International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:3048-7722
Volume 4, Issue 4, July-Aug-2026, PP: 01-5

A Machine Learning-Based Framework for Automatic Speech Emotion Recognition

S.Venkateswara Rao[1], K. Vaishnavi[2]

[1]Assistant Professor, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana –, India.

[2]MCA Student, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana –, India.

***Abstract* – Human speech conveys not only linguistic information but also valuable emotional cues that reflect an individual's psychological and behavioral state. Accurately identifying these emotions is becoming increasingly important for developing intelligent human–computer interaction systems capable of delivering natural, adaptive, and personalized communication experiences. Conventional speech emotion recognition methods often depend on handcrafted acoustic features and traditional classification techniques, which may struggle to capture the complex temporal and spectral characteristics of emotional speech. Recent advances in machine learning and deep learning have significantly improved the capability of automated systems to analyze speech signals and recognize emotional patterns with greater accuracy, enabling their application in domains such as virtual assistants, customer support, healthcare, education, and mental health monitoring. The effectiveness demonstrates that the CNN model achieves the highest recognition performance among the evaluated approaches, outperforming both SVM and LSTM in emotion classification accuracy while effectively learning complex acoustic representations from speech data. The integration of advanced preprocessing, discriminative feature extraction, and deep learning-based classification contributes to reliable and efficient emotion recognition across multiple emotional categories. The proposed framework provides a scalable solution for intelligent speech analysis and offers significant potential for enhancing emotion-aware applications in modern human–computer interaction systems by enabling machines to interpret and respond to human emotions more naturally and effectively.**

***Keywords* - Machine Learning, Speech Emotion Recognition, Automatic Emotion Recognition, Speech Signal Processing, Feature Extraction, Artificial Intelligence, Audio Classification**

I. INTRODUCTION

The ability to recognize human emotions from speech has become an essential component of intelligent human–computer interaction systems. Unlike textual communication, speech conveys both linguistic information and emotional cues through variations in pitch, tone, rhythm, energy, and speaking style. These vocal characteristics enable computers to infer a speaker's emotional state, allowing digital systems to communicate in a more natural, personalized, and context-aware manner. As artificial intelligence continues to evolve, Speech Emotion Recognition (SER) has emerged as a significant research area with applications spanning healthcare, customer relationship management, virtual assistants, online education, entertainment, call-center analytics, and mental health assessment. Developing reliable emotion recognition systems can substantially improve user experience by enabling machines to respond appropriately to human emotions rather than relying solely on spoken content.

Traditional speech emotion recognition methods. Although these approaches achieved acceptable performance under controlled conditions, they often struggled to model the complex temporal dynamics and spectral variations present in emotional speech. Human emotions are influenced by multiple factors, including speaking style, language, accent, recording conditions, and individual vocal characteristics, making accurate emotion recognition a challenging task. Consequently, modern SER research increasingly focuses on deep learning techniques capable of automatically learning discriminative representations directly from speech data while reducing the dependence on manually engineered features.

In parallel, Support Vector Machines (SVMs) continue to provide competitive classification performance, particularly when combined with informative acoustic features extracted from speech recordings. The integration of advanced feature extraction techniques with robust classification algorithms has enabled substantial improvements in emotion recognition accuracy across diverse speech datasets.

This study proposes an intelligent machine learning framework for Speech Emotion Recognition that integrates comprehensive audio preprocessing, acoustic feature extraction, and multiple classification models to identify emotions from speech signals.containing speech samples representing multiple emotional categories and diverse recording conditions.

The developedto compare the effectiveness of different machine learning models. Experimental analysis demonstrates that the CNN model provides superior recognition performance while effectively learning complex emotional representations from speech signals.

International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:3048-7722
Volume 4, Issue 4, July-Aug-2026, PP: 01-5

II. LITERATURE SURVEY

Although these methods demonstrated reasonable performance under controlled experimental conditions, their ability to generalize across different speakers, languages, and recording environments was often limited. As speech signals contain complex temporal and spectral variations, conventional approaches frequently struggled to capture the subtle emotional characteristics required for accurate classification.

To improve recognition performance, researchers introduced advanced feature extraction techniques capable of representing emotional information more effectively. Rate became widely adopted because they capture complementary acoustic characteristics associated with different emotional expressions. Several investigations demonstrated that combining multiple acoustic descriptors significantly enhances emotion recognition accuracy by providing richer representations of speech signals than individual features alone.

Recent studies have also investigated the effectiveness of Support Vector Machines (SVMs) for speech emotion classification. Despite the rapid advancement of deep learning models, SVMs continue to provide reliable classification performance when trained using carefully engineered acoustic features. Their strong generalization capability and relatively low computational requirements make them suitable for applications involving moderate-sized datasets and real-time emotion recognition systems. Comparative analyses consistently demonstrate that while SVMs remain competitive, deep learning architectures generally achieve superior recognition accuracy by automatically learning complex feature representations.

