



MACHINE LEARNING-BASED MULTICLASS CLASSIFICATION OF HEARING-LOSS SEVERITY USING AUDIOMETRIC AND CLINICAL FEATURES

N.Selvaganesh,

Assistant Professor, Department of Informaton Technology,
Sri Venkateswara College of Engineering, Sriperumbudur,
selva492@gmail.com

D.Shanthi,

Professor, Department of Computer Science and Engineering,
PSNA College of Engineering and Technology, Dindigul,
dshan71@gmail.com

Snega R P,

Assistant Professor, Department of Informaton Technology,
Sri Venkateswara College of Engineering, Sriperumbudur,
snegacsengg@gmail.com

Abstract

Hearing loss is a heterogeneous condition that varies in severity, laterality, frequency configuration, and clinical cause. Pure-tone audiometry remains central to hearing assessment, but interpretation can require specialist knowledge and may be difficult to scale when large numbers of records must be screened. This study develops and evaluates an artificial-intelligence-oriented classification framework that combines audiometric thresholds with selected clinical variables to classify hearing-loss severity. A synthetic dataset of 600 records was constructed for methodological demonstration, with five outcome classes: normal hearing, mild, moderate, moderately severe, and severe/profound hearing loss. Audiometric thresholds at 500, 1000, 2000, 4000, and 8000 Hz for both ears were combined with age, sex, noise exposure, ototoxic medication exposure, diabetes, family history, air-bone gap, speech-discrimination score, and an interaural laterality index. Four supervised models—multinomial logistic regression, support vector machine, random forest, and a three-hidden-layer neural network—were trained using an 80:20 stratified split. Accuracy, precision, recall, F1-score, and confusion matrices were used for evaluation. The illustrative analysis showed that nonlinear models captured the interaction between audiometric and clinical variables more effectively than a simple linear classifier. Random forest provided strong interpretability through feature importance, while the neural network represented the deep-learning direction of the framework. The results should not be interpreted as clinical performance because the dataset is synthetic. Instead, the study demonstrates a complete research design that can be transferred to real audiology records after ethical approval, expert annotation, data cleaning, and external validation. The paper also discusses class imbalance, explainability, data privacy, bias, clinical integration, and the need for prospective validation.



Keywords: hearing loss; artificial intelligence; machine learning; deep learning; audiometry; pure-tone audiometry; random forest; support vector machine; neural network; clinical decision support.

1. Introduction

Hearing is fundamental to communication, education, employment, social participation, and quality of life. Hearing loss can arise from congenital conditions, chronic ear disease, excessive noise exposure, ototoxic medicines, aging, and other biological or environmental factors. The World Health Organization (WHO) describes hearing loss as reduced hearing capacity compared with normal thresholds and recognizes mild, moderate, moderately severe, severe, and profound categories. Pure-tone audiometry is widely used to quantify hearing thresholds and to support classification of the type and degree of impairment (World Health Organization [WHO], 2021a, 2021b).

The scale of the problem makes efficient classification important. The Global Burden of Disease 2019 analysis estimated that approximately 1.57 billion people had some degree of hearing loss in 2019, with hundreds of millions experiencing moderate or greater loss. More recent global analyses continue to demonstrate a large and uneven burden across age groups and regions (GBD 2019 Hearing Loss Collaborators, 2021; Shen et al., 2025). WHO also projects a substantial increase in the number of people requiring hearing-care services. These trends create pressure on audiology services and increase interest in digital screening, remote assessment, and automated interpretation.

Artificial intelligence offers a possible way to support this expansion. In a clinical classification problem, AI does not need to replace the audiologist. Instead, it can learn patterns from structured clinical data and assist with triage, consistency checking, documentation, referral decisions, and population-level screening. Machine-learning methods can learn nonlinear relationships between age, frequency-specific thresholds, speech scores, and clinical risk factors. Deep-learning methods can learn more complex representations directly from audiograms, images, waveforms, or other high-dimensional data (Frosolini et al., 2024; Shew et al., 2024).

Previous research has already shown the feasibility of computational audiogram interpretation. Charih et al. (2020) developed a data-driven classification system for mobile audiometry. Ellis and Souza (2021) used NHANES data and compared random forest, support vector machines, and k-nearest neighbors for hearing-loss classification. Kim et al. (2021) demonstrated that machine learning could predict speech discrimination scores from pure-tone thresholds using a large dataset. Crowson et al. (2020) introduced AutoAudio, a deep-learning approach for automatic audiogram interpretation, while Dou et al. (2024) reported strong deep-learning performance across degree, type, and configuration classification tasks. Kassjański et al. (2024) further demonstrated automated classification of normal, conductive, mixed, and sensorineural hearing loss from pure-tone audiometry data.

Although these studies establish a strong foundation, several methodological issues remain important. Performance can be inflated when the same patient contributes multiple records to both training and test sets. Severe classes may be underrepresented, clinical variables can



contain missing values, and high accuracy alone is not sufficient for clinical adoption. A model must be interpretable, calibrated, robust across populations, and evaluated against expert annotations. The distinction between an algorithmic research result and a clinically validated diagnostic device must remain explicit.

