European Journal of Technique

journal homepage: https://dergipark.org.tr/en/pub/ejt

Vol.15, No.2, 2025

Research Article

Performance Analysis of YAMNet and VGGish Networks for Emotion Recognition from Audio Signals

Yunus KORKMAZ[1*]

[1*]Dicle University, Department of Computer Engineering, Diyarbakir, Turkey. (e-mail: yunus.korkmaz@dicle.edu.tr).

ARTICLE INFO

Received: Aug., 09. 2025
Revised: Sep., 24 2025
Accepted: Oct, 02. 2025

Keywords:
Audio emotion recognition
Transfer learning
Machine learning
Feature embeddings
Fine tuning

Corresponding author: *Yunus KORKMAZ*

ISSN: *2536-5010* | e-ISSN: *2536-5134*

DOI: https://doi.org/10.36222/ejt.1761640

ABSTRACT

Understanding human emotions through vocal cues is a key point for developing emotionally intelligent systems, particularly in fields such as human-computer interaction, healthcare, and virtual assistants. However, accurately recognizing emotions from speech remains a challenging task due to the variability in speaker traits, acoustic conditions, and the subtle, often overlapping nature of emotional states. In this study, a comparative analysis of transfer learning methods for speech emotion recognition (SER) was presented by employing pretrained audio-based neural networks. Specifically, YAMNet and VGGish models were employed both as static feature extractors and in a fine-tuning setup. The extracted embeddings were classified using traditional machine learning algorithms, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Random Forests (RF), and Logistic Regression (LR). Experiments were conducted on two widely used emotional speech datasets: RAVDESS and EmoDB. The results demonstrate that VGGish consistently outperforms YAMNet in both feature extraction and fine-tuning scenarios. The highest classification accuracy was achieved using VGGish features with LR on EmoDB (73.83%). Additionally, fine-tuning VGGish on EmoDB yielded a competitive accuracy of 72.90%. Also class-specific analysis showed that the highest AUC score of 0.9635 was obtained using the LR in VGGish + EmoDB setting, while fine-tuning both YAMNet and VGGish with EmoDB dataset has reached up to Recall score of 1 for the ‘Sadness’ emotion.

1. INTRODUCTION

The integration of emotion recognition from speech has become a pivotal aspect in the development of intelligent systems, particularly within sectors such as healthcare, education, customer engagement, and human-computer communication [1-3]. By enabling machines to infer users' affective states, Speech Emotion Recognition (SER) enhances interaction quality through improved contextual awareness and emotional responsiveness. Unlike textual content alone, vocal expressions convey nuanced acoustic characteristics that offer deep insight into the speaker’s psychological and emotional conditions [4].

Although substantial progress has been made in the field, accurately recognizing emotions from speech remains a complex task influenced by both intrinsic and extrinsic variables. Factors such as speaker diversity (encompassing variations in age, gender, accent, speech tempo, and emotional expressiveness) introduce non-trivial heterogeneity that complicates reliable classification [5,6]. Additionally, the acoustic overlap among different emotional categories and the limited accessibility to large, well-balanced emotional speech datasets present major obstacles to building generalizable SER frameworks. External disturbances, including ambient noise, channel artifacts, and uncontrolled recording environments, further obscure the emotional content embedded in speech signals, thereby impeding effective emotion recognition [7,8].

Recent advancements in deep learning have demonstrated strong potential in overcoming the inherent complexities of SER by enabling the automatic extraction of informative and emotion-relevant features from raw or preprocessed audio signals. Nonetheless, training deep architectures from the ground up typically demands large volumes of annotated data and high computational overhead (both of which pose significant barriers within the SER domain). To address these constraints, transfer learning has gained traction as an effective strategy. This approach leverages pretrained models (initially optimized for large-scale audio-related tasks) to repurpose their learned representations for emotion recognition, thereby reducing the need for extensive labeled data in target SER applications [9,10].

A growing body of research has examined the utility of transfer learning in SER, particularly through the use of convolutional and other deep neural architectures. While the predominant approach involves employing pretrained models as fixed feature extractors, more recent investigations have begun to explore fine-tuning techniques aimed at adapting

model parameters for emotion-specific objectives. Despite these advancements, comprehensive evaluations comparing feature extraction versus fine-tuning paradigms across well-established pretrained audio models remain scarce. Additionally, the performance of conventional machine learning algorithms (when applied to embeddings derived from these models) has yet to be systematically assessed [11-13].

To address these gaps, this study presents a comparative analysis of two widely used pretrained audio models, YAMNet and VGGish [14], in both feature extraction and fine-tuning settings. Their performance is evaluated using multiple classifiers on the RAVDESS [15] and EmoDB [16] datasets without applying class balancing. The study aims to provide insights into the suitability of these models for SER tasks and assess how traditional classifiers perform when coupled with transfer learning-based audio embeddings. A detailed evaluation using standard classification metrics is conducted to draw comprehensive conclusions about the effectiveness of each approach.

Although transfer learning has demonstrated potential in speech-related tasks, limited attention has been given to a detailed comparison between static feature extraction and fine-tuning strategies using pretrained audio models specifically for SER. Additionally, while end-to-end deep learning approaches dominate recent literature, the role of traditional classifiers, when coupled with high-level audio embeddings, remains underexplored. This study contributes to the literature by not only benchmarking two prominent audio-based neural networks (YAMNet and VGGish) across different learning paradigms, but also by examining the interplay between these embeddings and conventional classifiers under consistent experimental conditions. The findings aim to offer practical insights into optimizing SER pipelines, especially in resource-constrained environments where full-scale end-to-end training is not feasible.

The remainder of the paper is organized as follows: Section 2 reviews related work on SER and transfer learning approaches. Section 3 outlines the structure of the proposed SER system, including details of the pretrained models, datasets, and classifiers used. Section 4 presents the experimental results, followed by an in-depth discussion in Section 5. Finally, Section 6 concludes the study and outlines potential directions for future research.

2. LITERATURE REVIEW

Speech Emotion Recognition (SER) has emerged as a critical subfield within affective computing, focusing on enabling machines to interpret human emotional states from speech signals [17,18]. Emotions in speech can be inferred from prosodic, spectral, and temporal characteristics, reflecting changes in pitch, loudness, energy, and articulation [19,20]. Traditionally, SER systems have been employed in various applications such as intelligent virtual assistants, mental health monitoring, customer service automation, and e-learning platforms, where emotionally aware responses enhance human-machine interactions [21,22]. However, accurately detecting emotions remains a technically complex problem due to the variability in how emotions are expressed across different individuals, languages, and contexts [23,24].

In early SER studies, feature engineering played a central role. Handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), zero-crossing rate, energy, pitch contours, jitter, and shimmer were extracted from speech signals. These features were typically input to classical classifiers such as Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Hidden Markov Models (HMM), and Gaussian Mixture Models (GMM). Although these models offered valuable baselines, their performance was highly dependent on the selection and preprocessing of features, making them less generalizable across datasets and languages. Furthermore, they often struggled to model the sequential nature and hierarchical structure of speech signals [25-28].

