The Journal of Acoustical Society of India

Volume 52 Number 3 July 2025 pp 79-164

ISSN 0973-3302

THE JOURNAL OF ACOUSTICAL SOCIETY OF INDIA

Volume 52 Number 3 July 2025

ACOUSTICAL SOCIETY OF INDIA

1971

॥ नादाधीनमतोजगत् ॥

A Quarterly Publication of the ASI

https://acoustics.org.in

ASI

The Journal of Acoustical Society of India

The Refereed Journal of the Acoustical Society of India (JASI)

The Journal of Acoustical Society of India is a refereed journal of the Acoustical Society of India **(ASI)**. The **ASI** is a non-profit national society founded in 31st July, 1971. The primary objective of the society is to advance the science of acoustics by creating an organization that is responsive to the needs of scientists and engineers concerned with acoustics problems all around the world.

Manuscripts of articles, technical notes and letter to the editor should be submitted to the Chief Editor. Copies of articles on specific topics listed above should also be submitted to the respective Associate Scientific Editor. Manuscripts are refereed by at least two referees and are reviewed by Publication Committee (all editors) before acceptance. On acceptance, revised articles with the text and figures scanned as separate files on a diskette should be submitted to the Editor by express mail. Manuscripts of articles must be prepared in strict accordance with the author instructions.

All information concerning subscription, new books, journals, conferences, etc. should be submitted to Chief Editor:

B. Chakraborty, CSIR - National Institute of Oceanography, Dona Paula, Goa-403 004,
Tel: +91.832.2450.318, Fax: +91.832.2450.602, e-mail: bishwajit@nio.org

Annual subscription price including mail postage is Rs. 2500/= for institutions, companies and libraries and Rs. 2500/= for individuals who are not **ASI** members. The Journal of Acoustical Society of India will be sent to **ASI** members free of any extra charge. Requests for specimen copies and claims for missing issues as well as address changes should be sent to the Editorial Office:

ASI Secretariat, C/o Acoustics and Vibration Metrology, CSIR-National Physical Laboratory, Dr. KS Krishnan Road, New Delhi 110 012, Tel: +91.11.4560.8317, Fax: +91.11.4560.9310, e-mail: asisecretariat.india@gmail.com

ISSN 0973-3302

Printed at Alpha Printers, WZ-35/C, Naraina, Near Ring Road, New Delhi-110028 Tel.: 9810804196. JASI is sent to ASI members free of charge.

The Journal of Acoustical Society of India

A quarterly publication of the Acoustical Society of India

Volume 52, Number 3, July 2025

ARTICLES

INFORMATION

FOREWORD

Today, Artificial Intelligence (AI) and Machine Learning (ML) applications on sound (audio) data span almost every domain of life - science, industry, healthcare, environment, consumer tech, and the arts. In recent days, we have noticed a sea change in the number of presentations on AI/ ML applications of sound data during the NSA (National Symposium on Acoustics) meetings of ASI (Acoustical Society of India). In this issue, we chose six papers from the NSA-2024 meeting held at the Naval Science and Technology Laboratory (NSTL) (DRDO), Visakhapatnam. The selected researches are: Reshma *et al.* on "Deep learning application on sound data of human health"; Arunachalam and Nirupama on "Cochlear implant recipient care"; Kumar on "Bird song identification and classification"; Vijayasree and Venkatesham on "Use of Particle Swarm Application (PSO)"; Keerthivasan *et al.* on "Deep learning use of underwater passive acoustic data". Sanyal *et al.* provided new and interesting data on the application of AI models in classifying Hindustani music clips. Besides, we have also included three more papers in this issue: Biswas *et al.* on "Unsupervised domain adaptation for multi-label audio event detection"; Dipali Singh *et al.* on "Optimizing active noise control" and Param Preet Singh *et al.*, who highlighted the potential of graph-based semi supervised learning to democratize data annotation and accelerate progress in music information retrieval.

I am grateful to Dr. Ramakrishna. Scientist 'F' and Shri Ganesh Kumar, Former Outstanding Scientist, NSTL- DRDO, Visakhapatnam, who were the organizers of NSA-2024, for their allowance to look into the articles submitted for the NSA conference proceedings. Also grateful to Dr. Mahavir Singh, Managing Editor, JASI and President, ASI, for his help in many ways during the publication of this issue.

Bishwajit Chakraborty, Ph.D.
Former Chief Scientist and CSIR Emeritus Scientist,
CSIR-National Institute of Oceanography, Dona Paula, Goa-403 004
Guest Editor

The Journal of Acoustical Society of India : Vol. 52, No. 3, 2025 (pp. 79-92)

Unsupervised domain adaptation for multi-label audio event detection via maximum classifier discrepancy and confidence-based learning

Arkaprava Biswas[1*], Gaurav Tank[2] and Vipul Arora[3]

[1]Department of Electrical Engineering, IIT Kanpur, India

[2]Department of Computer Science Engineering, IIT Kanpur, India

[3]Department of Electrical Engineering, IIT Kanpur, India

e-mail: arkapravabiswas661@gmail.com

[Received: 00-00-2025; Accepted: 00-00-2025]

ABSTRACT

Audio Event Detection (AED) systems face significant challenges when transferring from synthetic training data to real-world audio, primarily due to domain shifts and lack of labeled target data. To address this, we first explore the Maximum Classifier Discrepancy (MCD) method for unsupervised domain adaptation (UDA) in AED, which aligns feature representations between labeled synthetic and unlabeled real domains by exploiting inter-classifier disagreement. Building on this, we propose a novel confidence-based learning module that refines the adaptation process by prioritizing high-confidence predictions in the unlabeled target domain. This selective learning mechanism mitigates the impact of noisy or uncertain predictions, leading to more stable and effective adaptation. Evaluated on the DESED dataset under the DCASE 2021 Task 4 setup, our method achieves consistent improvements over other baselines. The proposed framework demonstrates strong potential for bridging the synthetic-real gap in AED without requiring any labeled real-world data.

1. INTRODUCTION

Audio Event Detection (AED) plays a critical role in various real-world applications, such as public safety surveillance, smart home automation, wildlife monitoring, and multimedia indexing. The AED task comprises two sub-tasks: audio tagging (also known as weak detection), where the goal is to identify the presence of sound events in an audio clip, and strong detection, which aims to temporally localize those events at the frame level. Let $x \in \mathrm{R}^{T \times D}$ denote a feature sequence extracted from an audio clip, where T is the number of frames and D is the feature dimension. For C sound event classes, strong detection produces frame-wise outputs $\mathrm{y}_t \in [0,1]^{T \times C}$, while weak detection yields clip level outputs $\mathrm{y} \in [0,1]^{C}$. Figure 1 shows the log-Mel spectrogram of an audio clip. Figure 2 and 3 presents the frame-level strong annotations and clip-level weak annotations respectively.

Recent AED systems have benefited significantly from the use of Deep Neural Networks (DNNs), particularly Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and their hybrid variants such as Convolutional Recurrent Neural Networks (CRNNs)[1, 2, 3]. Several architectural

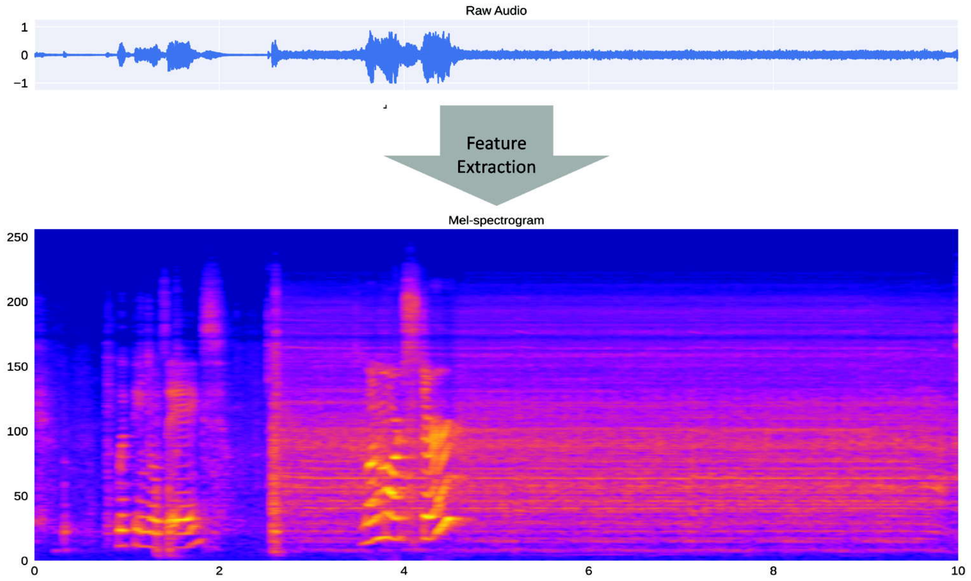
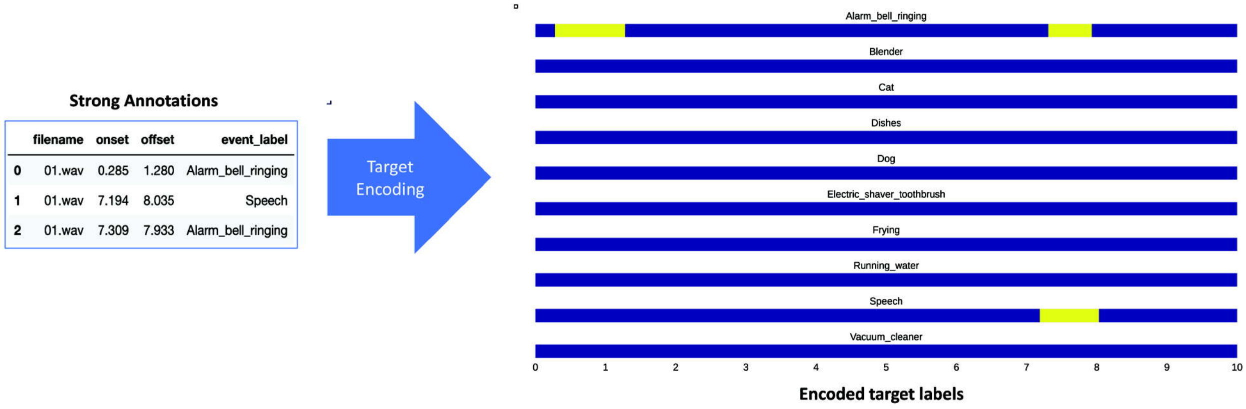
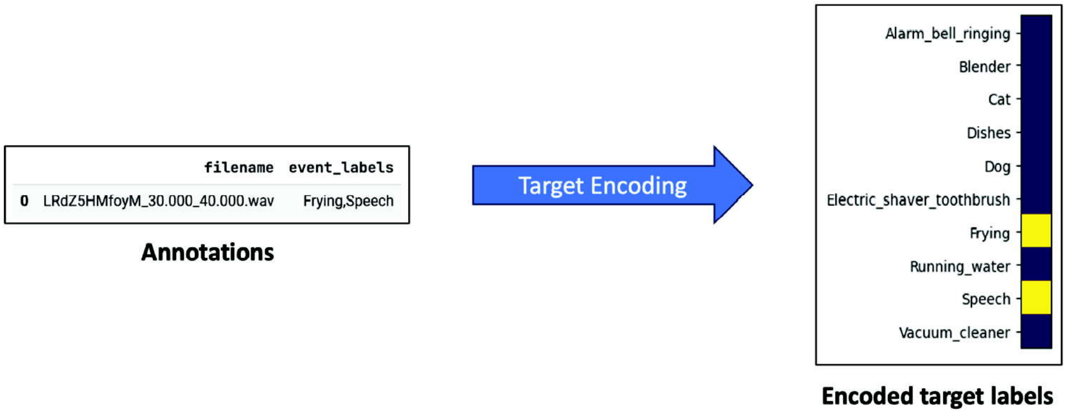


Fig. 1. A log-mel-spectogram representation of an audio.



	filename	onset	offset	event_label
0	01.wav	0.285	1.280	Alarm_bell_ringing
1	01.wav	7.194	8.035	Speech
2	01.wav	7.309	7.933	Alarm_bell_ringing

Fig. 2. Frame-level strong annotation of an audio.



	filename	event_labels
0	LRdZ5HMfoyM_30.000_40.000.wav	Frying,Speech

Fig. 3. Clip-level weak annotation of an audio.

innovations have been proposed to improve AED, including Frequency-Dynamic CNNs[4], Forward-Backward CRNNs[5], and transformer-based approaches[6].

Despite the advancements in AED architectures, a major challenge persists: the scarcity of strongly labeled real-world audio data. Strong labels, which provide frame-level temporal annotations for each sound event, are expensive and time-consuming to obtain at scale. To address this issue, recent research has embraced weakly supervised and semi-supervised learning strategies that make efficient use of both limited labeled data and abundant unlabeled or weakly labeled audio.

A prominent approach in this direction is the Mean Teacher (MT) framework, which has been successfully adopted in the DCASE Challenge series, particularly in Task 4[7]. The MT framework is a semi-supervised learning technique based on consistency training, originally proposed for image classification, and later adapted for audio event detection. It involves two models: a student and a teacher. The student model is trained via gradient descent, while the teacher model is an exponential moving average (EMA) of the student weights. The central idea is to encourage the student model to produce predictions that are consistent with those of the teacher under various input perturbations or augmentations.

In the AED context, the MT framework enables learning from unlabeled real audio by promoting prediction consistency between the student and teacher on these samples. The synthetic audio, which is fully labeled with strong annotations, is used to supervise the student with standard supervised losses, such as frame-wise binary cross-entropy. Meanwhile, the real audio, which lacks strong labels, is processed through both the student and teacher networks, and a consistency loss-typically Mean Squared Error (MSE) or KL-divergence-is applied between their predictions.

This approach was first introduced in[7] and has since served as a foundational baseline for many AED systems. Subsequent studies have extended the MT framework in several ways to improve its robustness and performance. For example, RCT[8] integrates residual connections and temporal convolution blocks to enhance representation learning in the student-teacher paradigm. Couple LF[9] couples local and global temporal features in the MT framework to capture multi-scale dependencies in audio. Guided Learning[10] introduces an additional guidance module that refines the pseudo-labels generated from the teacher to improve learning from noisy real-world data. A joint training framework combining Guided Learning with Mean Teacher was proposed in[11], showing improved generalization to unseen audio domains. Consistency regularization has been coupled with strong augmentations such as Spec Augment, mixup, or random masking to encourage the model to be invariant to these transformations[12, 13].

In summary, the Mean Teacher framework has become a cornerstone of weakly and semi-supervised AED research, enabling the effective utilization of synthetic and real data without requiring frame-level annotations for the latter. Its adaptability and ease of integration with other modules and losses have led to its widespread adoption in the community, particularly under the DCASE challenge benchmarks.

However, models trained predominantly on synthetic audio often suffer from degraded performance on real-world audio due to domain shift-a discrepancy between the distributions of training (source) and deployment (target) domains. This issue has received relatively limited attention in the AED community. Most domain adaptation research to date has focused on multi-class classification problems[14, 15], which do not directly translate to the multi-label nature of AED.

One of the few works addressing domain adaptation in AED is by Yang *et al.*[16], who applied domain adversarial training using the Domain-Adversarial Neural Network (DANN)[17] framework. Their approach aimed to learn domain in variant features through adversarial alignment but failed to preserve class discriminability due to its class-agnostic nature. Other methods include the application of Joint Distribution Optimal Transport (JDOT)[18], and hybrid systems combining Mean-Teacher with domain discriminator modules[19].

In this work, we explore a principled solution to the domain adaptation problem in AED by adapting the Maximum Classifier Discrepancy (MCD) method[20], which was originally developed for multi-class classification. MCD works by training two classifiers to maximize their disagreement on target (unlabeled real) data while simultaneously training a feature generator to minimize this discrepancy, thus encouraging discriminative and domain-invariant features. We extend this idea to the multi-label AED setting, demonstrating its efficacy in aligning synthetic and real-world audio distributions.

Furthermore, we introduce a novel confidence-based learning module designed to refine the MCD framework. While MCD encourages exploration of the target domain's decision boundary, it is susceptible to instability caused by noisy or low-confidence predictions. Our proposed module selectively focuses learning on high-confidence predictions in the target domain, effectively filtering out uncertain examples during training. This confidence-guided fine tuning not only stabilizes the adaptation process but also enhances the model's discriminability in complex acoustic scenes.

Our key contributions are summarized as follows:

1. We adapt the Maximum Classifier Discrepancy (MCD) method for the task of multi-label Audio Event Detection, demonstrating its effectiveness in addressing domain shift between synthetic and real audio data.
2. We propose a novel confidence-based fine-tuning module that leverages prediction certainty to selectively refine learning in the target domain, resulting in more stable and robust adaptation.
3. Our method requires no strong labels for real-world audio, operating in a fully unsupervised domain adaptation (UDA) setting, and shows consistent performance improvements over strong baselines on the DESED dataset under the DCASE 2021 Task 4 framework.

2. MATHEMATICAL FORMULATION

In this section, we outline the theoretical foundations of our approach to unsupervised domain adaptation (UDA) for audio event detection (AED). Our work builds upon the discrepancy-based framework proposed in[20], which leverages the concept of HΔH-divergence to measure distributional differences between the source and target domains. We begin by formally defining this divergence and then derive a generalization bound on the target domain error.

2.1 *HΔH-Divergence and Generalization Bounds*

Let X be the input space and H a hypothesis class of functions mapping $X \rightarrow \{0,1\}^C$, where C is the number of sound event classes. Given a source domain distribution D_S and a target domain distribution D_T, the goal of UDA is to minimize the expected error on the target domain $R_T(h)$, using labeled data drawn from DS and unlabeled data from D_T.

2.1.1 HΔH-Divergence : The symmetric difference hypothesis space is defined as :

$$H\Delta H = \{x \rightarrow h(x) \oplus h'(x) \mid h,h' \in H\},$$

where $\oplus$ denotes the XOR operation. Intuitively, the disagreement between two hypotheses on a given input measures the uncertainty in classification. The $H\Delta H$-divergence between two distributions D_S and D_T is defined as:

$$d_{H\Delta H}(D_S, D_T) = 2 \sup_{h,h' \in H} |\mathbb{E}_{x\sim D_S}[I(h(x) \neq h'(x))] - \mathbb{E}_{x\sim D_T}[I(h(x) \neq h'(x))]| \; .$$

This divergence quantifies the maximum difference in disagreement between pairs of hypotheses over the two domains. A larger $d_{H\Delta H}$ implies a higher domain shift.

2.1.2 A Tighter Upper Bound on Target Domain Error

Let $h \in H$ denote a hypothesis (*e.g.,* a sound event detector), and let $R_S(h)$ and $R_T(h)$ represent the expected errors of h on the source and target domains, respectively. Let $h^* \in H$ be the ideal joint hypothesis minimizing the combined error on both domains:

$$h^* = \arg\min_{h \in H} R_S(h) + R_T(h).$$

We present a bound on the target error in terms of the source error and the *HΔH*-divergence, following the theoretical framework introduced in[20].

[Relation Between Source and Target Disagreement] For any two hypotheses $h, h' \in H,$

$$|R_S(h,h') - R_T(h,h')| \leq \frac{1}{2} d_{H\Delta H}(D_S, D_T).$$

From the definition of *HΔH*-divergence,

$$d_{H\Delta H}(D_S,D_T) = 2 \sup_{h,h' \in H} |R_S(h,h') - R_T(h,h')|,$$

implying

$$|R_S(h,h') - R_T(h,h')| \leq \frac{1}{2} d_{H\Delta H}(D_S, D_T),$$

for all $h, h' \in H$.

[Triangle Inequality for Hypotheses] For any three hypotheses $h_1, h_2, h_3 \in H$,

$$R(h_1, h_2) \leq R(h_1, h_3) + R(h_2, h_3)$$

[Target Error Bound] For any hypothesis $h \in H$,

$$R_T(h) \leq R_S(h) + \frac{1}{2} d_{H\Delta H}(D_S, D_T) + \lambda,$$

where $\lambda = \min_{h \in H} \mathrm{R}_S(h) + R_T(h)$ is the joint error of the optimal hypothesis. Using Theorem 2.1.2,

$$R_T(h) \leq R_T(h^*) + R_T(h, h^*).$$

Applying the triangle inequality and Theorem 2.1.2,

$$R_T(h) \leq R_T(h^*) + R_S(h,h^*) + |R_T(h,h^*) - R_S(h,h^*)|$$

$$\leq R_T(h^*) + R_S(h,h^*) + \frac{1}{2} d_{H\Delta H}(D_S, D_T).,$$

Further, $R_S(h, h^*) \leq R_S(h) + R_S(h^*)$ (by triangle inequality),

$$\Rightarrow R_T(h) \leq R_S(h) + R_S(h^*) + R_T(h^*) + \frac{1}{2} d_{H\Delta H}(D_S, D_T).$$

Let $\lambda = R_S(h^*) + R_T(h^*)$, which is the error for ideal joint hypothesis.

2.1.3 Generalization with Finite Samples

In practical scenarios, we work with finite samples from the source and target domains. Let $\hat{d}_{H\Delta H}(S,T)$ be the empirical *HΔH*-divergence computed from these samples. Then the generalization bound becomes:

$$R_T(h) \leq R_S(h) + \frac{1}{2} \hat{d}_{H\Delta H}(S,T) + \lambda + \beta,$$

where β is a generalization error term that depends on the VC-dimension of the hypothesis class H, and the number of samples from D_S and D_T. This bound motivates methods like Maximum Classifier Discrepancy (MCD) for domain adaptation, which seeks to minimize the target error by optimizing the source loss and maximizing the discrepancy between classifiers on the target domain.

3. EXPERIMENTS

In this section, we present our proposed approach for unsupervised domain adaptation in audio event detection (AED) using the Maximum Classifier Discrepancy (MCD) framework. Originally developed for single-label image classification and semantic segmentation[20], we adapt and extend MCD to the multilabel, frame-level AED setting, where the goal is to align synthetic (source) and real (target) audio domains at a fine temporal resolution.

3.1 Maximizing Classifier Discrepancy for AED

We first aim to minimise $R_S(h)+\frac{1}{2}d_{H\Delta H}(D_S,D_T)$ for synthetic to real audio domain adaptation by adapting the MCD framework. We begin by training a feature extractor $F(.)$ and two independent classifiers $C_1(.)$ and $C_2(.)$ using strongly labeled synthetic audio. This initial training phase learns class boundaries in the source domain. To enable generalization to the target domain, we employ the discrepancy between C_1 and C_2's predictions on unlabeled real audio to guide the alignment of features learned by $F(.)$. The classifiers are optimized to push target samples toward decision boundaries, while the feature generator is optimized to pull them away, resulting in aligned yet class-discriminative target features.

3.1.1 Supervised Training on Source and Weak Target Labels

Let $x^S \in X_*^S$ denote an augmented source-domain audio sample with corresponding frame-level label $y^S \in Y_*^S$. The log-mel spectrogram of x^S is passed through $F(.)$ to obtain frame-level embeddings, which are further passed to $C_1(.)$ and $C_2(.)$, yielding predictions $\hat{y}_1^S$ and $\hat{y}_2^S$.

Similarly, for weakly labeled target-domain audio $x^W \in X^W$, with clip-level label $y^W \in Y^W$, the spectrogram is passed through the same pipeline to produce clip-level predictions $\hat{y}_1^W, \hat{y}_2^W$ obtained by aggregating frame-level predictions.

The frame-level supervised loss is defined as:

$$\mathrm{BCE}_{\mathrm{frame}} = \mathrm{BCE}\left(y^S, \hat{y}_1^S\right) + \mathrm{BCE}\left(y^S, \hat{y}_2^S\right),$$

while the clip-level supervised loss is :

$$\mathrm{BCE}_{\mathrm{clip}} = \mathrm{BCE}\left(y^W, \hat{y}_1^W\right) + \mathrm{BCE}\left(y^W, \hat{y}_2^W\right).$$

The total supervised loss $\mathcal{L}_S$ is the sum of both:

$$\mathcal{L}_s = BCE_{frame} + BCE_{clip}, \tag{1}$$

In the first stage, we optimize $F(.)$, $C_1(.)$, and $C_2(.)$ jointly by minimizing $\mathcal{L}_s$.

3.1.2 Maximizing Discrepancy Between Classifiers

Next, to expose the decision boundaries in the target domain, we compute the disagreement between C_1 and C_2 on unlabeled target data. Let $x^T \in X^T$ be an unlabeled target audio sample. The spectrogram of x^T is passed through $F(.)$, and then through C_1 and C_2, resulting in frame-level predictions $\hat{y}_1^T$ and $\hat{y}_2^T$, respectively.

The frame-level discrepancy at time step t is defined as:

$$\left|\hat{y}_1^T(t) - \hat{y}_2^T(t)\right|,$$

and the total discrepancy loss over all frames and samples is:

$$\mathcal{L}_d = \sum_{(y_1^T, y_2^T)} \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left|y_1^T(t) - y_2^T(t)\right| . \tag{2}$$

We now update the classifiers C_1 and C_2 by maximizing the discrepancy $\mathcal{L}_d$, while still ensuring their performance on the source data via $\mathcal{L}_s$. The combined objective for classifier update becomes:

$$\min_{C_1, C_2} \left(\mathcal{L}_s - \delta \mathcal{L}_d\right), \tag{3}$$

where $\delta \geq 0$ is a hyperparameter controlling the influence of the discrepancy term.

3.1.3 Feature Alignment via Discrepancy Minimization

To align target features with the source decision boundaries, we update the feature extractor $F(.)$ to minimize the discrepancy $\mathcal{L}_d$, while keeping C_1 and C_2 fixed. This encourages $F(.)$ to generate feature embeddings for target-domain inputs that yield consistent predictions across both classifiers.

We perform multiple updates (*e.g.*, K steps) to $F(.)$ for every update of $C_1(.)$ and $C_2(.)$, following the alternating optimization procedure:

- **Step 1 :** Minimize $\mathcal{L}_s$ over F, C_1, C_2.
- **Step 2 :** Fix F, minimize $\mathcal{L}_s - \delta\mathcal{L}_d$ over C_1, C_2.
- **Step 3 :** Fix C_1, C_2, minimize $\mathcal{L}_d$ over F for K steps.

Through this adversarial interplay, the model learns domain-invariant, class discriminative frame-level representations suitable for AED in real-world, unlabeled audio recordings.

3.2 Pseudo-label Based Learning for AED

Once the feature extractor $F(.)$ and classifiers $C_1(.)$, $C_2(.)$ have been trained using the maximum classifier discrepancy framework, we select and checkpoint the best-performing model based on validation performance. We then further refine the model by minimizing the combined error over both the source and target domains through pseudo-label based learning. This way we further minimise λ, which is the error of the hypothesis on the combined source and target domain.

To generate the pseudo-labeled set $\langle X_P, Y_P \rangle$, we consider all unlabeled target-domain samples $x^T \in X^T$. For each x^T, we obtain frame-level predictions $\hat{y}_1^T$ and $\hat{y}_2^T$ from classifiers $C_1(.)$ and $C_2(.)$, respectively. We retain those samples for which the classifiers strongly agree across frames, *i.e.*,

$$\sum_{t=1}^{\mathcal{T}} \left| y_1^T(t) - y_2^T(t) \right| < \gamma,$$

where γ is a threshold hyperparameter controlling the confidence level. For the retained samples, we define the soft pseudo-label as the mean prediction:

$$\hat{y}^T(t) = \frac{y_1^T(t) + y_2^T(t)}{2} .$$

To obtain discrete frame-wise labels, we apply a thresholding function $\phi_\alpha(\cdot)$ defined as:

$$\phi_\alpha(z(t)) = \begin{cases} 1 & \text{if } z(t) \geq \alpha, \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha \in (0,1)$ is a confidence threshold for binarizing the average prediction. In practice, to adapt the confidence condition over time, we follow a dynamic thresholding approach similar to single-label classification settings, where:

$$\gamma = \gamma' \cdot \beta_{\text{epoch}},$$

with γ' and $\beta \in (0,1)$ as hyperparameters controlling the initial strictness and decay rate.

The confidence-based loss $\mathcal{L}_{conf}$ is then defined over the pseudo-labeled samples as follows:

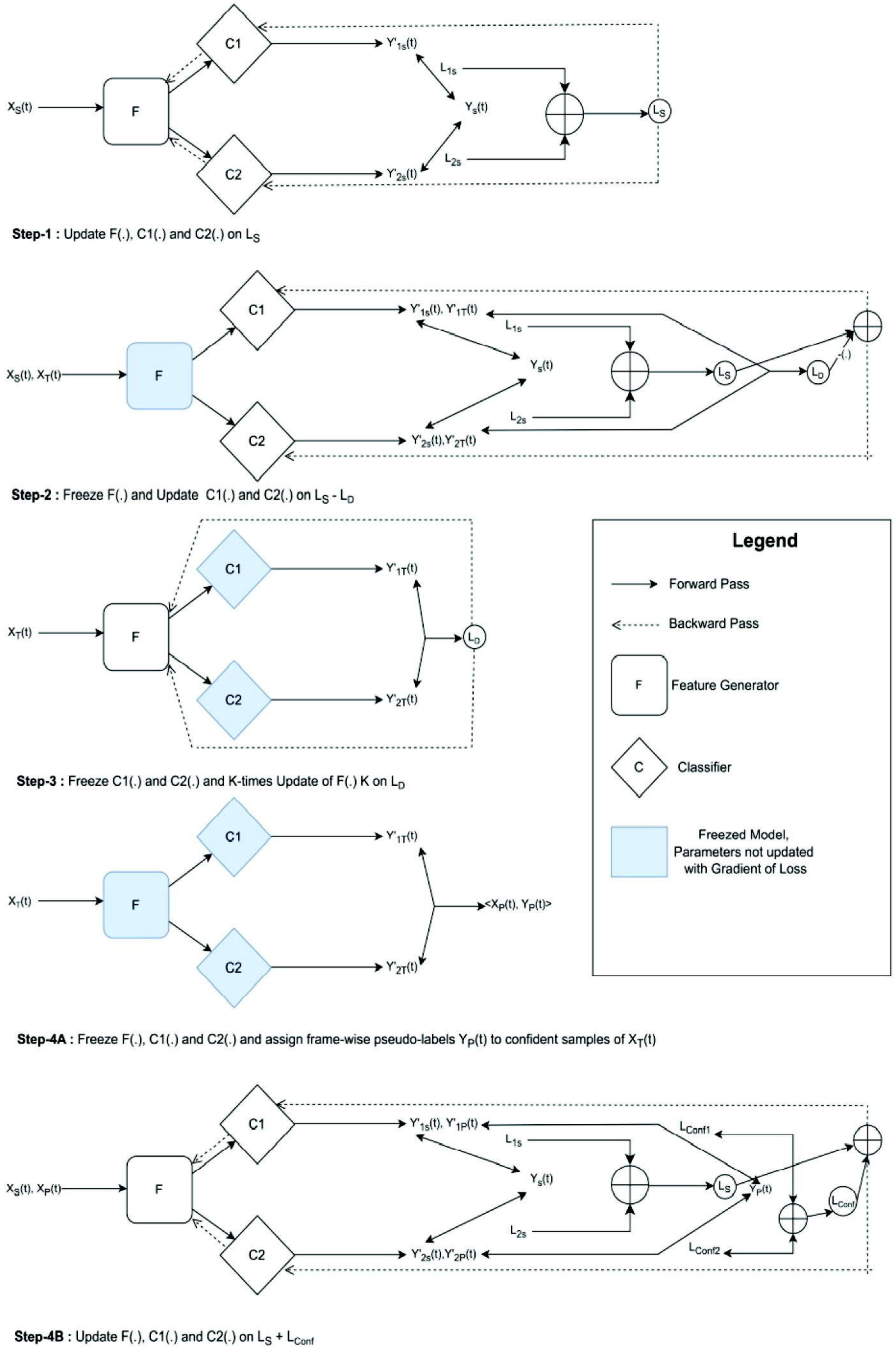
$$\mathcal{L}_{conf} = \frac{1}{B} \sum_{B} \mathbb{I} \left(\sum_{t=1}^{\mathcal{T}} \left| y_1^T(t) - y_2^T(t) \right| \leq \gamma \right) \cdot \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \left[y_1'^T(t) \log y_1^T(t) + y_2'^T(t) \log y_2^T(t) \right] \tag{4}$$

where $y_1'^T(t) = \phi_\alpha \left(\frac{y_1^T(t) + y_2^T(t)}{2} \right)$, and similarly for $y_1'^T(t)$. The indicator function $I(\cdot)$ filters out predictions that are not confidently aligned across classifiers.

Finally, we minimize the total loss, consisting of both the supervised loss $\mathcal{L}_s$ and the confidence-based loss $\mathcal{L}_{conf}$, jointly over $F(.)$, $C_1(.)$ and $C_2(.)$:

$$\min_{F, C_1, C_2} (\mathcal{L}_s + \mathcal{L}_{conf}). \tag{5}$$

This phase reinforces decision boundaries with high-confidence target predictions while continuing to leverage strongly and weakly labeled source data, improving model generalization across domains. Figure 4 presents a step-by-step illustration of the proposed solution.



Step-1 : Update F(.), C1(.) and C2(.) on L_S

Step-2 : Freeze F(.) and Update C1(.) and C2(.) on L_S - L_D

Step-3 : Freeze C1(.) and C2(.) and K-times Update of F(.) K on L_D

Step-4A : Freeze F(.), C1(.) and C2(.) and assign frame-wise pseudo-labels $Y_P(t)$ to confident samples of $X_T(t)$

Step-4B : Update F(.), C1(.) and C2(.) on L_S + L_{Conf}

Fig. 4. Diagram illustrating the proposed algorithm.

3.3 *Pseudo-code of Our Algorithm for Audio Event Detection*

Algorithm 1 Proposed Training Framework for AED with Domain shift

```
Require: Strongly labeled source data, $\langle X_*^S, Y_*^S \rangle$,
Weakly labeled target data $\langle X^W, Y^W \rangle$, Unlabeled target data $X^T$
Require : Number of feature extractor updates per discrepancy step $K$, discrepancy weight $\delta$, confidence threshold $\gamma$, thresholding function $\phi_\alpha$, dynamic scaling factor $\beta$
Initialize networks $F(.)$, $C_1(.)$, $C_2(.)$
while not converged do
    // Step 1: Supervised Training on Strong Synthetic Source and Weak Target
    Compute frame-level predictions $y_1^S$, $y_2^S$ from $X^S$
    Compute clip-level predictions from $X^W$ using aggregation over frames
    Compute supervised loss $\mathcal{L}_s$ as in Eq. 1
    Update $F(.)$, $C_1(.)$, $C_2(.)$ to minimize $\mathcal{L}_s$
    // Step 2: Maximize Classifier Discrepancy
    Compute $\mathcal{L}_d$ on unlabeled target audio $X^T$ as in Eq. (2)
    Update $C_1(.)$, $C_2(.)$ to maximize $\mathcal{L}_d$ while minimizing $\mathcal{L}_s$: $\min_{C_1, C_2} (\mathcal{L}_s + \delta \cdot \mathcal{L}_d)$.
    // Step 3: Feature Alignment
    for $i = 1$ to $K$ do
        Update $F(.)$ to minimize $\mathcal{L}_d$
    end for
end while
// Step 4: Pseudo-label Based Fine-tuning
Checkpoint best model $F(.)$, $C_1(.)$, $C_2(.)$ based on validation
Load checkpointed model
for each epoch do
    Update dynamic threshold: $\gamma \leftarrow \gamma' \cdot \beta_{epoch}$
    Initialize pseudo-labeled set $\langle X^P, Y^P \rangle \leftarrow \varnothing$
    for each $x^T \in X^T$ do
        Compute $T$ $\hat{y}_1^T$, $\hat{y}_2^T$
        if $\sum_{t=1} \left| y_1^T(t) - y_1^T(t) \right| < \gamma$ then
            Compute $\hat{y}^T = \phi_\alpha \left( \frac{\hat{y}_1^T + \hat{y}_2^T}{2} \right)$
            Add $\langle x^T, \hat{y}^T \rangle$ to $\langle X^P, Y^P \rangle$
        end if
    end for
    Compute $\mathcal{L}_{conf}$ on $\langle X^P, Y^P \rangle$ as in Eq. (4)
    Compute supervised loss $\mathcal{L}_s$ as in Eq. 1
    Fine-tune $F(.)$, $C_1(.)$, $C_2(.)$ to minimize $\mathcal{L}_s + \mathcal{L}_{conf}$
end for
```

4. EXPERIMENTS

4.1 Experimental settings

Dataset : We have experimented with our method on the DCASE-2021 task-4 dataset[21] or DESED Dataset. The dataset comprises 10 sound event classes: alarm_bell_ringing, blender, cat, dishes, dog, electric shaver/toothbrush, frying, running water, voice, and vacuum cleaner. 10000 synthetic highly labelled clips, 1578 weakly labeled real audio clips, and 14412 unlabeled real audio clips are used for the training. We use the DCASE-validation dataset as our evaluation dataset, which consists of 692 strongly labelled real audio samples. We split the evaluation dataset into validation and test set with 1:1 ratio. The synthetic audio dataset is a strongly labelled dataset that are synthetically generated by the Scaper library[22]. The real audio clips are extracted from Audioset[23].