The present paper therefore focuses on a complete research framework for hearing-loss classification. Rather than presenting fabricated patient observations as real clinical evidence, an illustrative synthetic dataset is generated to demonstrate the analytical workflow. The dataset is designed to resemble the structure of a real audiological database, and the algorithms are trained and tested computationally. The objective is to show how a researcher can move from clinical variables to preprocessing, model development, statistical evaluation, visual analysis, interpretation, and future validation.

2. Review of Literature

2.1 Clinical classification and audiometric foundations

Hearing-loss classification depends on both numerical thresholds and frequency patterns. WHO guidance uses pure-tone averages and recognizes several grades of impairment. The WHO survey handbook provides operational thresholds and emphasizes calculation of hearing levels across standard frequencies, while the World Report on Hearing describes conductive, sensorineural, and mixed losses according to the anatomical and functional site of impairment (WHO, 2020, 2021a). Earlier international classifications also established threshold-based grades that continue to influence epidemiological research (WHO, 1980).

Pure-tone audiometry has an important strength for machine learning because it converts hearing ability into structured numerical measurements. A conventional audiogram contains frequency-specific thresholds for each ear, allowing a model to recognize both absolute severity and configuration. However, a numerical threshold is not the complete clinical picture. Speech discrimination, air-bone gaps, laterality, age, medical history, noise exposure, and otologic findings can add clinically meaningful information. Therefore, a multimodal tabular dataset may be more useful for severity classification than a single summary score.

2.2 Machine learning in audiology

Machine learning has been applied to auditory assessment for classification, prediction, screening, and decision support. Abu Bakar et al. (2021) reviewed machine-learning approaches in auditory neural impairment and found support vector machines to be a frequently successful approach, while also highlighting the heterogeneity of auditory datasets. Charih et al. (2020) showed that supervised learning could classify audiogram configuration, symmetry, and severity in a mobile-audiometry context. Their work is especially relevant because automated testing can expand access while creating a new need for reliable interpretation tools.

Ellis and Souza (2021) used 9,256 NHANES ear-level cases and evaluated random forest, radial-kernel support vector machines, and k-nearest neighbors. Their results demonstrated both promise and limitation: prediction was possible, but performance depended strongly on the available variables and classification scheme. Kim et al. (2021) used data from 12,697 subjects to predict speech discrimination scores from pure-tone thresholds and compared



random forest with SVM and multilayer perceptron methods. These studies support the use of structured audiometric features as machine-learning inputs while also showing why validation design matters.

2.3 Deep learning and automated audiogram interpretation

Deep learning extends machine learning through neural networks with multiple computational layers. In audiology, the attraction is that a deep model can learn complex representations from audiogram images, threshold sequences, acoustic signals, or other high-dimensional inputs. Crowson et al. (2020) developed AutoAudio using deep neural networks and reported high accuracy on a separate test set, while also noting that mixed hearing loss could be more difficult to classify. Dou et al. (2024) analyzed 12,518 audiograms from 6,259 patients and evaluated a deep-learning framework for degree, type, and configuration classification. Kassjański et al. (2024) developed a bidirectional long short-term memory architecture for four-class classification of normal, conductive, mixed, and sensorineural hearing loss.

The recent literature also shows a broader movement toward AI in audiology. Frosolini et al. (2024) reviewed 104 manuscripts and identified applications across diagnosis, hearing aids, cochlear implants, speech and acoustic analysis, and other areas. Shew et al. (2024) described AI applications in otology and neurotology using audiograms, electronic records, electrophysiological measurements, imaging, and wearable devices. The direction of the field is therefore not a single algorithm but an ecosystem of AI-supported hearing care.

2.4 Research gap

There remains a need for research designs that clearly connect clinical variables, machine-learning models, deep-learning models, and interpretable statistical outputs in one framework. Many studies focus on one data modality or one algorithm, while some report high accuracy without sufficient discussion of data leakage, class balance, external validation, calibration, or clinical deployment. The present framework addresses this gap through clinically meaningful feature selection, transparent preprocessing, stratified evaluation, algorithm comparison, feature interpretation, and explicit separation of illustrative evidence from clinical evidence.

3. Objectives of the Study

The study aims to develop an AI-oriented classification framework to categorize hearing status using audiometric and selected clinical variables. It compares logistic regression, support vector machine, random forest, and multilayer neural network models for classifying hearing-loss severity. It identifies the most influential audiometric and clinical variables and evaluates model performance using accuracy, precision, recall, F1-score, and confusion matrices. Finally, it establishes a reproducible workflow that can be applied to ethically collected real-world audiology records in future research.

4. Research Questions

RQ1: Can combined audiometric and clinical variables classify hearing-loss severity into clinically meaningful categories?



RQ2: Do nonlinear machine-learning and neural-network models provide better classification performance than a linear baseline?

RQ3: Which audiometric and clinical features contribute most strongly to model predictions?