The emergence of deep learning introduced significant improvements in SER systems. Convolutional Neural Networks (CNNs) became popular for extracting local patterns from spectrograms or Log-Mel representations, while Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) architectures, showed strong performance in capturing the temporal dependencies of emotional speech [29]. These models eliminated the need for manual feature selection by learning task-specific representations directly from the data. However, deep architectures are typically data-hungry and computationally expensive, often requiring large annotated datasets for effective training-resources that are scarce in the emotion recognition domain.

To solve these limitations, transfer learning has been widely adopted as a practical approach for SER. Transfer learning leverages models that are pretrained on large-scale general-purpose audio datasets and repurposes them for emotion classification tasks [30]. This strategy allows SER systems to benefit from rich, generalizable feature representations without the need for training from scratch. Pretrained models such as VGGish and YAMNet have gained popularity in this context. These models, trained on datasets like AudioSet, can capture complex acoustic features across a broad range of sound events, making them useful for various downstream tasks, including SER. YAMNet, based on a MobileNet architecture, is pretrained on AudioSet and outputs embeddings that summarize high-level acoustic properties. It has been used in applications such as environmental sound recognition, audio event detection, and affective computing. VGGish, a variant of the VGG network adapted for audio, also extracts embeddings from Log-Mel spectrogram inputs and has been frequently employed in tasks requiring robust and transferable acoustic features. Both models support two transfer learning strategies: using their embeddings as fixed features (feature extraction) and fine-tuning their internal layers for adaptation to the target domain.

Another notable gap in existing literature is the lack of multi-dataset evaluations. Many studies validate their models on a single dataset, which limits the generalizability and practical applicability of their findings. The impact of data distribution, class imbalance, and domain shift across datasets is rarely addressed. In addition, few works analyze how different pretrained networks perform across both static and fine-tuned settings on multiple datasets using consistent evaluation metrics, such as accuracy, F1-score, precision, recall, ROC curves, and AUC scores. Given these limitations, a comprehensive evaluation that jointly investigates the performance of multiple pretrained audio models, both as feature extractors and under fine-tuning, across diverse emotional speech corpora is necessary. The current study addresses this need by comparing YAMNet and VGGish under both strategies, using two well-known datasets,

RAVDESS and EmoDB, without applying class balancing. Four traditional classifiers are used to explore the discriminative capacity of extracted embeddings, and detailed evaluations are conducted using a wide set of metrics. This study thus contributes to filling a critical gap in the literature by offering systematic, reproducible, and practical insights for building efficient SER systems using transfer learning.

3. SPEECH EMOTION RECOGNITION SYSTEM

This section presents the architecture and components of the Speech Emotion Recognition (SER) framework based on transfer learning. The system consists of several stages like preprocessing and dataset preparation, feature extraction using pretrained audio models, classification using conventional machine learning algorithms, and fine-tuning of pretrained networks. In this paper, pre-trained audio networks were used in two main transfer learning aspects which are using networks as feature extractor and fine-tuning the networks directly on two different datasets

Both aspects of transfer learning were performed to understand their relative performances in recognizing speech emotion using digital audio signals. The following sections provide a detailed explanation of the dataset used, the methods applied, experimental setup, and the comparative analysis of different classification techniques. An overview of the general pipeline was illustrated in Figure 1.

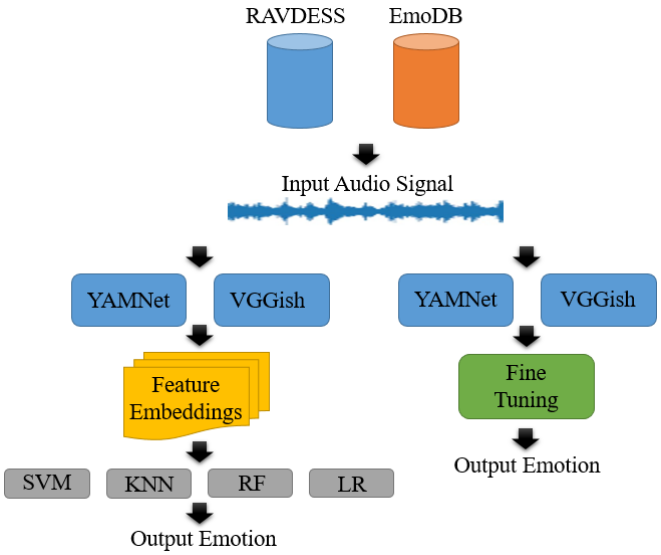

Figure 1. Pipeline of speech emotion recognition system

Two publicly available emotional speech datasets were utilized in this study: RAVDESS and EmoDB as given by Table 1. The RAVDESS dataset contains recordings from 24 professional actors vocalizing two sentences in eight emotional classes as its details were shown by Table 2. For the feature embedding phase of this study, because the neutral emotion class originally had 96 samples (different than other classes), it was excluded (downsapmling process) from RAVDESS, and a total of 1344 samples (192 per each of remaining class) were used, resulting in a balanced subset. On the other hand, the EmoDB dataset includes 535 recordings from 10 actors in seven emotional categories, with an inherently imbalanced distribution. In EmoDB, number of samples in each class of 7 classes were different as given by Table 3. So, any balancing strategy was not applied to EmoDB. All audio files have 16 kHz sampling frequency (Fs), they all converted to mono, and normalized for amplitude consistency. No augmentation was applied in order to evaluate models in a more realistic and unaltered setting.

TABLE I
THE DATASETS USED IN THIS STUDY

Dataset	Lang.	Speakers	Classes	Samples	Fs
RAVDESS	Eng.	24	7	1344	16KHz
EmoDB	Germ.	10	7	535	16KHz

TABLE II
STRUCTURE OF THE ORIGINAL RAVDESS DATASET

Emotion	Samples
Suprised	192
Disgust	192
Angry	192
Sad	192
Calm	192
Fearful	192
Happy	192
Neutral	96

TABLE III
STRUCTURE OF THE ORIGINAL EMODB DATASET

Emotion	Samples
Anger	127
Boredom	81
Neutral	79
Happiness	71
Fear	69
Sadness	62
Disgust	46

In the first experimental setup, feature extraction was performed using two pretrained models: YAMNet and VGGish. These models were used as static encoders, producing embeddings for each input audio sample without any internal weight updates. YAMNet which is based on the MobileNet-v1 architecture takes waveform audio inputs and outputs 1024-dimensional embeddings. YAMNet network architecture was illustrated by Figure 2.

When using YAMNet, each audio clip was segmented into overlapping frames of 0.975 seconds with a hop size of 0.48 seconds. Embeddings from all frames were averaged to form a single feature vector per sample. On the other side, VGGish which is adapted from the VGG architecture operates on Log-Mel spectrogram patches (96x64) and produces 128-dimensional embeddings. Similar to YAMNet, average pooling was applied to derive fixed-length representations. VGGish network architecture was shown in Figure 3.