Although significant progress has been achieved in Speech Emotion Recognition, several challenges remain unresolved. Variations in speaker characteristics, linguistic diversity, background noise, recording quality, and emotional intensity continue to influence recognition performance. Furthermore, many existing systems rely on limited datasets that may not accurately represent real-world communication scenarios. The proposed study addresses these challenges by integrating comprehensive audio preprocessing, multiple acoustic feature extraction techniques, and comparative evaluation of SVM, CNN, and LSTM models using diverse publicly available datasets. This integrated framework aims to improve recognition accuracy, model robustness, and practical applicability across a broad range of emotion-aware intelligent systems.

III. PROPOSED METHODOLOGY

The proposed framework presents an intelligent Speech Emotion Recognition (SER) system that automatically identifies human emotions from speech signals using advanced machine learning and deep learning techniques. Unlike conventional emotion recognition methods that rely primarily on manually designed acoustic descriptors and single classification models, the proposed framework combines comprehensive audio preprocessing, feature optimization, emotion classification, and emotion prediction.

The process begins with speech data acquisition, where audio recordings are collected from publicly available speech emotion datasets. The proposed framework utilizes datasets obtained from Kaggle and GitHub, containing speech samples representing multiple emotional categories recorded under different environmental conditions. The diversity of these datasets enables the model to learn various emotional expressions while improving its ability to generalize across different speakers, recording qualities, and speaking styles.

Following data acquisition, the speech recordings undergo audio preprocessing to improve signal quality before feature extraction. During preprocessing, background noise is minimized through noise reduction techniques, while normalization standardizes variations in recording amplitude and vocal intensity. The audio signals are subsequently segmented into short overlapping frames, allowing detailed analysis of local speech characteristics. These preprocessing operations reduce irrelevant variations and ensure that the extracted features accurately represent emotional information rather than recording artifacts.

The processed audio frames are then supplied to the feature extraction module, where multiple acoustic descriptors are computed to represent the emotional characteristics of speech. The framework extracts Mel-Frequency Cepstral Coefficients (MFCCs) to capture perceptually meaningful spectral information, Chroma Features to describe harmonic content, Spectral Contrast to measure differences between spectral peaks and valleys, and Zero-Crossing Rate to characterize signal frequency variations. These complementary features are combined into comprehensive feature vectors that provide rich representations of speech signals associated with different emotional states.

To improve computational efficiency and reduce redundant information, the extracted features undergo f Comparative evaluation enables the framework to identify the model that provides the highest recognition performance. The CNN model effectively learns hierarchical spatial representations from speech features, whereas the LSTM captures temporal dependencies within speech sequences, and the SVM performs robust classification using optimized acoustic descriptors.

Finally, the trained classification model is integrated into a Flask-based web application developed using the TensorFlow framework. Users can upload speech recordings through the web interface, where the system

International Journal for Novel Research in Economics , Finance and Management
www.ijnrefm.com
ISSN:3048-7722
Volume 4, Issue 4, July-Aug-2026, PP: 01-5

automatically performs preprocessing, feature extraction, and emotion prediction. The predicted emotional state is then displayed to the user, enabling applications inwhile maintaining computational efficiency, providing a practical and scalable solution for automatic speech emotion recognition.

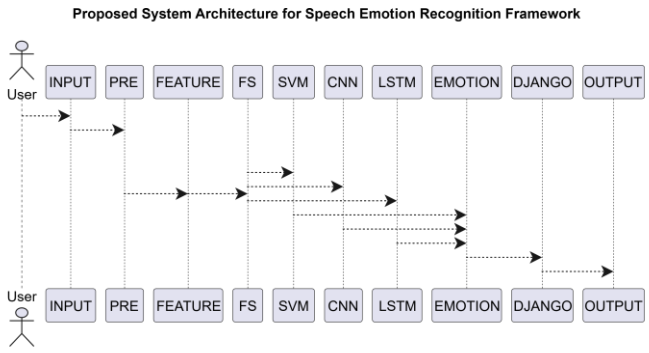

Fig. 1. Proposed system architecture of the Speech Emotion Recognition framework

IV. METHODOLOGY

This structured workflow enables the system to learn meaningful emotional characteristics while improving recognition accuracy and computational efficiency.

The first stage involves dataset preparation, where speech recordings representing multiple emotional categories are collected from publicly available datasets. The experimental framework utilizes speech emotion datasets obtained from Kaggle together with an additional GitHub Speech Emotion Detection dataset containing approximately 1500 audio samples. These datasets include recordings from speakers with diverse demographic backgrounds, recording environments, and emotional expressions, enabling the model to learn generalized emotional characteristics suitable for real-world applications.

