International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2817-2827
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0415

A Comparative Analysis of SVM, CNN and LSTM Models for Speech Emotion Recognition

Sridevi J[1], S L Jayalakshmi[2]
[1]Research Scholar, Dept. of Computer Science, School of Engineering and Technology., Pondicherry University, Puducherry, India.
[2]Assistant professor, Dept. of Computer Science, School of Engineering and Technology., Pondicherry University, Puducherry, India.
Email ID: *sridevijagdish2408@gmail.com[1], sathishjayalakshmi02@pondiuni.ac.in[2]*

Abstract
The Speech Emotion Recognition (SER) project aims to develop an intelligent system capable of recognizing human emotions from speech signals. SER plays a major role in applications such as Human-Computer Interaction (HCI), sentiment analysis and psychological research. In this project, we leverage machine learning techniques and signal processing methods to analyze speech signals and extract features that capture the emotional content, following a structured pipeline that includes data collection, preprocessing, feature extraction, model training and validation. To reduce high frequency noise and retain essential speech characteristics, a low-pass filter is applied and then Mel-Frequency Cepstral Coefficients (MFCCs) is applied to extract meaningful features from audio files, and employing machine learning models like Support Vector Machines (SVM), as well as deep learning models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM), facilitates emotion classification. The system's objective is to accurately distinguish between different emotions like anger, sadness, happiness, and neutral states from speech signals. Moreover, the inclusion of a user-friendly interface enhances accessibility and usability, enabling seamless interaction with the system. Through experimentation and rigorous evaluation, the efficacy of the proposed approach in recognizing emotions from speech is demonstrated. The SER project holds immense potential to contribute to various domains, including HCI, mental health assessment, and affective computing, thereby augmenting our comprehension and interaction with human emotions.

Keywords: *SVM, CNN, LSTM, Speech emotion recognition, RAVDESS*

1. Introduction
The Speech Emotion Recognition (SER) project endeavors to develop a sophisticated system capable of accurately detecting and categorizing emotions conveyed through human speech captured in audio recordings. With applications spanning diverse domains such as Human-Computer Interaction (HCI), customer service and healthcare, emotion recognition holds significant promise. By integrating machine learning algorithms and advanced signal processing techniques, this project meticulously extracts relevant features from speech signals. These features serve as inputs to train robust models, enabling precise classification of emotions with a high degree of accuracy. With the addition of a user-friendly interface, the system becomes accessible to a wider audience, facilitating seamless interaction and utilization across various contexts. Ultimately, by accurately discerning emotional nuances within spoken language, the system aims to provide invaluable insights, enhance user experiences, and foster improved communication dynamics. This study aims to develop an advanced system for emotion recognition in human speech using machine learning techniques. Like human perception of

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2817-2827
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0415

emotions through voice, we are training computers to achieve this capability. Initially, we collect speech recordings and apply noise reduction techniques to ensure clear data. Subsequently, we employ sophisticated algorithms to analyze various aspects of the recordings, such as speech rate and tone, to extract emotional cues. Leveraging this information, our system learns to classify different emotions, including happiness, sadness and neutral. We utilize a range of techniques to train the system effectively. Training and rigorous testing are conducted to evaluate the system's ability to accurately identify emotions in new recordings. Ultimately, our objective is to enhance the computer's proficiency in understanding emotions in speech, which holds significant potential for wide range of applications, including real-time mental health monitoring systems, emotionally intelligent virtual systems, adaptive learning platforms that respond to student's emotions and assistive technologies for individuals with communication disorders. The problem we aim to address is the accurate recognition of emotions conveyed through human speech. Emotions play a crucial role in communication and being able to detect them accurately can enhance various applications, including virtual assistants, customer service systems and mental health monitoring tools. However, recognizing emotions solely based on audio data poses significant challenges due to the complexity and variability of human speech. The remaining part of this paper is structured as follows: Section 2 presents literature review in SER. Section 3 highlights the proposed methodology, including architecture and model description. Section 4 discusses experimental studies and discussions. Finally, Section 5 illustrates the conclusion and future enhancement.