The mathematical representation of the embedding extraction process can be defined with formula shown by Equation 1.

$$e_i = \frac{1}{T}\sum_{t=1}^{T} f_\theta\left(x_{i,t}\right) \quad (1)$$

where $x_{i,t}$ represents the t-th segment of the i-th audio input, $f_\theta(\cdot)$ denotes the pretrained model and e_i is the final aggregated embedding vector.

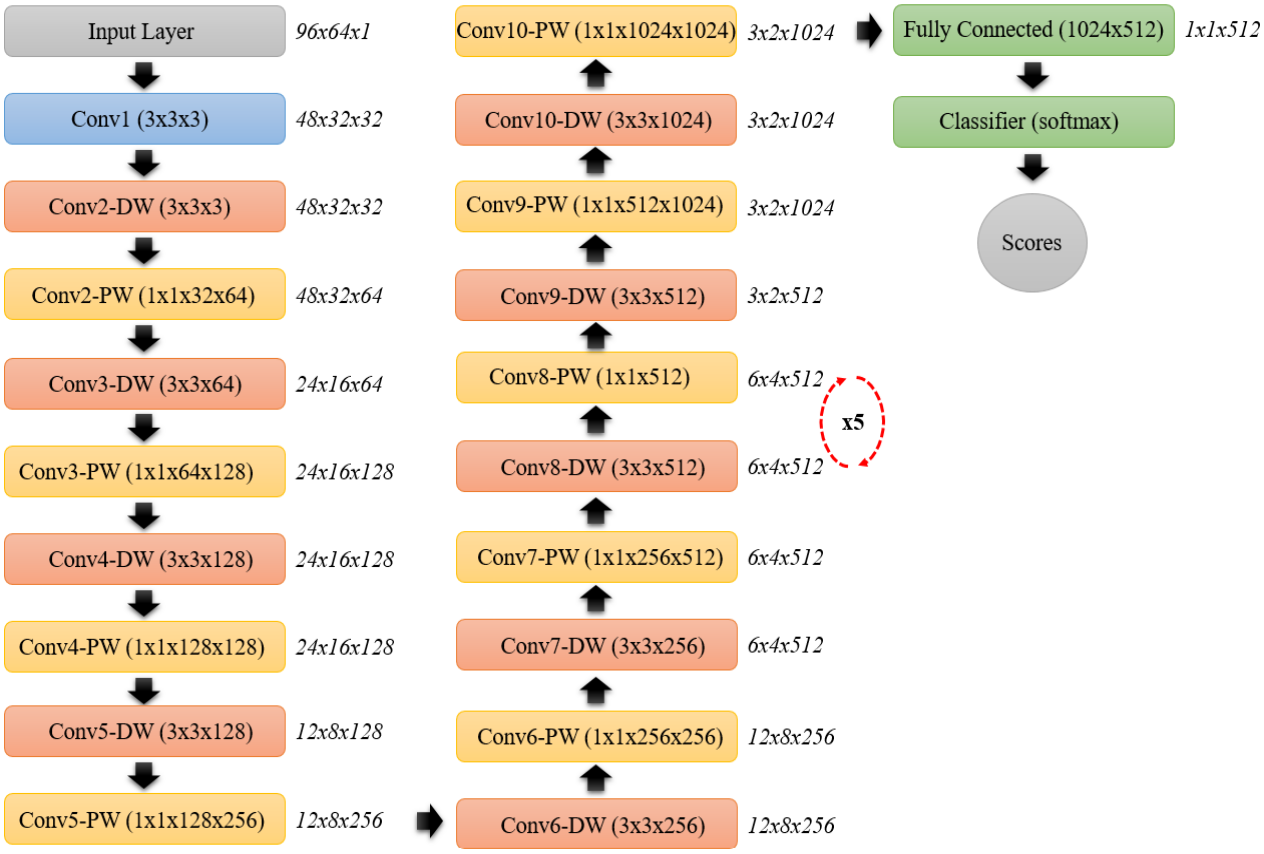

Figure 2. Original YAMNet architecture

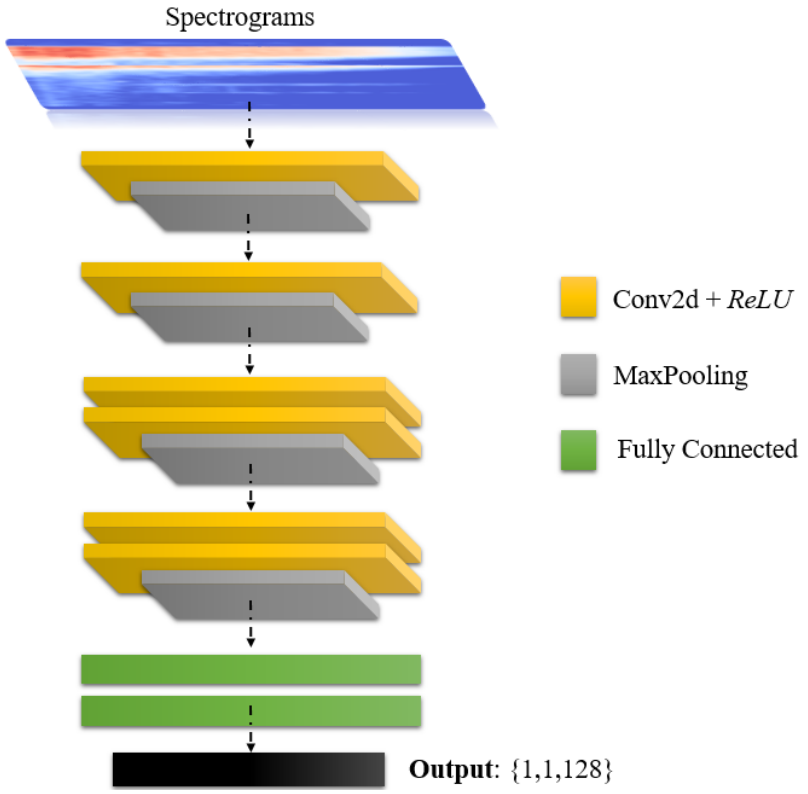

Figure 3. VGGish network architecture

Once the embeddings were obtained, they were classified using four standard supervised learning algorithms which are Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Random Forest (RF) and Logistic Regression (LR). SVM is supervised learning algorithm that constructs an optimal hyperplane to separate data into distinct classes by maximizing the margin between support vectors of different classes [31] while KNN is a non-parametric classification method that assigns a class label based on the majority vote of the k closest training examples in the feature space [32]. On another side, RF is an ensemble learning algorithm that builds multiple decision trees during training and outputs the mode of their predictions to enhance generalization and reduce overfitting [33], and LR is a linear model used for binary and multiclass classification that estimates the probability of class membership using the logistic (sigmoid) function [34]. The settings that applied to mentioned classifiers after feature embedding extraction process were given by Table 4.

