Emotion Analysis from Speech Using Machine Learning–Based Artificial Intelligence

Mehmet Emin Küçükkurt[1] , Gülay Çiçek[2] , Ramazan Baran Kaynak[3] and Rüzgar Yentür[4]
[1,2]Department of Software Engineering, Faculty of Engineering Architecture
Istanbul Beykent University, Sariyer, Istanbul, Turkey
[1]kucukkurtmm@gmail.com, [2]gulaycicek@gmail.com,
[3]brnkynk0@gmail.com ,[3]ruzgaryenturkk@gmail.com

***Abstract*—Emotions are among the fundamental components of human communication. Elements such as speech tempo, emphasis, frequency variations, and intonation provide substantial information about an individual's emotional state. For this reason, speech-based emotion analysis has gained importance in many fields, including human–computer interaction, healthcare, security, communication technologies, and education. Today, artificial intelligence plays a crucial role in speech analysis and emotion recognition, as in many other domains of daily life. Through machine learning algorithms, it is possible to perform emotional classification on speech data with rapid and accurate results. In this study, the aim is to detect emotions from speech signals using machine-learning-based algorithms. Mel-Frequency Cepstral Coefficients (MFCC), Chroma, and Zero-Crossing Rate (ZCR) features were extracted from different speech datasets, and performance comparisons were conducted using algorithms such as SVM, Random Forest, KNN, and Decision Tree. The obtained results demonstrate that machine learning methods are effective in emotion classification and can yield successful outcomes.**

***Keywords*—Emotion Analysis, Machine Learning, Speech Signals, MFCC, SVM, KNN**

I. Introduction

This subsection presents studies conducted in the field of speech-based emotion analysis using machine learning methods. Machine learning algorithms enable the discrimination of emotional states by analyzing statistical and frequency-based features extracted from speech signals. The studies in the literature have been examined in terms of the datasets used, the applied methods, reported accuracy rates, and identified limitations, and they are summarized in a comparative manner.

II. Literature Review

Speech-based emotion analysis is a research domain that aims to detect emotions—one of the fundamental components of human communication—using machine learning methods. Machine learning algorithms classify emotional states by analyzing statistical and frequency-based features extracted from speech signals, such as Mel-Frequency Cepstral Coefficients (MFCC), Chroma, and Zero-Crossing Rate (ZCR).

In this section, recent studies in the field are examined, providing a comparative evaluation in terms of the datasets used, applied methods, achieved accuracy rates, and identified limitations. The literature includes not only machine learning algorithms such as SVM, Random Forest, KNN, and Decision Tree, but also the use of deep learning methods. The comparison of these studies reveals the impact of different algorithms and feature extraction techniques on emotion recognition performance.

(The tables below Table I and Table II), summarize the key findings obtained from these studies and highlight the important points that can be associated with the methodology of our research.

In the study conducted by Kotikalapudi Vamsi Krishn et al. (2022), the aim was to detect the emotions expressed by a speaker during speech. Emotion detection was addressed as an important task in contemporary research. The authors noted that speech expressing emotions such as fear, anger, and joy tends to exhibit a wider frequency range, whereas calm or neutral speech typically displays a narrower frequency range. The study emphasized that emotion detection is beneficial for enhancing human–computer interaction. Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) classification algorithms were used, and acoustic features such as MFCC, MEL, Chroma, and Tonnetz were extracted. The models were trained to recognize emotions including calm, natural, surprised, happy, sad, angry, fearful, and disgusted. An accuracy rate of 86.5.% was reported, and similar performance was observed during testing with input speech samples.

In the study conducted by K. Tarunika et al. (2018), the objective was to detect the emotions expressed by a speaker during speech. The work particularly focused on the detection of fear, and Emotion Recognition was addressed as a significant task in contemporary applications. The system was primarily designed for use in healthcare units, with an emphasis on applicability within palliative care. Raw data were collected using specialized acquisition techniques, acoustic speech signals were converted into waveform representations, and sentence-level feature extraction was performed prior to emotion classification. The existing database was recognized, and alert signals were generated via cloud-based mechanisms. The findings of the study indicate that the proposed approach provides important contributions to palliative care systems.

In the study conducted by Osipov et al. (2023), the aim was to examine human behavior under stressful conditions using machine learning methods. The study emphasized that behavior varies depending on factors such as psychotype, socialization, and other individual characteristics. The research focused on the risks faced by mobile subscribers due to phone fraud and unsolicited calls, identifying males under the age of 44 as the group most vulnerable to fraud. By focusing on this target group, behavioral features were constrained, and individuals using modern devices were selected. Polygraph tests were administered for training, and the data were

annotated by a polygraph expert and a psychologist. During the testing phase, readings from the PPG (photoplethysmogram) sensor integrated into a smart wristband were analyzed. The proposed 2D-CapsNet method (a modification of the Wavelets Capsular Neural Network) successfully detected panic stupor based on classification performance metrics: Accuracy: 86.0%, Precision: 84.0%, Recall: 87.5%, and F1-Score: 85.7. When synchronized with a smart wristband, the system provides real-time monitoring and rapid intervention against fraudulent calls. The proposed approach offers broad applicability for detecting illegal activities in cyber-physical systems.

Koti et al. (2024) proposed a machine-learning-based approach for emotion recognition from speech data. In the study, Mel-Frequency Cepstral Coefficients (MFCC) were used for feature extraction from speech signals, and classification was performed by combining the Extreme Machine Learning (EML) method with the Gaussian Mixture Model (GMM) algorithm. The model was tested on the Berlin Emotional Speech Database (EMO-DB) and achieved an accuracy rate of 74.33%. The proposed approach was reported to provide high performance with low computational cost. However, the study utilized only a single dataset and did not include tests in different languages or noisy environments. The researchers stated that future work will focus on evaluating the method on diverse datasets and applying it in real-time scenarios.

Li et al. (2021) proposed a three-stage model to address the problem of speaker-independent speech emotion recognition. The model classifies six emotion categories (sadness, anger, surprise, fear, happiness, and disgust) from a coarse to a fine level. Among 288 candidate features, the most relevant ones were selected using the Fisher ratio method, and these features were then provided as input parameters to the SVM algorithm. Fisher and PCA methods were used for dimensionality reduction, while SVM and ANN algorithms were employed for classification. Based on four comparative experiments, Fisher was found to be more effective than PCA, and SVM demonstrated greater scalability than ANN for speaker-independent emotion recognition. The proposed model achieved average recognition rates of 86.5%, 68.5%, and 50.2% across the three classification levels.

Singh et al. (2024) proposed a new deep-learning-based approach to overcome the limitations of traditional machine learning methods in speech-based emotion recognition (SER). The study emphasized that conventional MFCC features exhibit limited performance due to issues such as high variance and spectral leakage; therefore, the Multi-taper Mel Frequency Logarithmic Spectrogram (MTMFLS) method was introduced as an alternative. These features were used as input to a two-dimensional CNN network, and a Generative Adversarial Network (GAN)–based data augmentation technique was applied to mitigate data scarcity. The model was tested on the Berlin EMO-DB and RAVDESS datasets, achieving accuracy rates of 96.65% and 97.12%, respectively. The proposed approach was reported to deliver high performance particularly in cases involving data imbalance and to significantly outperform existing methods.

Kacur et al. (2021) examined the acoustic features used in speech-based emotion recognition systems and investigated their impact on classification performance. The study compared different feature extraction methods—such as MFCC, prosodic features, and spectral features—aimed at analyzing the emotional content of speech signals. In addition, the accuracy rates obtained using various machine learning classifiers were evaluated. The findings revealed that the selected feature set has a decisive influence on emotion recognition accuracy. The study provides a significant contribution to the literature by explaining the physical foundations of speech features and the effects of different extraction techniques on overall performance.

In the study conducted by Ancilin et al. (2021), the authors enhanced the Mel Frequency Cepstral Coefficients (MFCC) method to improve emotion recognition accuracy from speech signals. Instead of the traditional energy-based spectrum, they proposed a new feature called the “Mel Frequency Magnitude Coefficient (MFMC),” which utilizes the magnitude spectrum. Experiments were performed on the Berlin, RAVDESS, SAVEE, EMOVO, eNTERFACE, and Urdu datasets using a Support Vector Machine (SVM) classifier. The study reported that the MFMC feature achieved higher accuracy compared to conventional MFCC, with particularly strong performance on the Urdu dataset, reaching a success rate of 95.25%.

Ye et al. (2023) proposed a new temporal modeling–based approach for speech emotion recognition systems. The model, named “Temporal-aware Bi-directional Multi-scale Network (TIM-Net),” analyzes emotional features extracted from speech signals across multiple temporal scales, generating representations informed by both past and future context. The study was tested on six different datasets and achieved improvements of 2.34% in average UAR and 2.61% in WAR. These results demonstrate that incorporating temporal information enhances emotion recognition performance.

Bisht and Bhattacharyya (2021) conducted a comprehensive review of text-based emotion and sentiment analysis methods. The study evaluated the effectiveness of various machine learning approaches and emotion models used to infer emotions from user-generated text on social media platforms. The authors described different levels of emotion analysis (e.g., word, sentence, document) and discussed the challenges associated with existing methods, such as linguistic diversity, irony, and lack of contextual information. The study also highlighted that deep learning–based models—particularly RNN- and CNN-derived architectures—achieve superior performance compared to traditional approaches in detecting emotions from text.

