International Journal of Advance and Innovative Research
Volume 13, Issue 1 (XIII): January – March 2026
ISSN 2394 - 7780

A HYBRID LSTM AND GRU BASED EMOTION DETECTION MODEL IN SPEECH ACOUSTIC SIGNALS

Manan Jain[1], Sumitkumar Tripathi[2] and Kanojia Mahendra[3]
[1]Information Technology, Sheth L.U.Jhaveri and Sir M.V. College, India, bscit.manan@gmail.com
[2]Information Technology, Sheth L.U.Jhaveri and Sir M.V. College, India, sumit11.tripathi@gmail.com
[3]Department of Computer Science, Sheth. L.U.J. and Sir M.V. College, India, kgkmahendra@gmail.com
*Corresponding author: Manan Jain, bscit.manan@gmail.com

ABSTRACT
Speech Emotion Recognition represents a key domain within affective computing, allowing intelligent systems to identify emotions of humans through vocal signals and understand their emotions and try to identify the emotional situation of an individual. Despite considerable improvement and advanced techniques in speech analysis and deep learning methods, obtaining consistent emotion detection in practical, real-world conditions continues to face difficulties, because there are differences among speakers and their vocal signals, background noise interference behind the speaker, and the complex time-dependent patterns in speech. The current study proposes a hybrid recurrent model that merges Long Short-Term Memory and Gated Recurrent Unit layers, aiming to improve the modeling of sequential emotional patterns while keeping the overall model straightforward and efficient and understanding the emotion of a human. Mel-Frequency Cepstral Coefficients serve as the main acoustic feature set with performance assessment conducted on the Ryerson Audio-Visual Database of Emotional Speech and Song, covering eight distinct emotion classes. To promote better generalization and reliable multiclass outcomes, dropout regularization along with a softmax activation in the output layer was been included. The experimental results demonstrated an overall accuracy of 74.1%, which reflects excellent prediction performance compared to the individual LSTM and GRU implementations. These findings suggest that the combination of such recurrent components effectively captures immediate emotional indicators and understands their emotions as well as extended contextual information, thereby rendering the suggested approach well-suited for real-world speech emotion recognition tasks and understand the emotions accurately of a speaker.

Keywords: *Affective Computing, Deep Learning, LSTM-GRU Networks, Mel-Frequency Cepstral Coefficients, RAVDESS Dataset, Recurrent Neural Networks, Speech Emotion Recognition.*

1. INTRODUCTION
Verbal communication acts as a crucial medium for human interaction., transmitting not only verbal content but also emotional information that reflects a speaker's inner condition. Advanced technologies such as virtual assistants, social robots, and human-computers that can interact with humans are evolving continuously, In the field of affective computing Speech Emotion Recognition (SER) has gained significant attention (El Ayadi et al., 2011; Schuller, 2018). With considerable improvement and advanced technique, achieving reliable emotion detection from speech remains difficult because of differences among speakers, differences in speaking styles, environmental noise, and the complex nature of emotional expression in natural conversations (El Ayadi et al., 2011; Schuller et al., 2011). In earlier times Speech Emotion Recognition systems was designed with manual acoustic feature combined with traditional classifier, but these approaches often lacked the capability to train nonlinear connections and long-term temporal dependencies present in emotional speech signals (El Ayadi et al., 2011; Fayek et al., 2017). The advancement of deep learning has empowered more effective representation learning for speech-based tasks (LeCun et al., 2015). We know that Recurrent Neural Networks (RNNs) are ideal for handling sequential data processing but standard RNNs struggle with vanishing gradient problems that limit their ability to capture retain long-term contextual information (Hochreiter & Schmidhuber, 1997).

To tackle these problem, gated recurrent architectures, Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) and Gated Recurrent Units (GRUs) (Cho et al., 2014) have been widely adopted in speech emotion recognition (SER). LSTMs are well suited for modeling long-range temporal dependencies, while GRUs provide a more compact structure with fewer parameters and typically faster training (Fayek et al., 2017). Hybrid designs that join both convolutional and recurrent layers can improve feature learning (Mao et al., 2014; Trigeorgis et al., 2016; Zhao et al., 2019), but adding convolutional components often increases computational cost and demands larger datasets for stable optimization. Focusing instead on an LSTM-GRU combination emphasizes temporal modeling, preserves architectural simplicity, and better balances performance with efficiency. Motivated by this, the present work proposes a hybrid LSTM–GRU framework for multiclass SER. Mel-Frequency Cepstral Coefficients (MFCCs) are adopted as the acoustic representation