Audio Pre-processing : For all studies, we pre-process log-melspectograms from raw audio recordings (sample rate 16000 Hz) with window size 128 ms, hop length 16 ms, and 128 mel-bins.

Model Architecture : Our experiments utilize a CRNN architecture with 2.234 M parameters. The feature extractor comprises 11 convolutional blocks with Relu activation and batch normalization. Five blocks include max pool layers with a pooling length of 2 across the frequency axis. The convolutional layers output channels are 16, 16, 32, 32, 64, 64, 128, 128, 256, 256, 256 respectively. The final convolutional layer yields a feature map of size (4, T, 256), which is reshaped to (T, 1024) by concatenating all channels' frequency contents. A Dense layer further down samples this to a 256-dimensional feature map per frame. A bidirectional RNN layer then generates a 128-dimensional frame-level embedding. One linear layer classifier maps the 128-dimensional frame-level embeddings into 10 classes. We take the mean or weighted mean of frame-level predictions across all time frames to get weak-level predictions.

Baseline Systems : For empirical comparison, we re-implement three established baselines: a fully supervised system, a domain adversarial discriminator model[16], and the official DCASE 2021 Task 4 baseline[24]. The fully supervised model is trained using a combination of strongly labeled synthetic audio and weakly labeled real audio samples. The DCASE 2021 baseline leverages the Mean Teacher framework[7], incorporating both synthetic strongly labeled and real weakly labeled data. It employs a supervised training scheme alongside a consistency loss that enforces temporal prediction stability for unlabelled real audio. This consistency loss is computed using audio-specific augmentations, including mix-up, time masking, and filter augmentations[25].

The domain discriminator baseline adopts a domain-adversarial neural network (DANN) approach[17] to align the latent feature distributions of labeled synthetic and unlabeled real audio at the frame level. Following this alignment, the model is fine-tuned using weakly labeled real audio. All models are trained under identical conditions, including the same network architecture and hyper parameter settings, to ensure a fair and controlled comparison.

In all baselines, the feature extractor and classifier are integrated into a unified model with 2.23M parameters. The domain discriminator module consists of a lightweight feed forward network with three linear layers, comprising approximately 0.031 million parameters.

4.2 Results

All systems are evaluated using the SED-Eval[26] library. Predictions are made by thresholding frame-level outputs to 0.5 and post-processing with a standard median filter setup for the DCASE dataset[27]. Performance is compared based on micro-averaged event-based, segment-based F_1-scores, PSDS-1[28] and PSDS-2, as shown in Table 1. The event-based F_1-score is measured by accurately identifying each sound event within a specified collar, and the segment-based F_1-score is measured by accurately determining events within a particular segment of the audio. We keep the thresholding hyperparameter $\alpha = 0.5$ for confidence-based fine-tuning and discrepancy loss weight $\delta = 1$. We have set the initial confidence level to 0.005 and β to 0.95 for the bench-marking performance.

Table 1. Performance Comparison on AED, result is presented as mean ± standard deviation, computed from 3 independent runs.

Method	Event-based F_1	Segment-based F_1	PSDS-1	PSDS-2
Fully Supervised	27.2 ± 0.1	56.4 ± 0.4	0.169 ± 0.004	0.289 ± 0.007
DCase 2021-task-4 baseline	36.7 ± 0.2	70.6 ± 0.8	0.246 ± 0.003	0.417 ± 0.005
Domain Discriminator	28.7 ± 0.3	60.2 ± 0.4	0.177 ± 0.005	0.309 ± 0.006
Ours	37.3 ± 0.1	63.6 ± 0.8	0.248 ± 0.005	0.410 ± 0.005

For the maximum classifier discrepancy module, we have used same Adam optimizer with learning rate 1 x 10^{-4} and betas = (0.9, 0.999) for $F(.)$, $C_1(.)$, and $C_2(.)$. For the confidence based learning module, we have used same Adam optimizer with learning rate 1 x 10^{-5}, betas = (0.9, 0.999) and weight decay = 0.005 for $F(.)$, $C_1(.)$, and $C_2(.)$.

Table 1 presents the performance comparison of different methods evaluated on the DESED dataset. For our proposed method, we fix the feature update hyper-parameter K to 8 and set the confidence threshold γ to 0.005. The results demonstrate that our approach achieves superior performance in terms of event-based F_1 score and PSDS-1 metrics, while delivering comparable results for segment-based F_1 score and PSDS-2. This highlights the effectiveness of our method in improving precise event detection and overall system robustness.

4.3 Ablation Studies

We examine the influence of confidence-based fine-tuning and varying confidence levels on model performance. Table 2 illustrates performance with the confidence-based learning module for different initial confidence levels, highlighting its impact on model performance. The table shows that the initial confidence level has a noticeable effect on various performance metrics.

Table 2. Table 2: Effect of various initial confidence level on performance, result is presented as mean ± *standard deviation,* computed from 3 independent runs.

Method	Event-based F_1	Segment-based F_1	PSDS-1	PSDS-2
0	36.1 ± 0.6	62.9 ± 0.5	0.233 ± 0.005	0.379 ± 0.003
0.001	35.8 ± 0.3	63.3 ± 0.4	0.236 ± 0.002	0.403 ± 0.003
0.002	36.6 ± 0.2	63.8 ± 0.5	0.262 ± 0.004	0.414 ± 0.006
0.005	37.3 ± 0.1	63.6 ± 0.8	0.248 ± 0.005	0.410 ± 0.005

When the initial confidence level is set to 0, the Event-based F_1 score is 36.1, the Segment-based F_1 score is 62.9, PSDS-1 is 0.233, and PSDS-2 is 0.379. Increasing the initial confidence level to 0.001 results in a slight decrease in the Event-based F_1 score to 35.8 but an increase in the Segment-based F_1 score to 63.3, with improvements in PSDS-1 and PSDS-2 to 0.236 and 0.403, respectively.

Setting the initial confidence level to 0.002 further improves the performance metrics, achieving an Event-based F_1 score of 36.6, the highest Segment-based F_1 score of 63.8, and the highest PSDS-1 and PSDS-2 scores of 0.262 and 0.414, respectively. At an initial confidence level of 0.005, the Event-based F_1 score reaches its peak at 37.3, while the Segment-based F_1 score slightly decreases to 63.6, and PSDS-1 and PSDS-2 scores are 0.248 and 0.410, respectively.

These results demonstrate that different initial confidence levels can significantly affect the performance of the model, we can tune the confidence level for our desired metric. For all the experiments, we have set β = 0.95.

We further conduct an ablation study to evaluate system performance without incorporating this module, focusing solely on the feature alignment phase. In this setting, we vary the hyperparameter K,

Table 3. K vs F1-scores w.o. pseudo-label based fine-tuning on AED, result is presented as mean ± standard deviation, computed from 3 independent runs.

Method	Event-based F_1	Segment-based F_1	PSDS-1	PSDS-2
2	34.1 ± 0.6	61.4 ± 0.4	0.218 ± 0.002	0.360 ± 0.004
4	35.6 ± 0.4	62.1 ± 0.6	0.222 ± 0.006	0.392 ± 0.008
8	36.1 ± 0.6	62.9 ± 0.5	0.233 ± 0.005	0.379 ± 0.003

which controls the number of update steps applied to the feature encoder $F(.)$ for each update of the classifiers $C_1(.)$ and $C_2(.)$. The performance results corresponding to different values of K are summarized in Table 3.

Our observations indicate that increasing K leads to consistent improvements in AED performance, demonstrating that deeper optimization of the feature encoder enhances the alignment of frame-level features between the source and target domains. This suggests that stronger feature alignment directly contributes to more effective adaptation in the absence of pseudo-label supervision.

5. CONCLUSION AND FUTURE WORKS

In this work, we proposed an effective domain adaptation framework for multilabel Audio Event Detection (AED) that addresses the distribution discrepancy between synthetic and real audio domains. Building upon the Maximum Classifier Discrepancy (MCD) method, we extended it to handle the challenges of multi-label frame-level AED by training two classifiers and a shared feature extractor to align feature distributions while preserving class boundaries. Further, we introduced a novel confidence-based pseudo-label learning strategy that selectively incorporates high-confidence target domain predictions to fine-tune the model, thereby improving detection accuracy on real-world data.

Our experimental results on the DESED dataset demonstrate that the proposed approach achieves comparable performance to state-of-the-art methods for AED and superior performance compared to other unsupervised domain adaptation methods for AED. The analysis also highlights the critical role of feature encoder update iterations and confidence thresholding in enhancing domain alignment and overall system robustness.

For future work, we plan to explore several promising directions. First, investigating adaptive or trainable confidence thresholding mechanisms and curriculum learning strategies may improve the pseudo-label quality and training stability. Lastly, extending the framework to handle domain shifts across other audio. speech and biomedical applications.

REFERENCES

[1] Y. Chen, Y. Zhang and Z. Duan, 2017. "Sound event detection using convolutional neural networks," DCASE.

[2] G. Parascandolo, H. Huttunen and T. Virtanen, 2016. "Recurrent neural networks for polyphonic sound event detection in real life recordings," in 2016 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6440-6444.

[3] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen and T. Virtanen, 2017. "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **25**(6): 1291-1303.

[4] H. Nam, S.-H. Kim, B.-Y. Ko and Y.-H. Park, 2022. "Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection," *INTERSPEECH 2022*.

[5] J. Ebbers and R. Haeb-Umbach, 2020. "Forward-backward convolutional recurrent neural networks and tag-conditioned convolutional neural networks for weakly labeled semi-supervised sound event detection".

[6] Z. Ye, X. Wang, H. Liu, Y. Qian, R. Tao, L. Yan and K. Ouchi, 2021. "Sound event detection transformer: An event-based end-to-end model for sound event detection".

[7] L. JiaKai, 2022. "Mean teacher convoluion system for dcase 2018 task 4," 2018.

[8] N. Shao, E. Loweimi and X. Li, 2022. "RCT: Random consistency training for semi-supervised sound event detection," in *Proc. Interspeech*, pp. 1541-1545.

[9] R. Tao, L. Yan, K. Ouchi and X. Wang, 2021. "Couple learning for semisupervised sound event detection," in *Interspeech.*

[10] L. Lin, X. Wang, H. Liu and Y. Qian, 2020. "Guided learning for weaklylabeled semi-supervised sound event detection," in ICASSP 2020 - 2020, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626-630.

[11] H. Yen, P.-J. Ku, M.-C. Yen, H.-S. Lee and H.-M. Wang, 2020. "Joint training of guided learning and mean teacher models for sound event detection".

[12] H. Dinkel, X. Cai, Z. Yan, Y. Wang, J. Zhang and Y. Wang, 2021. "The smallrice submission to the dcase 2021 task 4 challenge: A lightweight approach for semi-supervised sound event detection with unsupervised data augmentation".

[13] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong and Q. V. Le, 2019. "Unsupervised data augmentation for consistency training," *arXiv preprint arXiv:1904.12848*.

[14] E. Tzeng, J. Hoffman, K. Saenko and T. Darrell, 2017. "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[15] A. Kumagai and T. Iwata, 2019. "Unsupervised domain adaptation by matching distributions based on the maximum mean discrepancy via unilateral transformations," in *AAAI Conference on Artificial Intelligence*. [Online]. Available: https://api.semanticscholar.org/CorpusID:57506940

[16] L. Yang, J. Hao, Z. Hou and W. Peng, 2020. "Two-stage domain adaptation for sound event detection".

[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand and V. Lempitsky, 2017. Domain-Adversarial Training of Neural Networks. Cham: Springer International Publishing, pp. 189-209. [Online]. Available: https://doi.org/10.1007/978-3-319-58347-110

[18] G. G. Michel Olvera and Emmanuel Vincent, 2021. "Improving sound event detection with auxiliary foreground-background classification and domain adaptation".

[19] F.-C. Chen, K.-D. Chen and Y.-W. Liu, 2022. "Domestic sound event detection by shift consistency mean-teacher training and adversarial domain adaptation" *DCASE.*

[20] K. Saito, K. Watanabe, Y. Ushiku and T. Harada, 2018. "Maximum classifier discrepancy for unsupervised domain adaptation," in 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 3723-3732.

[21] N. Turpault, R. Serizel, A. Parag Shah and J. Salamon, 2019. "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events,* New York City, United States.

[22] J. Salamon, D. MacConnell, M. Cartwright, P. Li and J. P. Bello, 2017. "Scaper: A library for soundscape synthesis and augmentation," in 2017 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 344-348.

[23] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal and M. Ritter, 2017. "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017,* New Orleans, LA.

[24] L. Delphin-Poulat and C. Plapous, 2020. "Mean teacher with data augmentation for dcase 2019 task 4," DCase 2019 workshop.

[25] H. Nam, S.-H. Kim and Y.-H. Park, 2022. "Filteraugment: An acoustic environmental data augmentation method," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4308-4312.

[26] A. Mesaros, T. Heittola and T. Virtanen, 2016. "Metrics for polyphonic sound event detection," *Applied Sciences,* **6:** 162.

[27] H. Nam, S.-H. Kim, D. Min, B.-Y. Ko, S.-D. Choi and Y.-H. Park, 2022. "Frequency dependent sound event detection for dcase 2022 challenge task 4".

[28] Bilen, G. Ferroni, F. Tuveri, J. Azcarreta and S. Krstulović, 2020. "A framework for the robust evaluation of sound event detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 61-65.

The Journal of Acoustical Society of India : Vol. 52, No. 3, 2025 (pp. 93-108)

Human acoustic signature analyzer for acoustic audio analysis and classification using feature extraction and deep learning

Reshma P.[1*], S. Vijayan Pillai[2], Usha K.[3], Vipin S.S.[4] and Ajith S.[5]
Keltron, Trivandrum, Kerala, India
e-mail: reshmapnair333@gmail.com

[Received: 02-06-2025; Accepted: 09-08-2025]

ABSTRACT

The organs and mechanisms in human body generate sounds and vibrations called biomedical acoustic signals or Human Body Acoustics that reach the body surface through surrounding tissues, which can be converted into electrical signals by suitable sensors. This signal exhibits the state of the organs or mechanisms that generate it. The anomaly in these signals can be used to detect the functional irregularities, however, there exist challenges in the translation of this anomaly to a powerful and reliable clinical monitoring tool. The Human Body Acoustic Signature Analyzer is a comprehensive tool capable of capturing, storing, learning, analysing, and classifying acoustic signatures automatically. The tool uses feature extraction and deep learning techniques to predict functional abnormalities. The proposed system integrates a specialized acoustic sensor for data acquisition of normal and abnormal heart or lung acoustic signatures and extracts a wide range of features from the signals. The primary objective of the system is to bridge the gap between traditional stethoscope-based auscultation and modern digital analysis, providing a powerful tool that enhances the capabilities of medical practitioners and improves acoustic medical diagnosis. The comprehensive human acoustic system can be employed where limited access to specialists exists. This bridges the gap in rural area-based healthcare in a sustainable manner. The proposed system can assist medical personnel in facilitating more efficient and accurate diagnoses further leads to improved patient outcomes and more equitable distribution of healthcare.

1. INTRODUCTION

Traditional auscultation using a stethoscope lacks digital analysis, making it subjective and independent on practitioner experience. There is a need of an advanced system to capture, analyze, and classify heart and lung sounds for anomaly detection. The Human Body Acoustic Signature Analyzer system can be used to learn, analyze and classify acoustic signatures of a particular person or group of individuals and thereby predicting any deviation from the normal pattern. Analysis can be performed on individual subjects or on groups. The system level analysis tool also works with artificial intelligence or machine learning capabilities by analyzing acoustics related to cardiovascular system, respiratory system, gastrointestinal system, and nervous system[1]. The Human Body Acoustic signals encapsulate the functional state of the underlying organs and physiological mechanisms. Detailed analysis of these signals enables objective assessment of a patient's clinical condition. There is a requirement of having low-cost and user-friendly monitoring method, which makes it possible to have more frequent evaluation of the patient's status. There are different commercial monitoring devices like ECG machine being well in use.

The proposed system identifies the gap of an automated system in diagnosing human body acoustic signatures for disease prediction. Medical professionals in resource-limited settings can leverage the power of AI to gain valuable insights from body acoustic signatures and thereby enhances sustainable healthcare using advanced technology and in-depth acoustic insights. The proposed system primarily focuses on the preventive healthcare. The simple and early diagnosis treatments are possible with detecting the signs of the disease and thereby prevent major complications.

Acoustic signals including heart and lung signals can be collected using Human acoustic transducer and audio editor software. Audio Editor also helps in pre-processing the signals including deletion of selected portions, filtering, and amplification. During data acquisition, the captured signal can be visualized in the time-domain and played back real-time for verification. The collected human acoustic signals will be stored and given to analysis software. The signal will be analyzed against 22 different processes and the feature vectors are extracted. Among the feature values, best features are selected for further analysis. Machine learning and deep learning-based classification and prediction can be performed from the extracted features.

Artificial intelligence and deep learning have improved audio analysis and classification[5] and deep learning has the unique ability to learn complex patterns. In Deep learning-based approach, personal and clinical attributes of individuals or patients and the negative and positive samples will be analyzed intrinsically and the insights will be used for finding the variations in different patterns of normal and abnormal samples of acoustic signatures. This can be used to predict related diseases in the early stage itself. Deep learning models require large datasets for training the network and automatically learning features from each acoustic audio classes. The classification results affect how the classification task is chosen. Binary classification, multiple classification, and regression are mostly used in addressing classification problems[6]. In most cases, deep learning models cannot learn from completely arbitrary data and feature engineering can help in identifying and selecting features.

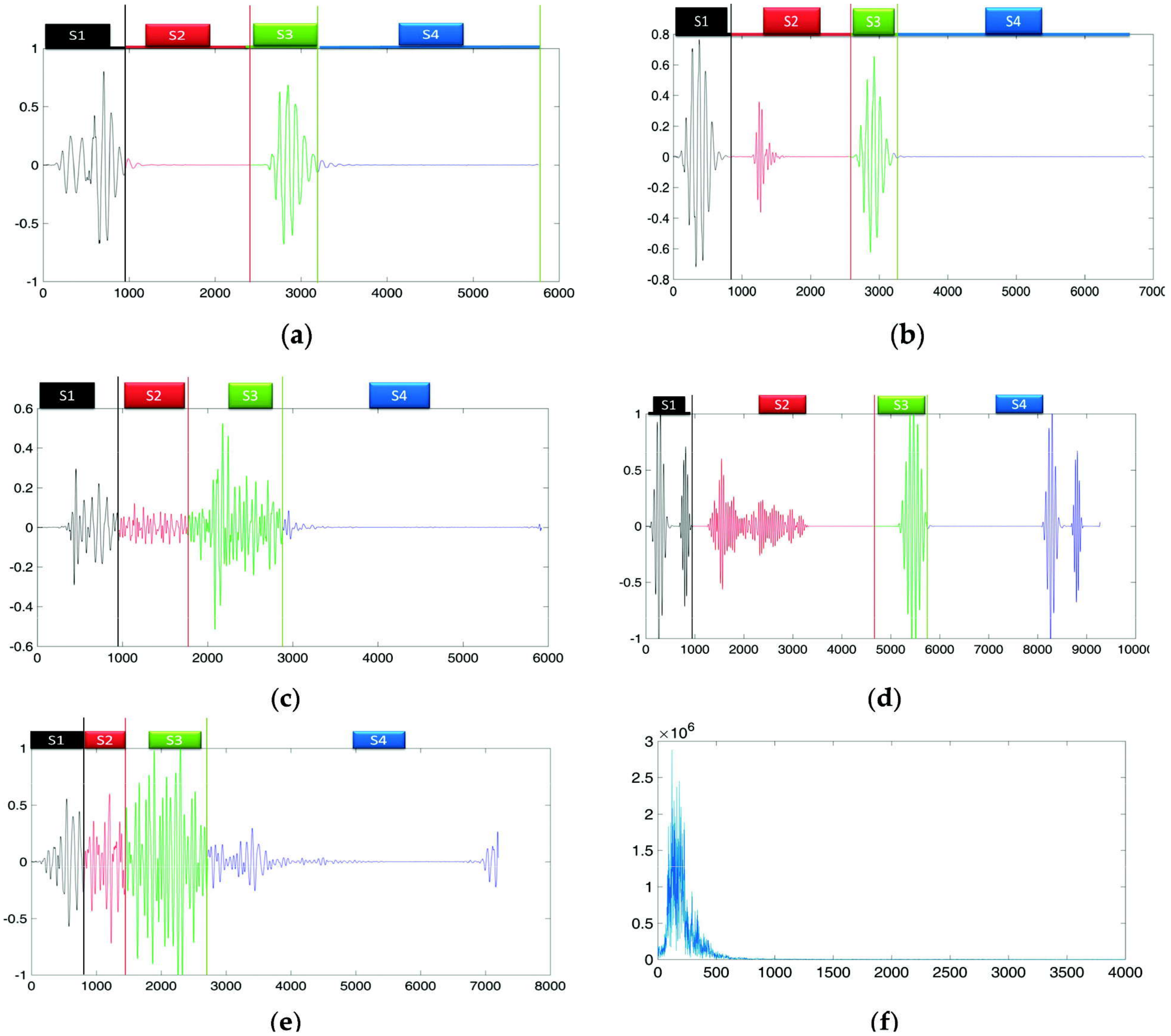
Acoustic audio analysis can be implemented in the diagnostics and preventive healthcare of cardiovascular, respiratory, and gastrointestinal diseases. Existing approaches can make use of acoustic signature classification techniques to assist the medical practitioners in daily clinical practice. The main scope of the proposed approach is to develop a system that can aid in classifying human body acoustics to address cardiovascular diseases, respiratory illnesses, and gastrointestinal diseases. The system is designed to provide a low-cost and user-friendly remote monitoring method, which makes it possible to have more-frequent evaluation of the patient's status. The proposed system employs an integrated approach to capture all biomedical acoustic signals with storage and analysis. In this paper, we mainly concentrate on the human heart acoustic signature analysis and prediction.

Diagnosis errors and delays harm over 43 million people annually, worsening patient outcomes and increasing healthcare costs. Missed early warnings, delayed tests, and slow specialist responses are major contributors. Effective healthcare systems could save up to 8 million lives each year. Early detection and continuous monitoring, though challenging outside clinical settings, are critical for managing chronic conditions and reducing hospital readmissions. AI in healthcare has the potential to revolutionize patient care, treatment planning, and drug development. However, challenges around data privacy, costs, and ethics must be addressed. Immediate adoption of AI-based healthcare solutions is essential to drive transformation.

In the coming sections we will discuss related works, proposed methodology, heart acoustic signature analysis and prediction, implementation details, experimental results, conclusion and future scope, and bibliography.

2. LITERATURE REVIEW

Several studies in the current technical literature have addressed the automatic detection of cardiac diseases[7]. Some researchers have evaluated different machine learning algorithms in predicting the presence of coronary artery disease. The authors proposed a predictive model for heart disease diagnosis using a fuzzy rule-based approach combined with decision trees. The human heart consists of two atrial chambers in the upper portion of the heart through which blood enters and two ventricular chambers in the lower portion of heart through which blood exits. The heart-related signal has frequency components from above 100 Hz and below 600 Hz. Figure. 1 illustrates different conditions of the heart sound signal in the temporal domain while Figure 1(f) is in its spectral domain[2].



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 1. Adopted from Ref.[2] (a) Normal Heart Sound (b) Murmur in Systole (c) Mitral Regurgitation (d) Mitral Stenosis (e) Aortic Stenosis (f) Spectrum of a PCG Signal.

For heart, acoustic signatures are valuable for diagnosing various cardiac conditions, fetal heart rate monitoring, automatic detection of heart murmurs and other heart diseases[7]. It is well known that respiratory system diseases are diagnosed by auscultation, percussion, and tactile fremitus using low-

Table 1. Respiratory Illnesses.

Diseases	Frequency range in Hz
Pneumonia	300 - 600
Asthma	165, 239, 329
Chronic obstructive pulmonary disorder	233-311
Pneumothorax	400-600

cost acoustic procedures. Current sensor technology and computational analysis methods enable to measure, analyze and classify lung acoustic signals that originate internally, such as breath or vocal sounds[3]. Table 1 illustrates the indicative frequency range for different diseases[13].

Digestive system generates sounds and vibrations called bowel sounds[4]. There are four categories based on their waveform and frequency content Single Burst (SB), Multiple Bursts (MB), Continuous Random Sound (CRS), and Harmonic Sound (HS). Table 2 illustrates the spectral features of different types of bowel sounds. These medical conditions include abdominal conditions such as appendicitis, intestinal obstruction, irritable bowel syndrome, inflammatory bowel disease and pyloric[14].

Table 2. Gastrointestinal Tract Diseases.

Type	Duration (mSec)	Spectral Centroid (Hz)	Spectral Flatness	Quantities	CIT Ratio
SB	18-58	347-681	0.056-0.39	240000	> (1.7-5)
MB	100-1030	345-753	0.073-0.47	2237	0.8-(1.7-5)
CRS	119-1637	316-609	0.026-0.37	836	0.2-0.8
HS	73-763	269-630	0.018-0.43	214	0-0.2

3. PROPOSED METHODOLOGY

The Human Acoustic Signature Analyzer identifies the need of an automated system in remote healthcare for diagnosing cardiovascular, respiratory, and gastrointestinal diseases. The system integrates a specially designed human acoustic data acquisition system consists of a human acoustic transducer device and a dedicated data collection software Audio Editor which makes data collection process more précised manner aiding an automated acoustic signature classification technique to help medical practitioners in daily clinical practice. The power of artificial intelligence and machine learning is leveraged to gain valuable insights from heart and lung acoustic signatures. The system architecture of the Human Body Acoustic Signature Analyzer is plotted in the Figure. 2.

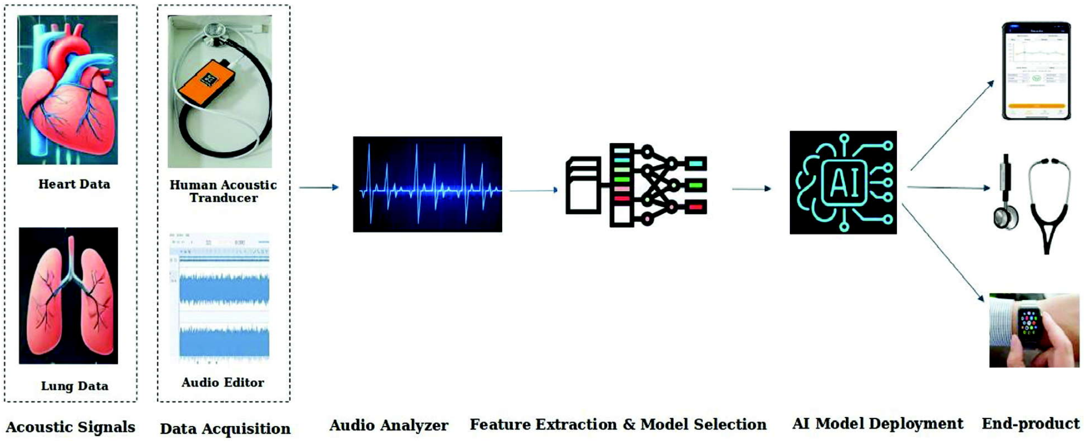


Fig. 2. System Architecture of the Proposed System.

The proposed system receives human acoustic signatures including heart, lungs or gastrointestinal signals collected using human acoustic transducer and audio editor software. During data acquisition, real-time plotting and playback makes easier to select the more precision signal and preprocessing is also performed. For feature extraction and analysis, a specialized Signal Analyzer software is employed where time-domain, frequency-domain and time-frequency domain features are extracted. This analysis helps in identifying the most essential and accurate features of human acoustics. With the selected features extracted, the human acoustic signals are fed into a deep learning model for analysis and prediction. Being the most essential and accurate application, classification and prediction using artificial intelligence serves mostly in predictive diagnostics and healthcare. Wide range of applications possible in classification and prediction of human acoustic signals and the product can be a mobile application or a smart stethoscope with AI capability or even a smart watch aiding real-time health monitoring and prediction. In the coming section we will discuss more on the heart acoustic signal acquisition, preprocessing, feature extraction and analysis.

4. HEART ACOUSTIC SIGNATURE ANALYSIS AND PREDICTION

According to the Centre for Disease Control and Prevention every 37 seconds one person loses his life due to cardiovascular diseases. Heart sound analyzer system that acquires, records, and analyzes the acoustic signals of the heart and delivers the results in an easy-to-interpret graphical display. It also supports physicians in analyzing their patients' heart sounds for suspected murmurs. Cardiovascular diseases are the leading cause of death worldwide, making early detection and continuous monitoring of heart function essential. Auscultation, the traditional method of listening to heart sounds via a stethoscope, is non-invasive but highly subjective, relying on clinician expertise. Advances in digital stethoscopes, signal processing, and machine learning now enable objective, reproducible analysis of heart sounds.

A system with laptop, electronic stethoscope and a printer can assist the doctors in real-time heart analysis. Data was taken from multiple heart locations of 167 individuals. The proposed system can assist physicians by providing feedback through the GUI, aiding the detection of murmurs as potential indicators of heart disease. The heart produces distinct sounds at five auscultation points aortic, pulmonic, erb's point, tricuspid, and mitral, each reflecting different physiological functions. However, it remains unclear which site provides the most predictive diagnostic value, particularly across demographics like age, gender, and medical history. Identifying the most informative auscultation point can enhance automated diagnostic models and guide clinicians toward more effective, targeted screening.

When you classify heart sound signals into normal and any abnormal sounds, based on murmur, they are classified into Ventricular Septal Defect (VSD), Aortic Stenosis (AS), AI Murmur, Pulmonic Stenosis (PS) Murmur and Innocent murmur. For the signal acquisition using auscultation procedure, we have used Human Acoustic Transducer device. The signal processing in such devices involve three phases. Analog signal processing converts acoustic sounds from the patient's body into electronic signals using microphones. As a preprocessing stage, acoustic signals are filtered and amplified first to remove the noise elements. For heart acoustic analyzer a specific frequency ranges from 20 to 150 Hz and for lung acoustic analyzer, a frequency range from 50 to 2500 Hz are considered for processing and the cleaned signals can be converted from analog to digital.

For predictive modelling or deep learning applications, it is recommended to prioritize auscultation data from the mitral area as the primary source due to its high diagnostic relevance in capturing low-frequency cardiac events, including those associated with mitral valve pathologies. Erb's point is suggested as the secondary choice, given its central location and utility in detecting both aortic and pulmonary murmurs, thereby offering a comprehensive acoustic representation of cardiac activity. Prioritizing these anatomical sites enhances the likelihood of capturing diagnostically significant features for model training and classification. There is a critical scarcity of large-scale, open-source heart sound datasets that are publicly available for researchers to develop and benchmark automated diagnostic algorithms. Furthermore, there exists no dedicated open accessed heart sound dataset that reflects regional-level or

Indian demographic and clinical diversity which hinders the development of population-specific diagnostic models, which are crucial for effective deployment of AI-driven cardiac screening tools in regional healthcare systems.

5. IMPLEMENTATION DETAILS

Vibration based human acoustic transducer is employed to pick up the sound and vibrations from the human body. The output from the transducer is subjected to pre-conditioning to eliminate bias and false alarms. The pre-conditioned output is processed to extract physical and statistical properties parameters. These parameters are used to build a data library and further analysis. The results are presented in the form of graphs and listing binary results. Feature engineering process of acoustic signal is depicted in Figure 3.

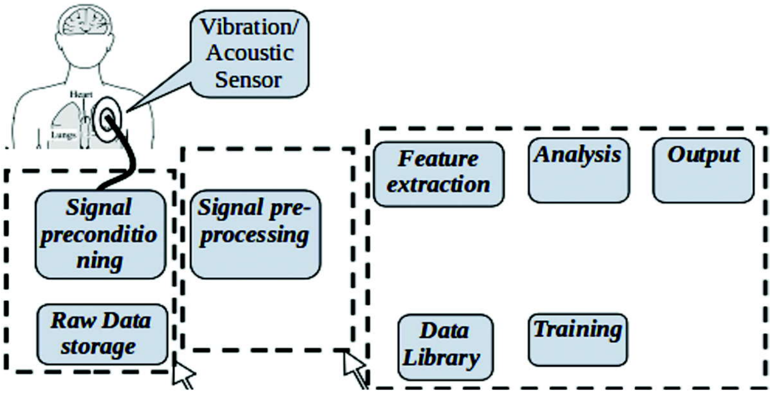
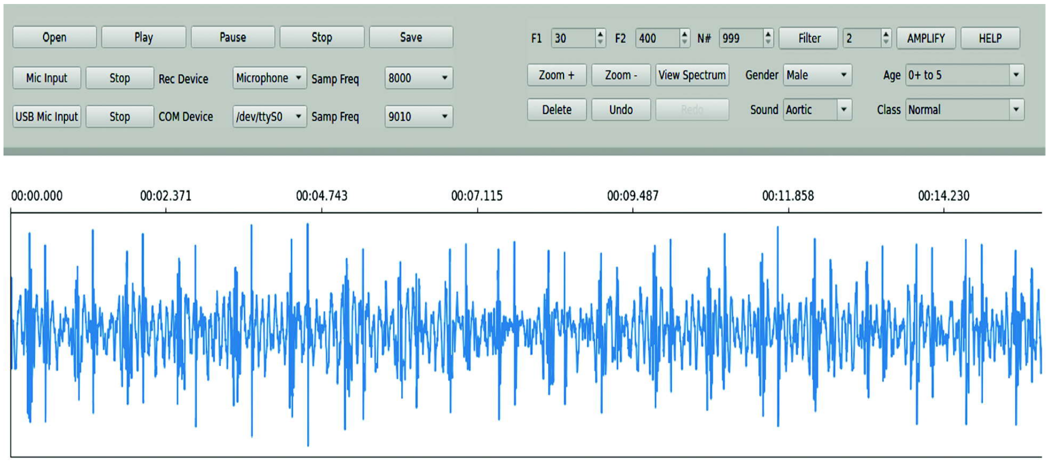


Fig. 3. Feature Engineering in Human Acoustic Signature Analysis.

Total system has hardware for precision data collection software module running in Windows or Linux OS. Signal sensing and preconditioning is done in hardware, which is controlled by audio editor software as shown in Figure 4. Through audio editor, raw data storage in an organised manner is carried out. The audio editor facilitates opening and saving audio files in audio editor, time series data visualization, audio playback options, recording, audio cropping, high pass or low pass filtering, undo or redo operations, time series graph, zoom in and zoom out feature and spectrum analysis.



Rec Device Microphone Samp Freq 8000
COM Device /dev/ttyS0 Samp Freq 9010
F1 30 F2 400 N# 999 Filter 2
Gender Male Age 0+ to 5
Sound Aortic Class Normal

Fig. 4. Audio Editor Interface.

The signal processing, feature extraction, analysis, data library creation training and audio-visual presentation is done by Analyzer software as shown in Figure 4. Analyzer facilitates the extraction of LOFAR, DEMON, Spectrum, Cepstrum, Spectral Centroid, Spectral Spread, Spectral Skewness, Spectral Crest Factor, Spectral Flatness, Spectral Flux, Spectral Decrease, Spectral Kurtosis, Spectral Slope, Spectral Roll off, Spectral Entropy, Spectral Pitch Chroma, Spectral Tonal Power Ratio, Standard Deviation, Root Mean Square Value, Zero Crossing Rate, Peak Envelope, Autocorrelation and Maximum Autocorrelation. It also supports computation of statistical parameters like Maximum (Max), Minimum (Min), Mean, Median and Standard Deviation.

5.1 Data Acquisition

A high-performance medical acoustic data acquisition system is designed to collect the acoustic signatures from the body. The Human Acoustic Transducer module is a non-invasive and high-precision device which can sense the acoustic signals from heart, lungs, bowels, joints, and other parts. For the data acquisition of heart acoustic signatures, data points from five auscultation points of heart including the aortic area, pulmonic area, erb's point, tricuspid and mitral area are concentrated. The signal processing in an electronic stethoscope involves three main phases analog, digital, and AI-based processing. Initially, analog signal processing converts acoustic sounds from the patient's body into electronic signals using microphones. These signals are then filtered and amplified to remove unwanted noise and the cleaned signals are then forwarded to an Analog-to-Digital Converter (ADC) for digital conversion.

The amplitude of heart sounds in digital recordings depends on the microphone gain, the environmental noise and the subject's body characteristics. Selecting ideal sample rate is very crucial in data acquisition phase. For standard clinical analysis, 1000 Hz or 1 KHz is mostly chosen to capture pathological murmurs. For phonocardiography, 4 KHz is essential and for high fidelity, future proofing and AI based diagnostics, 8000 Hz is ideal. If ultra-high precision is needed for high resolution auscultation, then 16 KHz is used. There are 167 healthy human subjects of Indian origin were involved in the signal acquisition process. Normal healthy human subjects between the age group 18 to 56 were participated. Human Acoustic Transducer with sampling rate 8000 Hz was used for data acquisition.