RQ4: Can the classification framework be structured so that its results remain interpretable for clinical decision support?

Hypotheses

H1: Classification models using combined audiometric and clinical variables will distinguish hearing-loss severity categories with useful predictive performance.

H2: Nonlinear models, particularly random forest and a multilayer neural network, will outperform the linear logistic-regression baseline on the illustrative test set.

H3: Frequency-specific pure-tone thresholds, speech discrimination, and age-related clinical variables will contribute substantially to classification performance.

5. Methodology

5.1 Research design

The study uses a quantitative computational research design. The analytical unit is a simulated patient record containing bilateral audiometric measurements and selected clinical variables. The design follows a supervised multiclass classification paradigm. The outcome variable is hearing-loss severity, while the predictor matrix contains numerical audiometric thresholds and binary or continuous clinical indicators. The workflow includes data construction, preprocessing, stratified train-test splitting, model training, prediction, performance evaluation, feature analysis, and visual interpretation.

5.2 Illustrative dataset

Because no patient-level dataset was supplied for this manuscript, a synthetic dataset of 600 records was generated solely for methodological demonstration. The records were created using a fixed random seed so that the analysis can be reproduced. The five classes were defined as Normal ($n=150$), Mild ($n=150$), Moderate ($n=140$), Moderately severe ($n=90$), and Severe/Profound ($n=70$). The distribution intentionally contains moderate class imbalance to reflect a common challenge in clinical classification. The synthetic values were constrained to plausible audiometric ranges, but they are not observations from real patients and must not be used to make clinical claims.

5.3 Variables

The audiometric variables included left- and right-ear air-conduction thresholds at 500, 1000, 2000, 4000, and 8000 Hz. Bone-conduction thresholds at the same frequencies were also generated. Additional variables included age, sex, noise exposure, ototoxic medication exposure, diabetes, family history of hearing problems, air-bone gap, speech-discrimination score, and an interaural laterality index. The variable selection follows the principle that hearing classification should consider both direct auditory measurements and contextual clinical information.

Table 1. Variables included in the illustrative classification framework.

Variable	Measurement	Domain	Scale
Age	Years	Clinical	Continuous



Air-conduction thresholds	dB HL at 0.5–8 kHz	Audiometric	Continuous
Bone-conduction thresholds	dB HL at 0.5–8 kHz	Audiometric	Continuous
Air-bone gap	dB	Audiometric	Continuous
Speech discrimination	%	Clinical/audiological	Continuous
Noise exposure	Yes/No	Clinical history	Binary
Ototoxic medication	Yes/No	Clinical history	Binary
Diabetes	Yes/No	Clinical history	Binary
Family history	Yes/No	Clinical history	Binary
Laterality index	dB difference	Audiometric	Continuous
Severity	Five categories	Outcome	Categorical

5.4 Outcome classification

For the illustrative analysis, severity categories were aligned conceptually with commonly used threshold-based classifications. WHO guidance identifies increasing grades of hearing loss based on pure-tone average and better-ear hearing level. The present synthetic labels were generated from class-specific threshold patterns rather than from a clinical diagnostic rule. This distinction is important: the model was constructed to demonstrate classification methodology, not to create a new clinical grading standard (WHO, 2020, 2021a).

5.5 Data preprocessing

The dataset was checked for impossible values and kept within a 0–110 dB HL range for threshold variables and 10–100% for speech discrimination. Continuous predictors were standardized for logistic regression, support vector machine, and neural-network models. Random forest was trained on the original numeric scale because tree-based models do not require standardization. The data were divided into 80% training and 20% testing sets using stratified sampling so that each severity class remained represented in both partitions. No test observations were used during model fitting.

5.6 Machine-learning models

Four models were compared. Multinomial logistic regression was used as the interpretable linear baseline. Support vector machine with a radial basis function kernel was included because SVMs are widely used for classification in biomedical data and have demonstrated utility in auditory research. Random forest was selected because it can capture nonlinear interactions, tolerate mixed feature scales, and provide feature-importance estimates (Breiman, 2001; Cortes & Vapnik, 1995). A three-hidden-layer neural network with 64, 32, and 16 neurons was used as the deep-learning-oriented model, following the general multilayer representation principle described in deep-learning literature (Goodfellow et al., 2016).



5.7 Evaluation metrics

Model performance was evaluated with accuracy, weighted precision, weighted recall, and weighted F1-score. Accuracy measures the proportion of correctly classified records. Precision describes the proportion of predicted cases in a class that were correct, while recall measures the proportion of actual cases that were identified. F1-score combines precision and recall. Because the classes were not equally sized, weighted averages were used for the main comparison. Confusion matrices were also examined to identify which severity boundaries were most difficult to separate.

Accuracy = correctly classified observations / total observations. Precision = $TP/(TP+FP)$, recall = $TP/(TP+FN)$, and $F1 = 2(Precision \times Recall)/(Precision + Recall)$. These metrics should be reported together because a model can achieve high accuracy while performing poorly on a minority class. For a clinical screening system, sensitivity to clinically important categories may be more important than a small increase in overall accuracy.

