number of test data. In addition, precision (2) quantifies the proportion of positive class predictions that correspond to the positive class. In contrast, recall (3) computes the number of positive class predictions from all positive examples. Finally, F_1 (4) provides a single score that addresses both precision and recall concerns in a single number.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

The probability that a classifier provides more weight to the correct class than to the incorrect class is graphically presented in the form of Area Under the Curve (AUC). It is the area under a Receiver Operating characteristic Curve (ROC) that compares the true-positive rate to the false-positive rate by varying the decision threshold of the classifier.

In order to statistically compare performance of models the McNemar's Test[93] was used. The primary purpose of this test is to examine the disparities between two classifiers, specifically in relation to the instances where they made divergent predictions. The initial step entails performing calculations to determine the subsequent values:

- n00: number of items misclassified by both A and B;
- n01: number of items misclassified by A but not by B;

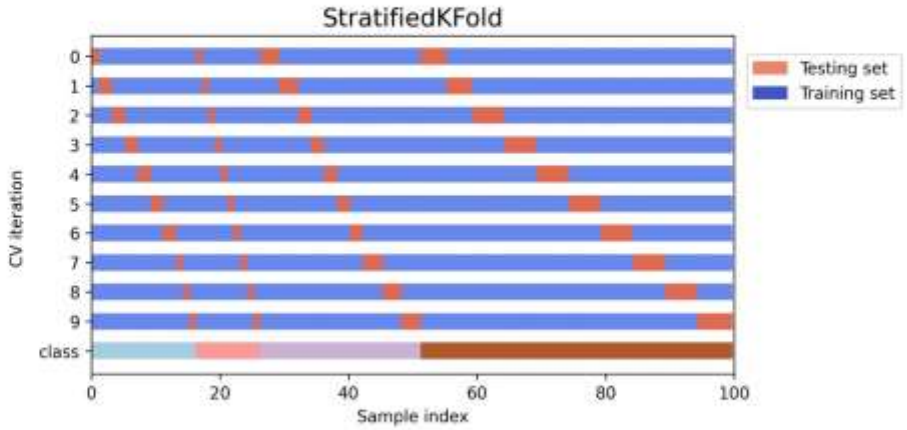


Figure 4. Schematic representation of a Stratified K-fold cross-validation, which uses proportional subsets of each class in every CV iteration.

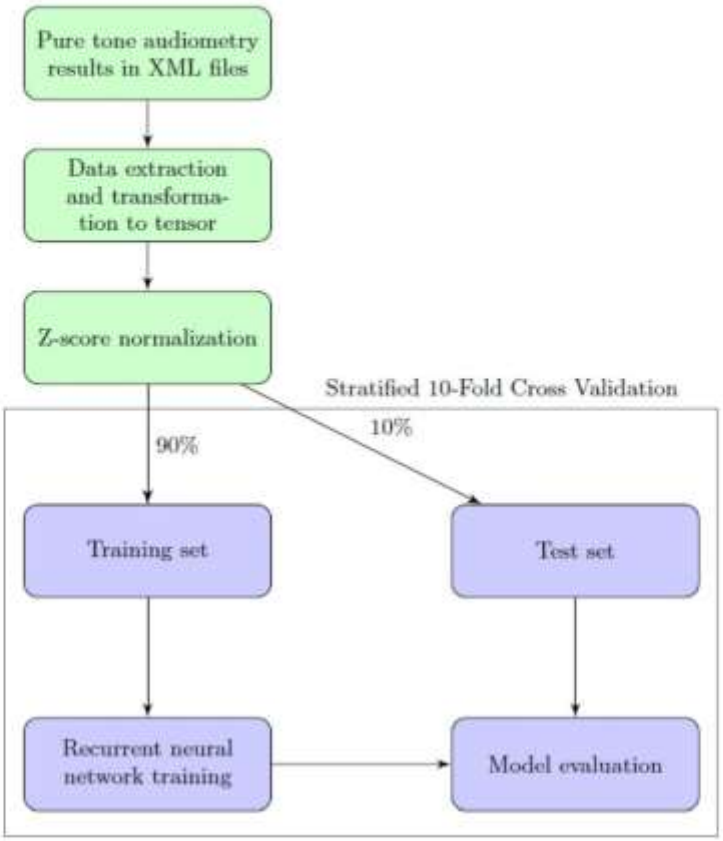


Figure 5. General workflow of the presented research.

- n10: number of items misclassified by B but not by A;
- n11: number of items classified correctly by both A and B.

The null hypothesis states that the error rates of A and B, denoted as n01 and n10 respectively, are equal.

Results

Initial classification results

The applied stratified tenfold cross-validation of the proposed initial LSTM model yielded the following results: the average classification accuracy is 98.29% (+/− 0.46%), the average precision is 98.30% (+/− 0.47%), the average recall is 98.29% (+/− 0.46%), and the average F_1 score is 98.27% (+/− 0.47%). Table 4 presents the detailed information of each phase of the SKCV procedure.

Metrics	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Accuracy (%)	97.67	98.34	97.87	98.87	97.94	98.14	98.34	99.14	97.87	98.74
Precision (%)	97.64	98.34	97.86	98.89	98.00	98.13	98.42	99.14	97.88	98.74
Recall (%)	97.67	98.34	97.87	98.87	97.94	98.14	98.34	99.14	97.87	98.74
F_1 score (%)	97.63	98.33	97.85	98.86	97.87	98.10	98.35	99.13	97.87	98.74

Table 4. Stratified tenfold cross validation score of proposed model.

The confusion matrix of the LSTM model is presented in Fig. 6. Normal hearing, mixed hearing loss, conductive hearing loss, and sensorineural hearing loss are represented by the N, M, C, and S indices, respectively.

Final classification results

The applied stratified tenfold cross-validation of the proposed Bi-LSTM model yielded the following results: the average classification accuracy is 99.33% (+/− 0.23%), the average precision is 99.32% (+/− 0.33%), the average recall is 98.85% (+/− 0.45%), and the average F_1 score is 99.08% (+/− 0.29%). Table 5 presents the detailed information of each phase of the SKCV procedure.

Table 6 displays the precision, recall, and F_1 score for each class of the proposed Bi-LSTM model.

The confusion matrix of the Bi-LSTM model is presented in Fig. 7.

Comparison of different normalization methods

The issue of choosing an optimal neural network architecture as well as data normalization technique for the presented problem has been investigated in detail in[41]. The accuracy of both initial LSTM and proposed Bi-LSTM models using different normalization techniques is presented in Table 7.

Comparison with current state-of-the-art

The current state-of-the-art in raw audiometry data classification uses the C4.5 algorithm[42] as a Decision Tree Classifier[8]. In order to facilitate a comparison, the C4.5 model was applied and evaluated on the presented dataset. Application of stratified tenfold cross-validation resulted in the following outcomes: the mean classification accuracy is 95.64% (+/− 0.69%), precision 95.69% (+/− 0.74%), recall 95.66% (+/− 0.75%), F_1 score 95.63% (+/− 0.77%). Detailed results are provided below in Table 8.

The confusion matrix of the C4.5 model is presented in Fig. 8.

In order to statistically compare performance of the proposed BI-LSTM model with the C4.5 classifier we used McNemar's Test. The significance level was determined at the value of $p = 0.05$. The null hypothesis can be

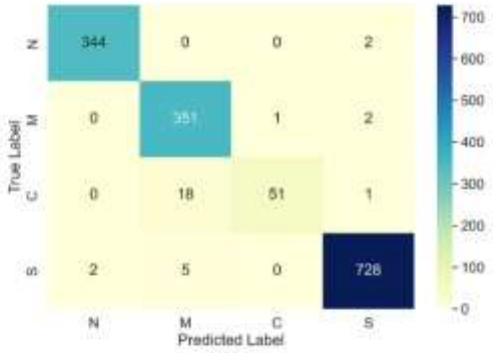


Figure 6. Confusion matrix of initial model.

Metrics	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Accuracy (%)	99.27	99.33	99.27	99.14	99.00	99.34	99.40	99.60	99.73	99.00
Precision (%)	99.22	99.43	99.52	99.10	98.84	99.58	99.66	99.69	99.47	98.67
Recall (%)	98.35	99.50	98.09	98.42	99.30	98.92	98.79	98.90	99.47	98.76
F_1 score (%)	98.76	99.46	98.78	98.75	99.07	99.25	99.22	99.29	99.47	98.71

Table 5. Stratified tenfold cross validation score of proposed model.

	Normal	Mixed hearing loss	Conductive Hearing Loss	Sensorineural hearing loss
Precision (%)	100.00	99.72	100.00	99.19
Recall (%)	99.42	99.15	98.57	99.86
F_1 score (%)	99.71	99.43	99.28	99.53

Table 6. Comparison of each class's precision, recall, and F_1 score.

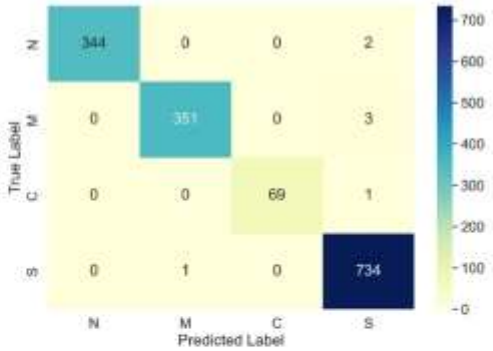


Figure 7. Confusion matrix of initial model.

	LSTM	Bi-LSTM
Linear normalization/Max–Min	97.21% (+/− 1.17%)	98.84% (+/−0.37%)
Robust scaler	97.61% (+/− 0.75%)	99.04% (+/−0.23%)
Max abs scaler	97.87% (+/− 1.34%)	99.27% (+/−0.22%)
Z-score normalization	98.29% (+/− 0.46%)	99.33% (+/−0.23%)

Table 7. An analysis of the accuracy of LSTM and Bi-LSTM models using various normalizing approaches.

Hearing type	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
Accuracy (%)	96.21	95.67	94.15	96.34	96.15	95.95	95.28	95.41	96.14	94.88
Precision (%)	96.20	95.70	94.17	96.81	96.13	95.93	95.43	95.49	96.14	94.93
Recall (%)	96.21	95.68	94.15	96.76	96.15	95.95	95.28	95.41	96.14	94.88
F_1 score (%)	96.17	95.67	94.07	96.76	96.12	95.92	95.27	95.39	96.12	94.83

Table 8. Stratified tenfold cross validation score of C4.5 model.

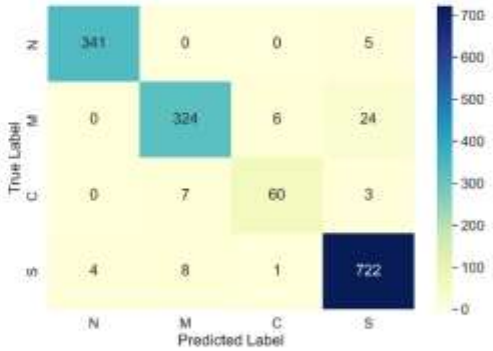


Figure 8. Confusion matrix of C4.5 model. The N, M, C, and S indices represent normal hearing, mixed hearing loss, conductive hearing loss, and sensorineural hearing loss, respectively.

rejected in a two-tailed test if the calculated chi-square value (X^2) exceeds the critical chi-square value ($X^2(0.05)$) at a significance level of 0.05. The chi-square statistic yielded a value of 127.21, while the p-value was determined to be 0.0000000001. The result provides evidence to reject the null hypothesis, indicating significant statistical difference between the two models.

The performance of solving a classification problem at different threshold settings is usually represented by the area under the receiver operating characteristic curve (AUC-ROC). The AUC-ROC is typically applicable to binary classification issues, however, the one-vs-all technique enables it to be extended to multiclass classification problems. The One-vs-the-Rest (OvR) multiclass strategy, also known as one-versus-all, involves computing a ROC curve for each class. In each stage, a specific class is viewed as the positive class, while the remaining classes are viewed as the negative class in the majority. The micro-averaged ROC curves with AUC parameters of models is shown in Fig. 9.

Discussion

The initial LSTM model satisfactorily classified normal hearing (N), sensorineural (S), conductive (C), and mixed hearing loss (M) in terms of average accuracy of 98.29% (+/− 0.46%) based on stratified tenfold cross validation. The average accuracy is superior than the existing state-of-the-art. However, the results displayed in Fig. 6 render the model unsuitable for clinical applications. The main concern is the presence of false negatives in individuals with normal hearing, which may lead to the patient being at risk of not obtaining appropriate medical treatment. In other words, the classification precision in the case of normal hearing ought to be 100%. That said, this requirement is not met as there are four instances where sensorineural hearing loss was incorrectly classified as normal hearing. Therefore, more efforts have been made to enhance the model. Several different architectures were investigated, but only one variant of the traditional LSTM, the Bidirectional LSTM, yielded improved results. This observation is supported by a number of published studies in which authors demonstrate that bidirectional LSTM models outperform conventional LSTM models. This insight is apparent in research on natural language processing[43,44], but it applies to other fields as well[45–48]. An example of an application outside of natural language processing is an article that concentrates on forecasting the spread of the COVID-19 pandemic using a Bi-LSTM model on time series data[49]. Out of the fifteen models tested, Bi-LSTM achieved the highest performance, exceeding that of LSTM and other RNN variants. This potential to further improve LSTM classification results motivated us to evaluate the performance of Bi-LSTM on our dataset.

