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 Koźmiński[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.

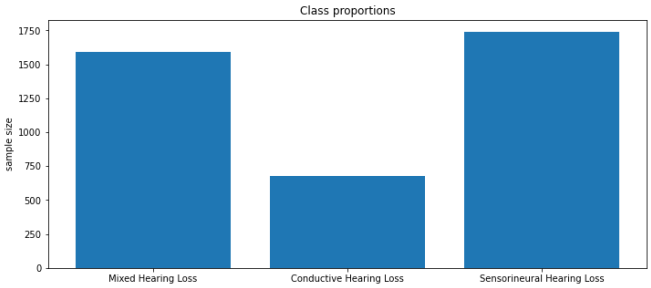

Figure 1. The three forms of hearing loss represented in the dataset, along with their respective proportions.

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.

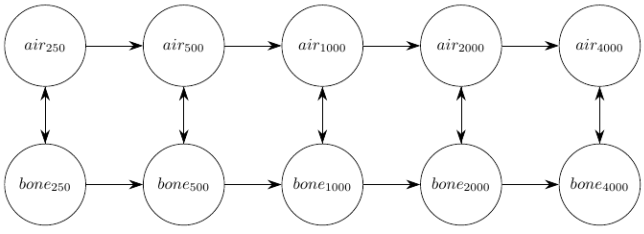

Figure 2. The GNN architecture's input graph structure.

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].

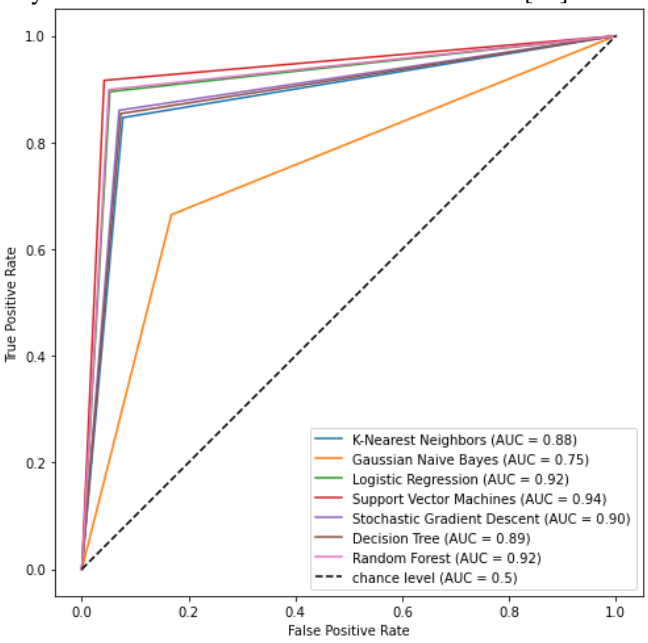

Figure 3. ROC curves with the AUC parameters for machine learning models.

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%)**

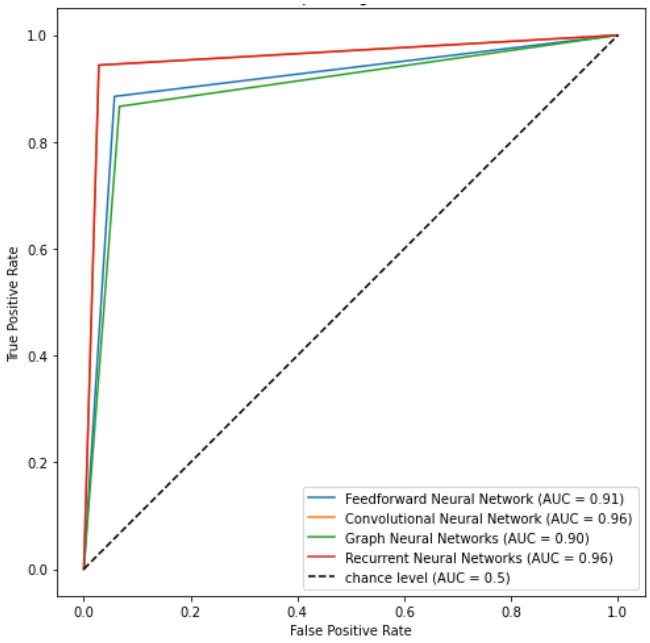

Figure 4. ROC curves with the AUC parameters for deep learning models.

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] Elbaşı, 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.