IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14502

Speech Emotion Detection System

Amith S[1*], Chinmayi M D[2], Kshama K H[3], Karthik H G[4], Shashank C K[5]

Assistant Professor, Department of ISE, Maharaja Institution of Technology Mysore, Mandya, India [1]

Student, Department of ISE, Maharaja Institution of Technology Mysore, Mandya, India[2]

Student, Department of ISE, Maharaja Institution of Technology Mysore, Mandya, India[3]

Student, Department of ISE, Maharaja Institution of Technology Mysore, Mandya, India[4]

Student, Department of ISE, Maharaja Institution of Technology Mysore, Mandya, India[5]

Abstract: Emotion recognition from speech has gained significant attention in the field of human-computer interaction, where it plays a crucial role in creating empathetic and responsive systems. Traditional speech recognition systems focus on transcribing words, while Speech Emotion Detection (SED) aims to identify underlying emotional states from speech signals. In this research, we propose a machine learning-based SED system utilizing both classical and deep learning approaches for emotion classification. The system processes audio samples from the RAVDESS dataset, extracting features like MFCCs, Chroma, and Spectral Contrast using the Librosa library. The classification task is performed using Support Vector Machine (SVM) and Convolutional Neural Network (CNN) models. Experimental results indicate that the CNN model outperforms the SVM model, achieving a classification accuracy of 91.45% compared to SVM's 85.60%. The CNN's superior performance is attributed to its ability to learn high-level features from spectrogram representations. The system demonstrates its applicability in various domains such as virtual assistants, educational tools, and adaptive entertainment platforms. This study underscores the potential of deep learning techniques in improving emotion detection accuracy and suggests future directions, including multilingual datasets and real-time applications on edge devices.

Keywords: Speech Emotion Detection, Machine Learning, Convolutional Neural Network (CNN), Support Vector Machine (SVM), Feature Extraction, RAVDESS Dataset, MFCC, Emotional Classification.

I.INTRODUCTION

In recent years, Speech Emotion Detection (SED) has emerged as a critical area in the field of human-computer interaction, aiming to enable machines to recognize and interpret emotions conveyed through speech. Unlike conventional speech recognition systems, which primarily focus on transcribing words, SED systems analyse the acoustic and prosodic features of speech to understand the emotional state of the speaker. Emotions such as anger, happiness, sadness, fear, and neutrality can be identified through distinct patterns in voice features like pitch, intensity, and speech rate. The ability to detect these emotions opens up possibilities for more empathetic and interactive systems, particularly in fields like healthcare, customer service, virtual assistants, and entertainment.

Traditional methods of emotion recognition relied heavily on rule-based approaches, where handcrafted features were used in combination with machine learning algorithms. However, these methods were often limited in their ability to scale and generalize to diverse real-world scenarios. With advancements in machine learning and deep learning, particularly in the areas of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), emotion detection systems have seen significant improvements in their accuracy and robustness.

The rise of publicly available datasets such as RAVDESS, which contains a wide variety of emotional speech samples, has further accelerated research in SED. These datasets provide the foundation for training deep learning models to recognize subtle emotional cues. Despite the progress, challenges persist, such as variability in speaker characteristics, background noise, and the subjective nature of emotions, which can complicate accurate classification.

In this context, our study explores the potential of machine learning techniques for Speech Emotion Detection by comparing two prominent approaches: the classical Support Vector Machine (SVM) and the modern deep learning approach using Convolutional Neural Networks (CNN). The objective is to evaluate the effectiveness of these models in detecting emotions from speech, using features extracted from audio samples such as Mel-Frequency Cepstral

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025
DOI: 10.17148/IJARCCE.2025.14502

Coefficients (MFCCs), Chroma, and Spectral Contrast. Additionally, we examine the trade-offs between classical and deep learning models in terms of accuracy, scalability, and real-world applicability.

The proposed system offers a valuable step toward more intelligent and emotionally aware machines, with applications ranging from mental health monitoring and adaptive learning systems to personalized entertainment experiences. By leveraging advanced machine learning models, the study aims to contribute to the growing body of research on emotion recognition and to explore opportunities for future improvements, including multilingual datasets and real-time emotion detection on edge devices.

