Comparative Analysis of ML and DL models for Speech Emotion Recognition

1st Preetam Teja
School of AI
Amrita Vishwa Vidyapeetham
Chennai, India
preetam_teja@outlook.com

2nd Samhitha S
School of AI
Amrita Vishwa Vidyapeetham
Chennai, India
Samhithas04@gmail.com

3rd Dhanush M C
School of AI
Amrita Vishwa Vidyapeetham
Chennai, India

***Abstract*—Speech Emotion Recognition plays a crucial role in human-computer interaction, enabling systems to interpret and respond to human emotions effectively. In this study, we perform a comparative analysis of traditional machine learning and deep learning techniques for SER using the RAVDESS dataset. For machine learning, we employ Support Vector Machines (SVM) and Extreme Gradient Boosting (XGBoost) with handcrafted features extracted from spectrograms. For deep learning, we implement a Deep Convolutional Neural Network (DCNN) and a Long Short-Term Memory (LSTM) network to learn feature representations directly from spectrogram inputs. Our evaluation considers multiple performance metrics, including accuracy, precision, recall, and F1-score, to determine the efficacy of each approach. The results provide insights into the trade-offs between model complexity, interpretability, and accuracy, highlighting the potential of deep learning models for robust emotion recognition. This study contributes to the advancement of SER systems by identifying optimal methodologies for improved emotion classification.**

***Index Terms*—Speech Emotion Recognition (SER), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Deep Convolutional Neural Network (DCNN), Long Short-Term Memory (LSTM), Spectrogram, Machine Learning, Deep Learning, RAVDESS, SAVEE, Feature Extraction, Emotion Classification.**

I. Introduction

Speech Emotion Recognition (SER) has emerged as a critical area in human-computer interaction, enabling machines to recognize and respond to human emotions effectively. It has diverse applications in fields such as virtual assistants, mental health monitoring, call center analytics, and intelligent tutoring systems. Given the complexity of emotional expression in speech, the choice of computational models plays a significant role in achieving accurate emotion classification.

In this study, we conduct a comparative analysis of traditional machine learning and deep learning techniques for SER using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). This dataset contains speech recordings with eight distinct emotions, providing a balanced and controlled environment for evaluating different SER methodologies. Spectrograms, which visually represent the frequency components of speech over time, are used as the primary input representation for all models.

For traditional machine learning, we extract handcrafted features from spectrograms and employ Support Vector Machines (SVM) and Extreme Gradient Boosting (XGBoost) classifiers. These models rely on statistical and spectral features to capture emotion-related variations in speech. On the other hand, deep learning models, specifically a Deep Convolutional Neural Network (DCNN) and a Long Short-Term Memory (LSTM) network, are used to automatically learn features directly from spectrograms without explicit feature engineering. DCNNs excel at extracting spatial patterns from images, while LSTMs are well-suited for modeling temporal dependencies in sequential data.

To evaluate the effectiveness of these approaches, we use multiple performance metrics, including accuracy, precision, recall, and F1-score. Our analysis provides insights into the trade-offs between interpretability, computational complexity, and classification accuracy, highlighting the strengths and limitations of each method in SER tasks.

The remainder of this paper is structured as follows: Section II reviews related work in SER. Section III describes the dataset and preprocessing techniques. Section IV details the methodologies for both traditional machine learning and deep learning approaches. Section V presents the experimental setup and evaluation metrics. Section VI discusses the results, and Section VII concludes the paper with insights and future research directions.

II. Related Works

Speech Emotion Recognition (SER) has been extensively explored using various machine learning and deep learning techniques. This section reviews prior research focusing on models, feature extraction methods, and the impact of dataset size and diversity on model performance.

A. Traditional Machine Learning Approaches

Early SER studies employed classifiers such as Support Vector Machines (SVM), Gaussian Mixture Models (GMM), and ensemble methods like XGBoost. For instance, SVMs have been utilized for their effectiveness in high-dimensional spaces, demonstrating reasonable performance in emotion

Identify applicable funding agency here. If none, delete this.

classification tasks [1]. GMMs have been applied to model the probabilistic distribution of speech features, capturing the underlying emotional states [1]. XGBoost, known for its efficiency and performance, has been used to enhance classification accuracy in SER by effectively handling complex feature interactions [2].