Based on stratified tenfold cross validation, the proposed Bi-LSTM model successfully classified normal hearing (N), sensorineural hearing (S), conductive hearing (C), and mixed hearing loss (M) with an average accuracy of 99.33%. Beginning at 99.00 and ending at 99.73, the accuracy remained stable, with standard deviation equal to 0.23%. Precision, recall, and F_1 score share similar characteristics with regard to accuracy. Table 6 revealed a diversity of outcomes, with performance parameters broken down by individual classes. The classification performance of normal hearing, mixed hearing loss and sensorineural hearing loss has shown to be notably high. In contrast, cases of conductive hearing loss have been classified with lower accuracy, particularly in terms of recall. The precise definition of recall for conductive hearing loss is the ratio of correctly predicted cases of conductive hearing loss (69) to actual number of conductive hearing loss cases ($69 + 1 = 70$), which is approximately 98.57%. There are a few explanations for this behaviour. To begin with, only 4.37% of the dataset represents conductive hearing loss (Table 2). This implies that any misclassification will have a greater effect on statistical calculations that take true positive examples into consideration. Furthermore, regardless of the weighting method used in

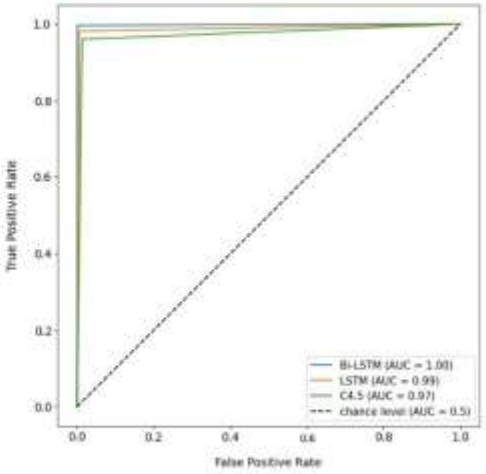


Figure 9. A ROC curve with the AUC parameter for each model.

the training process, NN has a lower chance of learning patterns from smaller amounts of data, as demonstrated by[49]. The lower number of conductive hearing loss samples in the used dataset is caused by the rules of patient treatment employed by medical institutes. In particular, conductive hearing loss is usually caused by pathology along the root of the ear, which essentially means that an object is blocking the air canal. This type of problem is typically diagnosed with an otoscope during the initial patient examination, therefore nullifying the need for conducting a pure-tone audiometry test. Currently there is no practical way to alleviate this problem, as performing audiometry tests on patients who can be diagnosed by simpler methods is not financially viable.

The results presented in Table 7 demonstrate that the selection of the optimal normalizing procedure considerably affects the final accuracy value. Irrespective of the model used, Z-score normalization proved to be the most effective strategy for scaling features, closely followed by Max Abs Scaler. The application of linear normalization (Max–Min) resulted in a decrease in the degree of accuracy, particularly the loss of over one percentage point in the LSTM model and around half a percentage point in the Bi-LSTM model.

The proposed solution significantly outperformed the current state-of-the-art in raw audiometry data classification, held by the C4.5 Decision Tree (DT-J48) method proposed by Elbaşı and Obali[9]. Application of the C4.5 classifier to the presented dataset demonstrated a level of accuracy which is in line with the one reported by the original publication, with a mean value of 95.64% and a standard deviation of 0.69%. This effectively makes the state-of-the-art model around 4 percentage points inferior compared to the proposed Bi-LSTM model. The dissimilarity is also evident in the confusion matrices of both models. Specifically, there is a notable distinction in the misclassification of mixed hearing loss (M) as sensorineural hearing loss (S). The C4.5 model exhibits 24 instances of this misclassification, whereas the Bi-LSTM model demonstrates only 3 instances. Furthermore, there are four cases in which hearing loss was misclassified as normal hearing in the C4.5 model. It is worth noting that such misclassifications were not observed in the Bi-LSTM model. The distinction between C4.5 and BI-LSTM is also evident in the ROC-AUC curves shown in Fig. 9. The C4.5 model with an AUC of 0.97 performed less effectively compared to the Bi-LSTM model with an AUC of 1.0. Finally, the findings obtained by McNemar's Test indicate a statistically significant difference in the classification outcomes between the C4.5 model and the Bi-LSTM model.

In general terms of the audiogram classification problem, the overall accuracy of the presented model (99.33%) exceeds the most performant of the existing approaches to hearing loss classification, presented by Crowson et al. for raster data[22], which is 97.5%. When compared directly, the difference in accuracy may not seem very large, however it should be noted that it was obtained on a significantly larger dataset (1007 samples in Crowson et al.[22] versus 15,046 in this study), which ensured proper variation of training as well as validation data and guaranteed that the obtained classification results are not overly optimistic. In machine learning, the larger and more diverse the dataset, the better it is for discovering general patterns, particularly in medical applications where specific cases occur infrequently but must be evaluated correctly by NN. Classification of hearing loss in audiograms is typically based on frequencies between 0.5 and 4 kHz, but hearing loss can also be detected in the upper pitch range of 4 to 8 kHz[15]. Consequently, it is crucial to train on a sufficient number of examples to illustrate these specific audiometry challenges. It should also be noted that the work presented by Crowson et al. has been based on interpretation of raster audiogram images. As these images are the outcome of pure-tone audiometry tests, working with them is the intuitive approach. Unfortunately, the majority of audiogram images are generated by specialised software provided by different hardware vendors. While the symbols appearing in audiograms are standardised by the American Speech-Language-Hearing Association[33], there are no strict rules regarding other aspects of creating audiograms. As a result, images from different sources can have a large variety of differences, ranging from small details such as variance in colour of plots and size of measurement point indicators, to changes which can significantly impede the performance of an automated classifier, such as placing the test results from both ears on a single plot. As a result, image-trained models such as those presented by Crowson et al.[22] and Charih et al.[20] will function properly only with specific sources of audiometry data. In comparison, the classifier developed during the presented study works on raw audiometry data, allowing it to bypass vendor-specific issues with data representation.

Moreover, every system that will be used in clinical settings must meet extremely stringent requirements, in order to ensure that it does not pose undue risk to patients. In this context, it is not only important that the developed classifier achieves a high level of overall accuracy, but the types of errors it may be prone to make are also crucial. In the context of the presented work, the most dangerous scenario is when a patient with hearing loss is misclassified as having normal hearing, which can result in them not receiving proper medical care. Therefore, a secondary goal of the presented research has been to eliminate this type of error. The results of this endeavour are visible in the confusion matrix of the presented model (Fig. 7). As it can be seen in Fig. 7, there are no instances in which conductive hearing loss, mixed hearing loss, or sensorineural hearing loss are categorised as normal hearing by the developed model. While the developed model is not completely error-free, its potential practical application should not put patients at risk. This is a significant step-up from the current state-of-the-art C4.5 model proposed by Elbaşı and Obali[9], which yielded five instances of misclassification that may result in patients not obtaining appropriate medical care. Moreover, while both classifiers exhibit instances when individuals with normal hearing were erroneously identified as having sensorineural hearing loss, the proposed model exhibits a lesser number of this sort of error compared to the state-of-the-art, with 2 errors in the Bi-LSTM model as opposed to 5 errors in the C4.5 model. However, this type of error is less significant, as it would result in the patient being directed to a qualified audiologist who would rectify the mistake.

Conclusion

This paper presents a Bi-LSTM-based model for classification of raw audiometry data into normal hearing and three types of hearing loss. The developed solution advances the classification of hearing loss types beyond the current state-of-the-art in several areas. First, the achieved classification accuracy (99.33%) is superior to that presented in current state-of-the-art in raw audiometry data classification, presented by Elbaşı and Obali[7], which achieved 95.5%. The findings obtained from the comparative analysis between the C4.5 model proposed by Elbaşı and Obali and the Bi-LSTM model presented in this study using the same dataset indicate that the Bi-LSTM model exhibits significantly higher accuracy and does not produce any errors that could negatively impact patient health.

Secondly, the proposed solution also managed to outperform the current state-of-the-art in raster audiogram classification presented by Crowson et al.[22] which achieves 97.5% accuracy.

Thirdly, the presented research was conducted on 15,046 audiometry test result samples, which is nearly 15 times larger than the largest dataset produced to date in terms of hearing loss type, which consists of 1007 audiograms and was established by Crowson et al.[22]. The high variety and representativeness of the used dataset ensures that the reliability of the obtained results also constitutes an improvement to the state-of-the-art.

Finally, working with raw audiometry data allows for a more flexible implementation in clinical settings. In contrast to the approach presented e.g. by Crowson et al.[22], the presented method is not limited to working with audiogram images produced by specific sources.

This being said, there are a few limitations associated with this study. Using an unbalanced dataset with only 4.37% instances of conductive hearing loss results in a lower value of F_1 score compared to other classes. Moreover, working with raw audiometry data means that the classifier can only be used in medical facilities, as patients are generally only presented with audiogram images of their test results. In consequence, further work is needed to integrate the presented model with an accurate audiogram image parser in order to make it more broadly available to patients as well as physicians.

Overall, the presented results suggest that the developed NN-based audiometry data classifier can be applicable to clinical practice, either in the form of a classification system for general practitioners or a support system for professional audiologists. Moreover, the model can work with all hardware systems that generate text results of audiometry tests. By allowing general practitioners to classify the results of pure tone audiometry tests, the developed model may help to significantly reduce the caseload of audiology specialists. In addition, the proposed solution gives professional audiologists access to an AI decision support system that has potential to reduce their workload while also increasing diagnostic precision and decreasing human error.

Data availability

The datasets analysed during the current study are not publicly available due to the confidentiality restrictions imposed by the approved ethics of study but are available from the corresponding author on reasonable request.

Received: 29 February 2024; Accepted: 7 June 2024
Published online: 20 June 2024

References

1. Blazer, D. G. & Tucci, D. L. Hearing loss and psychiatric disorders: A review. *Psychol. Med.* **49**(6), 891–897. https://doi.org/10.1017/s0033291718003409 (2018).
2. World Health Organization. *World Report on Hearing. WHO*. Available at: https://www.who.int/publications/i/item/9789240020481 (2021)
3. Glossary. *National Institute of Deafness and Other Communication Disorders*. Available at: https://www.nidcd.nih.gov/glossary
4. Ballantyne, J. C., Graham, J. M. & Baguley, D. *Ballantyne's Deafness* (Wiley, 2009).
5. Margolis, R. H. & Saly, G. L. Distribution of hearing loss characteristics in a clinical population. *Ear Hear.* **29**, 524–532. https://doi.org/10.1097/AUD.0b013e3181731e2e (2008).
6. Jin, D. *et al.* Artificial intelligence in radiology. *Artif Intell Med* https://doi.org/10.1016/b978-0-12-821259-2.00014-4 (2021).
7. Monshi, M. M. A., Poon, J. & Chung, V. Deep learning in generating radiology reports: A survey. *Artif. Intell. Med.* **106**, 101878. https://doi.org/10.1016/j.artmed.2020.101878 (2020).
8. Wasmann, J.-W.A. *et al.* Computational audiology: New approaches to advance hearing health care in the digital age. *Ear Hear.* **42**(6), 1499–1507. https://doi.org/10.1097/aud.0000000000001041 (2021).
9. Elbaşı, E., Obali, M. Classification of hearing losses determined through the use of audiometry using data mining, in *Conference: 9th International Conference on Electronics, Computer and Computation* (2012).
10. Walker, J. J., Cleveland, L. M., Davis, J. L. & Seales, J. S. Audiometry screening and interpretation. *Am. Fam. Phys.* **87**(1), 41–47 (2013).
11. Chandrasekhar, S. S. *et al.* Clinical practice guideline: Sudden hearing loss (Update). *Otolaryngol. Head Neck Surg.* **161**(1_suppl), 1-S45. https://doi.org/10.1177/0194599819859885 (2019).
12. Wardenga, N. *et al.* Hearing aid treatment for patients with mixed hearing loss. Part II: Speech recognition in comparison to direct acoustic cochlear stimulation. *Audiol. Neurotol.* **25**(3), 133–142. https://doi.org/10.1159/000504285 (2020).
13. Margolis, R. H. & Saly, G. L. Toward a standard description of hearing loss. *Int. J. Audiol.* **46**(12), 746–758. https://doi.org/10.1080/14992020701572652 (2007).
14. Audiometric classification of hearing impairments. *International Bureau for Audiophonology*. Available at: https://www.biap.org/en/recommandations/recommendations/tc-02-classification/213-rec-02-1-en-audiometric-classification-of-hearing-impairments/file (1996).
15. Margolis, R. H. & Saly, G. L. Asymmetric hearing loss: Definition, validation, and prevalence. *Otol. Neurotol.* **29**(4), 422–431. https://doi.org/10.1097/MAO.0b013e31816c7c09 (2008).
16. Lee, C.-Y., Hwang, J.-H., Hou, S.-J. & Liu, T.-C. Using cluster analysis to classify audiogram shapes. *Int. J. Audiol.* **49**(9), 628–633. https://doi.org/10.3109/14992021003796887 (2010).
17. Pasta, A., Petersen, M. K., Jensen, K. J., and Larsen, J. Rethinking hearing aids as recommender systems, in *CEUR Workshop Proceedings*, Vol. 2439, 11–17 (2019).