2. Literature Review

Recent advances in the Speech Emotion Recognition (SER) have increasingly utilized in machine learning, deep learning techniques, hybrid models and innovative feature representations to enhance the accuracy of emotion classification across a wide range of datasets. Researchers have focused on advancing the feature extraction methods, architectural design, and cross domain adaptability to overcome the key challenges such as data imbalance, nuanced emotional expressions, and the need for real-time performance. A two-phase framework incorporating Windowed Long-Term Average Spectrum (WLTAS) and a Logistic-Rectified Linear Unit (LoRLU) was introduced for effective spectral feature extraction and sparsity induction, respectively. Classification was performed using a Gibbs Restricted Boltzmann Machine (GRBM), enabling the model to capture high-dimensional dependencies and achieve an impressive 99% accuracy, 0.97 precision, and 0.98 F1 score. While the method shows strong potential for applications in mental health monitoring, its generalizability across datasets remains a limitation [1]. To address class imbalance, a Sparse Learning-Based Fusion Model (SLBF) was proposed, selectively retraining weaker nodes while preserving well-performing ones. This model combined 2D CNN-SEN and MLSTM-FCN architectures and utilized MFCC, RMS energy, and Zero Crossing Rate (ZCR) features. The model achieved an impressive accuracy of 99.3% on the SAVEE dataset and 97.92% on RAVDESS, surpassing traditional classifiers by effectively addressing the issue of underrepresented emotional states and its higher computational complexity may limit the scalability [2]. In another effort to model temporal dependencies, a Bi-directional Gated Recurrent Unit (Bi-GRU) architecture enhanced with multi-head attention was developed. This model captured both local and global dependencies, addressing issues like dispersed emotional cues and inconsistent feature scales. It demonstrated superior accuracy on benchmark datasets such as IEMOCAP and Emo-DB, and proved effective for related tasks like sentiment analysis. Nevertheless, the architectural complexity may increase the risk of overfitting [3]. An innovative use of visual representation was demonstrated in a method employing a VGG-based Deep Convolutional Neural Network (DCNN) optimized with the Beluga Whale Optimization (BWO) algorithm and Chaogram signal transformation. By converting speech signals into 2D images, the model enabled CNNs to exploit spatial patterns for emotion classification. Experiments on EMO-DB and eNTERFACE05 confirmed significant

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2817-2827
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0415

improvements in accuracy, although computational cost remains a consideration [4]. Building on transformer architectures, a lightweight Vision Transformer (ViT) model was introduced using mel-spectrogram inputs to extract spatial and global emotional features. Patch-based feature extraction and self-attention mechanisms enabled the model to attain up to 98% accuracy on datasets such as TESS and EMO-DB. Despite some sensitivity to configuration changes, the model's simplicity supports real-time implementation [5]. To enhance generalizability and robustness, a light weight deep neural ensemble model was proposed utilizing classical audio features such as MFCC, ZCR, Chroma STFT, and RMSE. Fine-tuning with learning rate schedulers and regularization techniques enabled superior performance across five datasets, RAVDESS, TESS, SAVEE, CREMA-D, and EmoDB when measured by accuracy, AUC-ROC, AUC-PRC, and F1-score. Future work aims to automate feature extraction and enhance model interpretability [6]. A hybrid system combining CNN and LSTM was designed for simultaneous keyword spotting and negative emotion detection, focusing on the wake-word "ON" and emotions like sadness and anger. Using MFCC, LPCC, and CHROMA features, the CNN achieved 97.23% accuracy for keyword detection, while the LSTM reached 88.94% for emotion recognition. Although promising, reliance on semi-simulated datasets constrains its real-world applicability [7]. Accent variability was addressed through a multi-model framework incorporating both conventional machine learning (Random Forest, SVM, KNN) and deep learning models (1D-CNN, LSTM). With nine acoustic features including MFCC, Chroma-STFT, and spectral descriptors, the 1D-CNN achieved up to 99% accuracy across regionally diverse English datasets, highlighting the model's cross-accent generalizability. Trade-offs were noted with dimensionality reduction methods such as KBest [8]. A CNN-LSTM hybrid architecture was developed to jointly extract spatial and temporal features from speech signals. Time-distributed CNN layers facilitated dynamic feature extraction, leading to 65% accuracy for seven emotions and 75% for six emotions on the RAVDESS dataset, outperforming SVM by 10%. Authors suggest integration with Hidden Markov Models (HMM) and larger datasets for further improvements [9]. Real-time emotion recognition was explored through a multimodal SER system combining audio and video inputs. MFCCs, power, tone, and composite features were extracted and processed using an RNN-based classifier. Although real-time implementation was effective, limitations included feature redundancy and insufficient dataset diversity [10]. Lastly, a traditional SVM-based SER system using MFCC and LPCC features was evaluated on the LDC and UGA datasets. The gender-dependent SVM classifier using MFCC achieved the highest accuracy 84.42%, outperforming the One Against All (OAA) strategy and LPCC-based models. Future enhancements include expanding emotional categories and incorporating broader datasets [11].

3. Methodology

This work presents an automated Speech Emotion Recognition (SER) using Python and machine learning techniques. It begins by preprocessing audio data to remove noise and extract features like MFCCs. These features are then scaled and used to train SVM, CNN and LSTM classifiers for emotion classification. Evaluation metrics and confusion matrices are generated to assess classifier performance. The system aims to provide accurate emotion detection in real-time, with potential applications in human-computer interaction and mental health monitoring.

MFCC: MFCC is a popular feature extraction technique most commonly used in speech and audio processing. It effectively represents the spectral properties of sound well-suited for speech recognition and music analysis. It is a set of coefficients that describes the shape of the sound signals power spectrum. In our SER system, we employ three different classifiers, SVM, CNN, and LSTM. Each classifier utilizes different techniques for emotion classification and has its own strengths and weaknesses.