B. Deep Learning-Based Approaches

With advancements in deep learning, models like Deep Convolutional Neural Networks (DCNN) and Long Short-Term Memory (LSTM) networks have been applied to SER. CNNs have been effective in automatically extracting hierarchical features from raw speech signals, leading to improved emotion recognition performance [3]. LSTMs, capable of capturing temporal dependencies in sequential data, have shown promise in modeling the dynamic nature of speech emotions [4].

C. Feature Extraction Techniques in SER

Feature engineering plays a crucial role in SER performance. Mel-Frequency Cepstral Coefficients (MFCCs) are among the most commonly used features. Studies have explored various MFCC configurations, such as MFCC 13, MFCC 13 with delta and delta-delta (MFCC 13$\Delta\Delta$), and MFCC 40. Research indicates that MFCC 40 provides a more comprehensive representation of the speech signal, leading to better emotion classification accuracy [5]. Combining spectral features like MFCC with prosodic features has also been investigated, showing enhanced performance in emotion recognition tasks [6].

Additionally, tools like OpenSMILE have been employed to extract a wide range of features, including 88-dimensional feature sets encompassing spectral, prosodic, and voice quality features. Feature selection methods, such as using XGBoost to identify the top contributing features, have been applied to reduce dimensionality and improve model performance [2].

D. Impact of Dataset Size and Diversity

The choice and size of datasets significantly affect SER model performance. Commonly used datasets include the Surrey Audio-Visual Expressed Emotion (SAVEE) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Studies have shown that models trained on larger and more diverse datasets tend to generalize better to unseen data [7]. However, challenges arise when dealing with smaller datasets like SAVEE, which may limit the model's ability to capture the variability in emotional expressions. Data augmentation techniques and combining multiple datasets have been proposed to address these limitations [8].

E. Comparative Studies in SER

Several comparative analyses have been conducted to evaluate the performance of different models and feature sets in SER. Research comparing traditional machine learning models (e.g., SVM, GMM) with deep learning models (e.g., DCNN, LSTM) suggests that deep learning approaches generally achieve higher accuracy, particularly when trained on large datasets. Furthermore, studies assessing the impact of various feature sets have found that combining MFCC with other spectral and prosodic features enhances classification performance across multiple classifiers [2].

This study builds upon existing research by conducting a comprehensive comparative analysis using the SAVEE and RAVDESS datasets. By evaluating the impact of different feature sets, dataset sizes, and classifiers (SVM, XGBoost, GMM, DCNN, LSTM) on SER performance, we aim to provide insights into the most effective approaches for emotion recognition from speech.

III. METHODOLOGY

This study conducts a comparative analysis of speech emotion recognition (SER) using multiple datasets, features, and models. The methodology consists of dataset selection, preprocessing, feature extraction, model architectures, training procedures, and evaluation metrics.

A. Datasets

Two publicly available datasets were utilized: the Surrey Audio-Visual Expressed Emotion (SAVEE) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Table I summarizes the dataset details.

TABLE I
DESCRIPTION OF THE DATASETS USED IN SER

Dataset	Speakers	Emotions	Total Samples
SAVEE	4 (Male)	7	480
RAVDESS	24 (12 Male, 12 Female)	8	1440

B. Preprocessing Techniques

Preprocessing is essential for ensuring consistent and high-quality input data. The following steps were applied:

1) Resampling: All audio files were resampled to a uniform sampling rate of 16 kHz to standardize the data.

2) Voice Activity Detection (VAD) and Trimming: VAD was applied to remove non-speech regions, ensuring only relevant speech segments were retained.

3) Data Visualization: Spectrograms and waveform plots were used to visualize the structure of the audio signals. Figure 1 illustrates an example spectrogram of an emotion-labeled speech signal.

