Speech Emotion Recognition Using ML Models and Audio Features

Tytiana James
School of Computing and Informatics
University of Louisiana at Lafayette
Lafayette, USA
tytiana.james1@louisiana.edu

T. Maheswara Reddy Yenumula
School of Computing and Informatics
University of Louisiana at Lafayette
Lafayette, USA
C00538416@louisiana.edu

Tanner Mergist
School of Computing and Informatics
University of Louisiana at Lafayette
Lafayette, USA
tanner.mergist1@louisiana.edu

Rafeeq Muhammad
School of Computing and Informatics
University of Louisiana at Lafayette
Lafayette, USA
rafeeq.muhammad1@louisiana.edu

Ali Harimi
School of Computing and Informatics
University of Louisiana at Lafayette
Lafayette, USA
ali.harimi1@louisiana.edu

Kasem Khalil
Electrical and Computer Engineering
University of Mississippi
Mississippi, USA
kmkhalil@olemiss.edu

Ashok Kumar
School of Computing and Informatics
University of Louisiana at Lafayette
Lafayette, USA
ashok.kumar@louisiana.edu

2025 IEEE International systems Conference (SysCon) | 979-8-3315-0818-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/SYSCON64521.2025.11014858

***Abstract*—The ability to recognize and respond to human emotions has become a crucial component in enhancing human-computer interaction, with the development of speech emotion recognition technologies playing a pivotal role in this advancement. While deep learning models have driven significant progress in the field of speech emotion recognition, traditional machine learning algorithms offer a practical alternative, balancing performance with lower computational requirements. This study presents a comprehensive approach to speech emotion recognition using the Berlin Emotional Speech Database, classifying seven emotional states: anger, sadness, anxiety or fear, disgust, boredom, happiness, and neutrality. The study employs a range of acoustic features, including pitch, RMS energy, MFCCs, and formants, which are combined with 14 statistical descriptors and extracted using tools like OpenSMILE. Preprocessing steps, such as normalization, noise reduction, and silence removal, are applied to enhance the quality and reliability of the data. The performance of traditional machine learning models, including Support Vector Machine, Random Forest, and k-Nearest Neighbors, is evaluated on the processed dataset. The results demonstrate the effectiveness of these traditional models, with Support Vector Machine achieving the highest classification accuracy of 90.65%, followed by Random Forest and k-Nearest Neighbors. The results of this study highlight the capacity of traditional machine learning techniques to effectively capture the complexities of emotional expression, while circumventing the computational burden associated with deep learning models. The practical relevance of this research extends to real-time applications across various domains, including healthcare, virtual assistants, and customer service, where the demand for efficient and reliable emotion recognition systems is paramount.**

***Index Terms*—Speech Emotion Recognition, Emotion Classification, Machine Learning, Audio Features**

I. Introduction

Recognizing and responding to human emotions has become a crucial component in enhancing human-computer interaction, enabling systems to comprehend, interpret, and adapt to users' emotional states. This capability enhances the naturalness, engagement, and personalization of various applications, including virtual assistants, telecommunications, healthcare, and customer service, where accurately detecting emotional cues can improve user satisfaction and service quality. By integrating emotion-aware mechanisms, Speech Emotion Recognition bridges the gap between technology and human-centric experiences, fostering more intuitive and context-aware interfaces.

Despite its growing importance, emotion recognition remains a challenging task due to the subtle and dynamic nature of emotional expression in speech. Variations in pitch, energy, and speech patterns are often nuanced and highly dependent on linguistic, cultural, and contextual factors, making it difficult to consistently extract discriminative emotional features. While deep learning approaches have demonstrated significant improvements in performance, they often come with high computational costs, limiting their feasibility for real-time and resource-constrained applications. This trade-off between accuracy and efficiency poses a critical bottleneck in deploying scalable and accessible solutions across diverse real-world scenarios.

To address these limitations, this study investigates traditional machine learning techniques as a computationally efficient alternative to deep learning. Using engineered acoustic features,

including pitch, Mel-frequency cepstral coefficients, root mean square, energy, and formants, in addition to optimized ML classifiers, this approach seeks to achieve a balance between accuracy and efficiency. Unlike deep learning models, which require extensive training data and computational resources, traditional ML methods offer interpretability, lower computational overhead, and adaptability to real-time environments.