The device was tested at two sampling rates 9010 Hz and 8000 Hz to evaluate frequency resolution and signal fidelity. Recordings were conducted in a soundproof and acoustically treated room to eliminate environmental noise. A standardized protocol was followed for all subjects, with data collected from five auscultation points including Aortic, Pulmonic, Erb's point, Tricuspid, and Mitral areas. This multi-point approach ensured comprehensive coverage of heart sound characteristics. The dataset was designed to capture substantial demographic variability, balanced class distributions, and clinically meaningful recording conditions. Informed consents were taken from each human subjects involved in the data acquisition phase. During data collection procedure objective and subjective features are collected. Objective features including age, gender, height, weight and subjective features including the presence or absence of cardiovascular diseases. Features including medical history like blood pressure, glucose, cholesterol, breathlessness, palpitations, and surgery related history also collected from the human subjects before auscultation procedure.

The heart sound dataset was collected from a total of 167 human subjects, including male and female representing a wide range of age groups and health conditions. Subjects were categorized into different classes including normal and multiple classes of other health conditions. Data acquisition was performed under controlled clinical conditions to ensure consistency and minimize noise interference. A human acoustic transducer, specifically engineered for biomedical acoustic signal capture, was used for recording.

5.2 Data Preprocessing

The captured acoustic signal is subjected to pre-processing, which involves a series of steps aimed at refining the raw audio data to ensure it is ready for detailed analysis. One of the first tasks in pre-processing is to remove any extraneous noise that may have been captured during recording. Figure-5 illustrates the outcome of pre-processing of heart acoustic signal. Pre-processing applies band-pass filtering to remove

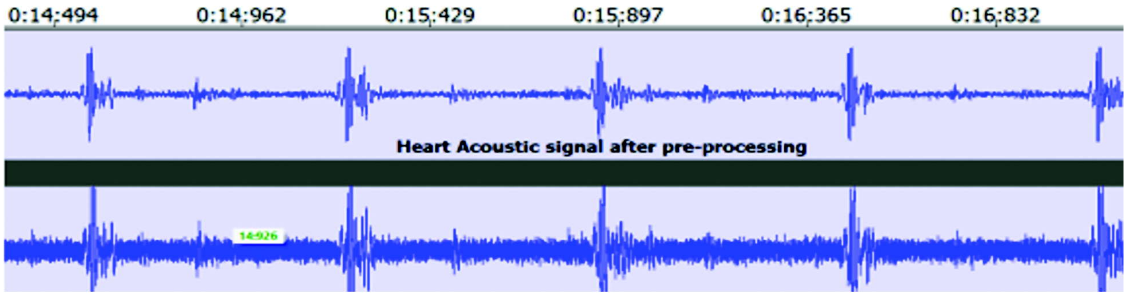


Fig. 5. Heart Signal after pre-processing.

noise, amplification to strengthen the signal, segmentation to split heart sounds and prepare data ready for feature extraction.

5.3 Feature Extraction, Learning and Feature Reduction

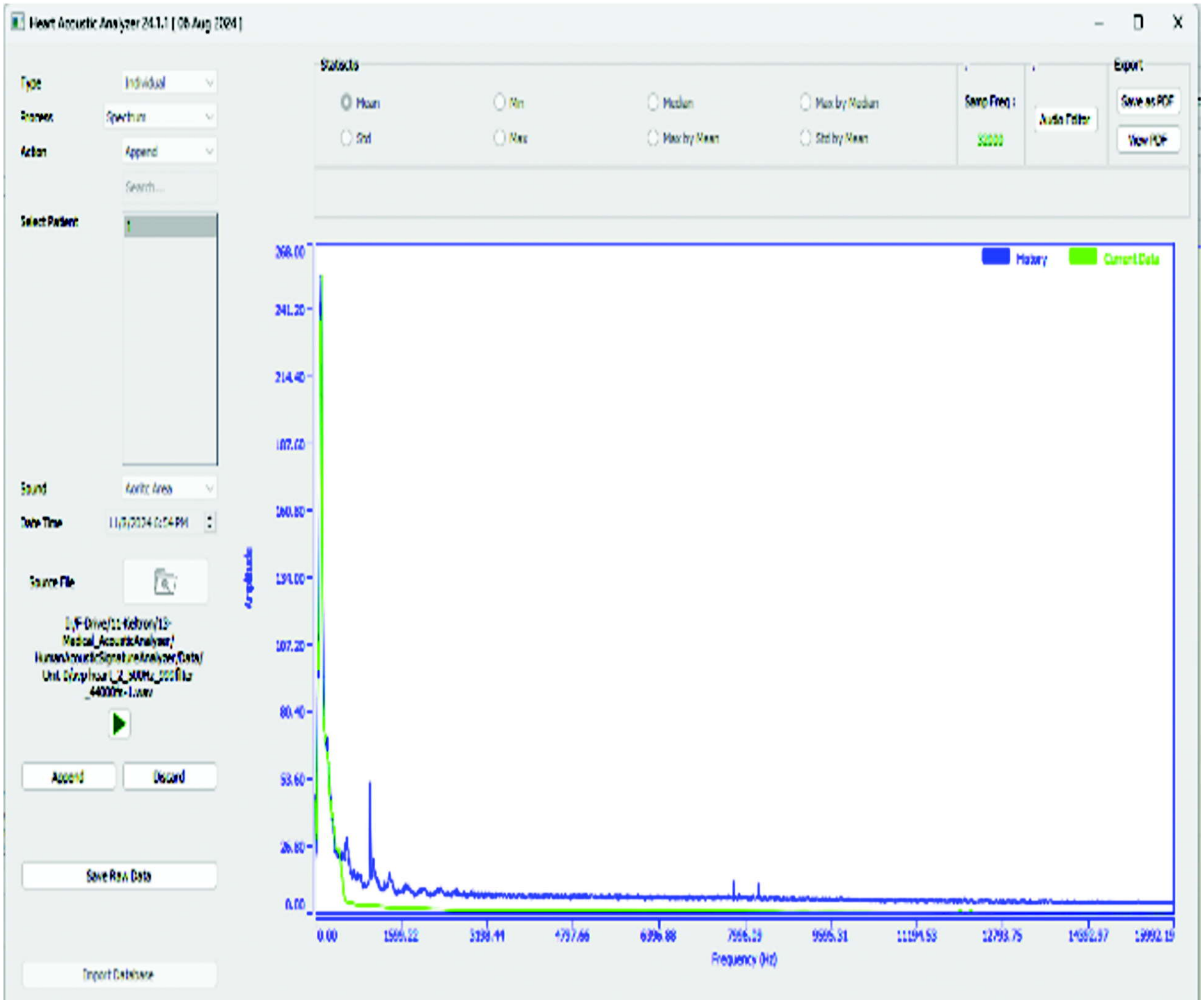
The parameters which have the potential to discriminate between various classes are identified, a few features are selected among them and the redundant features are removed. Feature reduction also reduces the dimensionality and computational load. Feature extraction technique involves twenty-two processes and eight statistical computations comprising of min, max, mean, and standard deviation analysed in two modes of operations, individual and group mode. The individual mode helps to track variations in the acoustic signature of a particular individual and the group mode helps to build an acoustic signature database of people in the same category and allows testing a new individual against the learned database.

The feature extraction process includes Mel-frequency Cepstral Coefficients (MFCC), Spectrum, Cepstrum, DEMON, LOFAR, and other spectral and temporal features. Spectral features including Spectral Centroid, Spectral Spread, Spectral Skewness, Spectral Crest Factor, Spectral Flatness, Spectral Flux, Spectral Decrease, Spectral Kurtosis, Spectral Roll Off, Spectral Slope, Spectral Entropy, Spectral Pitch Chroma, and Spectral Tonal Power.

Temporal features include Time Std, Time RMS, Time Zero Crossing and Time Peak Envelope. This extraction process involves advanced signal processing techniques to filter out redundancy of the audio signal, retaining only the core data that encapsulates the critical information. Data compression technique is also included wherein, instead of storing the entire audio waveform, parameters such as MFCC, statistical and physical features are stored. This compression, which stores only the key data, not only reduces the overall storage requirements but also enhances the performance of the system by making data retrieval and processing faster and more efficient. The compressed data retains the essential information needed for subsequent analysis, allowing the Human Acoustic Analyzer to perform its functions effectively.

Heart Sound Analysis helps in detecting normal and any variations in heart sounds. Low frequency peak from 10 to 200 Hz represents normal heart sounds and high frequency components from 200 to 1000 Hz indicate murmurs, valve issues or other abnormalities. Spectral differences help in diagnosing heart diseases. Amplitude spectrum or linear scale spectrum shows the magnitude of each frequency component in a signal and computed using FFT. Logarithmic or dB Spectrum converts the amplitude into a logarithmic scale and useful for detecting weak signals that are not visible in a linear scale. Figure 7 illustrates the filtered heart sound signal and its spectrum in linear scale and dB scale.

During feature extraction and analysis, a comparative study on the features of human acoustic transducer extracted signals and PhysioNet heart sound signals were analyzed. In baseline comparison, the abnormal heart beats have noise like feature in between every beat and comparatively higher frequencies than normal heart beats, while normal looks like more regular than any unusual heart beats and silence in between every beats. Figure 8 and 9 shows feature learning of normal and murmur signal in time-domain and frequency domain.



Heart Acoustic Analyzer 24.1.1 [08 Aug 2024]

Type: Individual

Process: Spectrum

Action: Append

Select Patient: 1

Sound: Aortic Area

Date Time: 11/7/2024 [illegible] PM

Source File: [illegible]

Statistics: ● Mean ○ Min ○ Median ○ Max by Median ○ Std ○ Max ○ Max by Mean ○ Std by Mean

Samp Freq : 32000

Fig. 6. Signal Analyzer Interface.

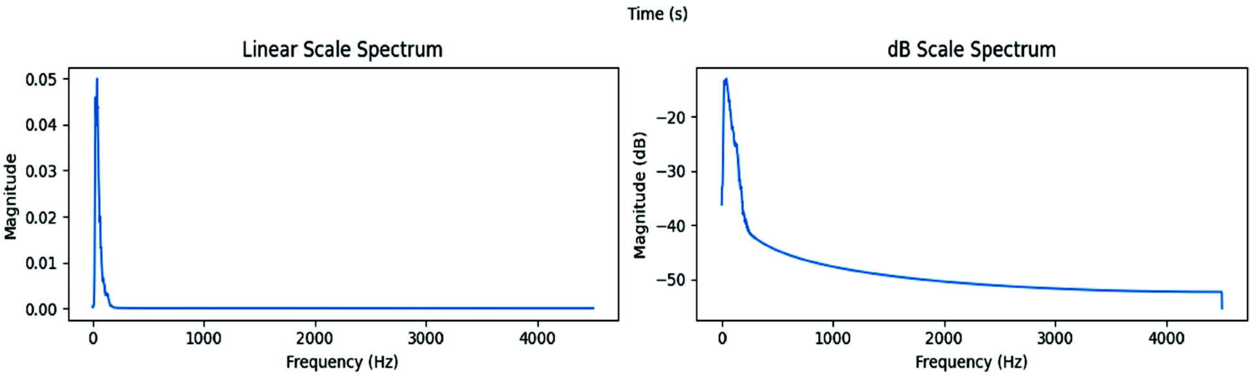


Fig. 7. Linear Scale and dB Scale Spectrum of Filtered Heart Signal.

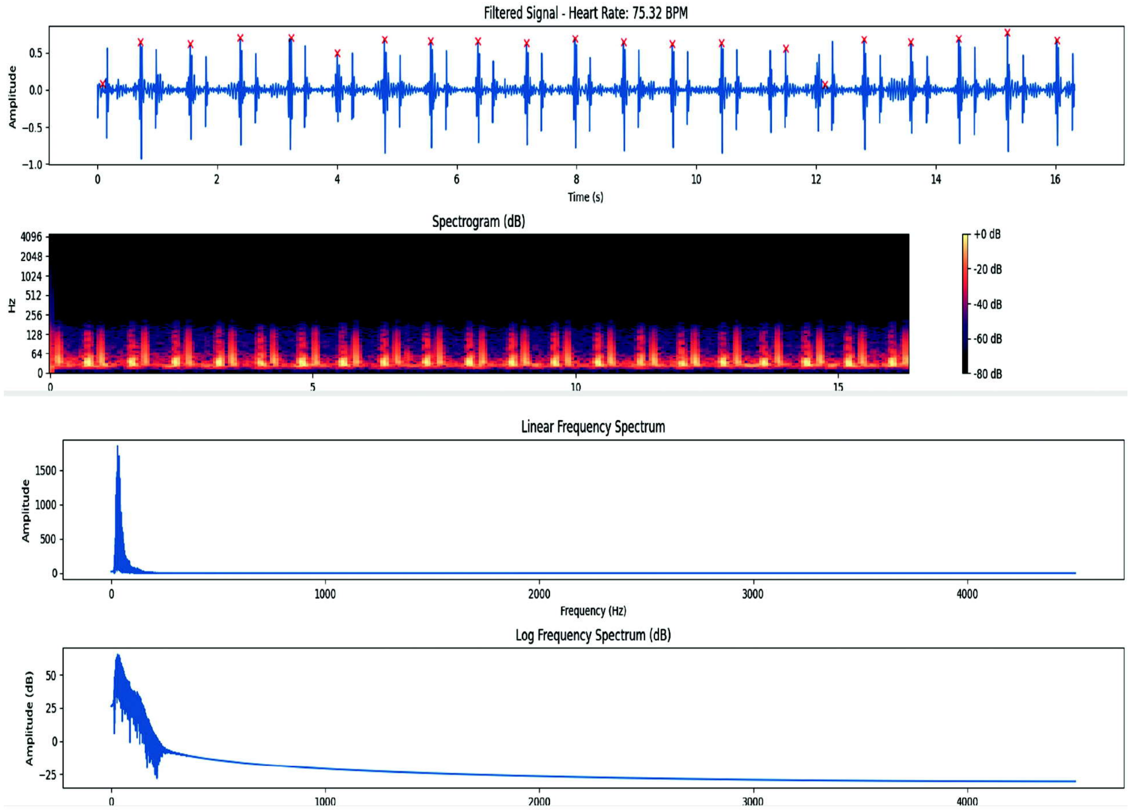


Fig. 8. Time-domain and Frequency domain analysis of a Healthy Heart Signal.

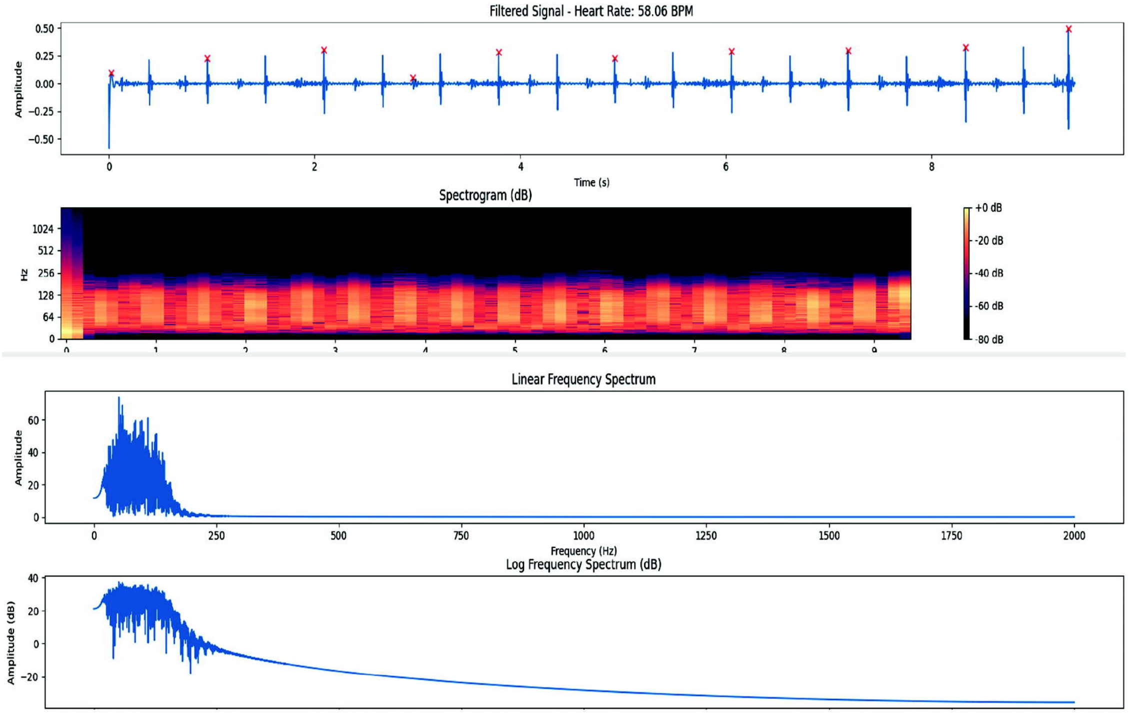


Fig. 9. Time-domain and Frequency-domain analysis of Heart Signal with murmurs.

For a quantitative evaluation of heart sound auscultation points with respect to gender-based acoustic variability, it is essential to analyze and compare signal characteristics across different anatomical recording sites. This approach enables the identification of auscultation locations that provide superior signal clarity and are more predictive for disease classification tasks.

Table 3. Auscultation Points.

Data Points	Locations	Associated Disease
Aortic	Right 2nd intercostal	Aortic stenosis, aortic regurgitation, systolic murmurs
Pulmonic	Left 2nd intercostal	Pulmonary stenosis, pulmonary hypertension, split S2 sounds
Erb's Point	Left 3rd intercostal	Best for aortic regurgitation, early diastolic murmurs, and general heart sound clarity
Tricuspid	Left 4th intercostal	Tricuspid regurgitation, ventricular septal defects
Mitral	Left 5th intercostal midclavicular	Mitral stenosis, mitral regurgitation, S1 clarity, apex beat

Heart beat rate of a newborn up to 4 weeks is assumed as 100 to 205 BPM, and for infants up to 1 year, the heart beat rate is between 100 to 180 BPM. Toddlers between 1 and 3 years will have 98 to 140 BPM and up to 5 years will have 80 to 120 BPM. School students between 5 to 12 years normally have 75 to 118 BPM and adolescents between 13 to 18 years have 60 to 100 BPM. Based on heart rate variation of individuals, several heart conditions can be identified. In some cases, individuals may report experiencing heart palpitations throughout the day. While most palpitations are brief and harmless, persistent symptoms warrant medical evaluation. Additionally, heart palpitations are relatively common during pregnancy due to an increase in heart rate and blood volume necessary to support fetal development.

Table 4. Quantitative Comparison of Heart Auscultation Points.

Frequency Range	Condition	Possible Disease
10-200 Hz	Normal heart sounds (S1, S2)	Healthy heart
200-400 Hz	Systolic Murmur	Aortic Stenosis, Mitral Regurgitation
400-800 Hz	Diastolic Murmur	Mitral Stenosis, Aortic Regurgitation
800-1000 Hz	High-frequency murmurs	Valve defects, Structural abnormalities

For a quantitative evaluation of heart sound auscultation points with respect to gender-based acoustic variability, it is essential that the following audio features should be extracted and analyzed at each auscultation sites including aortic, pulmonic, erb's, tricuspid and mitral for both male and female subjects. Peak amplitude captures the maximum instantaneous signal value, providing insight into the loudness and signal strength at each location. Root Mean Square Energy reflects the overall energy content of the signal, which can help differentiate between strong and weak heart sound recordings. Signal-to-Noise Ratio quantifies the proportion of useful heart sound signal relative to background noise, aiding in assessing the recording quality and Spectral Power represents the mean power distributed across frequency bins, allowing for a frequency-domain comparison of acoustic energy across sites. By statistically comparing these features across gender and auscultation points, one can determine which locations offer the most diagnostically valuable and acoustically clean signals, thereby informing both clinical practice and model input selection for heart sound analysis systems.

5.4 *Experimental Results*

Heart acoustic signatures of aortic, pulmonic, erb's, tricuspid and mitral are given for training. The below plots in Figure 10 provide insights into how well each individual feature can separate the five heart auscultation points. For mfcc_1_mean, all classes overlap to some extent, but pulmonic and erb's cluster around slightly different mean ranges. The tight distribution of this feature suggests it is a stable characteristic across recordings.

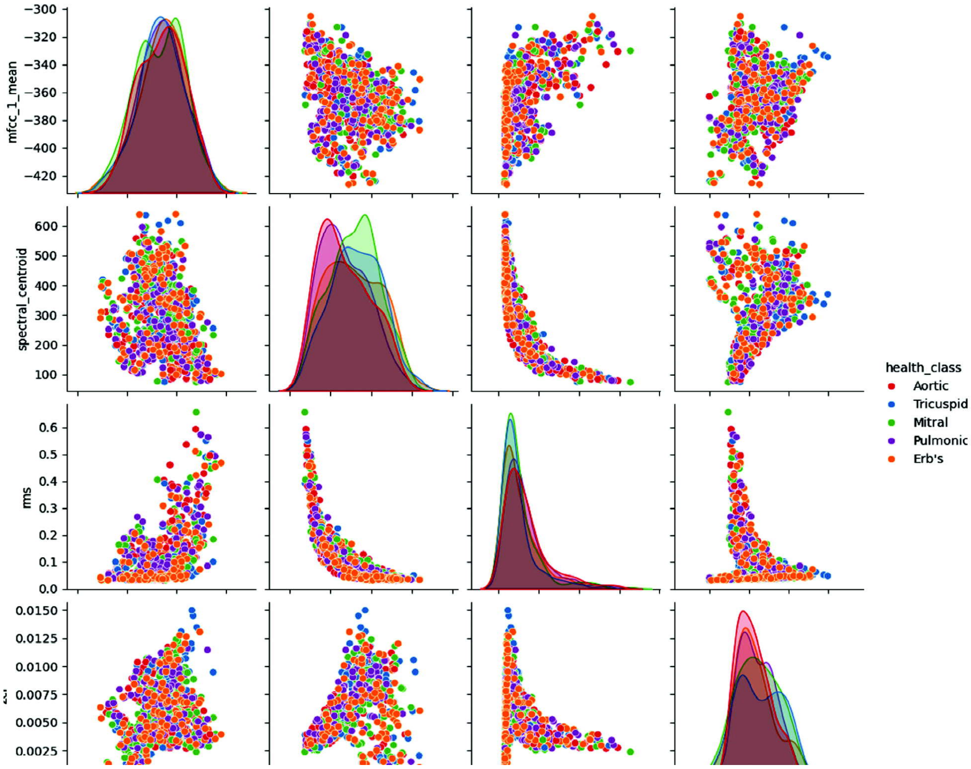
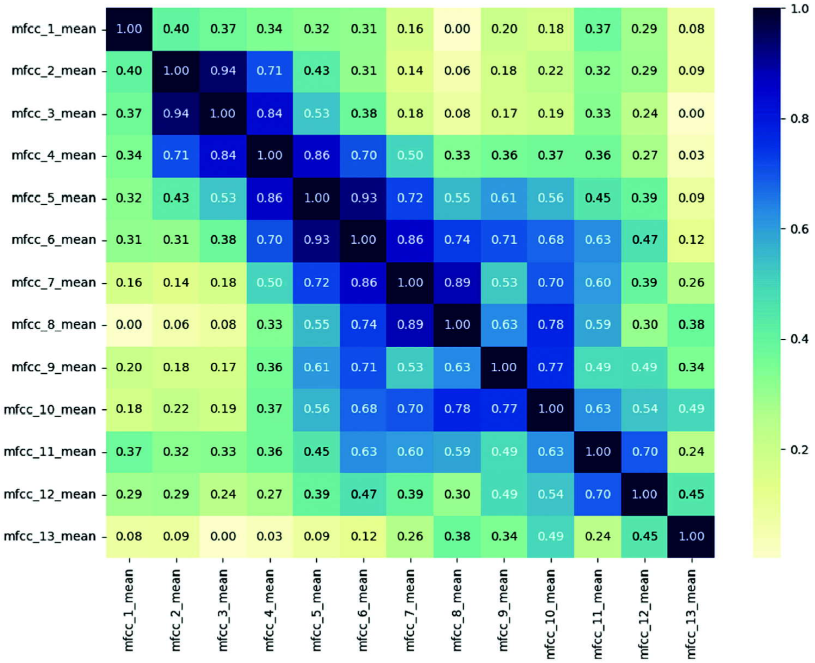


Fig. 10. Multivariate Feature Analysis and Class Separability in Heart Sounds.

The spectral centroid[7] feature shows that aortic and pulmonic lean toward lower centroid values, offering limited but possible separation. The rms feature exhibits distinctive patterns among classes, with mitral showing a notably higher energy range, making it a particularly useful feature, especially in combination with others. For Zero Cross Rate (zcr), aortic and pulmonic have distinct peaks, indicating a moderate degree of class separability. Figure 11 gives the view of correlation matrix of MFCC mean features.

The off-diagonal scatter plots reveal interactions between feature pairs and how they relate to different heart classes. In the mfcc_1_mean vs spectral centroid plot, scattered clusters emerge, with erb's and pulmonic forming more distinct boundaries. The mfcc_1_mean vs rms plot shows



	mfcc_1_mean	mfcc_2_mean	mfcc_3_mean	mfcc_4_mean	mfcc_5_mean	mfcc_6_mean	mfcc_7_mean	mfcc_8_mean	mfcc_9_mean	mfcc_10_mean	mfcc_11_mean	mfcc_12_mean	mfcc_13_mean
mfcc_1_mean	1.00	0.40	0.37	0.34	0.32	0.31	0.16	0.00	0.20	0.18	0.37	0.29	0.08
mfcc_2_mean	0.40	1.00	0.94	0.71	0.43	0.31	0.14	0.06	0.18	0.22	0.32	0.29	0.09
mfcc_3_mean	0.37	0.94	1.00	0.84	0.53	0.38	0.18	0.08	0.17	0.19	0.33	0.24	0.00
mfcc_4_mean	0.34	0.71	0.84	1.00	0.86	0.70	0.50	0.33	0.36	0.37	0.36	0.27	0.03
mfcc_5_mean	0.32	0.43	0.53	0.86	1.00	0.93	0.72	0.55	0.61	0.56	0.45	0.39	0.09
mfcc_6_mean	0.31	0.31	0.38	0.70	0.93	1.00	0.86	0.74	0.71	0.68	0.63	0.47	0.12
mfcc_7_mean	0.16	0.14	0.18	0.50	0.72	0.86	1.00	0.89	0.53	0.70	0.60	0.39	0.26
mfcc_8_mean	0.00	0.06	0.08	0.33	0.55	0.74	0.89	1.00	0.63	0.78	0.59	0.30	0.38
mfcc_9_mean	0.20	0.18	0.17	0.36	0.61	0.71	0.53	0.63	1.00	0.77	0.49	0.49	0.34
mfcc_10_mean	0.18	0.22	0.19	0.37	0.56	0.68	0.70	0.78	0.77	1.00	0.63	0.54	0.49
mfcc_11_mean	0.37	0.32	0.33	0.36	0.45	0.63	0.60	0.59	0.49	0.63	1.00	0.70	0.24
mfcc_12_mean	0.29	0.29	0.24	0.27	0.39	0.47	0.39	0.30	0.49	0.54	0.70	1.00	0.45
mfcc_13_mean	0.08	0.09	0.00	0.03	0.09	0.12	0.26	0.38	0.34	0.49	0.24	0.45	1.00

Fig. 11. The Correlation Matrix of MFCC Mean Features.

clearer clustering, particularly for the mitral class, which trends toward the upper-right region, likely reflecting stronger valve sounds. There is some correlation between spectral centroid and zcr, where erb's and pulmonic tend to occupy distinct zones. In the rms vs zcr plot, mitral and aortic can be reasonably separated, although erb's and pulmonic remain more intermingled.

5.5 The Baseline comparison of different Classification Algorithms

A classification model trained on Mel-frequency Cepstral Coefficients (MFCC) extracted from audio samples of four distinct classes of aahh sound, breath sound, heart sound and cough sound. Each audio class contains 300 audio samples and the dataset is split into 80% training data and 20% test data. An ML classifier trained on the extracted MFCC features can be used to learn the patterns specific to each class. The model's performance is evaluated using a confusion matrix which shows strong classification results across all the classes. The results indicate a high level of accuracy and minimal confusion between classes. The model strength indicates that MFCC features are effective for capturing the unique acoustic characteristics.

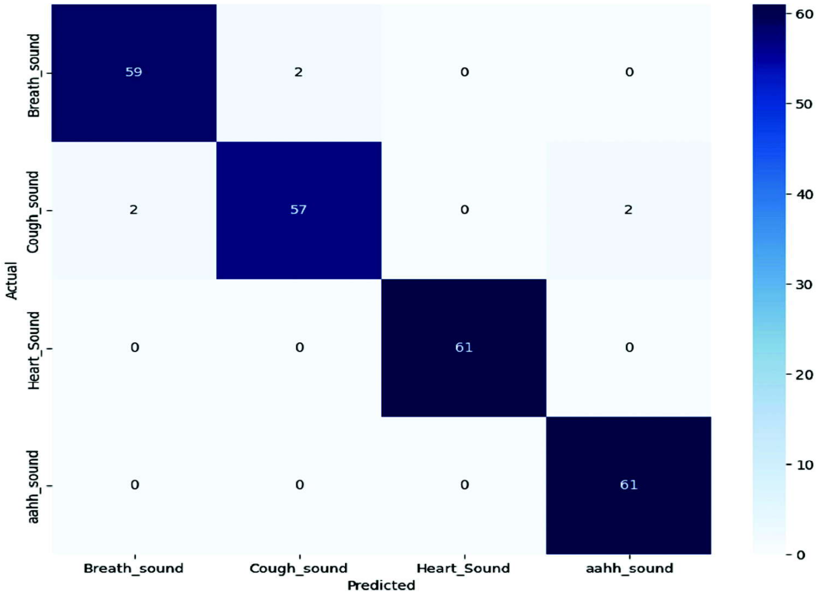


Fig. 12. Confusion Matrix.

To compare accuracy of different training models against MFCC features. The below bar chart in Figure 14 illustrates the accuracy of five models SVM, KNN, LSTM, Random Forest, and CNN on four different health classes. CNN and SVM shows better performance achieving 97% accuracy. Random Forest and LSTM shows good consistency and gives 95% while KNN gives only 93% on the real data samples. SVM is a good choice when speed and interpretability matter and CNN[10] is highly effective and scalable for deep learning workflows.

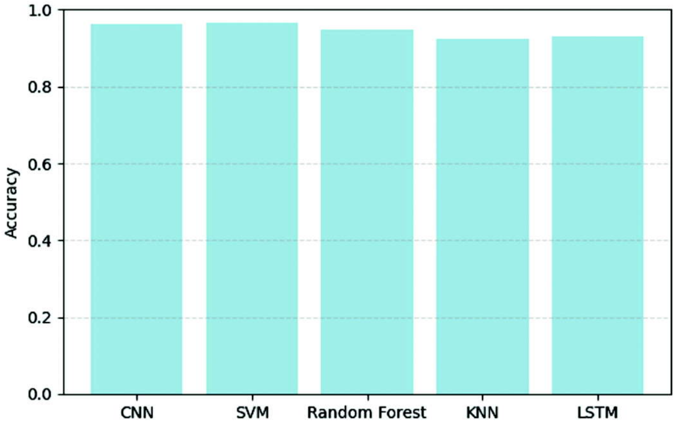


Fig. 13. Accuracy comparison of classifiers on MFCC Features.

5.6 Deep Learning based Classification Models

The dataset collected were separated into training, testing and validation sets to ensure the model efficiency. For the proposed approach, the model architecture selected was Convolutional Neural Network (CNN) and the model was trained by adjusting the weights and fine tuning of hyper parameters. After the completion of training process, the trained model will be saved and tested on a separate test dataset to calculate the performance evaluation metrics. Periodic training and continuous monitoring were enabled to ensure the model remains effective in classifying and predicting the class of signals. For the acoustic signature classification, the signatures can be analysed by using Mel Frequency Cepstral Coefficients[9] or Spectrograms and that can extract intrinsic signal features and translates them to 2D spectrum of audio signatures.

The deep learning-based acoustic signal classification aimed to identify the intrinsic features of body acoustic signatures including human voice, heart and lungs and the model would be able to find different variations and predict the category of diseases by analysing the complex patterns generated from human acoustic signature samples which was fed into the deep learning network. To train a highly precisioned acoustic classifier model, more quality and quantity data needed and based on that the model parameters can be finetuned. Mitral recordings tend to produce clearer waveforms, better MFCC or Spectrogram features, and higher classification accuracy in many deep learning models. Erb's point is also excellent due to minimal interference and overlapping valve sounds. Mitral and Erb's point are often the most informative for general heart sound classification. Mitral area captures apical sounds and Erb's Point gives a balanced overview of both aortic and pulmonic sounds and is great for diastolic murmurs[8].

For heart acoustic signal classification, we have collected signals from five data points including aortic, pulmonic, erb's point, tricuspid and mitral area. Initially, a CNN based classification was performed using a network trained on Mel-Frequency Cepstral Coefficients and Spectrograms, which are extracted features from heart sound recordings. The CNN learns patterns in the MFCC and spectrogram features to differentiate between various heart acoustic signatures. To ensure the model performance, class weights are computed and used during training to address data imbalance. The model is evaluated using a confusion matrix and classification report that shows how well the model predicts each class, along with metrics like precision, recall, and F1-score.

MFCC-based loss trends confirm severe overfitting in the model and the training loss drops sharply to nearly zero, indicating that the model is fitting the training data extremely well. However, the validation loss does not follow the same trend and it stabilizes at a significantly higher value and even slightly increases over epochs. This divergence between training and validation loss highlights the model's inability to generalize, suggesting it is memorizing training examples rather than learning meaningful patterns applicable to new data. Addressing this requires better data balancing across auscultation points and gender, regularization techniques like dropout and L2 penalties, and more robust validation strategies such as early stopping, data augmentation, or adopting more suitable architectures like CNN-LSTM for temporal feature extraction. Both MFCC and spectrogram models show overfitting and poor validation due to data imbalance, and limited model architecture. To address this, applied dropout, batch normalization, L2 regularization, early stopping, and feature standardization.

Deep learning models are designed to automatically learn features from raw data. Manual feature extraction in deep learning can hamper the efficiency of deep learning models[11]. For deep learning better approach is to train raw waveform or spectrograms or mel-spectrograms to preserve both time and frequency domain information leading to improved performance in deep learning models.

6. CONCLUSION AND FUTURE SCOPE

This paper presents a system for non-invasive biomedical acoustic monitoring that aims to capture heart, lung, and bowel sounds using specialized sensors. In heart acoustic analyzer, the system extracts a wide range of features from the time, frequency, and time-frequency domains and employs extensive feature extraction and deep learning techniques to analyse heart acoustic signals. The goal is to develop a low-cost, automated diagnostic tool, particularly suited for use in resource-limited settings. The proposed system bridges the gap between traditional auscultation and modern diagnostics.

The system captures, processes, and analyses the acoustic signatures produced by the human body. Extensive research on the requirement and social relevance of a stand-alone system for heart acoustic signal analysis was identified. By incorporating deep learning models [12], the system could interpret the intrinsic features of acoustic signatures which aids in health-related prediction. Future scope is to develop a more precisioned stand-alone system for acoustic signature classification and prediction of normal and abnormal signals that can assist the medical practitioners and health workers in preventive healthcare in a cost-effective and reliable manner.

7. ACKNOWLEDGMENTS

We would like to extend our sincere gratitude to KELTRON for providing the necessary resources for this research endeavour. We also gratefully acknowledge the contribution of Dr. Gokul V Pillai for reviewing the paper and providing valuable insights.

Conflicts of Interest : The authors declare no conflict of interests.

REFERENCES

[1] Jadyn Cook, Muneebah Umar, Fardin Khalili and Amirtaha Taebi, 2022. "Body Acoustics for the Non-Invasive Diagnosis of Medical Conditions" *Bioengineering*, **9:** 149.

[2] Yaseen, Son G.-Y. and Kwon S., 2018. "Classification of Heart Sound Signal Using Multiple Features" *Appl. Sci.,* **8:** 2344. doi:10.3390/app8122344.

[3] Rao A., Huynh E., Royston T.J., Kornblith A. and Roy S., 2019. "Acoustic Methods for Pulmonary Diagnosis". *IEEE Rev. Biomed. Eng.,* **12:** 221-239.

[4] Du X., Allwood G., Webberley K.M., Osseiran A., Wan W., Volikova A. and Marshall B.J., 2018. "A Mathematical Model of Bowel Sound Generation". *J. Acoust. Soc. Am.,* **144:** EL485-EL491. Doi:10.1121/1.5080528.

[5] Zaman, K., Sah, M., Direkoglu, C. and Unoki, M., 2023. "A survey of audio classification using deep learning". *IEEE Access.* doi: 10.1109/ACCESS.2023.3318015.

[6] Li, F., Zhang, Z., Wang, L. and Liu, W., 2022. "Heart sound classification based on improved mel-frequency spectral coefficients and deep residual learning". *Frontiers in Physiology,* **13:** 1084420. doi:10.3389/fphys.2022.1084420.

[7] Brunese, L., Martinelli, F., Mercaldo, F. and Santone, A., 2020. "Deep learning for heart disease detection through cardiac sounds". *Procedia Computer Science,* **176:** 2202-2211. doi: 10.1016/j.procs.2020.09.257 Oliveira, J., Renna, F., Costa, P., Nogueira, M., Oliveira, A. C., Elola, A., Ferreira, C., Jorge, A., Bahrami Rad, A., Reyna, M., Sameni, R., Clifford, G. and Coimbra, M., 2022. The CirCor DigiScope Phonocardiogram Dataset (version 1.0.3). PhysioNet. https://doi https://doi.org/10.13026/tshs-mw03.