ISSN 2394 - 7780

(Ancilin & Milton, 2021) and extracted using the Librosa toolkit (McFee et al., 2015). The model is evaluated on the RAVDESS dataset (Livingstone & Russo, 2018), with training stability aided by dropout regularization (Srivastava et al., 2014) and optimization via the Adam algorithm (Kingma & Ba, 2015). Results show the hybrid architecture improves recognition compared with individual recurrent models while keeping computational demands moderate, underscoring its suitability for practical SER systems (Fayek et al., 2017; Zhao et al., 2019).

2. LITERATURE REVIEW

In the field of affective computing Speech Emotion Recognition has become a central area, mainly due to its importance in developing emotionally intelligent computers that can interact with humans. Through speech we can reveal a speaker's emotions, intentions, and mental state through communication, transmitting not just words but also nonverbal communication signals. Adding emotional intelligence to smart systems can significantly enhance user interaction, system flexibility, and overall experience in areas such as voice assistants, companion robots, psychological health tools, and automated customer support (Schuller, 2018; El Ayadi et al., 2011). Even with the advanced techniques in speech analysis, detecting the emotions of humans in different languages accurately is difficult. This is because of the difference in how emotions are expressed individually, and background noises, and emotional boundaries (Schuller et al., 2011). At earlier times SER was mostly dependent on machine learning techniques which integrated with simple classifiers like hidden Markov model and support vector machines. While these techniques created important early benchmarks, they were restricted by their limited ability to handle the complex, nonlinear, and constantly changing nature of emotional speech (El Ayadi et al., 2011). Moreover, it showed weak performance when tested on real world data, which made the researchers to acquire more advance modeling approaches. Later reviews and experiments clearly showed that traditional classifiers have difficulty capturing extended temporal relationships, an essential element for dependable emotion detection (Fayek et al., 2017).

With the help of advanced technique, the process in deep learning is fast and has brought major changes to SER research. Deep neural network allows automatic extraction of layered features and helps the models to recognize meaningful patterns from raw data (LeCun et al., 2015). Recurrent Neural Networks (RNNs) were first used to extract local spectral information from time-frequency representations like spectrograms and Mel filterbanks (Mao et al., 2014; Badshah et al., 2017). These methods have performed better than the classic crafted feature but still facing difficulties with capturing dependencies over a longer period of time. Because of this limitation, the field gradually shifted toward RNNs, which are naturally designed for handling sequential information (Lim et al., 2016). In the field of acoustic feature types, Mel-Frequency Cepstral Coefficients (MFCCs) is still the most popular choice in SER. Their strength lies in how well they represent the speech spectrum in a way that matches human hearing perception and their robustness (Ancilin & Milton, 2021; El Ayadi et al., 2011). Several research findings indicate that Mel-Frequency Cepstral Coefficients (MFCCs) generally achieve stronger performance in emotion recognition tasks than unprocessed spectral representations (Ancilin & Milton, 2021).. However, MFCCs alone are limited in their ability to represent the gradual evolution of emotional states over time. To achieve improved performance, they are often integrated with sequence-based models capable of capturing temporal dependencies within speech signals (Fayek et al., 2017). MFCCs cannot fully capture changes in emotion gradually, so for better result it should be combine with other model that can handle sequence (Fayek et al., 2017).

Early attempts to capture temporal dynamics in speech processing relied on conventional Recurrent Neural Networks (RNNs), but their effectiveness was limited by the vanishing gradient issue when dealing with long input sequences (Hochreiter & Schmidhuber, 1997). To overcome this challenge, Long Short-Term Memory (LSTM) networks were introduced, incorporating memory cells along with gating mechanisms that regulate information flow and enable stable learning of long-range dependencies (Hochreiter & Schmidhuber, 1997). Later, Gated Recurrent Units (GRUs) emerged as a more compact alternative, reducing architectural complexity by using fewer gates, which resulted in faster training and lower computational demands while preserving comparable performance (Cho et al., 2014; Fayek et al., 2017). More recently, advancements in Speech Emotion Recognition (SER) have emphasized the integration of attention mechanisms within recurrent frameworks, allowing models to concentrate on emotionally significant segments of speech and thereby enhance classification accuracy (Li et al., 2021; Mirsamadi et al., 2017). In particular, directional self-attention has shown effectiveness in diminishing the influence of silent or less informative portions of the signal, improving overall robustness in emotion detection (Li et al., 2021). However, despite their performance benefits, attention-based architectures typically involve increased computational complexity, which may restrict their suitability for real-time systems or devices with limited processing resources (Schuller, 2018).

