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 Elbaşı & 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.

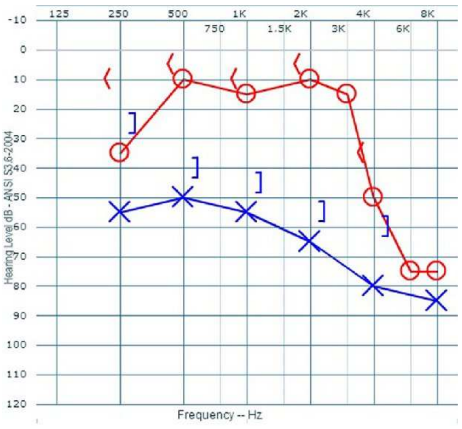

Fig. 1. A pure tone audiogram showing air and bone conduction thresholds for **both left and right ear** [8]. The "X" and "]" symbols are used to mark left-sided air and bone conduction, respectively. The "O" indicate air conduction, whereas the "<" denote bone conduction, both in the right ear.

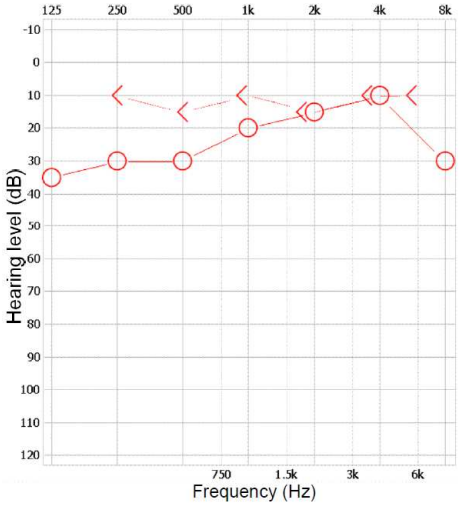

Fig. 2. A pure tone audiogram showing air and bone conduction thresholds **only for the right ear.** The "O" and "<" indicate left-sided air and bone conduction, respectively.

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.

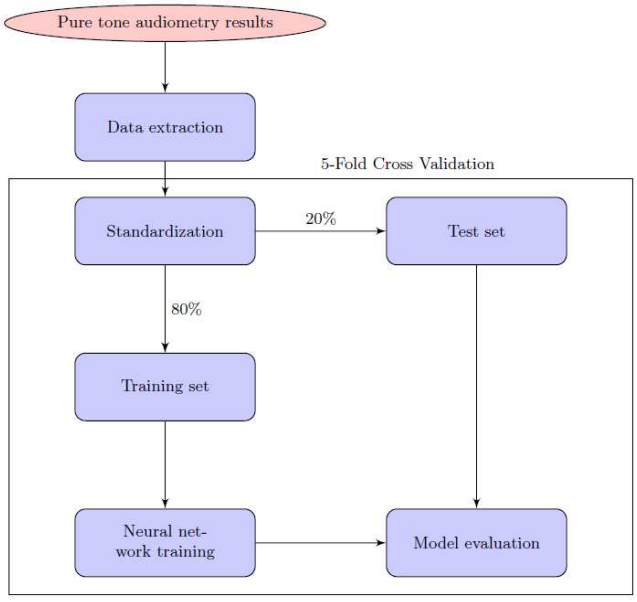

Fig. 3. Workflow of processes leading to model evaluation.

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

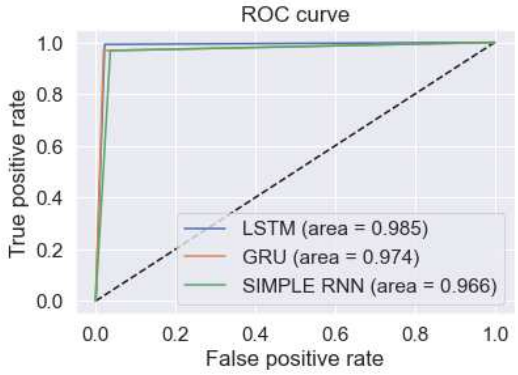

Fig. 4. ROC curve with AUC parameter of RNN models.

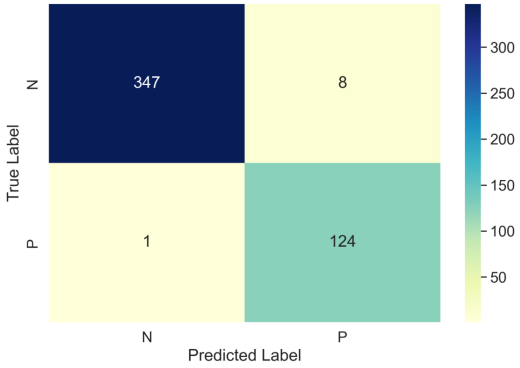

Fig. 7. Confusion matrix of LSTM.

- 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.

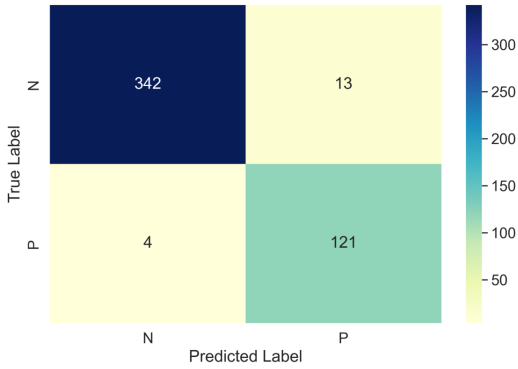

Fig. 5. Confusion matrix of simple RNN.

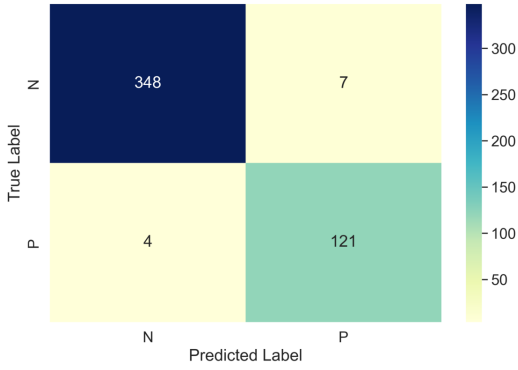

Fig. 6. Confusion matrix of GRU.

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. Koźmiński, 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 Bhaudaria, 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