18. Guo, R., Liang, R., Wang, Q. & Zou, C. Hearing loss classification algorithm based on the insertion gain of hearing aid. *Multimed. Tools Appl.* https://doi.org/10.1007/s11042-023-14886-0 (2023).
19. Belitz, C., *et al.* A machine learning based clustering protocol for determining hearing aid initial configurations from pure-tone audiograms, in *Interspeech*. https://doi.org/10.21437/interspeech.2019-3091 (2019).
20. Charih, F., Bromwich, M., Mark, A. E., Lefrançois, R. & Green, J. R. Data-driven audiogram classification for mobile audiometry. *Sci. Rep.* https://doi.org/10.1038/s41598-020-60898-3 (2020).
21. Abeer, E. *et al.* Data-driven audiogram classifier using data normalization and multi-stage feature selection. *Sci. Rep.* https://doi.org/10.1038/s41598-022-25411-y (2023).
22. Crowson, M. G. *et al.* AutoAudio: Deep learning for automatic audiogram interpretation. *J. Med. Syst.* https://doi.org/10.1007/s10916-020-01627-1 (2020).
23. Barbour, D. L. & Wasmann, J.-W.A. Performance and potential of machine learning audiometry. *Hear. J.* **74**(3), 40–44. https://doi.org/10.1097/01.hj.0000737592.24476.88 (2021).
24. Aziz, B., Riaz, N., Rehman, A., Ur Malik, M. I. & Malik, K. I. Colligation of hearing loss and chronic otitis media. *Pak. J. Med. Health Sci.* **15**(8), 1817–1819. https://doi.org/10.53350/pjmhs211581817 (2021).
25. Raghavan, A., Patnaik, U. & Bhandaria, A. S. An observational study to compare prevalence and demography of sensorineural hearing loss among military personnel and civilian population. *Indian J. Otolaryngol. Head Neck Surg.* **74**(S1), 410–415. https://doi.org/10.1007/s12070-020-02180-6 (2020).
26. Kassjański, M., Kulawiak, M. and Przewoźny, M. Development of an AI-based audiogram classification method for patient referral, in *Computer Science and Information Systems (FedCSIS), 2019 Federated Conference on*. https://doi.org/10.15439/2022f66 (2022).
27. *Guidelines for Manual Pure-tone Threshold Audiometry*, Vol. 20, 297–301 (ASHA, 1978).
28. Gao, L., Zhang, L., Liu, C. & Wu, S. Handling imbalanced medical image data: A deep-learning-based one-class classification approach. *Artif. Intell. Med.* **108**, 101935. https://doi.org/10.1016/j.artmed.2020.101935 (2020).
29. Zhu, M. *et al.* Class weights random forest algorithm for processing class imbalanced medical data. *IEEE Access* **6**, 4641–4652. https://doi.org/10.1109/ACCESS.2018.2789428 (2018).
30. Farid, DMd. & Mofizur Rahman, C. Assigning weights to training instances increases classification accuracy. *Int. J. Data Min. Knowl. Manag. Process* **3**(1), 13–25. https://doi.org/10.5121/ijdkp.2013.3102 (2013).
31. Hou, L., *et al.* Normalization helps training of quantized LSTM. *Neural Inf. Process. Syst.* (2019).
32. de Amorim, L. B. V., Cavalcanti, G. D. C. & Cruz, R. M. O. The choice of scaling technique matters for classification performance. *Appl. Soft Comput.* **133**, 109924. https://doi.org/10.1016/j.asoc.2022.109924 (2023).
33. Kassjański, M. *et al.* Detecting type of hearing loss with different AI classification methods: A performance review, in *Computer Science and Information Systems (FedCSIS), 2019 Federated Conference On*. https://doi.org/10.15439/2023f3083 (2023).
34. Banerjee, I. *et al.* Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. *Artif. Intell. Med.* **97**, 79–88. https://doi.org/10.1016/j.artmed.2018.11.004 (2019).
35. Wang, L., Wang, H., Song, Y. & Wang, Q. MCPL-based FT-LSTM: Medical representation learning-based clinical prediction model for time series events. *IEEE Access* **7**, 70253–70264. https://doi.org/10.1109/access.2019.2919683 (2019).
36. Sun, C., Hong, S., Song, M., & Li, H. (2020). A review of deep learning methods for irregularly sampled medical time series data.
37. Bengio, Y., Simard, P. & Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **5**(2), 157–166. https://doi.org/10.1109/72.279181 (1994).
38. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**(11), 2673–2681. https://doi.org/10.1109/78.650093 (1997).
39. Prusty, S., Patnaik, S. & Dash, S. K. SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Front Nanotechnol* https://doi.org/10.3389/fnano.2022.972421 (2022).
40. Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* **10**(7), 1895–1923. https://doi.org/10.1162/089976698300017197 (1998).
41. Kassjański, M. *et al.* Efficiency of artificial intelligence methods for hearing loss type classification: An evaluation. *J. Autom. Mob. Robot. Intell. Syst.* (in press).
42. Salzberg, S. L. C4.5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach. Learn.* **16**(3), 235–240. https://doi.org/10.1007/bf00993309 (1994).
43. Zhang, Y., Liu, Q. and Song, L. Sentence-state LSTM for text representation, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/p18-1030 (2018).
44. Kowsher, Md. *et al.* LSTM-ANN & BiLSTM-ANN: Hybrid deep learning models for enhanced classification accuracy. *Procedia Comput. Sci.* **193**, 131–140. https://doi.org/10.1016/j.procs.2021.10.013 (2021).
45. Khan, M., Wang, H., Ngueilbaye, A. & Elfatyany, A. End-to-end multivariate time series classification via hybrid deep learning architectures. *Pers. Ubiquit. Comput.* https://doi.org/10.1007/s00779-020-01447-7 (2020).
46. da Silva, D. G. *et al.* Comparing long short-term memory (LSTM) and bidirectional LSTM deep neural networks for power consumption prediction. *Energy Rep.* **10**, 3315–3334. https://doi.org/10.1016/j.egyr.2023.09.175 (2023).
47. Pirani, M., Thakkar, P., Jivrani, P., Bohara, M.H. and Garg, D. A comparative analysis of ARIMA, GRU, LSTM and BiLSTM on financial time series forecasting, in *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*. https://doi.org/10.1109/icdcece53908.2022.9793213 (2022).
48. Shahin, A. I. & Almotairi, S. A deep learning BiLSTM encoding-decoding model for COVID-19 pandemic spread forecasting. *Fractal Fract.* **5**(4), 175. https://doi.org/10.3390/fractalfract5040175 (2021).
49. Abdul Lateh, M. *et al.* Handling a small dataset problem in prediction model by employ artificial data generation approach: A review. *J. Phys. Conf. Ser.* **892**, 012016. https://doi.org/10.1088/1742-6596/892/1/012016 (2017).
50. *Guidelines for Audiometric Symbols. Committee on Audiologic Evaluation*, 25–30. (American Speech-Language-Hearing Association, 1990).

Author contributions

M.K.S., M.K.L., T.P. and D.T. designed the plan and goals of presented research; T.P., D.T., J.K., A.M., K.K., A.K., P.M.D. and M.G. performed data collection; T.P., D.T. and J.K. conducted data classification; M.K.S. and M.K.L. performed the initial literature review; M.K.S. and M.K.L. designed, trained and tested the neural network; M.K.S., M.K.L. and T.P. drafted the initial sections of the manuscript; M.K.S. and M.K.L. wrote the results, discussion and conclusions and prepared the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.K.

P5. Publication P5

Author Contribution Statement

I declare that in the publication:

M. Kassjański, M. Kulawiak, T. Przewoźny, D. Tretiakow, A. Molisz, "Development and testing of an open source mobile application for audiometry test result analysis and diagnosis support," Scientific Reports, 15, 14302. https://doi.org/10.1038/s41598-025-99338-5. (2025)

my contribution, in accordance with CRediT (Contributor Role Taxonomy) was as follows: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization. Furthermore, my contribution percentage to the development of the publication was 70%.

01.07.2025 Michał Kassjański

Date Michał Kassjański

I, the undersigned, hereby certify that the information given by Michał Kassjański is correct.

01.07.2025 Marcin Kulawiak

Date Marcin Kulawiak

16.07.2025

Date Tomasz Przewoźny

15.07.2025

Date Dmitry Tretiakow

16.07.2025

Date Andrzej Molisz

scientific reports

OPEN

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

Michał Kassjański[1], Marcin Kulawiak[1], Tomasz Przewoźny[2], Dmitry Tretiakow[3] & Andrzej Molisz[2]

Hearing impairments are typically assessed using pure tone audiometry, a diagnostic method that allows for the identification of the degree, type and configuration of hearing loss. The results of this assessment are generally displayed in the form of an audiogram, which graphically represents the softest sounds perceivable by an individual across a range frequencies. This paper presents a novel Open Source mobile application for the Android operating system that allows users to scan and analyse audiograms using a smartphone camera and subsequently classify the type of hearing loss. The application workflow is divided into three main stages: scanning, digitalization and classification of the audiogram. For this purpose, the application implements several artificial intelligence and image processing techniques, including YOLOv5, Optical Character Recognition (OCR) and Hough Transform. The scanned audiogram is analysed by a clinically validated AI model for classification of audiometric test results, providing clinicians with valuable assistance in formulating a diagnosis. All implemented algorithms and models were optimized for functionality on mobile devices. The application was evaluated on three distinct classes of smartphones across various price points, demonstrating its efficacy and consistent performance. The presented mobile application constitutes an advanced AI-driven decision support system that is readily accessible to general practitioners, otolaryngologists and audiologists. Its integration in medical facilities presents a substantial opportunity to decrease clinical workload, enhance diagnostic accuracy and reduce the likelihood of human error in hearing loss evaluations, which is particularly important in developing countries.

Keywords Audiogram, Hearing loss type, Mobile app, Audiogram digitalization, Public health

Hearing serves as a vital sensory function that is integral to our daily experiences. Any impairment of its functionality may significantly affect one's communication skills, relationships and the general understanding of their surroundings. Untreated hearing loss ranks as the third leading cause of long-term disability worldwide. This condition impacts a wide range of individuals, spanning various age groups and has consequences for individual persons and their families as well as entire economic systems. The global economy faces an estimated annual loss of around 1 trillion US dollars due to issues with inefficient diagnosis and treatment of hearing loss[1]. The pressing nature of this public health issue is heightened by forecasts suggesting a notable increase in the population experiencing hearing loss in the coming decades. At present, it is estimated that more than 1.5 billion individuals face different levels of hearing impairment, a number projected to rise to 2.5 billion by 2050, according to the World Health Organization (WHO)[1]. Tackling this looming crisis demands urgent focus and a unified approach to raise awareness, increase access to hearing healthcare and execute impactful intervention strategies.

The early identification and effective management of hearing impairment, particularly in children, play a crucial role in minimizing the impact of auditory deficiencies. Studies show that early detection can greatly reduce occurrences of childhood hearing loss, leading to improved developmental results[1]. Medical and surgical interventions for ear diseases have shown effectiveness in restoring hearing function, frequently enabling patients

[1]Department of Geoinformatics, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdańsk, Poland. [2]Department of Otolaryngology, Medical University of Gdańsk, Gdańsk, Poland. [3]Department of Otolaryngology, The Nicolaus Copernicus Hospital in Gdańsk, Copernicus Healthcare Entity, Gdańsk, Poland. email: michal.kassjanski@pg.edu.pl

to regain their original hearing capabilities. Nonetheless, the effective diagnosis and management of hearing loss are fundamentally connected to the presence of sufficient and sustainable hearing healthcare resources. A major obstacle to the effectiveness of hearing health systems is the lack of trained professionals who can provide essential audiological services[1]. This problem is especially evident in low-income countries, where the number of ear, nose, and throat (ENT) specialists is less than one for every million individuals. The limited availability of audiologists significantly intensifies the difficulties encountered in meeting the hearing health needs of these populations[1].

The issue is exacerbated by the fact that while skilled hearing healthcare practitioners can manually identify and treat certain types of hearing loss, many problems can only be properly identified with the use of pure tone audiometry - a method widely recognized as the gold standard for assessing auditory functionality. This evaluation procedure quantifies audiometric threshold shifts, aiding in the classification of hearing loss into specific categories: conductive, sensorineural or mixed. The extent of hearing loss can vary from mild to profound, greatly affecting a person's quality of life. Evaluating auditory function through pure tone audiometry is crucial not only for individual diagnosis but also for advancing epidemiological studies and developing effective rehabilitation strategies[1]. The outcomes of pure tone audiometry are generally illustrated through an audiogram, which serves as a graphical depiction showing the lowest sound intensity, quantified in decibels, that an individual is capable of perceiving across various frequencies. This data provides a thorough insight into an individual's hearing abilities and acts as an essential resource for professionals in crafting personalized strategies for individuals facing hearing challenges. However, precise classification of audiograms with respect to the specific location of hearing impairment poses a considerable challenge even for experienced clinicians. General practitioners, in particular, often encounter considerable obstacles when attempting to interpret audiograms accurately[2]. This is inter alia because there can be as many as 3.62 million potential configurations of audiograms derived solely from air conduction alone[3]. Such diverse representations of audiograms lead to substantial interpretive difficulties even for professionals in the field[4].

In literature, considerable attention has been focused on creating automated techniques for classifying audiograms for the purpose of hearing aid setup. Nevertheless, the analysis of audiometry test results for the purpose of diagnosing various forms of hearing loss has received considerably less attention. Notably, Crowson et al.[5] introduced an automated approach for classifying hearing loss types on audiogram images. Their research employed a ResNet model, achieving a classification accuracy of 97.5%. Although this solution, named AutoAudio, demonstrates impressive accuracy, there are certain obstacles to implementing it in a clinical environment, in particular in developing countries. The ResNet architecture, which includes deep convolutional neural networks, requires significant computational resources for analyzing audiograms, which presents challenges for the mobile environment. Furthermore, the model's training was performed on a relatively small dataset of 1007 audiograms, and limited to one specific type of audiogram, which constrains its applicability across different contexts. The structural features of audiograms generally exhibit a level of consistency; however, important discrepancies may arise from the differing hardware and software configurations employed during their creation. In addition, the authors have not published the trained model nor the training dataset, which significantly impairs independent validation and advancement of their methodology. In this context, Kassjanski et al.[6] have proposed a hearing impairment classifier which directly analyses audiometry test results. The presented AI model, trained on 15,046 audiometry test results, uses Bi-LSTM layers to analyse a patient's frequency responses and produce a diagnosis with over 99% accuracy. However, clinicians predominantly interpret auditory assessment outcomes in the form of printed audiograms instead of numerical test results. In consequence, the implementation of this model in clinical practice would require the users to manually convert each analysed audiogram into a set of numbers for classification, which would be very time consuming.