ISSN 2394 - 7780

In recent years, hybrid architectures and ensemble learning strategies have gained significant traction in Speech Emotion Recognition research. Approaches that integrate RNNs with LSTM and GRU units leverage the complementary advantages of these components to effectively capture both fine-grained spectral characteristics and long-term temporal dependencies in speech signals (Zhao et al., 2019; Lim et al., 2016). Additionally, end-to-end deep learning frameworks that operate directly on spectrogram representations or raw audio waveforms have minimized the reliance on handcrafted acoustic features and achieved competitive performance on benchmark datasets (Tzirakis et al., 2018; Trigeorgis et al., 2016). Despite their promising results, such end-to-end models typically require extensive labeled data and high computational capacity, which can restrict their deployment in resource-limited or real-world settings (Latif et al., 2018). Widely recognized benchmark datasets such as RAVDESS (Livingstone & Russo, 2018) and IEMOCAP (Busso et al., 2008) have significantly contributed to the progress of Speech Emotion Recognition research by enabling standardized evaluation and fair, objective comparisons among various models. Findings from studies using these datasets consistently highlight the trade-off between architectural complexity and practical efficiency: while attention-driven and ensemble approaches often achieve superior accuracy, they tend to increase computational cost and inference time, whereas simpler models may struggle to capture rich contextual and emotional nuances (Schuller et al., 2011; Fayek et al., 2017).

Advancements in Speech Emotion Recognition (SER) have significantly benefited from deep learning architectures tailored for sequential signal processing. Early implementations that combined Mel-Frequency Cepstral Coefficients (MFCCs) with basic RNN models achieved accuracy levels of approximately 68% on benchmark datasets such as RAVDESS. The incorporation of attention mechanisms into recurrent frameworks enhanced the modeling of temporal emotional patterns, leading to performance improvements of around 71%. Further gains were observed with hybrid architectures integrating RNN, LSTM, and GRU units, which more effectively captured both spectral characteristics and temporal dynamics, pushing accuracy to nearly 72–73%. Similarly, attention-based peephole LSTM variants demonstrated improved contextual learning capabilities, though at the expense of increased computational complexity. Overall, contemporary studies highlight the persistent trade-off between maximizing recognition performance and maintaining computational efficiency, particularly in real-time deployment scenarios. The LSTM-GRU framework introduced in this study attains an accuracy of 74.1% on the RAVDESS dataset while preserving a comparatively streamlined architecture. By combining LSTM's strength in modeling long-range dependencies with the computational efficiency of GRU gating mechanisms, the model achieves enhanced temporal representation with manageable resource requirements, indicating that blended recurrent structures can serve as a practical and scalable solution for emotion-sensitive speech applications.

3. DATA CORPUS

This research uses RAVDESS dataset, a dataset that is used globally for Speech Emotion Recognition studies (Livingstone & Russo, 2018). The dataset contains 1440 labels recorded by 24 professionally trained actors, with balanced representation of male and female speakers. Every audio sample is classified into eight emotional categories such as neutral, calm, happy, sad, angry, fearful, disgust, and surprised. All the recorded audio are present in WAV format, featuring a sampling rate of 48 kHz, and can be openly used for academic purpose by anyone. Feature extraction was performed before converting the audio signals into mono, amplitude-normalized, and subjected to silence removal to maintain uniformity across samples.

For extracting acoustic feature MFCC's is used, because it is very effective in capturing speech, distinction significant spectral features (Ancilin & Milton, 2021). From Librosa library MFCC extraction was carried out (McFee et al., 2015). MFCC extraction was carried out using the Librosa library (McFee et al., 2015), creating 20-dimensional MFCC vectors that are arranged as temporal sequences appropriate for modeling recurrent neural networks and are calculated at the frame level. During the process, damaged recording, low-energy utterances, and to ensure the accuracy and reliability of the data, samples with irregular temporal length were removed. A total of 864 labeled speech were saved for assessment through testing. To maintain the original class distribution and guarantee repeatable outcomes, the revised dataset was then split 80:20 across sets for training and testing using sampling stratification.