The primary objective of this study is to classify seven emotional states—anger, sadness, anxiety/fear, disgust, boredom, happiness, and neutrality—using the Berlin Emotional Speech Database. The methodology follows a structured pipeline, including preprocessing for noise reduction, feature extraction from acoustic properties, and classification using Support Vector Machines, Random Forest, and k-Nearest Neighbors. Feature engineering plays a pivotal role in this process, as carefully designed features can effectively capture the unique characteristics of emotional speech, directly influencing model accuracy and robustness.

Through this study, we demonstrate the effectiveness of traditional machine learning models in emotion recognition tasks, emphasizing a balanced approach that achieves high accuracy without compromising computational efficiency. Our findings aim to provide insights for future advancements in SER, particularly in applications that require real-time emotional detection and responsive human-computer interfaces.

II. Related Work

Over the years, the field of Speech Emotion Recognition has experienced remarkable progress, transitioning from systems capable of recognizing only isolated words to advanced, continuous recognition models that can operate independently of the speaker and accommodate large vocabularies. This evolution underscores the growing need for systems that can comprehend not only the spoken words but also the underlying emotional states. Despite these advancements, Speech Emotion Recognition continues to face significant challenges, including the confounding factors of background noise, cultural and environmental differences, and variations in individual speech characteristics, rendering accurate emotion detection a persistent challenge [1][2].

To overcome these challenges, deep learning methods have significantly enhanced SER performance, particularly on large-scale and complex datasets [3][4]. For instance, Yuan et al. developed a model that leverages adversarial learning to isolate emotional information from speaker-specific characteristics, enabling it to generalize more effectively across diverse speakers and languages [5].In one study, Idoko compared CNN and LSTM models for SER, using Mel-Frequency Cepstral Coefficients (MFCCs) and wavelet-based features. CNN outperformed LSTM, achieving 61% accuracy compared to LSTM's 56% [6]. In another study Pham et.al.,[7] implemented CNN and achieved 76% accuracy categorizing seven emotions. These results highlight the potential of deep learning architectures in capturing intricate emotional patterns; however, the computational complexity of such models remains a major concern, especially for real-time and resource-constrained applications. Furthermore, researchers have incorporated Automatic Speech Recognition (ASR) models to enhance SER performance, particularly in cases where data scarcity limits deep learning models' effectiveness. Self-supervised learning frameworks, such as Wav2Vec 2.0, have been adapted to improve SER accuracy and robustness in real-world scenarios [8][4]. Additionally, architectures like CNNs, Deep CNNs (DCNNs), and Recurrent Neural Networks (RNNs) excel in extracting spatial, spectral, and temporal features from speech signals, allowing models to learn local and global patterns crucial for emotion classification [10][11]. Feature selection plays a pivotal role in SER, as the choice of features directly affects model accuracy and robustness. Sakurai and Kosaka demonstrated the advantages of combining acoustic and linguistic features, while others have explored multi-level fusion techniques to integrate emotional characteristics across different feature sets [9][12]. Furthermore, pitch fusion-based models have proven particularly effective for tonal languages, where pitch variations encode distinct emotional cues, as demonstrated in Thanh et al.'s study on Vietnamese speech datasets [13].

Although researchers have explored model compression, quantization, and knowledge distillation to reduce the computational cost of deep learning models, these approaches often trade off accuracy or require additional fine-tuning. Consequently, there remains a need for lightweight and interpretable models that balance computational efficiency with classification performance. This study, therefore, focuses on traditional machine learning models that prioritize efficiency and interpretability, making SER feasible for settings with limited computational resources. By utilizing the Berlin Emotional Speech Database (EmoDB) and targeting essential acoustic features, this research seeks to find a balance between model simplicity and accuracy, ultimately broadening SER's applicability in real-world, practical scenarios.

III. Methodology

This research explores the use of traditional machine learning algorithms to classify diverse emotional states from audio speech samples in the domain of Speech Emotion Recognition. The methodology involves four main stages: dataset selection, preprocessing, feature extraction, and classification. For this study, we utilized the Berlin Emotional Speech Database (EmoDB), applying SMO (SVM), Random Forest, and k-Nearest Neighbors (k-NN) algorithms for classification. Each stage is elaborated upon below.

A. Dataset

The Berlin Emotional Speech Database (EmoDB) [14] was selected for its clear emotional annotations and high-quality recordings. This dataset includes 535 speech samples that cover seven emotional categories: anger (127 samples), boredom (81), fear/anxiety (69), happiness (71), sadness (62), disgust (46), and neutrality (79). The recordings were made by 10 professional speakers (five male and five female), ensuring expressive vocal quality. Although the controlled

nature of EmoDB may limit generalization, it provides a solid foundation for evaluating our SER models in a standardized setting.