Summing up, AutoAudio, which focuses on interpretation of audiogram images, may exhibit generalization issues due to being trained on a small dataset. Moreover, due to using a deep model, it is poorly fit for the hardware capabilities of mobile devices, particularly in low-income areas. On the other hand, the Bi-LSTM model, although trained on a much larger dataset and using a potentially much more performant architecture, cannot interpret imagery data. In conclusion, although recent years have seen advancements in the automated classification of audiometry data, especially in terms of accuracy, considerable challenges persist in implementing these methods for broad clinical application.

This paper presents a mobile application aimed at improving the efficiency of audiogram analysis by leveraging smartphone technology. The proposed application utilizes the functionalities of a smartphone to scan, process and classify audiograms using a model derived from the work of Kassjanski et al.[6]. The results enable direct classification of audiograms via a smartphone application, which could streamline the diagnostic process The application enables general practitioners to independently classify tonal audiometry results for the purpose of subsequent patient referrals. Consequently, this study not only seeks to minimize the number of uncomplicated cases referred to audiologists but also facilitates the allocation of greater time and resources to the management of complex cases that necessitate specialized care. Additionally, the application may also be used as an educational resource for budding audiologists, providing them with a platform to hone their diagnostic skills and deepen their understanding of audiometric evaluation. Through this approach, the presented application stands to contribute significantly to both clinical practice and education within the field of audiology.

Materials and methods

The review of literature relevant to the presented research is systematically organized into three primary domains: audiogram digitalization, classification of hearing loss types and mobile-based diagnostic decision support systems.

Audiogram digitalization

The digitization of audiograms signifies an essential progress required for the successful application of modern automated audiogram analysis models. Automated systems are generally trained on pre-processed digital data produced by specialized audiometric software, ensuring precision and uniformity in analysis. Nonetheless, considerable obstacles emerge when the operational setting transitions to clinical environments, where professionals may be limited to printed audiograms. In these situations, the unavailability of accessible digital data limits the model's usability and efficiency. The current literature on this specific issue is limited to the publications listed below.

The initial effort was carried out by Li et al.[7], who developed multiple convolutional neural networks to extract audiograms, symbols and axis labels from audiogram images. The integration of outcomes from all models yields a digital depiction of the audiogram. The system demonstrated 98% accuracy on scanned images, while achieving 84% accuracy on photographs captured with a camera.

Subsequently, Chairh and Green[8] introduced an advanced digitalization tool that employs YOLOv5 for symbol recognition and Tesseract for label identification. The dataset included 3,200 reports, in contrast to only 420 reports analyzed by Li et al.[7]. This investigation took into account all audiological symbols, encompassing those obscured from air and bone conduction. The audiogram, axis label and symbol models achieved mAP@0.5 scores of 84%, 34% and 39%, respectively.

The latest work was carried out by Yang et al.[9], who introduced a system similar to that of Chairh and Green[8], featuring a multi-stage integration of YOLOv5 models alongside an optical character recognition (OCR) model. The investigation analyzed both pure tone audiometry and sound field testing. The accuracy rate at each stage was around 98%, based on 2,535 samples used for audiogram detection and 2,214 records applied for symbol detection.

In summary, the process of audiogram digitalization can be delineated into two primary methodologies: the employment of convolutional neural networks and the integration of YOLO in conjunction with OCR models. Recent advancements in this field, as illustrated by Yang et al.[9], utilize the latter approach, attaining an accuracy rate of approximately 98%. The elevated accuracy highlights the effectiveness of integrating YOLO with OCR technologies to improve digitalization processes, especially in applications necessitating precise object detection and text recognition. In consequence, this approach was selected for implementation in the presented application for the purpose of label detection, providing improved detection accuracy even with noisy data.

Classification of hearing loss types

The majority of publications regarding audiogram classification have concentrated on the automation of hearing aid configuration, rather than identification of the various types of hearing loss.

In this context, Crowson et al.[5] employed the AutoAudio model to categorize audiogram images into four different types of hearing: normal hearing and three types of hearing loss (sensorineural, conductive, and mixed). A dataset comprising 1007 static audiograms was used for training and testing purposes. Due to the relatively small size of the training dataset, the authors opted for transfer learning. From all evaluated architectures the ResNet-101 model attained the highest classification accuracy of 97.5%. The weaknesses of this method include the relatively small size of the training dataset and the fact that it is highly dependent on the style and presentation of processed audiogram images.

Other attempt at classifying types of hearing loss have focused on processing raw audiometry data instead of audiogram images[10]. In this area, the best results have been achieved by Kassjanski et al.[6], who proposed a Bi-LSTM-based model for classification of audiometry test results into normal hearing and all three types of hearing loss. The study involved an analysis of 15,046 audiometry test result samples. The proposed model achieved a classification accuracy of 99.33%. In consequence, the proposed Bi-LSTM model demonstrated accuracy superior not only to other methods working with numerical audiometry data, but also that demonstrated by AutoAudio[5], making it the best choice for inclusion in the presented application.

Mobile-based diagnostic decision support systems

Clinical decision support applications for mobile operating systems hold significant promise for improving the access to medical services as well as their quality. However, the current challenge for most healthcare systems lies in the insufficient level of digitalization. Consequently, implementing an application that enhances the medical staff's work while simultaneously ensuring standards of patient data protection presents significant challenges[11]. Nonetheless, it is evident that efforts are underway in this domain and usage of this type of application is visibly increasing and expected to become significantly more prevalent in the future[12–14].

The incorporation of smartphone technology into medical imaging has transformed how healthcare professionals analyze and interpret medical data. One particular example is a mobile application that allows users to analyze medical images using a smartphone camera, as illustrated in[15,16]. This enables medical professionals to efficiently employ automated classification models on medical data, thereby optimizing the diagnostic process. In consequence, the accessibility and convenience of smartphones is used to expedite image analysis and enhance patient outcomes through prompt and precise diagnosis[17].

In the domain of otolaryngology, Kanimozhi et al.[18] proposed a mobile application for hearing impairment diagnosis via on-device audiometry test assessment with machine learning. A decision tree classifier was integrated into the application to enhance the precision of test result analysis and classification, facilitating the identification of different levels of hearing loss (normal, mild, moderate, moderately severe, severe, profound). Furthermore, the application provides recommendations based on the test outcomes, proposing potential subsequent actions for the user, such as pursuing a professional evaluation, consulting a healthcare provider or performing additional testing if hearing loss is identified. Unfortunately, the application is only capable of assessing a general degree of hearing loss, not its particular type (which is crucial for effective treatment).

The idea of performing audiometry testing on mobile devices has been implemented in several applications[19–21], however it has always encountered fundamental difficulties linked to acquiring precise audiometric measurements in non-clinical contexts, especially in a home setting and without the oversight of a qualified audiologist. This has been exemplified e.g. by Masalski and Kręcicki[22] who analysed pure tone audiometry results across three different testing environments: clinical evaluations performed with an audiometer, self-administered assessments conducted on a specially calibrated computer under the supervision of an audiologist, and independent self-tests carried out at home. The findings reveal a mean difference in hearing threshold values of -1.54 dB between the first and second testing series, with a standard deviation of 7.88 dB. Furthermore, the data generated from the first and third series exhibited a mean difference of -1.35 dB accompanied by a standard deviation of 10.66 dB. The findings indicate that while mobile-based audiometry offers a practical option for general hearing assessments, it cannot be used as a replacement for clinical tests.

In summary, while mobile-based diagnostic decision support systems have seen a wide range of applications, they do not present a viable alternative to professional audiometric evaluation. At the same time, there is presently no solution that integrates the capability to diagnose the type of hearing loss from an audiogram image captured with a smartphone camera. In this context, the presented research aims to develop such an application through integration of state-of-the-art methods for image recognition and analysis as well as audiometry test result classification. The application would perform audiogram digitization with the help of YOLO, Optical Character Recognition (OCR), and image processing techniques to extract raw audiometric data, which would subsequently serve as input for the Bi-LSTM audiometry test result classification model. To fulfil the functional requirements for operating on medical data, all computations would need to be executed on the mobile device. Furthermore, the application would feature a simple and intuitive user interface, enabling easy and efficient use in a clinical environment.

Ethics declarations

This study was approved by the Committee on Research Ethics at Medical University of Gdańsk (IRB KB/23/2024). Prior to participating in the hearing test and this study, the subjects provided informed consent. All methods were carried out according to the relevant guidelines and regulations.

Design of the application

The methodology employed for the processing and classification of pure-tone audiometry test results on mobile devices can be delineated into three distinct stages: scanning, digitization and classification of audiograms. The initial stage, scanning, entails the acquisition of audiometric data from conventional audiogram images, accomplished through the use of a smartphone camera. The digitization stage is essential, as it involves transforming the scanned audiograms into a digital format that is appropriate for analysis. The classification stage systematically organizes the digitized audiometry test results, enabling a thorough analysis of hearing loss type.

Audiogram scanning

The scanning process was realized using the ML Document Scanner from Google's ML Kit[23]. This library simplifies the scanning process by enabling the user to simply position their smartphone camera over the document for automated capture. Furthermore, perspective correction was applied along with automatic rotation detection to ensure that documents are displayed in an upright position. Furthermore, applications using the ML Kit do not need to ask for permission to use the smartphone camera. Instead, the system utilizes the camera permission from Google Play services, allowing users to manage which files they choose to share with the application. This design decision allows users to retain authority over the files they choose to share with the application, thus improving user privacy and trust in the scanning procedure.

The YOLOv5 architecture created by Ultralytics[24], which has been pre-trained on the COCO dataset[25], was used to train a model dedicated for audiogram detection. The dataset used for the training and testing process was derived from a previous study[6]. The dataset is comprised of 15,046 audiometric test results derived from 9,663 adult subjects who underwent evaluation at the Otolaryngology Clinic of the University Clinical Centre in Gdansk, Poland, between 2010 and 2022. Among the participants, 5,591 identified as female (57.86%) and 4,072 as male (42.14%), with ages ranging from 18 to 98 years. Only two test results (corresponding to the left and right ears) were included for each subject, thereby ensuring the data integrity through the prevention of duplication and promoting diversity within the dataset. To optimize the YOLOv5 model for the detection of audiograms, 100 test results were randomly selected from the entire pool. The hearing test reports were printed and photographed, followed by manual annotation, resulting in 200 sample audiograms, as each report contains two audiograms. During the selection of the audiogram area, slightly larger bounding boxes were intentionally chosen to avoid tight enclosures that might inadvertently cut off more peripheral text. This decision was made to ensure that critical contextual information is preserved for subsequent recognition tasks, which is crucial for ensuring the overall accuracy of the detection model.

Audiogram digitization

The process of digitizing an audiogram involves three essential stages: line detection, symbol detection and label detection. In the preliminary phase, line detection focuses on recognizing the graphical lines that illustrate the different thresholds of hearing at various frequencies. Subsequently, the emphasis on symbol detection involves identifying the particular symbols that align with these lines, which reflect auditory responses. Ultimately, label detection is essential for linking each identified line and symbol to their corresponding labels, which indicate the frequency and intensity levels. A thorough grasp of the spatial layout of these individual lines, along with their corresponding labels, enables an accurate interpretation of the symbol values represented in the audiogram.

Firstly, the detection of lines on the audiogram was accomplished using the Probabilistic Hough Transform[26] method, a variant of the original Hough Transform[27]. The development aimed to address certain limitations of the conventional approach, particularly the high computational costs associated with processing large images. The Probabilistic Hough Transform operates by randomly choosing a subset of edge pixels from the image and subsequently fitting lines to those selected pixels. This procedure is conducted repeatedly, with the expectation that each cycle will identify further pixels associated with the same line. A consensus set of pixels is ultimately established, followed by the application of least-squares regression to fit a line to that set. The Probabilistic Hough Transform offers a significant speed advantage over the traditional Hough Transform, which is essential for mobile applications. However, it may sacrifice some accuracy, particularly in images with numerous noisy or spurious edge pixels. To counteract this potential drawback, the system was further optimized through the incorporation of an interpolation method. This method calculates the position of any undetected lines by leveraging the spatial coordinates of the two closest detected lines, thereby enhancing the overall robustness of the line detection process in audiograms. The implementation of the Probabilistic Hough Transform begins with the crucial step of acquiring an edged image. The widely recognized Canny edge detection method[28] was utilized to accomplish this task. Prior to implementing the Canny algorithm, the image is converted to grayscale. This simplification improves edge detection by minimizing the complexities linked to color information. Moreover, the parameters employed in the Canny edge detector are determined by the median value of the image, which guarantees that the threshold settings are adaptively modified to align with the unique features of the input image. Figure 1 illustrates the line detection results.