ISSN 2394 - 7780

Table 1. Dataset Summary

Parameter	Value
Dataset	RAVDESS
Total Samples	1,440
Speakers	24 (12 Male, 12 Female)
Emotion Classes	8
Audio Format	WAV, 48 kHz
Features Extracted	20 MFCC
Final Samples Used	864
Data Split	80% Training, 20% Testing

Table 1 shows a summary of the dataset used for speech emotion recognition experiments. This study uses RAVDESS dataset, which contains 1,440 speech recordings of 24 professional actors with balanced gender representation and eight emotion categories, recorded in WAV format at a sampling rate of 48 kHz (Livingstone & Russo, 2018). Librosa library was used for extracting Mel-Frequency Cepstral Coefficients (MFCCs) in order to obtain acoustic representations of speech signals that are relevant to perception (Ancilin & Milton, 2021; McFee et al., 2015). After preprocessing and feature selection, 864 labeled samples were retained for experimental analysis. The dataset was divided into training and testing subsets using stratified sampling to maintain class balance and ensure reliable performance evaluation, which is consistent with best practices in speech emotion recognition research (El Ayadi et al., 2011; Schuller, 2018).

4. RESEARCH METHODOLOGY

A structured and systematic methodology is implemented to develop a multiclass Speech Emotion Recognition (SER) system based on a Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) deep learning architecture. The overall framework includes dataset selection, signal preprocessing, acoustic feature extraction, feature normalization, sequential modeling, model training, performance evaluation, and comparative analysis. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is utilized for experimental validation (Livingstone & Russo, 2018). The dataset comprises 1,440 speech recordings produced by 24 professional actors, including 12 male actors and 12 female actors, ensuring balanced gender representation. Each recording is labeled under one of eight emotion categories: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. All audio samples are provided in WAV format with a sampling rate of 48 kHz, offering high quality speech signals suitable for detailed acoustic analysis. Before feature extraction, several preprocessing steps are applied to standardize the input signals. The recordings are converted to mono format when required to ensure consistent channel representation. Amplitude normalization is performed to remove loudness variation between samples. Silence removal is applied to eliminate non-informative segments at the beginning and end of recordings. Basic noise reduction techniques are also applied when necessary to improve signal clarity. These preprocessing operations are implemented using the Librosa Python library (McFee et al., 2015), ensuring reproducibility and consistency across experiments.

Following preprocessing, acoustic feature extraction is performed using Mel-Frequency Cepstral Coefficients (MFCCs). MFCCs are widely used in SER research because they effectively capture perceptually relevant spectral characteristics of speech signals (Ancilin & Milton, 2021; El Ayadi et al., 2011). Each speech signal is segmented into short overlapping frames of approximately 20-40 milliseconds to preserve short-term temporal information. For each frame, MFCC coefficients are computed, resulting in a time-sequential feature representation defined in Equation (1)

$$X \in R^{T \times D} \quad (1)$$

where T represents the total number of temporal frames and D denotes the dimensionality of the MFCC feature vector. This sequential structure preserves the dynamic variation of speech over time and serves as suitable input for recurrent neural networks. To enhance pitch-related and harmonic information, chroma features are optionally incorporated. Chroma features encode the distribution of spectral energy across twelve pitch classes and provide complementary prosodic information relevant to emotional expression (El Ayadi et al., 2011; Schuller, 2018). When chroma features are concatenated with MFCCs, the updated feature dimensionality is defined as

$$D = D_{MFCC} + 12 \quad (2)$$

as shown in Equation (2). All feature sequences are padded or truncated to a fixed temporal length T to maintain consistent input dimensions. Feature scaling and normalization are applied to stabilize the training process.