C. Feature Extraction

Feature extraction plays a critical role in SER performance. The following feature sets were engineered:

1) Mel-Frequency Cepstral Coefficients (MFCCs): Three different MFCC feature sets were used:

- **MFCC-13**: The base 13 MFCCs.
- **MFCC-13$\Delta\Delta$**: Includes delta and delta-delta coefficients.
- **MFCC-40**: A 40-coefficient representation found to be more effective in classification.

Experiments demonstrated that MFCC-40 provided better accuracy.

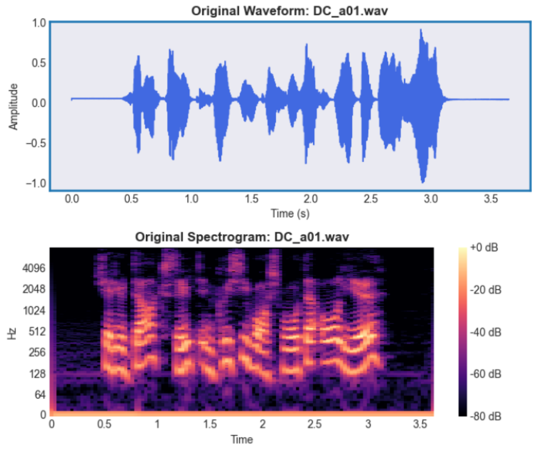

Fig. 1. Spectrogram of a sample emotional speech utterance

2) Spectral and Prosodic Features: Additional features like zero-crossing rate (ZCR), and root mean square energy (RMS), were combined with MFCC-40 to enhance classification performance.

3) OpenSMILE Features and Feature Selection: The OpenSMILE toolkit was used to extract 88 features. XGBoost-based feature selection identified the top 50 most significant features.

D. Model Architectures

The following models were employed for comparative analysis. Table II provides a summary.

TABLE II
MODELS USED IN SER ANALYSIS

Model	Description
SVM	Support Vector Machine, efficient in high-dim spaces
XGBoost	Gradient boosting, strong feature selection capability
GMM	Gaussian Mixture Model for probabilistic distribution
DCNN	Deep CNN for automatic hierarchical feature extraction
LSTM	LSTM for capturing temporal dependencies

E. Training Procedures

Each model was trained separately on SAVEE, RAVDESS, and a combined dataset to evaluate dataset size effects. Hyperparameters were tuned to optimize performance, and cross-validation was performed.

F. Hyperparameter Tuning

To optimize model performance, we employed both **Optuna** and **Grid Search** for hyperparameter tuning.

1) Optuna-based Hyperparameter Optimization: Optuna, a powerful Bayesian optimization framework, was used to efficiently explore the hyperparameter space, particularly for deep learning models like DCNN and LSTM. The optimization process followed these steps:

- Defined an objective function evaluating model performance based on validation accuracy.
- Sampled hyperparameters such as learning rate, dropout rate, number of layers, and hidden units.
- Used `Tree-structured Parzen Estimator (TPE)` for intelligent sampling.
- Conducted multiple trials, selecting the best configuration automatically.

2) Grid Search for Classical Models: For traditional machine learning models such as SVM, XGBoost, and GMM, we performed exhaustive hyperparameter tuning using Grid Search. The process involved:

- Defining a grid of hyperparameter values.
- Training models on all possible combinations.
- Selecting the best configuration based on cross-validation performance.

3) Best Hyperparameters Found: Table III summarizes the optimal hyperparameters found using Optuna and Grid Search.

TABLE III
OPTIMIZED HYPERPARAMETERS FOR DIFFERENT MODELS

Model	Optimal Hyperparameters
SVM	C=10, Gamma = 'scale', Kernel=RBF
XGBoost	Learning Rate=0.05, Max Depth=9, n_estimators=300
GMM	Components=8, Covariance Type=Full
DCNN	Learning Rate=0.001, Dropout=0.3, Layers=4
LSTM	Hidden Dim=128, Dropout=0.25, Learning Rate=0.002

This tuning process significantly improved model performance, ensuring optimal generalization and minimizing overfitting.

G. Evaluation Metrics

Models were evaluated using accuracy as the primary metric. The impact of different feature sets, dataset sizes, and model architectures on classification performance was analyzed.