Secondly, the architecture of YOLOv5s was employed to accurately recognize symbols on audiograms, paralleling the established methodology used in audiogram detection. To train the model, 200 audiogram images were extracted from the training dataset created for the audiogram detection stage. To address the limited availability of audiograms with masked symbols, the dataset was further augmented by incorporating 32 additional audiograms that included such symbols. This enhancement brought the total number of manually annotated audiograms to 232. Before the training process started, the entire audiogram dataset was subjected to binarization using the Otsu method. This operation aimed to enhance data recognition efficiency while reducing the effects of noise and the natural fluctuations found within the dataset. Additionally, the binarization process is essential for removing most of unnecessary audiogram lines that could be located near the symbols of interest. Such lines could potentially cause incorrect detections, which could undermine the precision of the symbol identification process. Figure 2 illustrates a sample audiogram before and after the binarization process.

A total of 8 distinct classes, each corresponding to various audiological symbols, were considered for the 4 different types of measurement. The symbols are presented in Table 1.

Thirdly, the application of Optical Character Recognition (OCR) technology, particularly the Machine Learning Kit Text Recognition v2 API created by Google, alongside the fine-tuned YOLOv5s model, has proven to offer a strong foundation for detecting labels on audiograms. Preliminary investigations utilized the Open Source Tesseract engine, however results were less than satisfactory, especially on mobile devices. This resulted in the implementation of a specialized OCR system designed specifically for mobile applications, thus avoiding the need for extensive model training, which is frequently essential when using Tesseract. Even with these

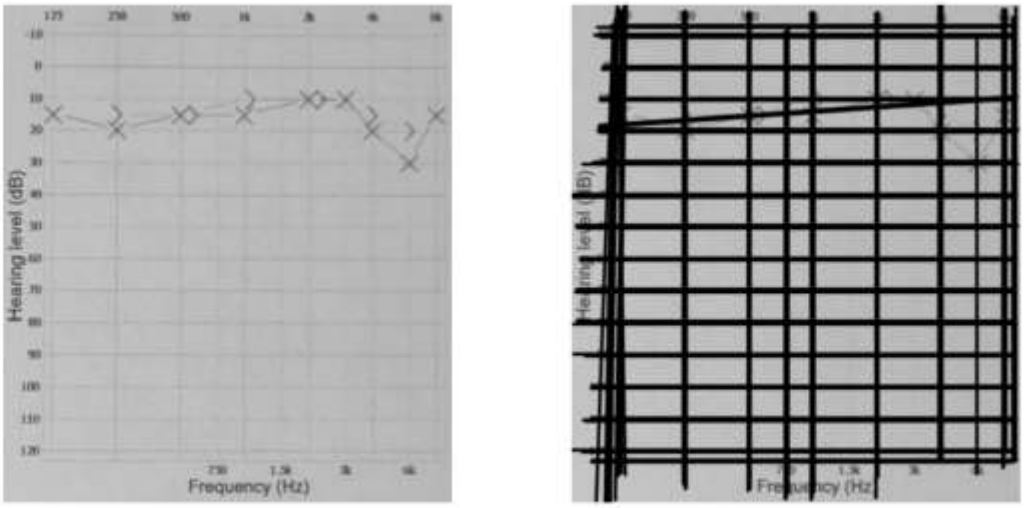


Fig. 1. The grayscale image of a photographed audiogram (left) alongside the lines detected using the Probabilistic Hough Transform method (right). Image represents data produced by the presented smartphone application.

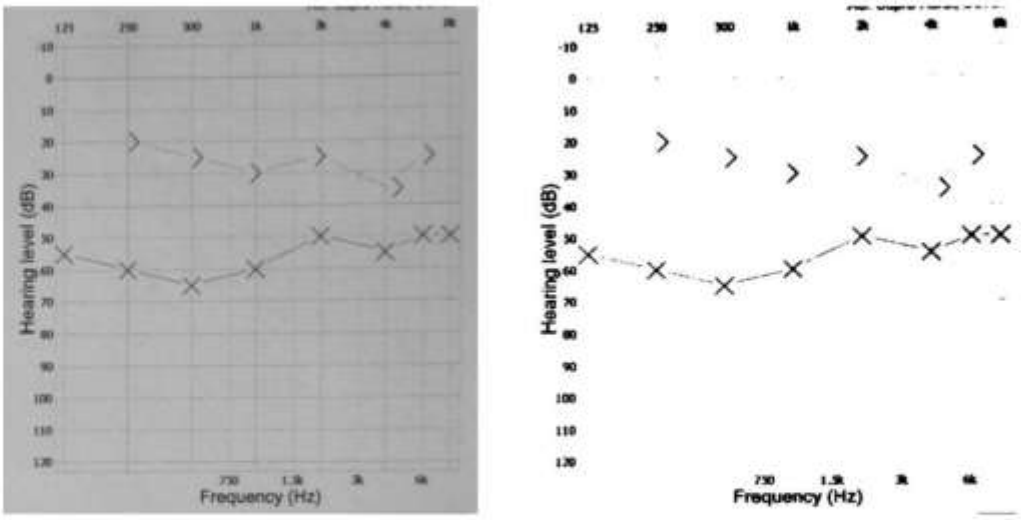


Fig. 2. A photographed audiogram image prior to (left) and subsequent to (right) Otsu binarization. Image represents data produced by the presented smartphone application.

Measurement type		Ear	
Stimulation	Masking	Left	Right
Air	No	×	○
Air	Yes	□	△
Bone	No	>	<
Bone	Yes	]	[

Table 1. Audiological symbols identified on audiograms during digitization.

advancements, the OCR's performance in noisy image conditions demonstrated inherent limitations, as it did not consistently recognize all labels, particularly when the text was blurred. An additional layer incorporating YOLOv5 detection was integrated into the process to tackle these challenges. The dataset comprising 232 audiograms, which had previously been utilized for symbol detection, was subjected to a meticulous manual annotation process. This resulted in a dataset specifically employed for label detection modal training. This model was developed to identify 22 unique categories, covering hearing levels from −10 dB to 120 dB and frequencies from 125 Hz to 8 kHz. Regarding image preprocessing techniques, a conscious decision was made to omit binarization, as initial tests indicated that this method resulted in a decline in recognition accuracy. Instead, the images were converted to greyscale, which was found to be more effective in enhancing the performance of the OCR and YOLOv5 integration.

Hearing loss type classification

In the context of final audiogram classification, the implementation of Bi-LSTM model, as detailed in the previous study[9], played a pivotal role in the stage of diagnosing hearing loss type. The model was trained on a comprehensive dataset comprising 15,046 audiometry test results. Each audiometric evaluation consists of seven distinct lists that correspond to air and bone conduction thresholds, measured in decibels across a range of frequencies - specifically, 125 Hz, 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz and 8000 Hz. Prior to inputting this data into the initial layer of the network, comprehensive normalization was performed using Z-score normalization. This crucial step ensures that the data is scaled appropriately, facilitating improved convergence during training. The first Bi-LSTM layer of the network utilizes dual Long Short-Term Memory (LSTM) units to examine temporal sequences in both forward and reverse orientations. This bidirectional approach facilitates a deeper comprehension of the fundamental patterns present in the audiometric data. Following the first Bi-LSTM layer, the model includes a dropout layer aimed at reducing the likelihood of overfitting by randomly turning off a portion of the neurons during the training process. Subsequent to this, further layers of LSTM and dropout are incorporated, which improve the model's ability to generalize from the training data. The architecture concludes with a dense layer that incorporates a softmax activation function, enabling the classification of four main types of hearing loss: normal hearing, sensorineural hearing loss, conductive hearing loss, and mixed hearing loss. By

employing a thorough assessment with stratified 10-fold cross-validation, the Bi-LSTM model attained average accuracy of 99.33%, highlighting its effectiveness in the precise classification of hearing loss types.

Optimization of neural networks for mobile use

This study utilized three detection models to analyze audiograms, symbols, and labels, employing the You Only Look Once (YOLO) model, first introduced in 2016 by Redmon et al.[29]. The YOLO framework leverages convolutional layers to predict both the bounding boxes and class probabilities of all objects depicted within an image. Since the YOLO algorithm is a single-shot detector, it only looks at the image once. The algorithm calculates a confidence score for each predicted bounding box by multiplying the probability of an object being present within a specified grid with the Intersection over Union (IoU) metric, which assesses the overlap between the predicted bounding box and the ground truth. Currently, YOLO has experienced numerous iterations, with versions spanning from 1 to 11. This study specifically utilized YOLO version 5, which has been previously validated in research involving its application in mobile platforms[30]. Moreover, YOLOv5 exhibits superior computational efficiency and diminished video memory consumption relative to its predecessors, while having a smaller memory footprint than its successors, rendering it advantageous for use in mobile applications.

Upon completion of the training process, all three YOLOv5 models were exported to TorchScript format. The procedure was performed using the function provided by Ultralytics. Furthermore, the models were refined for mobile usage[31], through the following optimizations:

- Conv2D + BatchNorm fusion - the Conv2d and BatchNorm2d operations are integrated into a single Conv2d operation within the forward method of the respective module and all its submodules;
- Insert and Fold prepacked ops - the computation graph is modified by substituting standard 2D convolutions and linear operations with their prepacked counterparts;
- ReLU/Hardtanh fusion - output activations within the convolutional kernel are clamped, thereby streamlining the computational process. This technique is applicable to both 2D convolution and linear operation kernels;
- Dropout removal - dropout and dropout_ nodes are eliminated from the module when the training mode is inactive;
- Conv packed params hoisting - convolution packed parameters are relocated to the root module. This adjustment enables the removal of convolution structures;
- Add/ReLU fusion - instances where ReLU operations follow addition operations are combined into a singular add_relu operation.

Through these optimizations, provided by PyTorch Mobile[31], the YOLOv5 models have been significantly refined, making them suitable for deployment in mobile applications. To successfully integrate the models into the Android application, the PyTorch Android Lite library, specifically version 1.13.1 was employed. This library is designed to streamline the process of loading and executing machine learning models on mobile devices, thereby enabling efficient inference capabilities.

Furthermore, the deployment of the hearing loss classification model also required optimization for mobile device compatibility. The original Bi-LSTM model[6] developed in Keras has been converted into TensorFlow Lite format using post-training quantization techniques. The conversion technique led to a notable decrease in model size and enhanced latency during execution on Central Processing Units (CPUs) and hardware accelerators, while still preserving an acceptable degree of accuracy degradation. This being said, it must be noted that TensorFlow Lite does not currently offer native support for converting Bi-directional LSTM models. As a result, the Multi-level Intermediate Representations (MLIR) conversion method was utilized as a feasible option, even though it has not yet received widespread acknowledgment and validation. A comparative analysis was conducted to evaluate the effectiveness of this conversion process, utilizing a dataset of 1,000 samples sourced from the test data. The performance of the original Bi-LSTM model developed in Keras and the quantized TensorFlow Lite model was assessed. The results from both models showed a significant level of agreement, matching to five decimal places across all test instances. Additionally, the input tensor for the TensorFlow Lite model was structured to replicate that of the Bi-LSTM Keras model, exhibiting a shape of (1, 7, 2) and utilizing a data type of float32. This configuration pertains to the examination of seven unique frequencies across two different conductions, highlighting the model's proficiency in managing intricate data effectively.

System architecture

The presented application realizes the functionality of diagnosing hearing type from an audiogram on a smartphone through several steps realized by individual modules. The first module supports scanning of hearing test reports using the smartphone camera. In this phase the scanned document is refined via perspective correction and, if necessary, automatic rotation. Subsequently, the YOLOv5 model is used to detect and extract audiograms from the scanned report. Following the successful identification of an audiogram, the process of digitalization is started. This process involves several essential operations, such as applying grayscale transformation, detecting labels using OCR alongside YOLOv5, and performing line detection through the Hough Transform method. Simultaneously, the audiogram is subjected to a binarization process, enabling further symbol analysis via the YOLOv5 model. This detailed examination gathers essential data from symbols, labels, and lines, which is then carefully processed and combined in order to recover the original values recorded during the tonal audiometry test. The final module of the system effectively utilizes these values to classify hearing type through the Bi-LSTM model. The classified results are then displayed in a simple user interface, ensuring user accessibility. The comprehensive structure and functional dynamics of the proposed system are depicted in Fig. 3.

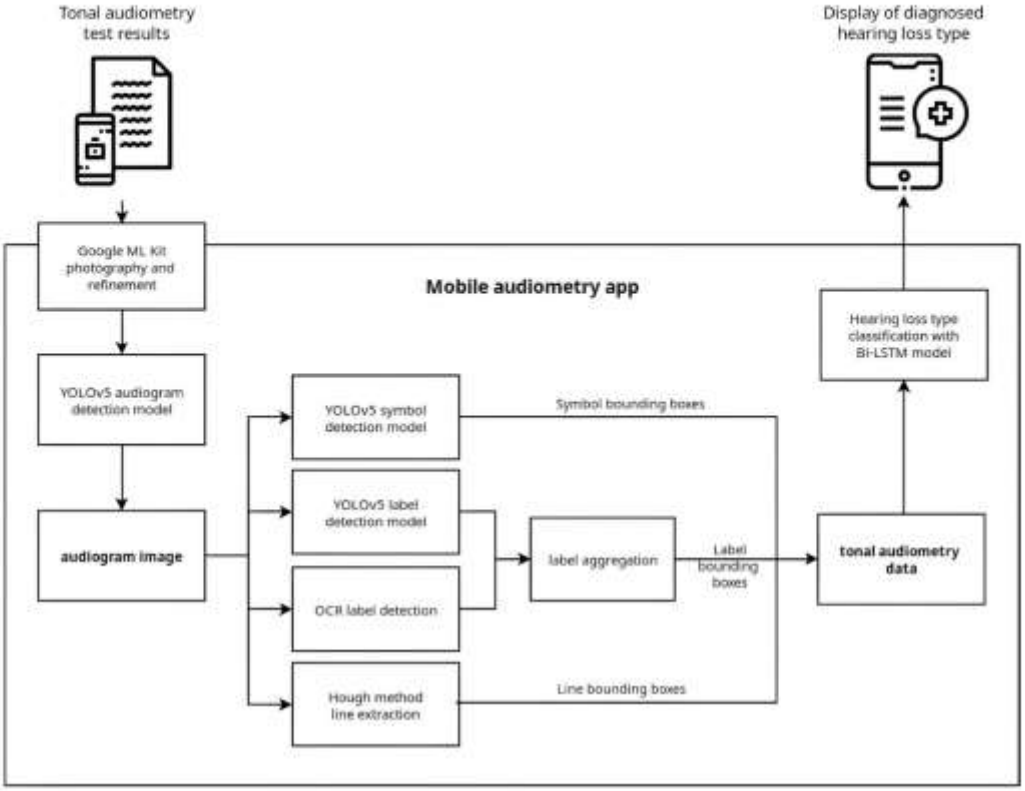


Fig. 3. General architecture of the presented system.