TABLE IV
HYPER PARAMETERS USED IN CLASSIFIERS OF SVM, KNN, RF AND LR

Classifier	Parameter	Value
SVM	kernel	'linear'
KNN	# of neighbors	5
RF	# of estimators	100
LR	Max Iteration	1000
All of 4	Train-Test rate	80%-20%

In the second experimental setup, both YAMNet and VGGish were fine-tuned end-to-end using the emotional speech datasets. A classification head consisting of a fully connected layer and softmax activation was appended to the final feature layer of the pretrained networks. The models were trained using categorical cross-entropy loss as formulated by Equation 2.

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \qquad (2)$$

where $y_{i,c}$ is the ground truth and $\hat{y}_{i,c}$ is the predicted probability for class c.

In the training process when using audio networks as feature extractor, training was performed for 50 epochs with a learning rate of 1e-4 using the Adam optimizer. Early stopping based on validation loss was applied to prevent overfitting. In fine-tuning phase of this work, train-test rate for training the network was choosen as 80%-20% for both YAMNet and VGGish. Moreover, number of epochs was assigned to 100 in fine-tuning process. In Classification procedure using extracted embeddings was given by Algorithm 1.

ALGORITHM I
EMBEDDING EXTRACTION ALGORITHM USED IN CLASSIFICATION

Input: E : Extracted embeddings from pretrained model
y : Corresponding class labels

Output: $\hat{y}$: Predicted class labels

Step 1. Split E and y into training and testing sets

Step 2. For each classifier $C \in \{SVM, KNN, RF, LR\}$
Train C on the training set
Predict $\hat{y}$ on the testing set

Step 3. Evaluate predictions using Acc., F1 Sc., AUC Sc. and Conf. Matrix

Algorithm 1 illustrates the feature embedding extraction and classification procedure employed in this study. Specifically, the embeddings obtained from the pretrained YAMNet and VGGish models are split into training and testing sets, classified using machine learning algorithms, and finally evaluated with standard performance metrics.

4. RESULTS

This section presents the experimental outcomes derived from the two distinct phases of the transfer learning approach employed in this study for speech emotion recognition. In the phase of using pre-trained audio neural networks as feature extractor, audio embeddings obtained from pretrained models were used as input to traditional machine learning classifiers. In phase of fine-tuning, the pretrained models themselves were retrained directly on the target datasets. For the first phase, the classification performance of four well-known algorithms - Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Random Forest (RF), and Logistic Regression (LR) - was evaluated using feature embeddings extracted from YAMNet and VGGish models. For the second phase, the fine-tuning process involved retraining YAMNet and VGGish models on the RAVDESS and EmoDB datasets separately. For both approaches, recognition performance was assessed and compared in terms of classification accuracy, F1 score and ROC curves together with AUC scores. Table 5 shows the classification accuracy performance of all classifiers over RAVDESS dataset when using audio networks as feature extractor.

TABLE V
RESULTS OF CLASSIFIERS WITH RAVDESS DATASET BY USING EMBEDDINGS FROM AUDIO NETWORKS

Audio Network	Classifier	Accuracy (%)
YAMNet	SVM	60.59
	KNN	40.15
	RF	50.56
	LR	62.45
VGGish	SVM	63.57
	KNN	49.44
	RF	52.42
	LR	63.20

When using RAVDESS, the best accuracy was reached with SVM classifier fed by feature embeddings from VGGish networks. It will be useful to remind that these results are raw results which means no data augmentation and no hyper parameter tuning processes were done over classification phase. Only 'neutral' was completely removed from dataset (7 classes left) in order to prevent any obvious imbalancing issue. Table 6 shows the classification accuracy performance of all classifiers over EmoDB dataset when using audio networks as feature extractor, again with no augmentation, hyper parameter tuning. Because all classes have different number of samples in EmoDB, balancing or removing any of classes was not prefered here as balancing strategy. Additionally, in an independent test, number of samples in each class of EmoDB was set to 46 in order to balance dataset considering the class having the least number of samples ('Disgust' class), but training process was cancelled after obtaining the worse results comparing to unbalanced EmoDB version.

To evaluate the classification performance of the proposed system in the feature extraction setting, F1 scores were calculated for each combination of pretrained model and dataset. Specifically, embeddings extracted from YAMNet and VGGish were used as input features to four traditional classifiers-SVM, KNN, Random Forest, and Logistic Regression-trained and tested on the RAVDESS and EmoDB datasets. The F1 score, which considers both precision and recall, provides a balanced measure of performance, particularly useful in multi-class classification problems with potential class imbalance. The results, summarized in Table 7, reveal comparative insights into the effectiveness of each classifier under different feature-model configurations.

TABLE VI
RESULTS OF CLASSIFIERS WITH EMODB DATASET BY USING EMBEDDINGS FROM AUDIO NETWORKS

Audio Network	Classifier	Accuracy (%)
YAMNet	SVM	68.22
	KNN	56.07
	RF	60.75
	LR	71.96
VGGish	SVM	69.16
	KNN	59.81
	RF	72.90
	LR	73.83

TABLE VII
F1 SCORES (%) OF CLASSIFIERS USING EXTRACTED FEATURE EMBEDDINGS

Classifier	YAMNet + RAVDESS	YAMNet + EmoDB	VGGish + RAVDESS	VGGish + EmoDB
SVM	60.40	67.41	63.46	68.62
KNN	39.51	53.90	48.05	57.13
RF	49.47	57.51	51.57	71.32
LR	62.20	71.21	62.88	72.99

.

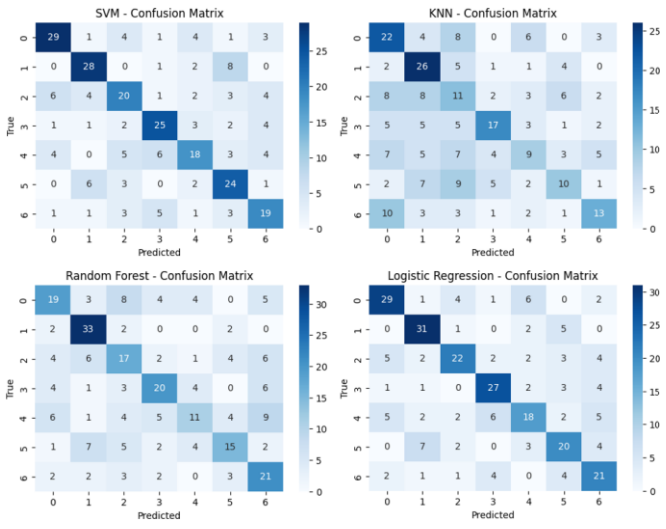

Figure 4. Confusion matrix of each classifier in YAMNet with RAVDESS setting

To provide a more detailed view of class-wise prediction performance, confusion matrices were generated for each classifier under the four configurations which are YAMNet with RAVDESS, YAMNet with EmoDB, VGGish with RAVDESS, and VGGish with EmoDB. These matrices offer insights into how well each emotional class was recognized and where misclassifications occurred, thus complementing the overall evaluation metrics such as accuracy and F1 score. The confusion matrices for each of YAMNet with RAVDESS, YAMNet with EmoDB, VGGish with RAVDESS, and VGGish with EmoDB settings were presented in the Figure 4, 5, 6 and 7, respectively