[8] J. H. Oliveira, F. Renna, P. Costa, D. Nogueira, C. Oliveira, C. Ferreira, A. Jorge, S. Mattos, T. Hatem, T. Tavares, A. Elola, A. Rad, R. Sameni, G. D. Clifford and M. T. Coimbra, 2021. "The CirCor DigiScope Dataset: From Murmur Detection to Murmur Classification". *IEEE Journal of Biomedical and Health Informatics,* https://doi.org/10.1109/JBHI.2021.3137048.

[9] Li. *et al.,* 2022. "Heart sound classification based on improved MFCC and deep residual learning." *Frontiers in Physiology,* **13:** 1084420. DOI:10.3389/fphys.2022.1084420- includes comparisons of SVM, RF, and ResNet models.

[10] Lundervold and Lundervold, 2019. "An overview of deep learning in medical imaging focusing on MRI." *Zeitschrift für Medizinische Physik,* **29**(2): 102-127. DOI: 10.1016/j.zemedi.2018.11.002-though MRI-focused, discusses XAI methods applicable to biomedical signals.

[11] Purwins *et al.,* 2019. "Deep learning for audio signal processing." *IEEE Journal of Selected Topics in Signal Processing,* **13**(2): 206-219. DOI:10.1109/JSTSP.2019.2891965.

[12] Qinghao Zhao, Shijia Geng, Boya Wang, Yutong Sun, Wenchang Nie, Baochen Bai, Chao Yu, Feng Zhang, Gongzheng Tang, Deyun Zhang 2, Yuxi Zhou, Jian Liu and Shenda Hong, 2024. "Deep Learning in Heart Sound Analysis: From Techniques to Clinical Applications", **9:** 4: 0182. doi: 10.34133/hds.0182. eCollection 2024.

[13] Brashier B. and Salvi S., 2015. "Measuring Lung Function Using Sound Waves: Role of the Forced Oscillation Technique and Impulse Oscillometry System." *Breathe,* **11:** 57-65.

[14] Hu Y., Kim E.G., Cao G., Liu S. and Xu Y., 2014. "Physiological Acoustic Sensing Based on Accelerometers: A Survey for Mobile Healthcare" *Ann. Biomed. Eng.,* **42:** 2264-2277. doi:10.1007/s10439-014-1111-8.

The Journal of Acoustical Society of India : Vol. 52, No. 3, 2025 (pp. 109-116)

Artificial intelligence advancements in cochlear implant recipient care: A systematic review of recent trends in preoperative assessment and postoperative management

Arunachalam A.S. and Nirupama S.
Department of Audiology, All India Institute of Speech and Hearing,
Mysuru, Karnataka, India
e-mail: arun.chalam476@gmail.com

[Received: 10-03-2025; Accepted: 10-07-2025]

ABSTRACT

Cochlear implants (CIs) have revolutionized hearing rehabilitation for individuals with severe-to-profound hearing loss. This systematic review explores recent advancements in the application of artificial intelligence (AI) for enhancing preoperative assessment and postoperative management of cochlear implant recipients. A total of 42 relevant studies employing AI techniques, including machine learning, deep learning, and natural language processing, in the care of CI were identified from a comprehensive search of 20 electronic databases. These studies were critically appraised to synthesize emerging trends and their impact on clinical practice. Preoperatively, AI-driven algorithms have considerably improved patient selection based on complex audiometric data analysis, speech recognition scores, and medical histories. These models further improve predictive accuracy in the selection of candidates for cochlear implantation and, by doing so, optimize surgical outcomes and minimize risks. AI-based mapping software, such as FOX, will provide tailored rehabilitation strategies in the postoperative period through real-time monitoring of auditory performance metrics. This AI tool will also provide calibration for the patient's sound processor, including unique recommendations for fitting based on appropriate pitch and volume adjustments. These include speech perception tests and device use patterns that enable timely modifications in programming and rehabilitative strategies. It is these individualized options that help increase both patient satisfaction and functional outcomes in daily communications. Conclusion: It is now clear that the emerging AI technologies mark a paradigm shift in the care of cochlear implant recipients by offering formidable tools for the practice of precision medicine and personalized intervention. Continued research and clinical validation will be necessary to tap into and utilize the full potential of AI in the optimization of outcomes throughout a patient's journey with a cochlear implant.

1. INTRODUCTION

Artificial Intelligence may completely revolutionize cochlear implant care by developing advanced solutions not only for preoperative assessment but also for postoperative management. AI technologies, such as machine learning algorithms and neural networks, have immense potential in further enhancing the accuracy of presurgical assessments and the optimization of postsurgical results (Ross, 1995; You *et al.*, 2020). While this has been an effective system in certain key areas, like automatic speech recognition and natural language processing (Lesica *et al.*, 2021), its application in hearing healthcare, especially in CI, is still at its infancy stage. The integration of AI with CI technology would permit the improved selection of candidates, refinement in electrode mapping, and predictions over postoperative performance, leading to better overall patient outcomes (Koyama, 2024; Aliyeva, 2023). The aim of this systematic review is to discuss recent development in AI in CI care, with its application throughout preoperative to postoperative stages to give an overview of the potential benefits arising therefrom.

2. METHODS AND RESULTS

A complete systematic review was performed using databases such as Google Scholar and PubMed using Boolean operators to find peer-reviewed articles in English from 2010 to 2024. Following PRISMA guidelines, two authors screened titles and abstracts, identifying relevant studies for full-text review. Data on AI advancements in cochlear implants, covering pre- and post-operative aspects, were extracted.

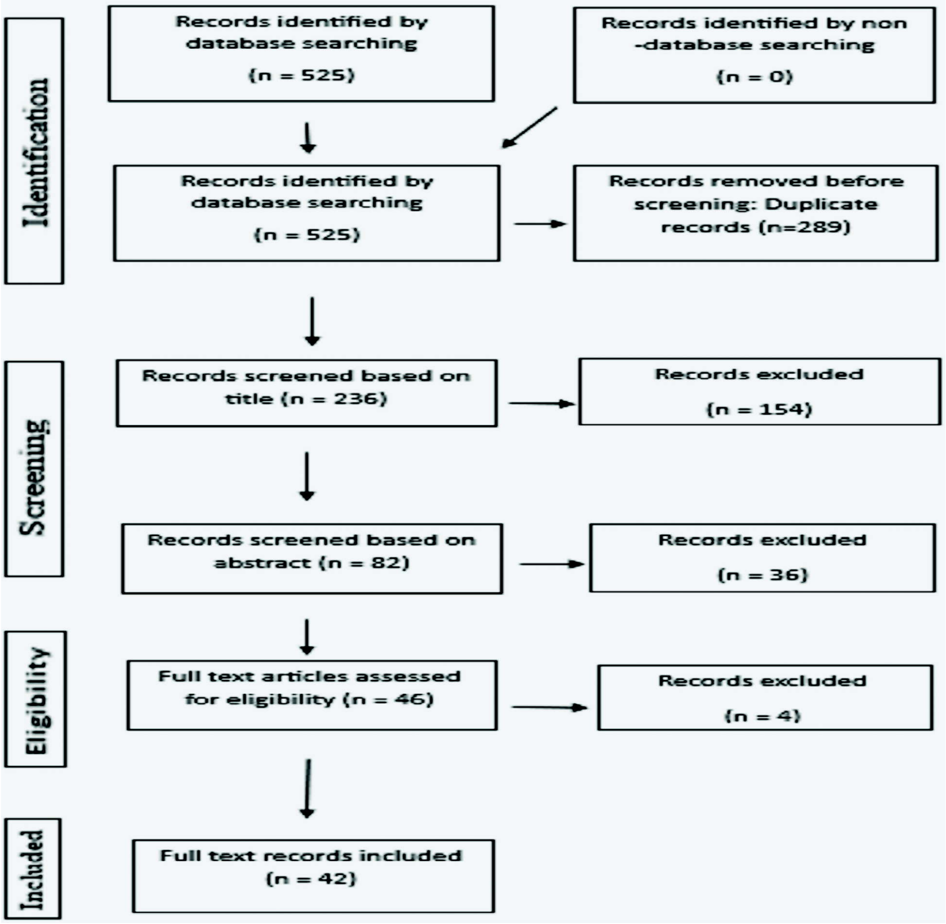

Identification
Records identified by database searching (n = 525)
Records identified by non -database searching (n = 0)
Records identified by database searching (n = 525)
Records removed before screening: Duplicate records (n=289)
Screening
Records screened based on title (n = 236)
Records excluded (n = 154)
Records screened based on abstract (n = 82)
Records excluded (n = 36)
Eligibility
Full text articles assessed for eligibility (n = 46)
Records excluded (n = 4)
Included
Full text records included (n = 42)

Fig. 1. PRISMA flow diagram of the study selection process.

Out of 525 results, 46 papers met the criteria after removing duplicates and disqualified studies, leading to the final selection of 42 publications: 18 experimental, 12 reviews, and 12 case studies, theses, and other study designs from various countries reflecting growing advancements and awareness in AI for cochlear implants.

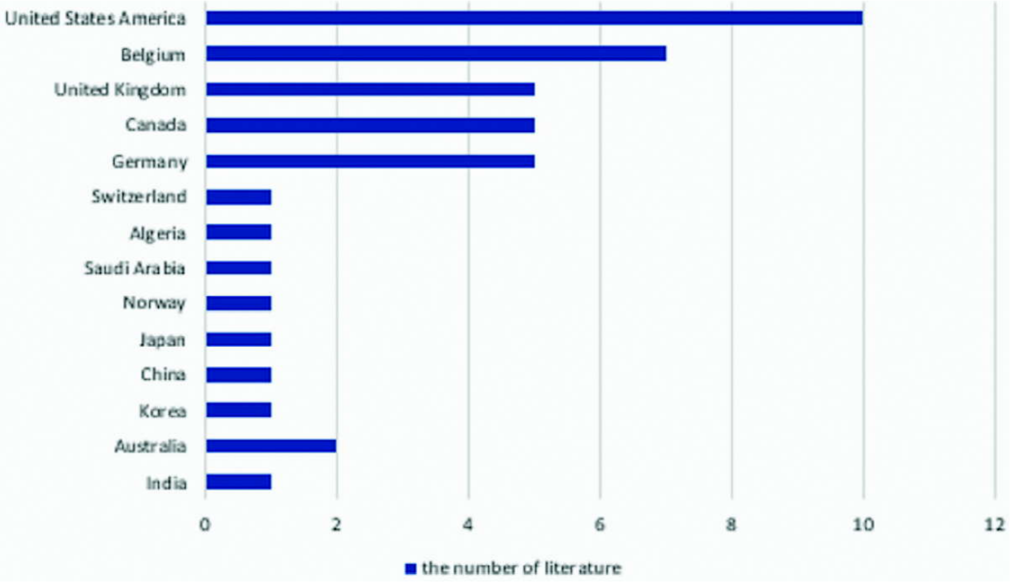


Fig. 2. Research location by the included articles.

The following are the outcomes regarding application of artificial intelligence in Preoperative Assessment and Postoperative Management. Some of the techniques which are mentioned in the review articles are:

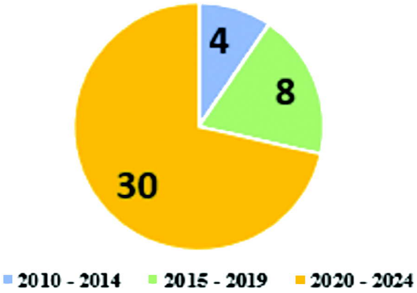
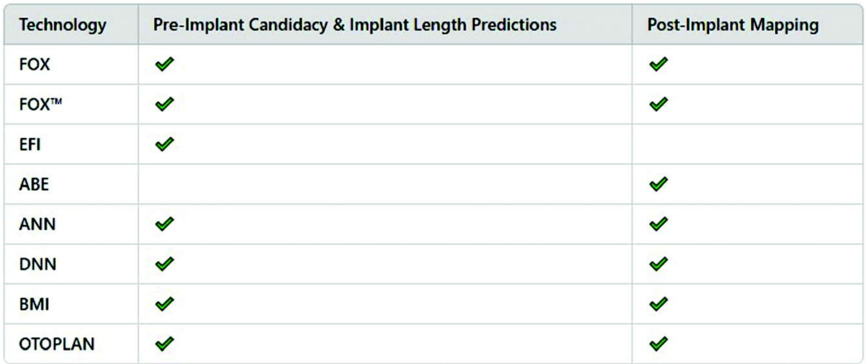


Fig. 3. Year of publications.

Vaerenberg *et al.* (2010) conducted a prospective study with 8 patients in Belgium, evaluating the "Fitting to Outcomes eXpert" (FOX) software for cochlear implant programming. The study found FOX to be feasible and effective, producing favourable results in psycho-acoustic tests. Botros (2010) developed AutoNRT, based on study performed on 42 participants, an AI-driven system for automating the measurement of ECAP thresholds, demonstrating significant clinical success with machine-learned decision trees. Vaerenberg *et al.* (2014) demonstrated a retrospective analysis on 255 participants with psychoacoustic targets and using FOX can effectively optimize cochlear implant programming. Iversen (2014) highlighted that while AI and ML in CI programming have limitations, they offer substantial potential, with future work needed to expand datasets and refine techniques. Iversen (2014) explored experimental study on 300 participants about how AI and machine learning can enhance cochlear implant programming, noting significant potential despite current limitations. Battmer *et al.* (2015) conducted experimental study on 27 Cochlear Implant recipients and demonstrated that the FOX software efficiently optimizes cochlear implant programming, reduces fitting time, and improves speech perception within the first six months compared to traditional methods. Based on experimental research on 10 Cochlear Implant Recipients ,Buechner *et al.* (2015) found that the FOX® software provides a standardized CI fitting approach based on outcome measures rather than user comfort, enhancing the fitting process and improving auditory outcomes through AI-driven adjustments. Kamar (2016) reviewed hybrid intelligence systems that combine human and AI inputs, highlighting their potential to enhance cochlear implant fittings and speech understanding through improved AI-human collaboration. Seeber & Bruce (2016) presented a review of the computational models in cochlear implants, including stages of their historical development, their impact on the development of CI technology, and stimulation strategies. In the retrospective study by Meeuws *et al.* (2017), 25 participants indeed found that FOX2G intelligent agent provides an optimal way of fitting cochlear implants, improving speech understanding and having the potential for further improvement with machine learning. Kim *et al.* performed an experimental investigation on a cohort of 120 subjects and found that state-of-the-art machine learning techniques, such as Random-Forest Regression, significantly improve the prediction of cochlear implantation outcome in postlingually deaf adults. This improves the precision of predictions against individual variability to support clinicians in providing better, more tailored advice, and optimizing the timing and methods of CI surgeries to attain more desirable outcomes. The study of Nogueira *et al.* showed that the ABE significantly enhanced the telephone speech intelligibility and quality for 9 cochlear implant participants. This improvement speaks to a better communication for users through enhanced clarity of speech on the phone. Saeed *et al.* provide an overview regarding the transformative effect AI and machine learning have caused in cochlear implantations and precision medicine, hence putting great stress on treatments tailored for improved outcomes. The breakthrough AI has brought to otology, You *et al.* (2020) discuss, focuses on various new developments and perspectives for future work in optimizing hearing aids and treating disorders affecting the audio system. Meeuws *et al.* (2020) have conducted a promising study on AI-driven remote fitting of cochlear-implanted recipients based on self-assessed psychoacoustic tests. However, there are still technical and regulatory challenges to be overcome. Wathour *et al.* 2020 demonstrated in the case of 2 Cochlear implant recipients that AI-assisted fitting, using FOX®, resulted in significantly better auditory outcomes for cochlear implant users compared to manual fitting. Crowson *et al.* (2020) presented the increasing role of machine learning in cochlear implants, focusing on improvement in speech and sound processing, and automated mapping for pediatric and adult recipients. Crowson *et al.* (2020) demonstrated the capability of supervised machine learning in analyzing the adverse event report for cochlear implants that enhanced the detection of pattern and trend for device safety and performance. Crowson *et al.* (2020) conducted an experimental study in 1604 patients and found that the determination of performance in cochlear implant patients can be predicted by the use of supervised machine learning and thereby assist preoperative education and improve decision-making in surgical outcomes.



Technology	Pre-Implant Candidacy & Implant Length Predictions	Post-Implant Mapping
FOX	✓	✓
FOX™	✓	✓
EFI	✓	
ABE		✓
ANN	✓	✓
DNN	✓	✓
BMI	✓	✓
OTOPLAN	✓	✓

Above table shows the results of AI algorithms used in pre-implantation to post-implantation.

3. DISCUSSION

By integrating AI in cochlear implants, it is a completely new dimension to auditory health-from the initial stages of pre-implantation prediction to post-fitting enhancements. AI-powered predictive analytics, as in Kim *et al.*, 2018, and Crowson *et al.*, 2020, not only quicken timely interventions but also optimize those conducted using cochlear implants. Such AI-driven tools as FOX™ ensure heightened fitting accuracy and satisfaction of the patients, according to the studies of Buechner *et al.* and Battmer *et al.* For example, Meeuws *et al.* (2020) studied remote fitting capabilities that improve accessibility for patients who are geographically isolated. Other new methods that go further to refine both the implant performance and the patient outcomes include 3D-printed biomimetic cochleae and machine learning co-modelling (Lei *et al.*, 2021). Carlson *et al.* also developed an artificial intelligence model, which ascertains cochlear implant candidacy from routine audiometry and demographic data, with 87% accuracy. It has been demonstrated to be well-matched with clinical outcomes and serves to support web-based and EMR tools. OTOPLAN software enhancements add to the precision of cochlear implant surgeries through very detailed 3D reconstructions, thorough surgical planning, and real-time integration with navigation systems. Chen *et al.* (2021) combined 3D printing and machine learning to continue the development of cochlear implants to provide biomimetic cochlear models with electric field imaging profiles. This would thereby provide customized CI electrode arrays and offer an avenue toward digital twin cochlear implants for a personalized gain or performance outcome. Essaid *et al.* (2023) go over AI development for cochlear implants and point out that FOX optimizes the initial fitting and rehabilitation, enhances speech recognition, and improves user satisfaction by citing challenges such as variability in patients and data requirements.

4. CONCLUSION AND ETHICAL CONSIDERATIONS

AI in cochlear implant care is rapidly changing both pre- and postoperative processes by optimizing patient selection and providing personalized treatment. AI algorithms increase the precision of candidacy predictions and surgical planning, while mapping software such as FOX enables post-surgical adjustments to be more precise and tailored for individual performance. This trend toward precision medicine improves functional outcomes and patient satisfaction. Further research will be required to extend the datasets, robustness of algorithms, and clinical applicability to further enhance AI-driven interventions. Lesica *et al.* (2021) emphasize the urgent need for robust ethical frameworks to address privacy, security, and algorithmic bias in the development and implementation of AI in auditory healthcare. While current research on AI in cochlear implants shows promise, it is essential to translate these findings into clinical practice through real-world trials to address practical challenges and validate efficacy. This transition will ensure that AI technologies are robust and beneficial for routine use, ultimately improving patient outcomes and satisfaction.

5. DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article, further inquiries can be directed to corresponding author/s.

5.1 *Disclaimer statements*

Contributors : Both the authors were actively involved in either planning, development, execution, monitoring, analysing and in the reviewing stage of this research.

Conflicts of interest : The author has no conflict of interest to disclose regarding this article.

Ethical Clearance : Not applicable.

6. ACKNOWLEDGEMENT

We thank Dr. M. Pushpavathi, Director, AIISH, Dr. Ajish K. Abraham, Dean Infrastructure, AIISH and Dr. Devi. N, HOD, Department of Audiology, AIISH for constant support and guidance throughout the study. Last, we thank NSA for the opportunity for presenting our research work.

REFERENCES

[1] Aliyeva, A., 2023. Transhumanism: Integrating Cochlear Implants with Artificial Intelligence and the Brain-Machine Interface. *Cureus.* https://doi.org/10.7759/cureus.50378

[2] Aliyeva, A., Sari, E., Alaskarov, E. and Nasirov, R., 2024. Enhancing Postoperative Cochlear Implant Care With chatgpt-4: A Study on Artificial Intelligence (AI)-Assisted Patient Education and Support. *Cureus.* https://doi.org/10.7759/cureus.53897

[3] Alohali, Y. A., Fayed, M. S., Abdelsamad, Y., Almuhawas, F., Alahmadi, A., Mesallam, T. and Hagr, A., 2023. Machine Learning and Cochlear Implantation: Predicting the Post-Operative Electrode Impedances. *Electronics (Switzerland),* **12**(12). https://doi.org/10.3390/electronics12122720

[4] Battmer, R. D., Borel, S., Brendel, M., Buchner, A., Cooper, H., Fielden, C., Gazibegovic, D., Goetze, R., Govaerts, P., Kelleher, K., Lenartz, T., Mosnier, I., Muff, J., Nunn, T., Vaerenberg, B. and Vanat, Z. (2015). Assessment of 'fitting to outcomes expert' FOXTM with new cochlear implant users in a multi-centre study. *Cochlear Implants International,* **16**(2): 100-109. https://doi.org/10.1179/1754762814Y.0000000093

[5] Botros, A., 2010. The application of machine intelligence to cochlear implant fitting and the analysis of the auditory nerve response. https://doi.org/10.26190/unsworks/4008

[6] Buechner, A., Vaerenberg, B., Gazibegovic, D., Brendel, M., De Ceulaer, G., Govaerts, P. and Lenarz, T. (2015). Evaluation of the 'Fitting to outcomes expert' (FOX®) with established cochlear implant users. *Cochlear Implants International,* **16**(1): 39-46. https://doi.org/10.1179/1754762814Y.0000000085

[7] Carlson, M. L., Carducci, V., Deep, N. L., dejong, M. D., Poling, G. L. and Brufau, S. R., 2024. AI model for predicting adult cochlear implant candidacy using routine behavioral audiometry. *American Journal of Otolaryngology - Head and Neck Medicine and Surgery*, **45**(4). https://doi.org/10.1016/j.amjoto.2024.104337

[8] Crowson, M. G., Dixon, P., Mahmood, R., Lee, J. W., Shipp, D., Le, T., Lin, V., Chen, J., & Chan, T. C. Y., 2020. Predicting Postoperative Cochlear Implant Performance Using Supervised Machine Learning. *Otology and Neurotology,* **41**(8): E1013-E1023. https://doi.org/10.1097/MAO.0000000000002710

[9] Crowson, M. G., Hamour, A., Lin, V., Chen, J. M. and Chan, T. C. Y., 2020. Machine learning for pattern detection in cochlear implant FDA adverse event reports. *Cochlear Implants International,* 313-322. https://doi.org/10.1080/14670100.2020.1784569

[10] Crowson, M. G., Lin, V., Chen, J. M. and Chan, T. C. Y., 2020. Machine Learning and Cochlear Implantation - A Structured Review of Opportunities and Challenges. In *Otology and Neurotology* **41**(1): E36-E45. Lippincott Williams and Wilkins. https://doi.org/10.1097/MAO.0000000000002440

[11] Dillon, M. T., Helpard, L., Brown, K. D., Selleck, A. M., Richter, M. E., Rooth, M. A., Thompson, N. J., Dedmon, M. M., Ladak, H. M. and Agrawal, S., 2023. Influence of the Frequency-to-Place Function on Recognition with Place-Based Cochlear Implant Maps. Laryngoscope, **133**(12): 3540-3547. https://doi.org/10.1002/lary.30710

[12] Essaid, B., Kheddar, H., Batel, N., Lakas, A. and Chowdhury, M. E. H., 2024. *Advanced Artificial Intelligence Algorithms in Cochlear Implants: Review of Healthcare Strategies, Challenges, and Perspectives.* http://arxiv.org/abs/2403.15442

[13] Hafeez, N., Du, X., Boulgouris, N., Begg, P., Irving, R., Coulson, C. and Tourrel, G., 2021. Electrical impedance guides electrode array in cochlear implantation using machine learning and robotic feeder. *Hearing Research,* 412. https://doi.org/10.1016/j.heares.2021.108371

[14] Iversen, A. H., 2014. *Use of Artificial Intelligence and Machine Learning algorithms to predict programming levels in Cochlear Implant patients.* http://www.duo.uio.no/

[15] Kamar, E. (n.d.). *Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence.*

[16] Kim, H., Kang, W. S., Park, H. J., Lee, J. Y., Park, J. W., Kim, Y., Seo, J. W., Kwak, M. Y., Kang, B. C., Yang, C. J., Duffy, B. A., Cho, Y. S., Lee, S. Y., Suh, M. W., Moon, I. J., Ahn, J. H., Cho, Y. S., Oh, S. H. and Chung, J. W., 2018. Cochlear Implantation in Postlingually Deaf Adults is Time-sensitive Towards Positive Outcome: Prediction using Advanced Machine Learning Techniques. *Scientific Reports,* **8**(1). https://doi.org/10.1038/s41598-018-36404-1

[17] Koyama, H., 2024. Machine learning application in otology. In *Auris Nasus Larynx,* **51**(4), 666-673. Elsevier Ireland Ltd. https://doi.org/10.1016/j.anl.2024.04.003

[18] Lei, I. M., Jiang, C., Lei, C. L., de Rijk, S. R., Tam, Y. C., Swords, C., Sutcliffe, M. P. F., Malliaras, G. G., Bance, M. and Huang, Y. Y. S., 2021. 3D printed biomimetic cochleae and machine learning co-modelling provides clinical informatics for cochlear implant patients. *Nature Communications,* **12**(1). https://doi.org/10.1038/s41467-021-26491-6

[19] Lesica, N. A., Mehta, N., Manjaly, J. G., Deng, L., Wilson, B. S. and Zeng, F. G., 2021. Harnessing the power of artificial intelligence to transform hearing healthcare and research. In *Nature Machine Intelligence,* **3**(10): 840-849. Nature Research. https://doi.org/10.1038/s42256-021-00394-z

[20] Lu, S., Xie, J., Wei, X., Kong, Y., Chen, B., Chen, J., Zhang, L., Yang, M., Xue, S., Shi, Y., Liu, S., Xu, T., Dong, R., Chen, X., Li, Y. and Wang, H., 2022. Machine Learning-Based Prediction of the Outcomes of Cochlear Implantation in Patients With Cochlear Nerve Deficiency and Normal Cochlea: A 2-Year Follow-Up of 70 Children. *Frontiers in Neuroscience,* 16. https://doi.org/10.3389/fnins.2022.895560

[21] Meeuws, M., Pascoal, D., Bermejo, I., Artaso, M., De Ceulaer, G. and Govaerts, P. J., 2017. Computer-assisted CI fitting: Is the learning capacity of the intelligent agent FOX beneficial for speech understanding? *Cochlear Implants International,* **18**(4): 198-206. https://doi.org/10.1080/14670100.2017.1325093

[22] Meeuws, M., Pascoal, D., Janssens de Varebeke, S., De Ceulaer, G., & Govaerts, P. J. (2020). Cochlear implant telemedicine: Remote fitting based on psychoacoustic self-tests and artificial intelligence. *Cochlear Implants International,* **21**(5): 260-268. https://doi.org/10.1080/14670100.2020.1757840

[23] Nogueira, W., Abel, J. and Fingscheidt, T., 2019. Artificial speech bandwidth extension improves telephone speech intelligibility and quality in cochlear implant users. *The Journal of the Acoustical Society of America,* **145**(3): 1640-1649. https://doi.org/10.1121/1.5094347

[24] Patro, A., Freeman, M. H. and Haynes, D. S., 2024. Machine Learning to Predict Adult Cochlear Implant Candidacy. In *Current Otorhinolaryngology Reports. Springer Science and Business Media B.V.* https://doi.org/10.1007/s40136-024-00511-7

[25] Ross, D., 1995. Artificial Intelligence: A Philosophical Introduction. *Philosophical Books,* **36**(3). https://doi.org/10.1111/j.1468-0149.1995.tb02493.x

[26] Saeed, H. S., Stivaros, S. M. and Saeed, S. R., 2019. The potential for machine learning to improve precision medicine in cochlear implantation. In *Cochlear Implants International,* **20**(5): 229-230). Taylor and Francis Ltd. https://doi.org/10.1080/14670100.2019.1631520

[27] Schurzig, D., Repp, F., Timm, M. E., Batsoulis, C., Lenarz, T. and Kral, A., 2023. Virtual cochlear implantation for personalized rehabilitation of profound hearing loss. *Hearing Research,* 429. https://doi.org/10.1016/j.heares.2022.108687

[28] Seeber, B. U. and Bruce, I. C., 2016. The history and future of neural modeling for cochlear implants. In Network: *Computation in Neural Systems,* 27(2-3): 53-66. Taylor and Francis Ltd. https://doi.org/10.1080/0954898X.2016.1223365

[29] Skidmore, J., Xu, L., Chao, X., Riggs, W. J., Pellittieri, A., Vaughan, C., Ning, X., Wang, R., Luo, J. and He, S. (2021). Prediction of the Functional Status of the Cochlear Nerve in Individual Cochlear Implant Users Using Machine Learning and Electrophysiological Measures. *Ear and Hearing,* **42**(1): 180-192. https://doi.org/10.1097/AUD.0000000000000916

[30] Umashankar, A., MN, A. and C. P., 2021. Applications of Artificialintelligence in Hearing Aids and Auditory Implants: A Short Review. *Journal of Hearing Science,* **11**(3): 20-23. https://doi.org/10.17430/jhs.2021.11.3.2

[31] Vaerenberg, B., De Ceulaer, G., Szlávik, Z., Mancini, P., Buechner, A. and Govaerts, P. J., 2014. Setting and reaching targets with computer-assisted cochlear implant fitting. *The Scientific World Journal,* 2014. https://doi.org/10.1155/2014/646590

[32] Vaerenberg, B., Govaerts, P. J., De Ceulaer, G., Daemers, K. and Schauwers, K., 2010. Experiences of the use of FOX, an intelligent agent, for programming cochlear implant sound processors in new users. *International Journal of Audiology,* **50**(1): 50-58. https://doi.org/10.3109/14992027.2010.531294

[33] Waltzman, S. B. and Kelsall, D. C., 2020. The Use of Artificial Intelligence to Program Cochlear Implants. *Otology and Neurotology,* **41**(4): 452-457. https://doi.org/10.1097/MAO.000000000000256

[34] Wathour, J., Govaerts, P. J. and Deggouj, N., 2020. From manual to artificial intelligence fitting: Two cochlear implant case studies. *Cochlear Implants International,* **21**(5): 299-305. Https://doi.org/10.1080/14670100.2019.1667574

The Journal of Acoustical Society of India : Vol. 52, No. 3, 2025 (pp. 117-122)

Machine learning for endemic bird song identification and classification

C.R.S. Kumar*
School of Computer Engineering and Mathematical Sciences,
Defence Institute of Advanced Technology, Pune-411 025, India
e-mail: suthikshnkumar@diat.ac.in

[Received: 00-00-2025; Accepted: 00-00-2025]

ABSTRACT

Endemic bird species are critical to biodiversity and ecosystem health, often serving as indicators of environmental changes. The identification and monitoring of these species through their vocalizations are essential for conservation efforts. Endemic Birds of Western Ghats (India's west coast region) need to be identified and protected. In this regard, bird songs play a role in identifying such precious birds. A ML model trained on bird songs dataset plays an important role in ecological conservation of these endangered birds. This paper presents a novel approach to classifying endemic bird songs using advanced machine learning techniques. Leveraging a dataset of bird calls from Wester Ghats (India's West coast region) endemic species, we explore the efficacy of machine learning algorithms (EfficientNet based on convolutional neural networks (CNNs)) in accurately identifying and classifying bird songs. Our methodology includes data preprocessing steps such as noise reduction, spectrogram generation, and augmentation to enhance the robustness of the model. Mel spectral coefficients are features derived from the Mel-frequency cepstrum, commonly used in audio signal processing, particularly in speech and music analysis. We utilize the EfficientNet model developed by Google Research which is a family of CNNs known for its performance efficiency and scalability. EfficientNet models achieve higher accuracy with fewer parameters compared to previous architectures. EfficientNet introduces a novel compound scaling method that uniformly scales all dimensions of depth, width, and resolution using a simple yet highly effective strategy. This allows the model to be scaled up or down efficiently, maintaining a balance between model complexity and accuracy. This paper discusses the details of datasets, pre-processing, feature engineering, Model fine tuning, hyper-parameter settings, post processing, metrics, scoring and comparison of the final implementation.

1. INTRODUCTION

Bird song identification and classification have long been critical components of ornithological research, aiding in species identification, behavioral studies, and conservation efforts[1,2]. Traditional methods, which rely heavily on expert knowledge and manual analysis, can be labor-intensive and limited by the availability and expertise of ornithologists. As the need for efficient and scalable solutions grows, machine learning (ML) emerges as a powerful tool to address these challenges, offering automated and accurate identification of bird species based on their vocalization[9].

Endemic birds, those species unique to specific geographic regions, present particular importance and complexity in conservation biology. Monitoring these species is crucial for biodiversity preservation, as they are often more vulnerable to habitat loss and environmental changes[4,5,10,11]. Bird population has been found to be beneficial and any decline in their numbers can also adversely affect human population[14]. Accurate identification and classification of their songs are essential for tracking population dynamics, understanding ecological roles, and implementing effective conservation strategies.

Machine learning, a subset of artificial intelligence (AI), has demonstrated remarkable success in various fields, including image and speech recognition, medical diagnosis, and autonomous systems[8]. In the context of bird song identification, ML algorithms can be trained to recognize patterns and classify audio signals with high precision. This approach leverages large datasets and sophisticated models to overcome the limitations of traditional methods, providing a scalable and efficient solution for analyzing vast amounts of audio data.

This paper explores the application of machine learning techniques to the identification and classification of endemic bird songs. We present a comprehensive overview of the methodology, including data collection, preprocessing, feature extraction, model selection, training, and evaluation. By integrating advanced ML models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), with audio signal processing techniques, we aim to achieve robust and accurate bird song classification.

The primary contributions of this work are twofold. First, we explore various feature extraction methods and ML models, highlighting their strengths and limitations in the context of bird song classification. Second, we demonstrate the effectiveness of our approach through empirical evaluations on a dataset of endemic bird songs, showcasing significant improvements in classification accuracy compared to traditional methods.

The findings of this study have significant implications for ornithological research and conservation efforts. By enabling automated and accurate identification of bird species, our approach can facilitate large-scale monitoring and analysis of avian populations, contributing to more informed conservation strategies. Additionally, the methodologies and insights presented in this paper can serve as a foundation for future research in the application of machine learning to bio acoustic studies.

2. ENDEMIC BIRD SONGS OF WESTERN GHATS DATASET

The dataset consist of bird songs to identify which birds are calling in recordings made in a Global Biodiversity Hotspot in the Western Ghats[1]. The total data set size is 23.43 GB with song recordings in ogg format for 182 bird species. The following files are provided in the dataset:

- **train_audio :** The training data consists of short recordings of individual bird calls generously uploaded by users of xenocanto.org. These files have been downsampled to 32 kHz where applicable to match the test set audio and converted to the ogg format.
- **test_soundscapes :** the test_soundscapes directory will be populated with approximately 1,100 recordings to be used for scoring. They are 4 minutes long and in ogg audio format. The file names are randomized but have the general form of soundscape_xxxxxx.ogg.
- **unlabeled_soundscapes :** Unlabeled audio data from the same recording locations as the test soundscapes.
- **train_metadata.csv :** A wide range of metadata is provided for the training data. The most directly relevant fields are:
  - o primary_label - a code for the bird species[2].
  - o latitude & longitude: coordinates for where the recording was taken. Some bird species may have local call 'dialects,' to provide geographic diversity in training data.
  - o author - The user who provided the recording.
  - o filename: the name of the associated audio file.

- **eBird_Taxonomy_v2021.csv :** Data on the relationships between different species.

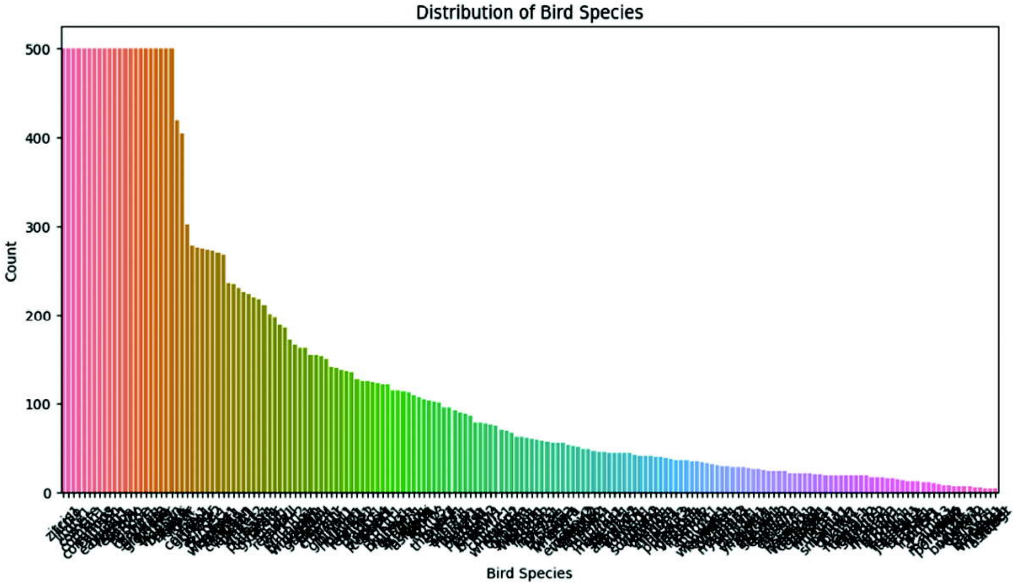


Fig. 1. Distribution of fles for bird species.