Evaluation of created mobile AI models

The k-fold cross-validation method[32] was utilized to evaluate the performance of the created mobile-optimized AI models. This methodical strategy involves dividing the dataset into k separate subsets, generated through a random shuffling of the data points to reduce any possible bias. In the process of k-fold cross-validation, the model undergoes iterative training on k-1 subsets, with one subset set aside as the validation set, allowing for a thorough evaluation of the model's predictive performance. This procedure is carried out k times, guaranteeing that every subset serves as a validation set precisely one time. The evaluation scores obtained from the various iterations provide a thorough and dependable assessment of the model's effectiveness. The k-fold cross-validation method reduces the potential for bias that could result from depending on a single train-test split by averaging the performance metrics over all k iterations. As a result, it offers a detailed insight into the model's ability to generalize, thus strengthening the reliability of the results in the realm of predictive modeling. The presented research used $k = 5$, which resulted in train to test dataset proportions of 80–20%, respectively.

The assessment of the YOLOv5 model's performance was carried out using the mean average precision (mAP) metric, which is a recognized standard for evaluating the effectiveness of object detection. The mAP metric integrates classification and localization elements, offering a thorough assessment of an algorithm's performance in identifying different objects. Additionally, mAP encompasses several essential elements, such as the precision-recall (PR) area under the curve (AUC), multiple object categories (MOC) and intersection over union (IoU). To achieve a suitable equilibrium in the precision-recall trade-off, the AUC is crucial in calculating mAP. The PR curve for each object category is produced by methodically adjusting the confidence thresholds linked to the model's predictions. The average precision (AP) for each individual class is then obtained from the PR curve, enabling effective classification and localization of multiple classes. Furthermore, the average precision is computed over various IoU thresholds, referred to as AP50 – 90. The average precision values (mAP50 – 90) are calculated to yield a comprehensive assessment of model performance across all specified IoU thresholds. Finally, the overall AP is computed by averaging the AP values calculated at each IoU threshold defined below:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i,$$

where AP_i is the average precision of each class and N is the total number of classes. The IoU threshold was set at 0.5 in this study. An IoU value of 0.5 or higher is classified as a True Positive (TP), signifying that the predicted bounding box adequately overlaps with the ground truth. Conversely, an IoU value falling below 0.5 is classified as a False Positive (FP), suggesting that the predicted bounding box does not accurately encompass the intended object. Additionally, cases where the model does not identify an object that exists in the ground truth are classified as False Negatives (FN). Lastly, True Negatives (TN) refer to the segments of the image background where no objects are identified, thus confirming the lack of any pertinent features in those regions. Following this, the evaluation of model performance was further enhanced by incorporating precision and recall (sensitivity) as metrics, calculated as follows:

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN}.$$

Mobile application testing

The evaluation of mobile applications involves assessing their functionality, usability, and performance across different devices. This process is essential to ensure that the apps meet quality standards and provide a seamless experience for users. This involves evaluating various elements including functionality, compatibility, security, and user interface to detect and resolve problems prior to the app's launch. In the case of the presented application, the testing was conducted manually in accordance to pre-planned scenarios.

The goal of the presented research was to make the application work correctly on a possibly wide range of smartphone devices from different price ranges. However, whereas a base hardware requirement for running the application could be defined based on the amount of system RAM needed to fit and execute the AI models, an important issue that required investigation was the impact of the quality of a smartphone camera on the functionality of the app. Consequently, an evaluation was conducted on three distinct smartphones across different budget categories, namely the Motorola Moto G82, Samsung Galaxy S21 5G, and Samsung Galaxy S23 Ultra. All of those devices operate on the Android system and possess a minimum of 6 GB of RAM. The distinction is evident in the primary camera specifications, beginning with the Motorola G82, which features a triple camera setup with a maximum resolution of 50 MP, while the Samsung S21 also boasts a triple camera configuration, but with a resolution of 64 MP. The most prominent camera system is found in the Samsung S23 Ultra, featuring a quad setup and a maximum resolution of 200 MP. It should be noted that although the devices were considered mid- to high-end when they debuted, all of them are over two years old at the time of writing, and thus may be acquired with discounts of up to 50% of their launch price. The comprehensive specifications of those devices can be found in Table 2.

Since the quality of a photograph is primarily determined by the lighting conditions under which it was taken, experiments were conducted under three distinct illumination levels selected according to the Common and Recommended Light Levels Indoors established by the National Optical Astronomy Observatory[56]: 50 lx, 500 lx, and 1000 lx, which correspond to "Dark surrounding", "Normal Office Work" and "Normal Drawing Work", respectively. After examining the outcomes from a range of tests, it became evident that there were no notable performance differences between the 500 lx and 1000 lx conditions. Consequently, the results presented further on will be narrowed to the illumination levels of 50 lx and 500 lx. The differences between images captured under those conditions are depicted in Fig. 4, which presents the same audiogram photographed at 50 and 500 lx.

The performance of different smartphone cameras was assessed by measuring the number of audiogram lines which have not been properly detected by the application due to insufficient image quality. This approach enabled concurrent evaluation of the performance of missing data interpolation methods implemented in the application.

The above-mentioned tests have been performed using a carefully selected set of audiograms which covered all types of hearing. Furthermore, for each classification of hearing type, the audiograms were categorized into three groups based on their complexity, which was defined by the number of presented symbols, the degree of symbol overlap, and the number of symbols overlapping with audiogram lines. A sample representation of all three groups, named Simple, Average and Complex is shown in Fig. 5. It should be emphasized that this

Device	Motorola Moto G82 5G[57]	Samsung Galaxy S21 5G[58]	Samsung Galaxy S23 Ultra[59]
Date of release	June 07 2022	January 29 2021	February 17 2023
OS version	Android 13	Android 14	Android 14
Chipset	Qualcomm Snapdragon 695	Exynos 2100	Qualcomm Snapdragon 8 Gen 2
RAM	6GB	8GB	8GB
Primary camera	50 MP with optical image stabilisation	12 MP with optical image stabilisation	200 MP with optical image stabilisation
Price on release	$272	$799	$1199

Table 2. The specification of devices on which the application was tested.

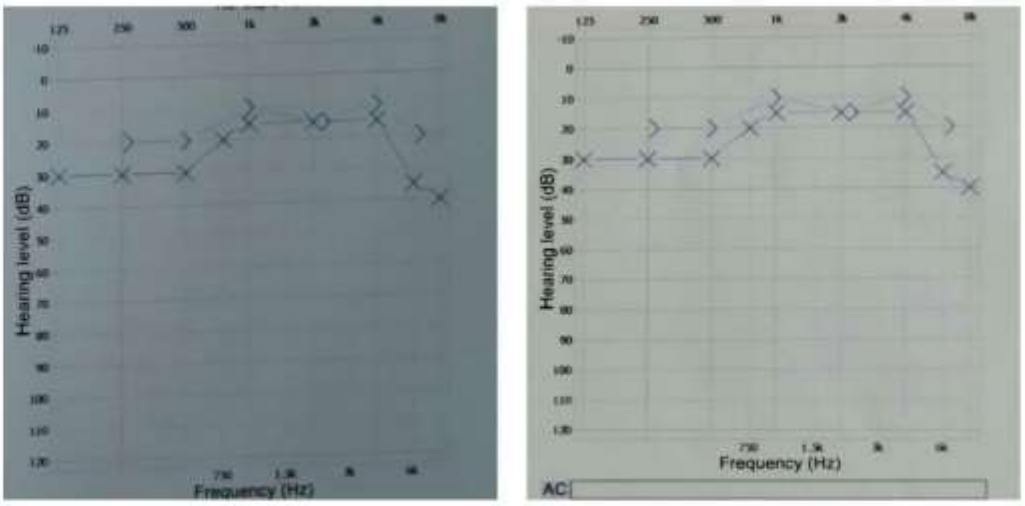


Fig. 4. The same audiogram photographed by the Motorola G82 at 50 lx (left), and 500 lx (right).

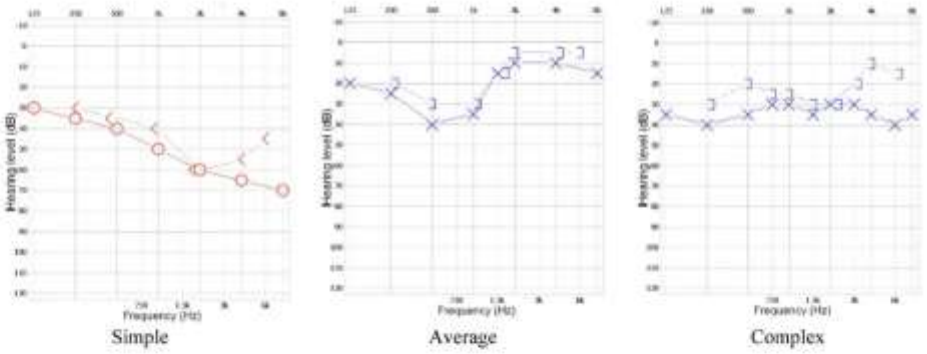


Fig. 5. Three audiograms representing levels of complexities.

gradation in audiogram complexity does not necessarily correspond to the challenges encountered in the medical classification of hearing types. This is because the presented evaluation focuses on the degree of difficulty posed by an audiogram to the detection system, particularly in relation to the number of objects present and the degree of observed overlap, instead of the values and relationships of the actual measurements.

All in all, a total of twelve audiograms of varying complexity were used across three different devices under two varying lighting conditions to evaluate the impact of smartphone camera quality on application performance. In addition, any detection errors identified during the tests were used to measure the effectiveness of the data interpolation methods implemented in the application.

Technical requirements for mobile application usage

The application's Software Development Kit (SDK) target has been designated as level 34 (Android 14) to comply with the prevailing standards for incorporating mobile applications onto the Google Play platform[37]. The minimum SDK level was set at 28 (Android 9) to fulfil the requirements of the Android packages employed within the application. From the hardware standpoint, a minimum of 6 GB of RAM is required to accommodate the execution of all implemented AI models as well as the ML Kit, which alone needs at least 1.7 GB of RAM. Moreover, at least 400 MB of storage space is essential for the app's installation due to the large size of the bundled AI models. Finally, a camera sensor with at least 12 MP resolution is also recommended. At the time of writing, new devices fulfilling these requirements can be purchased for less than 100 USD. Considering the fact

Metrics	mAP50	Precision	Recall	mAP50-95
K1	0.995	0.998	1.00	0.980
K2	0.995	0.999	1.00	0.983
K3	0.995	0.999	1.00	0.982
K4	0.995	0.999	1.00	0.985
K5	0.995	0.999	1.00	0.975
Average	0.995	0.999	1.00	0.981

Table 3. Audiogram detection performance in 5-Fold Cross-Validation.

Metrics	mAP50	Precision	Recall	mAP50-95
K1	0.995	0.989	1.00	0.681
K2	0.987	0.987	0.983	0.627
K3	0.995	0.995	1.00	0.637
K4	0.949	0.955	0.936	0.622
K5	0.995	0.957	0.971	0.661
Average	0.984	0.977	0.978	0.646

Table 4. Symbol detection performance in 5-Fold Cross-Validation.

Metrics	mAP50	Precision	Recall	mAP50-95
K1	0.995	0.997	1.00	0.681
K2	0.993	0.995	0.997	0.627
K3	0.995	0.994	0.999	0.637
K4	0.995	0.995	0.998	0.622
K5	0.992	0.996	0.998	0.661
Average	0.994	0.995	0.998	0.669

Table 5. Label detection performance in 5-Fold Cross-Validation.

that pre-owned devices are even more affordable, the hardware requirements for running the application should make it viable for use even in low-income areas.

Results

The evaluation of each independent YOLO-based object detector was conducted using a 5-fold cross-validation approach to guarantee reliable performance metrics. The calculated metrics included mean Average Precision at an IoU threshold of 0.50 (mAP50), as well as precision, recall and mAP across a range of IoU thresholds from 0.50 to 0.95 (mAP50-95). The comprehensive findings of these assessments of the audiogram, symbol and line detection models are outlined in Tables 3, 4 and 5, respectively.

Table 6. The results of testing the mobile application on different devices under varying lighting conditions. Values in the three right-most columns represent the number of audiogram lines which have not been detected and thus needed to be interpolated.

Figure 6 illustrates a sample product of the implemented line detection system on an audiogram photographed by the Motorola Moto G82 5G under 500 lx conditions. The audiogram represents mixed hearing loss, and its complexity is considered as "Simple". The lines identified using the Hough method are represented by white colour, while the approximated lines are highlighted in purple.