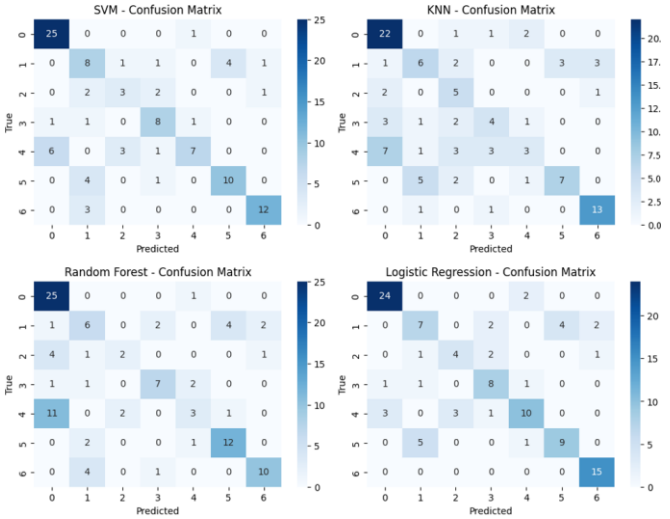

Figure 5. Confusion matrix of each classifier in YAMNet with EmoDB setting

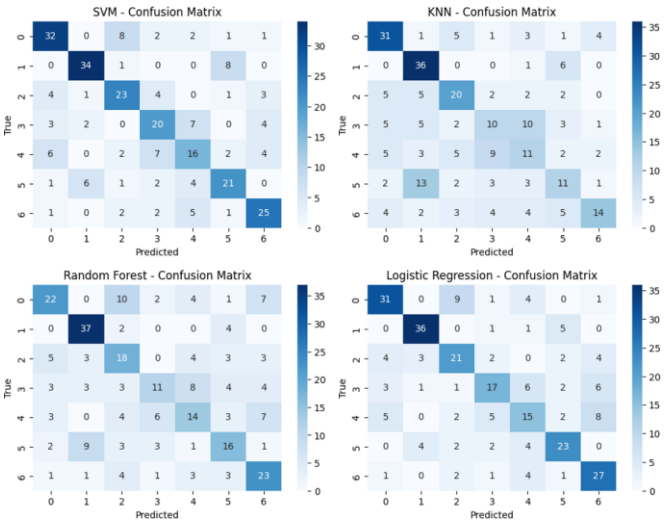

Figure 6. Confusion matrix of each classifier in VGGish with RAVDESS setting

Receiver Operating Characteristic (ROC) curves and corresponding Area Under the Curve (AUC) scores were utilized to further assess the discriminative power of the classifiers. ROC curves illustrate the trade-off between true positive and false positive rates across different classification thresholds, providing valuable insight into classifier behavior beyond accuracy or F1 score. AUC values, on the other hand, offer a single scalar metric reflecting the overall capability of the model to distinguish between classes. Figure 8 and Figure 9 present ROC curves for all classifiers in both YAMNet and VGGish configurations, respectively.

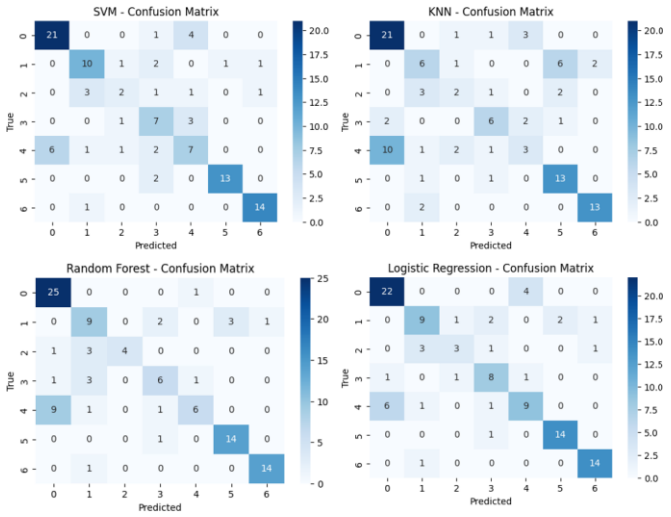

Figure 7. Confusion matrix of each classifier in VGGish with EmoDB setting

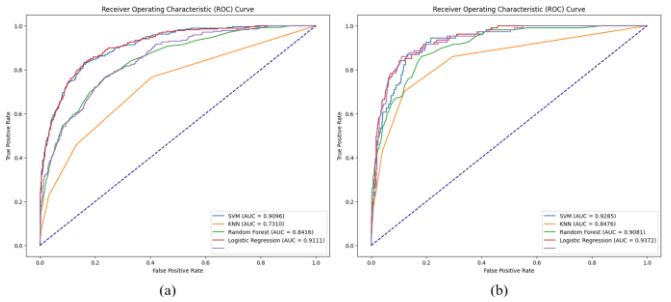

Figure 8. ROC curves of classifiers for YAMNet model with (a) RAVDESS dataset and (b) EmoDB dataset

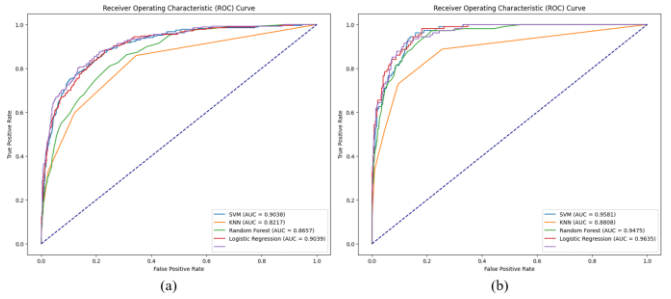

Figure 9. ROC curves of classifiers for VGGish model with (a) RAVDESS dataset and (b) EmoDB dataset

AUC scores provide the clear measure of a classifier's ability to distinguish between classes, independent of classification thresholds. Higher AUC values indicate better overall separability and are particularly useful in evaluating performance under class imbalance conditions. Table 8 presents the AUC scores obtained for each classifier across all model-dataset configurations (YAMNet with RAVDESS, YAMNet with EmoDB, VGGish with RAVDESS, and VGGish with EmoDB).

In addition to the feature extraction strategy, fine-tuning experiments were conducted to evaluate the classification performance of the pretrained audio networks when retrained on the RAVDESS and EmoDB datasets. The fine-tuned VGGish model achieved the highest accuracy of 72.90% on the EmoDB dataset, followed closely by the YAMNet model with 67.29%. On the RAVDESS dataset, the VGGish model again outperformed YAMNet, yielding an accuracy of 56.94% compared to 40.28%. These results indicate that while both pretrained models benefit from fine-tuning, VGGish demonstrates a better adaptability to emotional speech data, particularly when applied to the EmoDB corpus.