ISSN 2394 - 7780

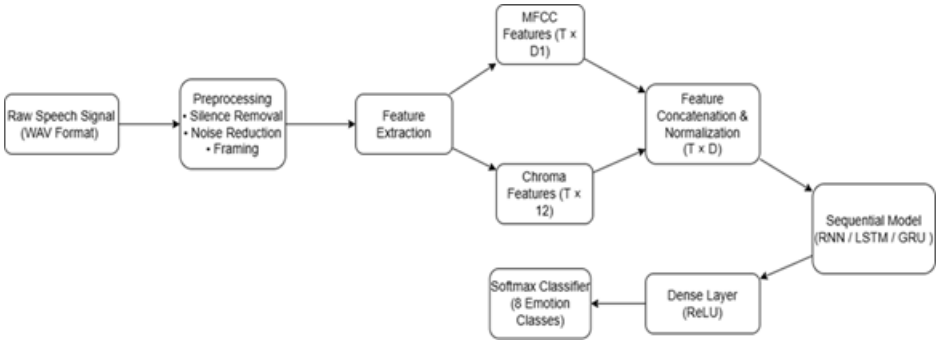

Figure 1: Workflow of the Proposed SER System

Figure 1 presents the overall workflow of the proposed SER framework. The diagram visually summarizes the transformation from raw speech input to emotion classification. It shows preprocessing operations, MFCC and chroma feature extraction, feature concatenation, normalization, and sequential modeling using recurrent architectures. The workflow representation aligns with the mathematical formulation defined in Equation (1) and provides a conceptual overview of the entire feature-to-classification pipeline, as shown in Figure 1. For sequential modeling, a recurrent architecture combining LSTM and GRU layers is designed. LSTM networks are used to capture long-term contextual dependencies in speech signals through gated memory mechanisms (Hochreiter & Schmidhuber, 1997). GRU layers are included to refine learned temporal representations while reducing model complexity (Cho et al., 2014). This combination allows the model to benefit from both extended memory capability and computational efficiency.

For a batch size B, the three-dimensional input tensor is represented as:

$$X \in R^{B \times T \times D} \quad (3)$$

as defined in Equation (3). Stacked LSTM layers first process the input sequences to model long-range emotional dependencies. The output of the LSTM block is then forwarded to a GRU layer, which further refines temporal features and improves training stability. Dropout regularization with a rate of 0.3 is applied after recurrent layers to reduce overfitting and improve generalization (Srivastava et al., 2014). The extracted high-level representation is then passed through a fully connected dense layer with Rectified Linear Unit (ReLU) activation, followed by a Softmax output layer that produces probability distributions across the eight emotion classes. The computational complexity of each recurrent layer is approximated as:

$$O(T \cdot (H^2 + H \cdot D)) \quad (4)$$

where H denotes the number of hidden units (Hochreiter & Schmidhuber, 1997), as defined in Equation (4). Further Figure 2 provides a detailed representation of the LSTM-GRU model architecture. The diagram shows the sequential flow of data from the input tensor defined in Equation (3) through stacked LSTM layers, a GRU refinement layer, dropout regularization, dense transformation, and final softmax classification. The architecture visually clarifies how temporal dependencies are modeled hierarchically before emotion prediction as shown in Figure 2. This hierarchical architecture enables progressive abstraction of temporal features, allowing the network to capture both short-term emotional variations and long-term contextual dependencies within speech signals. By combining complementary gating mechanisms across layers, the framework enhances representational richness while maintaining computational efficiency. This layered structure also improves gradient flow across time steps, supporting stable learning and effective convergence during training. Model training is performed using Backpropagation Through Time (BPTT). Categorical cross-entropy is used as the loss function for multiclass classification. The Adam optimizer with an initial learning rate of $1 \times 10^{-3}$is used for parameter optimization (Kingma & Ba, 2015). Early stopping based on validation loss is applied to prevent overtraining and improve model generalization (Fayek et al., 2017).