During the audio preprocessing stage, the speech recordings are enhanced before feature extraction. Noise reduction techniques remove unwanted environmental disturbances, while amplitude normalization standardizes variations in recording volume. Each speech recording is segmented into small overlapping frames that preserve local temporal information required for detailed acoustic analysis. These preprocessing operations improve signal quality and reduce the influence of recording inconsistencies on classification performance.

The processed speech frames are subsequently analyzed through feature extraction, where multiple acoustic descriptors are computed to represent emotional characteristics. The framework extracts from each speech segment. These features capture complementary aspects of speech, including frequency distribution, harmonic structure, spectral variation, and temporal signal behavior. The resulting feature vectors provide comprehensive representations that enable accurate discrimination among different emotional states.

Following feature extraction, feature optimization is performed to reduce dimensionality and eliminate redundant information. Feature selection techniques identify the most discriminative acoustic attributes, improving computational efficiency while preserving classification performance. The optimized feature vectors are then supplied to three classification models: Each model is trained independently to learn emotion-specific patterns present within the extracted speech features, allowing comparative evaluation of their recognition capabilities.

Finally, the trained models are the highest emotion recognition performance among the evaluated approaches. The optimized model is integrated into a Flask-based web application supported by the TensorFlow deep learning framework, enabling users to upload speech recordings and receive real-time emotion predictions through an interactive interface. This methodology establishes an efficient, scalable, and reliable framework for intelligent Speech Emotion Recognition suitable for numerous emotion-aware applications in healthcare, customer support, education, and human–computer interaction.

V. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed Speech Emotion Recognition (SER) framework was experimentally evaluated to investigate using a diverse collection of speech recordings obtained from publicly available datasets. The primary objective of the evaluation was to identify the most suitable classification model capable of providing reliable emotion recognition while maintaining computational efficiency for practical deployment.

The experimental dataset consisted of speech recordings collected from Kaggle together with an additional GitHub Speech Emotion Detection dataset containing approximately 1500 audio samples. The combined dataset represented multiple emotional categories recorded under different environmental conditions and speaker characteristics. Prior to model training, all speech recordings underwent preprocessing procedures including segmentation, and feature extraction to ensure consistent input quality. Multiple acoustic descriptors, namely were extracted from each audio sample to generate informative feature vectors suitable for emotion classification.

Three classification models—Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM)—were independently trained using the extracted acoustic features. The SVM classifier demonstrated reliable performance by effectively separating different emotional categories using optimized

decision boundaries. Although SVM provided competitive classification results, its ability to capture highly complex emotional variations remained comparatively limited when handling large and diverse speech datasets. The LSTM model successfully learned temporal relationships among speech sequences, enabling accurate recognition of emotions that evolved over time. However, its sequential processing resulted in relatively higher computational requirements during training compared with CNN-based approaches.

Among the evaluated models, the Convolutional Neural Network (CNN) achieved the highest overall recognition performance. Experimental analysis demonstrated that CNN effectively learned hierarchical acoustic representations directly from the extracted speech features, enabling more accurate discrimination among different emotional categories. According to the reported experimental results, the CNN model achieved an emotion recognition accuracy of approximately 99.6%, outperforming both the LSTM model (98%) and the SVM model. The confusion matrix further indicated that CNN successfully distinguished closely related emotional classes while minimizing misclassification between similar speech patterns. These findings demonstrate the capability of deep convolutional architectures to capture complex spectral variations associated with human emotional expressions.
Comparative analysis confirmed that integrating comprehensive preprocessing, multiple performance. Furthermore, the implementation of the framework within a Flask-based web application supported by the TensorFlow platform enabled efficient emotion prediction through an interactive user interface, demonstrating the practical applicability of the proposed system in real-world emotion-aware applications.

Overall, the experimental findings verify that the proposed Speech Emotion Recognition framework provides accurate, scalable, and computationally efficient emotion classification. The combination of advanced acoustic feature extraction with CNN-based deep learning enables reliable recognition of diverse emotional states while supporting applications in virtual assistants,

VI. FUTURE ENHANCEMENT

Although the proposed Speech Emotion Recognition framework demonstrates strong performance in recognizing emotions from speech signals, several improvements can further enhance its accuracy, robustness, and practical applicability. Future research may focus on expanding the training dataset by incorporating speech recordings from multiple languages, age groups, cultural backgrounds, and real-world acoustic environments. A larger and more diverse dataset would improve the model's ability to generalize across different speakers while reducing bias toward specific recording conditions.

Another promising direction involves integrating Vision Transformers (ViT) adapted for speech spectrogram analysis. These architectures can capture long-range contextual dependencies and complex acoustic relationships more effectively than conventional neural networks, potentially improving recognition accuracy for subtle emotional expressions.

