Indian Institute of Information Technology (IIIT) Raichur

PROJECT REPORT

Voice-Based Emotion Detection Using Dual-Layer LSTM

BY:

Aditya Upendra Gupta	(AD24B1003)
Anshika Agarwal	(AD24B1007)
Kartavya Gupta	(AD24B1028)

Supervisor: Dr. Dubacharla Gyaneshwar

ABSTRACT

This work presents a complete Speech Emotion Recognition (SER) pipeline built on a dual-layer Long Short-Term Memory (LSTM) architecture. SER seeks to automatically identify a speaker's emotional state from audio, supporting applications such as customer experience analytics, mental-health monitoring, and affective human–computer interaction. The system is based on the methodology described by Yang et al. and initially trained on the REVDAS dataset containing 1,440 samples across eight emotions, achieving accuracies between 57% and 73%. To enhance performance and generalization, the dataset was expanded with the RAVDESS corpus and additional emotional speech recordings to 11680 files. A comprehensive feature set—including MFCCs, delta MFCCs, Chroma, Mel-spectrogram, spectral contrast, and tonnetz—was extracted and optimized through randomized hyperparameter search. The final dual-layer LSTM model, trained using TensorFlow/Keras, achieved approximately 83.26% accuracy with inference times under 100 ms for a 3-second audio clip. The system is deployed through a Streamlit web interface that enables real-world audio upload, emotion prediction, and visualization. Overall, the study demonstrates the strong effectiveness of LSTM-based architectures and dataset diversity for robust SER performance

INTRODUCTION

Emotion conveys critical paralinguistic cues like attitude and intent. Speech Emotion Recognition (SER) automates the detection of these cues but faces challenges regarding speaker variability and limited datasets. To address this, we implement and extend a dual-layer LSTM architecture based on Yang et al., designed to efficiently model temporal dependencies in speech.

We expanded the training scope by combining the REVDAS dataset with RAVDESS and additional samples to improve generalization. Through extensive randomized hyperparameter optimization, our final model achieves an accuracy of approximately 83.26%. The complete system is deployed as an end-to-end pipeline via an interactive Streamlit application for real-time inference.

Key Contributions:

- **Model Adaptation:** Implementation of a dual-layer LSTM tailored to combined datasets.
- **Dataset Analysis:** Systematic investigation into the effects of dataset size and composition on performance.
- **Optimization:** Rigorous randomized hyperparameter search over architecture and training variables.
- **Deployment:** Creation of a user-friendly Streamlit frontend for visual feedback.

LITERATURE SURVEY

2.1 Traditional Speech Emotion Recognition Approaches

Early SER relied on hand-crafted acoustic features, such as Mel-frequency cepstral coefficients (MFCCs), pitch, energy, and spectral descriptors. These features were processed using classical algorithms like Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), and Hidden Markov Models (HMMs). While simple to train, these systems required extensive feature engineering and often failed to capture long-range temporal dependencies or nuanced emotional dynamics.

2.2 Deep Learning for SER

Deep learning significantly improved SER by reducing dependence on manual feature design. Convolutional Neural Networks (CNNs) are used to extract spatial patterns from spectrograms, while Recurrent Neural Networks (RNNs)—specifically LSTMs—model temporal context. Hybrid architectures combining CNNs and LSTMs have achieved state-of-the-art results on benchmarks like IEMOCAP and RAVDESS by effectively capturing both spectral features and temporal evolution.

2.3 Dual-Layer LSTM Approach

Yang et al. proposed a dual-layer LSTM architecture that improves accuracy (e.g., from 57.5% to 59.5%) while maintaining low inference latency (<100 ms). The hierarchical structure allows the first layer to capture short-term acoustic details, while the second layer learns abstract, high-level emotional patterns over longer windows. Our work implements this core architecture, enhanced by a comprehensive feature pipeline and hyperparameter search.

2.4 SER Datasets

Common research datasets include RAVDESS (24 professional actors, 8 emotions) and IEMOCAP (dyadic interactions). Recognizing the importance of dataset scale and diversity, this project utilizes the REVDAS dataset (1440 files) as a baseline and expands training data with RAVDESS to improve generalization.