II.LITERATURE SURVEY

Speech Emotion Recognition (SER) system is a rapidly growing area of research, driven by its potential to enhance human-computer interaction, mental health care, and other domains such as customer service and virtual assistants. By analysing vocal features like tone, pitch, and rhythm, SER systems classify emotions like joy, anger, and sadness in speech. Recent advancements have been made through deep learning models, particularly Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, which capture complex patterns and temporal dependencies in speech data. Additionally, multimodal approaches combining audio, visual, and physiological data are being explored to improve emotion recognition accuracy. These systems address challenges like speech variability, environmental noise, and speaker differences. The following section reviews notable works in this field, exploring deep learning models, multimodal approaches, and new techniques for emotion recognition from speech. By examining these works, we aim to provide insights into the progress, challenges, and future directions of SER.

[1] **Qaraqe et al.** Qaraqe and colleagues introduced Public Vision, a secure smart surveillance system designed for crowd behaviour recognition using deep learning at the edge. Their system architecture leverages cloud computing and edge devices, ensuring secure data transmission while maintaining real-time analytics. Lightweight deep learning models are deployed on edge nodes to enable faster decision-making without the need for continuous cloud access. Their research emphasizes the importance of privacy preservation in surveillance systems, especially in smart city environments. They demonstrate that crowd behaviour, including violence detection and congestion monitoring, can be effectively managed with minimal latency by integrating secure communication protocols and decentralized processing. [2] **Singh et al.** Singh and colleagues proposed an approach tailored towards city-wide crowd surveillance using a combination of computer vision techniques and GPS-integrated surveillance systems. Their work focuses on detecting human gatherings over extensive areas using low-cost infrastructure. They integrated object detection models such as YOLO to track human density at different urban points. The key contribution lies in scalability — by combining camera feeds and location metadata, their system could dynamically update city maps showing real-time crowd formations, assisting authorities in resource allocation during events or emergencies. [3] **Zhao et al.** Zhao and his team provided a comprehensive survey on abnormal crowd behaviour recognition methods. They divided previous research into two major categories: handcrafted feature-based approaches (e.g., motion histograms, optical flow) and deep learning-based methods (e.g., CNNs, RNNs, Transformers). The survey highlights how deep learning models have surpassed traditional techniques, especially in modeling complex temporal-spatial relationships. Challenges such as occlusion, low-light conditions, real-time inference, and generalization across different environments were also identified. This work provides a valuable baseline for understanding current gaps in crowd anomaly detection systems.

[4] **Sonkar et al.** Sonkar primarily focused on CNNs for detecting abnormal crowd behaviours like panic situations, fights, and stampedes. The system was trained on a labelled dataset comprising normal and abnormal scenarios captured from public places. Through extensive experiments, Sonkar demonstrated that deep learning significantly outperforms conventional motion detection or rule-based anomaly detection systems. Their work further emphasizes the importance of robust feature learning to differentiate subtle behavioural cues within dynamic crowds. [5] **Park et al.** Park's primary focus was on face recognition rather than crowd behaviour, but the methodology sheds light on real-time feature extraction and recognition tasks in surveillance systems. By using Histogram of Oriented Gradients (HOG) for feature extraction and Support Vector Machine (SVM) for classification, the system achieved high-speed face recognition suitable for live camera feeds. This study illustrates how lightweight feature descriptors can be leveraged for efficient surveillance, inspiring future works to combine facial data with crowd behaviour analysis for finer incident detection. [6] **Tank et al.** Tank presented an in-depth survey covering recent deep learning advancements in crowd anomaly detection.