3. MACHINE LEARNING MODEL

The ML model utilized in this implementation is EfficientNet[3]. EfficientNet is a family of convolutional neural networks (CNNs) developed by researchers at Google AI[7]. Introduced in 2019, EfficientNet aims to optimize both accuracy and efficiency (i.e., computational and memory resources) in image classification tasks. The key innovation of EfficientNet is its compound scaling method, which uniformly scales all dimensions of depth, width, and resolution using a simple yet effective formula. The key concepts are as follows:

- **Compound Scaling:** EfficientNet uses a compound coefficient ϕ to scale the depth, width, and resolution of the network uniformly:
  - o *Depth:* Number of layers (or the number of repeated blocks).
  - o *Width:* Number of channels (or the number of filters in each layer).
  - o *Resolution:* Size of the input image.
- **Baseline Network:** EfficientNet starts with a baseline network, EfficientNet-B0, which is optimized for a good balance of accuracy and efficiency. EfficientNet-B0 is built using a mobile architecture called MobileNetV2 as a foundation, incorporating techniques like squeeze-and-excitation (SE) optimization.
- **Family of Models:** The EfficientNet family includes multiple models, from EfficientNet-B0 to EfficientNet-B7, each scaled up progressively:
  - o *EfficientNet-B0:* The baseline model.
  - o *EfficientNet-B1 to B7:* Models with progressively higher values of ϕ, leading to larger and more accurate networks.
- **Performance :** EfficientNet models achieve state-of-the-art performance on several benchmark datasets like ImageNet while being more computationally efficient than previous architectures like ResNet and Inception. This efficiency allows for deployment in resource-constrained environments like mobile devices and embedded systems.

The main advantages of EffientNet are as follows:

- **High Accuracy :** EfficientNet models achieve top-tier accuracy on various image classification benchmarks.
- **Efficiency :** They require fewer parameters and less computational power compared to other high-performing models.
- **Scalability :** The compound scaling method provides a systematic way to scale models for different resource constraints.

The important applications of EfficientNet are in Image Classification, Transfer Learning and deployment model for Edge Devices.

4. IMPLEMENTATION

Using EfficientNet for bird song identification and classification leverages its high accuracy and efficiency for audio-based machine learning tasks. While EfficientNet is primarily a convolutional neural network designed for image classification, it can be adapted for audio classification tasks, such as bird song identification, by converting audio signals into a visual representation like spectrograms. The steps to implement EfficientNet for Bird Song Identification are as follows:

DataSet Partition	Score (mean-evraged ROC-AUC)	Remarks
35% (Public-LB)	0.658629	Sampling Rate = 32 Khz
65% (Private-LB)	0.651838	Learning Rate = 1e-03
		Optimizer = adan

- **Data Collection and Preparation**
  - *Collect Audio Data :* Gather a dataset of bird songs. Datasets like Xeno-canto[4] or Cornell Lab of Ornithology's Macaulay Library provide a rich source of bird song recordings.
  - *Label Data :* Ensure each audio file is labeled with the corresponding bird species.
  - *Data Augmentation :* Apply techniques like adding noise, pitch shifting, and time stretching to increase the diversity of your dataset.
- **Convert Audio to Spectrograms**
  - *Generate Spectrograms :* Convert each audio clip into a spectrogram. A spectrogram is a visual representation of the spectrum of frequencies of a signal as it varies with time. Libraries like librosa in Python can be used for this purpose[6].
  - *Preprocess Spectrograms :* Normalize and resize the spectrograms to match the input size expected by EfficientNet (*e.g.,* 224x224 pixels for EfficientNet-B0).
- **Model Training**
  - *Choose an EfficientNet Model :* Depending on your computational resources, select an appropriate EfficientNet model (e.g., EfficientNet-B0 for lower resources, EfficientNet-B7 for higher accuracy).
  - *Transfer Learning :* Use a pre-trained EfficientNet model and fine-tune it on your spectrogram dataset. Transfer learning helps leverage the pre-trained weights for faster and more accurate convergence.
  - *Model Architecture :* Modify the final classification layer of EfficientNet to match the number of bird species in your dataset.
- **Training Process**
  - *Split Data :* Divide your dataset into training, validation, and test sets.
  - *Compile Model :* Use a suitable loss function (e.g., categorical cross-entropy) and optimizer (e.g., Adam).

  - *Train Model :* Train the model on the training set while monitoring its performance on the validation set.
  - *Early Stopping and Checkpoints :* Implement early stopping and model checkpoints to avoid overfitting and save the best model.

- **Evaluation and Testing**
  - *Evaluate Model :* After training, evaluate the model's performance on the test set using metrics like accuracy, precision, recall, and F1-score.
  - *Confusion Matrix :* Analyze the confusion matrix to see how well the model distinguishes between different bird species.
- **Deployment**
  - *Export Model :* Save the trained model for deployment.
  - *Inference Pipeline :* Develop an inference pipeline that takes raw audio input, converts it to a spectrogram, and uses the trained EfficientNet model to classify the bird species.
  - *User Interface :* Optionally, create a user interface or API for users to upload audio files and get predictions.

The python code for the EfficientNet Model implementation is published on the BirdCLEF competition website[1] in the code section.

5. RESULTS AND DISCUSSION

Once the model is trained and evaluated, the results are analyzed and discussed based on performance metrics and observations. Model Performance Metrics: The evaluation metric the model is a version of macro-averaged ROC-AUC that skips classes which have no true positive labels[12]. The ROC (Receiver Operating Characteristics) curve is plotted with TPR against FPR. When AUC is close to 0.5, the model lacks discrimination capability to differentiate between positive class and negative class. For Multiclass classification the ROC-AUC is very useful metric[13].

The results obtained are comparable to top scoring model implementations in the competition. (Model 27 is based on EfficientNet as discussed).

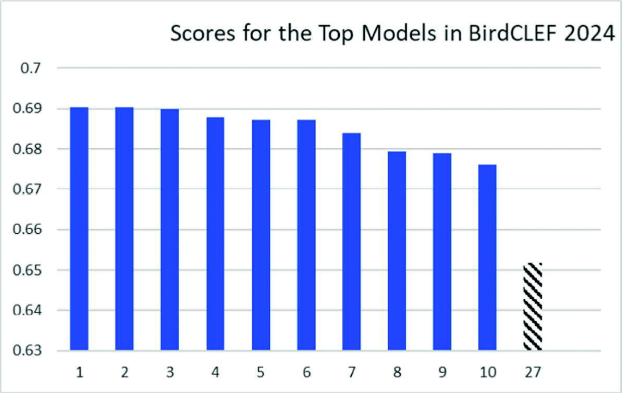


Fig. 2. Comparison of performance scores of different models.

6. SUMMARY AND CONCLUSION

EfficientNet, a family of convolutional neural networks developed by Google AI, has shown significant promise in various applications due to its high accuracy and computational efficiency. This study focused on utilizing EfficientNet for bird song identification and classification, a task that involves converting audio recordings of bird songs into spectrograms and applying the EfficientNet architecture to classify different bird species.

EfficientNet's application in bird song identification and classification has proven to be highly effective, leveraging its efficient architecture to achieve excellent performance while being computationally resource-efficient. The key findings and implications include:

- **High Performance and Efficiency :** The model's performance and efficiency in terms of parameter count and computational power makes it suitable for deployment in resource-constrained environments, such as field devices for real-time bird monitoring.

- **Data and Model Challenges :** Challenges such as data quality, background noise, and model interpretability were identified. Addressing these challenges will be crucial for further improving the model's robustness and reliability.

7. ACKNOWLEDGMENTS

Acknowledge organizers of BirdCLEF 2024 ML Competition for making available the dataset for training and testing the model. The python code with EfficientNet ML model was submitted for Global Machine Learning competition i.e BirdCLEF 2024, Kaggle (https://kaggle.com/competitions/birdclef-2024) . The EfficientNet based model submitted by authors was ranked 27th and achieved silver medal in the competition.

REFERENCES

[1] Holger Klinck, Maggie, Sohier Dane, Stefan Kahl, Tom Denton and Vijay Ramesh, 2024. BirdCLEF 2024. Kaggle. https://kaggle.com/competitions/birdclef-2024 (Last accessed on 12th July 2024).

[2] Bird Species data for individual birds by Cornell Lab of Ornithology: https://ebird.org/ (Last accessed on 12th July 2024).

[3] Tan M. and Le Q.V, 2019, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR,* **97:** 6105-6114.

[4] Xeno-canto: A community-driven platform with extensive bird song recordings. https://xeno-canto.org/ (Last accessed on 12th July 2024).

[5] Endemic birds of Western Ghats: ENVIS-BHNS website: http://www.bnhsenvis.nic.in/Database/Endemic-Birds-of-Western-Ghats_19786.aspx (Last accessed on 12th July 2024).

[6] McFee B., Raffel C. and Liang D., *et al.*, 2015, librosa: Audio and Music Signal Analysis in Python. *Proceedings of the 14th Python in Science Conference,* https://doi.org/10.25080/Majora-7b98e3ed-003.

[7] Transfer Learning with TensorFlow, Tensor Flow documentation on transfer learning and EfficientNet. TensorFlow Guide, Website: https://www.tensorflow.org/guide/keras/transfer_learning (Last accessed on 12th July 2024).

[8] Wikipedia on Machine Learning: https://en.wikipedia.org/wiki/Machine_learning (Last accessed on 30th Dec 2023).

[9] AurélienGéron, 2022, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, Third Edition, Paperback, OReilly Publishers, California.

[10] BirdLife International website: https://www.birdlife.org/ (Last accessed on 12th July 2024).

[11] Birding in India website: https://www.birding.in/ (Last accessed on 12th July 2024).

[12] S. Narkhede, "Understanding AUC- ROC Curve", https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (Last accessed on 12th July 2024).

[13] Hand D.J. and Till R.J., 2001, A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning,* **45:** 171-186. https://doi.org/10.1023/A:1010920819831

[14] Frank Eyal G. and Sudarshan Anant, 2022. "The Social Costs of Keystone Species Collapse : Evidence From The Decline of Vultures in India," The Warwick Economics Research Paper Series (TWERPS) 1433, University of Warwick, Department of Economics.

The Journal of Acoustical Society of India : Vol. 52, No. 3, 2025 (pp. 123-131)

Optimizing active noise control through GRU-based modeling: A comparative evaluation

Deepali Singh[1], Rinki Gupta[2], Arun Kumar[1*] and Rajendar Bahl[1]

[1]Centre for Applied Research in Electronics
Indian Institute of Technology Delhi, New Delhi, India
[2]Electronics and Communication Engineering Department
Amity University, Uttar Pradesh, Noida, India
e-mail: arunkm@care.iitd.ac.in

[Received: 00-00-2025; Accepted: 00-00-2025]

ABSTRACT

Active Noise Control (ANC) aims to minimize unwanted acoustic noise by generating an anti-noise waveform that destructively interferes with the unwanted noise. Traditional approaches have shown limitations in handling complex and time-varying noise patterns. To address this, recent advances have introduced deep learning models capable of capturing underlying noise characteristics. This paper presents a Recurrent Stacked Autoencoder (RSAE) architecture for ANC that combines convolutional feature extraction with Gated Recurrent Units (GRUs). The convolutional layers process spectral features of the noise, while the GRUs model its temporal evolution, enabling the system to adapt effectively to non-stationary noise environments. During training, the RSAE learns to predict the anti-noise using simulated noise, with the model's architecture optimized to account for the dynamics of both the secondary path and the acoustic propagation delay. Simulation studies demonstrate that the proposed RSAE achieves a noise reduction of 25.11 dB, outperforming conventional methods. These findings suggest that integrating GRUs significantly enhances the system's ability to generalize across varying noise conditions and improve real-time ANC performance.

1. INTRODUCTION

The growing problem of urban noise pollution poses significant threats to both environmental quality and human well-being. Active Noise Control (ANC) emerges as a promising solution by generating anti-noise, which ideally has the same but negative amplitude compared to the unwanted noise[1,2]. Through destructive interference, this anti-noise effectively lowers the overall noise level when paired with the original noise[3,4]. In contrast to passive noise control techniques like sound barriers and absorbers, ANC approaches are more compact, often less expensive, and more successful, especially when it comes to reducing low-frequency noise[5,6]. As a result, ANC has found widespread applications in various areas, including headphones, ventilation systems, and automotive environments[7,8].

Adaptive algorithms are generally used in conventional ANC systems to continually modify the coefficients of adaptive control filter to decrease error[9,10]. The filtered-X least mean square (FxLMS) method and its variations are commonly utilized among them because of their capacity to effectively handle the delays caused by the secondary path while preserving a high level of computational efficiency[11,12,13]. However, the overall efficacy of noise reduction can be jeopardized by adaptive ANC algorithms' long response times and divergence risk[14,15]. To enhance system stability and avoid the complexity of adaptive procedures, fixed-filter ANC algorithms are often preferred in practical applications.

Neural network-based ANC has gained significant attention over the past seven years[16]. Various architectures, including deep and recurrent networks, have been explored to enhance ANC performance in acoustic environments. In[16], the authors employed a convolutional neural network (CNN) to classify noise types with an accuracy of 96.43% for ANC. The system chooses the best filter for noise cancellation from a collection of 15 pre-trained filters based on the categorization. In[17], both the secondary path and the reverse secondary path have been modelled using convolutional recurrent networks (CRNs), which provide an error that is reduced using an adaptive filter for noise cancellation. In another approach, a deep learning method called DNoiseNet was used for feedback ANC designed for non-stationary noise in construction environments[18]. An MLP was utilized to predict the secondary path, and the system relied on simulated room impulse responses (RIRs) for modeling. In[19], the authors framed ANC as a supervised learning problem, where a CRN is trained on primary noise and its corresponding anti-noise, with simulations modeling loudspeaker non-linearities and RIRs. Additionally, identifying appropriate input and training signals for deep learning models in feed-forward ANC systems remains a significant challenge.

In our previous work, we developed a Stacked Autoencoder (SAE) based ANC algorithm that used two autoencoders: one for estimating the anti-noise from the loudspeaker and another for modeling the secondary path and the non-linearities of the secondary path and the speaker[20]. To strengthen the system's ability to handle the temporal dynamics of the noise and make the system more robust, Gated Recurrent Units (GRU) layers are used. The GRU layers will capture the temporal patterns of the noise for input signals and the evolving patterns of the anti-noise being learned. The GRU layers are crucial as they improve the performance of the learning system on a real-time basis by implicitly accounting for the simultaneous learning of the anti-noise and its evolving patterns[21,22]. While the SAE-based ANC technique developed in[16] effectively models the anti-noise and the secondary path, it primarily focuses on spatial feature extraction and non-linearities without explicitly addressing the temporal dynamics of noise. To address this limitation, we propose the integration of GRU layers between the encoder and decoder. This addition allows the system to learn and adapt to the temporal dependencies in the noise signals, improving the robustness and responsiveness of the ANC system in non-stationary noise environments[22]. The GRU layers, known for their ability to capture long-term dependencies in sequential data, improve the system's ability to predict and cancel noise by accounting for both current and past signal states[23]. This not only enhances the feature extraction capabilities of the autoencoder but also introduces advanced temporal modeling. The key contributions of this work include:

1. The design of an RSAE to predict the cancelling noise using primary noise as input.
2. A training strategy that involves training the RSAE on simulated noise.
3. Experimental results showing that GRU integration enhances noise reduction, particularly in unpredictable, real-world environments.

The remainder of this document is structured as follows: Section 2 details the design and methodology of the proposed ANC architecture, which is built upon the RSAE framework. Section 3 covers the experimental setup and data preprocessing steps. Section 4 presents the evaluation results along with comparative analyses. Finally, Section 5 provides the concluding remarks.

2. THE PROPOSED GRU-BASED RSAE FOR ANC

The framework for the proposed RSAE-based ANC setup is depicted in Fig. 1. The noise $x[n]$ is delayed and attenuated based on the known distance of the secondary path, to simulate the desired anti-noise $s[n]$ as described in (1) for a non-reflective scenario. The noise $x[n]$ and the corresponding anti-noise $s[n]$ are used as input and training target for RAE2 in Fig. 1.

$$s[n] = \alpha x[n - \Delta] \quad [1]$$

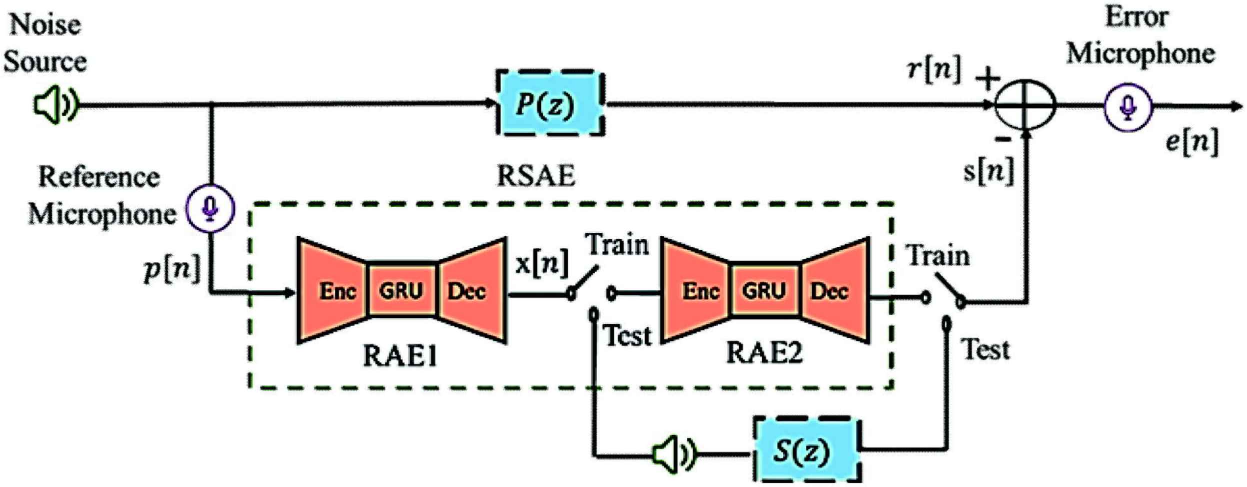


Fig. 1. Framework of the Recurrent Stacked autoencoder (RSAE) ANC.

In this equation, $x[n]$ represents the input noise used for training, while Δ denotes the delay, which can be fractional, and is calculated as t/f_s, where f_s, is the sampling rate, and t is the time delay, determined by the distance between the error microphone and the secondary noise source. The attenuation factor, α, models the reduction in noise level according to the free-field inverse square law, which says that the signal's power attenuates proportionally to the square of the distance from the sources. Also,

$$e[n] = r[n] - s[n] \quad [2]$$

A similar approach was applied for generating noise for the training of RAE1. An independent realization of white noise was used when simulating the primary noise source $p[n]$. Based on the estimated sound speed and the known distance of the primary path in a non-reflective room, the attenuation and delay for generating $p[n]$ from $r[n]$ was computed using (1).

ANC aims to predict and generate an anti-noise signal that, when combined with the original noise, results in significant noise reduction or cancellation. Mathematically, if $r[n]$ represents the primary noise at time n, and $s[n]$ is the anti-noise generated by the ANC system as given in Fig. 1, the objective is to reduce the residual noise.

The encoder employs convolutional layers in RAE1 to map the input noise $p[n]$ to a lower-dimensional latent space as the waveform $z[n]$. This process is represented as:

$$z[n] = f_{enc}(p[n];\ \theta_{enc}), \quad [3]$$

where f_{enc} is the encoding function parameterized by θ_{enc}[23]. The encoder captures local noise patterns, including frequency components and spectral features. The decoder reconstructs the anti-noise from the latent representation:

$$\hat{s}\,[n] = f_{dec}(z[n];\ \theta_{dec}), \quad [4]$$

where $\hat{s}\,[n]$ is the reconstructed signal, and f_{dec} is the decoding function parameterized by θ_{dec}[23].

Noise is inherently non-stationary, with evolving spectral or feature patterns over time. The GRU layer captures these temporal dependencies through the following equations:

$$z_n = \sigma(W_z.[x[n], h[n-1]] + b_z) \qquad [5]$$

$$r_n = \sigma(W_r.[x[n], h[n-1]] + b_r) \qquad [6]$$

$$\tilde{h}_n = tanh(W_h.[x[n], r_n \odot h[n-1]] + b_h) \qquad [7]$$

$$h[n] = (1-z_n) \odot h[n-1] + z_n \odot \tilde{h}_n), \qquad [8]$$

where z_n is the update gate, r_n is the reset gate and $\tilde{h}_n$ is the candidate hidden state. The symbol ⊙ denotes element-wise multiplication[22].

Combining GRU layers with a convolutional autoencoder benefits from both spatial and temporal modeling. The convolutional encoder models spectral features, reducing the noise to a latent representation $z[n]$. This representation encodes essential noise characteristics in a lower-dimensional space. The GRU processes the sequence of latent representations $z[n]$ over time:

$$h[n] = GRU(z[n], h[n-1]) \qquad [9]$$

The decoder reconstructs $\hat{s}[n]$ using the updated hidden state from the GRU. The anti-noise $r[n]$ is generated to reduce the residual noise $r[n] = \hat{s}[n]$.

The system reduces the residual error $e[n]$. The objective is to reduce $e[n]$, optimizing the parameters θ_{enc}, θ_{dec} and the GRU weights. The GRU captures temporal dependencies between consecutive latent representations, predicting future noise characteristics based on past information.

Architecture of the proposed RSAE ANC model : The architecture of the recurrent autoencoders (RAE1 and RAE2) used in the proposed RSAE model consists of an encoder with 5 1D convolutional layers, each with filters [2, 4, 8, 16, 32], which are increasing in number as shown in Fig. 2. These layers have a kernel size of 32, a stride of 2. After encoding, the latent representation is passed through a series of 4 GRU layers, each with 32 units, allowing the model to capture temporal dependencies in the input data. The decoder consists of 1D deconvolutional layers with decreasing filters [32, 16, 8, 4, 2, 1, 1] and the same padding, which reconstructs the output to match the input dimension. To concatenate feature maps from the encoder with the matching layers in the decoder, skip connections are used to retain spatial information from earlier layers and improve network performance.

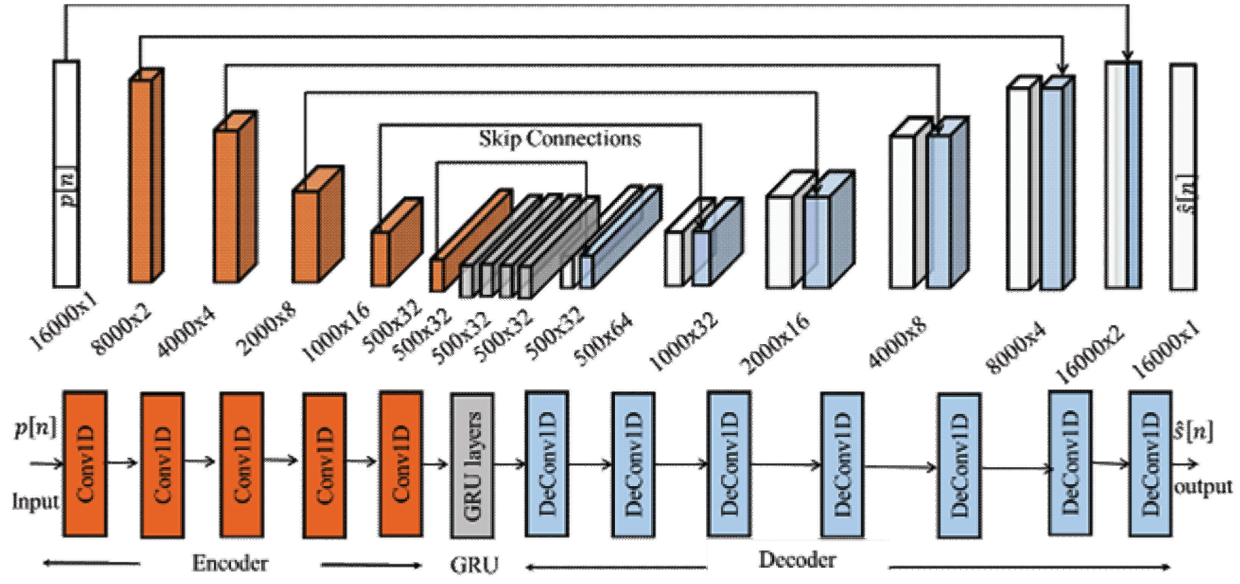


Fig. 2. Architecture of the Recurrent Stacked autoencoder (RSAE) ANC.

All hidden layers employ the rectified linear unit (ReLU) activation function, however the final output layer uses linear activation. The model is trained for 80 epochs using the Adam optimizer, with a learning rate of 0.002, and aims to minimize the mean square error (MSE) loss function. Training is conducted with a batch size of 64. RAE1 and RAE2 individually consist of 123,774 trainable parameters, combining convolutional, GRU, and deconvolutional layers. The network has been designed to optimise time-sequential data performance, leveraging convolutional feature extraction and temporal modeling with GRUs. The training and evaluation procedures are detailed in the following sections.

Training of the RSAE-based ANC model : The RSAE framework consists of two key components: RAE1 and RAE2. In this configuration, RAE2 is designed to replicate the non-linear behaviour of the secondary loudspeaker and its acoustic path. Concurrently, RAE1 learns to predict the input signal $x[n]$, which ultimately generates an anti-noise output $s[n]$. This output, after passing through the secondary path, minimizes the reference noise $r[n]$ at the error microphone through destructive interference.

Step 1 : Training RAE2 on Simulated Noise : First, RAE2 is trained using synthetically generated noise. A delayed and attenuated version of the input is computed to represent the anti-noise target. This process imitates a direct-path, non-reflective acoustic scenario and is mathematically described by Equation (1). The modeling accounts for free-space attenuation and propagation delay due to distance.

Step 2 : Training RAE1 on Simulated Noise : Once RAE2 has been trained, RAE1 is optimized. During this phase, the signal $p[n]$ is input to RAE1, and the target output is the inverse of the reference noise –$r[n]$. RAE1 generates an intermediate prediction $x[n]$, which is fed into RAE2. The output $s[n]$ is evaluated against the desired anti-noise. While training RAE1, the weights of RAE2 remain unchanged to preserve its learned secondary path model.

The target reference signal $r[n]$ is synthesized using Equation (1), where the delay and attenuation parameters are based on the known distance of the primary path. Once trained, RAE1 can be deployed in a physical environment to produce real-time anti-noise through a loudspeaker. The block diagram in Fig. 1 illustrates this RSAE-based system structure. Detailed simulation settings for generating the synthetic training data are outlined in the following section.

3. EXPERIMENT SETUP AND DATA PREPROCESSING

The ANC system is modelled within a rectangular enclosure, as is commonly done in many studies on noise cancellation within confined spaces. For this simulation, we adopt the dimensions of a physical room, measuring 3.34 m × 3.04 m × 2 m (width × length × height)[20]. Additionally, delay and reflection effects are incorporated into both the primary and secondary paths of the single-channel ANC system. The simulated noise dataset was divided into 1-second segments, which was a time series divided in an 80:20 ratio for training and testing. These 1-second frames, each containing 16,000 samples, were provided as input to the RAE1 and RAE2 models.

The proposed model is trained using computer-generated band-limited white noise. To synthesize this noise, 10 hours of white noise distributed normally, sampled at 16 kHz, is produced in MATLAB. Next, it is passed through a finite impulse response (FIR) bandpass filter using Kaiser window to yield a reference signal. Stopband frequencies are 125 Hz and 1800 Hz and passband frequencies between 150 to 1500 Hz are defined. The stopband attenuation is 0.002, and the passband ripple is 0.1145. We use the corresponding filtered noise for training and testing.

Also, some types of real-world noise are used to assess RSAE. These non-white noise samples were obtained from different databased such as DCASE (2017, 2018) [24,25]. These noise samples are used to create 140-minute audio files for each noise type. These noise files are sampled at 16 kHz. Of the 140 minutes per noise type, 80% are used for training while the remaining 20% is set aside for testing. The same band-pass filter of 150-1500 Hz was used to filter all noise samples. In the simulations, it is assumed that the speed of sound is 343 m/s.

4. RESULTS AND DISCUSSION

The following section presents the evaluation results of the RSAE-based ANC algorithm in comparison with the SAE and FxLMS methods. The noise reduction performance is assessed across various real-world noise environments, and the analysis includes both numerical and spectral evaluations.

Noise reduction in real-world environments : Initially, the RSAE ANC approach is evaluated against the SAE method described in[16] and the FxLMS algorithm with a step size of 0.001. The comparison is based on the noise reduction (NR) measured in dB, as follows:

$$NR = 10_{\log_{10}}\left[\frac{\sum_{n=1}^{L} e^2(n)}{\sum_{n=1}^{L} r^2(n)}\right] \qquad [10]$$

where L the signal vector's length.

Table 1. NR for RSAE, SAE and FxLMS ANC Methods.

ANC Method	NR (dB)			
	Band-limited White noise (150-1500 Hz)	**Car**	**Auditorium**	**Industry**
RSAE	25.11	19.60	20.21	20.99
SAE	23.65	16.83	17.52	17.93
FxLMS	1.54	5.43	6.92	7.15

Table 1 contains the NR comparison results for band-limited white, auditorium, car and industry noises. The table indicates that the RSAE method has better performance in all four types of noise than traditional FxLMS and SAE methods. This indicates that the RSAE, which has GRU layers in place of SAE can better capture temporal structure among dynamic noise patterns resulting in a superior performance than an untimed counterpart such as SAE. This is due to the fact that, through its GRU mechanism of carrying over memory from one time step to another, the RSAE can more effectively follow up and learn models for these occurring noise changes, as would both be hardly possible with a simple SAE or FxLMS. This simulation thus validates the efficacy of the RSAE method when dealing with real-world noise scenarios.

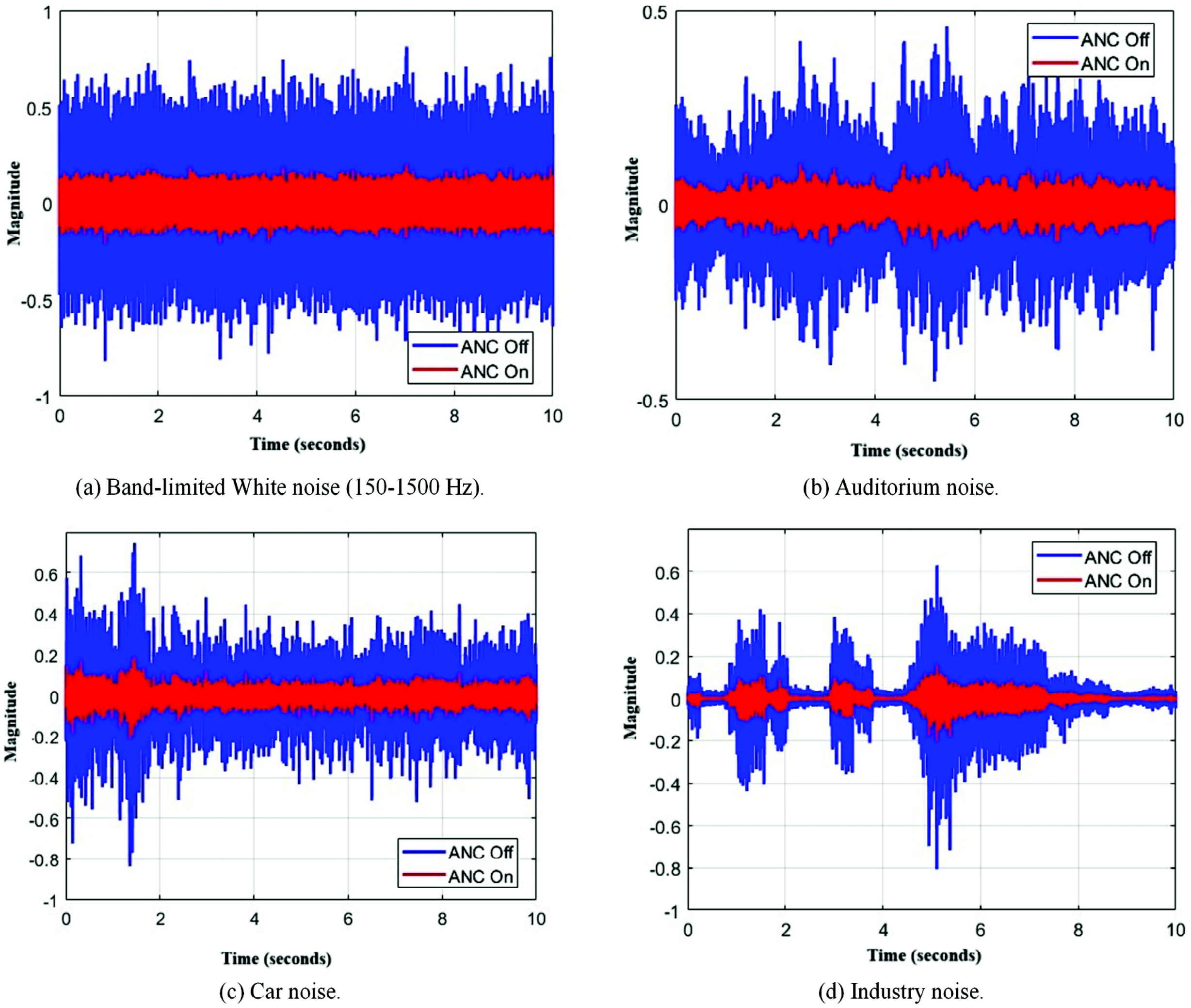
Fig. 3 represents the noise reduction outcomes for RSAE -ANC technique applied to one observation of band-limited white noise, and auditorium, car and industry noise. The noise reduction is observed to be uniform over all time instances. Fig. 4 represents Model loss vs epochs for training and testing data. No over-fitting or under-fitting is observed. The model loss converges well on test data.

Power Spectral Density Analysis : The performance of various ANC algorithms in reducing noise was evaluated using the power spectral density (PSD), as given in Fig. 5. PSD provides a detailed understanding of how power is distributed across various frequency ranges within a noise signal.

Specifically, in every band the RSAE-ANC and SAE ANC algorithms successfully suppress almost all components below the noise. This demonstrates the effectiveness of RSAE-ANC technique in cancelling noise over all frequency bands. We performed PSD calculations in a diverse set of noise environments - band-limited white, auditorium, car and industrial noises.

5. CONCLUSION AND FUTURE SCOPE

This work introduces a novel RSAE-based method and gives details of its performance in ANC. When training the model, we add GRU layers between the encoder and decoder to make it easier for the model to predict temporal dependencies within noise patterns. This allows the system to better predict and cancel out noise over time, improving overall ANC performance. The experimental evaluations demonstrate that the RSAE gives up to 25.11 dB noise reduction. This is more as compared to 1.54 dB and 23.65 dB achieved



(a) Band-limited White noise (150-1500 Hz).

(b) Auditorium noise.

(c) Car noise.

(d) Industry noise.

Fig. 3. Noise reduction results of RSAE-ANC algorithm on different noise types.

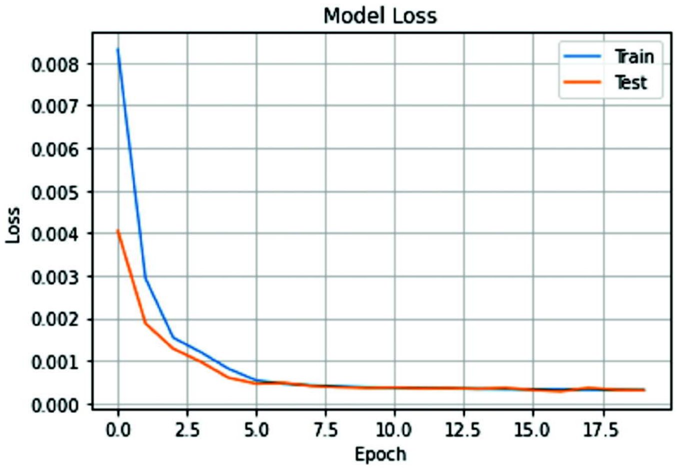
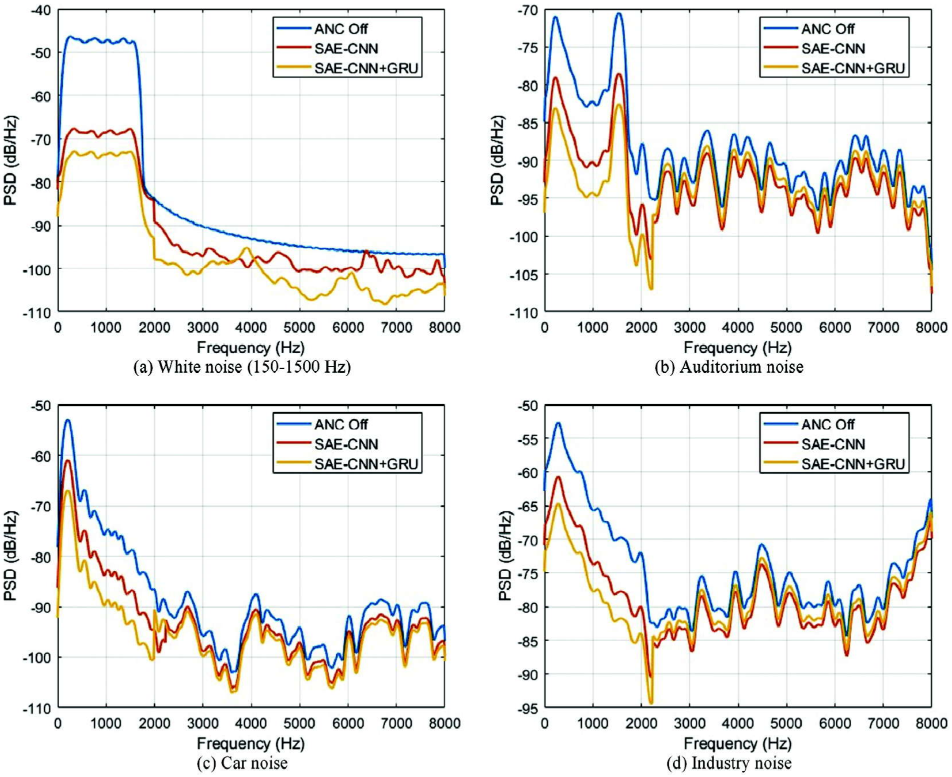


Fig. 4. Model loss vs epochs for train and test data.



