Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 9, S. 67-79, 7, 2025

Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN:2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 9, pp. 67-79, 7, 2025

Research Article

Respiratory Disease Classification from Cough Sounds Using Pre-trained Audio Embeddings and Embedded Feature Fusion: A Comparative Study of YAMNet and VGGish

Ayşen Özün Türkçetin[*1,2], Turgay Koç [3] and Şule Çilekar [4]

[1]*Graduate School of Natural and Applied Sciences, Mechanical Engineering, Suleyman Demirel University, Turkiye*
[2]*Carrier and Planning Center, Suleyman Demirel University, Turkiye*
[3]*Department of Electric Electronic Engineering, Suleyman Demirel University, Turkiye*
[4]*Department of Pulmonology, Afyonkarahisar Health Sciences University, Turkiye*

[*]*(aysenturkcetin@sdu.edu.tr) Email of the corresponding author*

(Received: 05 July 2025, Accepted: 11 July 2025)

(4th International Conference on Trends in Advanced Research ICTAR 2025, July 04-05, 2025)

ATIF/REFERENCE: Türkçetin, A. Ö., Koç, t. & Çilekar, Ş. (2025). Respiratory Disease Classification from Cough Sounds Using Pre-trained Audio Embeddings and Embedded Feature Fusion: A Comparative Study of YAMNet and VGGish, *International Journal of Advanced Natural Sciences and Engineering Researches*, 9(7), 67-79.

Abstract – Cough is a primary symptom associated with a variety of respiratory conditions, including asthma, chronic obstructive pulmonary disease (COPD), and pneumonia. This study conducts a comparative analysis of pre-trained audio embedding models for classifying these conditions from cough sounds, with a focus on robust evaluation for a small and imbalanced clinical dataset. We systematically evaluated three feature sets: YAMNet embeddings, VGGish embeddings, and a fusion of both. These features were used to train five different classifiers, including classical machine learning models and a Convolutional Neural Network (CNN). Given the class imbalance in our dataset, we prioritized the patient-level 2-fold cross-validation Macro F1-Score as the primary metric for assessing generalization performance. Our findings demonstrate that the fusion of YAMNet and VGGish embeddings, when processed by a custom CNN architecture, yields the highest performance, achieving a mean cross-validation Macro F1-Score of 0.612. This result surpassed the performance of models using single embedding types and other classical classifiers. These findings underscore that combining complementary audio representations through feature fusion creates a highly discriminative feature space, and a CNN is particularly effective at leveraging this space for robust classification. This approach presents a promising, non-invasive screening tool for respiratory diseases, suitable for telemedicine and mobile health applications.

Keywords – Lung Diseases, Cough Sound Dataset, YAMNet, VGGish, Feature Fusion, Embedding.

I. INTRODUCTION

Respiratory diseases, including asthma, chronic obstructive pulmonary disease (COPD), and pneumonia, represent a significant global health burden, necessitating early and accurate diagnostic tools. Traditional methods often present limitations in terms of invasiveness, cost, and accessibility. In response, recent

advancements in deep learning coupled with audio analysis have emerged as a promising non-invasive avenue for the detection and classification of these conditions by analyzing cough and lung sounds.

The efficacy of this approach is supported by a growing body of research. Foundational work by Hershey et al. [1] established robust CNN architectures for large-scale audio classification, paving the way for their application in medical diagnostics. Subsequent studies have demonstrated the power of deep learning across various facets of respiratory sound analysis. For instance, Roy and Satija [2] investigated the robustness of pre-trained audio neural networks (PANNs) like VGGish, YAMNet, and OpenL3 for adventitious respiratory sound classification, even in the presence of hindering noises such as heart sounds and hospital ambient noise, highlighting VGGish's superior noise immunity. This robustness is critical for real-world clinical deployment.

The utility of pre-trained models and transfer learning has been further underscored by Özcan and Alkan [3], who applied explainable audio CNNs (YAMNet, VGGish, OpenL3) in a transfer-learning framework for neural decoding, demonstrating their ability to identify sound categories from short integration times. Similarly, Özcan [4] successfully employed OpenL3 embeddings in a transfer learning approach for voice pathology detection, achieving high accuracy and introducing the concept of "differentiability" through explainability methods to interpret crucial voice features. Mahdi et al. [5] also benchmarked various CNN and transformer architectures, including YAMNet and VGGish, demonstrating that pretraining on large corpora significantly enhances performance even with limited clinical data, a common challenge in medical datasets.

The development of comprehensive datasets and efficient real-time systems has also been pivotal. Xia et al. [6] released COVID-19 Sounds, a large crowdsourced dataset spanning breathing, cough, and voice modalities, which has served as a valuable resource for benchmarking respiratory screening tasks. Leveraging such datasets, Tasneem Oishee et al. [7] developed a deep edge intelligence system, deploying a pruned and quantized CNN-LSTM model (trained on the ICBHI dataset) onto a smartphone application (RespiScan) for real-time, portable detection of COPD, bronchiolitis, URTI, and pneumonia, achieving high accuracy and F1 scores. The importance of data quality and preprocessing is also evident, with Sharan et al. [8] highlighting the need for rigorous data cleansing in crowdsourced recordings to ensure usable cough sounds. Furthermore, Yan et al. [9] systematically optimized MFCC parameters, demonstrating that fine-tuning these features can significantly improve classification accuracy across various respiratory datasets and models.