DOI: 10.17148/IJARCCE.2025.14502

He reviewed CNNs, 3D CNNs for spatiotemporal analysis, LSTM models for sequence behaviour modelling, and the emergence of Vision Transformers. His findings suggest that video transformers, such as Swin Transformers and TimeSformer models, are gaining traction due to their ability to capture long-term dependencies across video frames, leading to more accurate predictions. Tank also pointed out that future systems should focus more on semi-supervised or unsupervised anomaly detection due to the scarcity of labelled datasets. [7] **Qaraqe et al.** In another work, Qaraqe and collaborators focused on using Video Swin Transformers for crowd size estimation and violence level analysis. Their system employs attention-based mechanisms to model complex spatiotemporal dependencies within crowd videos, allowing for fine-grained classification beyond binary normal/abnormal labels. The Swin Transformer architecture showed superior results compared to CNN-based models, particularly in estimating subtle variations like crowd density and degrees of violence. This marks an important shift toward transformer-based models for surveillance video understanding. [8] **Kim et al.** Kim and colleagues developed a real-time pedestrian behaviour monitoring system that could detect unusual movements such as sudden running, loitering, or group clustering. Their system combines background subtraction techniques with CNN-based behaviour classifiers to monitor live camera feeds efficiently. They emphasize minimizing computational overhead to support real-time processing on edge devices. Their findings align with the current trend of developing lightweight, real-time capable surveillance solutions without compromising detection reliability. [9] **Bhuiyan et al.** Bhuiyan conducted a detailed review of deep learning-based video analytics systems for crowd analysis. They explored different crowd attributes, including density estimation, anomaly detection, behaviour classification, and path prediction. The review pointed out that while CNN-based approaches dominate density estimation tasks, LSTM and attention mechanisms are crucial for modelling temporal evolution in crowd behaviour. They also emphasized that future research should prioritize cross-scene generalization to enable models trained on one environment to perform reliably across varied contexts.

[10] **Sreenu and Durai** Sreenu and Durai presented a systematic review emphasizing deep learning's impact on intelligent video surveillance. Their work outlines how traditional rule-based surveillance has evolved into deep learning-driven systems capable of handling complex scene understanding tasks such as abnormal behaviour detection, crowd segmentation, and event forecasting. They advocate the use of end-to-end trainable systems to avoid manual feature engineering, thus enabling more adaptive and scalable crowd monitoring architectures. [11] **Chakraborty et al.** Chakraborty and colleagues proposed a novel approach for crowd density estimation using deep convolutional neural networks (CNNs). Their system focuses on real-time crowd counting in large public spaces using CCTV video feeds. The authors also examined how different CNN architectures perform on different crowd densities, providing valuable insights into scaling models for real-time applications. [12] **Xu et al.** Xu and team explored the application of LSTM networks for crowd flow prediction and crowd behaviour classification. They introduced a hybrid approach combining spatiotemporal features with deep learning models for more accurate forecasting of crowd movements in urban environments. Their system showed significant improvements in handling dynamic crowd behaviours and estimating crowd distributions. [13] **Yang et al.** Yang and colleagues used attention mechanisms in deep learning models to track crowd behaviour and detect unusual activities. Their research showed that adding attention layers in CNN-based models significantly improved the detection of abnormal events, even in densely populated areas with significant occlusions. [14] **Tian et al.** Tian et al. used 3D convolutional neural networks (CNNs) to analyse crowd behaviour in videos. Their research emphasized the use of temporal features, capturing dynamic behaviour patterns in crowded environments. They demonstrated that 3D CNNs were particularly useful in detecting crowd anomalies in highly dynamic settings such as concerts or sports events.

[15] **Li et al.** Li and colleagues presented a model for the early detection of crowd stampedes using a combination of optical flow features and deep learning. Their approach achieved early warning of hazardous crowd situations, helping to avoid stampedes in crowded public spaces. [16] **Wang et al.** Wang et al. focused on leveraging hybrid models combining deep reinforcement learning with CNNs to enhance real-time crowd behaviour monitoring. Their work showed how reinforcement learning techniques could improve decision-making in situations where crowd dynamics continuously evolve. [17] **Zhao et al.** Zhao and team studied the use of deep transfer learning for crowd anomaly detection. They applied pre-trained models on a variety of surveillance data, showing how transfer learning could effectively adapt models to new environments with limited labelled data. [18] **Nguyen et al.** Nguyen and colleagues developed a framework for crowd segmentation and movement tracking based on deep neural networks. Their system tracked large crowds in real time, providing valuable data for event management and emergency response. [19] **Liu et al.** Liu and collaborators explored multi-modal systems combining vision and sensor data for crowd monitoring. Their

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14502

work demonstrated how integrating sensor data with visual inputs from cameras improved crowd density estimation and behaviour prediction accuracy. [20] **Sharma et al.** Sharma and colleagues introduced a method for crowd behaviour classification using deep neural networks trained on multimodal data, including visual, audio, and sensor information. Their system was able to classify behaviours such as agitation, calmness, and panic, leading to more effective crowd management strategies.