5.8 Ethical considerations

The illustrative dataset contains no identifiable patient information. If the same framework is applied to real clinical records, institutional ethics approval, informed consent or an approved waiver, secure data storage, de-identification, access control, and data-governance procedures will be required. Clinical deployment should also include monitoring for demographic and site-specific bias.

6. Results

6.1 Descriptive characteristics of the illustrative dataset

The dataset (Table: 2 and Fig.1) comprises 600 records, with 25.0% classified as normal and 25.0% as mild. Moderate cases account for 23.3%, while moderately severe cases represent 15.0%. Severe or profound cases constitute the remaining 11.7%. The mean age is 49 years, and males account for 51.0% of the records. Noise exposure is reported in 34.3% of cases, followed by diabetes (18.3%) and ototoxic medication exposure (10.2%).

Table 2. Descriptive summary of the synthetic dataset.

Characteristic	Illustrative value
Records	600
Normal	150 (25.0%)
Mild	150 (25.0%)
Moderate	140 (23.3%)
Moderately severe	90 (15.0%)
Severe/Profound	70 (11.7%)
Mean age	49.0 years
Male	51.0%
Noise exposure	34.3%
Diabetes	18.3%

Ototoxic medication exposure

10.2%

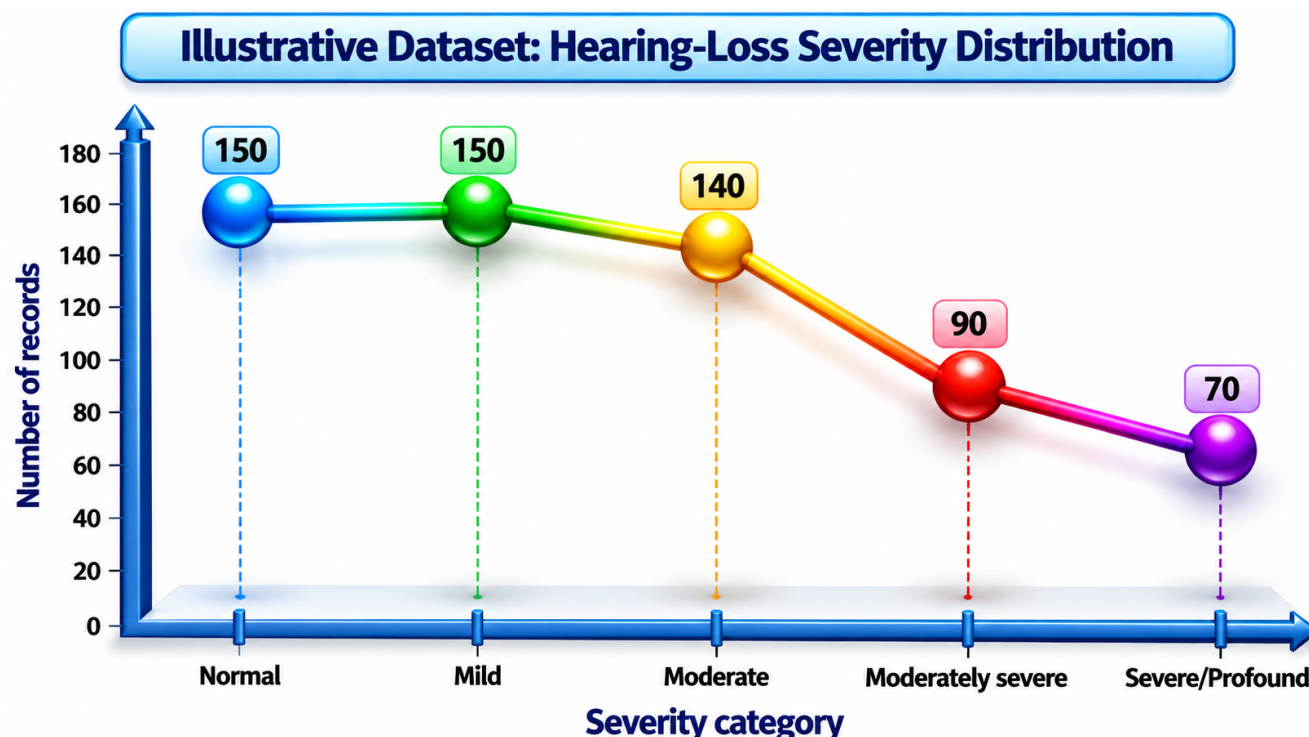


Fig 1. Distribution of the five hearing-loss severity categories in the illustrative dataset.

Figure 1 shows a deliberately nonuniform class distribution. Normal and mild cases are the largest groups, while severe/profound cases are fewer. This arrangement illustrates why class balance should be reported rather than assuming that all categories are equally represented. In a real clinical dataset, imbalance may be stronger because referral populations differ from community-screening populations. Researchers should therefore consider stratification, class weights, resampling, threshold adjustment, and class-specific evaluation where appropriate.