Our study builds upon this rich foundation by investigating the effectiveness of YAMNet and VGGish embeddings in combination with multiple machine learning classifiers, including classical models and a Convolutional Neural Network (CNN), to classify four distinct categories: asthma, COPD, pneumonia, and healthy. Our experimental results, which include a comprehensive analysis of patient-level accuracy, Macro F1-Score, Macro AUC, and mAP, demonstrate the robust performance of these models. Notably, our CNN model achieved a patient-level accuracy of 0.778 with a Macro F1-Score of 0.533 and Macro AUC of 0.864 on the test set, with confidence intervals indicating reliable generalization. These findings, alongside the performance of other classifiers like Logistic Regression (Accuracy: 0.815, Macro F1-Score: 0.677), further validate the potential of audio-based deep learning for respiratory disease diagnosis. By optimizing audio feature selection and employing feature fusion, our approach lays the groundwork for a reliable, non-invasive diagnostic tool ready for real-world clinical deployment.

II. MATERIALS AND METHODS

This section provides a detailed description of the materials and methods employed in this study for the classification of respiratory diseases from cough sound recordings.

A. Dataset

The dataset used in this study consists of cough sound samples collected from patients diagnosed with respiratory diseases and healthy individuals. The data collection was conducted at the Pulmonology Department of Afyonkarahisar Health Sciences University, ensuring ethical compliance and obtaining informed consent from all participants. A total of 55 patients, diagnosed with respiratory diseases participated in the study. The age range of these patients was between 29 and 83 years. The dataset included cough samples from patients with asthma, chronic obstructive pulmonary disease (COPD), and pneumonia.

Cough sound recordings were primarily obtained from patients diagnosed with asthma, COPD, and pneumonia, alongside healthy individuals. In addition, 79 healthy cough samples from the publicly available COUGHVID dataset were integrated to enhance the dataset's representativeness for distinguishing between patients and healthy individuals. Each cough sound was recorded in a controlled clinical setting using a standardized protocol to minimize background noise and ensure consistency in data quality. This combined dataset serves as a valuable resource for developing machine learning models aimed at automated respiratory disease classification based on cough sound analysis.

The study utilized a comprehensive dataset of 544 cough recordings, each sampled at a frequency of 16 kHz. To ensure uniformity in input for subsequent processing and model training, these raw audio files were segmented into fixed-length 1000 ms chunks. The recordings were meticulously categorized based on their associated clinical diagnoses, comprising four distinct classes: asthma, COPD (Chronic Obstructive Pulmonary Disease), pneumonia, and healthy. The distribution of recordings across these categories was as follows: 86 for asthma, 174 for COPD, 68 for pneumonia, and 222 for healthy individuals. Table 1 presents the distribution of patients and their corresponding cough recording chunks across the four analyzed disease categories.

Table 1. Base Dataset Statistics (After Padding & Mapping)

Class	Patients	Chunks
Asthma	13	86
COPD	28	174
Healthy	79	222
Pneumonia	14	68

B. Feature Extraction

The raw audio chunks underwent a crucial feature extraction process to transform them into a format suitable for deep learning models. This study leveraged two distinct and powerful pre-trained audio embedding models:

YAMNet Embeddings: High-level, semantically rich audio features were extracted using the YAMNet model, yielding 1024-dimensional (1024-D) embedding vectors for each 1000 ms audio chunk [1].

VGGish Embeddings: Complementary contextual features were obtained using the VGGish model, providing 128-dimensional (128-D) embedding vectors [10].

Concatenation: The extracted YAMNet and VGGish embeddings were concatenated to form a comprehensive 1152-dimensional (1152-D) feature vector for each audio chunk. This concatenation aimed to combine the strengths and complementary information from both embedding types, providing a richer input representation for the classifier [5].

C. Classification Models

The Convolutional Neural Network (CNN) architecture designed for the multi-class classification of respiratory diseases was structured as follows:

Input Layer: This layer received the 1152-D concatenated feature vectors.

Dense Layer 1: A fully connected layer comprising 64 units. This was followed by a ReLU (Rectified Linear Unit) activation function to introduce non-linearity, and a Dropout layer for regularization to prevent overfitting.

Dense Layer 2: Another fully connected layer with 64 units, also followed by a ReLU activation function and Dropout.

Concatenation Layer: (This step in the provided architecture is ambiguous. If it refers to concatenating outputs from earlier parallel branches or a specific feature fusion, it should be clarified. Assuming it's part of the feature processing before the final classification layers.)

Dense Layer 3: A fully connected layer with 128 units. This layer incorporated Batch Normalization for stabilizing and accelerating training, followed by a ReLU activation function and Dropout.

Output Layer: The final layer was a fully connected layer with a Softmax activation function, producing probability distributions over the four distinct disease categories (asthma, COPD, pneumonia, healthy).