ISSN 2394 - 7780

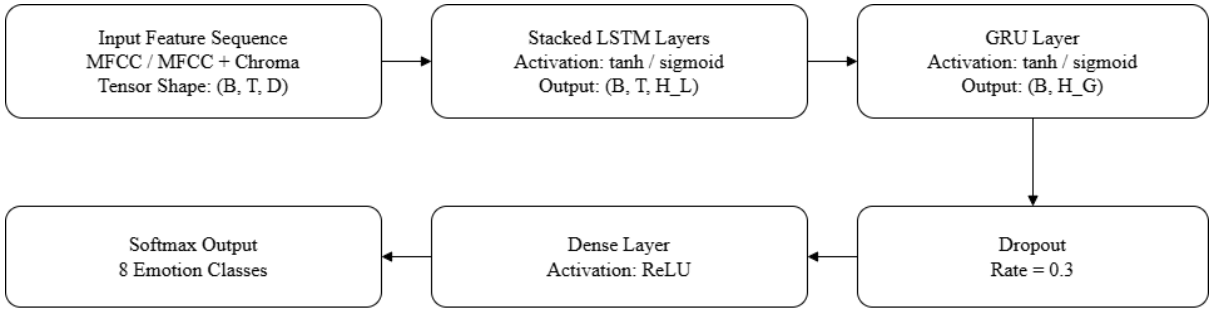

Figure 2: Proposed Model for Speech Emotion Recognition

Performance evaluation is conducted using accuracy, precision, recall, and F1-score. Confusion matrix analysis is performed to examine class-wise prediction behavior. Statistical significance testing is conducted using the Shapiro-Wilk test, followed by paired t-tests or Wilcoxon signed-rank tests when appropriate, at a significance level of $\alpha = 0.05$. Finally, the LSTM-GRU model is compared with individual LSTM and GRU architectures under identical experimental conditions to assess improvements in temporal emotion modeling performance.

5. RESULT AND DISCUSSION

This section presents the experimental evaluation of the proposed LSTM-GRU model for Speech Emotion Recognition (SER). The effectiveness of the model is evaluated using standard performance metrics, namely accuracy, precision, recall, and F1-score. Accuracy represents the overall proportion of correctly classified samples. Precision reflects the reliability of predicted emotion labels, whereas recall indicates the model's ability to correctly identify relevant emotional instances. The F1-score provides a balanced measure of precision and recall and is particularly suitable for multiclass classification problems.

To validate the effectiveness of the proposed architecture, its performance is compared with individual LSTM and GRU models trained under identical experimental settings. As presented in Table 2, the proposed LSTM-GRU model achieves an overall accuracy of 74.1 percent. In comparison, the LSTM model achieves an accuracy of 71.6 percent and the GRU model achieves 72.9 percent. Furthermore, the proposed architecture records a precision value of 0.75, a recall value of 0.74, and an F1-score of 0.74, indicating consistent improvement across all evaluation metrics.

Table 2. Performance Comparison of Speech Emotion Recognition Models

Model	Accuracy (%)	Precision	Recall	F1-Score
LSTM	71.6	0.72	0.71	0.71
GRU	72.9	0.73	0.73	0.72
Proposed Model	74.1	0.75	0.74	0.74

The comparative results demonstrate that integrating LSTM and GRU layers enhances temporal feature learning by combining long-term contextual modeling with efficient gating mechanisms. Class-wise analysis shows strong recognition performance for emotions such as angry, happy, and surprise. Limited confusion is observed between acoustically similar emotion pairs such as calm and neutral, and sad and fear, primarily due to overlapping spectral and prosodic characteristics. Overall, the results confirm that the proposed hybrid architecture provides stable and consistent improvements across multiple evaluation metrics.

Training and validation performance curves further demonstrate stable convergence behavior. The model shows rapid improvement during early epochs, followed by smooth stabilization without divergence between training and validation accuracy. The use of dropout regularization and early stopping effectively controls overfitting and improves generalization capability. Overall, the experimental findings indicate that the LSTM-GRU framework provides improved classification robustness, balanced emotion recognition performance, and computational efficiency compared to individual recurrent models.

6. CONCLUSION

This study presents a LSTM-GRU deep learning framework for multiclass speech emotion recognition that integrates the long-term contextual modeling capability of Long Short-Term Memory networks with the computational efficiency of Gated Recurrent Units (Hochreiter & Schmidhuber, 1997; Cho et al., 2014). Mel-Frequency Cepstral Coefficients were utilized as the primary acoustic representation, and the model was evaluated using the RAVDESS emotional speech dataset (Livingstone & Russo, 2018). Experimental findings indicate that the proposed architecture achieved an overall classification accuracy of approximately 74.1%, performing better than individual models LSTM and GRU baseline models (Fayek et al., 2017). Confusion

ISSN 2394 - 7780

matrix evaluation demonstrated stable class-wise recognition, while training dynamics confirmed effective convergence with minimal overfitting due to dropout regularization and early stopping (Srivastava et al., 2014). The results highlight that combining complementary recurrent structures improves temporal emotion modeling without introducing excessive computational complexity. Consequently, the proposed framework provides an efficient solution for real-time emotion-aware speech processing and can be applied in intelligent virtual assistants, automated customer support systems, mental health monitoring tools, and human-robot interaction platforms (Schuller, 2018).