B. Pre-Processing

The preprocessing phase prepares raw audio files for analysis, enhancing the accuracy of feature extraction and classification. Each audio sample underwent the following steps:

- Normalization: We standardized the volume across all samples to ensure that variations in loudness did not introduce bias, allowing a consistent focus on emotional content.
- Silence Removal: Non-speech segments and minor background noise were removed, refining the data to focus on voiced regions that are critical for emotion identification.
- Padding: For consistency, zero-padding was applied to the end of shorter audio samples. This step ensured that each file had a uniform length, facilitating efficient processing and model training.
- Noise Reduction: Background noise was minimized to improve clarity, ensuring that emotional cues were preserved and pronounced in each audio sample.

These preprocessing steps collectively enhanced data quality, setting a reliable base for the subsequent feature extraction and classification stages.

C. Feature Extraction

The core of SER lies in capturing nuanced acoustic features that represent different emotional expressions. For this study, we employed the OpenSMILE, Librosa, and Parselmouth libraries to extract a comprehensive set of 238 features, focusing on essential acoustic and statistical properties that contribute to emotion recognition.

Four primary types of features were extracted:

- Pitch (Fundamental Frequency): Pitch, as a key acoustic feature, reflects emotional intensity. For instance, high-pitch frequencies often indicate emotions like happiness or anger, while lower pitch may signal sadness.
- Root Mean Square (RMS) Energy: RMS energy provides insight into the loudness of speech, which correlates with emotional arousal. Higher energy levels are typical in intense emotions, such as anger, aiding in understanding emotional intensity.
- Mel-Frequency Cepstral Coefficients (MFCCs 1–12): MFCCs are widely recognized in audio processing for capturing short-term spectral nuances, making them essential for identifying vocal subtleties across emotions.
- Formants (F1, F2, F3): Formants represent vocal tract resonances that shift with different emotional states. These resonances are critical for capturing emotional expressions embedded in speech.

To deepen analysis, we calculated 14 statistical descriptors for each feature, including metrics such as maximum, minimum, range, mean, RMS, standard deviation, skewness, kurtosis, percentiles, and inter-quartile range. These descriptors capture the variability of features, providing a comprehensive view of the emotional characteristics present in the dataset.

D. Classification

Following preprocessing and feature extraction, we utilized three machine learning models—SMO (SVM), Random Forest, and k-NN—chosen for their unique capabilities in handling different aspects of the data. Each model is described below:

- Support Vector Machine (SVM): We applied SVM with the Sequential Minimal Optimization (SMO) algorithm, which efficiently handles high-dimensional data. This model is particularly effective for distinguishing emotions that have distinct acoustic profiles. Formants were excluded from this model to reduce complexity, allowing faster computation without compromising classification accuracy.
- Random Forest: The Random Forest algorithm combines multiple decision trees trained on random subsets of data. This ensemble approach reduces the likelihood of over-fitting and provides insights into feature importance, which is valuable for interpreting emotional patterns. Hyper parameters, including the number and depth of trees, were tuned to balance accuracy and computational efficiency.
- k-Nearest Neighbors (k-NN): The k-NN algorithm classifies samples based on their proximity to labeled neighbors, which works well in SER, where similar emotions cluster in feature space. Optimal performance was achieved by tuning the number of neighbors (k) to enhance classification accuracy based on local patterns.

Each model was fine-tuned using cross-validation to identify the most effective hyperparameters, thereby optimizing model accuracy and performance. This approach allowed us to leverage each model's strengths, achieving reliable and robust emotion classification across the EmoDB dataset. By employing hyper-parameter selection and tuning using CVParameterSelection in WEKA, accuracy, precision, and recall were higher across all models with tuning than without tuning.

IV. Results

To assess the effectiveness of our SER approach, we evaluated the Support Vector Machine (SVM), Random Forest, and k-Nearest Neighbors (k-NN) models based on their performance in classifying the seven emotions within the EmoDB dataset. We used several metrics, including accuracy, precision, recall, and F-measure, to quantify each model's effectiveness. Precision reflects the accuracy of positive predictions. Higher precision indicates fewer misclassified emotions. In other words, it measures how many of the predicted instances of an emotion is actually correct. The calculation to compute this is:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

. Recall indicates the model's ability to identify actual positives. Higher recall indicates fewer missed emotions. This

measures how many actual instances of an emotion are correctly identified The calculation to compute this is:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

. F-measure provides a balanced assessment by combining precision and recall. Higher F-measure values demonstrate the model's capability to maintain both accuracy and consistency in emotion recognition. Additionally, confusion matrices were created to reveal each model's specific strengths and areas for improvement across different emotions.