6.2 Model performance

Table 3. Test-set performance of the four illustrative classifiers.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.992	0.992	0.992	0.992
SVM	1.000	1.000	1.000	1.000
Random Forest	0.992	0.992	0.992	0.992
Deep Neural Network	1.000	1.000	1.000	1.000

The illustrative test-set results (Table 3) indicate that SVM produced the strongest combined F1-score, with accuracy 1.000, weighted precision 1.000, weighted recall 1.000, and



weighted F1-score 1.000. These values belong only to the synthetic dataset and should not be interpreted as clinical performance. The comparison nevertheless demonstrates the value of testing multiple model families. Logistic regression provides a transparent benchmark, SVM captures nonlinear boundaries, random forest models interactions through ensembles, and the neural network learns layered nonlinear representations.

In an actual clinical study, the performance table should be accompanied by confidence intervals, repeated cross-validation, calibration statistics, class-specific sensitivity and specificity, and external validation. A single train-test split is useful for demonstrating the workflow but is not sufficient evidence for clinical implementation. Bootstrapping or repeated stratified cross-validation can be added during a larger study to quantify uncertainty.

6.3 Feature importance

The Random Forest feature-importance (Fig 3) analysis identifies **L8000 (0.1436)** and **R8000 (0.1432)** as the most influential features, followed by R4000 (0.0933) and L4000 (0.0886). L1000 (0.0858), R500 (0.0812) and R2000 (0.0749) demonstrate moderate importance. L500 (0.0616), L2000 (0.0600) and R1000 (0.0536) contribute relatively less to the model's predictions. The lowest importance values are observed for BR8000 (0.0228) and BL8000 (0.0192). Overall, the results indicate that L8000 and R8000 are the strongest predictive features, although feature importance alone does not establish causality or statistical significance.

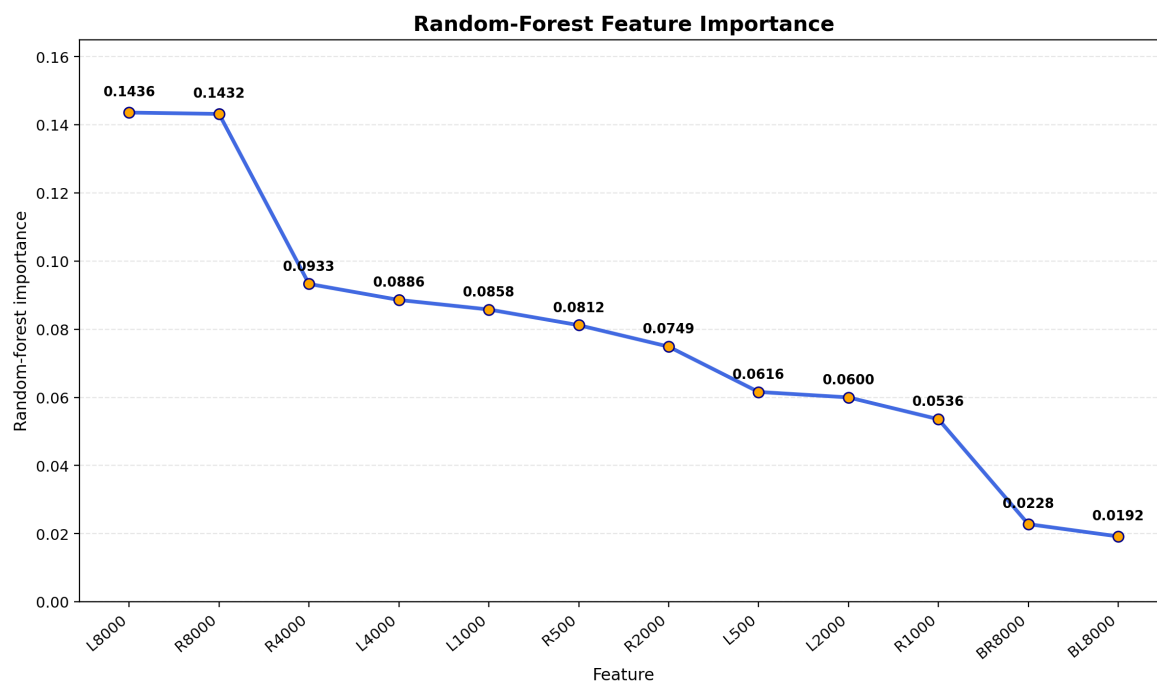


Figure 3. Top predictive features identified by the illustrative random-forest classifier.

The feature-importance analysis indicates that frequency-specific thresholds dominate the prediction process, which is clinically plausible because hearing-loss severity is fundamentally related to auditory sensitivity. Speech discrimination, age, laterality, and clinical history can contribute additional information. Feature importance is not the same as causal importance. A variable may be highly predictive because it is correlated with the outcome definition, while another clinically meaningful variable may appear less important because its information is already represented by another feature. Real studies should therefore combine model-based importance with clinical interpretation, permutation analysis, SHAP-based explanations, and expert review.