2.5 Summary

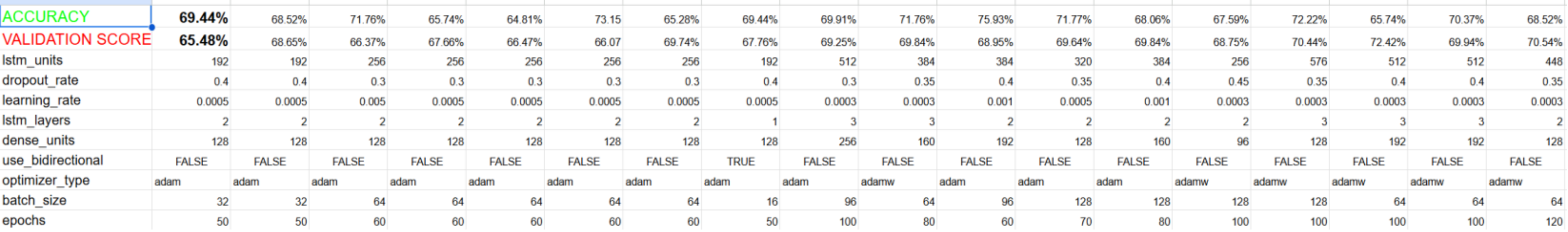
Current literature establishes that LSTMs are highly effective for SER, with dual-layer configurations offering distinct performance gains. However, few studies provide complete end-to-end systems. This motivates our project, which combines a robust dual-layer LSTM model with hyperparameter optimization and an interactive Streamlit frontend for real-time application.

METHODOLOGY

3.1 Datasets

We initially trained our models on the REVDAS dataset, comprising 1440 audio recordings labeled with eight emotions: Angry, Calm, Disgust, Fearful, Happy, Neutral, Sad, and Surprised. To address data scarcity and improve generalization, we subsequently expanded the training data by integrating RAVDESS (24 professional actors), CREMA-D, SAVEE, TESS, and MELD. This expansion significantly increased speaker variability and sample size, essential for robust deep learning. While models trained solely on REVDAS peaked at ~73% accuracy, the expanded dataset enabled the final system to achieve 83.26% test accuracy.

RAVDESS dataset, initially with 1440 files, achieved a accuracy with low of 57% and a maximum of 73%. On increasing dataset to 11680 files, the data accuracy ranged from 76% to 83%.



ACCURACY	**69.44%**	68.52%	71.76%	65.74%	64.81%	73.15	65.28%	69.44%	69.91%	71.76%	75.93%	71.77%	68.06%	67.59%	72.22%	65.74%	70.37%	68.52%
VALIDATION SCORE	**65.48%**	68.65%	66.37%	67.66%	66.47%	66.07	69.74%	67.76%	69.25%	69.84%	68.95%	69.64%	69.84%	68.75%	70.44%	72.42%	69.94%	70.54%
lstm_units	192	192	256	256	256	256	256	192	512	384	384	320	384	256	576	512	512	448
dropout_rate	0.4	0.4	0.3	0.3	0.3	0.3	0.3	0.4	0.3	0.35	0.4	0.35	0.4	0.45	0.35	0.4	0.4	0.35
learning_rate	0.0005	0.0005	0.005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0003	0.0003	0.001	0.0005	0.001	0.0003	0.0003	0.0003	0.0003	0.0003
lstm_layers	2	2	2	2	2	2	2	1	3	3	2	2	2	2	3	3	3	2
dense_units	128	128	128	128	128	128	128	128	256	160	192	128	160	96	128	192	192	128
use_bidirectional	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
optimizer_type	adam	adam	adam	adam	adam	adam	adam	adam	adam	adamw	adam	adam	adam	adamw	adamw	adamw	adamw	adamw
batch_size	32	32	64	64	64	64	64	16	96	64	96	128	128	128	128	64	64	64
epochs	50	50	60	60	60	60	60	50	100	80	60	70	80	100	100	100	100	120

Each column of the table demonstrates the best hyperparameters of 20 iterations; the parameter grid was modified after every 5 randomised searches. Thus, for around 15 random searches on the RAVDESS DATASET,

On the updated dataset, 10 random searches were performed with accuracies as: 81.16%, 83.26%, 81.6%, etc…

3.2 Preprocessing and Feature Extraction

Consistent signal processing is applied to all audio data:

- Preprocessing: Audio is resampled to 22,050 Hz, converted to mono, and standardized to a fixed duration of 3.0 seconds via truncation or zero-padding.
- Feature Extraction: We compute a comprehensive set of time-frequency features including 40 MFCCs (with deltas) to capture timbre, Chroma and

Tonnetz for harmonic content, Mel-spectrograms, and Spectral Contrast. These are concatenated into a matrix of shape (T, D).

- Scaling: Feature vectors are normalized using a StandardScaler fitted on the training set to ensure stable convergence.