A. Model Performance Overview

All three models exhibited strong performance, although to varying extents. The SVM model demonstrated the highest classification accuracy, achieving an overall score of 90.65% and an F-measure exceeding 90%, underscoring its impressive capability in accurately distinguishing emotions across the comprehensive dataset. The Random Forest model closely followed with an accuracy of 87.48%, while the k-NN model attained 87.1%. These results suggest that the SVM model's exceptional capacity to handle high-dimensional data contributes to its distinct advantage in recognizing even the most subtle emotional variations within the dataset. Additionally, the execution times of these machine learning models do not exceed a mere 0.3 seconds, highlighting their efficiency and suitability for practical applications.

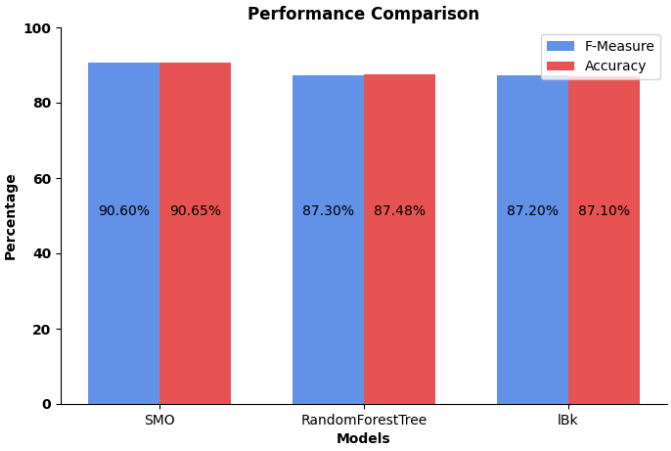

Fig. 1. A comparison of F-measure and Accuracy across each model

B. Precision and Recall Analysis by Emotion

A closer examination of precision and recall values for each emotion across the models, illustrated in Figures 2-4, reveals distinct trends in model performance:

SVM: This model performed well in accurately identifying emotions such as anger, neutrality, and sadness, suggesting that it effectively utilizes distinct acoustic features associated with these emotions. However, it encountered challenges with the anxiety/fear category, where recall was lower due to overlapping characteristics with sadness and neutrality. SVM's ability to manage high-dimensional data likely contributed to its strong performance on more distinct emotional states.

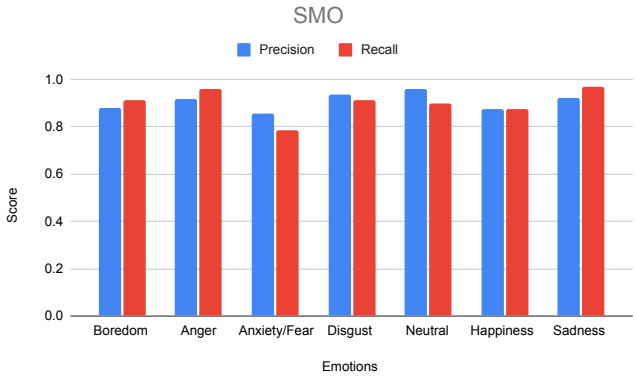

Fig. 2. Precision and Recall for each emotion using SVM model

Random Forest: Random Forest demonstrated strong precision and recall for emotions like happiness and neutrality, benefiting from its ensemble learning approach, which captures a broad range of emotional patterns.

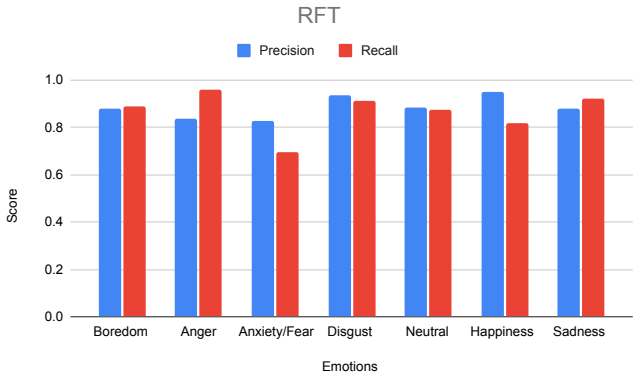

Fig. 3. Precision and Recall for each emotion using RFT model

However, similar to SVM, Random Forest struggled with emotions that shared similar acoustic properties, such as anxiety/fear and disgust share overlapping spectral features, particularly in pitch and energy variation, making them harder to differentiate. Random Forest, which relies on decision tree splits, may struggle when feature distributions are similar.