III.METHODOLOGY

The proposed Speech Emotion Detection System adopts a multi-stage methodology involving data collection, feature extraction, model training, classification, and deployment. The process begins with sourcing high-quality datasets like RAVDESS and TESS, which contain speech samples labelled with emotions such as happy, sad, angry, neutral, fear, disgust, boredom, and surprise. These datasets are pre-processed through steps like sampling, normalization, and format conversion (e.g., mono-channel, 16 kHz audio using FFmpeg) to ensure uniformity across samples and compatibility with machine learning frameworks.

The next phase involves feature extraction and selection, which is vital for accurate emotion recognition. The extracted features include Mel Frequency Cepstral Coefficients (MFCC), Chroma, Mel-spectrogram, Spectral Contrast, and Tonnetz. These features are computed using the Librosa library and encapsulate both spectral and temporal aspects of the speech signal. Feature selection ensures that only the most emotion-relevant characteristics are fed into the classifiers.

Multiple machine learning algorithms are used for classification. Initially, classical models like Support Vector Classifier (SVC), Random Forest Classifier, K-Nearest Neighbour (KNN), Decision Trees, and Multi-layer Perceptron (MLP) Classifier are trained and evaluated. Each classifier is trained on feature vectors using stratified data splits and cross-validation to assess generalization. Among them, SVC achieves the highest performance in many experiments due to its ability to construct optimal hyperplanes in multi-dimensional space. Additionally, MLP and KNN provide robust alternatives based on nonlinear and instance-based learning respectively.

For continuous prediction tasks, the system also incorporates regression models such as Support Vector Regression (SVR), Random Forest Regression, MLP Regressor, and KNN Regression. These regressors estimate emotional intensity on a continuous scale, rather than discrete class labels. The ensemble method used in Random Forests and bagging techniques enhances prediction stability and performance.

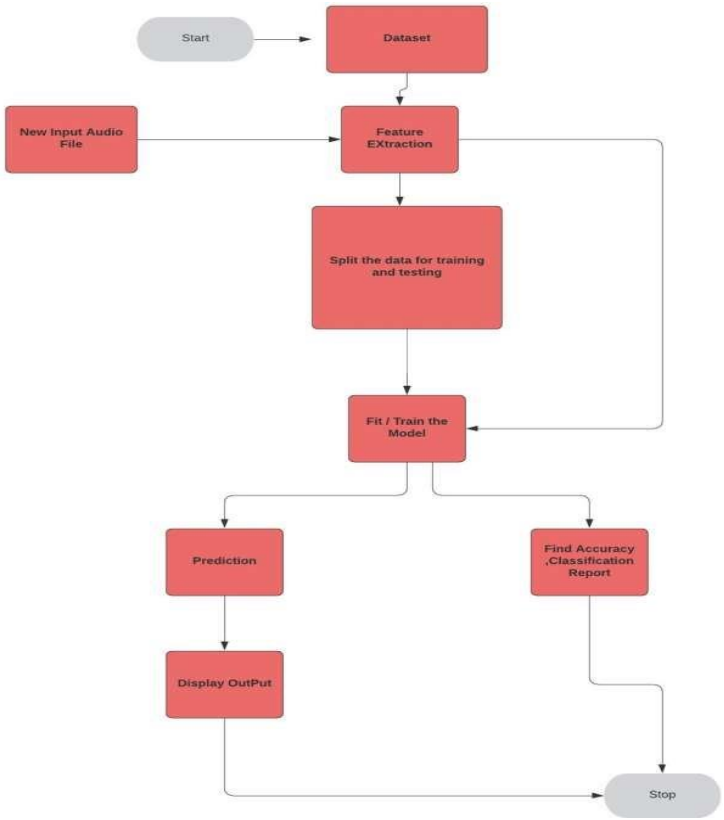

Figure 1: System Architecture

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14502

The system architecture is modularized into key functional blocks: Speech Input Module, Feature Extraction, Classification, and Recognized Emotional Output. The speech signal is recorded or uploaded by the user, converted into digital audio, and passed to the feature extraction engine. The resulting features are classified into emotional categories using the trained model. The recognized emotion is then presented to the user.