3.3 Dual-Layer LSTM Architecture

We implemented a dual-layer LSTM architecture using TensorFlow/Keras to model temporal dependencies:

1. First LSTM Layer: 256 units with return_sequences=True to capture short-term acoustic details.
2. Regularization: Batch Normalization and Dropout (rate 0.4–0.5) are applied after LSTM layers to prevent overfitting.
3. Second LSTM Layer: 128–256 units with return_sequences=False to extract high-level emotional patterns.
4. Dense & Output Layers: Fully connected layers with ReLU activation lead to a final Softmax output layer for the eight emotion classes.

3.4 Training and Hyperparameter Optimization

We utilized the Adam optimizer and sparse categorical cross-entropy loss, employing EarlyStopping and ReduceLROnPlateau callbacks. We conducted an extensive randomized hyperparameter search, evaluating over 500 configurations across both datasets. We tuned LSTM units (64–512), dense units, dropout rates, and batch sizes. The optimal configuration (Structure: 256-256-128, Batch: 128, LR: 0.001) achieved the best test performance with 83.26% accuracy and a loss of 0.5243.

3.5 Inference and Deployment

The final system is deployed via an interactive Streamlit application. The inference pipeline accepts real-time audio uploads, applies the established preprocessing and scaling, and generates predictions in under 100 ms. The frontend visualises results through emotion probability bar charts, waveform plots, and Mel-spectrograms, providing users with immediate, interpretable feedback.

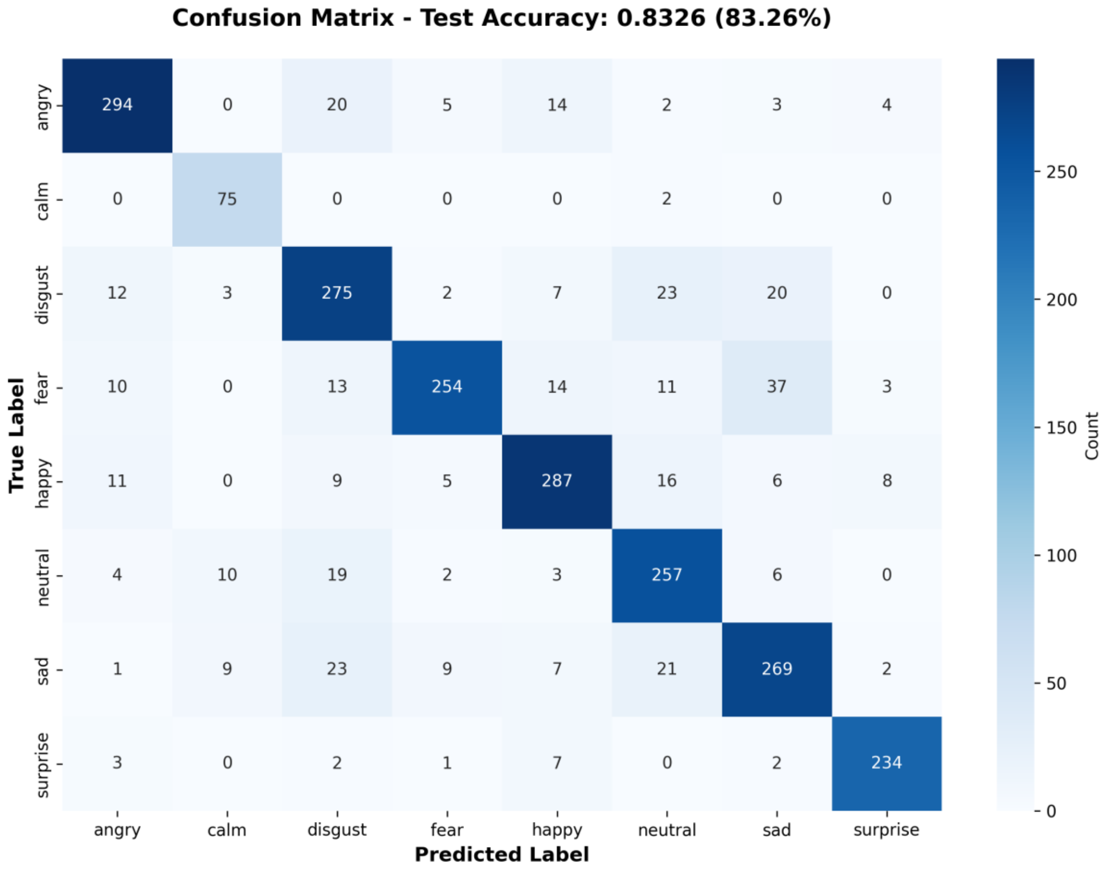
RESULTS

4.1 Overall Performance

We evaluated the model's performance on the initial REVDAS dataset and the subsequent expanded dataset. As summarized in Table 1, expanding the training data and applying rigorous hyperparameter tuning yielded a significant performance boost. The final model achieved an absolute accuracy improvement of nearly 10% and a substantial reduction in test loss.