k-NN: The k-NN model performed well with emotions like sadness and neutrality but was more sensitive to noise, resulting in reduced accuracy for anxiety/fear. This sensitivity highlights k-NN's limitations due to its reliance on proximity-based classification, which makes emotions without distinct clusters in feature space more prone to misclassification. Although tuning the number of neighbors improved performance to some degree, k-NN remained vulnerable to subtle variations in audio features.

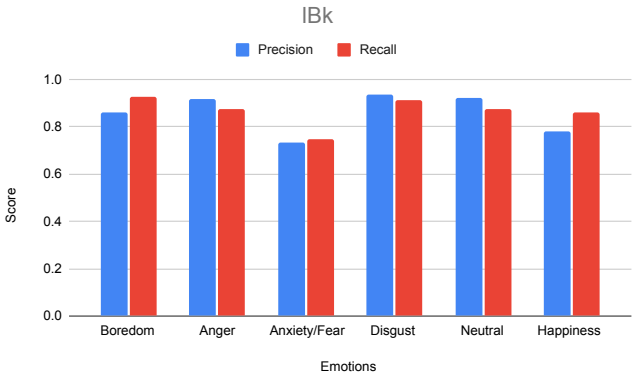

Fig. 4. Precision and Recall for each emotion using K-NN model

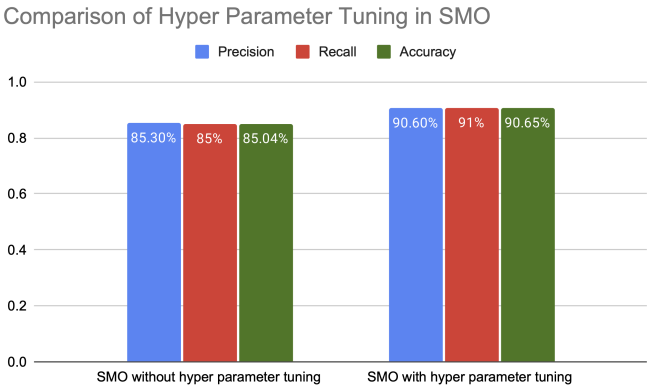

Fig. 5. SMO Hyper Parameter Tuning Comparison

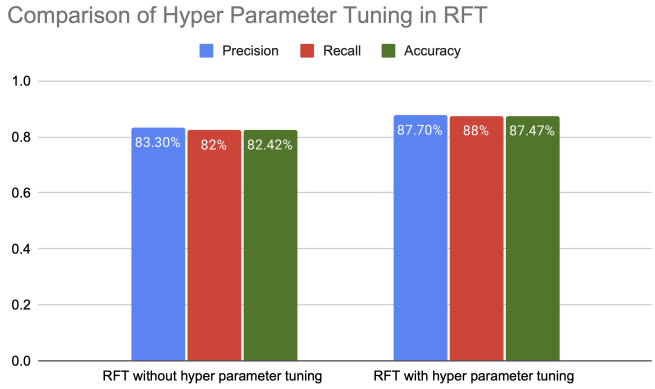

Fig. 6. RFT Hyper Parameter Tuning Comparison

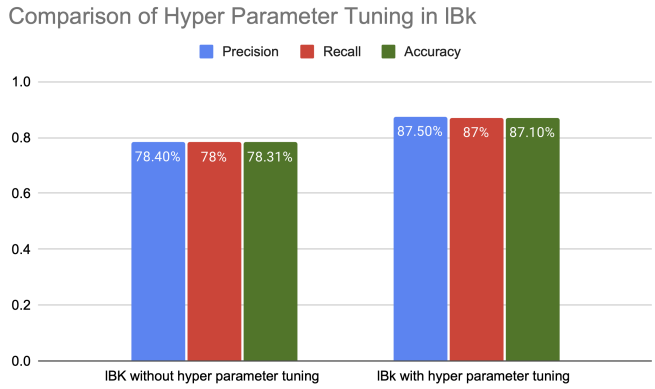

Fig. 7. K-NN Hyper Parameter Tuning Comparison

C. Hyperparameter Tuning Analysis

A comparison of these models shows that hyperparameter tuning using CVParameterSelection greatly benefited each model's accuracy, precision, and recall. Thus, leading to the most optimal performance for classifying emotions. The hyperparameter tuning allows the models to better fit the complex patterns in the data, reducing overfitting or underfitting.