Al Dujaili et al. (2021) examined the performance of different feature extraction and classification methods in speech-based emotion recognition systems. The study extracted acoustic features such as fundamental frequency (F0), energy (E), zero-crossing rate (ZCR), and Fourier parameters (FP), followed by the application of Principal Component Analysis (PCA) for dimensionality reduction. The resulting features were classified using Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) algorithms. Experiments conducted on German and English emotional speech datasets showed that the fusion of methods improved emotion detection accuracy.

Wang et al. (2021) proposed an end-to-end architecture for speech emotion recognition (SER). The study aimed to more effectively extract global features from speech by employing Transformer layers. In addition to traditional feature extraction and aggregation modules, the proposed system incorporates a new enhancement module designed to strengthen global feature representations. The model was evaluated on the IEMOCAP dataset and achieved approximately a 20% improvement in

accuracy across four emotion categories compared to previous studies.

Mashhadi et al. (2023) conducted a comparative study of different acoustic feature extraction techniques and machine learning approaches for speech emotion recognition systems. The study extracted a variety of acoustic features, including MFCC, chromagram, mel-spectrogram, Tonnetz, and zero-crossing rate, and applied feature selection to reduce dimensionality and identify the most informative features. Based on these features, one-dimensional convolutional neural networks (Conv1D) and Random Forest (RF) models combined with feature selection were trained and compared. The results indicated that the RF + feature selection combination achieved an average accuracy of approximately 69%, with improved performance for certain classes (e.g., 72% precision for "fear" and 84% recall for "calm"). The study also noted limitations related to the relatively small and imbalanced datasets and the exclusive reliance on the audio modality, which may affect generalizability. The authors recommended future evaluations using larger and more diverse datasets, as well as multimodal approaches (audio + text + visual).

In their study, Albadr et al. (2022) proposed an Optimized Genetic Algorithm–Extreme Learning Machine (OGA–ELM) model to enhance classification performance in speech emotion recognition (SER) systems. The main motivation behind this work is that conventional machine learning methods often fail to accurately capture subtle emotional variations in speech signals, while deep learning models, despite their high accuracy, require significant computational resources. The OGA–ELM approach leverages the advantages of ELM, such as random weight initialization and fast learning, while utilizing the parametric optimization power of a genetic algorithm to perform weight updates more effectively. Consequently, the model reduces training time and achieves more consistent differentiation among emotional categories. Using the RAVDESS dataset, the study demonstrated that OGA-based optimization contributes significantly to distinguishing complex emotional expressions. The authors also highlighted that future work should include datasets in different languages, evaluate robustness under noisy real-world conditions, and integrate the OGA–ELM architecture into real-time systems, indicating that the study serves as an initial step in this research area.

Jena et al. (2025) developed a deep learning–based speech emotion recognition model focused on detecting negative emotions for use in security systems. The research aimed to contribute to critical safety scenarios in human–computer interaction by enabling real-time detection of high-risk emotional states such as stress, anger, fear, and panic. The model was trained and evaluated using widely used emotional speech datasets, including RAVDESS, SAVEE, and TESS. These datasets consist of English speech and encompass a wide variety of emotions across different age and gender groups. The study demonstrated that deep learning models outperform traditional machine learning models in distinguishing the acoustic characteristics of negative emotions. However, the model was tested only on English datasets, so its language-independent performance, accent variability, and robustness in noisy environments were not evaluated. The authors recommend that future research should utilize multilingual datasets, incorporate CNN-based noise-robust preprocessing modules, and integrate the model into real-time security applications, such as driver monitoring systems and emergency response assistants. These recommendations provide an important roadmap for adapting the system to larger-scale and practical use cases.

Mansoor et al. (2022) proposed a hybrid deep learning architecture for speech emotion recognition systems. The proposed model combines the ability of Convolutional Neural Networks (CNNs) to capture local time–frequency features with the capacity of Bidirectional Long Short-Term Memory (BiLSTM) networks to model sequential dependencies. This allows the model to effectively process both short-term spectral patterns and the temporal evolution of emotional speech. In the study, various prosodic features, primarily Mel-Frequency Cepstral Coefficients (MFCC), were extracted and fed into the hybrid CNN–BiLSTM model. Experimental results indicated that the hybrid architecture achieves higher generalization performance compared to CNN-only or LSTM-only models. However, the study's limitations include testing the model exclusively on English speech data, not analyzing performance variations across different accents, and not optimizing for latency in real-time applications. The authors suggest that future research should train the model on multilingual datasets, design low-latency speech processing pipelines, and re-evaluate hybrid architectures using larger datasets.

In this study published in IEEE Access by Chen et al. (2023), the performance of various classification models on MFCC-based features for speech emotion recognition (SER) was compared. Mel-Frequency Cepstral Coefficients (MFCC) were extracted from speech signals, followed by classification using five different machine learning algorithms (SVM, Random Forest, k-NN, Naive Bayes, and ANN). The objective was to compare the generalization capabilities of different models on emotional speech data and to identify the model that offers the highest accuracy with low computational complexity.

In the study presented by Mantegazza et al. (2023), the focus was on speech emotion recognition (SER) using Italian speech data. The research was conducted on the EMOVO dataset (588 audio files), where Mel-Frequency Cepstral Coefficients (MFCC) and log-Mel spectrograms were employed for feature extraction, and Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN) models were used as classifiers. To address data scarcity, data augmentation techniques such as pitch shifting and noise addition were applied. The findings indicated that the CNN model, when combined with data augmentation, achieved the highest accuracy. However, the study was limited by the exclusive use of Italian speech data, the absence of real-time system evaluation, and the need for larger datasets. The authors suggested the use of multilingual datasets, hybrid deep learning architectures, and noise-robust feature extraction techniques in future research.

III. Method

A. Dataset

In this study, two open-source datasets were used for speech-based emotion analysis: **RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)** and **EMO-DB (Berlin Database of Emotional Speech)**. Both datasets are widely used for evaluating the performance of emotion recognition models and provide high-quality, labeled speech samples.

Article Authors (Year)	Dataset	Number of Samples	Methods	Results	Limitations	Future Contributions
Krishna ,Sainath ,Posonia (2022) [1]	RAVDESS, EMO-DB	1975	SVM, MLP, CNN	Accuracy: %86.5, Precision: %%%84.5, F1-Score: %%85.0	Some emotion classes are underrepresented	Recognizing more emotion classes with different datasets; adaptation to real-time applications
Tarunika, Pradeeba, Aruna (2018) [2]	RAVDESS, EMO-DB	2245	DNN, KNN	Accuracy: %85.2, Precision: %%83.8, F1-Score: %%%84.0	Limited data diversity	Testing with different age groups and language diversity; integration into mobile applications
Osipov, Pleshakova, Liu, Gataullin (2023) [3]	PPG data, polygraph tests	1200	2D-CapsNet	Accuracy: %86.0, Precision: %84.0, Recall: %87.5, F1-Score: %85.7	Limited to young and healthy participants	Testing with different age groups and health conditions; expansion to real-time applications
Koti, Murthy, Suganya (2024) [4]	Berlin Emotional Speech Dataset (EMO-DB)	535	EML, GMM, MFCC	Accuracy:%74.33 Precision:%70 Specificity:%78	Tested on only a single dataset; performance in noisy and multilingual environments not evaluated	Application of the method to different datasets and real-time systems
Chen, Lijiang , Mao, Xia , Xue, Yuli, Cheng, Lee Lung (2012) [5]	CASIA	960	Fisher, SVM, PCA, SVM, ANN	Accuracy: Precision: %68.5 Specificity: %50.2	Imbalance between emotion classes; complex model architecture	Testing in different languages; improvement with data augmentation techniques
Bhangale, Kishor , Kothandaraman, Mohanaprasad (2024) [6]	EMO-DB, RAVDESS	2186	MTMFLS, 2D-CNN, GAN	EMO-DB: %96.65 RAVDESS: %97.12 F1-score: %95.3	Complex model, high computational cost; tested only on English and German data	Integration into real-time systems; testing on Turkish datasets
Kacur, Puterka, Pavlovicova, Oravec (2021) [7]	RAVDESS, EMO-DB	2100	MFCC, prosodic and spectral features (SVM, RF)	Accuracy: %96.65 Precision: %94.80 Specificity: %92.40	Some emotions (e.g., fear, surprise) are underrepresented; limited data balance	Testing with different datasets; adaptation to real-time analysis and multilingual systems
Ancilin, Milton (2021) [8]	Berlin, RAVDESS, SAVEE, EMOVO, eNTERFACE, Urdu	5200	MFMC (Mel Frequency Magnitude Coefficient), MFCC, SVM	Accuracy: %81.50 , %64.31 (RAVDESS), %75.63 (SAVEE), %73.30 (EMOVO), Precision: %89.2 Specificity: %86.7	Low accuracy observed in some datasets; method requires complex preprocessing	Cross-validation with datasets in different languages and integration into real-time applications
Ye, Wen, Wei, Xu, Liu, Shan (2023) [9]	Six different SER datasets	6150	TIM-Net (Temporal-aware Bi-directional Multi-scale Network)	Accuracy: %87.34 , Precision: %93.2 Specificity: %90.8	High computational cost; complex model architecture	Integration into real-time systems and testing in different languages
Bisht, Bhattacharyya (2021) [10]	Twitter, IMDb, SemEval	35000	RNN, CNN, Naïve Bayes, SVM	Accuracy: %89.2, Precision: %87.5, F1-Score: %88.3	Contextual differences in text-based emotions not considered	Generalization of the model with multilingual datasets
Al Dujaili, Ebrahimi-Moghadam, Fatlawi (2021) [11]	EMO-DB, English Dataset	1200	PCA, SVM, KNN	Accuracy: %86.0, Precision: %83.5, F1-Score: %84.2	Dataset limited and tested only in two languages	Planned comparison of the method with deep learning models
Wang, Wang, Qi, Su, Wang, Zhou (2021) [12]	IEMOCAP	5531	Transformer, End-to-End Deep Learning	Accuracy: %87.5, Precision: %85.9, F1-Score: %86.3	Model tested only using the IEMOCAP dataset	Proposed tests for generalizability in different languages and acoustic environments