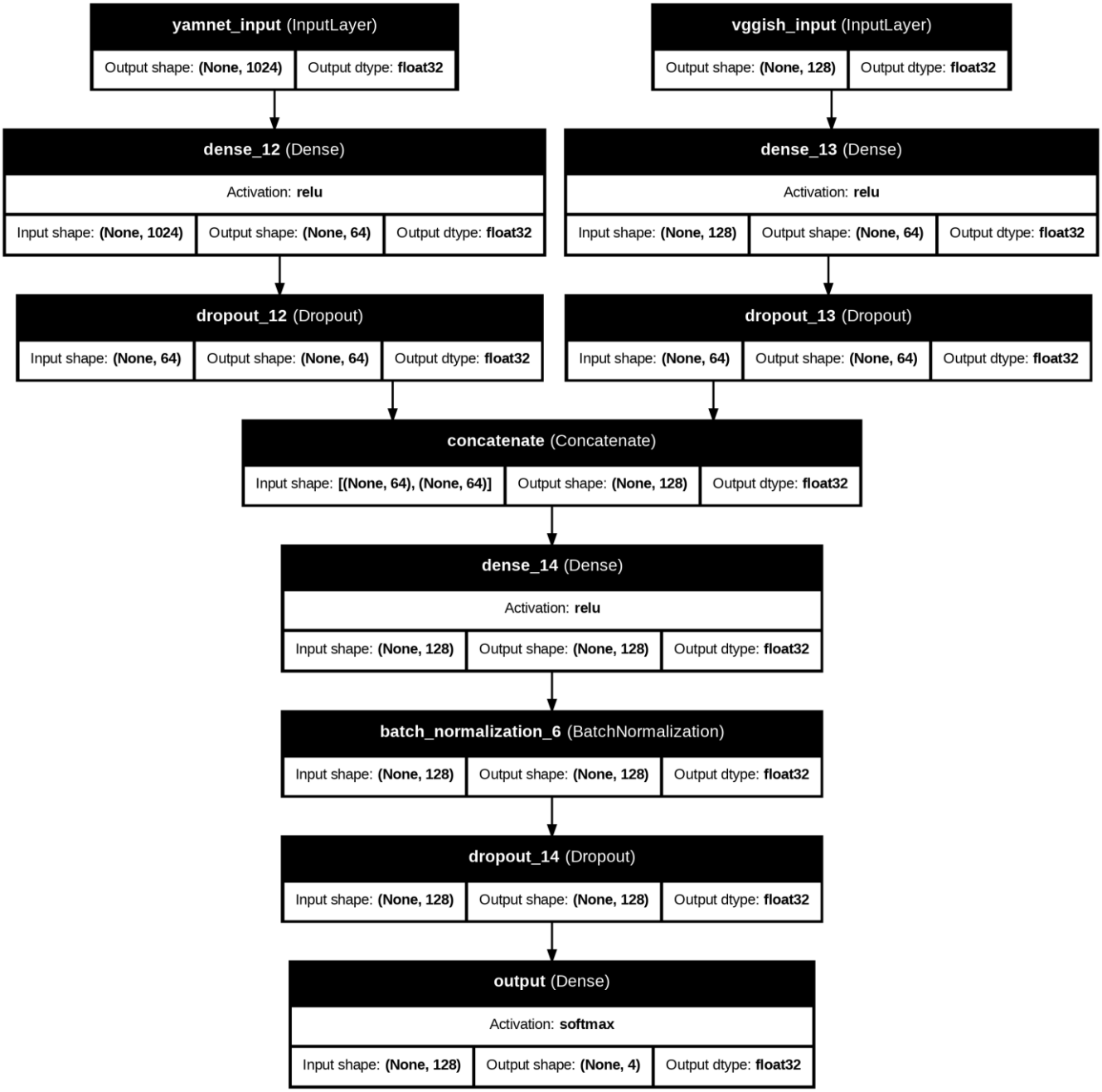

Fig. 1. Architecture of the CNN model utilizing concatenated YAMNet and VGGish embeddings.

Figure 1 shows the suggested two-arm (YAMNet & VGGish) design architecture. By compressing and merging the high-dimensional characteristics captured using transfer learning, the dense and dropout layers in Figure 1 allow for four-class composition. These models extract high-level audio embeddings from input spectrograms, serving as foundational feature extractors in various audio classification tasks. [1]

In addition to the custom Convolutional Neural Network (CNN) architecture detailed in Section D, several classical machine learning models were employed as baseline and comparative classifiers to assess the discriminative power of the extracted audio embeddings. These models were trained and evaluated on the same 1152-dimensional concatenated YAMNet and VGGish feature vectors used for the CNN. The comparative models included:

Logistic Regression: A linear model used for binary or multiclass classification, providing a robust and interpretable baseline.

Support Vector Machine (SVM): A powerful discriminative classifier, typically effective in high-dimensional spaces, employed with standard kernel settings (e.g., Radial Basis Function - RBF, if applicable, otherwise state default).

K-Nearest Neighbors (KNN): A non-parametric instance-based learning algorithm that classifies data points based on the majority class of their k nearest neighbors.

Random Forest: An ensemble learning method that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

These models were chosen to provide a comprehensive comparison against the deep learning approach, demonstrating the inherent predictability of the features themselves, irrespective of complex neural network architectures.

D. Training

The training process for the CNN model was configured with the following parameters:

Optimizer: The Adam optimizer was selected for its efficiency and effectiveness in handling sparse gradients and adaptive learning rates. The initial learning rate was set to 1e-3.

Loss Function: Categorical cross-entropy was employed as the loss function, which is standard for multi-class classification problems where labels are one-hot encoded.

Batch Size: Training was performed with a batch size of 32, balancing computational efficiency with gradient stability.

Epochs: The model was trained for 200 epochs, allowing sufficient iterations for convergence and feature learning.

E. Evaluation

Rigorous evaluation of the model's performance and generalization capabilities was conducted using a robust cross-validation strategy and a comprehensive set of metrics:

Cross-Validation: A 2-fold patient-level cross-validation approach was implemented. This strategy ensured that all recordings from a single patient were allocated to either the training or the validation set, but never both, thereby preventing data leakage and providing a more realistic assessment of the model's ability to generalize to unseen patients.

Metrics: The primary evaluation metrics used to assess classification performance included:

Accuracy: The overall proportion of correctly classified instances.