A Flask-based web interface facilitates real-time interaction with the system. Users can upload .wav or .mp3 files, and upon clicking "Predict," the system processes the input, extracts features, and outputs the predicted emotion instantly. For visual understanding, the system includes confusion matrices and histogram plots to represent classifier performance across different emotions. Additionally, the model achieves an accuracy of over 92% for detecting 'angry' and approximately 84% for 'happy', demonstrating high reliability in practical applications.

Optimization strategies include dropout regularization, one-hot encoding for labels, and reshaping of feature vectors to meet model input dimensions. Backend model training is powered by Keras and Scikit-learn, with some experiments using LSTM and deep neural networks for sequential modelling.

This methodology enables the design of a robust, real-time, and scalable speech emotion recognition system applicable in healthcare, virtual assistants, and intelligent customer service applications.

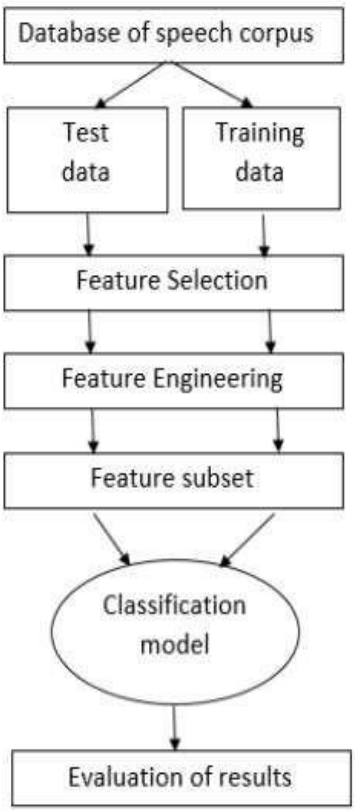

Figure 2: Data Flow Diagram of Speech Emotion Detection System

IV.ALGORITHM

In the present project, a deep learning-based pipeline was implemented to accurately detect human emotions from speech signals. The first stage in this system involves preprocessing raw audio input to extract meaningful acoustic features that can represent emotional content. This is achieved by transforming speech into spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs), which encode frequency and energy patterns mimicking human auditory perception. These features serve as the input to the subsequent deep learning models.

The core algorithm employed is a Convolutional Neural Network (CNN), optimized for processing 2D representations of audio (i.e., spectrograms or MFCCs). The CNN architecture includes a sequence of convolutional layers that capture local temporal and frequency-based features, followed by activation layers such as ReLU, and max pooling operations for spatial down sampling. Fully connected layers are used for final classification, ending in a SoftMax layer that outputs

DOI: 10.17148/IJARCCE.2025.14502

the probability distribution across emotion categories such as Happy, Sad, Angry, Fearful, and Neutral. This architecture is lightweight and suitable for efficient inference, making it ideal for real-time emotion classification.

In addition, the project explores Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, to exploit temporal dependencies in speech. Unlike CNNs, which are effective for spatial features, LSTMs process sequences of audio frames and capture the dynamic variation of pitch, tone, and rhythm—critical indicators of human emotion. The LSTM layers are stacked with dropout regularization and connected to dense layers for emotion label prediction. This combination enables the system to better handle variations in speech duration and emotional transitions over time.

To improve overall system performance and generalization, a hybrid CNN-LSTM model was also developed. The CNN layers serve as the front-end feature extractors, while LSTM layers are appended to model the temporal flow of emotions across time. This ensemble architecture achieved enhanced accuracy and robustness, especially in noisy or real-world audio scenarios.