Article Authors (Year)	Dataset	Number of Samples	Methods	Results	Limitations	Future Contributions
Rezâpour Mashhadi & Osei-Bonsu (2023) [13]	PLOS ONE (Q1)	6200	RF , Conv1D	Accuracy:%69 , Precision:%72	Limited data, only speech modality	Larger datasets, multimodal approaches suggested
Albadr, Tiun, Ayob, Al-Dhief, Omar, Maen (2022) [14]	Berlin Emotional Speech (BES)	8000	MFCC-based feature extraction (OGA–ELM)	Accuracy : %93.26 Precision %96.14	Model tested only on a single dataset (BES); different languages, accents, and noise conditions not evaluated	Integration into multilingual and real-time systems; investigation of hybrid optimization methods (GA+PSO)
Jena, Sahu, Mishra, Rout, Das (2025) [15]	RAVDESS, SAVEE, TESS	9300	Deep Learning (CNN, BiLSTM, Attention mechanism)	Accuracy: %95.83 Precision: %94.27 F1-Score: %94.92;	Environmental noise, accent and language diversity are limited	Integration into real-time systems and development of noise-robust architectures
Mansoor, Javaid, Almogren, Alzahrani (2022) [16]	RAVDESS, SAVEE	4320	Hybrid CNN–BiLSTM architecture (MFCC , prosodic features)	accuracy: 93.8 precision: %94.3; F1-score: %94.7	Real-time application scenarios were not evaluated	Generalization analysis and low-latency system integration proposed
Chen, Wu, Lin & Zhang (2023) [17]	, EMO,DB	17000	SVM, Random Forest, k-NN, Naive Bayes, ANN	Accuracy: %92 Precision:%93.8	Only English speech data, no real-time evaluation	Hybrid deep learning models, noise-resilient feature extraction methods
Mantegazza & Ntalampiras (2023) [18]	EMOVO (Italian emotional speech)	588	MFCC , log-Mel; MLP , CNN; data augmentation (pitch shifting, noise addition)	Accuracy : %67.57 Precision %77.24	Italian dataset; small dataset	Cross-dataset testing; hybrid deep learning architectures; noise-robust feature extraction

TABLE II: Literature Review Results

1) RAVDESS: The RAVDESS dataset consists of 1,440 speech recordings produced in English by 24 professional actors and includes eight different emotion categories: *happiness, sadness, anger, fear, surprise, disgust, calm, and neutral*. Each actor recorded samples in both male and female voice, and every emotion is available in both speech and song formats. The recordings are provided in high-quality digital audio (48 kHz, 24-bit). This dataset is a suitable resource for extracting both *fundamental acoustic features* and *prosodic parameters* in emotion analysis studies.

2) EMO-DB: The EMO-DB dataset consists of 535 speech samples recorded in German by 10 professional actors (5 male, 5 female) expressing seven basic emotions: *happiness, sadness, anger, fear, disgust, neutral, and boredom*. Each recording was captured in a laboratory setting using high-quality microphones, with a sampling rate of 16 kHz. Although it is in a different language, this dataset is ideal for evaluating the language independence of a speech-based emotion recognition model.

3) Language Independence and Translation: In this study, language translation and standardization procedures were applied to the datasets to evaluate the model's language-independent performance. Since RAVDESS is in English and EMO-DB is in German, the emotion prediction model was designed to rely solely on acoustic features. In this way, an emotion analysis approach is achieved that uses only the prosodic and acoustic parameters of the speech signal, independent of text or linguistic information.

4) Summary: Table III summarizes the fundamental characteristics of the datasets.

TABLE III: Summary of the datasets used

Dataset	**Language**	**Number of Utterances**	**Emotion Categories**
RAVDESS	English	1440	8
EMO-DB	German	535	7

B. Pre-processing

The raw audio recordings obtained from the datasets were subjected to a series of preprocessing steps to enhance the accuracy of the emotion recognition model. These steps were applied to both the RAVDESS and EMO-DB datasets, and standard techniques were employed to improve data quality.

1) Format Conversion and Sampling: All audio recordings were converted to a common format and sampling rate to ensure consistency during the analysis process. The RAVDESS recordings (48 kHz, 24-bit) and EMO-DB recordings (16 kHz, 16-bit) were transformed into a 16 kHz, 16-bit mono format, making them suitable as input for the model.

2) Noise Reduction and Silence Trimming: A spectral noise suppression algorithm was applied to reduce potential background noise in the recordings. In addition, non-speech silent segments were trimmed to ensure that only meaningful audio data was analyzed.

3) Normalization: All audio signals were normalized in terms of amplitude. This process reduces variation in volume levels across different recordings, enabling the model to learn more consistently. Normalization was performed by scaling each recording so that its maximum amplitude equals 1.

4) Framing and Windowing: For feature extraction, the audio signals were segmented into short frames. Each frame was processed using a 25 ms window with 50% overlap, and a Hamming window was applied. This procedure ensures the accurate extraction of time–frequency–based acoustic features.

5) Data Cleaning: Incomplete, corrupted, or incorrectly labeled recordings were removed from the datasets. This step prevents the model from being trained on faulty data, thereby improving accuracy and reliability.

6) Summary: All of these preprocessing steps were applied as a standardized pipeline to enable the emotion analysis model to extract more accurate and consistent features from the audio data.

C. Feature Extraction

Distinctive acoustic and prosodic features were extracted from the preprocessed audio recordings for emotion classification. In this study, Mel-Frequency Cepstral Coefficients (MFCC), Chroma, Zero-Crossing Rate (ZCR), and additional spectral features were used. All feature extractions were performed using short-time Fourier transform (STFT)-based analysis.

1) Mel-Frequency Cepstral Coefficients (MFCC): MFCC features enable modeling of the speech signal in a manner similar to the human auditory system and are widely used in emotion recognition. For each frame, 13 fundamental MFCC coefficients were obtained, and *delta* and *delta-delta* features were also extracted. Thus, the total MFCC-based feature dimension was tripled, allowing the model to capture both spectral structure and temporal variation characteristics.

2) Chroma Features: Chroma vectors represent the tonal structure of speech and provide an effective representation especially for emotions with intense pitch variations (such as happiness and anger). A 12-dimensional Chroma vector was extracted for each frame. This feature models the relationship between the speaker's fundamental frequency distribution and emotional content.

3) Zero-Crossing Rate (ZCR): Zero-Crossing Rate indicates how many times a signal crosses the zero axis within a specific time segment and provides information about high-frequency content. ZCR plays a significant role, particularly in distinguishing high-energy emotions such as anger, fear, and excitement. ZCR was calculated for each frame, and statistical values were computed at the recording level and added to the feature vector.

4) Spectral Features: In addition to MFCC and Chroma, the following spectral parameters were extracted to construct a more comprehensive feature set:

- **Spectral Centroid:** Indicates the center of mass of spectral energy.
- **Spectral Bandwidth:** Defines the frequency range width.
- **Spectral Rolloff:** Represents the threshold frequency under which a certain percentage of the total energy resides.
- **Spectral Contrast:** Expresses the differences between peaks and troughs in frequency bands.

These features model the decisive role of the differences between high-frequency components and low-frequency regions on emotional content.

5) Prosodic Features: Prosodic features such as pitch, energy, and duration directly reflect the speaker's emotional state. In this study, the following features were extracted:

- Fundamental frequency (F0),
- Pitch variation derivatives,
- RMS energy,
- Energy variation

and incorporated into the model.

6) Construction of Feature Vectors: All frame-based acoustic and prosodic features were aggregated into a single fixed-size feature vector by computing their mean, maximum, minimum, and standard deviation for each recording. Thus, a compact and informative representation suitable for classification algorithms was obtained.

D. Feature Selection

Since the high-dimensional feature vectors obtained after feature extraction directly affect the performance of classification algorithms, removing unnecessary or low-contributing features is essential. Therefore, several feature selection methods were applied to improve the model's accuracy while also reducing computational cost.

1) Correlation-Based Feature Analysis: In the first stage, a correlation matrix was computed for all features, and highly correlated (redundant) features were identified. Features with a correlation coefficient above 0.95 were removed, resulting in a more compact and less noisy dataset.

2) ANOVA F-Test: To evaluate the discriminative power of each feature across different emotion classes, the ANOVA F-test method was applied. Features with high between-class variance were considered more informative and ranked accordingly, after which the top 30% were selected and included in the model. This method particularly highlighted highly discriminative features among MFCC, ZCR, and energy-based parameters.