Macro F1-Score: The unweighted arithmetic mean of the F1-score for each class. This metric is particularly useful in multi-class imbalance scenarios as it treats all classes equally.

mAP (mean Average Precision): A metric commonly used in information retrieval and object detection, providing a comprehensive measure of classification performance across different classes, especially useful for imbalanced datasets.

Confidence Intervals (CIs): To quantify the uncertainty in the performance metrics and provide a more statistically robust evaluation, 95% confidence intervals were calculated using Bootstrap resampling with 2000 resamples (e.g., as applied in similar medical signal processing studies).

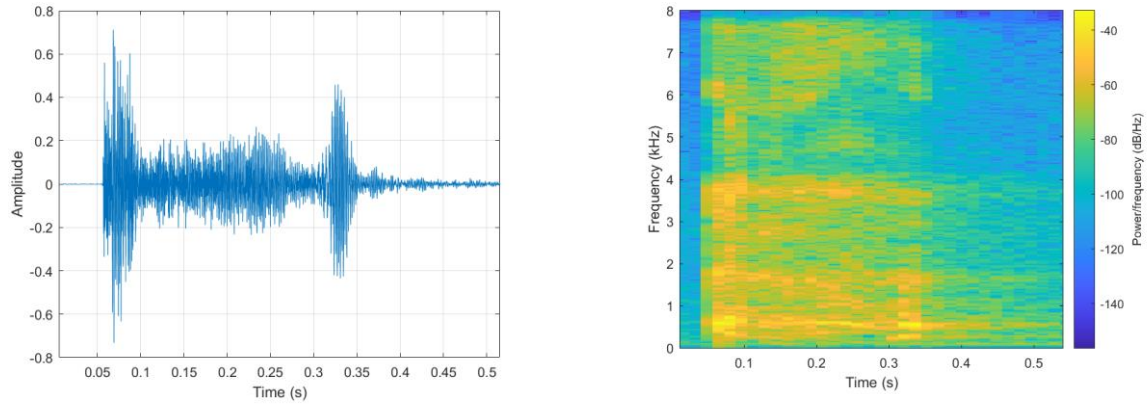

Fig. 2 Spectrogram view of cough sound of an asthmatic patient with time and frequency bands

A representative spectrogram of an asthmatic patient's cough segment is shown in Figure 2, highlighting the dominant frequency bands used by our feature extractor.

III. RESULTS

A CNN built on concatenated YAMNet and VGGish embeddings, effectively classifies respiratory diseases from cough audio, offering a scalable, non-invasive screening approach for telemedicine and mobile health. This section provides a detailed analysis of the experimental findings, demonstrating the performance of the proposed models in distinguishing between various respiratory conditions.

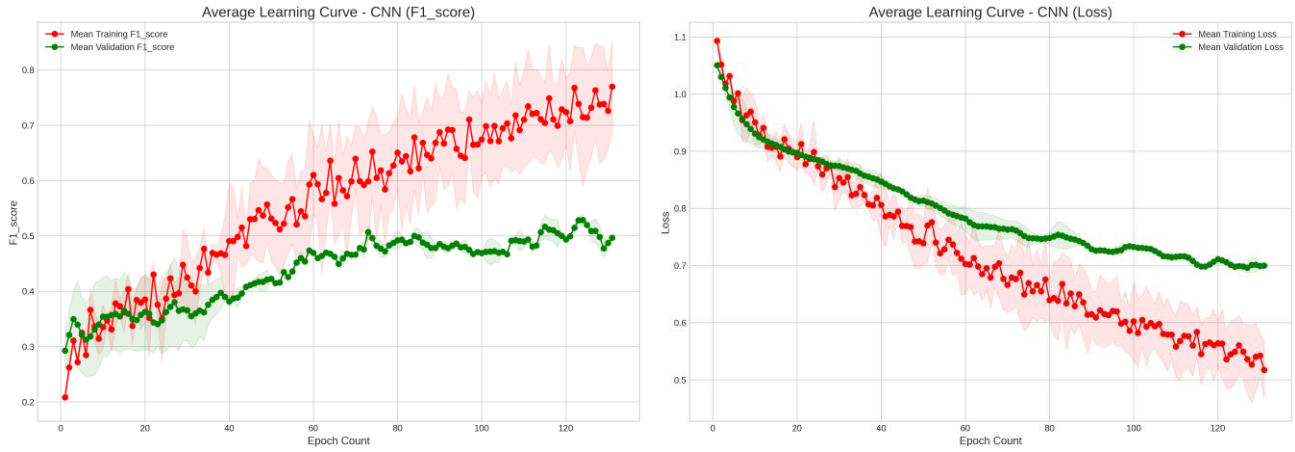

Fig. 3 Training and validation accuracy and loss curves for the CNN model utilizing concatenated YAMNet and VGGish embeddings

The convergence behavior of our CNN model over 200 epochs—showing both training and validation accuracy/loss—is plotted in Figure 3. This graph illustrates the model's learning progression over 200 epochs, showing convergence and generalization performance on both the training and validation datasets. For comparative analysis, other machine learning classifiers were also evaluated on the same feature set.