Support Vector Machine (SVM): It is a supervised learning model used for both classification and regression tasks. It identifies the optimal hyperplane that effectively separates the data points of different

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2817-2827
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0415

classes within the feature space. SVM is effective in high-dimensional spaces and is robust against overfitting. However, it may not perform well with large datasets and complex nonlinear relationships between features.

Convolutional Neural Network (CNN): It is a deep learning model widely employed for image recognition and processing tasks. It consists of convolutional layers that extract spatial features from input data. CNN can be adapted for analyzing one-dimensional data like audio signals by treating them as spectrograms or time-frequency representations. While CNN is efficient in learning hierarchical representations from data, it may require larger datasets for training and can be sensitive to variations in input data.

Long Short-Term Memory (LSTM): LSTM is a type of recurrent neural network (RNN) architecture designed to model sequential data. It is well-suited for analyzing time-series data and has memory cells that can maintain information over time steps. LSTM is effective in capturing long-term dependencies in sequential data, making it suitable for analyzing audio signals. However, training LSTM models can be computationally expensive and requires careful tuning of hyperparameters.

3.1.Architecture

The simple SER architecture as shown in Figure 1consists of a few essential components. The process begins with capturing an input speech signal and the signal is then processed by a speech processing module for further analysis [1]. Raw acoustic data typically requires preprocessing (Butterworth Low-pass Filter), as it is not suitable for direct feature extraction. Without this step, the effectiveness of AI algorithm may be compromised due to issues such as noise, distortion, or insufficient information affecting the accuracy of learning models. Feature extraction plays a critical role in compressing and summarizing raw voice data by identifying the most relevant information. In signal processing, signals are generally categorized as stationary, where their properties remain constant over time, where these properties change continuously. Voice signals fall into the latter category due to ongoing variations in speech characteristics such as pitch, volume, and emotional tone. Capturing accurate frequency information from such dynamic signals is a challenge; therefore, feature extraction (MFCCs) is typically performed by segmenting the audio into short frames that can be treated as approximately stationary [2][3]. Feature scaling (StandardScaler) is essential after extracting the features and normalize the range of independent variables in a dataset, by transforming features to a common scale. It ensures that each variable contributes proportionally to the learning process, which is critical when features differ significantly in magnitude or measurement units. Without proper scaling, algorithms may incorrectly interpret features with larger numerical ranges as being more significant than those with smaller ranges, regardless of their actual importance. Once the features are scaled and normalized, the dataset is split into training and testing subsets to build and evaluate the predictive model. During the training phase, machine learning and deep learning algorithms learn from the patterns and relationships in the training data to classify different emotional states accurately. Commonly used models in SER include SVM, CNN, and LSTM. These models are trained iteratively by minimizing a suitable loss function and adjusting internal parameters to improve prediction performance. After training the models generate emotion prediction for unseen test samples, typically categorizing them into predefined emotional classes such as happiness, sadness, anger, or neutral. The performance is the system is measured using evaluation metrics such as accuracy, precision, F1-score, and confusion matrices, indicating the model's effectiveness in recognizing individual emotional states. (Figure 1)

3.2.Module description

The experimental workflow began with loading the training data, followed by separating the features and target variables. The target variable was encoded to a suitable format for classification. For the CNN model, the features were reshaped accordingly, and a CNN architecture was defined, compiled, and trained on the processed data. The trained CNN model was then saved. Similarly, features were reshaped for the LSTM model, and an LSTM network was defined, compiled, and trained using the same dataset, with the

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2817-2827
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0415

trained model subsequently saved. For the SVM classifier, the training data was reloaded, features and target variables were again separated and encoded, and the SVM model was initialized and trained. The resulting SVM model was saved for inference. During the prediction phase, the SVM, CNN, and LSTM models, along with the label encoder, were loaded. Input audio was received along with its sampling rate, and preprocessing steps, including noise reduction and audio segmentation, were performed if necessary. MFCC features were extracted and scaled. The processed features were then passed through each trained model, SVM, CNN, and LSTM to predict the emotion. Finally, the predicted outputs were mapped back to their corresponding emotion labels using the label encoder.