TABLE VIII
AUC SCORES OF CLASSIFIERS USING EXTRACTED FEATURE EMBEDDINGS

Classifier	YAMNet + RAVDESS	YAMNet + EmoDB	VGGish + RAVDESS	VGGish + EmoDB
SVM	0.9096	0.9285	0.9038	0.9581
KNN	0.7310	0.8476	0.8217	0.8808
RF	0.8416	0.9081	0.8657	0.9475
LR	0.9111	0.9372	0.9039	0.9635

To gain deeper insight into the classification behavior of the fine-tuned models, confusion matrices and ROC curves were analyzed for all experimental configurations. The confusion matrices, presented in Figure 10, reveal the distribution of correct and incorrect predictions across emotion classes, highlighting the strengths and weaknesses of each model-dataset pair. Complementarily, Figure 11 displays the ROC curves, illustrating the true positive and false positive trade-offs for each class. These visualizations provide a more nuanced understanding of how each fine-tuned model performs under varying conditions and contribute to the interpretability of the overall classification performance.

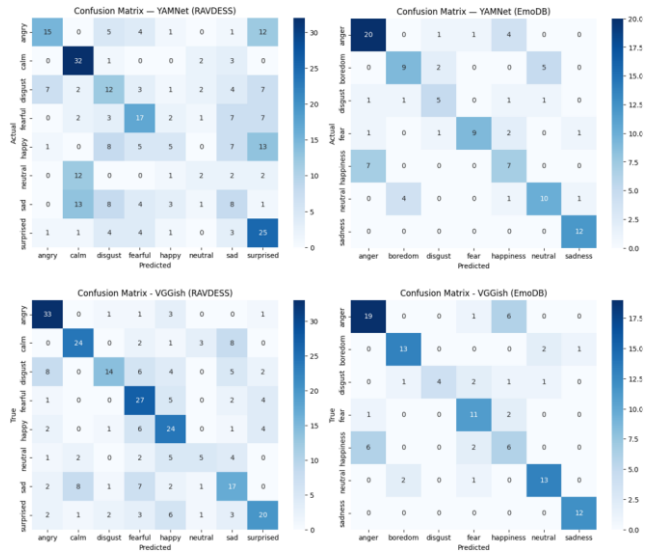

Figure 10. Confusion matrices of YAMNet and VGGish used with RAVDESS and EmoDB in fine tuning scenario

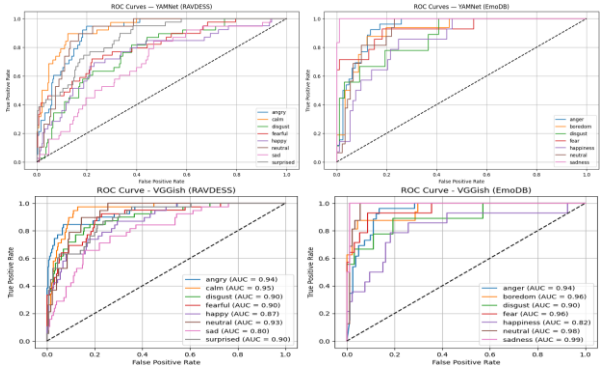

Figure 11. ROC curves and AUC scores of YAMNet and VGGish used with datasets in fine tuning scenario

The experimental results presented in this section demonstrate the usage of transfer learning for speech emotion recognition, considering both feature extraction and fine-tuning approaches. Feature embeddings derived from pretrained models, YAMNet and VGGish, were classified using four traditional machine learning algorithms, and performance metrics such as accuracy, F1 score, confusion matrices, ROC curves, and AUC scores were thoroughly analyzed. In parallel, fine-tuning the pretrained models directly on the RAVDESS and EmoDB datasets provided further insights into their adaptability and generalization capabilities. Overall, the results reveal that model performance varies depending on the dataset and the method used, offering valuable perspectives for future research and application development in emotion recognition systems.

TABLE IX
CLASS-WISE PRECISION(P), RECALL(R) AND F1 SCORE(F) FOR FINE-TUNING OF YAMNET(Y) AND VGGISH(V) WITH RAVDESS DATASET

Emotion	P(Y)	R(Y)	F(Y)	P(V)	R(V)	F(V)
Angry	0.62	0.39	0.48	0.67	0.85	0.75
Calm	0.52	0.84	0.64	0.69	0.63	0.66
Disgust	0.29	0.32	0.30	0.74	0.36	0.48
Fearful	0.46	0.44	0.45	0.50	0.69	0.58
Happy	0.36	0.13	0.19	0.48	0.63	0.55
Neutral	0.25	0.11	0.15	0.50	0.26	0.34
Sad	0.23	0.21	0.22	0.42	0.45	0.44
Suprised	0.37	0.64	0.47	0.65	0.53	0.58

TABLE X
CLASS-WISE PRECISION(P), RECALL(R) AND F1 SCORE(F) FOR FINE-TUNING OF YAMNET(Y) AND VGGISH(V) WITH EMODB DATASET

Emotion	P(Y)	R(Y)	F(Y)	P(V)	R(V)	F(V)
Anger	0.69	0.77	0.73	0.73	0.73	0.73
Boredom	0.54	0.56	0.60	0.81	0.81	0.81
Disgust	0.56	0.56	0.56	1.00	0.44	0.62
Fear	0.90	0.64	0.75	0.65	0.79	0.71
Happiness	0.47	0.50	0.48	0.40	0.43	0.41
Neutral	0.62	0.62	0.62	0.81	0.81	0.81
Sadness	0.86	1.00	0.92	0.92	1.00	0.96

The experimental results presented in this section demonstrate the usage of transfer learning for speech emotion recognition, considering both feature extraction and fine-tuning approaches. Feature embeddings derived from pretrained models, YAMNet and VGGish, were classified using four traditional machine learning algorithms, and performance metrics such as accuracy, F1 score, confusion matrices, ROC curves, and AUC scores were thoroughly analyzed. In parallel, fine-tuning the pretrained models directly on the RAVDESS and EmoDB datasets provided further insights into their adaptability and generalization capabilities. Overall, the results reveal that model performance varies depending on the dataset and the method used, offering valuable perspectives for future research and application development in emotion recognition systems.

5. DISCUSSION

This section provides a comprehensive analysis and interpretation of the results obtained through two distinct transfer learning strategies which are utilizing pretrained audio networks (YAMNet and VGGish) as feature extractors, fine-tuning the same models directly on two benchmark speech emotion datasets, RAVDESS and EmoDB. Rather than introducing a novel model, this study emphasizes a comparative exploration of the impact of pretrained feature representations versus model fine-tuning on classification performance. Accordingly, classification metrics including accuracy, F1 score, AUC, and confusion matrices were examined to evaluate and contrast the effectiveness of both approaches. This discussion aims to highlight meaningful patterns, dataset-specific behaviors, and algorithmic tendencies that emerged throughout the experiments.

