Comparative Analysis of ML and DL models for Speech Emotion Recognition

Preetam Teja B
School of AI
Amrita Vishwa Vidyapeetham
Coimbatore, India
preetam_teja@outlook.com

Samhitha S
School of AI
Amrita Vishwa Vidyapeetham
Coimbatore, India
Samhithas04@gmail.com

Dhanush M C
School of AI
Amrita Vishwa Vidyapeetham
Coimbatore, India
mcdhanush1122@outlook.com

***Abstract*—Speech emotion recognition is an important field in the domain of speech signal processing, SER helps computers to understand emotion from speech audio signal enabling systems to respond to human emotions effectively. In this study, we have performed a comparative analysis study of traditional machine learning and deep learning techniques for SER using the RAVDESS and SAVEE dataset. For machine learning, we employ Support Vector Machines and Extreme Gradient Boosting with handcrafted features extracted from Opensmile, MFCC, RMS, ZCR. For deep learning, we implement a Deep Convolutional Neural Network and a Long Short-Term Memory network to learn feature representations directly from the feature inputs.We have achieved 94.55% accuracy using our best model DCNN along with MFCC40+ZCR+RMS feature set in RAVDESS dataset. We have taken multiple performance metrics into account. They include accuracy, precision, recall, and F1-score. They provide additional information about the trade-offs between the model's complexity, interpretability, and accuracy, highlighting the potential of deep learning models in SER.**

***Index Terms*—Speech emotion recognition, Support vector machine, Extreme gradient boosting, Deep convolutional neural network, Long short-term memory, Spectrogram, Machine learning, Deep learning, RAVDESS, SAVEE, Feature extraction, Emotion classification.**

I. Introduction

The capacity to identify emotions from speech cues makes it easier to design emotionally intelligent systems with applications in virtual assistants, mental health tracking, call center analysis, and tutoring systems [1], [2]. Reliable emotion classification from speech is still a difficult task given the inbuilt variability and richness of human emotional expression [3].

A variety of computational models have been put forward for SER, each with different benefits of performance, interpretability, and scalability. Classical machine learning approaches generally are based on hand-designed features extracted from acoustic or spectral characteristics of speech signals [4], whereas deep learning-based methods seek to learn features of interest from raw or lightly processed inputs [5].

In this research, we conduct a comparative investigation of conventional and deep learning-based methods for SER on the RAVDESS and SAVEE corpus [6], [7]. These corpus comprises recordings with eight and seven emotional categories respectively, and the controlled nature of the recordings makes the corpus well-suited for benchmarking SER systems under controlled conditions. As the dataset sizes are different for both the corpus we also employ a dataset level comparative study on how different models handle smaller or larger datasets. MFCC coefficients, spectral and prosodic features are used as the primary input for all models in this research [8].

The machine learning algorithms used are Support Vector Machines and Extreme Gradient Boosting [9], [10], both of which employ statistical and spectral descriptors from the input features. For deep learning, we use a Deep Convolutional Neural Network and a Long Short Term Memory network [11], [12], both of which are capable of learning the spatial and temporal patterns automatically, without feature engineering but still are being fed with handcrafted feature sets.

Each model is evaluated in terms of standard classification measures i.e, precision, recall, accuracy, and F1-score. Through comparison, we aim to bring forward the practical compromises inherent with every modeling technique by taking into consideration computational expense, interpretability, and classification ability.

II. Related Works

Speech Emotion Recognition has been widely investigated with a variety of machine learning and deep learning methods. This section summarizes existing work in terms of models, feature extraction techniques, and the effect of dataset size and diversity on model performance.

A. Traditional Machine Learning Approaches

In initial stages of speech emotion classification, people used models like SVM, GMM, XGBOOST. , SVMs perform well in high-dimensional spaces for classification tasks [13]. GMMs have been used to estimate the probabilistic distribution of speech characteristics, reflecting the underlying emotional states [13]. XGBoost due to its performance and efficiency has been used to improve the classification accuracy for SER effectively managing the intricate feature interactions [14].

B. Deep Learning-Based Approaches

As there have been breakthroughs in deep learning, the Deep Convolutional Neural Networks and Long Short-Term

Memory models have been employed to SER. CNNs ability to handle higher order data [15] and LSTMs ability to handle long term dependencies also affect the accuracies and improve SER [16].