Discussion

Performance of audiogram extraction and digitalization models

The task of audiogram detection has been shown to be significantly less complex compared to the identification of symbols and labels. The proposed model achieved 99% mAP5, 99% precision, 100% recall and 98% of mAP50-95 in averaged results for this task (Table 3). When juxtaposed with the findings of Chairh and Green[8], who utilized a 3-fold cross-validation approach, the proposed model exhibited a notable enhancement in mAP50, improving by 15% points. At the same time, the precision and recall metrics remained consistent. In comparison, the obtained results were slightly lower than that of Yang et al.[9], who demonstrated an impressive 100% accuracy, precision, and recall in their audiogram detection model. It should be noted, however, that Yang et al. did not provide their mAP50 values and chose not to use K-fold cross-validation, making it challenging to conduct direct comparisons. Therefore, although both the presented model as well as that of Yang et al.[9] show very high

			Motorola Moto G82 5G	Samsung Galaxy S21 5G	Samsung Galaxy S23 Ultra
Normal hearing	Simple	50 lx	5	4	1
		500 lx	0	0	0
	Average	50 lx	0	1	1
		500 lx	0	0	0
	Complex	50 lx	9	2	0
		500 lx	3	0	1
Conductive hearing loss	Simple	50 lx	4	7	3
		500 lx	0	8	0
	Average	50 lx	4	0	2
		500 lx	0	1	1
	Complex	50 lx	7	0	0
		500 lx	0	0	0
Mixed hearing loss	Simple	50 lx	2	4	1
		500 lx	3	0	0
	Average	50 lx	7	0	1
		500 lx	4	0	0
	Complex	50 lx	12	11	2
		500 lx	5	0	0
Sensorineural hearing loss	Simple	50 lx	7	0	1
		500 lx	2	1	0
	Average	50 lx	1	0	1
		500 lx	1	0	0
	Complex	50 lx	8	0	5
		500 lx	0	1	0
Sum in 50 lx			66	29	18
Sum in 500 lx			18	11	2
Sum			84	40	20

Table 6. Presents the results obtained from testing the application on different smartphones under varied lighting conditions, with the main objective being to assess the effectiveness of the line detection method. In instances where specific lines were not correctly identified, the analysis was broadened to evaluate the accuracy of interpolating those lines. The numerical values displayed in the table correspond to the number of lines that have not been properly detected, and thus were subject to approximation.

effectiveness, the lack of comparable metrics and variations in methodology hinder the ability to determine the superior model.

When considering the symbol detection model, the performance metrics are also notable, attaining mAP50 of 98% while sustaining precision and recall at 97% (Table 4). When assessed using the more rigorous criterion of mAP50-95, the model's performance was markedly weaker, resulting in a score of 65%. This discrepancy indicates that the model performs well in simpler detection scenarios but faces difficulties in more complex instances of symbol recognition. Analysis of confusion matrices revealed that the lowest performance scores were linked to the masked air conduction symbols, denoted by square and triangle shapes. In contrast, the predicted classifications for all remaining symbols consistently attained scores of 90% or above. This notable discrepancy is primarily due to the infrequent presence of masked air conduction symbols in the dataset. This issue is a product of a broader problem, where masked air conduction symbols are significantly less prevalent in tonal audiometry test results than their counterparts. These results show that the dataset augmentation comprising an additional 32 audiograms with masked symbols has proven inadequate in addressing this imbalance. When comparing the obtained results to those of other symbol detection models, the study by Chairh and Green[8] reveals a score of only 39% mAP@50. This outcome is notably inferior to the performance metrics recorded in the presented model, especially when taking into account the more stringent mAP across the spectrum of 50 to 95 (mAP50-95), which is still higher than that of Chairh and Green[8]. It should be noted, however, that Chairh and Green's study also included the examination of handwritten audiograms, an area where the presented model has not been specifically evaluated. When comparing to the results presented by Yang et al.[9], the demonstrated level of accuracy aligns closely with that of the presented model, reaching 98.11%. At the same time, it should be noted that the authors did not clearly state the value of the mAP@50 metric in their findings and also did not implement k-fold cross-validation, which brings into question the robustness and generalizability of their results.

As far as the label detection model is concerned, the findings demonstrate a slight improvement in its effectiveness in relation to the symbol detection model. The presented label detection model attained a mAP@50 of 99%, with nearly flawless precision and recall in the averaged results, as outlined in Table 5. The comparative

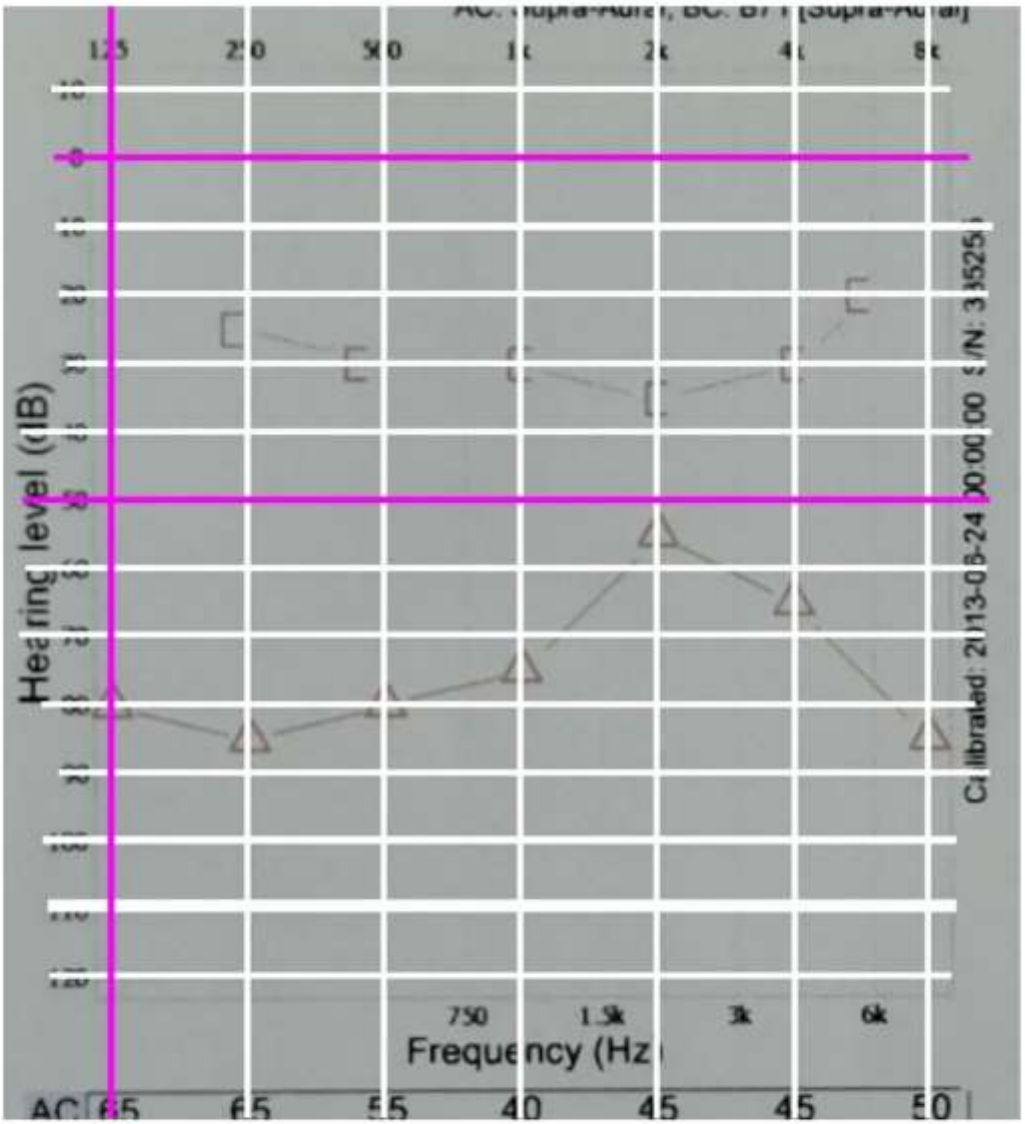


Fig. 6. Example of line detection in an audiogram (mixed hearing loss, 500 lx, Simple, Motorola Moto G82 5G). The lines identified using the Hough method are white, while interpolated lines are indicated in purple. Image represents data produced by the presented smartphone application.

analysis reveals that the performance metrics of the proposed model greatly surpass those of Chairh and Green[8], whose model achieves only a 34% mAP@50. Furthermore, the investigation carried out by Yang et al.[9] departs from the training of a YOLO-based model, concentrating solely on an OCR system that attains an accuracy close to 99%. Although this level of accuracy is very good and aligns with the 99% mAP@50 of the presented model, it should be noted that the presented application does not rely solely on this model and instead integrates its results with the output of OCR for a more comprehensive label detection.

Data interpolation methods and general application performance

The results presented in Table 6 demonstrate that the requirement for line interpolation was significantly higher under 50 lx lighting conditions than under 500 lx for all devices analyzed, which was to be expected. A more interesting observation is that the number of unrecognized lines was relatively high on certain devices even in good lighting conditions. The Motorola Moto G82 5G, which should be considered an antry-level smartphone at this point, exhibited greatest sensitivity to different light conditions. Under 50 lx it reached 66 approximated lines, while at 500 lx this figure decreased to a still significant 18 lines. Comparative analysis of photographs uncovered notable differences in low-light image quality between this device and its more expensive counterparts. The images taken with the Motorola exhibited a noticeably lower contrast and increased noise compared to those produced by both of the Samsung devices under 50 lx. This can be attributed to the fact that both the Samsung S21 and Samsung S23 Ultra possess advanced image processing algorithms that improve the clarity of photographs captured in low-light conditions. Aside from lighting conditions, another important factor influencing line recognition was perspective distortion of the captured image. In particular, optimal alignment of the lens directly towards the document was observed to correlate with a reduction in the number of required line approximations. Another interesting discovery which pertains to on-device image processing algorithms is that the Samsung Galaxy S23 Ultra applied a sharpening filter over its photographs, which in some cases improved the overall effectiveness of the detection system by generating more defined symbols and lines, but could also hinder the detection process when symbols overlapped. This is particularly visible for the Sensorineural hearing loss class of audiograms, where in several cases the Samsung Galaxy S21 achieved better results than the S23 Ultra under the same lighting conditions, despite having a significantly inferior camera hardware. Furthermore, the testing results indicate that, according to expectations, the audiograms categorized as "Complex" presented the greatest challenge to the detection algorithms, particularly in the locations where symbols overlapped with lines. As a result, 46% of all interpolated lines across all devices and conditions come from the "Complex" audiogram category. The most significant outcome of the performed tests is that even in cases where 8 or 11 lines needed to be interpolated, the application still managed to properly classify the audiogram. This proves the effectiveness of the presented solution, in particular regarding the implemented line interpolation algorithms.

In summary, the functionality of all modules comprising the application was reliably showcased across all evaluated devices. In more complex scenarios, the line interpolation functionality became increasingly critical, yet the system maintained a capability for accurate audiogram classification. As anticipated, more expensive devices have shown better performance in the Hough line detection system. Interestingly, this was more due to better image processing algorithms than camera hardware alone, which is exemplified by comparing the results of the Moto G82 and the Galaxy S21. This being said, even on an affordable device like the Motorola Moto G82 5G the application demonstrated commendable performance, proving its usability across a range of devices.

Application of the smartphone app within clinical setting

The application has been designed to simplify the process of analysing hearing test results conducted in audiology and otolaryngology clinics as well as in hearing aid fitting companies, thus allowing general practitioners to classify the results of pure tone audiometry tests. The app can analyze graphical audiometric results (audiograms) and provide automatic interpretations based on hearing threshold outcomes for frequencies spanning from 250 to 8000 Hz. Existing applications with similar functionality, such as HearX[36] and Sonic Sound[39], focus solely on identifying air conduction thresholds, known as the air threshold curve, which can reveal the extent of hearing loss (such as mild, moderate, profound, and severe), and recommending additional diagnostic measures. These tools mainly concentrate on fundamental hearing assessments that individuals can perform at home with standard headphones or speakers. However, they fail to provide professional-grade analysis of audiometric results. This presents an opportunity for the proposed application to fill a significant gap in the market, offering healthcare professionals a sophisticated and reliable means of interpreting audiometric data.

The proposed mobile application aims to address an essential aspect of audiological assessment: the localization and evaluation of hearing damage, differentiating between impairments in the outer and middle ear and dysfunctions in the inner ear and auditory nerve. This differentiation can be achieved through a detailed analysis of two key audiometric curves—air conduction and bone conduction—and their interrelationship. However, in routine clinical practice, such analysis often presents challenges due to the inherent subjectivity involved; different professionals may interpret identical test results differently, leading to inconsistent diagnoses. Margolis and Saly[3], established that among a sample of 231 audiograms assessed by five audiology experts, consensus regarding the type of hearing loss was reached in only 50% of cases. By incorporating precise criteria for classifying types of hearing loss—conductive, sensorineural, and mixed— the application can offer a clear and definitive point of reference. In consequence, the application may enable general practitioners to easily identify a patient's hearing loss type. By dealing with simple cases themselves, general practitioners will be able to not only accelerate the patient's treatment and recovery, but they will also reduce the workload of professional audiologists, who will only receive more complex cases. While experienced otolaryngologists and audiologists will likely not find the application as useful in their everyday practice, it may still offer them the opportunity to get a second opinion on intricate cases, as well as serve as an educational resource. As a result, implementation of this application in clinical practice could considerably alleviate the workload of audiologists by concentrating their efforts on cases that require expedited and specialized intervention, while more straightforward cases can be managed by general practitioners. Moreover, the ability to access supplementary opinions is could also reduce the potential for human error in the diagnostic process.