(a) White noise (150-1500 Hz)

(b) Auditorium noise

(c) Car noise

(d) Industry noise

Fig. 5. The power spectral density (PSD) of the band-limited white noise, auditorium, car and industry noise by SAE, RSAE ANC algorithms.

by conventional FxLMS and SAE ANC, respectively, in non-reflective environments. Additionally, the model trained on synthetic noise signals can be applied to real recorded noise, demonstrating its robustness, transferability, and practical utility. This highlights the potential of the unsupervised learning approach in real-world scenarios.

REFERENCES

[1] S. M. Kuo and D. R. Morgan, 1999. "Active noise control: a tutorial review," *Proceedings of the IEEE,* **87**(6): 943-973.

[2] S. J. Elliott and P. A. Nelson, 1993. "Active noise control," *IEEE Signal Processing Magazine,* **10**(4): 12-35.

[3] C. N. Hansen, 2002. Understanding active noise cancellation, *CRC Press.*

[4] T. Schumacher, H. Krüger, M. Jeub, P. Vary and C. Beaugeant, 2011. "Active noise control in headsets: A new approach for broadband feedback ANC," in 2011 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 417-420.

[5] Y. Kajikawa, W.-S. Gan and S. M. Kuo, 2012. "Recent advances on active noise control: Open issues and innovative applications," *APSIPA Transactions on Signal and Information Processing,* **1:** e3.

[6] J. Cheer and S. J. Elliott, 2015. "Multichannel control systems for the attenuation of interior road noise in vehicles," *Mechanical Systems and Signal Processing,* **60:** 753-769.

[7] P. R. Benois, R. Roden, M. Blau and S. Doclo, 2022. "Optimization of a fixed virtual sensing feedback and controller for in-ear headphones with multiple loudspeakers," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 8717-8721.

[8] C.-Y. Chang, C.-T. Chuang, S. M. Kuo and C.-H. Lin, 2022. "Multi-functional active noise control system on headrest of airplane seat," *Mechanical Systems and Signal Processing,* **167:** 108552.

[9] F. Yang, J. Guo and J. Yang, 2020. "Stochastic analysis of the filtered-x LMS algorithm for active noise control," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **28:** 2252-2266.

[10] S. Liebich, C. Anemüller, P. Vary, P. Jax, D. Rüschen and S. Leonhardt, 2016. "Active noise cancellation in headphones by digital robust feedback control," in 2016 *24th European Signal Processing Conference (EUSIPCO).* IEEE, pp. 1843-1847.

[11] M. Pawełczyk, 2002. "Analogue active noise control," *Applied Acoustics,* **63**(11): 1193-1213.

[12] N. Pan, J. Chen and J. Benesty, 2022. "DNN based multiframe single-channel noise reduction filters," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, pp. 8782-8786.

[13] H. Sun, J. Zhang, T. Abhayapala and P. Samarasinghe, 2022. "Spatial active noise control with the remote microphone technique: An approach with a moving higher order microphone," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, pp. 8707-8711.

[14] D. Shi, W.-S. Gan, B. Lam and S. Wen, 2020. "Feedforward selective fixed-filter active noise control: Algorithm and implementation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **28:** 1479-1492.

[15] C. Shi, R. Xie, N. Jiang, H. Li and Y. Kajikawa, 2019. "Selective virtual sensing technique for multichannel feedforward active noise control systems," in 2019 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, pp. 8489-8493.

[16] D. Shi, B. Lam, K. Ooi, X. Shen and W.-S. Gan, 2022. "Selective fixed-filter active noise control based on convolutional neural network," *Signal Processing,* **190:** 108317.

[17] D. Chen, L. Cheng, D. Yao, J. Li and Y. Yan, 2021. "A secondary path-decoupled active noise control algorithm based on deep learning," *IEEE Signal Processing Letters,* **29:** 234-238.

[18] Y. J. Cha, A. Mostafavi and S. S. Benipal, 2023. "Dnoisenet: Deep learning-based feedback active noise control in various noisy environments," *Engineering Applications of Artificial Intelligence,* **121:** 105971.

[19] H. Zhang and D. Wang, 2021. "Deep ANC: A deep learning approach to active noise control," *Neural Networks*.

[20] D. Singh, R. Gupta, A. Kumar and R. Bahl, 2024. "Enhancing active noise control through stacked autoencoders: Training strategies, comparative analysis, and evaluation with practical setup," *Engineering Applications of Artificial Intelligence,* **135:** 108811.

[21] S. T. Rajamani, K. T. Rajamani, A. Mallol-Ragolta, S. Liu and B. Schuller, 2021. "A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, pp. 6294-6298.

[22] K. Irie, Z. Tüske, T. Alkhouli, R. Schlüter and H. Ney, 2016. "LSTM, GRU, highway and a bit of attention: An empirical overview for language modeling in speech recognition," in *Interspeech,* pp. 3519-3523.

[23] R. Rana, 2016. "Gated recurrent unit (GRU) for emotion classification from noisy speech," *arXiv preprint arXiv:1612.07778*.

[24] S. R. Park and J. Lee, 2016. "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*.

[25] A. Rakotomamonjy and G. Gasso, 2014. "Histogram of gradients of time-frequency representations for audio scene classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **23:** 142-153.

The Journal of Acoustical Society of India : Vol. 52, No. 3, 2025 (pp. 132-138)

Optimizing acoustic enclosures using particle swarm optimization (PSO)

N.K. Vijayasree and B. Venkatesham*
Dept. of Mechanical and Aerospace Engineering, IIT Hyderabad, India
e-mail: venkatesham@mae.iith.ac.in

[Received: 19-03-2025; Accepted: 30-04-2025]

ABSTRACT

Acoustic enclosures are essential for attenuating noise from individual sources in various industrial applications. These enclosures are constructed using panels made of acoustic materials, coupled with structural elements, to achieve the desired noise reduction. The performance of these panels is measured in terms of Insertion Loss (IL), which quantifies the reduction in noise levels achieved by the enclosure. An effective method to evaluate IL is Statistical Energy Analysis (SEA). SEA divides the system into multiple subsystems, describing the flow of energy-input, storage, transmission, and dissipation within each subsystem. The absorption and transmission loss of the acoustic panels can be calculated using either analytical equations or experimental data, as a function of frequency. Traditionally, designing these enclosures involves a practice of design sufficiency, where materials and panel thickness are selected based on achieving sufficient IL. This process often relies on trial and experience, which can be time-consuming and may not yield optimal results. To address this, the current paper proposes a Particle Swarm Optimization (PSO) algorithm to optimize panels' selection and thickness. PSO is an iterative method that improves a potential solution by evaluating its performance parameters. Here, a group of possible solutions, called a swarm, move towards the optimum point through the search space, adjusting their positions based on calculated velocity and direction, influenced by both local best and global best positions. The proposed methodology is demonstrated for a particular chosen case. This approach offers a systematic and efficient alternative to traditional design methods, ensuring better performance and resource utilization in the design of acoustic enclosures.

1. INTRODUCTION

Acoustic enclosures are widely used in many engineering fields to mitigate noise from different sources like running machines, engines, etc. The purpose of enclosure is to surround the noise source with a set of specific materials, to limit the transmission outwards. Materials include a combination of acoustic materials along with structural components. Acoustic materials contribute to noise reduction by absorbing sound, while structural components provide transmission loss. The overall noise reduction, which is a combination of absorption and transmission losses, is referred to as insertion loss. This can be determined through acoustic measurements or theoretical models.

Mole et al. used a hybrid approach consisting of experimental DOE and simulations based on Boundary Element Methods (BEM) to design enclosures for DG sets[1]. Kosala et al. carried out experiments to predict the insertion-loss for enclosures with various combinations of structural and acoustic materials [2]. Szemela et.al. provided an analytical solution for closed cavities considering different sub-regions and continuity conditions on the region's coupling interface[3]. Though many different approaches and mathematical formulations have been explored for evaluating enclosures and noise reduction, Statistical Energy Analysis (SEA) has proven to be a very useful technique.

The approach of SEA is to break up the given system into subsystems. Subsystems are a division of several physical elements so that the vibro-acoustic characteristics are similar over them like damping, excitation and coupling properties. SEA then models the entire system and the energy distribution over the subsystems with the help of power balance equations. Pavan Gupta et.al. implemented SEA techniques to evaluate transmission loss across cylindrical enclosure and validated the results obtained experimentally[4]. An improved SEA model including the non-resonant response and more accurate transmission coefficient of finite panels is presented by Lie et al., for rectangular structures, followed by measurement[5].

Apart from the analysis, attempts have been made to optimize the design of enclosures. Mushiri et al. designed an optimum enclosure for diesel generator set overheat aspects by experimentation with different available materials[6]. Prasad et al. also attempted optimization of enclosures for high-capacity diesel generators by studying the influence of various factors through parametric studies with SEA techniques[7].

Though multiple studies have been carried out for SEA techniques and their implementation of acoustic enclosure performance evaluations, proper reported literature is not available on multiple parameter optimization of the same. The current paper describes a method to design an optimum acoustic enclosure to meet the desired acoustic specifications with minimum cost using Particle Swarm Optimization Algorithms.

2. PROBLEM FORMULATION FOR OPTIMIZATION

For given dimensions of the acoustic enclosure for a known equipment (source of noise), the objective function for optimization is minimizing the cost of the enclosure meeting the objective and specific acoustic requirements.

- Objective - Minimum Cost - The cost of enclosure is defined by the choice of two materials - The structural material (C_s) and the acoustic material (C_a).

$$Minimize\ C,\ where\ C = C_s + C_a \tag{1}$$

$$Cs = mass * C_{per\ kg},\ C_a = C_{per\ area} \times S_a \tag{2}$$

- Parameter - Acoustic and structural material selection - From the given database, algorithm has to find the suitable materials meeting the objective and acoustic requirements.
- Parameter - Area of acoustic material lining - One of the variables in the optimization is the area over which the acoustic material is lined (S_a). Though the entire surface area of the enclosure is available for lining, area for material can be obtained based on the performance parameters.

$$S_{a,\ surface} \leq S_{surface} \tag{3}$$

- Constraint - Sufficient insertion loss - The algorithm has to ensure sufficient insertion loss as per requirements. For the current work, limit has been defined for average Insertion loss across the frequencies considered along with limits for the minimum Insertion loss obtained over the distinct frequencies.

$$IL_{avg} \geq IL_{avg\_lim},\ IL_{min} \geq IL_{min\_lim} \tag{4}$$

3. SEA METHOD FOR ACOUSTIC ENCLOSURE EVALUATION

The acoustic property of the enclosures is quantified by insertion loss, which is defined as the sound power level difference between internal and external sound fields,

$$IL = SWL_{in} - SWL_{out} \tag{5}$$

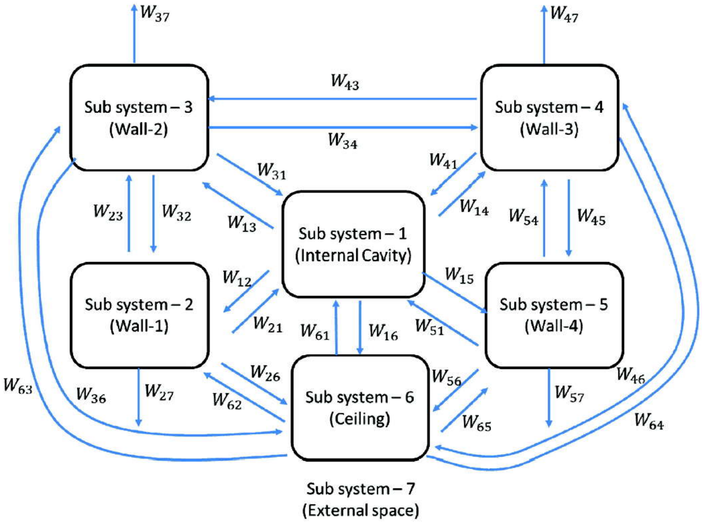


Fig. 1. Representation of subsystems and exchange of powers between them.

Where, SWL_{in}, SWL_{out} represent the source power and power radiating out from the enclosure respectively. For implementing SEA, the model needs to be divided into individual subsystems with an assumption of diffuse energy in each subsystem. Accordingly, the enclosure is divided into seven subsystems: four walls, ceiling, the internal air cavity and the external air space. For the current analysis, the floor is not considered for transmission nor treatment. The numbering of subsystems is established in the direction of flow of energy from the internal air cavity into the structures of enclosure and transmitted to the external space as shown in Fig. 1.

There are four essential parameters in the study of SEA:

- Internal loss factor - The energy lost due to internal losses within the air cavity. The absorption of material contributes to this loss factor.
- Dissipation Loss factor - Measure of the energy loss rate of a mode of oscillation in a dissipative system, which is a combination of structural damping and radiation loss factor.
- Coupling loss factor - The coupling loss factor is the fraction of energy transmitted from one subsystem to another. It is associated with the energy transmitted from one subsystem to another. The formulation of coupling loss factors depends on the type of junctions and the properties of the subsystems. The coupling loss factor is a frequency dependent function, and the formulations differ from the frequency of calculation and the critical frequency of the subsystem.
- Non-resonant coupling loss factor - The basic assumption of the SEA model is that the energy in each subsystem is contained in resonant modes so that the energy is proportional to the damping. However, when the proportionality ceases, the excited behavior is non-resonant, where transmission occurs through the system at frequency below the resonant frequency.

The total loss factor is determined by the summation of all the above-mentioned factors. From the loss factor *(η)*, the energy from one subsystem *(i)* to another *(j)* is calculated, which is further used to evaluate the total sound power emanating from the system. The details of the equations involved in calculating the various loss factors and estimating the Insertion Loss is provided in Reference [8].

4. DESCRIPTION OF PSO ALGORITHM

Particle Swarm Optimization (PSO) is a bio-inspired algorithm where each member of the swarm learns from both its own experiences and those of other members. This collective learning process helps the swarm to develop an effective search pattern. The flow chart of a basic PSO algorithm is given in Fig. 2.

The initial swarm of particles is generated randomly, with particles that meet the acoustic parameters and constraints being included in the swarm. Once the swarm is generated, the particles are then updated using the velocity update equation:

$$v_{i,(k+1)} = \omega v_{i,k} + c_1 r_1(\chi_{lb\_i,k} - \chi_{i,k}) + c_2 r_2(\chi_{gb} - \chi_{i,k}) \tag{6}$$

Here $\chi_{lb\_i,k}$ denotes the local best of each particle in the swarm over the iterations; χ_{gb} denotes the best particle across the swarm over the iterations; ω is the chosen inertia weight, c_1 and c_2 are the cognitive and social factors. r_1, r_2 are randomly chosen numbers between [0,1]. Velocity calculation from Equation (5) consists of three terms on the RHS.

- The first term denotes the momentum/inertia part which gives weightage to the particle's current direction ($v_{i,k}$), thus avoiding drastic changes in the particle movement.

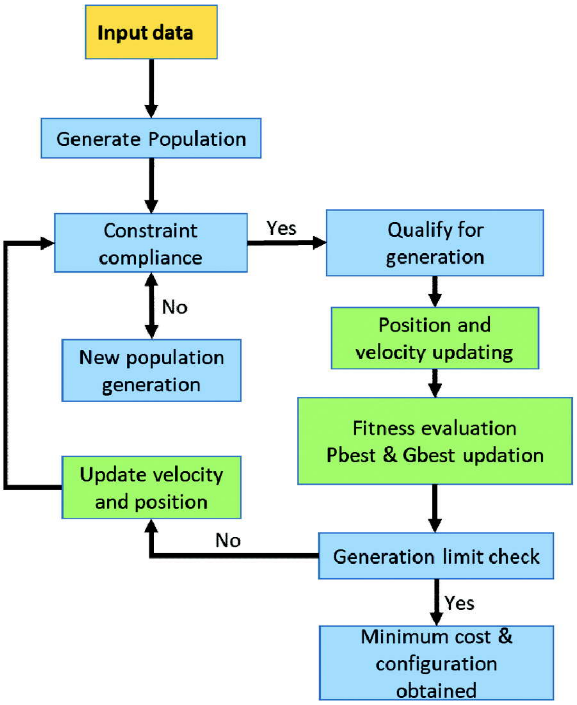

Input data
Generate Population
Constraint compliance
Yes
Qualify for generation
No
New population generation
Position and velocity updating
Fitness evaluation Pbest & Gbest updation
Update velocity and position
No
Generation limit check
Yes
Minimum cost & configuration obtained

Fig. 2. PSO Algorithm.

- The second term denotes the cognitive part which quantifies the current position of the particle relative to its past best position, $(\chi_{lb,i,k})$.
- The third term denotes the social part which quantifies the current position of the particle relative to the best particle obtained across the particles over the iterations (χ_{gb}).

From the calculated velocity, the updated position of the swarm is defined as :

$$\chi_{i,k+1} = \chi_{i,k} + v_{i,k+1} \tag{7}$$

During each iteration, particles are checked for compliance with parameter and constraint sufficiency. Only healthy particles are carried forward to the next iterations. New particles are introduced as needed to maintain swarm size. The process continues until convergence criteria are met.

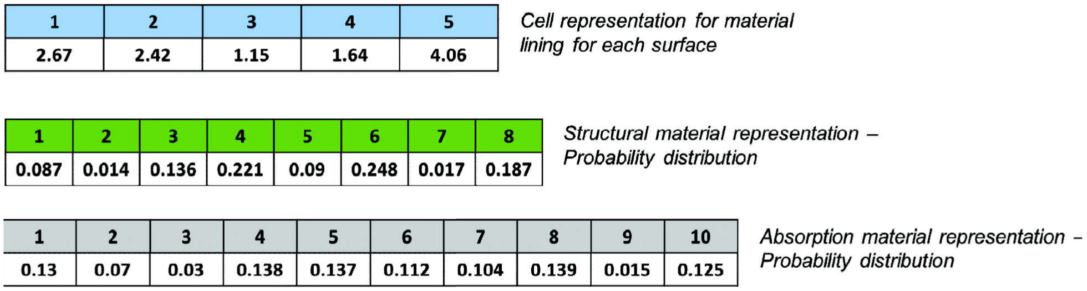
5. IMPLEMENTATION OF PSO ALGORITHM TO CURRENT PROBLEM STATEMENT

For the current case, the variables considered are:

- Area representation - A cell is represented for each available surface for lining. Accordingly, a total of five cells are represented for four walls and one ceiling respectively.
- Choice for absorption material - From the given database of materials, most suitable material needs to be selected by the algorithm. As this needs discrete variable representation, the same is converted to continuous domain by a random probability distribution across the entire set of available materials. For example, for the current case, a dataset of 10 acoustic materials is considered. Accordingly, one cell with a random probability is assigned for each material. This array of cells shall be used for all future iterations to carry out velocity and position updates. After the iteration, the material with highest probability is chosen as the suitable material.

- Choice for structural material - As explained for absorption material selection, cells are defined with probability distribution for structural material selection as well. For the current study, a material database consisting of eight materials is considered for evaluation.

With the above-mentioned variables, each particle of swarm is represented as given in Fig. 3.



1	2	3	4	5
2.67	2.42	1.15	1.64	4.06

Cell representation for material lining for each surface

1	2	3	4	5	6	7	8
0.087	0.014	0.136	0.221	0.09	0.248	0.017	0.187

Structural material representation – Probability distribution

1	2	3	4	5	6	7	8	9	10
0.13	0.07	0.03	0.138	0.137	0.112	0.104	0.139	0.015	0.125

Absorption material representation – Probability distribution

Fig. 3. Swarm particle representation

Once the particles are defined, the insertion loss is evaluated for each particle through SEA formulation. The fitness and ranking of each particle are determined and improved over iterations using PSO algorithm.

6. RESULTS AND DISCUSSION

The discussed PSO algorithm has been implemented over a known acoustic source of 120 dB power level (taken constant across frequencies). The dimensions of the enclosure considered are 4 x 1.25 x 2m. The insertion loss for each of the combinations of area and material selection, is evaluated using SEA techniques discussed. The computations were executed on a personal laptop with an Intel i5 1.8 GHz processor and 8GB of RAM. The results obtained for an average insertion loss of 15 dBA and minimum insertion loss of 10 dBA are provided in Table 1.

Table 1. Optimized acoustic lining and structural material chosen by algorithm.

Surface	Wall-1	Wall-2	Wall-3	Wall-4	Ceiling
Available area (m^2)	8	2.5	8	2.5	5
Lining area (from algorithm)	0.1442	0.69	0.11	0.065	0.0839
Acoustic material identified	Roxul Rockwool Safe n Silent Pro Acoustic Insulation Boards				
Cost of acoustic lining	₹ 620				
Structural material identified	16gage stainless steel				
Cost of structure	₹ 59530				
Total cost	₹ 60150				

The insertion loss plot is given in Fig. 4.

7. SUMMARY

The current paper discusses PSO algorithm for exploring cost optimization aspects while designing enclosures for meeting required acoustic parameters. The formulation for acoustic parameters and constraints along with the objective function of cost minimization has been described. Acoustic and structural material database suitable for the enclosure design has also been prepared (Annexure-I). The algorithm has been implemented to a known acoustic source and predetermined enclosure dimensions, where the choice of materials along with lining areas is provided as output. The algorithm implementation is also planned to be extended to design of enclosures with multiple acoustic layers.

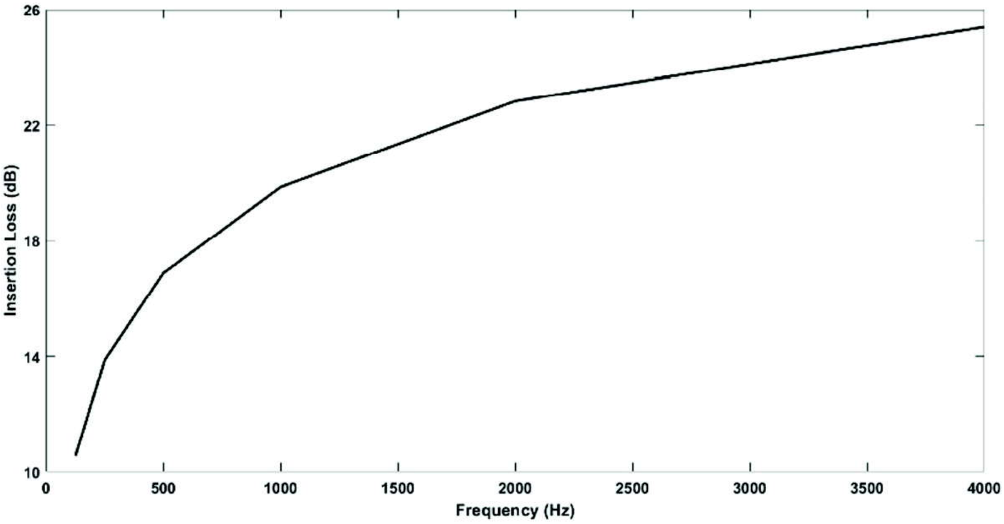


Fig. 4. Calculated Insertion loss of enclosure with optimized treatment

Annexure – I

Acoustic and Structural Material database

Material	Cost (per m^2) / Thickness (mm)	Absorption coefficient at frequency (Hz)					
		125	250	500	1000	2000	4000
Roxul Rockwool Pro Acoustic Insulation Board	₹560 / 50mm	0.15	0.8	1	1	1	1
Foam Acoustic Panel	₹1000 / 50mm	0.17	0.29	0.61	0.95	1	1
Foam Acoustic Panel	₹722 / 25mm	0.12	0.23	0.44	0.75	0.9	0.91
Knauf acoustic slabs	₹1502 / 50mm	0.16	0.38	0.64	0.86	0.93	0.95
Knauf ECOSE acoustic roll	₹950 / 50mm	0.19	0.51	0.74	0.89	0.88	0.88
Polyfill Acoustic Board	₹797.5 / 50mm	0.61	0.68	0.84	0.97	0.99	1
Thermafleece wool roll	₹749 / 50mm	0.2	0.4	0.65	0.75	0.85	0.95
Thermafleece wool roll	₹1145 / 100mm	0.25	0.6	0.85	0.9	0.95	1
Thermafleece wool roll	₹1468 / 140mm	0.6	0.9	0.95	1	1	1

Structural material	Cost (per kg)/ thickness	Transmission coefficient at frequency (Hz)					
		125	250	500	1000	2000	4000
SS 316 L sheet	₹297 / 4mm	0.03	0.015	0.008	0.004	0.002	0.001
Galvanized sheet	₹70 / 0.5mm	0.25	0.13	0.06	0.032	0.016	0.008
SS 304 sheet	₹233 / 0.64mm	0.19	0.1	0.05	0.02	0.012	0.006
SS 304 sheet	₹180 / 1.59mm	0.08	0.04	0.02	0.01	0.005	0.003
14 gage SS sheet	₹260 / 1.98mm	0.07	0.03	0.02	0.008	0.004	0.002
Galvanized Iron	₹53 / 2mm	0.06	0.03	0.016	0.008	0.004	0.002
Sheet Molding Compound (SMC)	₹195 / 1mm	0.56	0.28	0.14	0.07	0.04	0.018
SMC	₹250 / 5mm	0.11	0.06	0.03	0.01	0.007	0.004

REFERENCES

[1] Mole, D., Yadav, P., Kandalkar, M., Karanth, N. *et al.,* 2013. Design Optimization of Acoustic Enclosure for Noise Reduction of Diesel Generator Set, SAE Technical Paper 2013-26-0108.

[2] Krzysztof Kosala, Leszek Majkut and Ryszard Olszewski, 2020. Experimental Study and Prediction of Insertion Loss of Acoustical Enclosures, *Vibrations in Physical systems,* **31:** 202029.

[3] Krzysztof Szemela, Mirosław Meissner, Wojciech P. Rdzanek, 2023. Efficient analytical method for computing the acoustic field inside enclosures with a mixed rectangular-cylindrical geometry, *Applied Acoustics*, 201.

[4] Pavan Gupta and Anand Parey, 2022. Prediction of sound transmission loss of cylindrical acoustic enclosure using statistical energy analysis and its experimental validation, *Journal of Acoustical Society of America,* **151:** 544-560.

[5] Y. Lei, J. Pan and M.P. Sheng, 2012. Investigation of structural response and noise reduction of an acoustical enclosure using SEA method, *Applied Acoustics,* (73), 348-355.

[6] Tawanda Mushiria, Nyasha Madzirob and Charles Mbohwac, 2016. Design of an optimum acoustic enclosure for an open frame diesel generator, 14th Global Conference on Sustainable Manufacturing.

[7] Prasad Yadav, Harshal Bankar and Nagesh Voderahobli Karanth, 2017. Acoustic Enclosure Optimization for a Higher Capacity Diesel Generator Set Using Statistical Energy Analysis (SEA) Based Approach, Symposium on International Automotive Technology.

[8] Himanshu S. Malushte, Evaluation of Statistical Energy Analysis for Prediction of Breakout Noise from Air Duct, PhD Thesis, University of Nebraska- Lincoln

The Journal of Acoustical Society of India : Vol. 52, No. 3, 2025 (pp. 139-144)

Deep learning method for underwater sound classification using passive acoustic data

Keerthivasan R.*, Raghul M., Sanjana M.C. and Latha G.
National Institute of Ocean Technology, Chennai-600 100, India
e-mail: keerthivasan1206@gmail.com

[Received: 00-00-2025; Accepted: 00-00-2025]

ABSTRACT

In recent years, deep learning has emerged as a powerful tool for analyzing complex datasets, including underwater passive acoustic data. This paper presents a study on the application of Convolutional Neural Networks (CNNs) for the identification and classification of underwater biological sources. CNNs, a type of deep neural network (DNN) within supervised learning, have shown great promise in such tasks. The primary objective of this research is to develop an efficient and accurate method for classifying underwater bioacoustic signals using CNNs. Underwater audio recordings were pre-processed and converted into Mel-spectrograms, which served as input for the CNN. Spectrograms provide a visual representation of the frequency spectrum of the audio signal over time, allowing the CNN to effectively capture and analyze patterns within the data. CNNs can automatically extract hierarchical features from raw audio data using convolutional layers, which detect local patterns and temporal structures crucial for sound classification. The CNN architecture consists of several convolutional and pooling layers, followed by fully connected layers, designed to optimize feature extraction and classification performance. The acoustic datasets were split into training and validation sets to evaluate the performance of the model. The model was trained using the Adam optimizer and categorical cross-entropy loss function to achieve optimal results. A detailed analysis of the model's performance was conducted using metrics such as accuracy, precision, recall, and F1 score. Confusion matrices were generated to visualize the classification results and highlight areas for potential improvement. The findings suggest that CNNs are suitable for underwater sound classification of biological sources, offering a robust solution for passive acoustic monitoring in marine environments. The results underscore the potential of CNNs in facilitating more accurate and efficient monitoring of marine ecosystems and underwater activities.

1. INTRODUCTION

In recent years, the application of deep learning techniques has revolutionized various fields of data analysis, offering powerful tools for interpreting complex datasets[1]. One domain that stands to benefit significantly from these advancements is the analysis of underwater passive acoustic data, which plays a crucial role in marine ecosystem monitoring[2,3,4]. Accurate identification and classification of underwater biological sources are imperative for marine biology, environmental monitoring, and conservation. Traditionally, the analysis of underwater sounds has relied on manual inspection and expert knowledge, a process that is both time-consuming and susceptible to human error. As the volume of acoustic data collected from marine environments continues to grow, there is an urgent need for automated methods that can efficiently and accurately analyze these sounds[5,6,7].

Convolutional Neural Networks (CNNs), a class of deep neural networks (DNNs) within supervised learning, have demonstrated remarkable efficacy in image and signal processing tasks[1,8]. CNNs are particularly adept at recognizing hierarchical patterns, making them well-suited for the analysis of spectrograms and other time-frequency representations of audio signals. Recent studies have shown that CNNs can outperform traditional machine learning techniques in bioacoustic classification tasks [9,10]. This research aims to harness the capabilities of CNNs to develop an efficient and accurate method for classifying underwater bioacoustic signals[11]. The methodology involves preprocessing underwater audio recordings and converting them into Mel-spectrograms, which serve as inputs to the CNN, enabling the network to extract meaningful features and patterns from the raw audio data[12,13].

The CNN architecture designed for this study comprises convolutional and pooling layers, followed by fully connected layers, optimized for feature extraction and classification performance. The model is trained using the Adam optimizer and categorical cross-entropy loss function, with its performance evaluated through metrics such as accuracy, precision, recall, and F1 score[14]. Confusion matrices are generated to visualize classification results and identify areas for improvement. The findings of this research highlight the suitability of CNNs for underwater sound classification of biological sources, offering a robust and automated solution for passive acoustic monitoring in marine environments. The results underscore the potential of CNNs to enhance the accuracy and efficiency of marine ecosystem monitoring and contribute to improved conservation efforts[7,9].

2. DATA AND METHODOLOGY

This study leverages a comprehensive dataset of underwater audio recordings sourced from the Deepwater Ambient Noise Measurement System (DANMS) by the National Institute of Ocean Technology and the Macaulay Library, capturing biological sounds from dolphins. The dataset comprises a total of 564 recordings and the summary of the dataset is given in Table 1. While the sampling rate and duration vary across different recordings and collection periods, the recordings from DANMS are standardized with a sampling rate of 25 kHz and a duration of 75 seconds per recording. The dataset is categorized into four classes of dolphin vocalizations, labelled D1, D2, D3, and D4 as shown in Fig 1. D1, D3, and D4 are represented as individual, separate signals, while D2 is characterized as a mixed signal, typically observed when sounds are produced simultaneously by a group of dolphins.

Table 1. Summary of dataset.

Class	Description	Number of Samples
D1	Individual dolphin signal	64
D2	Mixed/group signal	230
D3	Individual dolphin signal	172
D4	Individual dolphin signal	98

To prepare these raw audio recordings for analysis, the data underwent a series of preprocessing steps. Initially, the audio recordings were segmented into smaller clips of

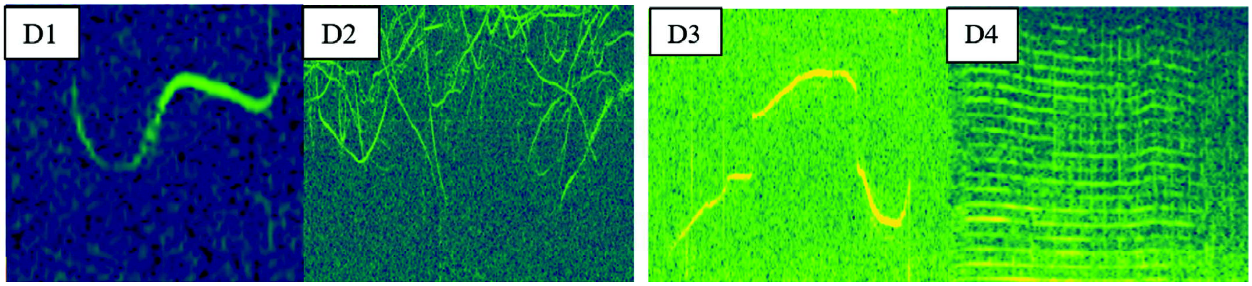


Fig. 1. The types of dolphin signals which are trained using CNN

uniform length to facilitate manageable input sizes for the CNN. These audio segments were then transformed into Mel-spectrograms, which visually represent the frequency spectrum over time, enhancing the richness of the input data[7,8].

The CNN architecture shown in Fig. 2, includes several layers tailored to optimize feature extraction and classification performance. The architecture starts with an image input layer that accepts spectrograms. This is followed by five convolutional layers (each with 2D convolution), where each convolutional layer is immediately followed by a batch normalization layer and a Rectified Linear Unit (ReLU) activation function to introduce non-linearity. After the first, second, third, and fifth convolutional blocks, max pooling layers are used to reduce the spatial dimensions of the data, decreasing computational load and helping to prevent overfitting. A dropout layer is included after the final pooling layer to further reduce overfitting. The final stage consists of a fully connected (dense) layer that interprets the features extracted by the convolutional layers and makes the final classification decisions, concluding with a softmax activation function to generate probability distributions over the predefined classes, followed by the classification output layer[10,15].

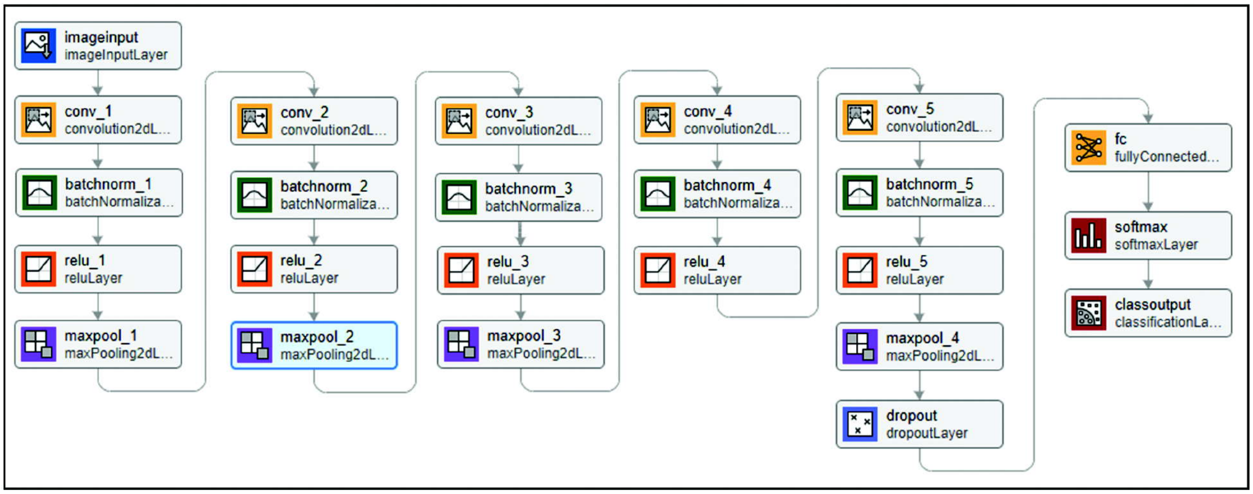


Fig. 2. The Flowchart of CNN architecture

For the training process, the dataset was divided into 80% for training and 20% for validation to ensure a robust evaluation of the model. The training employed the Adam optimizer, known for its efficiency and adaptive learning rate capabilities, and used categorical cross-entropy as the loss function, appropriate for multi-class classification tasks[14]. To enhance model generalization and prevent overfitting, techniques such as early stopping and dropout were applied[16]. The model's performance was rigorously evaluated using metrics including accuracy, precision, recall, and F1 score, with confusion matrices generated to visualize the classification outcomes.

