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.

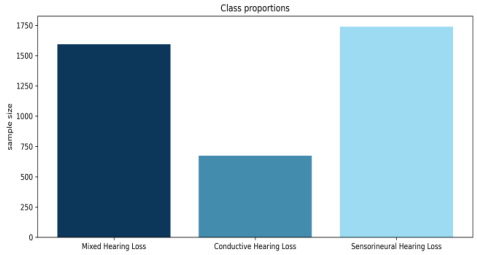

Figure 1. The class proportions in the input dataset

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:

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

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

$$Z_{\text{minmax}} = \frac{x - min}{max - min} \tag{2}$$

$$Z_{maxabs} = \frac{x}{|max|} \tag{3}$$

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.

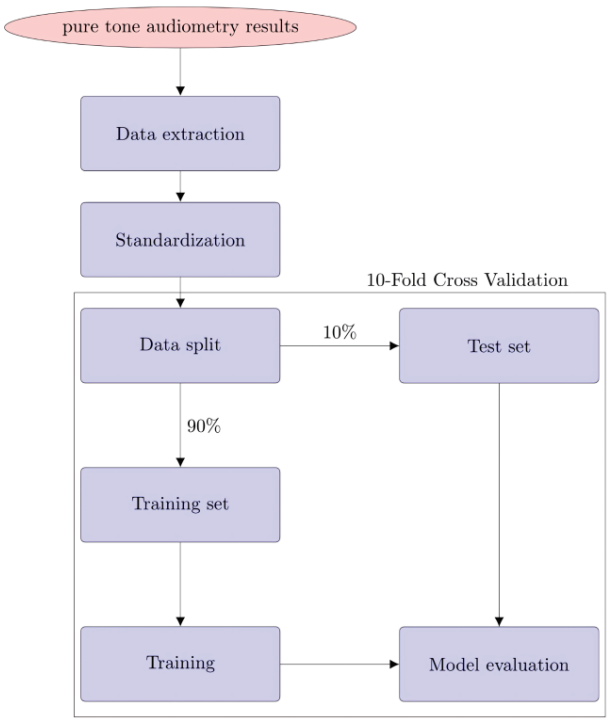

Figure 2. The workflow of the presented research into application of machine learning methods for the classification of hearing loss types based on pure-tone audiometry data

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.

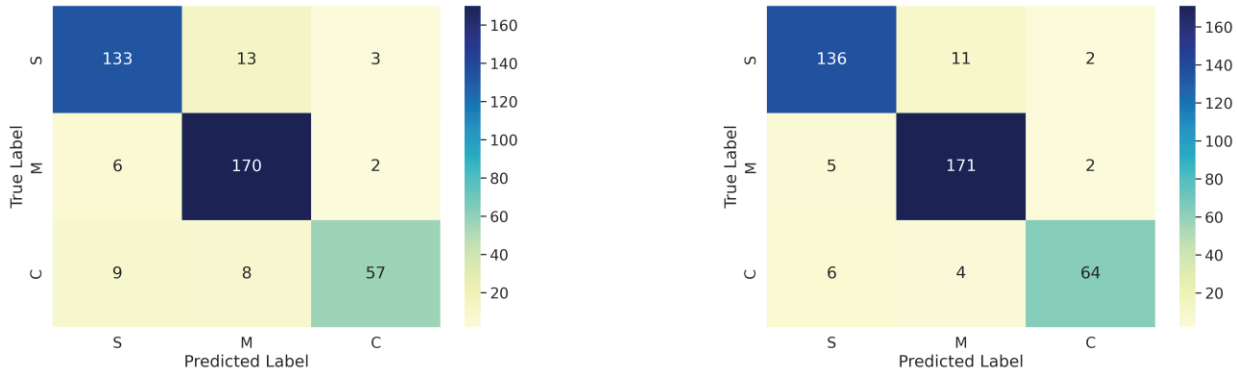

Figure 3. Confusion matrices of the logistic regression model trained without CGAN (left) and with CGAN (right)