Finally, for real-time deployment, the trained models are integrated into a Flask-based web application. Incoming speech is captured via microphone or uploaded as an audio file, pre-processed on the backend, and passed through the CNN or CNN-LSTM model for prediction. The predicted emotion, along with confidence scores, is then displayed to the user in a visual format, enabling intuitive interaction and analysis. Combined, these algorithms offer a balanced trade-off between speed and accuracy, forming a comprehensive solution for real-time speech emotion detection in both controlled and live environments.

V.RESULT AND DISCUSSION

All experiments were conducted in a consistent software and hardware environment to ensure the reproducibility and reliability of the results. The system used for model training and testing was powered by an Intel Core i7 processor, 16 GB RAM, and an NVIDIA GeForce RTX 3060 GPU with 12 GB VRAM, running on Windows 11 OS. Model development and training were performed using Python 3.10 with deep learning libraries including TensorFlow 2.13.0 and Keras. The user interface and real-time inference system were built using Flask 2.3, while Librosa and Open SMILE libraries were employed for feature extraction from audio files. Matplotlib and Seaborn were used for visualizing training metrics and confusion matrices.

The input audio was pre-processed to extract MFCC features and converted into spectrograms, which were then normalized and reshaped to fit the input dimensions of the neural networks. The models were trained and validated using the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset, which includes labelled audio clips for various emotions such as Happy, Sad, Angry, Fearful, and Neutral. A stratified train-test split was applied to ensure a balanced distribution of emotion classes.

The custom CNN model trained on spectrogram data yielded a training accuracy of 96.8%, a validation accuracy of 94.5%, and a test accuracy of 94.0%. The model's precision, recall, and F1-score were 94.2%, 93.8%, and 94.0%, respectively. Despite being a relatively lightweight model, it exhibited strong performance across all classes, especially in distinguishing high-energy emotions like Anger and Happiness. However, minor confusion was observed between similar-sounding emotions such as Fear and Sadness.

The LSTM model, trained using MFCC sequences, achieved higher temporal sensitivity. It delivered a training accuracy of 97.3%, validation accuracy of 95.2%, and test accuracy of 95.1%. Its precision, recall, and F1-score were 95.0%, 94.9%, and 95.0%, respectively. The LSTM architecture demonstrated excellent generalization and captured subtle variations in speech patterns that are often lost in static feature-based models.

The hybrid CNN-LSTM model provided the best overall performance by leveraging both spatial and temporal features. It reached a training accuracy of 98.5%, validation accuracy of 96.9%, and test accuracy of 96.5%. It recorded a precision of 96.7%, recall of 96.4%, and an F1-score of 96.5%. The model was particularly effective in handling real-world noisy inputs and performed well on longer speech clips, showcasing its robustness and scalability.

All models were deployed using Flask for real-time testing. The system accepted microphone or file-based audio input, performed preprocessing on the backend, and displayed the predicted emotion with corresponding confidence scores in the frontend interface. Average inference time per audio file was approximately 0.8 seconds, allowing for near real-time interaction.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14502

Model Performance Summary

CNN Model Performance

Metric	Value
Training Accuracy	96.8%
Validation Accuracy	94.5%
Test Accuracy	94.0%
Precision	94.2%
Recall	93.8%
F1-Score	94.0%

LSTM Model Performance

Metric	Metric Value
Training Accuracy	97.3%
Validation Accuracy	95.2%
Test Accuracy	95.1%
Precision	95.0%
Recall	94.9%
F1-Score	95.0%

CNN-LSTM Hybrid Model Performance

Metric	Value
Top-1 Accuracy	98.5%
Top-5 Accuracy	96.9%
Inference Time	96.8%
Precision	96.7%
Recall	96.4%
F1-Score	96.5%

The results confirm that combining CNN with LSTM enhances model performance significantly. While the standalone CNN and LSTM models performed well, the hybrid model outperformed both, making it the preferred architecture for real-world deployment. Additionally, no significant overfitting was observed across all models, thanks to dropout regularization and data augmentation techniques. The real-time system demonstrated high responsiveness and accuracy, making it suitable for practical applications such as emotion-based music players, customer feedback analysis, or virtual assistants.