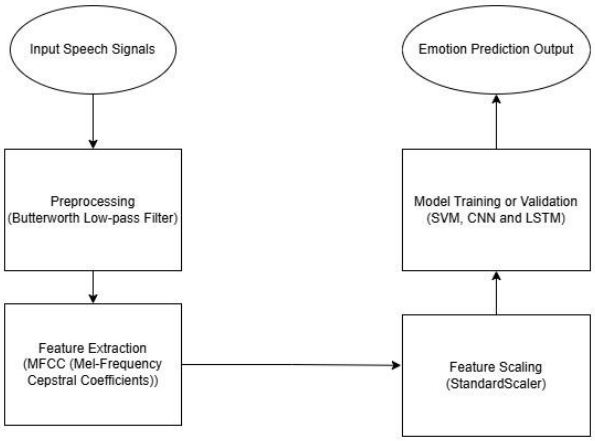

Figure 1 Speech Emotion Recognition Architecture

4. Experimental Studies and Discussion

4.1.Datasets

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is utilized as the dataset. It contains 1440 files derived from 60 recording trials per actor across 24 professional actors 12 males and 12 females. Each actor vocalizes two lexically matched statements using a neutral North American accent. The dataset captures a range of emotions including happy, sad, angry, fearful, surprise, calm, neutral, and disgusted, each expression is at two levels of emotional intensity (normal and strong). Every audio file in the dataset is uniquely named using a 7-part numerical identifier eg., (03-01-06-01-02-12.wav), which encodes specific stimulus characteristics [12]. The dataset used in this study comprises a total of 120 audio files, for the model training and testing, by dividing it into two parts 80% of the data (96 samples) was allocated for training, while the remaining 20% (24 samples) for testing. This split ensures the model is trained and tested for evaluating its generalization performance as shown in (Table 1)

Table 1 Distribution of Dataset for Training and Testing

DATASET	SIZE
Total dataset	120
Training	80%
Testing	20%

4.2.Implementations

Noise Reduction: In the noise reduction phase of the speech processing pipeline, a Butterworth low-pass filter was applied to mitigate high-frequency noise typically present in the audio signals. The filter was configured with a cutoff frequency of 2000 Hz and an order of 3 to ensure a smooth attenuation of undesired frequencies while preserving the important speech components. Audio signals were loaded using the Librosa library to retain the original sampling rate, and the low-pass filter was implemented via the scipy.signal module. After filtering, the processed audio was saved as cleaned audio as illustrated in (Figure 2)

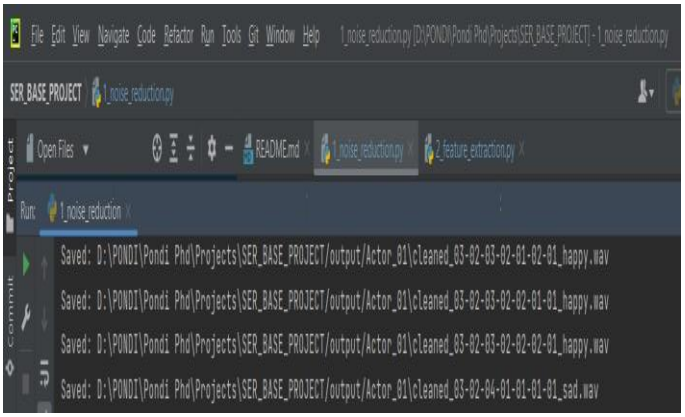

Figure 2 Noise reduction

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2817-2827
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0415

Feature Extraction: Following the noise reduction phase, effective feature extraction was carried out to SER. MFCCs were extracted from each audio signal, capturing key spectral properties that align with human auditory perception. The audio files were loaded using the Librosa library while preserving their original sampling rate to maintain data integrity. A total of 13 MFCCs were extracted per frame, and the mean value of each coefficient across all frames was calculated to form a consistent feature vector for each sample. Emotion labels were derived from the filenames using a predefined naming convention. The emotional labels were systematically organized and stored in an Excel file with the help of the panda's library and served as the foundation for training and evaluating the emotion classification model. as illustrated in (Figure 3)

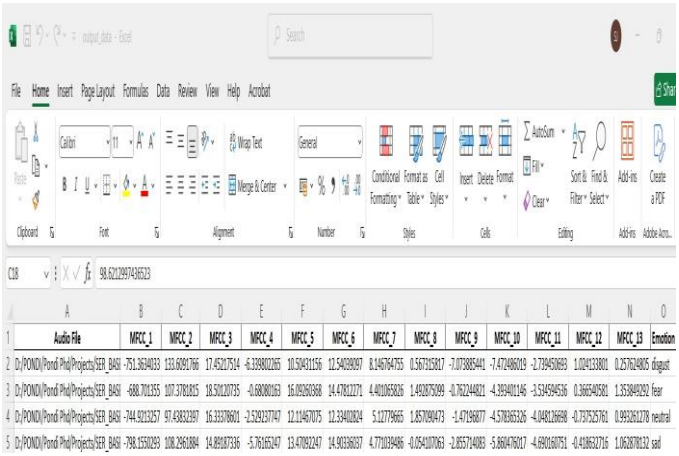

Figure 3 Feature Extraction

Feature Scaling: Following the feature extraction phase, MFCCs underwent normalization to standardize the data distribution and enhance the performance of subsequent classifiers. The extracted MFCC features were loaded from the feature extraction Excel file, and standardization was performed using the StandardScaler. This process transformed the MFCC values to have zero mean and unit variance, mitigating the impact of feature scale disparities. The normalized features were then compiled into a structured DataFrame along with their corresponding audio file names and emotion labels. The resulting dataset was saved as a new feature scaling Excel file and utilized for splitting into training and testing sets to enable robust model evaluation. as illustrated in (Figure 4)