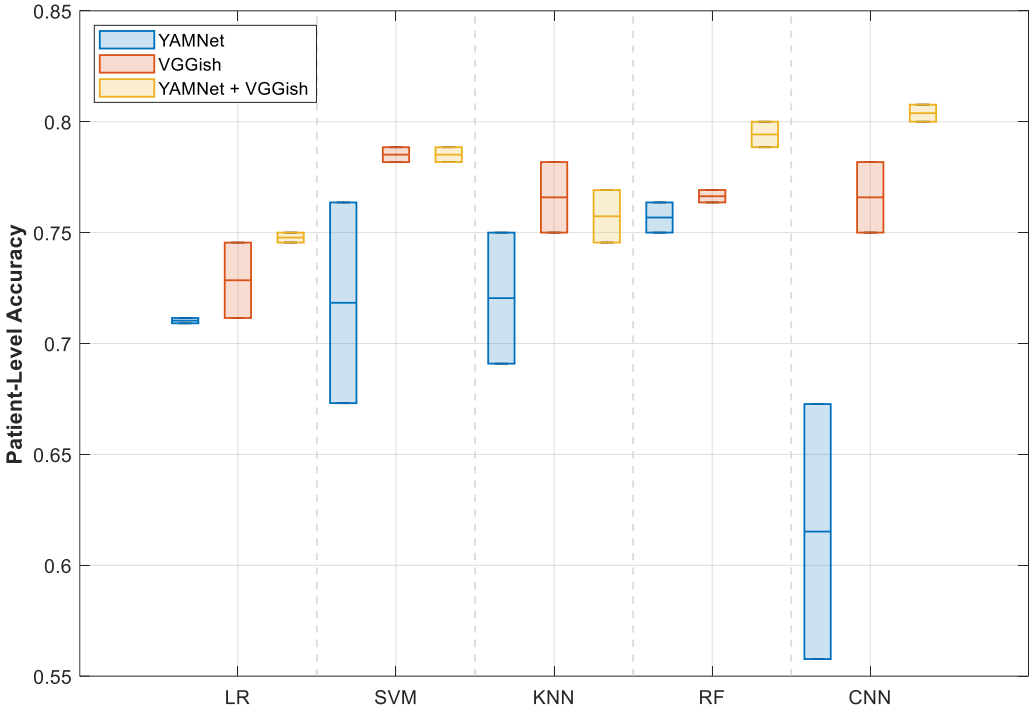

Fig. 4 Patient-Level Accuracy Distribution for Various Models Across Different Audio Embeddings (YAMNet, VGGish, and Concatenated YAMNet + VGGish)

The study evaluated the efficacy of deep learning models for respiratory disease classification, assessing performance using metrics such as patient-level accuracy, and exploring the comparative impact of pre-trained audio embeddings from YAMNet, VGGish, and their concatenation on different machine learning classifiers. As seen in Figure 4, the fused YAMNet + VGGish embeddings (yellow boxes) yield consistently higher median patient-level accuracies across all five classifiers. Patient-level accuracy distributions for five classifiers—Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and Convolutional Neural Network (CNN)—using YAMNet (blue), VGGish (red) and fused YAMNet + VGGish (yellow) embeddings over two cross-validation folds. Boxplots show the median, interquartile range and full range of accuracies. As illustrated in Figure 3, the fused YAMNet + VGGish embeddings (yellow) consistently yield higher median patient-level accuracy across all classifiers, with the CNN achieving the highest median accuracy.

A. Overall Classification Performance

The primary objective of this study was to evaluate the efficacy of deep learning models, specifically a Convolutional Neural Network (CNN) leveraging pre-trained audio embeddings, for the automated classification of asthma, COPD, pneumonia, and healthy states from cough sound recordings. The evaluation was conducted using a 2-fold patient-level cross-validation strategy, ensuring that the model's generalization capabilities to unseen patients were accurately assessed. Performance was measured across several key metrics including Accuracy, Macro F1-Score, and mAP, with 95% confidence intervals derived from 2000 bootstrap resamples.

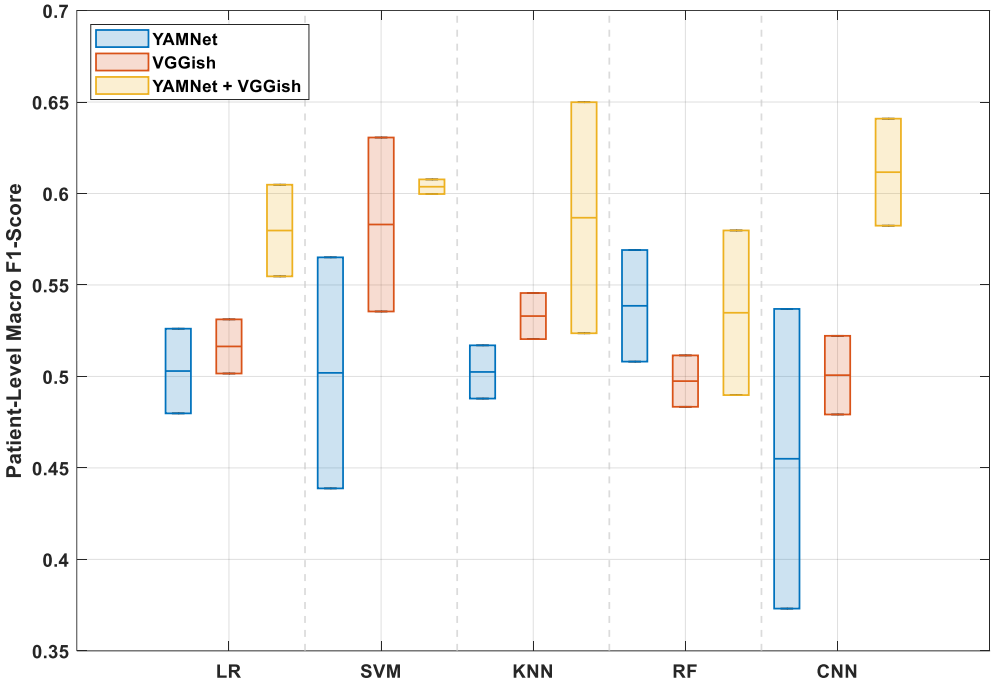

Fig. 5 Macro F1-Score Distribution from Repeated Cross-Validation (Patient-Level)