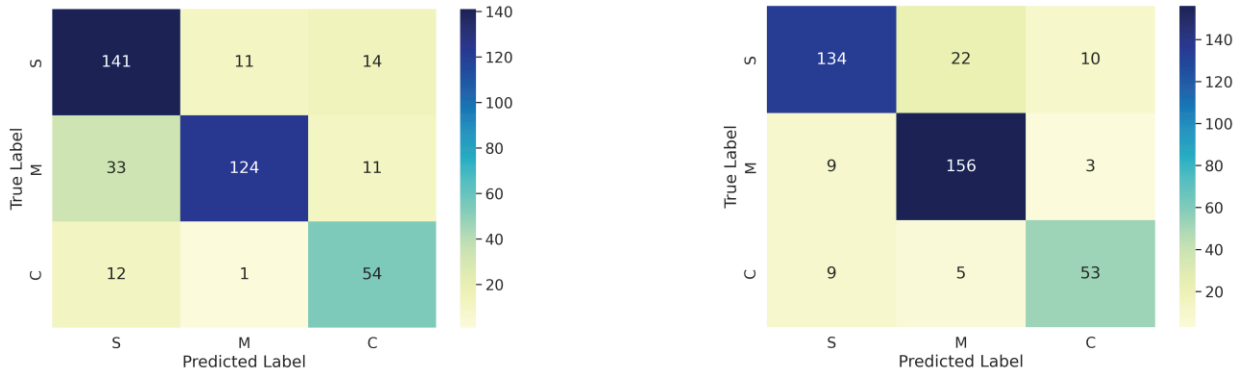

Figure 4. Confusion matrices of the stochastic gradient descent model trained without CGAN (left) and with CGAN (right)

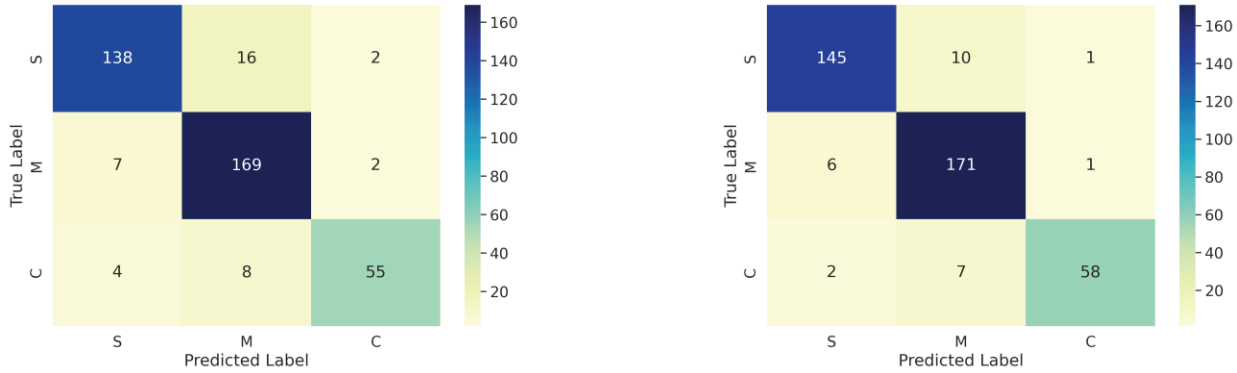

Figure 5. Confusion matrices of the support vector machines model trained without CGAN (left) and with CGAN (right)

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

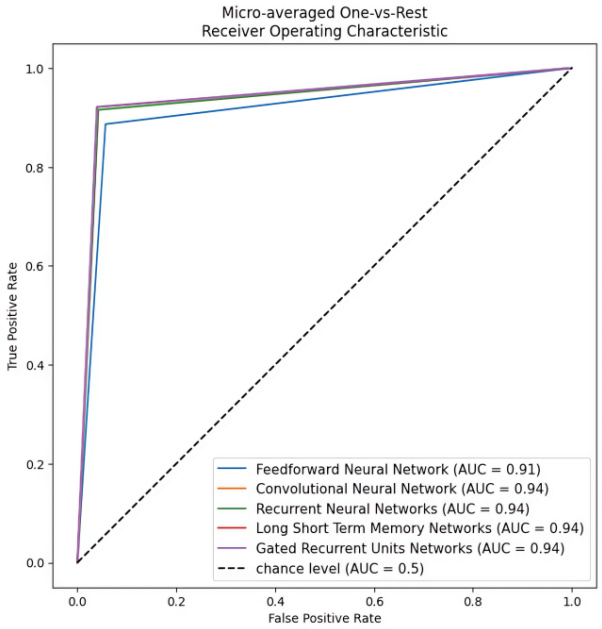

Figure 6. ROC curves with the AUC parameters for tested deep learning models during 10-Fold validation

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.