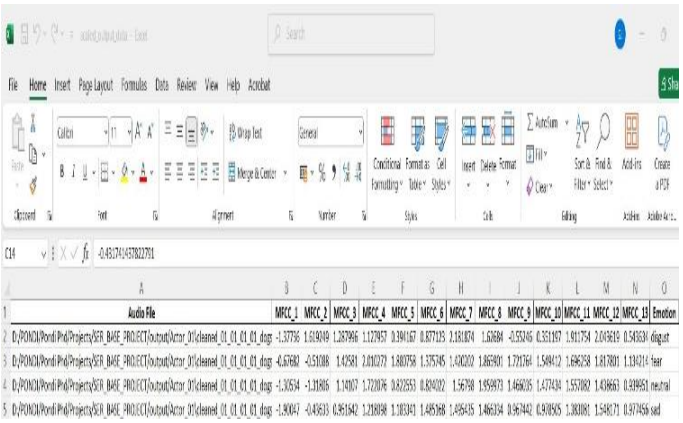

Figure 4 Feature Scaling

Data Splitting: Following the scaling of the MFCC features, the dataset is split into training and testing to support the training and evaluation of the emotion classification models. The data was randomly split into 80% as training and 20% as testing, ensuring a fair evaluation of model performance. This split was performed using the train_test_split function from scikit-learn, with a fixed random seed for reproducibility. The training and testing data were then saved into a train_test_data Excel file, with separate training and testing sheets for each subset, ensuring easy accessibility for future model training and validation tasks. This approach allows for a robust evaluation of model generalization by testing the trained models on unseen data. as illustrated in (Table 2)

Table 2 Data Split Overview (RAVDESS Subset)

Dataset	Number of Files	Percentage
Training Set	96	80%
Testing Set	24	20%
Total	120	100%

Classifier Selection: To recognize the emotion, we selected three classifiers SVM, CNN, and LSTM to assess the performance, with each model bringing distinct advantages suited to the characteristics of SER. SVM was used with a linear kernel, which is effective for high-dimensional spaces, such as MFCC

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2817-2827
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0415

features extracted from speech signals. The SVM classifier in this study was implemented using the SVC class from the sklearn.svm module with the kernel set to 'linear', making it well-suited for linearly separable datasets, which is often the case with carefully extracted features like MFCCs. The model was trained on the scaled MFCC features, which helped improve the performance by ensuring that all features contributed equally. Opting for a linear kernel allowed the model to achieve good balance between complexity and performance. As, SVM is known for its strong performance in high dimensional feature spaces and ability to manage both linear and non-linear decision boundary effectively. The CNN model was constructed using the Keras library and designed to capture spatial hierarchies in the input features. The model architecture consisted of two convolutional layers (Conv1D), each followed by max-pooling layers (MaxPooling1D), to reduce spatial dimensions while retaining important feature representations. The network architecture included 64 and 128 filters in the convolutional layers, with a kernel size of 3 and ReLU activation, followed by flattening the output and passing it through fully connected layers (Dense). The final layer employed a softmax activation function to classify the emotion categories. The CNN was compiled with the Adam optimizer and categorical crossentropy loss, and trained for 20 epochs with a batch size of 64. CNNs are well-suited for capturing local dependencies in data and performing well in tasks like SER, where spectral information plays a significant role in emotional recognition. LSTM was employed to model temporal dependencies in the speech signals. The LSTM model used in this study was designed with a single LSTM layer of 128 units, followed by a densely connected output layer with a softmax activation function. The LSTM model was compiled with the Adam optimizer and categorical cross-entropy loss, which were trained for 10 epochs with 32 batch size. This model excels at capturing sequential dependencies in time-series data, such as speech, which makes it highly effective for SER, where the temporal sequence of features play a crucial role in accurately identifying the emotions.

Model Training: Followed by the classifier selection, we train the selected classifier using the training data to learn patterns and relationships between input features and emotion labels. Prior to training, preprocessing steps such as feature scaling were applied to ensure uniform contribution from each feature, and target labels were encoded into numerical values to facilitate training. For each classifier, key hyperparameters such as batch size, learning rate, and the size of epochs were carefully selected based on empirical tuning and prior research. The models were optimized using loss functions suitable for classification tasks, such as categorical cross-entropy, and the Adam optimizer was employed to reduce the loss. The models were trained over several epochs with a batch size determined through experimentation. During training, the performance on the validation set was monitored to avoid overfitting. In the case of deep learning models like CNN and LSTM, techniques such as early stopping or dropout can be used to further mitigate overfitting. Once the models were trained, their effectiveness in recognizing emotion was evaluated on the testing dataset, and performance metrics such as precision, recall, f1-score, and support were evaluated to assess the model's performance in predicting emotion on unseen data.