IV. RESULTS AND DISCUSSION

This section compares the performance of two machine learning models—**SVM** and **XGBoost**—on two widely used emotional speech datasets: **SAVEE** and **RAVDESS**. Four different feature sets are evaluated:

- MFCC 13
- MFCC 40
- MFCC DD (Delta and Delta-Delta)
- MFCC 40 + ZCR + RMS
- OpenSmile Feature set

Each configuration is evaluated using **Accuracy**, **F1 Score**, **Precision**, and **Recall**.

A. SAVEE Dataset

Dataset Info: 480 utterances from 4 male speakers, covering 7 emotions.

TABLE IV
SVM RESULTS ON SAVEE DATASET

Features	Accuracy (%)	F1 Score	Precision	Recall
MFCC 13	42.71	0.33	0.36	0.35
MFCC 40	52.08	0.47	0.65	0.45
MFCC DD	25.00	0.16	0.17	0.18
MFCC 40 + ZCR + RMS	**55.20**	**0.51**	**0.68**	**0.48**

SVM Model Performance on SAVEE: **Observation:** The SVM model shows significant improvement when transitioning from MFCC 13 to MFCC 40, highlighting the benefit of using a higher-dimensional spectral representation. However, performance drops sharply with MFCC DD alone, suggesting that second-order dynamics may not be effective in isolation. Combining MFCC 40 with prosodic features (ZCR, RMS) yields the highest performance, indicating that **SVM benefits from multi-domain features in smaller datasets**.

Best performance: MFCC 40 + ZCR + RMS (Accuracy: **55.2%**, F1: **0.51**)

TABLE V
XGBOOST RESULTS ON SAVEE DATASET

Features	Accuracy (%)	F1 Score	Precision	Recall
MFCC 13	44.79	0.35	0.38	0.38
MFCC 40	57.29	0.53	0.58	0.52
MFCC DD	38.00	0.28	0.33	0.30
MFCC 40 + ZCR + RMS	**59.38**	**0.54**	**0.67**	**0.53**

XGBoost Model Performance on SAVEE: **Observation:** XGBoost exhibits clear performance gains over SVM on the SAVEE dataset across all feature types. The model leverages hierarchical decision-making to extract meaningful patterns even from MFCC DD. The best results are obtained with MFCC 40 + ZCR + RMS, showing that **XGBoost effectively captures both spectral complexity and prosodic variations in limited data scenarios**.

Best performance: MFCC 40 + ZCR + RMS (Accuracy: **59.38%**, F1: **0.54**)

B. RAVDESS Dataset

Dataset Info: 1440 utterances from 12 male and 12 female speakers, covering 8 emotions.

TABLE VI
SVM RESULTS ON RAVDESS DATASET

Features	Accuracy (%)	F1 Score	Precision	Recall
MFCC 13	59.03	0.59	0.60	0.59
MFCC 40	61.46	0.61	0.62	0.62
MFCC DD	58.33	0.58	0.59	0.58
MFCC 40 + ZCR + RMS	**62.85**	**0.62**	**0.63**	**0.63**

SVM Model Performance on RAVDESS: **Observation:** On the RAVDESS dataset, SVM performs consistently well across all feature sets. The improvements from MFCC 13 to MFCC 40 and further with ZCR + RMS emphasize the robustness of spectral-prosodic fusion even at scale. The overall higher scores compared to SAVEE indicate that **SVM scales better with more data and remains effective when paired with rich feature sets**.

Best performance: MFCC 40 + ZCR + RMS (Accuracy: **62.85%**, F1: **0.62**)

TABLE VII
XGBOOST RESULTS ON RAVDESS DATASET

Features	Accuracy (%)	F1 Score	Precision	Recall
MFCC 13	**53.12**	**0.52**	**0.54**	**0.52**
MFCC 40	48.26	0.47	0.47	0.47
MFCC DD	48.00	0.46	0.48	0.47
MFCC 40 + ZCR + RMS	49.65	0.47	0.48	0.48

XGBoost Model Performance on RAVDESS: **Observation:** Contrary to its performance on SAVEE, XGBoost struggles on the RAVDESS dataset. Accuracy and F1 scores decline across the board, especially with MFCC 40 and combined features. This suggests **possible overfitting or insufficient generalization when applying the same hyperparameters to a larger dataset**. MFCC 13 performs best, indicating simpler representations may be more stable in such cases.