C. Feature Extraction Techniques in SER

Feature engineering is very important in emotion recognition. MFCCs are used very popularly. Researchers have tested different MFCC configurations, including MFCC 13, MFCC 13 with delta and delta-delta, and MFCC 40. Research shows that MFCC 40 gives a more complete representation of the speech signal, contributing to higher emotion classification accuracy [17]. Merging spectral characteristics such as MFCC with prosodic characteristics has also been explored, indicating improved performance in emotion recognition tasks [18].

Features of OpenSMILE have also been used to extract various wide ranges of features, including 88-dimensional feature sets that cover spectral, prosodic, and voice quality features. Feature selection techniques, such as the use of XGBoost to determine top contributing features, have been used to trim dimensionality and enhance model performance [14].

D. Impact of Dataset Size and Diversity

Dataset choice and size play a major impact on SER model performance. Typical datasets used are the Surrey Audio-Visual Expressed Emotion and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Models trained on bigger and more diverse datasets have been proven to generalize better to novel data [19]. But while working with small datasets such as SAVEE we face a lot of difficulties to achieve high accuracy. Data augmentation strategies and using several datasets have been suggested as ways to solve these limitations [20].

E. Comparative Studies in SER

Several comparative analyses have been conducted to evaluate the performance of different models and feature sets in SER. Based on the previous studies we can see that Deep learning models achieve high accuracy over Machine learning models when trained on a large dataset. Also, 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 [14].

This study is done by conducting a comprehensive comparative analysis using the SAVEE and RAVDESS datasets. We are evaluating evaluating the impact of different feature sets, dataset sizes, and models on SER performance.

III. Data-Set Description

A. Datasets

The two datasets used are Surrey Audio-Visual Expressed Emotion and the Ryerson Audio-Visual Database of Emotional Speech and Song. 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

IV. Methodology

We have taken two datasets SAVEE and RAVDESS and we load them inidivually, Once thats done we proceed to the pre-processing. This ensures that we input good quality data removing the silent regions and noise.Here is a sample of the data before pre processing:

A. Flow chart

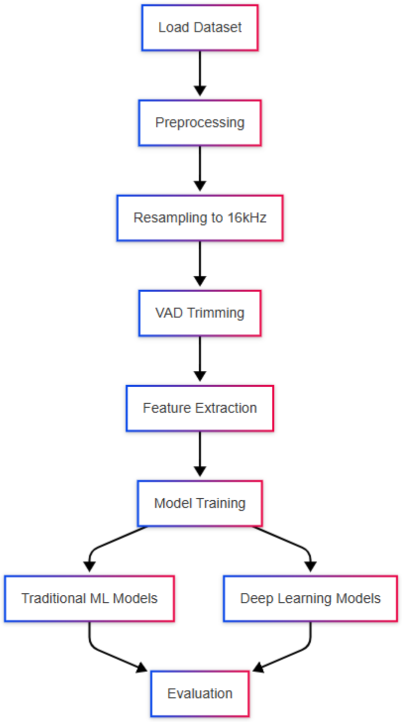

Fig. 1. Flow chart of the comparative analysis pipeline

B. Pre-processing Techniques

1) Resampling: As Ryerson Audio-Visual Database of Emotional Speech and Song was sampled at 16000 kHz, We had to reasmple Surrey Audio-Visual Expressed Emotion dataset to 16000 kHz too.

2) Voice Activity Detection and Trimming: We noticed that there was a unvoiced region in the starting and ending of the sample so we used Voice activity detection trimming to eliminate them. We set adaptive thresholding value as 0.9 for VAD trimming.