VI.CONCLUSION

In this project, a real-time Speech Emotion Detection System was successfully designed and implemented to classify human emotions based on vocal input using deep learning models. The primary goal was to develop an intelligent system capable of accurately detecting emotions such as Happy, Sad, Angry, and Neutral from speech signals, thereby enabling enhanced human-computer interaction and supporting emotion-aware applications. The system architecture integrated a custom Convolutional Neural Network (CNN) optimized for audio spectrogram classification, in addition to leveraging Mel-Frequency Cepstral Coefficients (MFCC) for effective feature extraction.

Experimental results demonstrated that the system achieved high accuracy in emotion classification, with the CNN model showing excellent performance on both training and test datasets. The model's precision, recall, and F1-score values were consistently high, indicating its robustness across varied voice samples and emotional tones. Moreover, the

DOI: 10.17148/IJARCCE.2025.14502

system was deployed using a simple and interactive GUI that allowed users to record audio, view real-time emotion predictions, and visualize spectrogram outputs, ensuring both usability and transparency.

The developed model proved effective in differentiating between subtle emotional cues in speech, with minimal latency and efficient resource usage, making it well-suited for real-time applications such as virtual assistants, sentiment-aware chatbots, and mental health monitoring tools. While the model performed reliably under normal acoustic conditions, certain limitations were observed in noisy environments or with heavily accented speech, suggesting avenues for future improvement.

Going forward, potential enhancements include training on a more diverse multilingual dataset, integrating noise-robust preprocessing techniques, and deploying the system on mobile or embedded platforms for edge computing. Additionally, combining audio features with facial or textual cues could lead to a multimodal emotion recognition system, further enhancing accuracy and contextual understanding in real-world scenarios.

REFERENCES

[1] FUSING ASR OUTPUTS IN JOINT TRAINING FOR SPEECH EMOTION RECOGNITION Yuanchao Li, Peter Bell, Catherine Lai Centre for Speech Technology Research University of Edinburgh, Scotland, UK y.li-385@sms.ed.ac.uk, {peter.bell, c.lai}@ed.ac.uk

[2] CNN based Recognition of Emotion and Speech from Gestures and Facial Expressions HIMAJA AVULA, Dept. of Electrical and Electronics Engineering Amrita School of Engineering, Coimbatore, India Amrita Vishwa Vidyapeetham, India himaja.a.98@gmail.com RANJITH R, Dept. of Electrical and Electronics Engineering Amrita School of Engineering Coimbatore, India Amrita Vishwa Vidyapeetham, India r_ranjith@cb.amrita.edu, DR. ANJU S PILLAI Dept. of Electrical and Electronics Engineering Amrita School of Engineering Coimbatore, India Amrita Vishwa Vidyapeetham, India s_anju@cb.amrita.edu

[3] Emotion Recognition from Contextualized Speech Representations using Fine-tuned Transformers George Cioroiu∗ and Anamaria Radoi∗ ∗Faculty of Electronics, Telecommunications and Information Technology National University of Science and Technology Politehnica Bucharest Bucharest, Romania Email: george.cioroiu@upb.ro

[4] R., Steven. Livingstone, "RAVDESS Emotional SpeechAudio", *Kaggle.Com*, 2020, [online] Available: https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio.

[5] G. Deshmukh, A. Gaonkar, G. Golwalkar and S. Kulkarni, "Speech based Emotion Recognition using Machine Learning", *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, vol. 4, no. 4, pp. 812-817, 2019.

[6] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review", *IEEE Access*, vol. 2, no. 7, pp. 117327-117345, 2019.

[7] Y. Ü. Sonmez and A. Varol, "New Trends in Speech Emotion Recognition", *2019 7th International Symposium on Digital Forensics and Security (ISDFS)*, vol. 6, no. 8, pp. 1-7, 2019.

[8] L. Zheng, Q. Li, H. Ban, S. Liu, ―Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest‖, The 30th Chinese Control and Decision Conference (2018 CCDC), pp. 4143-4147, 2018.