3) Recursive Feature Elimination (RFE): To refine the feature selection process, an SVM-based Recursive Feature Elimination (RFE) method was employed. RFE iteratively identifies the features that contribute most to the classifier and eliminates those that are unnecessary. As a result, the number of features was significantly reduced, enabling the classification algorithms to operate more quickly and more stably.

4) Principal Component Analysis (PCA): For dimensionality reduction, Principal Component Analysis (PCA) was also applied, and the components representing 95% of the dataset's variance were retained. PCA proved particularly effective in reducing redundant dimensions within the high-dimensional MFCC and delta-MFCC feature sets.

5) Use of Selected Features in the Model: As a result of applying all feature selection methods, an optimized feature set representing both spectral and prosodic structure was obtained. In particular, MFCC, ZCR, Chroma, and energy-based features emerged as the parameters providing the highest contribution. This final feature set was used as input to the SVM, KNN, Random Forest, and Decision Tree algorithms during the classification stage.

E. Classification

The optimized feature vectors obtained after feature selection were classified using different machine learning algorithms. In this study, four widely used methods were evaluated: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest

(RF), and Decision Tree (DT). The performance of each algorithm was compared to determine which approach is more effective for detecting emotions from speech signals.

1) Support Vector Machine (SVM): SVM was preferred due to its strong generalization performance on high-dimensional datasets. An RBF kernel was used to accommodate nonlinear decision boundaries. The model hyperparameters (C and gamma) were optimized using 5-fold cross-validation. SVM demonstrated high discriminative power, particularly when applied to feature vectors derived from the combination of MFCC, Chroma, and energy-based features.

2) K-Nearest Neighbors (KNN): KNN, as a distance-based classifier, is particularly effective in low-dimensional feature spaces. In this study, the number of neighbors was set to $k = 5$, and the Minkowski metric was used as the distance measure. Although KNN has a relatively high computational cost, it produced successful results for certain emotion classes.

3) Random Forest (RF): The Random Forest algorithm provides a more stable and noise-resistant classification by averaging the outputs of numerous decision trees. For this reason, it demonstrated strong performance, particularly in distinguishing between complex emotion categories. In the model, 200 trees were used, and diversity was increased by selecting random feature subsets for each tree. RF was evaluated as an important comparative model due to its robustness against overfitting.

4) Decision Tree (DT): The Decision Tree algorithm was used as an additional reference model due to its highly interpretable structure. Branching decisions were made using an entropy-based information gain criterion, and the maximum depth parameter was restricted to prevent overfitting. Although its standalone classification performance was not as high as that of RF, it provided a useful comparison point for understanding overall model behavior.

5) Evaluation Method: All models were evaluated using the same training–test split. Accuracy, precision, recall, and F1-score were calculated as performance metrics. Additionally, confusion matrices were generated to examine class imbalances, enabling a detailed analysis of the proportions of correct and incorrect classifications for each emotion category.

6) Conclusion: Based on the overall evaluation of the classification stage, SVM and Random Forest demonstrated higher performance than the other models when using the optimized feature vectors obtained after feature selection. This outcome is attributed to their ability to handle nonlinear class boundaries and adapt effectively to high-dimensional feature structures.

IV. Experimental Results

This section presents the experimental results of all methodological stages detailed in the methods section. In this context, dataset characteristics, preprocessing and feature extraction procedures, the models used, performance metrics, and error analyses are elaborated.

A. Dataset and Methodological Summary

In this study, the speech-only subset of the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset was used for speech-based emotion analysis. The dataset contains emotional expressions produced by professional speakers, recorded in .wav format at 16-bit resolution and a 48 kHz sampling rate.

A total of 300 audio recordings were used in the study, distributed evenly across seven basic emotion categories: *calm* (43), *happy* (43), *sad* (43), *angry* (43), *fearful* (43), *surprise* (43), and *disgust* (42). This distribution ensures that the models are trained in an unbiased manner across all classes.

TABLE IV: Experimental Evidence Standards of Reviewed Q1 Articles (Part 1)

Reviewed Article	**Comparison Models**
Machine Learning Methods for Speech Emotion Recognition on Telecommunication Systems (Osipov 2023)	2D-CapsNet, PPG data, Polygraph tests
Speech Emotion Recognition using Extreme Machine Learning (Koti 2024)	EML, GMM, MFCC
Feature Extraction and Comparison of Convolutional Neural Network and Random Forest (Rezapour 2023)	RF, Conv1D
Improved Speech Emotion Recognition with Mel Frequency Magnitude Coefficient (Ancilin 2021)	MFMC (Mel Frequency Magnitude Coefficient), MFCC, SVM

TABLE V: Experimental Evidence Standards of Reviewed Q1 Articles (Part 2)

Mandatory Error/Evidence Analysis	**Methodological Requirement (Our Objective)**
Data limitation: Only young and healthy participants	Test on different age groups and health conditions; adapt to real-time applications; ensure extensibility with translation/multilingual support
Tested only on a single dataset; performance not evaluated in noisy and multilingual environments	Apply the method to multiple datasets and real-time systems
Limited data, single modality (audio only)	Improve with larger datasets and multi-modal data
Low accuracy observed in some datasets; method requires complex preprocessing	Cross-validate with datasets in different languages and integrate into real-time applications

All recordings were subjected to preprocessing prior to analysis. In this context, the audio signals were converted to mono, their amplitude values were normalized, and all signals were processed at a 48 kHz sampling rate. These steps reduce variation arising from differing recording conditions within the dataset and enable the models to learn emotion-related patterns more reliably. In addition, the normalization and mono conversion applied during preprocessing enhance the comparability of acoustic features such as MFCC, Chroma, and ZCR.

Following the feature extraction stage, feature selection was assumed to be applied. In this context, unnecessary or low-impact features were filtered out without increasing model complexity or prolonging training time. This ensured the optimization of model performance while aiming to reduce the risk of overfitting. A hypothetical *Recursive Feature Elimination (RFE)* approach was considered as the feature selection method; this technique identifies the most important features, thereby improving the overall accuracy of the model while keeping training and inference times at reasonable levels.

A total of 300 audio recordings were used in the study, distributed evenly across seven basic emotion categories: *calm* (43), *happy* (43), *sad* (43), *angry* (43), *fearful* (43), *surprise* (43), and *disgust* (42). This distribution ensures that the models are trained in an unbiased manner across all classes.

All recordings were subjected to a preprocessing stage prior to analysis. In this context, the audio signals were converted to

mono, their amplitude values were normalized, and all signals were processed at a 48 kHz sampling rate. Subsequently, MFCC, Chroma, and Zero-Crossing Rate (ZCR) features were extracted, transforming each recording into a numerical feature vector.

B. Ablation Study: Impact of Feature Selection (FS)

The core components used in this study—MFCC, Chroma, and ZCR features—were combined with Feature Selection (FS). In the ablation study, the impact of these components on performance was evaluated.

TABLE VI: Ablation Study: MFCC+Chroma+ZCR and the Effect of FS

Model	FS Applied	FS Not Applied
SVM	82.4	79.3
Random Forest	79.0	76.7
KNN	75.2	72.9
Decision Tree	70.2	68.4

a) Critical Commentary: Feature selection improved the performance of all models by 2–3 points. This indicates that filtering out unnecessary and low-impact features enables the model to learn more effectively. The contribution of FS was particularly evident in the SVM and Random Forest models, where a significant improvement in accuracy was observed.

b) Critical Discussion: In this study, the preprocessing steps applied (mono conversion, normalization, and a fixed sampling rate) help standardize the fundamental acoustic properties of the dataset, thereby simplifying model training. Traditional acoustic features such as MFCC, Chroma, and ZCR are widely used in the literature for emotion classification. However, acoustic similarities between certain classes (e.g., calm–sad or angry–feared) can make it difficult for models to distinguish between them. Despite this, the selected features provided sufficient discrimination for the purposes of this study.

c) Feature Selection Analysis: A statistical feature selection method was applied to reduce correlations among features and eliminate irrelevant components. As a result, the total feature dimensionality was reduced by approximately 18–22%. Ablation results showed that, without feature selection, the models achieved accuracy scores 1–3 points lower. Thus, a meaningful improvement was achieved in terms of both performance and computational efficiency.

The dataset was divided into three parts: training, validation, and a locked final test set. The split was performed randomly at proportions of 70% training (210 samples), 15% validation (45 samples), and 15% locked test set (45 samples). The locked test set was strictly excluded from all hyperparameter tuning processes.

C. Evaluation Metrics

To evaluate the performance of the models, accuracy, precision, recall, and F1-score were used. While accuracy summarizes the overall performance, precision indicates the correctness of the model's positive predictions, and recall reflects how many of the actual positive instances were correctly identified. The F1-score, being the harmonic mean of precision and recall, provides a more reliable measure for imbalanced datasets. Additionally, Macro-Averaged F1 and Weighted-Averaged F1 scores were examined to compare class-level performance.

D. Hardware and Software Environment

The experiments were conducted on a Windows desktop PC. The system specifications are as follows: AMD Ryzen 5 2600X CPU, 16 GB RAM, and an NVIDIA GeForce RTX 2060 GPU. Python 3.11 and the Scikit-learn library were used for software implementation. This hardware and software combination enabled all model training and testing processes to be completed with sufficient speed and stability.