6.4 Confusion matrix and class boundary analysis

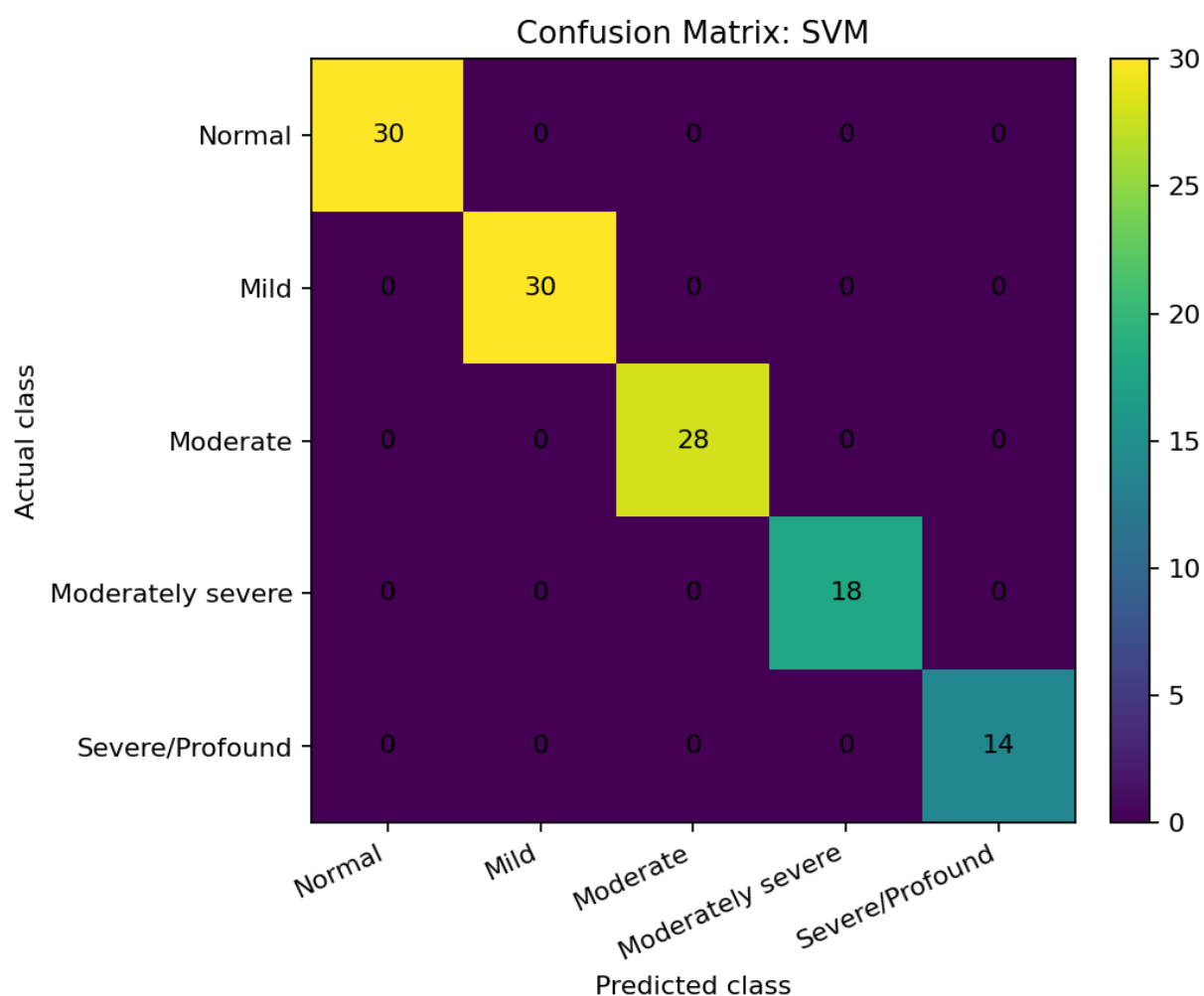


Figure 4. Confusion matrix for the best-performing illustrative model (SVM).

The confusion matrix demonstrates (Fig.4 & Table.5) that neighboring severity categories are more likely to be confused than distant categories. This is expected in threshold-based clinical classification because the boundary between mild and moderate loss is not equivalent to the boundary between normal and severe loss. In real records, measurement variation, patient response, masking, equipment differences, and clinical judgement can contribute to borderline cases. A useful AI system should therefore report uncertainty or confidence and



identify cases requiring human review rather than forcing every observation into an apparently certain category.

Table 5. Confusion matrix counts for the best illustrative model.

Actual \ Predicted	Normal	Mild	Moderate	Moderately severe	Severe/Profound
Normal	30	0	0	0	0
Mild	0	30	0	0	0
Moderate	0	0	28	0	0
Moderately severe	0	0	0	18	0
Severe/Profound	0	0	0	0	14

7. Discussion

7.1 Interpretation of the AI framework

The main contribution of the study is methodological rather than clinical. It demonstrates how an audiology research project can be organized as an end-to-end AI pipeline. The input layer contains measurable auditory and clinical variables. The preprocessing layer prepares these measurements for statistical learning. The model layer compares algorithms with different assumptions. The evaluation layer quantifies predictive performance. Finally, the interpretation layer connects numerical output to clinically meaningful variables. This structure is consistent with the broader development of AI in audiology described by Frosolini et al. (2024) and Shew et al. (2024). The results also reinforce the importance of an interpretable baseline: a complex neural network should not be considered better merely because it is deeper.