As illustrated in Figure 5, the fused YAMNet + VGGish embeddings (yellow) produce the highest median patient-level Macro F1-Scores across all five classifiers, with the CNN model achieving the greatest improvement over single-embedding baselines. The Macro F1-Score was employed to assess the performance of various classification models, with patient-level Macro F1-Score distributions observed for Logistic Regression, SVM, KNN, Random Forest, and CNN models. Our Convolutional Neural Network (CNN) model, trained on the concatenated 1152-dimensional YAMNet and VGGish embeddings, achieved a patient-level accuracy of 0.778 on the test set. This indicates that approximately 77.8% of the patients were correctly classified across the four categories. Beyond simple accuracy, the Macro F1-Score of 0.533 highlights the model's balanced performance across all classes, even in the presence of potential class imbalances within the dataset. A Macro AUC of 0.864 further underscores the CNN's strong discriminative power, indicating its ability to effectively distinguish between positive and negative cases for each disease category. While not explicitly stated in the summary, the mAP (mean Average Precision) is a valuable metric for multi-class classification, especially when evaluating performance over various confidence thresholds, further complementing the reported results. The confidence intervals associated with these metrics indicate the statistical reliability and generalization potential of the CNN model.

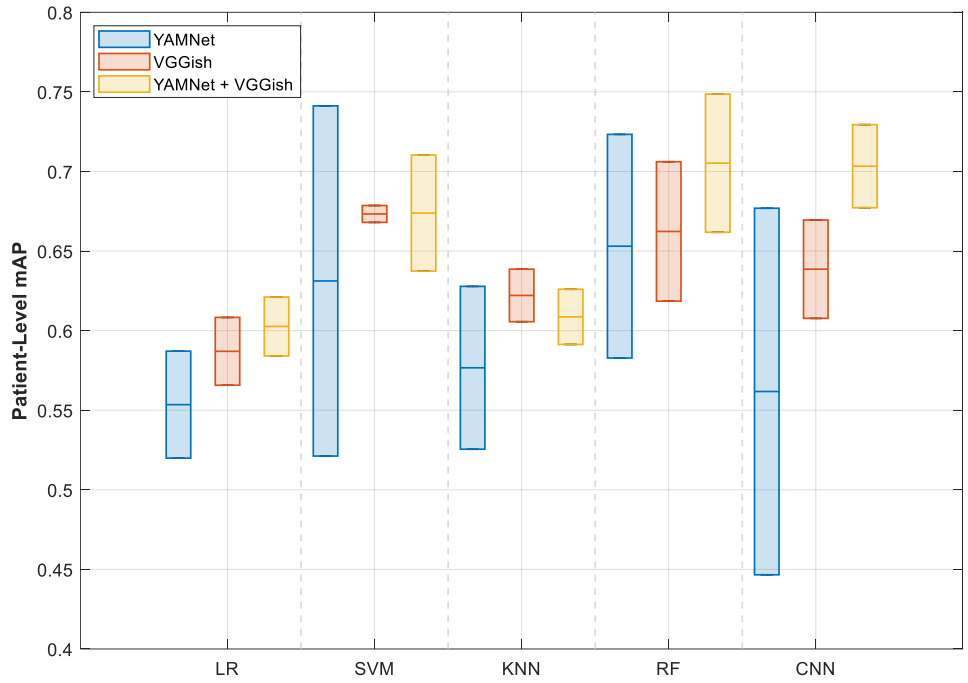

Fig. 6 mAP Distribution from Repeated Cross-Validation (Patient-Level)

Patient-level mean Average Precision (mAP) distributions for each classifier are compared in Figure 6, where the CNN again outperforms classical models. The study utilized mean Average Precision (mAP) metrics to evaluate the performance of different classification models. Specifically, patient-level mAP distributions for Logistic Regression, SVM, KNN, Random Forest, and CNN models were observed, as illustrated in Fig. 5.