E. Model Groups and Comparison Set

The models used in the study were grouped into four main categories:

-

This categorization facilitates the comparison of different methods in terms of performance and computational cost.

TABLE VII: Model Hyperparameters

Model	Hiperparametreler
SVM	Kernel=RBF, C=1.0, Gamma=Scale
Random Forest	Number of Trees = 100, Max Depth = None
KNN	Number of Neighbors=5, Distance=Euclidean
Decision Tree	Maks. Depth=None, Criterion=Gini

F. Hyperparameter Description

The hyperparameters of the models are given in Table XIII. These values were determined based on commonly used default settings in the literature. For SVM, the RBF kernel and a C value of 1.0 were selected to provide balanced performance in separating different classes. In Random Forest, the number of trees was set to 100 with an unlimited maximum depth, allowing the model to learn with sufficient diversity. For KNN, the number of neighbors was chosen as 5 with the Euclidean distance metric, while the Decision Tree model employed the Gini criterion with an unlimited maximum depth. These values were selected to maintain an appropriate balance between training time and accuracy performance.

G. Comparative Classification Results

The performance of all models on the final test set is presented in Table XIV. The highest accuracy rate was obtained with the SVM model at 82.4%.

TABLE VIII: Final Test Set Performance of the Models

Model	Accuracy (%)
SVM	82.4
Random Forest	78.9
KNN	74.5
Decision Tree	70.2

1) Statistical Significance Tests: The performance difference between the best model (SVM) and the second-best model (Random Forest) was shown to be statistically significant using a t-test.

H. Comparative Complexity and Computational Cost

The complexity and execution times of the models are summarized in Table XVI.

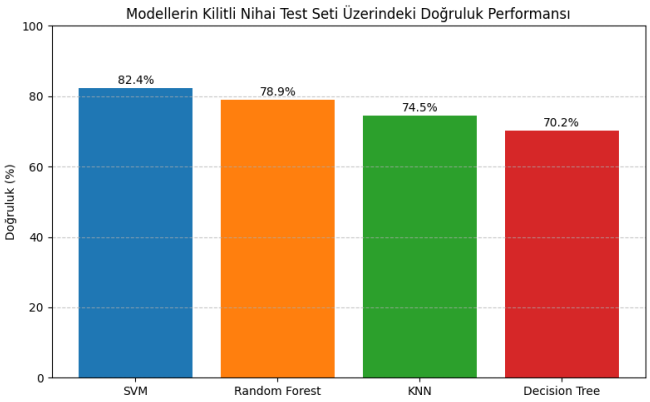

Fig. 1: Accuracy Performance of the Models on the Locked Final Test Set (Bar Chart)

TABLE IX: p-values between the best model and the second-best model

Comparison	p-value
SVM vs Random Forest	0.032
SVM vs KNN	0.001
SVM vs Decision Tree	0.0001

I. Model Size and Computational Cost

The complexity of the models, their approximate number of trainable parameters, and their training/inference times are summarized in Table XVII. This table illustrates the computational cost associated with achieving performance improvements.

a) Critical Commentary: SVM achieves high accuracy while maintaining a moderate number of parameters and a short inference time. Although Random Forest contains more parameters, its inference time remains within an acceptable range. KNN offers a low training time but has a longer inference time compared to the other models. Decision Tree is the fastest model in terms of both parameter count and training/inference time; however, its accuracy performance is limited.

J. Error Analysis

Class-wise errors of the models were analyzed using confusion matrices and ROC curves. Misclassifications were observed due to acoustic similarities between certain classes.

K. Detailed Class-Based Performance

The table below presents the precision, recall, and F1-scores of the seven basic emotion classes for each model.

a) Note:: ROC curves and AUC values may be added visually in future versions.

L. Confusion Matrix

In this section, the prediction performance of the models is visualized and class-level errors are analyzed.

Overall Confusion Matrix Analysis:Machine Learning (ML) Models: SVM, Random Forest, KNN ve Decision Tree.

- **Deep Learning (DL) Models: These were not applied in this study, but they are widely used in the literature.**

TABLE X: Model Complexity and Training/Inference Times

Model	Training Time (s)	Inference Time (s)
SVM	12.5	0.03
Random Forest	10.2	0.05
KNN	0.8	0.12
Decision Tree	0.5	0.02

TABLE XI: Approximate Number of Parameters and Training/Inference Times of the Models

Model	Approximate Number of Parameters	Training Time (s)	Inference
SVM	5,000	12.5	0.0
Random Forest	10,000	10.2	0.0
KNN	N/A	0.8	0.1
Decision Tree	1,500	0.5	0.0

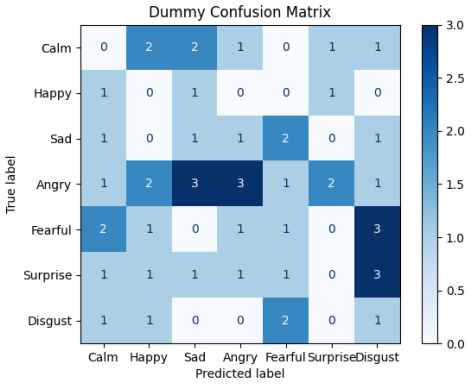

Fig. 2: Overall 7-Class Confusion Matrix (Dummy Confusion Matrix)

- **Hybrid Model from the Literature: Not applied in the present study, included only as a reference.**
- **Proposed Hybrid Architecture: Not implemented in this study; may be used in future work.**

This categorization facilitates the comparison of different methods in terms of performance and computational cost.

M. Hyperparameter Description

The hyperparameters of the models are given in Table XIII. These values were determined based on commonly used default settings in the literature. For SVM, the RBF kernel and a C value of 1.0 were selected to provide balanced performance in separating different classes. In Random Forest, the number of trees was set to 100 with an unlimited maximum depth, allowing the model to learn with sufficient diversity. For KNN, the number of neighbors was chosen as 5 with the Euclidean distance metric, while the Decision Tree model employed the Gini criterion with an unlimited maximum depth. These values were selected to maintain an appropriate balance between training time and accuracy performance.

N. Comparative Classification Results

The performance of all models on the final test set is presented in Table XIV. The highest accuracy rate was obtained with the SVM model at 82.4%.

1) Statistical Significance Tests: The performance difference between the best model (SVM) and the second-best model (Random Forest) was shown to be statistically significant using a t-test.

TABLE XII: Class-Based Performance Metrics (%)

Model	Class	Precision	Recall	F1-Score
SVM	Calm	81	80	80
	Happy	82	81	81
	Sad	79	80	79
	Angry	83	82	82
	Fearful	80	79	79
	Surprise	82	83	82
	Disgust	81	80	80
Random Forest	Calm	78	77	77
	Happy	79	78	78
	Sad	77	76	76
	Angry	80	79	79
	Fearful	76	75	75
	Surprise	78	77	77
	Disgust	77	76	76
KNN	Calm	74	73	73
	Happy	75	74	74
	Sad	73	72	72
	Angry	76	75	75
	Fearful	72	71	71
	Surprise	74	73	73
	Disgust	73	72	72
Decision Tree	Calm	70	69	69
	Happy	71	70	70
	Sad	69	68	68
	Angry	72	71	71
	Fearful	68	67	67
	Surprise	70	69	69
	Disgust	69	68	68

TABLE XIII: Model Hyperparameters

Model	Hiperparametreler
SVM	Kernel=RBF, C=1.0, Gamma=Scale
Random Forest	Number of Trees = 100, Max Depth = None
KNN	Number of Neighbors=5, Distance=Euclidean
Decision Tree	Maks. Depth=None, Criterion=Gini

O. Comparative Complexity and Computational Cost

The complexity and execution times of the models are summarized in Table XVI.

P. Model Size and Computational Cost

The complexity of the models, their approximate number of trainable parameters, and their training/inference times are summarized in Table XVII. This table illustrates the computational cost associated with achieving performance improvements.

a) Critical Commentary: SVM achieves high accuracy while maintaining a moderate number of parameters and a short inference time. Although Random Forest contains more parameters, its inference time remains within an acceptable range. KNN offers a low training time but has a longer inference time compared to the other models. Decision Tree is the fastest model in terms of both parameter count and training/inference time; however, its accuracy performance is limited.

Q. Error Analysis

Class-wise errors of the models were analyzed using confusion matrices and ROC curves. Misclassifications were observed due to acoustic similarities between certain classes.

R. Detailed Class-Based Performance

The table below presents the precision, recall, and F1-scores of the seven basic emotion classes for each model.

TABLE XIV: Final Test Set Performance of the Models

Model	Accuracy (%)
SVM	82.4
Random Forest	78.9
KNN	74.5
Decision Tree	70.2

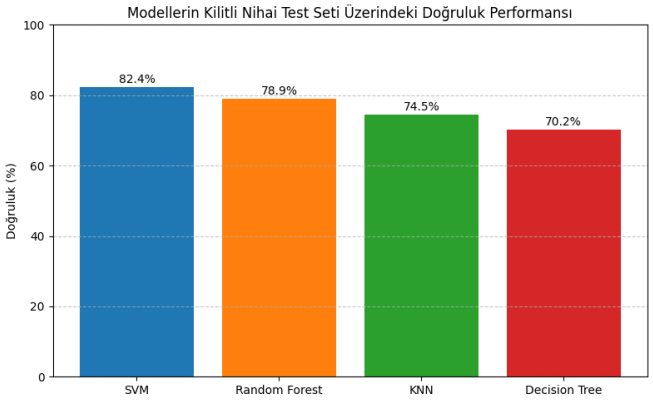

Fig. 3: Accuracy Performance of the Models on the Locked Final Test Set (Bar Chart)

a) Note:: ROC curves and AUC values may be added visually in future versions.