- SMO: With the SMO model, all evaluation metrics increased roughly 5%. The parameters included: 10 folds for cross validation, 1 random seed, a complexity parameter set to 1.0, tolerance parameter set to 0.001, epsilon for loss function set to 1.0E-12, and the kernel type being Poly Kernel. Using CVParameterSelection, adjusting the RBF kernel and optimizing the C parameter allowed the model to effectively handle complex decision boundaries, while adjusting gamma balanced over and under fitting.
- RFT: With the RFT model, all evaluation metrics increased roughly 5% as well. The parameters included: 10 folds for cross validation, 1 random seed, and 100 trees. Tuning the number of tress and maximum depth improved classification and prevented over and under fitting.
- lBk (K-NN): lBk benefited from hyperparameter tuning more that SMO and RFT, having a roughly 9% increase across all evaluation metrics. The parameters included: using 10-folds for cross validation and 1- nearest neighbor. Optimizing the number of neighbors to classify emotions reduces sensitivity and produces better stability and generalization for data.

D. Confusion Matrix Analysis

The confusion matrices (Figures 5-8) offer additional insights into each model's strengths and common misclassifications across emotions:

- SVM: The confusion matrix for SVM shows high accuracy in classifying anger, neutrality, and sadness, with minimal misclassifications. However, anxiety/fear was frequently misclassified as either sadness or neutrality, reflecting the acoustic similarities among these emotions. This pattern aligns with SVM's strong ability to handle well-separated emotional categories.
- Random Forest: The Random Forest confusion matrix underscores its effectiveness in classifying happiness and neutrality. However, emotions such as anxiety/fear and disgust were more prone to misclassification, indicating that while Random Forest effectively captures general patterns, it may lack the sensitivity required to distinguish