Best performance: MFCC 13 (Accuracy: **53.12%**, F1: **0.52**)

C. Deep Learning Model Analysis on RAVDESS Dataset

TABLE VIII
PERFORMANCE OF DCNN ON RAVDESS DATASET

Features	Accuracy (%)	F1 Score	Precision	Recall
MFCC 40 + ZCR + RMS	**94.44**	**93.80**	**94.55**	**93.40**
OpenSMILE Feature Set	91.15	91.05	89.45	91.00

Observation: The DCNN model achieved the best performance using the MFCC 40 + ZCR + RMS feature set with 94.44% accuracy. OpenSMILE features also yielded competitive results, but the handcrafted features led to superior learning when paired with CNN's hierarchical structure.

Observation: The DCNN model achieved the highest accuracy of **94.44%** when trained on MFCC 40 combined with ZCR and RMS features, outperforming the OpenSMILE feature set by a margin of over 3%. Furthermore, DCNN consistently yielded high values across all evaluation metrics, demonstrating its strength in learning complex patterns from engineered acoustic features. The OpenSMILE set still produced competitive results, highlighting its effectiveness as a compact, pre-selected feature representation.

TABLE IX
PERFORMANCE OF LSTM ON RAVDESS DATASET

Features	Accuracy (%)	F1 Score	Precision	Recall
MFCC 40 + ZCR + RMS	**85.76**	**0.86**	**0.86**	**0.85**
OpenSMILE Feature Set	80.86	0.81	0.80	0.80

Observation: LSTM showed better accuracy with the MFCC 40 + ZCR + RMS feature set (85.76%). While OpenSMILE still delivered decent results, the temporal modeling ability of LSTM benefited more from MFCC-based features.

Observation: The LSTM model also performed better with MFCC 40 + ZCR + RMS features, reaching an accuracy of **85.76%**, whereas performance dropped to **80.86%** with OpenSMILE features. While both feature types are temporal in nature, the MFCC-based representation appears to preserve emotional nuances more effectively for sequential models like LSTM. Nevertheless, LSTM's performance remained lower than that of the DCNN model across all metrics.

Inference: From these results, it is evident that:

- Deep learning models significantly benefit from handcrafted low-level descriptors such as MFCCs combined with ZCR and RMS.
- DCNN outperforms LSTM in this experimental setup, suggesting that convolutional architectures are more effective at capturing spectral variations present in emotional speech.
- OpenSMILE, although feature-rich, shows slightly lower performance compared to task-specific features when used with deep learning models.

References

[1] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, pp. 155–177, 2015.

[2] L. He and Y. Ren, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 577–588, 2018.

[3] Q. Ouyang, "Speech emotion detection based on mfcc and cnn-lstm architecture," *arXiv preprint arXiv:2501.10666*, 2025.

[4] T. D. Dhamale, "On the evaluation and implementation of lstm model for speech emotion recognition using mfcc," in *Proceedings of International Conference on Computational Intelligence and Data Engineering: ICCIDE 2021*. Springer, 2022.

[5] O. Atila and A. Engür, "A novel concatenated 1d-cnn model for speech emotion recognition," *Biomedical Signal Processing and Control*, vol. 79, p. 104057, 2023.

[6] B. Li, "Speech emotion recognition based on cnn-transformer with different loss function," *Journal of Computer and Communications*, vol. 13, pp. 103–115, 2025.

[7] C. Barhoumi and Y. BenAyed, "Real-time speech emotion recognition using deep learning and data augmentation," *Artificial Intelligence Review*, vol. 58, p. 49, 2024.

[8] S. Garg and S. Aggarwal, "Speech emotion recognition for multiclass classification using hybrid cnn-lstm," *ResearchGate*, 2023.