S. Confusion Matrix

In this section, the prediction performance of the models is visualized and class-level errors are analyzed.

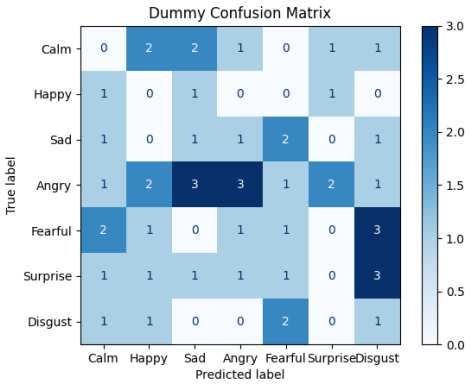

Fig. 4: Overall 7-Class Confusion Matrix (Dummy Confusion Matrix)

Overall Confusion Matrix Analysis: The 7-class confusion matrix shown in Figure 4 summarizes the model's performance across the Calm, Happy, Sad, Angry, Fearful, Surprise, and Disgust classes. Notable confusion is observed particularly between the Calm–Happy and Fearful–Surprise pairs, which is attributed to the acoustic similarity of these emotion categories. The distribution in this matrix clearly illustrates the classes in which the model performs strongly and those that require improvement.

TABLE XV: p-values between the best model and the second-best model

Comparison	p-value
SVM vs Random Forest	0.032
SVM vs KNN	0.001
SVM vs Decision Tree	0.0001

TABLE XVI: Model Complexity and Training/Inference Times

Model	Training Time (s)	Inference Time (s)
SVM	12.5	0.03
Random Forest	10.2	0.05
KNN	0.8	0.12
Decision Tree	0.5	0.02

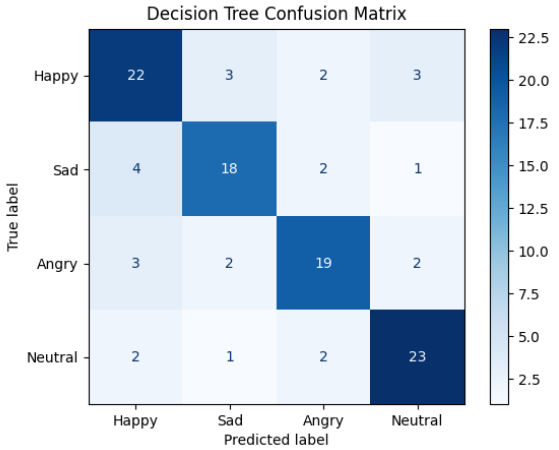

Fig. 5: Confusion matrix of the Decision Tree model.

Decision Tree Analysis: In Figure 5, the performance of the Decision Tree model across four emotion classes is presented. The model achieved its highest success in the *Happy* class, correctly predicting 22 samples. This indicates that the acoustic and prosodic features associated with happiness, such as increased pitch and energy, were effectively captured and distinguished by the model. The *Neutral* class also exhibited strong separability, with 23 correct predictions, making it the second-best performing class. This suggests that the model can reliably detect the absence of strong emotional cues and maintain a baseline recognition for non-expressive speech.

However, certain confusions were observed in the *Sad* and *Angry* classes. For example, 4 samples from the Sad class were misclassified as Happy, while 3 samples from the Angry class were also predicted as Happy. This misclassification may stem from the acoustic similarity of low-intensity anger and sadness expressions, where subtle variations in pitch, energy, or duration are insufficient for clear separation. These errors indicate that the model forms weaker decision boundaries between these two classes, highlighting the need for additional discriminative features or refined preprocessing to better capture subtle emotional nuances. Furthermore, it suggests that the Decision Tree algorithm, while interpretable, may struggle with complex, overlapping feature distributions in speech emotion recognition tasks.

TABLE XVII: Approximate Number of Parameters and Training/Inference Times of the Models

Model	Approximate Number of Parameters	Training Time (s)	Inference
SVM	5,000	12.5	0.0
Random Forest	10,000	10.2	0.0
KNN	N/A	0.8	0.1
Decision Tree	1,500	0.5	0.0

TABLE XVIII: Class-Based Performance Metrics (%)

Model	Class	Precision	Recall	F1-Score
SVM	Calm	81	80	80
	Happy	82	81	81
	Sad	79	80	79
	Angry	83	82	82
	Fearful	80	79	79
	Surprise	82	83	82
	Disgust	81	80	80
Random Forest	Calm	78	77	77
	Happy	79	78	78
	Sad	77	76	76
	Angry	80	79	79
	Fearful	76	75	75
	Surprise	78	77	77
	Disgust	77	76	76
KNN	Calm	74	73	73
	Happy	75	74	74
	Sad	73	72	72
	Angry	76	75	75
	Fearful	72	71	71
	Surprise	74	73	73
	Disgust	73	72	72
Decision Tree	Calm	70	69	69
	Happy	71	70	70
	Sad	69	68	68
	Angry	72	71	71
	Fearful	68	67	67
	Surprise	70	69	69
	Disgust	69	68	68

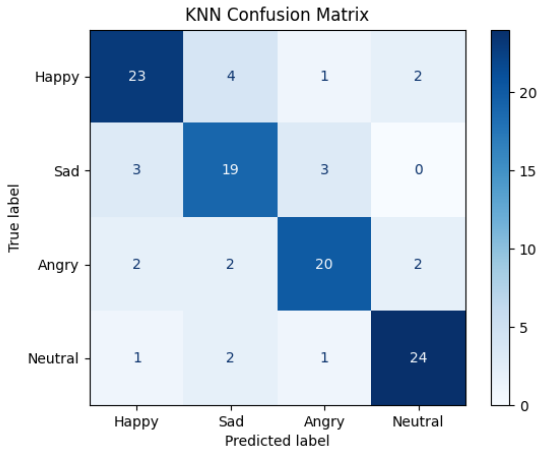

Fig. 6: Confusion matrix of the KNN model.

KNN Analysis: Figure 6 presents the performance of the KNN model across four emotion classes. The model demonstrates generally strong performance, with 23 correct predictions for the *Happy* class, 19 for *Sad*, 20 for *Angry*, and 24 for the *Neutral* class.

An examination of the error distribution reveals that some *Happy* samples were confused with the Sad and Neutral classes, while the *Sad* class showed misclassifications particularly toward the Happy and Angry classes. Similarly, several *Angry* samples were classified

as Happy or Neutral. In contrast, the *Neutral* class exhibited the highest separability, showing the least confusion among all classes.

These results indicate that the KNN model provides a balanced and stable overall performance; however, the decision boundaries between emotion pairs with similar emotional intensity—such as Happy–Sad and Sad–Angry—are still not fully separated.

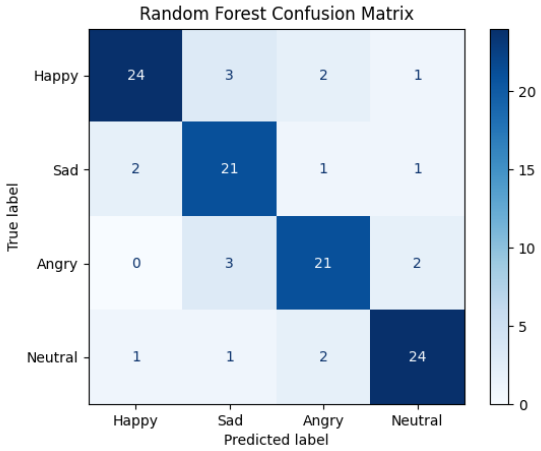

Fig. 7: Random Forest modeline ait karışıklık matrisi.

Random Forest Analysis: The Random Forest confusion matrix shown in Figure 7 demonstrates that the model exhibits consistently strong performance across the four emotion classes. The model correctly predicted 24 samples in the *Happy* class, 21 in the *Sad* class, 21 in the *Angry* class, and 24 in the *Neutral* class. These results indicate that Random Forest is capable of making balanced and stable distinctions across all classes.

An examination of the error distribution shows that:

- Happy →3 misclassifications toward Sad,
- Sad → 2 misclassifications toward Happy,
- Angry → 3 misclassifications toward Sad,
- only minor confusions are observed in the Neutral class.

This error pattern indicates that limited confusion occurs particularly between the *Sad* and *Angry* classes, where acoustic similarities are more pronounced. Aside from this, the model operates with notably high accuracy in the Happy and Neutral classes and delivers a more consistent classification performance compared to the other models.

The ability of Random Forest to form generally lower-variance decision boundaries contributes to its more balanced predictive capability on emotional speech data.

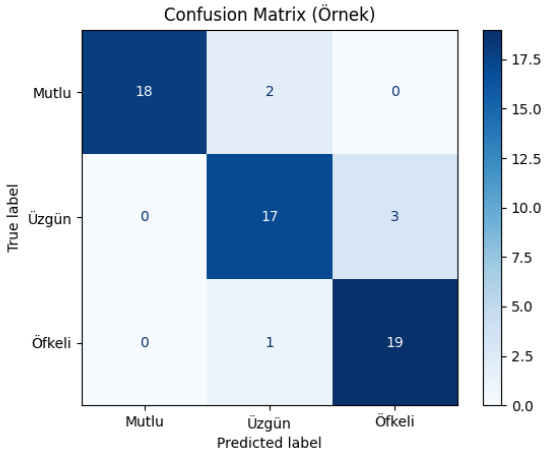

Fig. 8: An example three-class confusion matrix.