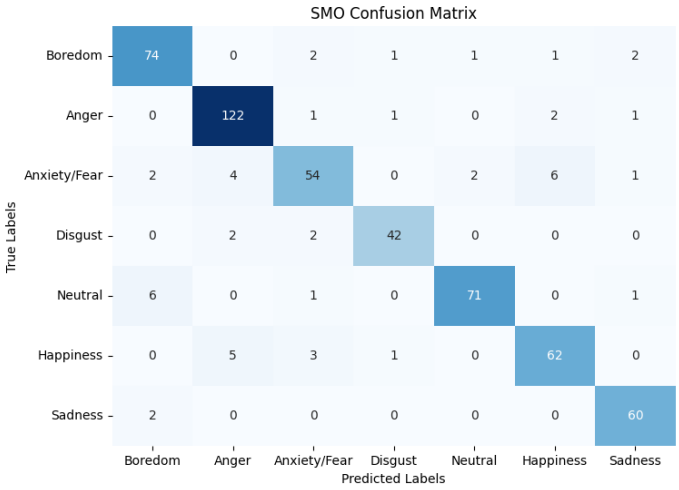

Fig. 8. SVM Confusion Matrix

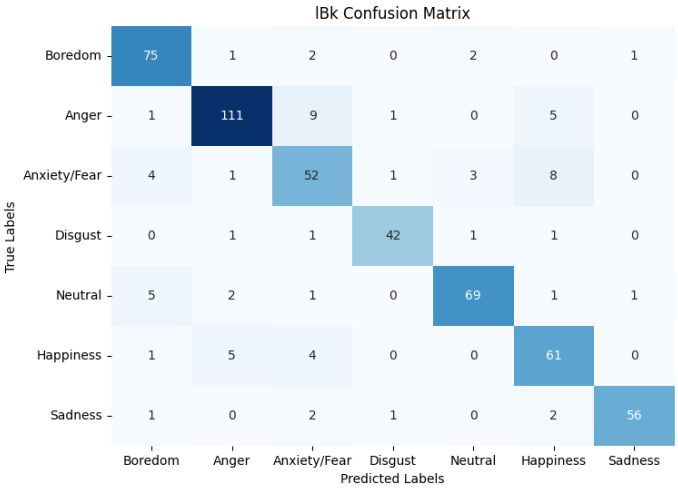

Fig. 10. K-NN Confusion Matrix

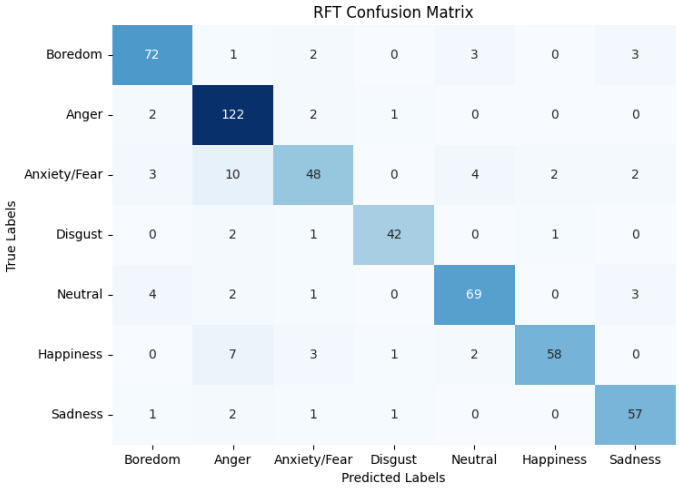

Fig. 9. RFT Confusion Matrix

finer nuances in similar emotional categories.

- k-NN: The confusion matrix for k-NN reveals its strength in classifying neutrality and sadness, but it shows higher rates of misclassification for emotions with less distinct acoustic characteristics, such as anxiety/fear and disgust. This suggests that k-NN's reliance on localized proximity in feature space does not adequately handle emotions with overlapping or subtle feature distributions.

E. Comparative Analysis and Discussion

While deep learning models such as those by Pham et al. [7] (76% CNN) and Idoko [6] (61% CNN, 56% LSTM) have demonstrated competitive performance in SER tasks, our study finds that traditional machine learning models, particularly SVM (90.65%) and Random Forest (87.48%), can achieve superior accuracy with lower computational overhead. The performance gap may be attributed to differences in feature engineering and dataset choice, as our approach leverages handcrafted acoustic features optimized for classification. These findings suggest that traditional ML remains a viable alternative for SER, especially in real-time and resource-constrained settings.

Comparing the models reveals that SVM's capability to process high-dimensional data made it particularly effective for SER tasks requiring recognition of subtle emotional distinctions. Although hyperparameter tuning was applied across all models to optimize performance, accurately classifying emotions with overlapping acoustic features—such as anxiety/fear and disgust—remained a challenge for all classifiers.

The confusion matrices (Figures 8–10) highlight specific trends in misclassification, offering insights into areas for potential improvement. SVM and Random Forest consistently performed well in classifying anger, neutrality, and sadness, suggesting that these emotions possess distinct acoustic markers that these models could effectively leverage. However, anxiety/fear was frequently misclassified as sadness or neutrality, likely due to overlapping spectral properties in pitch and energy distribution. Similarly, Random Forest and k-NN struggled with disgust, reinforcing the notion that some emotions share subtle acoustic similarities that challenge even well-optimized models.

These misclassification trends indicate that further refinements in feature engineering—such as incorporating higher-order statistical descriptors, prosodic features, or frequency modulation patterns—could enhance accuracy

V. Conclusion

This study successfully demonstrated the effectiveness of traditional machine learning models, specifically the Support Vector Machine (SVM), Random Forest, and k-Nearest Neigh(k-NN), for classifying emotions in speech. By carefully selecting and tuning features such as pitch, RMS energy, and MFCC from the Berlin Emotional Speech Database (EmoDB), the SVM model achieved a classification accuracy of 90.65%. This result highlights that, with optimized feature engineering and preprocessing, traditional models can yield high accuracy

in speech emotion recognition (SER), even in the face of overlapping emotional characteristics. Each model performed well in distinguishing distinct emotions like anger and neutrality, though, challenges persisted in classifying emotions with similar acoustic profiles, such as anxiety and fear. The study illustrates that traditional machine learning techniques, with appropriate feature selection and tuning, provide an efficient and interpretable approach to SER. These models are computationally lighter than deep learning alternatives, making them suitable for real-time applications where resource constraints are a consideration. For example, virtual assistants and call centers can use this to improve customer interactions by detecting such emotions to dynamically adjust responses based on their emotions. In education, these models can be used for teachers to keep track of students emotional responses and engagement during learning to adapt their teaching for the students. Additionally, this can be used for law or criminal investigations to analyze interrogations, emotions during trials and emergency calls to detect certain signals in speech such as deception or anxiety in speech. The practical implications suggest that lightweight machine learning models can enhance HCI in real world settings where emotion detection is useful.