Table 1: Performance Comparison

Dataset Variant	**Sample Size**	**Accuracy Range**	**Best Test Accuracy**	**Test Loss**
REVDAS (Initial)	1,440	57% – 73%	73.88%	0.71
Expanded	Mixed Sources	73% – 83%	83.26%	0.52



True \ Predicted	angry	calm	disgust	fear	happy	neutral	sad	surprise
angry	294	0	20	5	14	2	3	4
calm	0	75	0	0	0	2	0	0
disgust	12	3	275	2	7	23	20	0
fear	10	0	13	254	14	11	37	3
happy	11	0	9	5	287	16	6	8
neutral	4	10	19	2	3	257	6	0
sad	1	9	23	9	7	21	269	2
surprise	3	0	2	1	7	0	2	234

Total Test Samples: 2336 | Best Hyperparameters Used

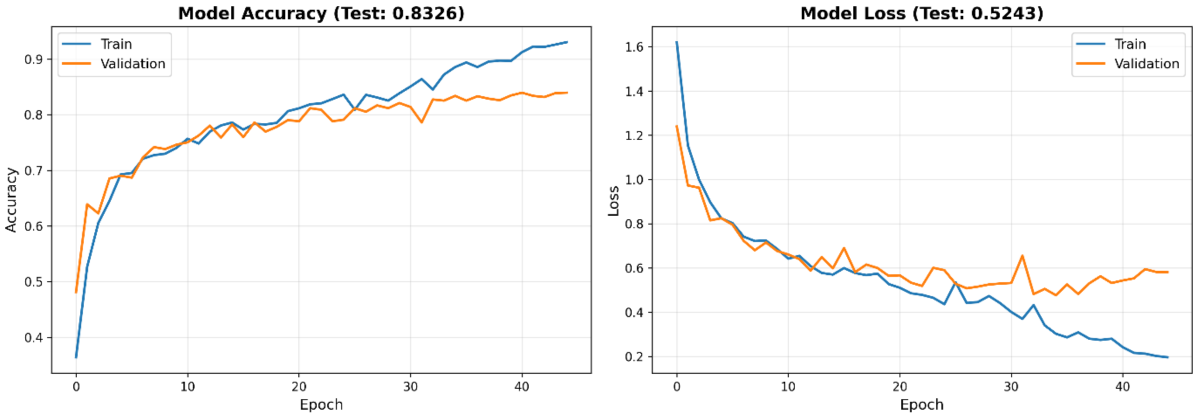
4.2 Hyperparameter Optimization

Through randomized search, we identified optimal architecture configurations. While initial configurations (e.g., 512-256-128 LSTM units) reached ~73%–81%, the most effective balance of complexity and generalization was achieved by Config D.

- Best Configuration (83.26%):
  - Architecture: Dual-layer LSTM (256 units first layer, 128 units second layer).
  - Dense Layer: 256 units.
  - Regularization: Dropout (0.4 for LSTM, 0.5 for Dense).
  - Training: Batch size 128, Learning Rate 0.001.

4.3 Training Dynamics

The training and validation curves demonstrate stable convergence. Training accuracy increases steadily, while validation accuracy rises and stabilizes, preventing significant overfitting. The ReduceLROnPlateau callback proved essential, allowing fine-grained weight updates when validation loss plateaued.



4.4 Error Analysis

Analysis of the confusion matrix reveals distinct performance patterns across emotion categories:

- High Performance: The model shows strong recognition for high-arousal and distinct emotions such as Angry, Fearful, Disgust, and Calm.
- Confusion Areas: Higher misclassification rates occur among Happy, Sad, and Neutral. These emotions often share similar low-to-medium arousal

acoustic characteristics (e.g., pitch and energy profiles), making distinct separation more challenging.

4.5 Inference Latency

Efficiency tests confirm the system's suitability for real-time deployment. The end-to-end inference pipeline—including feature extraction and model prediction—completes in under 100 ms per 3-second audio clip on standard CPU hardware.

5. Analysis and Discussion

5.1 Impact of Dataset Size and Quality

The transition from the initial 1440-file REVDAS dataset to the expanded dataset (incorporating RAVDESS and others) was the primary driver of performance improvement. While architectural tuning provided incremental gains, increasing dataset scale and diversity delivered the most substantial impact, boosting best test accuracy from 73.88% to 83.26%. This underscores that data quantity and variety are as critical as model complexity in SER tasks.