Table 3. Comparison with Recent Speech Emotion Recognition Studies

Reference	Dataset	Method	Results
Mao et al., 2014	IEMOCAP	CNN-based SER	64.0% Accuracy
Lim et al., 2016	IEMOCAP	CNN + RNN	68.8% Accuracy
Trigeorgis et al., 2016	IEMOCAP	CNN + LSTM (End-to-End)	69.2% Accuracy
Li et al., 2021	RAVDESS	Attention-based RNN	71.2% Accuracy
Zhao et al., 2019	RAVDESS	Deep 1D & 2D RNN-LSTM	72.0% Accuracy
Proposed LSTM-GRU	RAVDESS	Hybrid LSTM-GRU	74.1% Accuracy

As shown in Table 3, the proposed LSTM-GRU framework demonstrates competitive and consistent performance when compared with recent deep learning-based speech emotion recognition approaches, including convolutional neural network (CNN), attention-based, and hybrid recurrent architectures. While CNN-based models are effective in extracting spatial or spectrogram-level representations, they generally introduce additional architectural complexity and a higher number of trainable parameters. Attention-based mechanisms further enhance feature weighting but often increase computational overhead and training time. In contrast, the proposed LSTM-GRU architecture emphasizes efficient temporal sequence modeling by integrating the long-term contextual learning capability of LSTM networks with the streamlined gating structure of GRUs. This design enables effective modeling of both short-term emotional cues and long-range dependencies in speech signals without relying on convolutional layers or complex attention modules. As a result, the framework achieves improved classification accuracy while maintaining moderate parameter size and lower computational cost. The balanced trade-off between performance and efficiency makes the proposed model suitable for real-time and resource-constrained speech emotion recognition applications.

7. FUTURE WORK AND RECOMMENDATIONS

Future research can further enhance the proposed LSTM-GRU framework by incorporating attention mechanisms that allow the model to dynamically focus on emotionally significant temporal regions within speech signals. Attention-based recurrent models have demonstrated improved contextual representation in recent speech emotion recognition studies, and integrating such mechanisms may reduce confusion between acoustically similar emotion pairs by emphasizing discriminative segments. In addition, exploring bidirectional recurrent architectures may strengthen contextual learning by capturing both past and future temporal dependencies within an utterance. Extending evaluation to multiple benchmark datasets and performing cross-corpus validation would provide a more comprehensive assessment of robustness, speaker-independence, and adaptability across diverse linguistic and acoustic environments.

Another important direction involves integrating multimodal affective information, such as visual facial cues, to enhance recognition reliability through complementary emotional representations. Transfer learning strategies could be investigated to adapt pretrained speech models to emotion recognition tasks with limited labeled data, while semi-supervised learning approaches may improve performance in low-resource settings. Conducting broader statistical validation using repeated experiments and cross-validation would increase confidence in the reported findings. Furthermore, incorporating explainable artificial intelligence techniques could provide better insight into model decision-making processes, thereby improving interpretability and user trust in practical deployments. Future investigations may also analyze model behavior under real-time streaming conditions to evaluate latency and stability in dynamic environments. Additionally, exploring adaptive learning mechanisms could enable the system to personalize emotion recognition based on individual speaker characteristics over time.

Acknowledgement
The authors sincerely thank the Department of Information Technology, Sheth L.U. Jhaveri College of Arts and Sir M.V. College of Science and Commerce for their support, guidance, and resources that contributed to the successful completion of this research work.

Ethical Statement
This study did not involve experiments with human participants or animal subjects. All datasets used in this research were obtained from publicly available sources and processed in accordance with ethical research guidelines.

Conflicts of Interest
The authors declare that there are no conflicts of interest related to this research work.

Data Availability Statement
The dataset used in this study is derived from the publicly available Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), obtained from Kaggle. The experimental analysis utilized a processed feature dataset (features.csv) generated from raw audio recordings using MFCC-based feature extraction with the Librosa library, along with preprocessing steps such as normalization and silence removal. The processed dataset and trained model outputs are available from the corresponding author upon reasonable request.