Finally, the successful development and implementation of the presented application necessitated a collaborative effort among experts in medicine, information technology and regulatory frameworks governing medical devices, ensuring that the tool is both effective and compliant with industry standards. Thus, the application has been meticulously engineered to ensure optimal protection of the processed patient data. First

of all, the application does not collect or process sensitive data such as patient name or age. Secondly, no data is retained on permanent storage. The scanned document as well as extracted data are held exclusively in device RAM for the duration of the analytical process, after which they are irretrievably deleted upon reception of the classification results. Thirdly, the application performs all processing locally and is isolated from internet access, ensuring that patient data remains protected from external entities. Moreover, the application is executed in a kernel-level Application Sandbox, which ensures that its memory and files cannot be accessed by another process (thus providing an additional level of protection e.g. from spying applications). In the above context, the most important issue related to implementing the application in clinical practice is likely the need to inform the patient that their medical data may be analysed by an artificial intelligence system, for which the patient should provide explicit consent.

Limitations

While the presented application implements state-of-the-art AI models to provide freely available and accurate tool for assisting hearing loss diagnosis, it also comes with some inherent limitations. In particular, the application has been designed to operate with certain types of electronically generated audiograms, thus excluding any that are presented in handwritten formats. Additionally, the examination is limited to the outcomes illustrated in two separate audiograms—one corresponding to each ear—rather than considering situations where information from both ears is combined into a single audiogram. This being said, the application has been designed in a modular fashion, so that any new functionalities could be added as additional modules slotting between or in parallel to existing ones. By making this project Open Source, we would like to encourage its further development towards the incorporation of diverse audiogram types. To achieve this goal, it is vital to obtain a broader range of audiogram datasets, which will be key for fine-tuning the YOLO model employed for audiogram and symbol detection.

Conclusion

This study presents a mobile application designed for comprehensive classification of hearing loss types using audiograms captured through a smartphone camera on the Android operating system. The application implements state-of-the-art methods for scanning, digitalization and classification of audiograms. Scanning is achieved with a YOLOv5 AI model tuned to identify audiograms on photographed hearing test reports with 98% accuracy. The digitalization step involves the use of YOLOv5 models, OCR and the Hough Transform method for the detection of symbols, labels, and lines with over 98% accuracy. The classification step employs a Bi-LSTM model which categorizes digitalized audiograms into one of four distinct classes: normal hearing, conductive hearing loss, mixed hearing loss and sensorineural hearing loss with 99% accuracy. All the employed AI models have been specifically optimized and adapted for operation on mobile devices. 5-fold cross validation has been employed to verify that the created AI models yield results that are comparable or better to those presented in literature. Furthermore, the label detection phase uses an innovative integration of the YOLOv5 model and OCR.

The performance of the application has been evaluated across three distinct devices, under two varying lighting conditions, and with different levels of audiogram complexity. The results indicate that, thanks to the implemented data interpolation methods, the application performs well even on a relatively low-end device and under unfavourable lighting.

Current limitations of the application include the type of processed audiogram, which excludes those drawn by hand and/or displaying results from both ears on a single plot, as well as the supported operating systems, which are limited to Android version 9 and later. We hope to overcome these limitations opening the source code of the application to the public. This way researchers with access to different audiogram types will be able to adapt the detection networks to their data, and iOS users should be able to transfer all of the created AI models to their platform with relative ease.

Overall, the presented application has the potential to serve as an approachable and comprehensive diagnosis support system for doctors in clinical practice. By enabling the classification of pure tone audiometry test results by general practitioners, it may decrease the number of uncomplicated cases that are subsequently referred to an audiologist. Moreover, the mobile AI decision support system may further benefit audiologists and otolaryngologists by reducing their workload, improving diagnostic precision and decreasing the likelihood of human error.

Further work may involve a more precise classification of test results, which would include the probability of particular hearing disorders (e.g. otitis media, otosclerosis, noise-induced hearing loss, Ménière's disease, acoustic schwannoma, etc.). This way the application could generate suggestions for additional testing, advanced surgical interventions or the possible use of hearing aids, thus further increasing the efficiency of medical practitioners. Moreover, the training dataset for audiogram digitalization could be expanded with more instances of masked air conduction symbols.

Data availability

The datasets analysed during the current study are not publicly available due to the confidentiality restrictions imposed by the approved ethics of study but are available from the corresponding author on reasonable request.

Code availability

All code written in support of this publication is publicly available at https://github.com/michal-kass/AudiogramScan.

Received: 22 January 2025; Accepted: 18 April 2025
Published online: 24 April 2025

References

1. World Health Organization. World report on hearing. WHO. Available at: https://www.who.int/publications/i/item/9789240020481 (2021).
2. Kokkonen, J. & Varonen, S. Reliability of primary health care audiograms by non-qualified examiners—An analysis of 1,224 cases. *Otol. Neurotol.* **42**, e261 (2020).
3. Margolis, R. H. & Saly, G. L. Toward a standard description of hearing loss. *Int. J. Audiol.* **46**, 746–758 (2007).
4. Brennan-Jones, C. G., Eikelboom, R. H., Bennett, R. J., Tao, K. F. & Swanepoel, D. W. Asynchronous interpretation of manual and Automated Audiometry: Agreement and reliability. *J. Telemed. Telecare* **24**, 37–43 (2016).
5. Crowson, M. G. et al. AutoAudio: Deep learning for automatic audiogram interpretation. *J. Med. Syst.* https://doi.org/10.1007/s10916-020-01627-1 (2020).
6. Kassjański, M. et al. Automated hearing loss type classification based on pure tone audiometry data. *Sci. Rep.* **14**, 14203. https://doi.org/10.1038/s41598-024-64310-2 (2024).
7. Li, S. et al. Interpreting Audiograms with Multi-stage Neural Networks. https://arxiv.org/abs/2112.09357 (2021).
8. Charih, F. & Green, J. R. Audiogram digitization tool for Audiological Reports. *IEEE Access* **10**(110761), 110761–110769 (2022).
9. Yang, T.-W. et al. A novel method for audiogram digitization in audiological reports. *IEEE Access* **12**, 37862–37872 (2024).
10. Elbaşı, E. & Obali, M. Classification of hearing losses determined through the use of audiometry using data mining. In *Conference: 9th International Conference on Electronics, Computer and Computation* (2012).
11. Watson, H. A., Tribe, R. M. & Shennan, A. H. The role of medical smartphone apps in clinical decision-support: A literature review. *Artif. Intell. Med.* **100**, 101707. https://doi.org/10.1016/j.artmed.2019.101707 (2019).
12. Carter, J., Sandall, J., Shennan, A. H. & Tribe, R. M. Mobile phone apps for clinical decision support in pregnancy: A scoping review. *BMC Med. Inform. Decis. Mak.* **19**, 1–13. https://doi.org/10.1186/s12911-019-0954-1 (2019).
13. Daley, B. J. et al. MHealth apps for gestational diabetes mellitus that provide clinical decision support or Artificial Intelligence: A scoping review. *Diabetic Med.* **39**, e14735. https://doi.org/10.1111/dme.14735 (2021).
14. Dodson, C. H., Baker, E. & Bost, K. Thematic analysis of nurse practitioners use of clinical decision support tools and clinical mobile apps for prescriptive purposes. *J. Am. Assoc. Nurse Pract.* **31**, 522–526. https://doi.org/10.1097/JXX.0000000000000170 (2019).
15. Bourouis, A., Feham, M., Hossain, M. A. & Zhang, L. An intelligent mobile based decision support system for retinal disease diagnosis. *Decis. Support Syst.* **59**, 341–350. https://doi.org/10.1016/j.dss.2014.01.005 (2014).
16. Shreyas, S. K. & Rao, J. V. K. Diagnostic decision support for medical imaging and COVID-19 image classification on Arm Mali GPU. 2021 IEEE Globecom Workshops (GC Wkshps), https://doi.org/10.1109/gcwkshps52748.2021.9682104 (2021).
17. Ventola, C. L. Mobile devices and apps for health care professionals: Uses and benefits. *P & T: Peer-Rev. J. Formul. Manag.* **39**(5), 356–364 (2014).
18. Kanimozhi, P. et al. Revolutionizing hearing health: Mobile-based audiometry assessment enhanced by machine learning integration. In *2024 8th international conference on inventive systems and control (ICISC)* 60–66 https://doi.org/10.1109/icisc62624.2024.00017 (2024).
19. Irace, A. L., Sharma, R. K., Reed, N. S. & Golub, J. S. Smartphone-based applications to detect hearing loss: A review of current technology. *J. Am. Geriatr. Soc.* **69**, 307–316. https://doi.org/10.1111/jgs.16985 (2020).
20. Chen, C.-H. et al. Diagnostic accuracy of smartphone-based audiometry for hearing loss detection: Meta-analysis. *JMIR mHealth uHealth* **9**, e28378. https://doi.org/10.2196/28378 (2021).
21. Yesantharao, L. V., Donahue, M., Smith, A., Yan, H. & Agrawal, Y. Virtual audiometric testing using smartphone mobile applications to detect hearing loss. *Laryngoscope Investigat. Otolaryngol.* **7**, 2002–2010. https://doi.org/10.1002/lio2.928 (2022).
22. Masalski, M. & Kręcicki, T. Self-test web-based pure-tone audiometry: Validity evaluation and measurement error analysis. *J. Med. Internet Research* **15**, e71. https://doi.org/10.2196/jmir.2222 (2013).
23. ML Kit | google for developers. *Google* Available at: https://developers.google.com/ml-kit/. (Accessed: 16th January 2025).
24. Ultralytics. Ultralytics/yolov5. *GitHub* Available at: https://github.com/ultralytics/yolov5. (Accessed: 16th January 2025).
25. Common objects in context. *COCO* Available at: https://cocodataset.org/#home. (Accessed: 16th January 2025).
26. Kiryati, N., Eldar, Y. & Bruckstein, A. M. A probabilistic hough transform. *Pattern Recogn.* **24**, 303–316. https://doi.org/10.1016/0031-3203(91)90073-E (1991).
27. Illingworth, J. & Kittler, J. A survey of the hough transform. *Comput. Vis. Graph. Image Process.* **44**, 87–116. https://doi.org/10.1016/S0734-189X(88)80033-1 (1988).
28. Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-8**, 679–698. https://doi.org/10.1109/TPAMI.1986.4767851 (1986).
29. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 779–788, https://doi.org/10.1109/cvpr.2016.91 (2016).
30. Dluzhevskij, D., Stefanovič, P. & Ramanauskaitė, S. Investigation of yolov5 efficiency in iPhone supported systems. *Baltic J. Modern Comput.* 9, https://doi.org/10.22364/bjmc.2021.9.3.07 (2021).
31. Torch.utils.mobile_optimizer. *torch.utils.mobile_optimizer - PyTorch 2.5 documentation* Available at: https://pytorch.org/docs/stable/mobile_optimizer.html. (Accessed: 16th January 2025).
32. Berrar, D. Cross-validation. In *Encyclopedia of Bioinformatics and Computational Biology* 542–545, https://doi.org/10.1016/b978-0-12-809633-8.20349-x (2019).
33. Motorola. moto G82 5G. *motorola* Available at: https://www.motorola.com/we/smartphones-moto-g-82-5g/p?skuId=450. (Accessed: 16th January 2025).
34. Specs: S21 5G. *Samsung uk* (2021). Available at: https://www.samsung.com/uk/smartphones/galaxy-s21-ultra-5g/specs/. (Accessed: 16th January 2025).
35. Specs: Samsung Galaxy S23 Ultra: Samsung Caribbean. *Samsung latin_en* (2023). Available at: https://www.samsung.com/latin_en/smartphones/galaxy-s23-ultra/specs/. (Accessed: 16th January 2025).
36. *National Optical Astronomy Observatory: Recommended Light Levels (Illuminance) for Outdoor and Indoor Venues* Available at: https://web.archive.org/web/20210706034730/https://www.noao.edu/education/QLTkit/ACTIVITY_Documents/Safety/LightLevels_outdoor+indoor.pdf. (Accessed: 16th January 2025).
37. Target API level requirements for google play apps - play console help. Google Available at: https://support.google.com/googleplay/android-developer/answer/11926878?hl=en. (Accessed: 10th March 2025)
38. Hearwho. *World Health Organization* Available at: https://www.who.int/teams/noncommunicable-diseases/sensory-functions-disability-and-rehabilitation/hearwho. (Accessed: 16th January 2025).
39. Sonitum. SonicCloud sound check. *App Store* (2023). Available at: https://apps.apple.com/us/app/soniccloud-sound-check/id1478996953. (Accessed: 16th January 2025).

Acknowledgements
Computational resources for network training were provided by the Centre of Informatics - Tricity Academic Supercomputer & Network (CI TASK) under grant No. PT01167.

Author contributions
M.K.S., M.K.L., T.P. and D.T. designed the plan and goals of presented research; M.K.S. and M.K.L. performed the initial literature review; M.K.S. and M.K.L. trained and tested the AI-models; M.K.S. and M.K.L. developed and tested the mobile app; T.P., D.T. and A.M. provided medical expertise; M.K.S., M.K.L. and T.P. drafted the initial sections of the manuscript; M.K.S., M.K.L. and T.P. wrote the results, discussion and conclusions; M.K.S., M.K.L. prepared the final version of the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to M.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.