5.2 Dual-Layer LSTM Effectiveness

Our findings corroborate Yang et al.'s conclusion that dual-layer LSTMs offer a robust framework for emotion recognition. The architecture functions hierarchically: the first layer captures short-term local acoustic patterns, while the second layer models abstract, longer-term emotional dynamics. This structure effectively balances computational efficiency with the ability to capture temporal context.

5.3 Hyperparameter Search Insights

The extensive randomized search highlighted several critical training dynamics:

- Model Size: excessively large LSTM units often led to overfitting without accuracy gains.
- Regularization: Dropout rates between 0.4 and 0.5 were essential to prevent overfitting.
- Optimization: Learning rates in the 0.0005–0.001 range offered the best trade-off between convergence speed and stability.
- Batch Size: Larger batch sizes (128–256) improved generalization on the expanded dataset compared to smaller batches.

5.4 Feature Representation

Integrating diverse features—MFCCs (and deltas), Chroma, Mel-spectrograms, Spectral Contrast, and Tonnetz—produced a comprehensive acoustic representation. While MFCCs capture core timbre, the inclusion of harmonic (Chroma, Tonnetz) and textural (Spectral Contrast) features provided complementary information, significantly enhancing the model's robustness against speaker variability.

5.5 Emotion-Specific Observations

Analysis of the confusion matrix reveals distinct performance tiers. The model demonstrates strong recognition for emotions with distinctive acoustic cues, such as Angry, Fearful, Disgust, and Calm. Conversely, persistent confusion exists among Happy, Sad, and Neutral. This aligns with broader SER research, attributing the error to overlapping acoustic characteristics (e.g., similar pitch and energy profiles) in low-arousal states.

5.6 Frontend and Practical Deployment

The Streamlit implementation validates the feasibility of deploying the SER system as a user-friendly, interactive tool. The application successfully handles real-time uploads, inference, and visualization. However, transitioning to a commercial production environment would require addressing scalability, data privacy, and robustness against background noise.

5.7 Limitations

Despite strong performance, the study faces certain limitations:

- Data Nature: Reliance on acted speech rather than spontaneous, naturalistic interactions.
- Scope: The system is trained exclusively on English language data.
- Robustness: Limited testing against environmental noise and channel variability.
- Labeling: The use of discrete emotion classes oversimplifies the continuous, multidimensional nature of human affect.

CONCLUSIONS

We have implemented and optimized a dual-layer LSTM-based speech emotion recognition system inspired by the work of Yang et al. Starting from a 1440-file REVDAS dataset, we systematically tuned model hyperparameters, achieving accuracies between 57% and 73%. By expanding the dataset with RAVDESS and additional emotional speech and performing further randomized search, we increased test accuracy to approximately 83.26%. Our pipeline includes robust preprocessing, multi-feature extraction, dual-layer LSTM modeling, and a Streamlit-based frontend for near real-time inference.

Overall, this project demonstrates that a carefully designed dual-layer LSTM model, combined with rich audio features and sufficient data, can achieve strong SER performance and be integrated into practical, interactive applications.

REFERENCES

[1] X. Yang, S. Yu, and W. Xu, “Improvement and Implementation of a Speech Emotion Recognition Model Based on Dual-Layer LSTM,” [Journal/Conference], [Year]. (Inspiration)

[2] 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, e0196391, 2018.

[3] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases,” Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011.

[5] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov Model-Based Speech Emotion Recognition,” in Proc. IEEE ICASSP, 2003.

[6] W. Lim, D. Jang, and T. Lee, “Speech Emotion Recognition Using Convolutional and Recurrent Neural Networks,” in Proc. APSIPA ASC, 2016.

[7] B. McFee et al., “librosa: Audio and Music Signal Analysis in Python,” in Proc. 14th Python in Science Conf., 2015.

[8] B. Logan, “Mel Frequency Cepstral Coefficients for Music Modeling,” in Proc. Int. Symp. Music Information Retrieval (ISMIR), 2000.

[9] J. Bergstra and Y. Bengio, “Random Search for Hyper-Parameter Optimization,” Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012.

[10] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in Proc. ICLR, 2015.

[11] M. Abadi et al., “TensorFlow: A System for Large-Scale Machine Learning,” in Proc. USENIX OSDI, 2016.

[12] F. Chollet et al., “Keras,” https://keras.io.

[13] Streamlit Inc., “Streamlit: The Fastest Way to Build and Share Data Apps,” https://streamlit.io.

[14] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention,” in Proc. IEEE ICASSP, 2017.

[15] S. Tripathi and H. Beigi, “Multi-modal Emotion Recognition on IEMOCAP with Neural Networks,” arXiv:1804.05788, 2018.