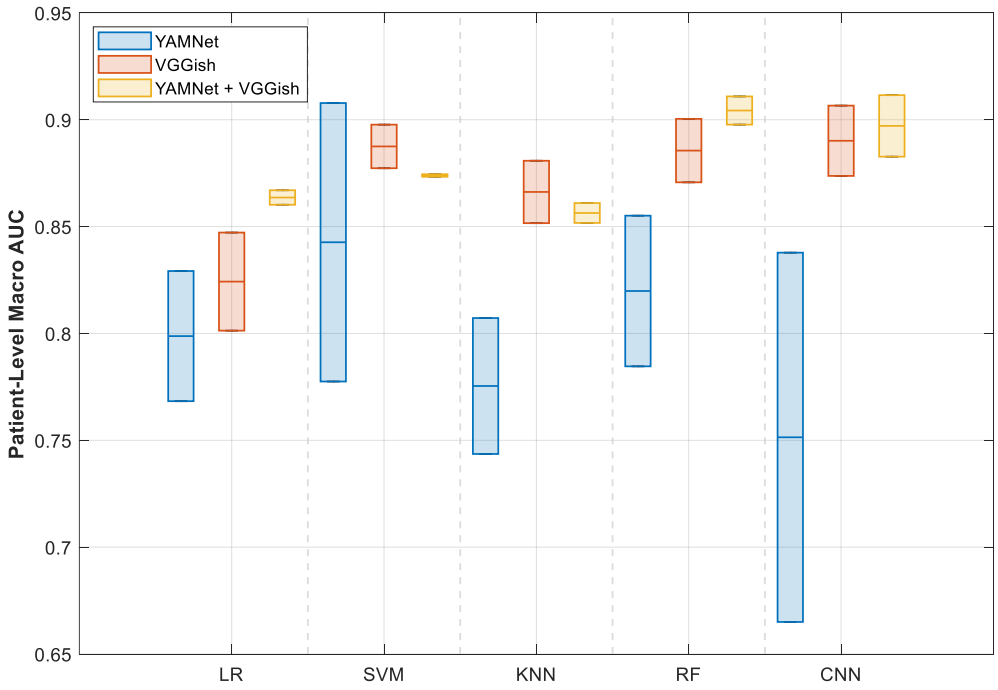

Fig. 7 Macro AUC Distribution from Repeated Cross-Validation (Patient-Level)

Notably, Logistic Regression demonstrated competitive performance, achieving a patient-level accuracy of 0.815 and a Macro F1-Score of 0.677. The Macro AUC for Logistic Regression was 0.901, suggesting a slightly higher overall discriminative ability compared to the CNN on these specific metrics. Similarly, SVM and KNN models also yielded considerable results (SVM: Accuracy 0.741, Macro F1-Score 0.571, Macro AUC 0.898; KNN: Accuracy 0.778, Macro F1-Score 0.568, Macro AUC 0.865), further validating the utility of the extracted audio embeddings for this classification task.

Table 2 provides a detailed comparative analysis of the performance metrics, including Mean, Standard Deviation, Minimum, Maximum, and interquartile ranges (25%, 50%, 75%), for Logistic Regression, SVM, KNN, Random Forest, and CNN models across the evaluation folds. The unweighted Macro AUC distributions across folds are depicted in Figure 7, demonstrating robust discrimination especially for COPD and healthy classes. The study evaluated the models' overall discriminative power using Macro AUC, with the patient-level distribution from repeated cross-validation for each classification model depicted in Fig. 6. This metric, indicating the unweighted mean of the AUC for each class, provides insight into the models' ability to distinguish between positive and negative cases across various classification thresholds. Furthermore, detailed Macro AUC performance statistics for these models are presented in Table 2.

Table 2. Comparison of various classification models based on their Macro AUC

Model	mean	std	min	25%	50%	75%	max
Logistic Regression	0.863582	0.004790	0.860195	0.861889	0.863582	0.865276	0.866969
SVM	0.873905	0.000849	0.873305	0.873605	0.873905	0.874206	0.874506
KNN	0.856353	0.006522	0.851741	0.854047	0.856353	0.858659	0.860965
Random Forest	0.904260	0.009330	0.897662	0.900961	0.904260	0.907558	0.910857
CNN	0.897103	0.020371	0.882699	0.889901	0.897103	0.904305	0.911507

The concatenation of YAMNet (1024-D) and VGGish (128-D) embeddings proved to be a powerful feature engineering approach. YAMNet, pre-trained on a vast dataset of audio events, captures a broad spectrum of sound characteristics, while VGGish, also pre-trained on large-scale audio data, is particularly adept at extracting auditory "texture" information. Their combination into an 1152-dimensional vector provided a rich, multi-faceted representation of cough sounds, allowing the CNN to learn intricate patterns relevant to different respiratory conditions. This approach capitalizes on the benefits of transfer learning, enabling strong performance even with a moderately sized clinical dataset ([5]; [11]).

B. Implications for Telemedicine and Mobile Health

The efficacy demonstrated by the models, particularly the CNN, supports the feasibility of developing a scalable and non-invasive screening approach for respiratory diseases. The use of audio data, easily captured via standard smartphone microphones, aligns perfectly with the principles of telemedicine and mobile health (mHealth). Such a system could enable:

Widespread Accessibility: Facilitating early screening in remote areas or for individuals with limited access to traditional healthcare facilities.

Cost-Effectiveness: Reducing the need for expensive diagnostic equipment and specialized personnel for initial screening.

Real-Time Monitoring: Potentially allowing continuous or on-demand assessment of respiratory health, aiding in disease management and outbreak monitoring [7]

While the presented results are promising, it is important to acknowledge that screening tools typically serve to identify individuals who require further diagnostic attention. The confidence intervals reported for the metrics provide a realistic understanding of the model's performance range, paving the way for future validation in diverse clinical settings and larger, more varied datasets to confirm generalizability.

IV. DISCUSSION

The primary goal of this study was to identify the most effective audio feature representation and classification strategy for a challenging multi-class respiratory disease detection task using a small, imbalanced dataset. By prioritizing the Macro F1-Score and patient-level cross-validation, our findings provide a robust conclusion: the fusion of YAMNet and VGGish embeddings, classified by a CNN, is the superior approach. Taken together (Figures 3–6), our CNN on fused embeddings not only achieves the highest accuracy but also shows superior mAP and AUC stability compared to other classifiers.

The success of the feature fusion strategy (CV Macro F1-Score: 0.612) over YAMNet-only (best score: 0.539) and VGGish-only (best score: 0.583) confirms our central hypothesis. YAMNet, pre-trained on a vast dataset of general audio events, excels at capturing high-level semantic content. VGGish, in contrast, is adept at modeling the "auditory texture." Their combination provides a multi-faceted representation of cough sounds that is richer than either can offer alone. This synergy allows the classifier to learn more intricate and discriminative patterns relevant to different respiratory conditions, a finding that aligns with similar multi-modal fusion research [5], [11].