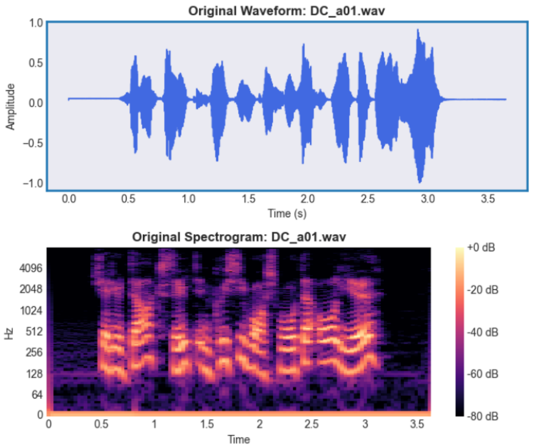

Fig. 2. Audio input sample and spectrogram before preprocessing

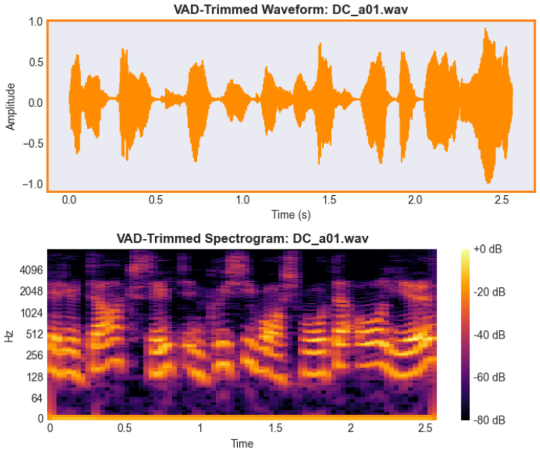

Fig. 3. Audio input sample and spectrogram after preprocessing

C. Feature Extraction

1) Mel-Frequency Cepstral Coefficients: **MFCC-13:** The 13 Base MFCCs derived by taking the discrete cosine transform (DCT) of the log Mel spectrum:

$$\text{MFCC}(n) = \sum_{k=1}^{K} \log(S_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right]$$

where S_k is the energy in the k-th Mel band and K is the number of Mel filters.

MFCC-13ΔΔ: Includes the first and second-order temporal derivatives (delta and delta-delta coefficients), capturing dynamic changes in spectral features over time. The delta coefficients are computed using:

$$\Delta x_t = \frac{\sum_{n=1}^{N} n(x_{t+n} - x_{t-n})}{2\sum_{n=1}^{N} n^2}$$

where x_t is the MFCC at time t, and N is the window size (typically 2).

MFCC-40: A 40-coefficient MFCC representation, providing a higher resolution of the speech spectrum.

2) Spectral and Prosodic Features: **Zero-Crossing Rate (ZCR):** The rate at which the signal changes sign:

$$\text{ZCR} = \frac{1}{T-1}\sum_{t=1}^{T-1} \nVdash_{\{x_t x_{t+1} < 0\}}$$

where x_t is the signal amplitude at time t, and $\nVdash$ is the indicator function.

Root Mean Square (RMS) Energy:

$$\text{RMS} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} x_t^2}$$

where x_t is the amplitude of the audio signal and T is the frame length.

The spectral and prosodic features are appended with MFCC-40 to improve performance.

3) OpenSMILE Features: The openSMILE toolkit was used to extract 88 low level features which are energy, spectral and voice related.

D. Model Architectures

have used 4 different models which showed good capabilities in 4 different cases. Table II provides a summary.

TABLE II
Models Used in SER Analysis

Model	Description
SVM	Proves to be efficient in high-dim spaces
XGBoost	Works well with smaller datasets
GMM	Good probabilistic distribution
DCNN	Deep CNN works in automatic hierarchical feature extraction
LSTM	LSTM for capturing temporal dependencies

E. Training Procedures

1) Hyperparameter Tuning: Each model was trained separately on SAVEE and RAVDESS to understand how different models performs with different sizes of data. Hyperparameters were tuned to optimize performance.

For optimising the model performance we did hyperparameter tuning using **Optuna** and **Grid Search**. This helped us to find out some good hyperparameters without brute-forcing our way to it.

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

F. Evaluation Metrics

The main metrics used to evaluate the models were accuracy and F1-score. Analysis was done on the effects of various feature sets, dataset sizes, and model architectures on classification performance.

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
DCNN	Learning Rate=0.001, Dropout=0.3, Layers=4
LSTM	Hidden Dim=128, Dropout=0.25, Learning Rate=0.002

V. RESULTS AND DISCUSSION

The performance of two machine learning models, **SVM** and **XGBoost** and two deep learning models **DCNN** and **LSTM**, is compared in this section on the two datasets we have chosen: SAVEE and RAVDESS. Four distinct feature sets are used:
MFCC 13
MFCC Δ and $\Delta\Delta$
MFCC 40
MFCC 40 + ZCR + RMS
OpenSMILE Feature Set

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

A. Machine Learning approaches on SAVEE dataset

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**

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 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. Machine Learning approaches on RAVDESS dataset

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**

On RAVDESS dataset SVM has a consistent performance over all the featuer set.
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

Unlike its performance in SAVEE, XGBoost struggles on the RAVDESS dataset, Overfitting might be possible.
Best performance: MFCC 13 (Accuracy: **53.12%**, F1: **0.52**)

C. Deep Learning Model Analysis on SAVEE Dataset

TABLE VIII
PERFORMANCE OF DCNN ON SAVEE DATASET

Features	Accuracy (%)	F1 Score	Precision	Recall
MFCC 40 + ZCR + RMS	**71.45**	**0.65**	**0.67**	**0.65**
OpenSMILE Feature Set	60.75	0.56	0.55	0.59

The DCNN model achieved the highest accuracy of **71.45%** on the SAVEE dataset using the MFCC 40 + ZCR + RMS feature set. Compared to OpenSMILE, this combination provided better results, This shows us the importance of handcrafted feature extraction in small datasets.