Kaggle. (n.d.). *RAVDESS Emotional Speech Audio Dataset*. Retrieved from https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

REFERENCES

1. J. Ancilin and A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient," *Appl. Acoust.*, vol. 179, Aug. 2021, Art no. 108046, doi: 10.1016/j.apacoust.2021.108046. https://www.sciencedirect.com/science/article/pii/S0003682X21001390

A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. 2017 Int. Conf. Platform Technol. Service (PlatCon)*, Busan, Korea (South), 2017, pp. 1-5, doi: 10.1109/PlatCon.2017.7883728. https://ieeexplore.ieee.org/document/7883728

B. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resources Eval.*, vol. 42, no. 4, pp. 335-359, Dec. 2008, doi: 10.1007/s10579-008-9076-6. https://link.springer.com/article/10.1007/s10579-008-9076-6

2. K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empir. Methods Natural Lang. Process. (EMNLP)*, Doha, Qatar, 2014, pp. 1724-1734, doi: 10.3115/v1/D14-1179. https://aclanthology.org/D14-1179/

3. M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572-587, Mar. 2011, doi: 10.1016/j.patcog.2010.09.020. https://doi.org/10.1016/j.patcog.2010.09.020

4. H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," Neural Networks, vol. 92, pp. 60-68, Aug. 2017, doi: 10.1016/j.neunet.2017.02.013. Correct https://www.sciencedirect.com/science/article/pii/S089360801730059X

5. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735-1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735. https://direct.mit.edu/neco/article/9/8/1735/6109/Long-Short-Term-Memory

6. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980 https://arxiv.org/abs/1412.6980

7. S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," in *Proc. Interspeech 2018*, Hyderabad, India, 2018, pp. 257-261, doi: 10.21437/Interspeech.2018-1514. https://arxiv.org/abs/1801.06353

8. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, May 2015, doi: 10.1038/nature14539. https://www.nature.com/articles/nature14539

9. Li, J. Liu, Z. Yang, L. Sun, and Z. Wang, "Speech emotion recognition using recurrent neural networks with directional self-attention," *Expert Syst. Appl.*, vol. 173, Jul. 2021, Art no. 114683, doi: 10.1016/j.eswa.2021.114683. https://www.sciencedirect.com/science/article/pii/S095741742100124X

10. W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Jeju, Korea (South), 2016, pp. 1-4, doi: 10.1109/APSIPA.2016.7820699. https://ieeexplore.ieee.org/document/7820699

11. S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art no. e0196391, doi: 10.1371/journal.pone.0196391. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196391

12. Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," IEEE Trans. Multimedia, vol. 16, no. 8, pp. 2203-2213, Dec. 2014, doi: 10.1109/TMM.2014.2360798. https://ieeexplore.ieee.org/document/6913013

13. B. McFee et al., "librosa: Audio and music signal analysis in Python," in Proc. 14th Python Sci. Conf., Austin, TX, USA, 2015, pp. 18-25, doi: 10.25080/Majora-7b98e3ed-003. https://proceedings.scipy.org/articles/Majora-7b98e3ed-003.pdf

14. S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, New Orleans, LA, USA, 2017, pp. 2227-2231, doi: 10.1109/ICASSP.2017.7952172. https://ieeexplore.ieee.org/abstract/document/7952552

15. B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90-99, May 2018, doi: 10.1145/3126594. https://dl.acm.org/doi/abs/10.1145/3129340

16. B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," Speech Commun., vol. 53, no. 9-10, pp. 1062-1087, Nov. 2011, doi: 10.1016/j.specom.2011.01.011. https://www.sciencedirect.com/science/article/pii/S0167639311000185

17. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929-1958, 2014. https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf

18. G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, 2016, pp. 5200-5204, doi: 10.1109/ICASSP.2016.7472669. https://ieeexplore.ieee.org/document/7472669

19. P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Calgary, AB, Canada, 2018, pp. 5089-5093, doi: 10.1109/ICASSP.2018.8461368. https://ieeexplore.ieee.org/abstract/document/8462677

20. P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," in *Proc. Interspeech 2018*, Hyderabad, India, 2018, pp. 3688-3692, doi: 10.21437/Interspeech.2018-1811. https://www.isca-archive.org/interspeech_2018/yenigalla18_interspeech.html

21. J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D and 2D RNN LSTM networks," Biomed. Signal Process. Control, vol. 47, pp. 312-323, Jan. 2019, doi: 10.1016/j.bspc.2018.12.004. https://www.sciencedirect.com/science/article/pii/S1746809418302337