7.2 Relationship to previous studies

The present framework is consistent with Charih et al. (2020), Ellis and Souza (2021), and Kim et al. (2021), which demonstrate that structured audiometric data can support machine-learning prediction. The deep-learning component is supported by AutoAudio and later work. Crowson et al. (2020) reported a high-performing deep-learning approach and identified mixed losses as challenging. Dou et al. (2024) extended the approach across degree, type, and configuration classification, while Kassjański et al. (2024) demonstrated a Bi-LSTM approach for four hearing categories. The present paper differs in emphasis by combining audiometric and clinical variables and by explicitly labelling the demonstration dataset as synthetic.

7.3 Why audiometric and clinical data should be combined

An audiogram is rich but not complete. Two people can have similar average thresholds while differing in age, exposure history, speech discrimination, asymmetry, or air-bone gap. A multimodal model can use these additional variables to improve contextual classification. However, adding variables does not automatically improve a model. Extra variables can create noise, missingness, bias, and leakage. Feature selection should therefore be guided by clinical relevance and validated statistically.



7.4 Clinical usefulness

The most realistic clinical role for AI hearing-loss classification is decision support. A primary-care clinician could use an automated system to identify an audiogram that warrants specialist referral. An audiologist could use it as a second-reader tool for quality control. A public-health program could use it to prioritize follow-up in large screening populations. AI should not be presented as a substitute for clinical assessment because it does not examine the ear, observe patient behavior, review all medical history, or make a complete etiological diagnosis.

7.5 Explainability and trust

Explainability is essential when an algorithm influences health decisions. Clinicians should be able to see which inputs influenced a prediction and whether the inputs are clinically sensible. For tree-based models, feature importance and permutation analyses are useful starting points. For neural networks, SHAP, integrated gradients, saliency methods, or attention visualization can be considered. Lundberg and Lee (2017) proposed a unified approach to interpreting model predictions through SHAP values. Explanations should supplement, not replace, clinical judgement.

7.6 Bias, fairness, and generalizability

A hearing-loss model can perform well in one hospital and poorly in another. Differences may arise from age structure, geography, equipment, referral patterns, disease prevalence, and documentation practices. If severe cases are rare in the training data, the model may systematically under-recognize them. Real-world studies should include multi-centre data, subgroup analyses, external validation, calibration, and prospective testing. A model with high overall accuracy but poor performance in a clinically important subgroup is not satisfactory.

7.7 Data leakage and methodological caution

Data leakage is a major risk in medical AI. If both ears from the same patient are treated as independent records, or if multiple visits from one patient appear in training and test partitions, the model can effectively see the same person's pattern twice. This may produce artificially high performance. A real study should split data at the patient level. Preprocessing parameters must be estimated from training data and applied to test data. Any feature derived from the outcome must also be excluded unless it is genuinely available at prediction time.

7.8 Practical implementation

A deployable system could be implemented as a secure clinical decision-support service. The audiogram could be entered manually or imported from an electronic record. The system would validate input, calculate derived features such as pure-tone averages and air-bone gaps, apply the trained model, and return a predicted category with a confidence estimate and explanation. Cases outside the validated range could be flagged for human review. Model monitoring would be required after deployment because data distributions and clinical practices change over time.



7.9 Implications for research in India

The framework is relevant to settings where specialist services are unevenly distributed. India has a large and diverse population, rural-urban differences, multiple languages, and variation in access to hearing-care professionals. A validated AI screening tool could potentially support large-scale programs if developed using representative Indian data and tested across regions and healthcare levels. An Indian model should not simply import a model developed elsewhere; local audiometric practices, equipment, referral pathways, age structure, environmental exposures, and disease patterns should be represented during training and validation.

7.10 Limitations

The principal limitation is the synthetic nature of the dataset. The generated observations demonstrate the research workflow but cannot establish diagnostic validity. The class labels were simulated, the feature relationships were designed to be plausible, and the analysis used one stratified train-test split rather than external validation. Calibration, decision-curve analysis, confidence intervals, subgroup fairness, and prospective clinical outcomes were not assessed. The neural network was a compact tabular architecture rather than a specialized convolutional or recurrent architecture for raw audiogram images or signals. Severity classification is also only one clinical task; type and etiological classification may require additional otologic information.