3. RESULTS AND DISCUSSION

3.1 Model Performance

The performance of the Convolutional Neural Network (CNN) model was evaluated using the validation dataset. The key metrics used to assess the model's performance included accuracy, precision, recall, and F1 score. These metrics provide a comprehensive evaluation of the model's classification capabilities shown in Table 2.

The CNN model achieved an accuracy of 94.7%, indicating a high level of overall correctness in the classification of underwater biological sounds. Precision and recall values were also robust, suggesting

Table 2. Performance Metrics.

Metric	Value (%)
Accuracy	94.7
Precision	97.7
Recall	88.5
F1 Score	91.3

that the model was both accurate in its positive predictions and efficient in identifying true positives. The F1 score, which balances precision and recall, further corroborates the model's reliability.

3.2 Analysis of Results

1. *Confusion Matrix Insights:* The confusion matrix provides deeper insights into the classification performance. Most classes were accurately identified, with minor misclassifications occurring in acoustically similar classes. This suggests that while the model effectively captures distinct features, there is still room for improvement in distinguishing subtle acoustic differences. Figure 3. Shows the illustration of confusion matrix.
2. *Feature Importance:* The mel-spectrograms as input features significantly contributed to the accuracy of the model. Spectrograms, in particular, enabled the CNN to effectively capture temporal and spectral patterns, which are crucial for sound classification.
3. *Training Dynamics:* The training process showed a steady improvement in accuracy and loss reduction over epochs. Early stopping and dropout techniques were effective in preventing overfitting, ensuring that the model generalizes well to unseen data.

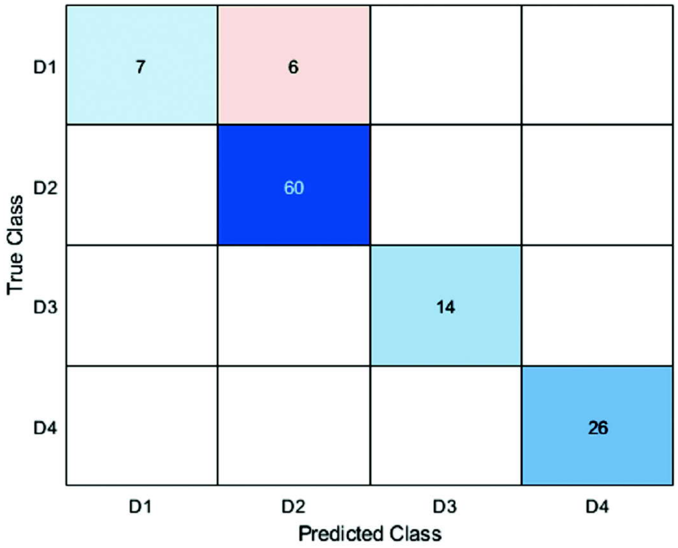


Fig. 3. A confusion matrix illustrating the true vs. predicted classifications, highlighting areas of high accuracy and common misclassifications.

3.3 Potential for Improvement

While the model demonstrated strong performance, several areas for potential improvement were identified:

1. *Data Augmentation:* Implementing data augmentation techniques, such as adding noise or shifting audio clips, could enhance the model's robustness to variations in the data.

2. *Advanced Architectures:* Exploring more advanced CNN architectures, such as ResNet or DenseNet, may further improve feature extraction and classification performance.
3. *Transfer Learning:* Utilizing pre-trained models on similar datasets could accelerate the training process and improve accuracy, especially when dealing with limited data.
4. *Integration with Other Methods:* Combining CNNs with other deep learning techniques, such as Recurrent Neural Networks (RNNs), LSTM or attention mechanisms, could improve the model's ability to capture temporal dependencies in the audio data.

3.4 Implications for Marine Ecosystem Monitoring

The successful application of CNNs to classify underwater biological sounds has significant implications for marine ecosystem monitoring. Automated classification systems can facilitate continuous and large-scale monitoring of marine environments, providing valuable data for researchers and conservationists[4,9]. The ability to accurately identify and monitor marine species can aid in understanding population dynamics, behaviours, and the impact of environmental changes.

4. CONCLUSION

The results of this study underscore the potential of Convolutional Neural Networks in advancing the field of underwater acoustic monitoring. The developed model offers a robust and efficient solution for classifying underwater biological sounds, contributing to more accurate and comprehensive marine ecosystem assessments. Future work will focus on refining the model by incorporating data augmentation, exploring advanced architectures, integrating additional data sources, and conducting statistical significance testing to further enhance performance and reliability.

5. ACKNOWLEDGMENTS

The authors are thankful to the Director, National Institute of Ocean Technology, for his encouragement in carrying out this work. The authors gratefully acknowledge the funding support from the Ministry of Earth Sciences, Government of India. The authors sincerely thank the Ocean Acoustics Group of NIOT for their help in calibration of the hydrophones. We acknowledge The Macaulay Library at the Cornell Lab of Ornithology for providing the media used in this study.

REFERENCES

[1] Y. LeCun, Y. Bengio and G. Hinton, 2015. Deep learning, *Nature,* **521**(7553): 436-444.

[2] F. Caruso *et al.,* 2020. Passive acoustic monitoring of marine ecosystems, *Annual Review of Marine Science,* **12:** 205-226.

[3] M. Goodwin *et al.,* 2022. Advances in underwater acoustic monitoring technologies, *Frontiers in Marine Science,* **9:** 789543.

[4] M.A. Roch et al., 2011. Marine mammal call classification using spectral shape features, *J. Acoust. Soc. Am.,* **130**(5): 2972-2983.

[5] Aide et al., 2013. Automated acoustic monitoring of animal populations, *Methods in Ecology and Evolution,* **4**(7): 675-681.

[6] R. Gibb et al., 2019. Emerging opportunities and challenges in passive acoustic monitoring, *Ecological Informatics,* **51:** 1-11.

[7] J. Salamon and J.P. Bello, 2017. Deep convolutional neural networks and data augmentation for environmental sound classification, *IEEE Signal Process. Lett.,* **24**(3): 279-283.

[8] K.J. Piczak, 2015. Environmental sound classification with convolutional neural networks, 2015 *IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1-6.

[9] D. Stowell *et al.,* 2019. Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge, *Methods Ecol. Evol.,* **10**(3): 368-380.

[10] E.C. Knight *et al.,* 2017. Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs, *Avian Conservation and Ecology,* **12**(2): 14.

[11] J.N. Kather *et al.,* 2024. Deep learning in bioacoustics: Approaches and applications, *Bioacoustics,* **33**(1): 45-67.

[12] P.R. White *et al.,* 2022. Mel-spectrogram enhancement for bioacoustic classification, *J. Acoust. Soc. Am.,* **151**(4): 2563-2572.

[13] A.N. Allen *et al.,* 2021. Convolutional neural networks for marine mammal vocalization detection, *Marine Mammal Science,* **37**(2): 553-573.

[14] D.P. Kingma and J.L. Ba, 2015. Adam: A method for stochastic optimization, *Proceedings of the 3rd International Conference on Learning Representations (ICLR).*

[15] Goodfellow, Y. Bengio and A. Courville, 2016. Deep Learning, MIT Press, Cambridge.

[16] N. Srivastava *et al.,* 2014. Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.,* **15**(1): 1929-1958.

The Journal of Acoustical Society of India : Vol. 52, No. 3, 2025 (pp. 145-154)

Learning from limited labels: Transductive graph label propagation for Indian Music Analysis

Parampreet Singh[1*], Akshay Raina[2], Sayeedul Islam Sheikh[3] and Vipul Arora[4]
[1,2,4]*Department of Electrical Engineering, Department of Chemical Engineering*
[1,2,3,4]*Indian Institution of Technology, Kanpur, India*
e-mail: params21@iitk.ac.in

[Received: 02-06-2025; Accepted: 15-07-2025]

ABSTRACT

Supervised machine learning frameworks rely on extensive labeled datasets for robust performance on real-world tasks. However, there is a lack of large annotated datasets in audio and music domains, as annotating such recordings is resource-intensive, laborious, and often require expert domain knowledge. In this work, we explore the use of label propagation (LP), a graph-based semi-supervised learning technique, for automatically labeling the unlabeled set in an unsupervised manner. By constructing a similarity graph over audio embeddings, we propagate limited label information from a small annotated subset to a larger unlabeled corpus in a transductive, semi-supervised setting. We apply this method to two tasks in Indian Art Music (IAM): Raga identification and Instrument classification. For both these tasks, we integrate multiple public datasets along with additional recordings we acquire from Prasar Bharati[1] Archives to perform LP. Our experiments demonstrate that LP significantly reduces labeling overhead and produces higher-quality annotations compared to conventional baseline methods, including those based on pretrained inductive models. These results highlight the potential of graph-based semi-supervised learning to democratize data annotation and accelerate progress in music information retrieval.

1. INTRODUCTION

Most modern Machine Learning (ML) methods require extensive supervision for robust performance on real-world problems. However, the limited availability of labeled data and the resource-intensive cost of annotating instances pose significant bottlenecks to the scalability and deployment of ML solutions. This underscores the need for robust systems capable of assigning high-quality labels to raw examples.

This challenge is especially acute in audio or music domains, where labeling those recordings requires listening for long durations and maintaining precision for frame-level annotations. Although a wealth of audio/music datasets exist online[1-13], most suffer from scaling issues due to associated annotation and collection costs, which may lead to mislabeling[14]. This limitation corresponds to the limited size of such datasets. For instance,[11] released the TablaSolo dataset of only 38 solo tabla (Indian percussion instrument)

[1] *Prasar Bharati is India's public broadcasting agency, comprising Doordarshan Television Network and All India Radio. It maintains an extensive archive of Indian classical music recordings.*

compositions. Similarly,[3,5] are datasets for sound event classification with only around 3 hours and 9 hours of recordings, respectively. There is a corpus of music databases[6], containing most datasets suffering from similar limitations.

It is therefore important to use systems capable of automatic annotation of audio or music recordings, while maintaining high label quality. Traditional ML systems that train on a labeled set and infer labels on the unlabeled set are ineffective solutions in such scenarios, as they require rich supervision[10, 26, 27]. One such solution is Label Propagation (LP), a semi-supervised technique that aims to learn from a sparsely labeled set and propagate labels onto a large unlabeled set using transductive learning. This approach has immense potential and has recently attracted substantial research interest[15-17]. It exploits two assumptions: (1) data points closer to each other are likely to have the same label; and (2) most data points on the same manifold should have the same label[20]. Unlabeled data points are labeled based on the similarity between their features and those of the labeled data points. Most LP algorithms are graph-based, where the edges between two nodes (instances) encode the affinity between them. This allows for effectively exploiting the underlying structure of the data to propagate labels through the graph, even with limited data. There have been notable works on LP for images[16, 18-21], but this has rarely been explored for audio/music datasets.

LP offers an effective solution for large-scale metadata expansion, particularly in domains where labeled data is scarce but unlabeled data is abundant. Its task-agnostic nature allows it to be applied across a wide range of applications, making it ideal for annotating massive audio and music corpora sourced from platforms like YouTube, Spotify, or public archives. By leveraging the inherent structure in the data, LP provides a scalable and efficient alternative to manual annotation.

In this study, we employ a LP framework for two key tasks in Music Information Retrieval: Raga Identification[10, 26, 27] and Instrument Recognition. We utilize multiple publicly available datasets in combination with a large corpus of unlabeled audio recordings from the Prasar Bharati Archives for carrying out LP across diverse musical content. Our results demonstrate that the proposed approach yields high-quality annotations, often outperforming several baselines, including fully supervised inductive learning approaches.

2. RELATED WORKS

The scarcity of labeled audio data has been a significant bottleneck in advancing machine learning applications in audio and music processing. Despite the growing interest in machine learning for audio and music processing, progress is often hindered by the limited size and scope of available labeled datasets. Many widely used resources, such as TinySOL[9], TablaSolo[11], and IRMAS[22], are restricted by their size or the number of samples. Larger datasets like AudioSet[1] and FSD50K[2] offer broader coverage but often suffer from weak labeling and insufficient annotation detail for specialized music information retrieval tasks. For Indian classical music, datasets like the IAM Raga Recognition Dataset[7], Saraga[8] and PIM[10] provide valuable resources but are limited in size, require lots of manual labeling, and often focus on specific aspects like raga recognition without broader applicability.

These limitations in dataset size, diversity, and annotation quality highlight the need for approaches that can maximize the utility of limited labeled data. In this context, LP offers a promising solution by enabling the automatic extension of labels from a small annotated subset to much larger unlabeled collections, thus addressing a key bottleneck in music and audio machine learning research.

Label Propagation is a semi-supervised learning method that leverages the structure of the data manifold to propagate labels from a small set of labeled examples to a larger unlabeled set. Early work by Zhu and Ghahramani[19] introduced a LP algorithm using a fully connected graph where edge weights are determined by the Gaussian kernel of the Euclidean distance between data points. This method iteratively propagates labels while keeping the labeled data fixed. Zhou *et al.*[18] extended this idea by incorporating both local and global consistency in the graph-based framework. They introduced a normalized graph Laplacian and formulated the LP as a closed-form solution, leading to efficient computation. In recent years, Iscen *et al.*[20] proposed a method that combines deep learning with LP. They use embeddings from a neural network to construct a sparse affinity matrix, which is then used in a diffusion process to propagate labels. This approach benefits from the representation power of deep networks and the structural information captured by the graph.

Our work leverages these advancements in label propagation and applies them to the music domain, specifically addressing the challenges of large-scale unlabeled datasets.

3. DATASETS

For both our tasks of Raga classification and Instrument classification, we leverage a combination of curated and publicly available datasets. Below, we describe the data sources and composition relevant to each task.

3.1 Raga Classification

For the Raga classification task, we use the PIM dataset introduced in[10], which consists of annotated audio recordings from Hindustani classical music performances. The dataset contains over 501 manually labeled audio files, totalling 23,365 audio chunks of 30 seconds, corresponding to 141 unique Ragas. It also includes additional metadata, including Raga, Tonic, Tala, Gharana, and performer annotations. Raga and Tonic labels have been manually annotated and verified for the dataset. It serves as the primary labeled source for our experiments. We apply LP to extend these annotations across a larger unlabeled corpus of audio recordings from Prasar Bharati archives.

Table 1. Sampling from various public datasets for Instrument Recognition. The numbers represent the number of audio samples for each instrument. The instruments are: Accordian (Acc.), Cymbals (Cym), D.K. (Drum Kit), Guitar (Guit.), Organ, Piano, Tabla (Tab.), Trumpet (Trum.), Sitar (Sit.), Flute (Flu.), Violin (Vio.)

Dataset	Acc.	Cym.	D.K.	Guit.	Organ	Piano	Tab.	Trum.	Sit.	Flu.	Vio.
FSD50K	99	835	351	2185	339	844	96	632	-	-	-
AudioSet	-	-	-	-	-	-	949	-	851	2697	-
IRMAS	-	-	-	-	-	721	-	577	-	451	580
TinySOL	689	-	-	-	-	-	-	96	-	118	284
Tabla Solo	-	-	-	-	-	-	38	-	-	-	-

3.2 Instrument Recognition

For the Instrument Recognition task, we curate a dataset comprising recordings primarily sourced from the Prasar Bharati Archives, supplemented with samples from several publicly available datasets. The Prasar Bharati audios predominantly feature instruments which are commonly found in real-life Indian Classical Music performances, such as Sitar, Tabla, Veena, Pakhawaj, and Flute. A notable challenge in these audios is the imbalance across instrument classes-with approximately 65% of the total duration concentrated in the top five most frequent classes. This imbalance can hinder the performance of machine learning models, especially during label propagation. To address this, we augment the Prasar Bharati audios by including class-specific samples from various open-source datasets, thereby improving class. Specifically, we sample: TablaSolo[11] for Tabla recordings, FSD50K[2] for Guitar, Drum-kit, and Tabla, TinySOL[9] and IRMAS[22] for Flute and Violin, and AudioSet[1] for Flute and Drum-kit samples. The complete source-wise distribution of instrument durations, including contributions from Prasar Bharati and external datasets, is shown in Figure 1, while the number of samples acquired from other datasets is provided in Table 1. The combined dataset includes a total of 20 instrument classes as shown in the legends in Figure 1.

We divide the whole dataset into labeled and unlabeled sets and curate a gold set of 200 manually labeled and verified audio recordings out of the unlabeled set for evaluation purposes.

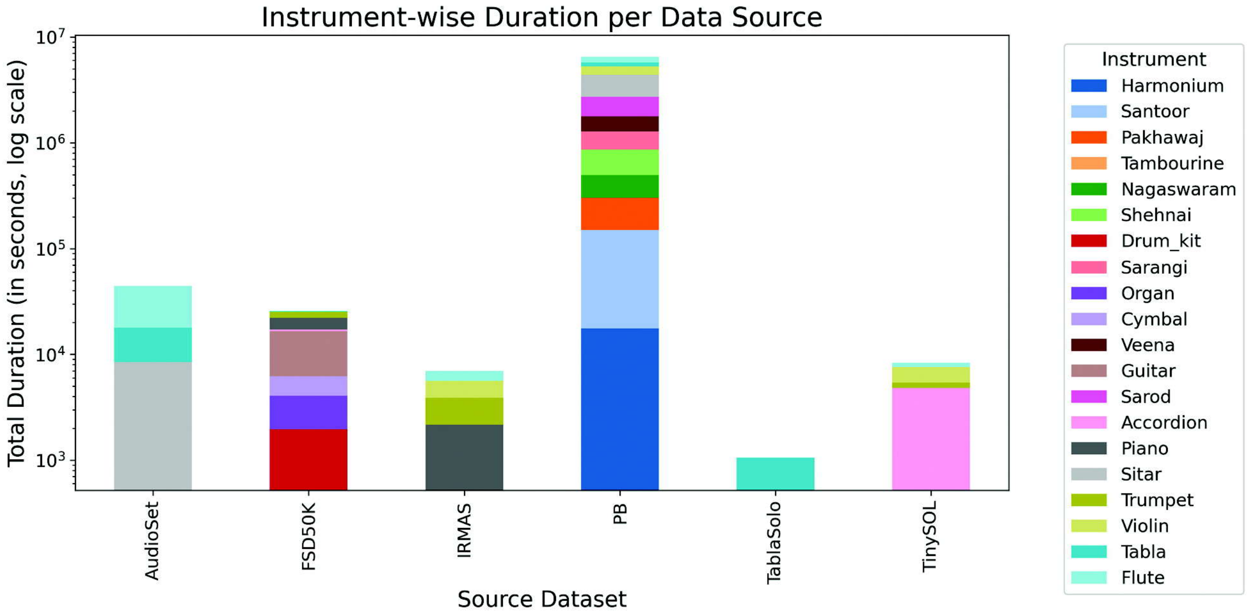


Fig. 1. Source-wise duration of all instrument samples used from various datasets (in seconds). PB represents Prasar Bharati audios.

4. LABEL PROPAGATION

The scarcity of reliable labeled data necessitates the use of efficient transductive LP methods. We utilize pseudo-labels for unlabeled data to train a classifier and construct a graph by exploiting the embeddings obtained from the network[20]. This is a two-fold method: (1) Train the network using the entire dataset (with pseudo labels for unlabeled data points), and (2) Construct a nearest-neighbor graph using the embeddings from the network. This flowchart in Figure 2 illustrates the process of LP. Initially, a small

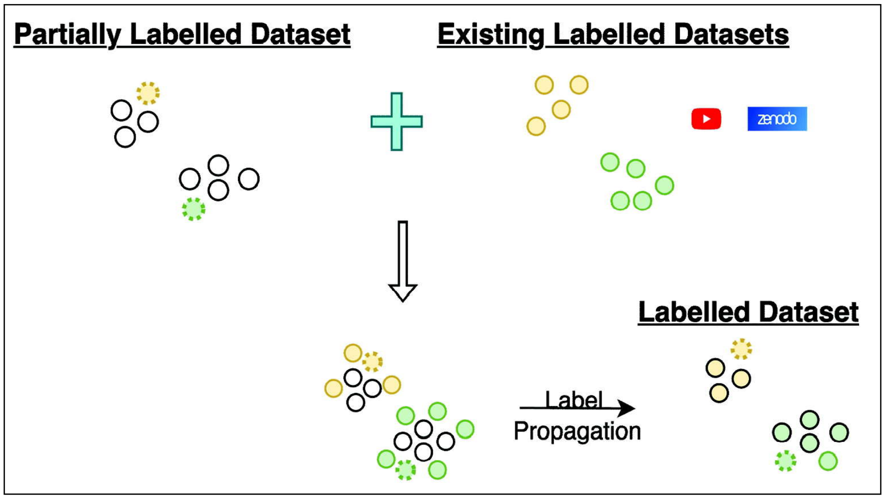


Fig. 2. Flowchart illustrating the Label Propagation process for automatic annotation of unlabeled audio samples. A small labeled dataset is combined with a larger set of partially labeled or unlabeled samples. Label Propagation is then applied using a similarity graph, enabling the transfer of label information from the labeled subset to the unlabeled data, resulting in the entire corpus being annotated in a transductive, semi-supervised manner.

set of audio samples with sparse labels (left) is combined with external labeled datasets sourced from platforms like YouTube and Zenodo (right). These datasets are merged to form a unified collection containing both labeled and unlabeled instances. Through the application of LP, label information from the annotated examples is extended to the unlabeled samples, resulting in a fully labeled dataset (bottom). This approach significantly reduces manual annotation effort while maximizing the utility of available data for downstream machine learning tasks.

4.1 *Classifier Training*

Let $X = \{x_1, x_2, ..., x_l, x_{l+1}, ..., x_\mathrm{n}\}$ with $x_i \in \chi$ be a collection of n data points, where the first l are labeled with class labels $y_i \in C$, and the remaining $u = n - 1$ points are unlabeled. The label space is denoted by $C = \{1, 2, ..., c\}$. We employ a deep network consisting of a feature extractor h_{θ_1}: $\chi \rightarrow \mathbb{R}^d$ that maps each input x_i to a d-dimensional embedding $z_i = h_{\theta_1}: (x_i)$, and a classifier g_{θ_2}: $\mathbb{R}^d \rightarrow \mathbb{R}^c$ that outputs class-wise confidence scores. The overall model is represented by $f_\theta(x) = g_{\theta_2}(h_{\theta_1}(x))$, where $\theta = (\theta_1, \theta_2)$.

The training objective involves multiple components:

$$L_s(X_l, Y_l; \theta) = \sum_{i=1}^{l} l_s(f_\theta(x_i), y_i),$$

$$L_s(X_u, \hat{Y}_u; \theta) = \sum_{i=l+1}^{n} l_s(f_\theta(x_i), \hat{y}_i),$$

$$L_u(X; \theta) = \sum_{i=1}^{n} l_u(f_\theta(x_i), f_{\hat{\theta}}(\hat{x}_i)),$$

Here, L_s is the supervised loss (typically cross-entropy) over the labeled data. L_p is a pseudo-labeling loss computed using labels $\hat{y}_i$ predicted by LP. L_u is the unsupervised loss term applied on $X_l \cup X_u$ to make the embeddings consistent across different transformations of an input.

4.2 *Transductive Label Propagation*

To generate labels $\hat{y}_i$ for the unlabeled examples, we use a transductive LP approach inspired by diffusion processes[18]. The core idea is to construct a similarity graph among all data points based on their embeddings and propagate known labels across the graph structure.

Let $W \in \mathbb{R}^{n \times n}$ be a symmetric adjacency matrix with zero diagonals, where each entry w_{ij} captures the similarity between embeddings z_i and z_j. This matrix is constructed using a symmetric k-nearest neighbor (k-NN) graph. We define $W = M + M^T$, where:

$$m_{ij} = \begin{cases} [(z_i^T zj)^y]_+, & if\ i \neq j \text{ and } z_j \in NN_k(z_i), \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathrm{NN_k}(z_i)$ denotes the k nearest neighbors of z_i in embedding space, γ is a sharpness hyperparameter, and $[\cdot]_+$ denotes just the positive part (ReLU function).

Next, we compute the symmetric normalized affinity matrix S using:

$$S = D^{-1/2} W D^{-1/2,}$$

where $D = \mathrm{diag}(W\mathbf{1}_n)$ is the degree matrix and $\mathbf{1}_n$ is an all-ones vector of length n.

We now construct a label matrix Y of shape $n \times c$ such that $Y_{ij} = 1$ if x_i is labeled and its class is j, and $Y_{ij} = 0$ otherwise. Thus, labeled examples are one-hot encoded, and unlabeled rows are all zeros. LP computes soft labels over all nodes via the closed-form solution:

$$P = (I - \alpha S)^{-1} Y,$$

where $P \in \mathbb{R}^{n \times c}$ contains the propagated label distributions, and $\alpha \in [0, 1]$ controls the strength of label diffusion. The rows of P can be interpreted as class probability scores. Finally, we assign pseudo-labels using:

$$\hat{y}_i = \underset{j}{\mathrm{argmax}}\ P_{ij,}$$

This assigns to each example x_i the class with the highest propagated score. Here, it is noteworthy that computing the matrix inverse $(I - \alpha S)^{-1}$ is computationally infeasible for large n because it is not sparse. Instead, following[22], we solve the linear system:

$$(I - \alpha S)Z = Y$$

using the conjugate gradient method, which yields $Z \approx P$. We iteratively propagate these pseudo-labels while jointly optimizing the loss functions described earlier, ensuring that the overall loss continues to decrease across iterations. Once convergence is achieved, the final pseudo-labels obtained from the diffusion process are used as predictions for evaluation.

Table 2. Performance of the PANNs[26] model on the labeled portion of the PIM dataset for the Speech vs. Music classification task. The model is evaluated on 501 Hindustani classical music recordings annotated with speech and music segments. Metrics are computed at the 30-second chunk level after excluding ambiguous segments containing both speech and music.

	Precision	**Recall**	**F1**
Speech	1.0	0.93	0.97
Music	1.0	0.98	0.99

5. EXPERIMENTS

We utilized the LP scheme discussed in Section 4 to expand the metadata for Music Instrument Recognition and Raga Identification tasks. First, we train a Deep Neural Network in a fully supervised setup on 80% of the labeled set X_l. We then infer labels from this trained network for the remaining 20% labeled recordings and all unlabeled recordings X_u. To assess the performance on the unlabeled set, we manually annotate and verify a subset of recordings from X_u. Finally, we report the accuracy obtained on the 20% held-out set (labeled) and the manually annotated recordings from the unlabeled set. We now explain experimental details for both tasks in detail.

5.1 *Music Instrument Recognition*

For the instrument detection task, we pre-process the audio recordings by first discarding all files shorter than 1 second. The remaining files are segmented into 5-second chunks, with each chunk inheriting the instrument label of its source recording. Mel-spectrograms are extracted from each chunk using a window size of 1024, hop length of 512, and 64 mel bins. These serve as input features to the network.

As a baseline, we use a modified ResNet-18[24] trained in a fully supervised fashion. For our proposed approach, we apply LP as described in Section 4, leveraging a small labeled subset and a larger pool of unlabeled examples. Both models are trained for a maximum of 50 epochs using the Adam optimizer with Stochastic Gradient Descent. For evaluation, we reserve a manually annotated gold test set of 200 test audio recordings, and accuracy is computed on the held-out test set to assess performance.

5.2 *Raga Identification*

For the Raga Identification task, we work with a total of 61,705 audio chunks, 30 seconds each, out of which 13,075 are labeled and taken from PIM[10] dataset and span 41 known Raga classes, along with a 42nd *Others* class. The remaining 48630 unlabeled recordings are sourced from the Prasar Bharati Archives. To construct the evaluation set, for each of the 42 Raga classes, we select at least one full audio file out of the labeled set based on their representation. These selected recordings are then split into 30-second chunks, and all resulting chunks are included in the evaluation set, resulting in a total of 3,210 evaluation chunks. During LP training, these chunks are merged with the unlabeled set by discarding their true labels, creating a transductive learning setting.

To filter out non-musical (speech) segments, we first annotate the 501 recordings present in the PIM [10] dataset with speech and music timings and evaluate the performance of the PANNs model [23] for automatic segmentation. The PANNs model is tested on 501 Hindustani classical music files, comprising 23,224 audio chunks of 30 seconds each. Chunks containing overlapping speech and music are excluded due to labeling

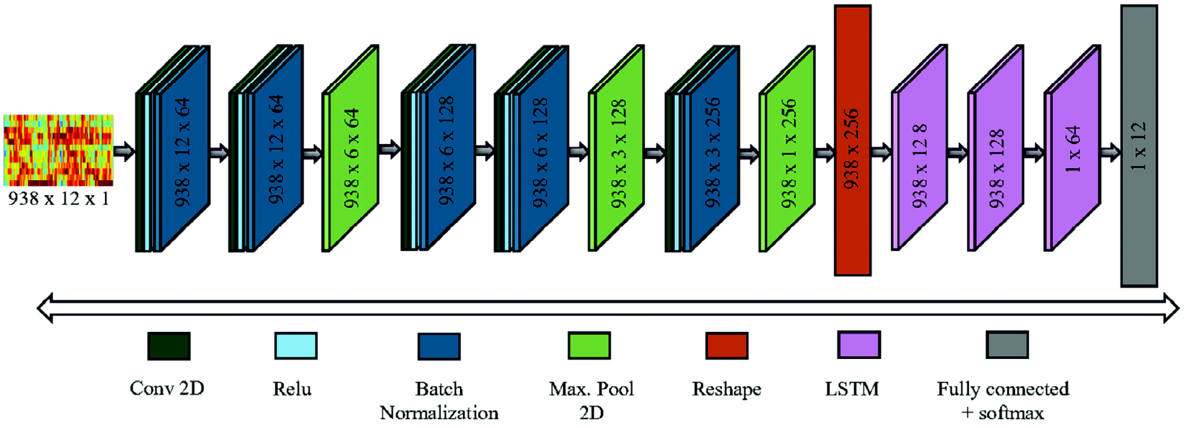


Fig. 3. Model architecture used for the Raga Identification task, comprising convolutional layers followed by LSTM layers to capture temporal dependencies in the audio data.

ambiguity. Predictions of PANNs are evaluated at the chunk level. For cases where the model's top prediction is not Music, we consider it music if one of the top predicted classes includes a relevant musical tag (*e.g.,* tabla, sitar) and Music is the second-highest class. Out of 23,224 chunks, 21,691 are classified as music, 1,453 as speech, and 80 as ambiguous. As shown in Table 2, the PANNs model shows robust and reliable performance for this task, and can be used for labeling all other audio files from PB recordings.

We train our model for a 42-class classification task with 41 known Ragas (Having the most number of audio files in the labeled dataset) and an additional Others class, which includes all other remaining Ragas and speech segments extracted from the audio files themselves using the PANNs model. For tonic normalisation, we use the tonic values provided in[10], and for the remaining audios, we use the *CompIAM* package[25] for computing tonic values, which is known to be very useful for the task, as explained in[10].

We employ a CNN-LSTM model architecture for our classification task, as shown in Fig. 3. Initially, we train it in a fully supervised manner using the given labels and evaluate it on the test set extracted from the labeled data. Next, we use the same architecture, pre-train it for just 10 epochs, and then use it for feature extraction and deploy the LP method to train it using both the labeled and unlabeled examples. For evaluation, we assess our models at both the audio chunk level and the audio file level. At the chunk level, we measure the accuracy of classification for each audio chunk, while at the audio file level, we determine the majority vote among all chunks sourced from the same audio file for class prediction.

6. RESULTS AND DISCUSSION

6.1 *Instrument Recognition*

Table 3 shows the performance comparison between the supervised baseline and the LP method. Accuracy is reported on two sets: Acc.1 refers to a held-out portion of the labeled set (X_l), and Acc.2 refers to a manually annotated subset of the originally unlabeled set (X_u). The LP method achieves 84.6% on Acc.1 and 91.7% on Acc.2, compared to 48.3% and 63.2% respectively for the supervised baseline. This demonstrates that the propagated labels are of high quality and the model predictions after LP are quite trustworthy, even on data not seen during training.

Table 3: Performance Analysis of Label Propagation with Baseline on Instrument Recognition. Acc.1 and Acc.2 correspond to the accuracies on subsets of X_l and X_u, respectively.

Method	Acc.1 (%)	Acc.2 (%)
Baseline	48.3	63.2
Label Propagation	84.6	91.7

6.2 *Raga Classification*

Table 4 presents the performance metrics for Raga classification, evaluated at both the chunk level (CL) and file level (FL. The supervised model achieves an F1-score of 0.60 (CL) and 0.77 (FL). With the application of LP, these scores improve to 0.62 (CL) and 0.82 (FL). While chunk-level gains are modest, the improvement at the file level is significant (5% absolute increase in F1-score). This is crucial for our application, which involves labeling full-length music recordings. Although the model is trained on a fixed set of Raga classes, it is designed to handle unseen or ambiguous cases by assigning such instances to a generic *Others* category.

Table 4: Performance Comparison for Raga Classification task. CL: Chunk Level, FL: File Level.

Method	Precision	Recall	F1 Score
Supervised (CL)	0.63	0.57	0.60
Supervised (FL)	0.76	0.80	0.77
Label Propagation (CL)	0.65	0.59	0.62
Label Propagation (FL)	0.83	0.82	0.82

These results reinforce the effectiveness of the LP framework in generating meaningful pseudo-labels for the two IAM tasks. The technique can be easily implemented to any downstream MIR classification tasks. These performance improvements after LP have practical use cases such as automatic cataloging or metadata generation in music archives. Overall, the LP method consistently enhances performance across tasks and demonstrates its potential as a reliable and scalable labeling strategy for large, partially labeled music datasets.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we explore the use of label propagation, a graph-based semi-supervised learning technique, to address the challenge of limited labeled data for Music Information Retrieval tasks. We focus on two key tasks within the domain of Indian Art Music (IAM): Raga identification and Instrument classification. By constructing a similarity graph over audio segments, we exploit the inherent structure in the feature space to propagate labels from a small, labeled subset to a much larger unlabeled corpus in a transductive, semi-supervised manner. Our results demonstrate that label propagation is an effective alternative to traditional supervised approaches, especially in domains like IAM research, where expert annotations are expensive, time-consuming, and require deep domain expertise. The method yields high-quality pseudo-labels and enables scalable annotation, making it well-suited for settings where labeled data is scarce, but unlabeled data is abundant.

This study highlights the broader utility of graph-based semi-supervised learning for developing robust MIR systems with minimal manual labeling. Future directions include exploring adaptive or dynamic graph construction, incorporating temporal and structural musical features into the propagation process, and extending the framework to support open-set recognition, where previously unseen labels are automatically detected and modeled rather than being grouped under a generic "Others" category. Such extensions would make the system more realistic and applicable in real-world scenarios where new categories continuously emerge.

8. ACKNOWLEDGMENTS

We would like to thank Prasar Bharati (PB), India for providing the audio recordings and metadata. We also acknowledge the contributions of Renu Chavan and Kajal Heer for their efforts in data annotation and validation.

REFERENCES

[1] J. F. Gemmeke *et al., 2017.* "Audio set: An ontology and human-labeled dataset for audio events," in 2017 *IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE*, pp. 776-780.

[2] E. Fonseca, X. Favory, J. Pons, F. Font and X. Serra, 2021. "Fsd50k: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **30:** 829-852..

[3] K. J. Piczak, 2015. "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on multimedia,* pp. 1015-1018.

[4] J. J. Bosch, F. Fuhrmann and P. Herrera, 2018. "IRMAS: a dataset for instrument recognition in musical audio signals." *Zenodo,* doi: 10.5281/zenodo.1290750.

[5] J. Salamon, C. Jacoby and J. P. Bello, 2014. "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on multimedia,* pp. 1041-1044.

[6] X. Serra, 2014. "Creating research corpora for the computational study of music: The case of the compmusic project," in *AES 53rd international conference: Semantic audio;* pp. 27-29; london, UK. New york: Audio engineering society. Article number **1-1** [9 p.]., Audio Engineering Society, 2014.

[7] S. Gulati, J. Serrà, K. K. Ganguli, S. Sentürk and X. Serra, 2016. "Indian art music raga recognition dataset (audio)." *Zenodo*. doi: 10.5281/zenodo.7278511.

[8] B. Bozkurt, A. Srinivasamurthy, S. Gulati and X. Serra, 2018. "Saraga: Research datasets of indian art music." *Zenodo.* doi: 10.5281/zenodo.4301737.