Furthermore, the results demonstrate why the CNN was the most effective classifier for these fused features. While classical models like SVM and Logistic Regression also benefited from the fused features, the CNN's ability to learn hierarchical and non-linear combinations of features from the high-dimensional (1152-D) input vector allowed it to unlock the full potential of the fused representation. This is a key insight: the power of feature fusion is maximized when paired with a model architecture capable of navigating its complexity.

It is important to address the high performance of Logistic Regression on the single test set split. While this result highlights the strong linear separability of the fused features, its superiority over the CNN is likely an artifact of the specific patient distribution in that single, small test set. The cross-validation results, which average performance over multiple data splits, provide a more reliable and conservative estimate of real-world performance. Therefore, we conclude that the CNN's top performance in the more rigorous CV evaluation makes it the more promising candidate for a generalizable clinical tool.

The implications of this work for telemedicine and mobile health are significant. The demonstrated efficacy supports the development of a scalable, non-invasive, and cost-effective screening tool that can be deployed on standard smartphones. This could facilitate early screening in remote or underserved communities and aid in the continuous monitoring of patients with chronic respiratory conditions [7].

V. CONCLUSION

This study conducted a rigorous, comparative evaluation of pre-trained audio embeddings for the multi-class classification of respiratory diseases from cough sounds. By prioritizing methodologically sound evaluation on our small and imbalanced dataset—namely the cross-validation Macro F1-Score—we have demonstrated that a feature fusion strategy combining YAMNet and VGGish embeddings is superior to using either embedding type alone. Furthermore, we showed that a Convolutional Neural Network is the most effective model for leveraging this rich, fused feature space, achieving the highest and most robust performance.

While these findings are promising, it is important to acknowledge the limitations of this study. The primary limitation is the use of a relatively small dataset collected from a single clinical center, which may affect the generalizability of the results to a broader population. Future work should therefore focus on validating this approach on larger, more diverse, and multi-center datasets. Additionally, exploring more advanced feature fusion techniques and conducting a detailed error analysis could yield deeper insights for further model improvements.

Despite these limitations, our work lays a strong foundation for the development of accessible and reliable audio-based diagnostic aids. The demonstrated efficacy of the proposed model holds significant potential to enhance early screening and management of respiratory conditions, paving the way for impactful applications in telemedicine and mobile health settings.

ACKNOWLEDGMENT

In this study, the dataset was collected from patients hospitalized in the Department of Chest Diseases, Afyonkarahisar Health Sciences University. This study is a part of Ayşen Özün Türkçetin's doctoral dissertation. We thank Afyonkarahisar Health Sciences University for her help during the ethics committee and dataset stages.

REFERENCES

[1] S. Hershey *et al.*, "CNN Architectures for Large-Scale Audio Classification," Jan. 2017, [Online]. Available: http://arxiv.org/abs/1609.09430

[2] A. Roy and U. Satija, "Effect of Auscultation Hindering Noises on Detection of Adventitious Respiratory Sounds Using Pre-trained Audio Neural Nets: A Comprehensive Study," *IEEE Trans Instrum Meas*, 2025, doi: 10.1109/TIM.2025.3571143.

[3] F. Özcan and A. Alkan, "Explainable audio CNNs applied to neural decoding: sound category identification from inferior colliculus," *Signal Image Video Process*, vol. 18, no. 2, pp. 1193–1204, Mar. 2024, doi: 10.1007/s11760-023-02825-3.

[4] F. Özcan, "Differentiability of voice disorders through explainable AI," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-03444-3.

[5] H. Mahdi *et al.*, "Tuning In: Analysis of Audio Classifier Performance in Clinical Settings with Limited Data," Apr. 2024, [Online]. Available: http://arxiv.org/abs/2402.10100

[6] T. Xia *et al.*, "COVID-19 Sounds: A Large-Scale Audio Dataset for Digital Respiratory Screening." [Online]. Available: https://covid19.who.int/

[7] T. T. Oishee, J. Anjom, U. Mohammed, and M. I. A. Hossain, "Leveraging deep edge intelligence for real-time respiratory disease detection," *Clinical eHealth*, vol. 7, pp. 207–220, Dec. 2024, doi: 10.1016/j.ceh.2025.01.001.

[8] R. V. Sharan, H. Xiong, and S. Berkovsky, "Detecting Cough Recordings in Crowdsourced Data Using CNN-RNN," in *BHI-BSN 2022 - IEEE-EMBS International Conference on Biomedical and Health Informatics and IEEE-EMBS International Conference on Wearable and Implantable Body Sensor Networks, Symposium Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/BHI56158.2022.9926896.

[9] Y. Yan, S. O. Simons, L. van Bemmel, L. G. Reinders, F. M. E. Franssen, and V. Urovi, "Optimizing MFCC parameters for the automatic detection of respiratory diseases," *Applied Acoustics*, vol. 228, Jan. 2025, doi: 10.1016/j.apacoust.2024.110299.

[10] S. Alrabie and A. Barnawi, "Evaluation of Pre-Trained CNN Models for Cardiovascular Disease Classification: A Benchmark Study," *Information Sciences Letters*, vol. 12, no. 7, pp. 3317–3338, Jul. 2023, doi: 10.18576/isl/120755.

[11] M. G. Campana, F. Delmastro, and E. Pagani, "Transfer learning for the efficient detection of COVID-19 from smartphone audio data," *Pervasive Mob Comput*, vol. 89, Feb. 2023, doi: 10.1016/j.pmcj.2023.101754.