As presented in Table 5, when feature embeddings were extracted from YAMNet and used with classical classifiers on the RAVDESS dataset, LR achieved the highest accuracy with 62.45%, followed by SVM with 60.59%. KNN, in contrast, showed the weakest performance at 40.15%. When using VGGish embeddings, performance generally improved across classifiers. In particular, SVM and LR yielded competitive accuracies of 63.57% and 63.20%, respectively, while KNN again lagged with 49.44%. The evaluation on the EmoDB dataset, shown in Table 6, further reinforces the superior performance of VGGish over YAMNet in feature extraction tasks. LR achieved the highest classification accuracy of 73.83% using VGGish embeddings, followed closely by RF with 72.90%. YAMNet also showed reasonable effectiveness with LR, obtaining 71.96% accuracy. These results clearly indicate that the EmoDB dataset was better classified overall, with higher accuracy scores across all classifiers compared to RAVDESS. The overall trend observed in both datasets suggests that LR consistently outperformed the other classifiers, particularly when paired with VGGish features. This highlights the effectiveness of simpler, linear models when combined with rich pretrained embeddings, especially in tasks involving compact emotional speech datasets.

The F1 scores reported in Table 7 provide further insight into the balance between precision and recall for each classifier across different configurations. For the RAVDESS dataset, the highest F1 score was achieved by Logistic Regression (62.20%) when using YAMNet features, closely followed by LR with VGGish embeddings (62.88%). In contrast, KNN performed poorly, with F1 scores of 39.51% (YAMNet) and 48.05% (VGGish), confirming its underperformance also seen in the accuracy results. A more distinct improvement was observed on the EmoDB dataset. Here, the combination of VGGish features and LR yielded the highest F1 score of 72.99%, with RF also performing robustly at 71.32%. YAMNet embeddings also supported decent classifier performance, particularly with LR (71.21%) and SVM (67.41%). Again, KNN showed the lowest F1 scores across both audio networks, reinforcing its relative inefficiency in this context. The results in Table 7 support the earlier accuracy findings and further validate the reliability of VGGish feature embeddings, especially when used with LR and RF classifiers. Moreover, the strong F1 scores achieved with EmoDB suggest that emotional expression in the German-language dataset is more distinguishable for the models than in the English-language RAVDESS dataset.

The Area Under the Curve (AUC) metric is a crucial indicator of a model's capability to distinguish between different emotional classes. As illustrated in Table 8, the highest AUC scores were consistently obtained by Logistic Regression across all configurations, particularly when paired with VGGish embeddings and the EmoDB dataset, reaching an

outstanding value of 0.9635. This highlights the classifier's robust performance in correctly ranking positive and negative examples, even in multiclass scenarios. For the RAVDESS dataset, LR with YAMNet (0.9111) and VGGish (0.9039) also achieved high AUC values, confirming the consistency observed in the earlier accuracy and F1 score metrics. The Random Forest classifier followed closely, especially in the VGGish + EmoDB setting with an AUC of 0.9475. SVM classifiers also maintained relatively high AUC values, although slightly lower than LR, particularly in the VGGish + RAVDESS case (0.9038). Conversely, KNN yielded the lowest AUC scores in all scenarios, further confirming its limited discriminative ability compared to the other classifiers. The performance gap was especially notable with the YAMNet + RAVDESS combination (0.7310). These results reinforce the conclusion that VGGish features, particularly when used with LR and RF classifiers, enable more effective emotion classification with higher confidence across multiple metrics.

To gain a more granular understanding of model performance in the fine-tuning approach, class-wise precision, recall, and F1 scores were analyzed for each emotional category using both YAMNet and VGGish models, as presented in Table 9. These metrics reveal significant differences in the models' ability to distinguish between specific emotions. When fine-tuned on RAVDESS, VGGish outperformed YAMNet across almost all emotional classes. For example, in recognizing the Angry emotion, VGGish achieved an F1 score of 0.75, markedly higher than YAMNet's 0.48. Similarly, for Disgust, VGGish maintained a high precision of 0.74, while YAMNet lagged behind with a precision of only 0.29. Emotions such as Neutral and Sad proved challenging for both models, but VGGish still showed a better balance, with F1 scores of 0.34 and 0.44, respectively. A closer look reveals that YAMNet tended to achieve higher recall in some cases, such as Calm (0.84), yet this often came at the cost of low precision, resulting in moderate overall F1 scores. On the contrary, VGGish provided a more balanced trade-off between precision and recall, which contributed to more stable and higher F1 scores across most categories. These results further underscore the robustness of the VGGish architecture when directly fine-tuned on emotional speech data, suggesting that it captures more generalizable and discriminative features compared to YAMNet, particularly for complex emotional categories.

The fine-tuning results on the EmoDB dataset revealed compelling differences in emotional recognition performance between YAMNet and VGGish, as detailed in Table 10. Overall, both models performed strongly across most emotions, but VGGish consistently achieved higher F1 scores, highlighting its superior generalization in emotion classification when fine-tuned on a smaller, high-quality dataset like EmoDB. For instance, VGGish demonstrated robust and consistent recognition for Neutral and Boredom emotions, each with an F1 score of 0.81. It also achieved the highest F1 score of 0.96 for Sadness, matching YAMNet's perfect recall of 1.00 while maintaining high precision. Despite YAMNet performing notably well on Fear (F1 = 0.75) and Sadness (F1 = 0.92), its performance on other emotions such as Happiness and Boredom was lower compared to VGGish. Interestingly, while YAMNet achieved higher recall for several classes, its precision was generally lower, leading to lower F1 scores in cases like Disgust and Happiness. Conversely, VGGish maintained a better balance between precision and recall, even though it occasionally sacrificed recall (e.g., Disgust recall = 0.44), still resulting in competitive F1 scores due to very high precision. These findings confirm that VGGish outperforms YAMNet in fine-tuning scenarios for both datasets, but especially for EmoDB, where emotion expressions are more distinct. The class-wise evaluation helps highlight which emotions are better captured and where further model refinement or data augmentation may be needed.

The confusion matrix of the VGGish + Logistic Regression model with the EmoDB dataset (accuracy: 73.83%) was illustrated in Figure 7. As shown, the model effectively distinguishes the Anger (22 correct predictions) and Sadness (14 correct predictions) classes with high accuracy. However, moderate confusion was observed between some emotion pairs, for instance, Fear and Neutral, and Happiness and Anger. These misclassifications highlight the challenge of overlapping acoustic characteristics among certain emotional expressions. Nonetheless, the overall distribution reflects a robust and consistent performance, further validating the potential of VGGish embeddings when coupled with Logistic Regression for speech emotion recognition tasks.