Model Evaluation: The performance of the SVM, CNN, and LSTM classifiers was assessed using standard evaluation metrics: precision, recall, F1-score, and accuracy. Among the three models, the SVM demonstrated the most balanced and reliable performance, achieving an overall accuracy of 75%. It attained a macro-averaged precision of 0.79, recall of 0.75, and F1-score of 0.73, indicating effective generalization across multiple emotion classes. In contrast, the CNN model performed moderately with an accuracy of 50%, a macro-averaged precision of 0.52, a recall of 0.52, and an F1-score of 0.47, reflecting its relatively weaker capacity to capture distinctive features for emotion recognition. The LSTM model yielded the lowest performance, with an overall accuracy of approximately 25% and macro-averaged precision, recall, and F1-score of 0.17, 0.24, and 0.17, respectively. This underperformance may be attributed to insufficient training data or suboptimal hyperparameters for

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2817-2827
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0415

sequence learning. Notably, SVM consistently achieved high precision and recall for most emotion classes, including disgust, calm, and sad, whereas CNN and LSTM struggled particularly with classes like disgust and sad, often failing to identify them correctly. From the below Figure 5 A for SVM, Figure 5 B for CNN, and Figure 5 C for LSTM, suggest that although deep learning models have strong potential, traditional classifiers such as SVM can achieve better performance in scenarios with limited data or when the input features are effectively engineered. (Figure 5)

Metric/ Emotion	precision	recall	f1-score	support
angry	1	0.5	0.666667	4
calm	0.8	1	0.888889	4
disgust	1	1	1	2
fear	1	0.5	0.666667	2
happy	0.666667	0.8	0.727273	5
neutral	0.333333	0.5	0.4	2
ps	0.5	1	0.666667	2
sad	1	0.714286	0.833333	7
accuracy	0.75	0.75	0.75	0.75
macro avg	0.7875	0.751786	0.731187	28
weighted avg	0.828571	0.75	0.755664	28

Figure 5 SVM Model Evaluation A

Metric/Emotion	precision	recall	f1-score	support
angry	1	0.75	0.857143	4
calm	0.5	0.5	0.5	4
disgust	0	0	0	2
fear	0.25	1	0.4	2
happy	0.571429	0.8	0.666667	5
neutral	1	0.5	0.666667	2
ps	0.5	0.5	0.5	2
sad	0.333333	0.142857	0.2	7
accuracy	0.5	0.5	0.5	0.5
macro avg	0.519345	0.524107	0.47381	28
weighted avg	0.52466	0.5	0.47483	28

Figure 6 CNN Model Evaluation B

Metric/Emotion	precision	recall	f1-score	support
angry	0.2	0.5	0.285714	4
calm	0.5	0.75	0.6	4
disgust	0	0	0	2
fear	0.166667	0.5	0.25	2
happy	0	0	0	5
neutral	0	0	0	2
ps	0	0	0	2
sad	0.5	0.142857	0.222222	7
accuracy	0.25	0.25	0.25	0.25
macro avg	0.170833	0.236607	0.169742	28
weighted avg	0.236905	0.25	0.199943	28

Figure 7 LSTM Model Evaluation C

Training Models: Three machine learning models, SVM, CNN, LSTM are trained to recognize emotions from speech data. The training process involves loading training data, separating features and labels, encoding the target variable, and fitting the models to the training data. Once trained, these models are saved for integration into the overall SER system. An ensemble model, also known as a combination model, is a machine learning technique that combines the predictions from multiple individual models to produce a single prediction. In the context of the SER system, an ensemble model could be constructed by combining the predictions from the SVM, CNN, and LSTM models. This ensemble approach often results in improved prediction accuracy and robustness compared to individual models.

4.3.Gradio User interface

As illustrated in Figure 6 A and Figure 6 B Gradio is used as a user interface framework for creating an interactive web-based interface for the SER system. Gradio simplifies the process of building and deploying machine learning models by providing a high-level interface that allows developers to create user-friendly interfaces with minimal code.

4.4.Utilization of Gradio

Interface Creation: Users can upload an audio file to the Gradio user interface to predict emotions.

Input Handling: Gradio handles the input received from the user in the form of an audio file.

Output Display: Gradio displays the predicted emotion as the output of the SER system.

Launching the Interface: Gradio's launchb() function is used to start the web-based interface, making it accessible to users via a web browser.

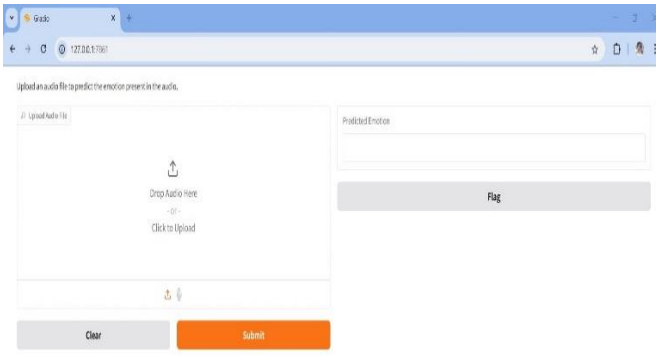

Figure 8 Gradio User Interface A

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2817-2827
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0415