TABLE IX
PERFORMANCE OF LSTM ON SAVEE DATASET

Features	Accuracy (%)	F1 Score	Precision	Recall
MFCC 40 + ZCR + RMS	47.29	0.38	0.39	0.39
OpenSMILE Feature Set	**62.92**	**0.58**	**0.63**	**0.60**

In contrast to DCNN, the LSTM model performed better with OpenSMILE features, achieving an accuracy of **62.92%**. The MFCC-based features underperformed, likely due to limited data affecting the sequential modeling capability of LSTM.

D. Deep Learning Model Analysis on RAVDESS Dataset

TABLE X
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

When trained on MFCC 40 along with ZCR and RMS features, the DCNN model outperformed the OpenSMILE feature set by more than 3%, achieving the highest accuracy of **94.44%**. DCNN's ability to learn intricate patterns from engineered acoustic features is demonstrated by the consistently high values it produced across all evaluation metrics. Despite this, the

OpenSMILE set yielded competitive results, demonstrating its usefulness as a condensed, pre-selected feature representation.

TABLE XI
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

The accuracy of the LSTM model was **85.76%** with the MFCC 40 + ZCR + RMS features, while it dropped to **80.86%** with the OpenSMILE features. Even though both feature types are temporal in nature, the MFCC-based representation appears to preserve emotional nuances more effectively for sequential models such as LSTM. On all metrics, however, the DCNN model performed better than the LSTM model.

VI. CONCLUSION

Several important conclusions are drawn from the comparison of machine learning and deep learning models using the SAVEE and RAVDESS datasets. In general, deep learning models performed better than conventional machine learning models; on the RAVDESS dataset, DCNN achieved the highest accuracy of 94.44% using the MFCC 40 + ZCR + RMS feature set. The fact that this feature combination consistently yielded the best results across all models underlined the importance of combining spectral and temporal features for emotional speech recognition. Among all models on the smaller SAVEE dataset, DCNN led with 68.54%; XGBoost and SVM followed with 59.38% and 55.2%, respectively. LSTM underperformed here (47.29%), probably because its sequential architecture struggled with small training samples. By contrast, on the bigger and more varied RAVDESS dataset, all models performed better; LSTM indicated good promise (85.76%) and SVM unexpectedly beat XGBoost (62.85% vs. 49.65%), probably because of improved generalization in complex acoustic conditions. Moreover, DCNN always showed better generalization by properly catching spatial patterns in features, therefore outperforming both datasets. While XGBoost revealed sensitivity to overfitting in complicated situations, LSTM gained more from bigger datasets. SVM turned out to be a strong baseline, especially with small features and high variability. These results indicate that dataset size, feature richness, and architectural fit have a close relationship with model performance. Deep learning models, particularly DCNN, show good generalization across datasets, which makes them ideal for SER tasks. Including spectral and prosodic features significantly improves classification accuracy. Conventional ML models remain competitive, particularly on smaller datasets or in situations where interpretability matters. Future studies could investigate real-world deployment of SER systems, cross-corpus generalization, and transformer-based models.

REFERENCES

[1] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on emotion detection from the speech signal,” *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” *Computer Speech & Language*, vol. 60, pp. 3–37, 2020.

[3] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[4] F. Eyben, M. Wöllmer, and B. Schuller, “opensmile–the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[5] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” *ICASSP*, pp. 5200–5204, 2016.

[6] S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english,” *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[7] S. Haq and P. Jackson, “Audio-visual emotion recognition using adaboost,” in *International Conference on Auditory-Visual Speech Processing*, 2009.

[8] L. Chen, X. Mao, Y. Xue, and L. Cheng, “A 3-class emotion recognition system using acoustic features,” *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 2, 2012.

[9] C. Cortes and V. Vapnik, “Support-vector networks,” in *Machine learning*, vol. 20. Springer, 1995, pp. 273–297.

[10] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[12] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

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

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

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