6. CONCLUSION & FUTURE DIRECTIONS

This study conducted a comparative analysis of speech emotion recognition (SER) using pretrained audio neural networks, focusing on both feature extraction and fine-tuning approaches. YAMNet and VGGish were employed as the backbone models, and their effectiveness was evaluated using two widely-used emotional speech datasets: RAVDESS and EmoDB. In the feature extraction phase, embeddings were obtained from both networks and classified using traditional machine learning algorithms, including SVM, KNN, Random Forest, and Logistic Regression. Among all configurations, the highest accuracy of 73.83% was achieved with VGGish embeddings and Logistic Regression on the EmoDB dataset, highlighting the synergy between rich audio representations and linear classifiers. In addition to overall accuracy, a comprehensive set of evaluation metrics, including F1 score, AUC, and confusion matrices, was used to examine the models' behavior in depth. The fine-tuning phase further explored the potential of end-to-end learning by retraining YAMNet and VGGish directly on the target datasets. Results showed that fine-tuning VGGish led to substantial improvements in class-wise performance, particularly in emotional categories such as Angry, Fearful, and Sad. This indicates that while feature extraction offers strong baselines, fine-tuning can refine representations in emotion-specific contexts. For future work, expanding the analysis to include additional pretrained models (e.g., TRILL, OpenL3) may offer further insights into the robustness of various architectures in SER tasks. Moreover, experimenting with larger and more diverse emotional speech datasets can help evaluate the generalizability of these models across languages, cultures, and acoustic conditions. Finally, exploring hybrid models that combine handcrafted features with deep embeddings could serve as a promising direction to enhance recognition performance in real-world applications.

REFERENCES

[1] Singh, Y.B., & Goel, S. (2022). A systematic literature review of speech emotion recognition approaches. Neurocomputing, 492, pp. 245-263.

[2] Llruba, C., & Palau, R. (2024). Real-Time Emotion Recognition for Improving the Teaching–Learning Process: A Scoping Review. Journal of Imaging, 21, 10(12), 313.
[3] Lope, J.d., & Grana, M. (2023). An ongoing review of speech emotion recognition. Neurocomputing, 528, pp. 1-11.
[4] Alhussein, G. et al. (2025). Speech emotion recognition in conversations using artificial intelligence: a systematic review and meta-analysis. Artificial Intelligence Review, Vol. 58, Article No: 198.
[5] Chakhtouna, A. et al. (2023). Speaker and gender dependencies in within/cross linguistic Speech Emotion Recognition. International Journal of Speech Technology, Vol. 26, pp. 609–625.
[6] Foggia, P. et al. (2024). Identity, Gender, Age, and Emotion Recognition from Speaker Voice with Multi-task Deep Networks for Cognitive Robotics. Cognitive Computation, Vol. 16, pp. 2713–2723.
[7] Kakuba, S., & Han, D.S. (2025). Addressing data scarcity in speech emotion recognition: A comprehensive review. ICT Express, 11(1), pp. 110-123.
[8] He, Z. (2025). Research Advanced in Speech Emotion Recognition based on Deep Learning. Theoretical and Natural Science, 86, pp. 45-52.
[9] Nguyen, D. et al. (2023). Meta-transfer learning for emotion recognition. Neural Computing and Applications, Vol. 35, pp. 10535–10549.
[10] Padi, S. et al. (2021). Improved Speech Emotion Recognition using Transfer Learning and Spectrogram Augmentation. arXiv:2108.02510.
[11] Phukan, O.C. et al. (2023). A Comparative Study of Pre-trained Speech and Audio Embeddings for Speech Emotion Recognition. arXiv:2304.11472.
[12] Liu, K. et al. (2024). Improving Pre-Trained Model-Based Speech Emotion Recognition From a Low-Level Speech Feature Perspective. IEEE Transactions on Multimedia, Vol. 26, pp. 10623-10636.
[13] Hassan, A. et al. (2024). Benchmarking Pretrained Models for Speech Emotion Recognition: A Focus on Xception. Computers, 13(12), 315.
[14] Hershey, S. et al. (2017). CNN architectures for large-scale audio classification. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131-135.
[15] Livingstone, S.R., & Russo, F.A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. Plos One, 13(5).
[16] Burkhardt, F. et al. (2005). A database of German emotional speech. Proc. Interspeech, pp. 1517-1520, doi: 10.21437/Interspeech.2005-446.
[17] Schuller, B. et al. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, 53(9-10), pp. 1062-1087.
[18] Ayadi, M.A. et al. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3), pp. 572-587.
[19] Lee, C.M. et al. (2005). Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing, 13(2), pp. 293-303.
[20] Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. Speech Communication, 48(9), pp. 1162-1181.
[21] Picard, R.W. (2003). Affective computing: challenges. International Journal of Human-Computer Studies, 59(1-2), pp. 55-64.
[22] Tao, J., & Tan, T. (2005). Affective Computing: A Review. Affective Computing and Intelligent Interaction, 3784.
[23] Schuller, B. et al. (2013). Computational paralinguistics challenge: social signals, conflict, emotion, autism. Proc. Interspeech, pp. 148-152.
[24] Burkhardt, F., & Sendlmeier, W.F. (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. Proc. ITRW on Speech and Emotion, pp. 151-156.
[25] Madanian, S. et al. (2023). Speech emotion recognition using machine learning - A systematic review. Intelligent Systems with Applications, 20.
[26] Feng, K., & Chaspari, T. (2020). A Review of Generalizable Transfer Learning in Automatic Emotion Recognition. Frontiers in Computer Science, 20.
[27] Belkacem, S. (2023). Speech Emotion Recognition: Recent Advances and Current Trends. 22nd International Conference on Artificial Intelligence and Soft Computing: (ICAISC), Proceedings, Part II.
[28] George, S.M., & Ilyas, P.M. (2024). A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. Neurocomputing, 568.
[29] Sonmez, Y.U., & Varol, A. (2019). New Trends in Speech Emotion Recognition. 7th International Symposium on Digital Forensics and Security (ISDFS), pp. 1-7.
[30] Jakubec, M. et al. (2024). Speech Emotion Recognition Using Transfer Learning: Integration of Advanced Speaker Embeddings and Image Recognition Models. Applied Sciences, 14(21).
[31] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), pp. 273-297.
[32] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), pp. 21-27.
[33] Breiman, L. (2001). Random forests. Machine Learning, 45(1), pp. 5-32.
[34] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), pp. 215-232.

BIOGRAPHIES

Yunus KORKMAZ received his B.Sc. degree in Computer Engineering from Cyprus International University in 2013. He earned his M.Sc. (2018) and Ph.D. (2023) in Computer Engineering from Fırat University, Turkey. Previously, he worked as a Software Developer at Yapi Kredi Bank and held research positions at the University of Turkish Aeronautical Association and Fırat University. His works focus on AI-driven solutions in audio/speech-based applications. He is currently an Assistant Professor in Department of Computer Engineering at Dicle University, Turkey.