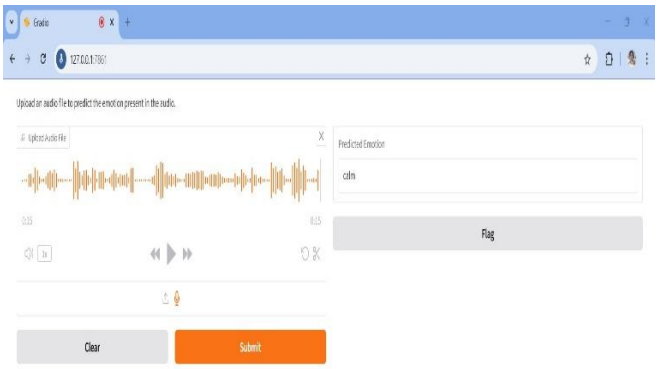

Figure 9 Gradio User Interface B

4.5.Discussion

Table 3 provides a summary of the evaluation metrics, and as discussed in the model evaluation section, the SVM classifier outperformed CNN and LSTM across all evaluation metrics. To further understand the behavior of each model, confusion matrices were analyzed for deeper insights into their prediction pattern across different emotion classes.

The corresponding confusion matrix Figure 7 reveals that SVM correctly classified the majority of samples for key emotions such as calm, happy, and sad, with minimal misclassifications. This indicates its robustness and reliability in distinguishing between different emotional states. SVM achieved particularly good performance in this study due to its ability to handle high-dimensional, linearly separable data effectively. The MFCC features, which represent the spectral properties of speech, are often well-separated for different emotions, especially when preprocessing steps like feature scaling are applied. The linear kernel makes it capable of learning the boundaries of decisions between various emotions and not overfitting to noisy data, as required for real-world SER applications. Its performance on the testing showed a good balance between bias and variance and is a good choice for this application. On the other hand, the CNN model delivered moderate performance, with 50% accuracy and an F1-score of 47%. As shown in its confusion matrix Figure 8, the model exhibited notable confusion across classes, especially misclassifying sad and calm as fear and neutral as calm. While CNNs are capable of learning spatial hierarchies from features, the small dataset likely limited their ability to effectively generalize emotional patterns. The LSTM model recorded the lowest performance, with only 25% accuracy and a 17% F1-score. The confusion matrix Figure 9 illustrates that the model frequently misclassified several emotions, most notably sad as fear or calm , and happy as fear. Even though LSTM is capable at capturing temporal dependencies in sequence data, the small sample size might have made it difficult for the model to learn time-dependent patterns of emotional expression effectively, resulting in poor performance. This comparative analysis suggests that in resource-limited environments with limited labeled data, traditional classifiers a such as SVM, when combined with effective preprocessing and feature extraction can outperform more complex deep learning models. (Figure 10)

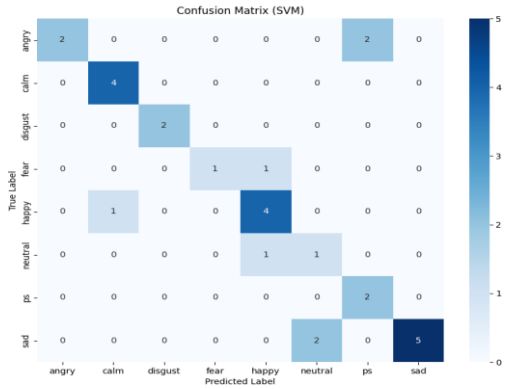

Figure 10 Confusion Matrix of SVM Classifier

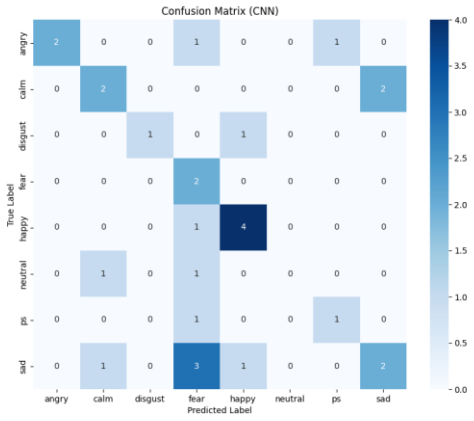

Figure 11 Confusion Matrix of CNN Classifier

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2817-2827
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0415