Future implementations may also incorporate d textual information. Integrating multiple modalities would enable the system to analyze emotional states more comprehensively, particularly in situations where speech alone provides limited emotional cues. Such multimodal frameworks could significantly improve recognition reliability in healthcare, psychological assessment, and intelligent virtual assistant applications.

The proposed framework can further be enhanced through real-time deployment on mobile devices, cloud platforms, and edge computing environments. Optimizing the trained deep learning models for lightweight execution would enable efficient emotion recognition with reduced latency while supporting continuous monitoring applications such as smart customer service systems, online education platforms, call-center analytics, and emotion-aware conversational agents.

Finally, future studies may explore Providing interpretable explanations for emotion predictions would assist researchers, clinicians, and application developers in understanding model decisions while facilitating wider adoption of Speech Emotion Recognition technologies across diverse real-world applications. These enhancements would contribute to the development of highly accurate, adaptive, and intelligent emotion-aware systems capable of providing more natural and empathetic human–computer interactions.

VII. CONCLUSION

This study presented an intelligent advanced audio preprocessing, comprehensive acoustic feature extraction, and multiple machine learning models to accurately recognize human emotions from speech signals. The proposed system utilizes effective approach for emotion classification. By combining diverse feature representations with modern machine learning techniques, the proposed framework provides an efficient and reliable solution for automatic emotion recognition from speech.

The experimental evaluation was conducted using publicly available speech emotion datasets collected from Kaggle and GitHub, representing multiple emotional categories, speakers, and recording environments. Comparative analysis demonstrated that the Convolutional Neural Network (CNN) achieved the highest recognition performance among the evaluated models, outperforming both SVM and LSTM in overall classification accuracy. The results further indicate that comprehensive audio

preprocessing and multi-feature extraction significantly enhance recognition capability by providing richer acoustic representations for machine learning models. The implementation of the framework through a Django-based web application integrated with the TensorFlow platform further demonstrates its practical applicability for real-time emotion prediction.

The proposed states from speech, the system contributes toward more adaptive, personalized, and empathetic interactions between humans and intelligent systems. The lightweight architecture and efficient feature extraction process also support practical deployment across different computing environments.

In conclusion, the integration of advanced acoustic feature engineering with SVM, CNN, and LSTM classification models establishes a robust and scalable framework for intelligent Speech Emotion Recognition. The experimental findings confirm that deep learning-based approaches, particularly CNN, provide superior performance in recognizing complex emotional patterns from speech signals.

REFERENCES

1. R. W. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997.
2. C. M. Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” IEEE Transactions on Speech and Audio Processing, vol. 13, no. 2, pp. 293–303, 2005.
3. M. Anjum, “Emotion recognition from speech for an interactive robot agent,” in Proc. IEEE/SICE Int. Symp. System Integration (SII), 2019, pp. 363–368.
4. S. T. Saste and S. M. Jagdale, “Emotion recognition from speech using MFCC and DWT for security systems,” in Proc. Int. Conf. Electronics, 2017, pp. 701–704.
5. Z. Yang, C. Zhang, Y. Xu, and Y. Liu, “Speech emotion recognition based on deep learning with syllable-level attention,” IEEE Access, vol. 9, pp. 7867–7879, 2021.
6. G. Hinton et al., “Deep neural networks for acoustic modeling in speech recognition,” IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012.
7. A. Mohamed, G. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 14–22, 2012.
8. F. Eyben, M. Wöllmer, and B. Schuller, “OpenSMILE: The Munich versatile and fast open-source audio feature extractor,” in Proc. ACM Int. Conf. Multimedia, 2010, pp. 1459–1462.
9. B. Schuller et al., “The INTERSPEECH 2010 paralinguistic challenge,” in Proc. INTERSPEECH, 2010, pp. 2794–2797.
10. C. Busso et al., “IEMOCAP: Interactive emotional dyadic motion capture database,” Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.
11. J. Gideon et al., “Progressive neural networks for transfer learning in emotion recognition,” in Proc. INTERSPEECH, 2017, pp. 1098–1102.
12. S. Yoon, S. Byun, and K. Jung, “Multimodal speech emotion recognition using audio and text,” in Proc. Int. Conf. Signal Processing and Communication Systems (ICSPCS), 2018.
13. H. Kaya and A. A. Karpov, “A novel multimodal approach for speech emotion recognition,” in Proc. INTERSPEECH, 2015.
14. M. Abadi et al., “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2016.
15. F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
16. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
17. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.
18. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
19. A. Graves, A.-R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6645–6649.
20. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learning Representations (ICLR), 2015.
21. K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” in Proc. INTERSPEECH, 2014.
22. D. P. Ramesh Babu and P. Krishna Rao, “Artificial intelligence-driven speech emotion recognition using deep learning for intelligent human–computer interaction,” International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 5, pp. 3652–3665, 2024.