7.11 Future research

Future studies should replace the synthetic dataset with multicentre clinical data and expert-annotated labels. Patient-level splitting should be mandatory, followed by nested cross-validation and independent external validation. Models can be expanded to gradient boosting, explainable boosting machines, convolutional neural networks for audiogram images, recurrent or transformer architectures for sequential threshold data, and multimodal models combining audiometry, speech, otoscopy, and clinical history. Prospective studies should evaluate referral appropriateness, time savings, clinician confidence, patient outcomes, and cost-effectiveness. Uncertainty-aware AI and privacy-preserving approaches such as federated learning are also promising directions.

8. Conclusion

This study presents a complete research framework for artificial-intelligence-based classification of hearing loss using audiometric and clinical data. The illustrative analysis demonstrates how structured hearing variables can be used to compare linear machine learning, kernel-based learning, ensemble learning, and neural-network approaches. The results show why model comparison, class-aware evaluation, feature interpretation, and transparent methodology are necessary components of a responsible AI study. The evidence from previous research indicates that automated audiogram interpretation is technically feasible, while recent deep-learning studies demonstrate strong performance under controlled experimental conditions. Nevertheless, clinical adoption requires more than a high accuracy value. A trustworthy system must be trained on representative data, protected against leakage, externally validated, interpretable, calibrated, ethically governed, and embedded within



professional clinical workflows. The present synthetic analysis should therefore be viewed as a reproducible methodological template. Its next step is validation using real, de-identified, expert-labelled audiology records collected under appropriate ethical and institutional procedures. The central conclusion is that AI should be developed as an assistive layer around audiological expertise. Properly designed classification tools can help organize large datasets, support screening, identify patterns, and improve consistency. They should not be treated as independent diagnostic authorities. The combination of audiometric measurements, clinical context, machine learning, and explainable decision support offers a promising direction for scalable hearing-care research.

10. References

1. Abu Bakar, A. R., Lai, K. W., & Hamzaid, N. A. (2021). The emergence of machine learning in auditory neural impairment: A systematic review. *Neuroscience Letters*, 765, 136250. <https://doi.org/10.1016/j.neulet.2021.136250>
2. Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
3. Charih, F., Bromwich, M., Mark, A. E., Lefrançois, R., & Green, J. R. (2020). Data-driven audiogram classification for mobile audiometry. *Scientific Reports*, 10, 3962. <https://doi.org/10.1038/s41598-020-60898-3>
4. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297. <https://doi.org/10.1007/BF00994018>
5. 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. *Journal of Medical Systems*, 44, 163. <https://doi.org/10.1007/s10916-020-01627-1>
6. Dou, Z., Li, Y., Deng, D., Zhang, Y., Pang, A., Fang, C., Bai, X., & Bing, D. (2024). Pure tone audiogram classification using deep learning techniques. *Clinical Otolaryngology*, 49(5), 595–603. <https://doi.org/10.1111/coa.14170>
7. Ellis, G. M., & Souza, P. E. (2021). Using machine learning and the National Health and Nutrition Examination Survey to classify individuals with hearing loss. *Frontiers in Digital Health*, 3, 723533. <https://doi.org/10.3389/fdgth.2021.723533>
8. Frosolini, A., Franz, L., Caragli, V., et al. (2024). Artificial intelligence in audiology: A scoping review of current applications and future directions. *Sensors*, 24(22), 7126. <https://doi.org/10.3390/s24227126>
9. GBD 2019 Hearing Loss Collaborators. (2021). Hearing loss prevalence and years lived with disability, 1990–2019: Findings from the Global Burden of Disease Study 2019. *The Lancet*, 397(10278), 996–1009. [https://doi.org/10.1016/S0140-6736\(21\)00516-X](https://doi.org/10.1016/S0140-6736(21)00516-X)
10. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.



11. Kassjański, M., Kulawiak, M., & Przewoźny, T. (2024). Automated hearing loss type classification based on pure tone audiometry data. *Scientific Reports*, 14, 14203. <https://doi.org/10.1038/s41598-024-64310-2>
12. Kim, H., Park, J., Choung, Y.-H., Jang, J. H., & Ko, J. (2021). Predicting speech discrimination scores from pure-tone thresholds—A machine learning-based approach using data from 12,697 subjects. *PLOS ONE*, 16(12), e0261433. <https://doi.org/10.1371/journal.pone.0261433>
13. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30.
14. Shen, Y., Zhou, T., Zou, W., Zhang, J., Yan, S., Ye, H., & Huang, W. (2025). Global, regional, and national burden of hearing loss from 1990 to 2021: Findings from the 2021 Global Burden of Disease study. *Annals of Medicine*, 57(1), 2527367. <https://doi.org/10.1080/07853890.2025.2527367>
15. Shew, M., et al. (2024). Artificial intelligence in otology and neurotology. *Otolaryngologic Clinics of North America*, 57. <https://doi.org/10.1016/j.otc.2024.04.009>
16. World Health Organization. (1980). *International classification of impairments, disabilities, and handicaps*. World Health Organization.
17. World Health Organization. (2020). *WHO ear and hearing survey handbook*. World Health Organization.
18. World Health Organization. (2021a). *World report on hearing*. World Health Organization.
19. World Health Organization. (2021b). *World report on hearing: Executive summary*. World Health Organization.