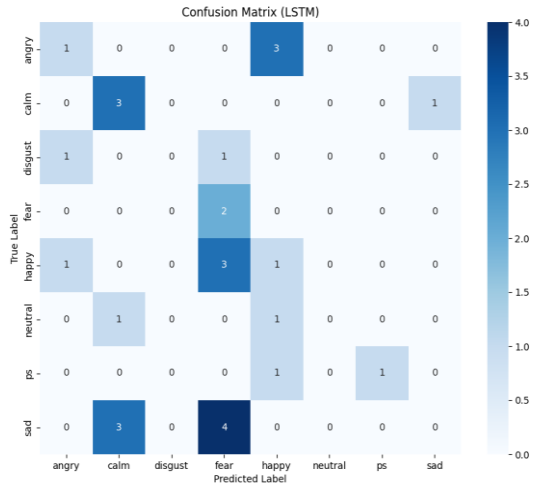

Figure 12 Confusion Matrix of LSTM Classifier

These results show that the SVM model is more effective at handling the classification task, likely due to its ability to handle smaller datasets and high-dimensional feature spaces efficiently. On the other hand, the relatively lower performance of CNN and LSTM suggests that these deep learning models may require a larger amount of training data and potentially further hyperparameter tuning to outperform traditional machine learning approaches such as SVM. (Table 3)

Table 3 Performance Comparison of SVM, CNN and LSTM Models

CRITERIA	SVM	CNN	LSTM
Accuracy	75%	50%	25%
Precision	79%	52%	17%
Recall	75%	52%	24%
F1-Score	73%	47%	17%
Support	28%	28%	28%

5. Conclusion and Future Enhancement

5.1.Conclusion

Based on the comparative accuracy results of different models, Support Vector Machine (SVM) has proven to be the most accurate classifier for our SER task, with an accuracy of 0.75, outperforming both CNN's and LSTM model, which attained accuracies of 0.5 and 0.25, respectively. This high accuracy indicates that SVM effectively separates different emotional states in the feature space and performs well on our dataset. Given its superior performance, a robust SER system was developed exclusively based on the SVM model, this system aims to improve the accuracy and reliability of emotion detection from audio signals. The SVM model's high accuracy ensures that it effectively captures and distinguishes the emotional nuances present in speech. To conclude, the SVM model has proven to be effective in enhancing the accuracy of SER. The Gradio interface makes the system user-friendly and accessible, allowing users to easily interact with the model and obtain emotion predictions. This project demonstrates the potential of using advanced machine learning techniques for real-world applications in emotion detection from audio signals. Future enhancement For future enhancements, the SER system could explore real-time emotion recognition capabilities, integrating multimodal inputs such as facial expressions with speech analysis. Additionally, incorporating adaptive learning algorithms to personalize emotion detection for individual users would enhance accuracy and user engagement. Emotion generation techniques could be explored to enable the system to respond empathetically. Cross-cultural emotion recognition models could be developed to ensure applicability across diverse populations. Furthermore, integrating privacy-preserving techniques and continuous monitoring mechanisms would strengthen user trust and data security. These enhancements aim to elevate the system's effectiveness, applicability and user experience in various domains.

References

[1]. Kanna, P. R., & Kumararaja, V. (2024). Enhancing Speech Emotion Detection With Windowed Long-Term Average Spectrum and Logistic-Rectified Linear Unit. Engineering Applications of Artificial Intelligence, 137, 109103.

[2]. Min, D. J., & Kim, D. H. (2024). Speech Emotion Recognition via Sparse Learning-based Fusion Model. IEEE Access.

[3]. Xu, C., Liu, Y., Song, W., Liang, Z., & Chen, X. (2024). A new network structure for speech emotion recognition research. Sensors, 24(5), 1429.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 06 June 2025
Page No: 2817-2827
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0415

[4]. Deepika, C., & Kuchibhotla, S. (2024). Deep-CNN based knowledge learning with Beluga Whale optimization using chaogram transformation using intelligent sensors for speech emotion recognition. Measurement: Sensors, 32, 101030.
[5]. Akinpelu, S., Viriri, S., & Adegun, A. (2024). An enhanced speech emotion recognition using vision transformer. Scientific Reports, 14(1), 13126.
[6]. Chowdhury, J. H., Ramanna, S., & Kotecha, K. (2025). Speech emotion recognition with light weight deep neural ensemble model using hand crafted features. Scientific Reports, 15(1), 11824.
[7]. Jena, S., Basak, S., Agrawal, H., Saini, B., Gite, S., Kotecha, K., & Alfarhood, S. (2025). Developing a negative speech emotion recognition model for safety systems using deep learning. Journal of Big Data, 12(1), 54.
[8]. Ahmad, R., Iqbal, A., Jadoon, M. M., Ahmad, N., & Javed, Y. (2024). XEmoAccent: Embracing Diversity in Cross-Accent Emotion Recognition using Deep Learning. IEEE Access.
[9]. Banga, A., Baheti, B., Sachdev, D., & Jajoo, Y. Speech Emotion Detection Using State of the Art CNN and LSTM.
[10]. Leelavathi, R., Deepthi, S. A., & Aruna, V. (2021). Speech emotion recognition using lstm. International Research Journal of Engineering and Technology.
[11]. Jain, M., Narayan, S., Balaji, P., Bhowmick, A., & Muthu, R. K. (2020). Speech emotion recognition using support vector machine. arXiv preprint arXiv:2002.07590.
[12]. Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https:// doi.org/ 10.1371/journal.pone.0196391