[9] C. Emanuele, D. Ghisi, V. Lostanlen, F. Lévy, J. Fineberg and Y. Maresz, 2020. "TinySOL: an audio dataset of isolated musical notes (5.0)." *Zenodo*. doi: 10.5281/zenodo.3685331.

[10] P. Singh and V. Arora, 2024. "Explainable deep learning analysis for raga identification in indian art music," arXiv preprint arXiv:2406.02443, 2024.

[11] S. Gupta, A. Srinivasamurthy, M. Kumar, H.A. Murthy and X. Serra, 2015. "Discovery of syllabic percussion patterns in tabla solo recordings." in 16th *international society for music information retrieval conference (ISMIR),* pp. 385-391.

[12] A. Shankar, G. Plaja-Roglans, T. Nuttall, M. Rocamora and X. Serra, 2024. "Saraga audiovisual: A large multimodal open data collection for the analysis of carnatic music," in *Proceedings of the 25th international society for music information retrieval conference, ISMIR,* pp. 61-69. doi: 10.5281/ zenodo.14877279.

[13] S. Kumar, P. Singh and V. Arora, 2025. "Recognizing ornaments in vocal indian art music with active annotation," arXiv preprint arXiv:2505.04419, 2025.

[14] L. Schmarje *et al.,* 2022. "Is one annotation enough?-a data-centric image classification benchmark for noisy and ambiguous label estimation," *Advances in Neural Information Processing Systems,* **35:** 33215-33232.

[15] T. Xie, B. Wang and C.-C. J. Kuo, 2023. "GraphHop: An enhanced label propagation method for node classification," *IEEE Transactions on Neural Networks and Learning Systems,* **34**(11): 9287-9301.

[16] T. Cai, R. Gao, J. Lee and Q. Lei, 2021. "A theory of label propagation for subpopulation shift," in Proceedings of the 38th *international conference on machine learning,* M. Meila and T. Zhang, Eds., in *Proceedings of machine learning research, PMLR,* **139:** 1170-1182.

[17] S. Tankasala *et al.,* 2023. "Cross-utterance ASR rescoring with graph-based label propagation," in *ICASSP* 2023.

[18] D. Zhou, O. Bousquet, T. Lal, J. Weston and B. Schölkopf, 2003. "Learning with local and global consistency," *Advances in neural information processing systems,* **16**.

[19] X. Zhu and Z. Ghahramani, 2002. "Learning from labeled and unlabeled data with label propagation," *ProQuest number: information to all users.*

[20] A. Iscen, G. Tolias, Y. Avrithis and O. Chum, 2019. "Label propagation for deep semi-supervised learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5070-5079.

[21] H. Zhu and P. Koniusz, 2023. "Transductive few-shot learning with prototype-based label propagation by iterative graph refinement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23996-24006.

[22] J. J. Bosch, J. Janer, F. Fuhrmann and P. Herrera, 2012. "A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals." in *ISMIR*, pp. 559-564.

[23] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang and M. D. Plumbley, 2020. "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **28:** 2880-2894.

[24] K. He, X. Zhang, S. Ren and J. Sun, 2016. "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 770-778.

[25] Genís Plaja-Roglans and Thomas Nuttall and Xavier Serra, compIAM. 2023. Available: https:// mtg.github.io/compIAM/

[26] P. Singh, A. Mishra, A. Raina and V. Arora, 2025. "Ontology-driven hierarchical learning for raga identification," in *2025 national conference on communications (NCC),* pp. 1-6.

[27] P. Singh, A. Gupta, A. Mishra and V. Arora, 2025. "Identification and clustering of unseen ragas in indian art music," in *Proceedings of the 26th international society for music information retrieval conference,* pp. 811-818.

The Journal of Acoustical Society of India : Vol. 52, No. 3, 2025 (pp. 155-164)

Towards evaluating multimodal large language models on Hindustani Classical Music dataset

Shankha Sanyal[1*], Enakshmi Ghosh Kundu[1], Medha Basu[1, 2], Sayan Nag[1,3] and Dipak Ghosh[1]

[1]*Sir C.V. Raman Centre for Physics and Music, Jadavpur University, Kolkata, India*
[2]*Department of Physics, Jadavpur University, Kolkata, India*
[3]*Department of Medical Biophysics, University of Toronto, Toronto, Canada*
e-mail: ssanyal.ling@jadavpuruniversity.in

[Received: 11-05-2025; Accepted: 12-07-2025]

ABSTRACT

The case of classifying music on the basis of genre, emotions, and acoustic features has taken centre stage of Music Information Retrieval (MIR) with the advent of different machine learning techniques. Such studies in the genre of Hindustani music is scarce owing primarily to the limited available database of annotated Hindustani music clips and also to the mixed emotions generated from Hindustani music. The present pilot study takes the help of a publicly available database of 2400 music clips - ***JUMusEmoDB*** which presently has 2400 audio clips (approximately 30 s each) where 400 clips each correspond to happy, sad, calm and anxiety emotional scales. The clips have been taken from different conventional 'raga' renditions played in three Indian instruments - *sitar, sarod* and *flute* by eminent maestros of ICM and digitized in 44.1 kHz frequency. From the available database, we look to compare the robustness of three open source and two closed source AI models in accurately predicting the various features of classical music. The three open source AI models used here are Audio-LLM,Qwen 2.5 Omni and Gamma, while the two close source models used are Gemini-1.5-Pro and GPT-4o. Using a total of 60 clips from the dataset, all the chosen models were asked four types of questions - a) identifying the raga, b) identifying the instrument, c) identifying the emotions portrayed by the particular clip and d) country of origin of the music clip. An accuracy index was computed for each of the questions to identify which model worked best for the case of Hindustani music. The results provide new and interesting data about the application of AI models in classifying Hindustani music clips, which would lead the way for more such studies in the future.

1. INTRODUCTION

The advent of Artificial Intelligence (AI) coupled with Machine Learning (ML) architectures into the musical world has completely changed the ways we used to perceive and enjoy music. From the era of physically owned long playing vinyl records, we have moved to an age where music has largely become a service accessed through platforms like Spotify, Apple Music, and YouTube Music, rather than a product that is physically owned. This on-demand access to vast libraries of music has created an entirely new dimension of music consumption which is in stark contrast to previous generations, where access was more constrained by physical formats and broadcast limitations. AI algorithms present in various music streaming platforms have become the most important tools for music listening for the present generation. These algorithms analyze huge amounts of user data and personal preferences to curate personalized playlists that introduce us to a diverse spectrum of musical genres and artists. This phenomenon has completely changed how this generation perceives and forms connections through music[1-3]. Various popular music streaming platforms take advantage of AI models to scrutinize user behavior, listening patterns, and even physiological responses, continuously refining their recommendations to align with individual tastes. Personalized playlists, which form the basic foundation of AI-driven music curation, play a critical role in shaping, utilize machine learning models to analyze extensive user data, thus striking a balance between familiar tracks and novel suggestions. The recent introduction and expansion of Spotify's "AI Playlist" further underscores this trend, enabling users to generate playlists based on creative prompts, effectively putting AI at the forefront of personalized music curation. In this scenario comes the effectiveness and utility of promoting Hindustani Classical Music in the forefront of AI generated music playlists and algorithms[4-7, 28].

Hindustani Classical Music is a rich and complex art form built upon the concepts of Raga (melodic framework) and Tala (rhythmic cycle). The automatic classification of HCM elements, particularly Ragas and genres, using AI and Machine Learning (ML) techniques is a significant area within Music Information Retrieval (MIR). This field aims to develop computational tools for organizing, analyzing, retrieving, and interacting with musical data. Unlike Western classical music, HCM relies heavily on improvisation, intricate ornamentation (gamakas, meends), microtones (shrutis), and a relative pitch system, presenting unique challenges and opportunities for AI applications. The basic classification task in AI for HCM is raga identification, which has been studied mostly on the basis of different features like Mel Frequency Cepstral Coefficients (MFCCs), Pitch class profiles, chroma features, spectral and temporal features[8-11]. Deep learning models such as RNN, CNN, CNN coupled with LSTMs have been successfully applied in different studies for the purpose of raga identification[12,13]. Studies report significant success, with deep learning models often achieving state-of-the-art results. For instance, accuracies above 88% on broader datasets and up to 97% on smaller subsets (e.g., 10 Ragas) have been reported using LSTM-based approaches on datasets like CompMusic (though primarily Carnatic, methods are transferable)[14]. Research using the PIM-v1 dataset achieved an F1-score of 0.89 for a 12-Raga subset using CNN-LSTM models[15]. Besides Raga, AI models are used to classify different Hindustani genres or styles (e.g., Dhrupad, Khayal, Thumri, Tappa)[16,17]. But these studies suffer from the availability of publicly available large datasets of HCM[18]. Also the emotional aspect of HCM is not studied conventionally, due to the complex nature of emotional arousal associated with each raga of HCM. Another important hurdle is isolating the main melody (vocal or instrumental) from accompanying instruments (like Tabla, Tanpura, Harmonium). In this scenario, it becomes important to study the ability of different existing models to classify different aspects of Hindustani classical music such as emotion, instrument, genre and raga.

The present study makes the use of two closed source and three open source AI models to study the validity of an open source publicly available annotated dataset of HCM to classify ragas, emotion, instruments and origin. The main aim of this study is to measure the readiness of these five multimodal foundation models, Qwen 2.5 Omni, GAMA, Audio LLM, GPT-4o, and Gemini 1.5 Pro in classifying tasks related to Hindustani Classical Music. For this, each model was fed with emotionally loaded instrumental clips of HCM from previously annotated dataset ***JUMusEmoDB***[19,20] and asked open-ended questions related to identification of emotion, raga and instrument. It was found that most of the models are not ready for classification of Hindustani music, specially in case of ragas, while Gemini 1.5 Pro reported adequate accuracy while identifying instruments of Indian origin. The study posits the need to reassess the training and application of machine learning models in classifying parameters of Hindustani classical music.

2. METHODOLOGY

2.1 Dataset Curation and Task Formulation

JUMusEmoDB[20] consists of 2400 music clips from the genre of Indian Classical Music, out of which 600 clips have been categorized to each of the four classes of emotions, namely, anxiety, calm, happy, and sad. Each clip is of 30 s length which is long enough for introducing an emotional imposition[21]. Out of the 2400 clips, each 800 clips are parts of different 'raga' renditions improvised in sitar, sarod and flute and by three eminent maestros of ICM. Each raga in ICM evokes not only a particular emotion (rasa), but a superposition of different emotional states such as joy, sadness, anger, disgust, fear and so on. In order to decipher the predominant emotions which were conveyed in the chosen ragas, emotion annotations were performed by 100 participants based on a 5-point Likert scale. From the annotated clips For the present study, a total of 12 clips were randomly selected belonging to 3 different instruments and 4 different emotions. So, the Ground Truth Table for the clips chosen corresponding to different ragas and emotions looks like the following Table 1.

Table 1. Ground Truth Table (Instrument,Emotion, Raga).

GT Instrument	GT Emotion	GT Raga
Flute	Anxiety	Dhun
Sarod	Anxiety	Ramkali
Sitar	Anxiety	Todi
Flute	Calm	Bilaskhani Todi
Sarod	Calm	Jog
Sitar	Calm	Bhatiyaar
Flute	Happy	Pahadi
Sarod	Happy	Bilaval Dhun
Sitar	Happy	Desh
Flute	Sad	Bhimpalasi
Sarod	Sad	Bageshree
Sitar	Sad	Malkauns

For the 5 ML models put to test, the open ended questions or tasks put forward have been shown in Fig. 1.

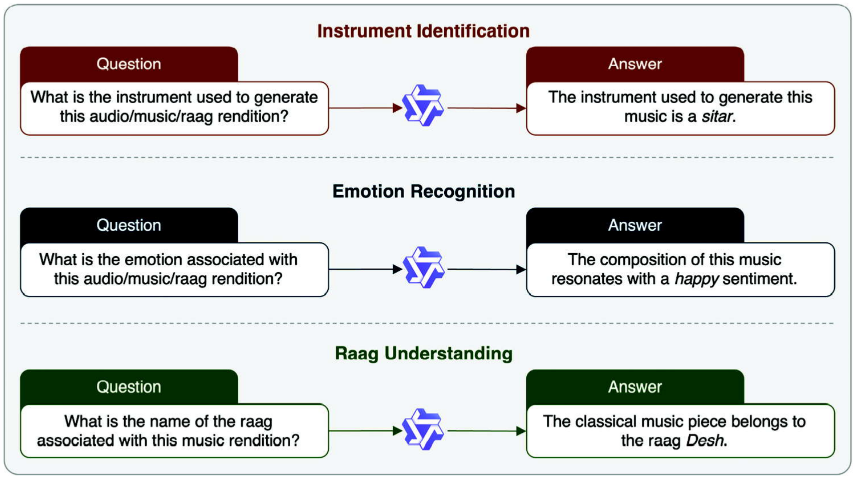

Instrument Identification
Question
What is the instrument used to generate this audio/music/raag rendition?
Answer
The instrument used to generate this music is a *sitar*.
Emotion Recognition
Question
What is the emotion associated with this audio/music/raag rendition?
Answer
The composition of this music resonates with a *happy* sentiment.
Raag Understanding
Question
What is the name of the raag associated with this music rendition?
Answer
The classical music piece belongs to the raag *Desh*.

Fig. 1. Overview of the proposed task formulation (Instrument Identification, Emotion Recognition, and Raga Understanding). We evaluate 2 closed-source and 3 open-source MLLMs which can handle audio. As an example, we show Qwen2.5-omni here.

The responses given by each of the models have been analyzed and corresponding accuracy values have been computed.

2.2 Models

We investigated the Hindustani classical music-based question answering skills of four prominent multimodal foundation models known for their diverse input processing: Qwen 2.5 Omni, GAMA, GPT-4o, and Gemini 1.5 Pro. The subsequent descriptions outline each model and its potential for this task, emphasizing their audio understanding capabilities.

Qwen-2.5 omni: Alibaba's Qwen 2.5 Omni is a unified, large AI model capable of reasoning across text, images, audio, and video simultaneously. It uses a shared embedding space and cross-modal alignment to effectively understand different types of data. Notably, it has shown strong results in audio-text tasks, particularly in understanding emotions and subtle auditory details[22].

GAMA: GAMA (Generalist Audio-Multimodal Assistant) is a recently developed open-source framework specifically for tasks where instructions are given through audio first[23]. It combines an Audio Spectrogram Transformer (AST) along with an Audio Q-Former and Multi-LayerAggregator for processing audio inputs which is then passed to an LLM and is trained on various multimodal datasets (audio-language) to ensure it works well across different audio-language tasks. GAMA is good at Audio Question Answering, Dense Captioning, and audio comprehension tasks involving complex reasoning.

GPT-4o: GPT-4o, where "o" stands for "omni," is OpenAI's latest multimodal model that natively handles text, vision, and audio[24]. Unlike earlier methods that used separate processing for different inputs, GPT-4o directly processes audio within its core architecture. This leads to faster, more accurate, and more consistent responses. It can understand the nuances of speech like tone and rhythm, as well as the meaning, making it very effective for real-time question answering based on audio.

Gemini-1.5-Pro: Google DeepMind's Gemini 1.5 Pro is a large multimodal AI model that can process long sequences of text, images, audio, and video[25]. Its design allows it to maintain information and connect facts across different types of data. In audio-related tasks, Gemini excels at understanding how sounds change over time and their meaning in extended audio, enabling it to follow instructions and generate responses with high accuracy.

Audio LLM: AudioLM, is another framework for high-quality audio generation with long-term consistency[26]. AudioLM maps the input audio to a sequence of discrete tokens and casts audio generation as a language modeling task in this representation space.

2.3 Evaluation

Since our task formulation follows an open-ended question answering setup, we adopt the LLM-as-a-Judge evaluation paradigm[27], where a powerful language model (GPT) serves as an automated evaluator. This approach enables a more nuanced assessment of the generated responses, particularly in tasks that lack definitive ground truth answers or where multiple valid outputs are possible. The GPT-based evaluator is prompted with both the task-specific question and the candidate answer, along with the Ground Truth reference, and is asked to determine the correctness, completeness, and relevance of the response in a structured and consistent manner.

We employ this evaluation method across all our proposed tasks to ensure a uniform and semantically aligned metric of quality. The judgments returned by the LLM are then aggregated to compute accuracy scores, indicating the proportion of model responses deemed correct by the evaluator. This framework allows us to compare the performance of different models on complex, open-ended tasks with greater fidelity than traditional exact match or BLEU-style metrics.

We report the resulting accuracies for all the models across all proposed tasks, offering a comprehensive view of performance under this LLM-guided evaluation scheme.

3. RESULTS AND DISCUSSION

The experiment has been conducted on 5 Predictive AI platforms namely - *Qwen-2.5 omni,* Audio LLM, *GAMA,* GPT-4o and *Gemini-1.5-Pro*. The features selected for comparison were emotion, raga and instrument used as already discussed in the Methodology section. Table 1 shows a snapshot of actual emotions of the ragas played and instruments on which the selected audio clips are being played. Tables 2 to 5 shows the generated output captured by the Individual AI platforms when the clip samples were uploaded for prediction.The input parameters were the sample clips of ragas played on different instruments.The output captured from generative AI platforms were Name of Instruments, Raaga played and Emotions as understood by AI. The term GT or GroundTruth refers to actual or the base fact. Table 1 being the pivot or base table for comparative study using AI models.

Table 2. Output generated by ***Qwen-2.5 AI Platform*** for the above selected audio clips

Instrument	Emotion	Raga
Flute (Bansuri)	Joyful, uplifting	Bhimpalasi
Veena	Melancholic/reflective	Kedar
Flute (Bansuri)	Melancholic/Sad	Bhimpalasi
Veena	Melancholic/Sad	Kalyan
Veena	Melancholic/Sad	Kedar
Flute (Bansuri)	Melancholic/Sad	Abhogi
Veena	Melancholic/Sad	Kedar
Veena	Melancholic/Sad	Kedar
Flute (Bansuri)	Melancholic/Sad	Abhogi
Sitar	Joyful/Lively	Yaman Kalyan
Harmonium	Melancholic/Sad	Bhimpalasi

Table 3. Output generated by *GAMA* for the same selected audio clips

Instrument	Emotion	Raga
Wind instrument	Calm and soothing	Raga
sitar	melancholic	Raga Jogi
Musical instrumen	Emotionally charged.	Raga - not specified
Wind instrument and woodwind instrument	A sense of calmness and tranquilit	Not provided
Musical instrument	Emotionally charged.	Raga - not specified
sitar	melancholic	Raga Jhajhar
sitar	melancholic	Raga Jogi
sitar	passion	Bageshree
Musical instrument	melancholic	Raga
sitar	calm and soothing	Raga Jog
Quota Exhausted	Quota Exhausted	Quota Exhausted
Quota Exhausted	Quota Exhausted	Quota Exhausted

From the comparison of performances of two open source AI platforms, it can be observed that Qwen 2.5 showed higher efficiency for raga and emotion prediction. Also, the instrument identification was nearly accurate, with few anomalies such as mixing sitar with harmonium. But the interesting observation being all the mix ups was with instruments of Indian origin only. The restriction on user specific quota could be a hindrance to bulk data gathering.

Table 4. Output generated by GPT-4o for the same selected audio clips

Instrument	Emotion	Raga
String instrument (likely a sitar or veena)	Calm and meditative emotion	Raga Yaman or a similar evening raga
Percussion instrument (likely a tabla or mridangam)	Energetic and rhythmic emotion	a tala-based composition rather than a specific raga
String instrument (likely a sitar or sarod)	Reflective and serene emotion	Raga Kafi or a similar late evening raga
String instrument (likely a cello or double bass)	Mournful and deep emotion	Not identifiable with certainty
String instrument (likely a cello or violin)	Calm and introspective	Not identifiable with certainty
String instrument (possibly sitar or veena)	Serene and meditative	Yaman or Bhairav
String instrument (possibly sitar or veena)	Tranquil and contemplative	Yaman or Bhairav
Stringed instrument (possibly sitar or sarod).	Melancholic and introspective emotion	An Indian classical raga
Stringed instrument (possibly sitar or veena)	Serene and meditative emotion	An Indian classical raga
String or wind instrument (e.g., sitar, flute)	Calm and meditative	Unknown
String or wind instrument (e.g., sitar, flute)	Reflective or romantic	Unknown
String or wind instrument (e.g., sitar, flute)	Reflective or romantic	Unknown

The widely used platform of GPT 4.0 gave better response for Emotion identification but could not show much accuracy for Raag and Instrument. The data upload quota was also restricted for a particular user. The quota renewal time was 12 hrs for a single user.

Gemini showed good accuracy for instrument, emotion and origin prediction. But for Raga prediction it failed to some extent. But Gemini did not have much restriction for user on the basis of data bandwidth.

Audio LLM had no restriction on data capture but showed very little accuracy in the Instrument or Raaga Prediction. Though it showed some accuracy while predicting the emotional appraisal of the clips and country of Origin.

Table 5. Output generated by Gemini 1.5 Pro for the same selected audio clips

Instrument	Emotion	Raga
Santoor	Joy/Happiness/Peace	Raga Yaman
Sitar	Tranquility/Peacefulness/Spiritual Devotion	Not enough information
Shehnai	Joyful/Celebratory/Devotional (depending on context, could also be mournful)	Not enough information
Vocal	Devotional/Spiritual/Solemn	Not enough information
Rudra Veena	Contemplative/Spiritual/Peaceful	Not enough information
Nadaswaram	Joyful/Celebratory/Auspicious	Not enough information
Vocal	Devotional/Spiritual/Solemn	Could be Dhrupad
Santoor	Peaceful / Contemplative / Meditative	Not enough information
Violin	Meditative/Contemplative/Spiritual	Not enough information
Vocal	Devotional/Spiritual/Peaceful	Not enough information
Sitar	Peaceful / Contemplative / Introspective	Not enough information
Sarod	Joyful/Lively/Optimistic	Not enough information
Sarod	Energetic/Lively/Virtuosic (demonstrating technical skill)	Fast taans particularly difficult to use for Raga identification

Table 6. Output generated by AudioLLM for the same selected audio clips

Instrument	Emotion	Raga
Sitar	sadness	Bhairavi
Sitar	sadness	Bhairavi
Sitar	sadness	Bhairavi
Flute	sadness	Bhairavi
Sitar	sadness	Bhairavi
Sitar	sadness	Bhairavi
Sitar	sadness	Bhairavi
Sitar	sadness	Bhairavi
Sitar	sadness	Bhairavi
Sitar	sadness	Bhairavi

Using the evaluation protocol outlined in the Methodology Section, the generated accuracy output have been shown in Table 6:

Table 7. Accuracy Table corresponding to the 5 AI models used

Model	Instrument	Emotion	Raga	Source
Audio LLM	30%	25%	0%	100%
GPT-4o	40%	20%	0%	83.33%
Gemini-1.5-Pro	25%	40%	0%	100%
Gamma	35.42%	31.67%	0%	87.23%
Qwen-2.5-omni	38.33%	18.33%	21.67%	100%

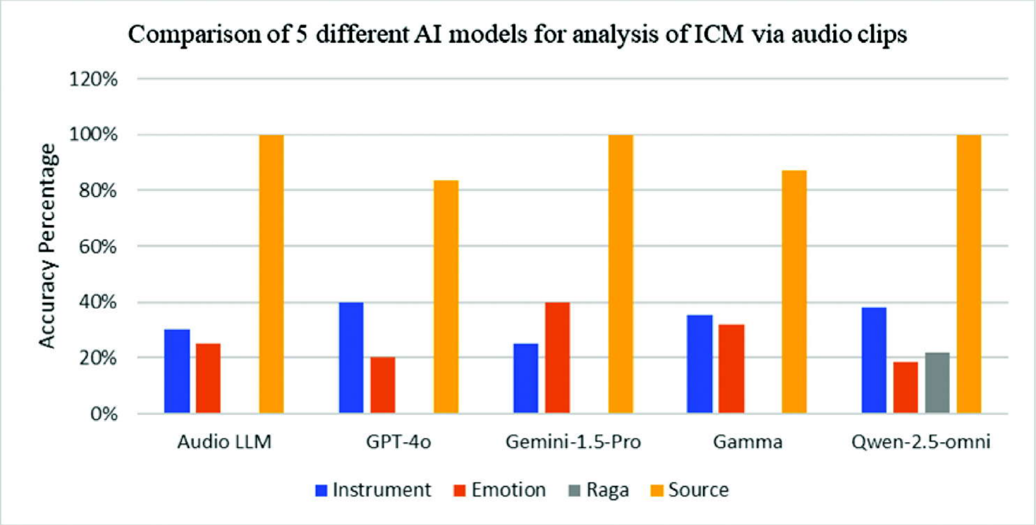


Fig. 2. Accuracy of the different AI models used in predicting different features of Hindustani Music.

Table 6 above shows the Comparative Study of accuracy for the output captured. Gemini and Gamma have shown good performance on Instrument and emotion analysis respectively. The well known ChatGPT AI indeed was good at all the features selected but could not perform in the realm of Raga. Qwen model proved to be a more dependable model for Raga prediction than all the other models. Below are some observations that were noted at the time of experiments for the different AI platforms during the time of data collected.

1. The proficiency and working methodology for Gamma & Gemini is on the higher scale while .ChatGPT & Audio LLM was less flexible for audio analysis on ICM
2. Raga prediction was better for Gamma than Gemini while Chat GPT and LLM was less proficient in this regards
3. All the models can predict the country of origin except LLM. The country of origin for Gemini showed errors compared to actual data. Also Gemini failed to list tuned models .It showed less proficiency to handle bulk data with frequent problems of user quota getting exceeded.
4. Audio LLM gave all the same input with no variation ,while Gamma could not make any difference between wind instruments and stringed ones. Instrument prediction was better for ChatGPT and Gemini.
5. The problem with Gamma is that there are certain usage restriction.So large numbers of samples above 50 would take longer time at an interval of 1 day for quota restriction policy.

In a nutshell,Gamma and Gemini can be considered for further introspection for the classification work.

4. CONCLUSION AND FUTURE DIRECTIONS

The present work is a first of its kind study which compares the predictive capacities of two widely used closed source AI models with that of three open source AI models in regard to various parameters generated from Hindustani music clips such as emotion, raga and instruments used. For this, a previously annotated database of HCM JUMusEmoDB was put to use to test and compare the predictive capacity of each of the models.

The study has the following interesting conclusions:

1. In case of instrument identification, closed source model GPT-4o outperforms others, as in most cases, the output prediction for GPT varied within the two stringed instruments put to test - sitar and sarod.
2. In case of emotion identification, Gemini 1.5 Pro performed the best, as the different emotions predicted by Gemini 1.5 was more or less in tandem with the co-elicited emotions found in case of Hindustani music. The accuracy values of open source platforms Gamma and Audio LLM in emotion prediction is almost at par with that of Gemini 1.5 Pro.
3. None of the models could predict raga accurately except for open source model Qwen-2.5 omni, which showed some expertise in identifying the complex nuances present in the Indian ragas.
4. Almost all the models showed sufficient expertise in identifying the source or origin of the clips, while some confused the clips with other countries of the Indian subcontinent.

The study posits the need for further training and validation of the neural network architectures, when it comes to assessment of various features of Hindustani classical music. The ragas provide an unique case study for the assessment and applicability of the audio models, as the ragas are continuously evolving from one artist to another, which makes their identification computationally all the more challenging. We are continuing this pilot study with close ended questions being asked to the models and comparing how they perform. The present study opens up new opportunities for future research using Artificial Intelligence in the unexplored domain of Hindustani Classical Music.

REFERENCES

[1] Civit M., Civit-Masot J., Cuadrado F. and Escalona, M. J., 2022. A systematic review of artificial intelligence-based music generation: Scope, applications, and future trends. *Expert Systems with Applications,* **209:** 118190.

[2] Bryce D., 2024. Artificial Intelligence and Music: Analysis of Music Generation Techniques Via Deep Learning and the Implications of AI in the Music Industry.

[3] Chu H., Kim J., Kim S., Lim H., Lee H., Jin S. and Ko S., 2022, October. An empirical study on how people perceive AI-generated music. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management,* pp. 304-314.

[4] Candusso S., 2024. *Exploring the impact of generative AI on the music composition market: a study on public perception, behavior, and industry implications* (Doctoral dissertation, Politecnico di Torino).

[5] Olayeni S., 2023. The impact of artificial intelligence (AI) in music business industry.

[6] Björklund G., Bohlin M., Olander E., Jansson J., Walter C. E. and Au-Yong-Oliveira M., 2022, April. An exploratory study on the Spotify recommender system. In *World Conference on Information Systems and Technologies,* pp. 366-378. Cham: Springer International Publishing.

[7] Gao X., Chen D. K., Gou Z., Ma L., Liu R., Zhao D. and Ham J., 2024. *AI-Driven Music Generation and Emotion Conversion. Affective and Pleasurable Design,* **123**(123).

[8] Bisht A. S., Negi C. M. S. and Singh R., 2022. Classification of Indian Classical Music (Hindustani Music) Genres through MFCCs Features using RNN-LSTM Model..

[9] Tejaswi Madhusudhan S. and Chowdhary G., 2024. *DeepSRGM--Sequence Classification and Ranking in Indian Classical Music with Deep Learning*. arXiv e-prints, arXiv-2402.

[10] Singh P. and Arora V., 2024. *Explainable Deep Learning Analysis for Raga Identification in Indian Art Music*. arXiv preprint arXiv:2406.02443.

[11] Shikarpur N. and Huang C. Z. A., 2024. *Exploratory Study Of Human-AI Interaction For Hindustani Music*. arXiv preprint arXiv:2411.13846.

[12] Das D. and Choudhury M., 2005, January. *Finite state models for generation of Hindustani classical music. In Proceedings of International Symposium on Frontiers of Research in Speech and Music,* pp. 59-64.

[13] Humse K., HS R. K., Veeraprathap V., Raju K., Ramachandra L.S., Yathiraj G. R. and Bhagyalakshmi R., 2025. Automated raga recognition in Indian classical music using machine learning techniques. *Journal of Integrated Science and Technology,* **13**(1), 1011-1011.

[14] Srinivasamurthy A., Gulati S., Repetto R. C. and Serra X., 2021. Saraga: Open datasets for research on indian art music. *Empirical Musicology Review,* **16**(1): 85-98.

[15] Gulati S., Serrà Julià J., Ganguli K. K., Sentürk S. and Serra X., 2016. Time-delayed melody surfaces for raga recognition.

[16] Verma V., 2017. Automatic mood classification of Indian Popular music. *International Journal for research in applied science and Engineering Technology (IJRASET),* **5**(VI).

[17] Sood A., Rathee T., Bansal P., Garg H., Aggarwal A. and Tyagi A., 2024. *Music Genre Classification using Artificial Neural Networks*.

[18] Banerjee S., 2017. A survey of prospects and problems in Hindustani classical raga identification using machine learning techniques. In *Proceedings of the first international conference on intelligent computing and communication,* pp. 467-475). Springer Singapore.

[19] Sarkar U., Nag S., Basu M., Banerjee A., Sanyal S., Sengupta R. and Ghosh D., 2021. *Neural network architectures to classify emotions in indian classical music.* arXiv preprint arXiv:2102.00616.

[20] Nag S., Basu M., Sanyal S., Banerjee A. and Ghosh D., 2022. On the application of deep learning and multifractal techniques to classify emotions and instruments using Indian Classical Music. *Physica A: Statistical Mechanics and its Applications,* **597:** 127261.

[21] Turnbull Douglas, *et al.,* Towards musical query-by-semantic-description using the cal500 data set, in: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 2007.

[22] Yang A., Yang B., Zhang B., Hui B., Zheng B., Yu B. and Qiu Z., 2024. *Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115.*

[23] Ghosh S., Kumar S., Seth A., Evuru C. K. R., Tyagi U., Sakshi S. and Manocha D., 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768.*

[24] Zhang H. and Shao H., 2024. Exploring the Latest Applications of OpenAI and ChatGPT: An In-Depth Survey. *CMES-Computer Modeling in Engineering & Sciences,* **138**(3).

[25] Carlà M. M., Giannuzzi F., Boselli F. and Rizzo S., 2024. Testing the power of Google DeepMind: Gemini versus ChatGPT 4 facing a European ophthalmology examination. *AJO International,* **1**(3), 100063.

[26] Li D., Tang C. and Liu H., 2024, July. Audio-LLM: Activating the Capabilities of Large Language Models to Comprehend Audio Data. *In International Symposium on Neural Networks,* pp. 133-142. Singapore: Springer Nature Singapore.

[27] Gu J., Jiang X., Shi Z., Tan H., Zhai X., Xu C. and Guo J., 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594.*

[28] Chowdhury S., Nag S., Joseph K. J., Srinivasan B. V. and Manocha D., 2024. Melfusion: Synthesizing music from image and language cues using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26826-26835.

INFORMATION FOR AUTHORS

ARTICLES

The Journal of Acoustical Society of India (JASI) is a refereed publication published quarterly by the Acoustical Society of India (ASI). JASI includes refereed articles, technical notes, letters-to-the-editor, book review and announcements of general interest to readers.

Articles may be theoretical or experimental in nature. But those which combine theoretical and experimental approaches to solve acoustics problems are particularly welcome. Technical notes, letters-to-the-editor and announcements may also be submitted. Articles must not have been published previously in other engineering or scientific journals. Articles in the following are particularly encouraged: applied acoustics, acoustical materials, active noise & vibration control, bioacoustics, communication acoustics including speech, computational acoustics, electro-acoustics and audio engineering, environmental acoustics, musical acoustics, non-linear acoustics, noise, physical acoustics, physiological and psychological acoustics, quieter technologies, room and building acoustics, structural acoustics and vibration, ultrasonics, underwater acoustics.

Authors whose articles are accepted for publication must transfer copyright of their articles to the ASI. This transfer involves publication only and does not in any way alter the author's traditional right regarding his/her articles.

PREPARATION OF MANUSCRIPTS

All manuscripts are refereed by at least two referees and are reviewed by the Publication Committee (all editors) before acceptance. Manuscripts of articles and technical notes should be submitted for review electronically to the Chief Editor by e-mail or by express mail on a disc. JASI maintains a high standard in the reviewing process and only accept papers of high quality. On acceptance, revised articles of all authors should be submitted to the Chief Editor by e-mail or by express mail.

Text of the manuscript should be double-spaced on A4 size paper, subdivided by main headings-typed in upper and lower case flush centre, with one line of space above and below and sub-headings within a section-typed in upper and lower case understood, flush left, followed by a period. Sub-sub headings should be italic. Articles should be written so that readers in different fields of acoustics can understand them easily. Manuscripts are only published if not normally exceeding twenty double-spaced text pages. If figures and illustrations are included then normally they should be restricted to no more than twelve-fifteen.

The first page of manuscripts should include on separate lines, the title of article, the names, of authors, affiliations and mailing addresses of authors in upper and lowers case. Do not include the author's title, position or degrees. Give an adequate post office address including pin or other postal code and the name of the city. An abstract of not more than 200 words should be included with each article. References should be numbered consecutively throughout the article with the number appearing as a superscript at the end of the sentence unless such placement causes ambiguity. The references should be grouped together, double spaced at the end of the article on a separate page. Footnotes are discouraged. Abbreviations and special terms must be defined if used.

EQUATIONS

Mathematical expressions should be typewritten as completely as possible. Equation should be numbered consecutively throughout the body of the article at the right hand margin in parentheses. Use letters and numbers for any equations in an appendix: Appendix A: (A1, (A2), etc. Equation numbers in the running text should be enclosed in parentheses, i.e., Eq. (1), Eqs. (1a) and (2a). Figures should be referred to as Fig. 1, Fig. 2, etc. Reference to table is in full: Table 1, Table 2, etc. Metric units should be used: the preferred from of metric unit is the System International (SI).

REFERENCES

The order and style of information differs slightly between periodical and book references and between published and unpublished references, depending on the available publication entries. A few examples are shown below.

Periodicals:

[1] S.R. Pride and M.W. Haartsen, 1996. Electroseismic wave properties, *J. Acoust. Soc. Am.*, **100** (3), 1301-1315.

[2] S.-H. Kim and I. Lee, 1996. Aeroelastic analysis of a flexible airfoil with free play non-linearity, *J. Sound Vib.*, **193** (4), 823-846.

Books:

[1] E.S. Skudzryk, 1968. *Simple and Comlex Vibratory Systems*, the Pennsylvania State University Press, London.

[2] E.H. Dowell, 1975. *Aeroelasticity of plates and shells*, Nordhoff, Leyden.

Others:

[1] J.N. Yang and A. Akbarpour, 1987. Technical Report NCEER-87-0007, Instantaneous Optimal Control Law For Tall Buildings Under Seismic Excitations.

SUMISSIONS

All materials from authors should be submitted in electronic form to the JASI Chief Editor: B. Chakraborty, CSIR - National Institute of Oceanography, Dona Paula, Goa-403 004, Tel: +91.832.2450.318, Fax: +91.832.2450.602,(e-mail: bishwajit@nio.org) For the item to be published in a given issue of a journal, the manuscript must reach the Chief Editor at least twelve week before the publication date.

SUMISSION OF ACCEPTED MANUSCRIPT

On acceptance, revised articles should be submitted in electronic form to the JASI Chief Editor (bishwajit@nio.org)