VI. Future Work

This study achieved strong results with traditional learning models; however, there are several avenues for future research. First, utilizing an uncontrolled dataset would allow for testing models with real conversations to classify emotions more effectively. Additionally, exploring deep learning techniques, such as convolution neural networks (CNNs) or recurrent neural networks (RNNs), could lead to further improvements in emotion classification, especially for subtle or overlapping emotions like anxiety and fear. Another essential area for exploration is the real-time implementation and testing of these models in models in embedded devices and in real-world applications such as healthcare or virtual assistants, to assess their practical utility.

References

[1] Y. C. Pan, M. X. Xu, L. Q. Liu, and P. F. Jia, "Emotion-detecting Based Model Selection for Emotional Speech Recognition," in *Proc. of the Multiconference on Computational Engineering in Systems Applications*, Beijing, China, 2006, pp. 2169-2172. doi: 10.1109/CESA.2006.4281997.

[2] A. B. Kandali, A. Routray, and T. K. Basu, "Comparison of Features Based on MFCCs and Eigen Values of Autocorrelation Matrix for Cross-Lingual Vocal Emotion Recognition in Five Languages of Assam," in *2009 Annual IEEE India Conference*, Ahmedabad, India, 2009, pp. 1-4. doi: 10.1109/INDCON.2009.5409400.

[3] M. I and M. A. Mukunthan, "The Evolution of Emotion Recognition in Speech: A Deep Learning Perspective," in *2023 International Conference on Energy, Materials and Communication Engineering (ICEMCE)*, Madurai, India, 2023, pp. 1-6. doi: 10.1109/ICEMCE57940.2023.10434020.

[4] L.-W. Chen and A. Rudnicky, "Exploring Wav2vec 2.0 Fine Tuning for Improved Speech Emotion Recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, 2023, pp. 1-5. doi: 10.1109/ICASSP49357.2023.10095036.

[5] Z. Yuan, C. L. Philip Chen, S. Li, and T. Zhang, "Disentanglement Network: Disentangle the Emotional Features from Acoustic Features for Speech Emotion Recognition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, Korea, 2024, pp. 11686-11690. doi: 10.1109/ICASSP48485.2024.10448044.

[6] Idoko, A. (2025). Comparative Analysis Of Mel-Frequency Cepstral Coefficients And Wavelet Based Audio Signal Processing For Emotion Detection And Mental Health Assessment In Spoken Speech. https://doi.org/10.2139/ssrn.5053717

[7] Pham, M. H., Noori, F. M., & Torresen, J. (2021, December). Emotion recognition using speech data with convolutional neural network. In 2021 IEEE 2nd international conference on signal, control and communication (SCC) (pp. 182-187). IEEE.

[8] Y. Li, "Enhancing Speech Emotion Recognition for Real-World Applications via ASR Integration," in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, Cambridge, MA, USA, 2023, pp. 1-5. doi: 10.1109/ACIIW59127.2023.10388136.

[9] M. Sakurai and T. Kosaka, "Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, Kyoto, Japan, 2021, pp. 824-827. doi: 10.1109/GCCE53005.2021.9621810.

[10] S. Taware and A. D. Thakare, "Deep Learning based Speech Emotion Recognition using Multiple Acoustic Features," in *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, Bangalore, India, 2024, pp. 1-5. doi: 10.1109/ICITEICS61368.2024.10625003.

[11] Ainurrochman and U. L. Yuhana, "Improving Performance of Speech Emotion Recognition Application using Extreme Learning Machine and Utterance-level," in *2024 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, Mataram, Indonesia, 2024, pp. 466-470. doi: 10.1109/ISITIA63062.2024.10668153.

[12] X. Yan, S. Shen, Z. Liao, and Z. Li, "Research on Speech Emotion Recognition based on Multi-Level Acoustic Information Fusion," in *2024 43rd Chinese Control Conference (CCC)*, Kunming, China, 2024, pp. 8423-8428. doi: 10.23919/CCC63176.2024.10661389.

[13] P. V. Thanh, N. T. T. Huyen, P. N. Quan, and N. T. T. Trang, "A Robust Pitch-Fusion Model for Speech Emotion Recognition in Tonal Languages," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, Korea, 2024, pp. 12386-12390. doi: 10.1109/ICASSP48485.2024.10448373.

[14] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech." in INTERSPEECH, 2005, pp. 1517–1520.