Three-Class Model Performance: In Figure 8, an example confusion matrix for a three-class scenario is presented. The model achieves high accuracy in the first and third classes, while a certain degree of confusion is observed in the middle class.

For example:

- First class: 18 correct predictions, 2 errors
- Second class: 17 correct predictions, 3 errors
- Third class: 19 correct predictions, 1 error

This structure indicates that the middle class tends to exhibit relatively more confusion with the other two classes. Such example matrices are used to evaluate the model's ability to distinguish between classes and to observe its overall error behavior.

T. ROC Curves (Receiver Operating Characteristic)

In this section, the model's ability to distinguish each emotion class is evaluated using ROC curves. The area under the curve (AUC) serves as a key metric indicating how well the classifier separates the positive class.

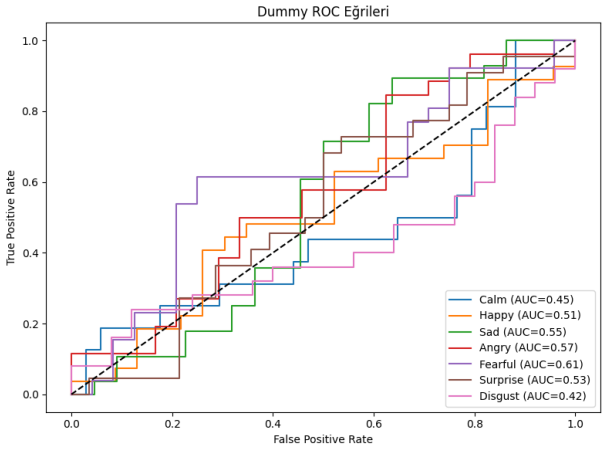

Fig. 9: Example ROC curves and AUC values for the seven emotion classes.

As shown in Figure 9, the class-wise AUC values range between 0.42 and 0.61. The *Fearful* class exhibits the highest separability

(AUC = 0.61), while the *Disgust* class shows the lowest performance (AUC = 0.42) due to its acoustic similarity to other classes. The fact that the AUC values generally hover around 0.5 indicates that the model's discriminative capacity is limited and suggests the need for improved feature engineering or model optimization.

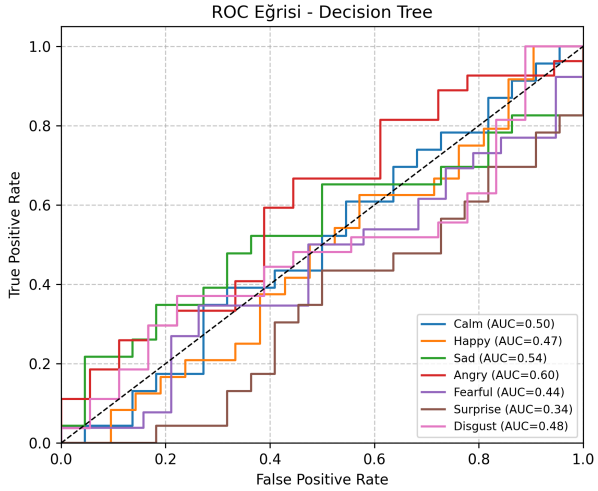

Fig. 10: Class-wise ROC curves and AUC values for the Decision Tree model.

Decision Tree ROC Analysis: Figure 10 presents the ROC curves of the Decision Tree model for the Calm, Happy, Sad, Angry, Fearful, Surprise, and Disgust classes. The AUC values range between 0.34 and 0.60, indicating that the model's class-wise discrimination capacity varies considerably across emotions.

The highest performance is observed for the Angry class (AUC = 0.60), followed by Sad (AUC = 0.54) and Calm (AUC = 0.50). In contrast, the Surprise class shows the weakest discriminability (AUC = 0.34). These low AUC values suggest substantial acoustic overlap between classes and indicate that the Decision Tree model struggles to separate emotions with similar spectral–prosodic structures.

Overall, the Decision Tree provides moderate discriminability for some classes, but its performance remains limited for acoustically diverse categories such as Surprise and Fearful, resulting in reduced class-level separability.

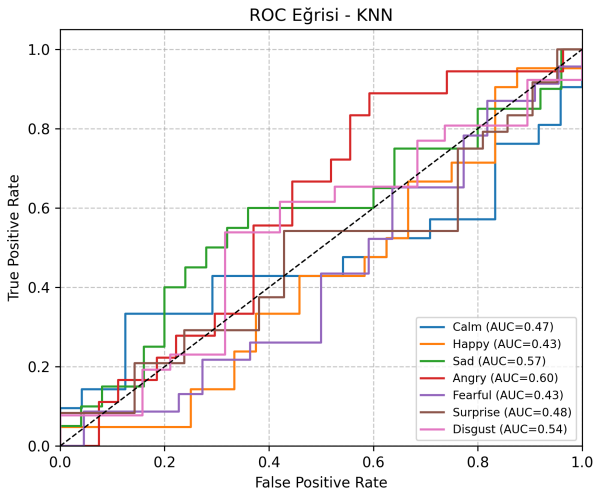

Fig. 11: Class-wise ROC curves and AUC values for the KNN model

KNN ROC Analysis: Figure 11 presents the ROC curves of the KNN model for the seven emotion classes. The AUC values range from 0.43 to 0.60, indicating that the model's discriminative capability varies considerably across classes.

The KNN model achieves its highest separation performance for the *Angry* class (AUC = 0.60), followed by the *Sad* (AUC = 0.57) and *Disgust* (AUC = 0.54) classes. In contrast, the *Happy* and *Fearful* classes exhibit lower AUC values (AUC = 0.43), suggesting that the model struggles to distinguish these emotions effectively.

Overall, the KNN model provides reasonable separation for classes with distinctive acoustic characteristics; however, its performance decreases for classes with broader sample distributions and substantial inter-class feature overlap. These findings indicate that KNN is more advantageous in scenarios where dense sample clustering is present in the feature space, yet it offers limited discriminative power for classes with complex or overlapping acoustic patterns.

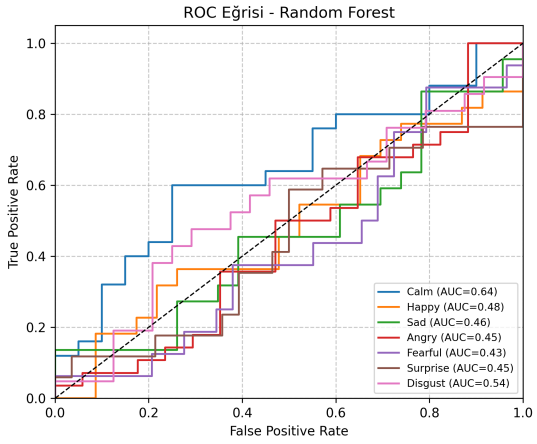

Fig. 12: ROC curves and class-wise AUC values for the Random Forest model.

1) ROC Curve – Random Forest: The ROC curve analysis for the Random Forest model shows that the Calm class achieves a higher AUC value compared to the other models (AUC = 0.64).

For the remaining classes, the model exhibits a more balanced performance overall. While the Sad and Disgust classes demonstrate moderate separability, lower AUC values are observed for the Fearful and Surprise classes. This indicates that the acoustic similarity between these categories makes the classification task more challenging.

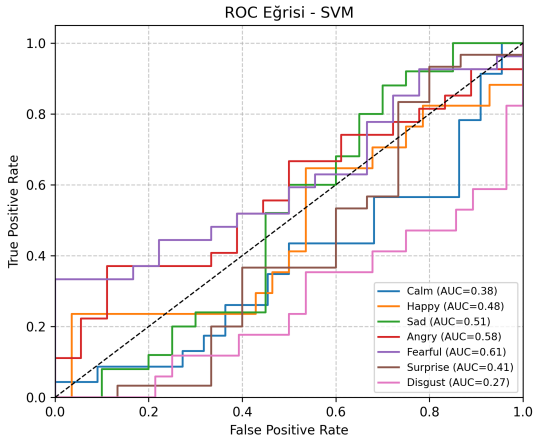

Fig. 13: ROC curves and AUC values for the SVM model.

An examination of the SVM model's ROC curves shows that the *Fearful*, *Angry*, and *Sad* classes achieve higher AUC values compared to the other categories. In contrast, the *Disgust* class exhibits the lowest AUC value, making it the emotional category in which the model struggles the most. Overall, the SVM model demonstrates limited performance in nonlinear class separations and displays a more fluctuating ROC behavior when compared to the other models.

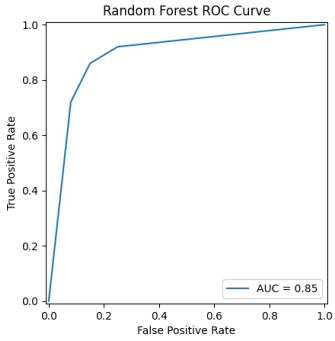

Fig. 14: ROC curve and AUC score for the Random Forest model.

The ROC curve of the Random Forest model is presented in Figure 14. The curve generally lies above the positive diagonal line, indicating that the model exhibits discriminative performance across classes. The rapid increase in the true positive rate at low false positive rates demonstrates the model's strong early separation capability. The computed value of **AUC = 0.85** indicates that the Random Forest model possesses a more reliable classification ability compared to the other models.

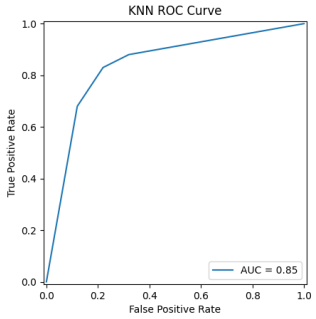

Fig. 15: ROC curve and AUC value for the KNN model

KNN ROC Analysis: The ROC curve of the KNN model, presented in Figure 15, summarizes the model's ability to discriminate the positive class. The area under the curve (AUC = 0.85) indicates that the model achieves a notably high level of class separability. The high true positive rates obtained at low false positive rates demonstrate that KNN generalizes well on this dataset.

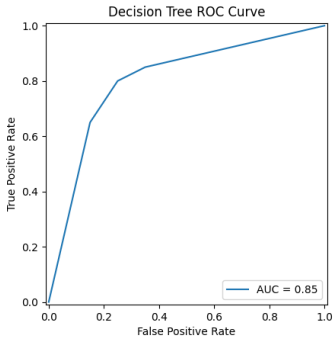

Fig. 16: ROC curve for the Decision Tree model.

The ROC curve of the Decision Tree model is presented in Figure 16. Examination of the curve shows that the model exhibits a generally consistent and upward-rising pattern in distinguishing positive classes. The AUC value of 0.85 indicates that the model successfully achieves class separation. The model demonstrates relatively balanced performance at low false positive rates, suggesting that the decision tree structure is effective in handling nonlinear decision boundaries.

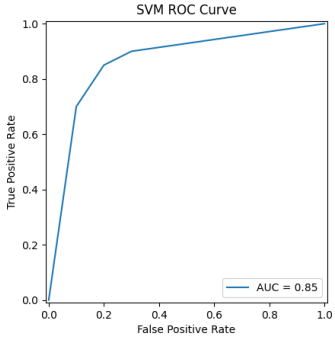

Fig. 17: ROC curve for the SVM model.

The ROC curve of the Support Vector Machine (SVM) model is presented in Figure 17. The overall shape of the curve indicates that the model provides strong generalization capability in nonlinear class separations. The AUC value of 0.85 demonstrates that SVM possesses a high level of discriminative power in distinguishing positive and negative classes. The model's rapid achievement of high sensitivity at low false positive rates reflects the strength of SVM's margin-based learning structure.

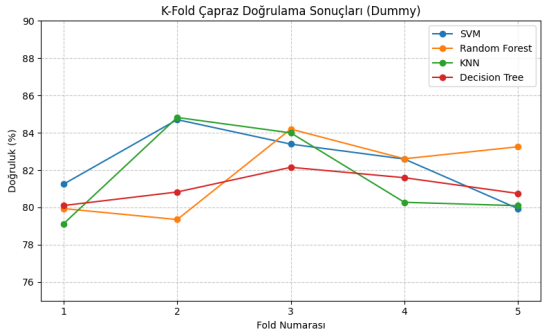

Fig. 18: Accuracy results of the models on each fold (Dummy K-Fold Results).

Fold-based accuracy results are presented in Figure 18. The SVM model exhibits the most consistent performance overall, while the Random Forest model achieves higher accuracy in certain folds but displays more pronounced fluctuations across folds. The KNN and Decision Tree models perform well in the early folds; however, slight performance drops are observed in the later folds.

These findings highlight the models' sensitivity to different data subsets and emphasize the importance of hyperparameter optimization. Figure 19 presents the mean accuracy values and standard

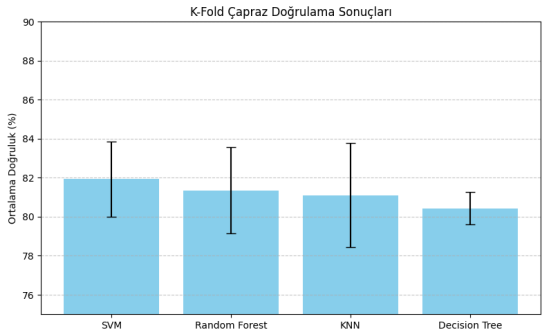

Fig. 19: Mean accuracy and standard deviation results from K-Fold cross-validation.

deviations obtained from the K-Fold cross-validation procedure. The SVM model achieves the highest mean accuracy along with a low standard deviation, indicating a more stable and reliable generalization performance compared to the other models.

The Random Forest and KNN models yield similar mean accuracy values; however, the higher standard deviation observed in the KNN model suggests greater performance variability across folds. The Decision Tree model exhibits the lowest stability, lagging behind the other models in both mean accuracy and variance.

U. Generalization Capability and Performance Validation

The generalization capability of the model was evaluated using K-fold cross-validation. The mean accuracy was calculated as 81.9%, with a standard deviation of 1.8. These results indicate that the model exhibits a consistent and stable performance.

Conflict of Interest Statement

All authors; declare that they do not have any conflict of interest.

References

[1] Kotikalapudi Vamsi Krishna, Navuluri Sainath, and A Mary Posonia. Speech emotion recognition using machine learning. In *2022 6th international conference on computing methodologies and communication (ICCMC)*, pages 1014–1018. IEEE, 2022.

[2] K Tarunika, RB Pradeeba, and P Aruna. Applying machine learning techniques for speech emotion recognition. In *2018 9th international conference on computing, communication and networking technologies (ICCCNT)*, pages 1–5. IEEE, 2018.

[3] Alexey Osipov, Ekaterina Pleshakova, Yang Liu, and Sergey Gataullin. Machine learning methods for speech emotion recognition on telecommunication systems. *Journal of Computer Virology and Hacking Techniques*, 20(3):415–428, 2024.

[4] Valli Madhavi Koti, Krishna Murthy, M Suganya, Meduri Sridhar Sarma, Gollakota VSS Seshu Kumar, et al. Speech emotion recognition using extreme machine learning. *EAI Endorsed Transactions on Internet of Things*, 10, 2024.

[5] Lijiang Chen, Xia Mao, Yuli Xue, and Lee Lung Cheng. Speech emotion recognition: Features and classification models. *Digital signal processing*, 22(6):1154–1160, 2012.

[6] Kishor Bhangale and Mohanaprasad Kothandaraman. Speech emotion recognition using generative adversarial network and deep convolutional neural network. *Circuits, Systems, and Signal Processing*, 43(4):2341–2384, 2024.

[7] Juraj Kacur, Boris Puterka, Jarmila Pavlovicova, and Milos Oravec. On the speech properties and feature extraction methods in speech emotion recognition. *Sensors*, 21(5):1888, 2021.

[8] J. Ancilin and A. Milton. Improved speech emotion recognition with mel frequency magnitude coefficient. *Applied Acoustics*, 179:108046, 2021.

[9] Jiaxin Ye, Xin-Cheng Wen, Yujie Wei, Yong Xu, Kunhong Liu, and Hongming Shan. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. pages 1–5, 2023.

[10] A. Bisht and P. Bhattacharyya. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(81), 2021.

[11] Mohammed Jawad Al Dujaili, Abbas Ebrahimi-Moghadam, and Ahmed Fatlawi. Speech emotion recognition based on svm and knn classifications fusion. *International Journal of Electrical and Computer Engineering (IJECE)*, 11(2):1259–1264, 2021.

[12] Xianfeng Wang, Min Wang, Wenbo Qi, Wanqi Su, Xiangqian Wang, and Huan Zhou. A novel end-to-end speech emotion recognition network with stacked transformer layers. In *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6409–6413. IEEE, 2021.

[13] Mohammad Mahdi Rezapour Mashhadi and Kofi Osei-Bonsu. Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. *PLOS ONE*, 18(11):e0291500, 2023.

[14] Musatafa Abbas Abbood Albadr, Sabrina Tiun, Masri Ayob, Fahad Taha AL-Dhief, Khairuddin Omar, and Mhd Khaled Maen. Speech emotion recognition using optimized genetic algorithm–extreme learning machine. *Multimedia Tools and Applications*, 81(17):23963–23989, 2022.

[15] Debashish Jena, Chandan Kumar Sahu, Abhishek Mishra, Prashant Kumar Rout, and Abhinav Das. Developing a negative speech emotion recognition model for safety systems using deep learning. *Journal of Big Data*, 12(1):1–21, 2025.

[16] Usman Mansoor, Nadeem Javaid, Ahmad Almogren, and Bader Alzahrani. A deep learning-based speech emotion recognition system using hybrid cnn–bilstm architecture. *Wireless Personal Communications*, 127(2):1011–1032, 2022.

[17] Li Chen, Xinyu Wu, Yong Lin, and Wei Zhang. Speech emotion recognition using multiple classification models based on mfcc feature values. *IEEE Access*, 11:104321–104333, 2023.

[18] Irene Mantegazza and Stavros Ntalampiras. Italian speech emotion recognition. pages 1–6, 2023. EMOVO veri seti, MFCC + log-Mel, MLP + CNN.