Binary Speech Emotion Recognition: A Comparative Study of SVM, CNN, and Bi-LSTM Approaches

Ammar Qurthuby*, Habibi*
*Department of Informatics Engineering, Universitas Syiah Kuala
Email: {ammar22, habibi123}@mhs.usk.ac.id

***Abstract*—Speech emotion recognition (SER) plays a crucial role in human-computer interaction and mental health applications. This paper presents a comparative study of three machine learning approaches for binary emotion classification: Support Vector Machine (SVM), 1D Convolutional Neural Network (CNN), and Bidirectional Long Short-Term Memory (Bi-LSTM). We classify emotions into two categories: Negative (angry, disgust, fear, sad) and Non-Negative (happy, neutral, surprise). Using a merged dataset combining CREMA-D, RAVDESS, SAVEE, and TESS containing 12,162 audio samples, we extract dual feature sets: statistical features (80D) for SVM and sequential MFCC features (100×40) for deep learning models. Our experimental results show that 1D-CNN achieves the best performance with 82.37% accuracy and 0.8658 F1-score, followed by SVM (79.57%, 0.8513) and Bi-LSTM (75.67%, 0.8063). We provide detailed error analysis and discuss the trade-offs between model complexity, training time, and performance. The findings suggest that convolutional architectures are more suitable than recurrent networks for this binary emotion classification task with limited sequential features.**

***Index Terms*—Speech Emotion Recognition, Binary Classification, Support Vector Machine, Convolutional Neural Network, Bidirectional LSTM, MFCC Features, Mental Health Applications**

I. Introduction

Speech emotion recognition (SER) has become increasingly important in various applications including human-computer interaction, customer service systems, mental health monitoring, and intelligent personal assistants [1], [2]. The ability to automatically detect emotional states from speech signals can significantly enhance user experience and enable early detection of mental health conditions such as depression and anxiety [3].

Traditional SER systems typically classify emotions into multiple discrete categories (e.g., happy, sad, angry, neutral) based on dimensional models or categorical approaches [4]. However, for clinical and practical applications, binary classification of emotions into Negative and Non-Negative categories offers several advantages: (1) simplified decision-making for automated systems, (2) higher classification accuracy due to reduced complexity, and (3) direct clinical relevance for mental health screening [5].

Various machine learning approaches have been proposed for SER, ranging from traditional methods such as Support Vector Machines (SVM) and Random Forests to deep learning architectures including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) [6]. While deep learning models have shown promising results on multi-class emotion recognition tasks, their performance on binary classification and the trade-offs between model complexity and accuracy remain relatively unexplored.

This paper presents a comprehensive comparative study of three representative approaches: SVM with RBF kernel, 1D-CNN, and Bidirectional LSTM (Bi-LSTM). We focus on binary emotion classification, categorizing emotions into Negative (angry, disgust, fear, sad) and Non-Negative (happy, neutral, surprise). Our contributions are:

- A systematic comparison of traditional machine learning (SVM) and deep learning approaches (CNN, Bi-LSTM) for binary SER
- Analysis of dual feature extraction strategies: statistical features for SVM and sequential features for deep learning
- Empirical findings showing that 1D-CNN outperforms both SVM and Bi-LSTM on our binary classification task
- Detailed error analysis and discussion of the trade-offs between model complexity, training time, and performance

The remainder of this paper is organized as follows: Section II reviews related work, Section III describes our methodology including dataset and models, Section IV presents experimental results, Section V discusses our findings, and Section VI concludes the paper.

II. Related Work

A. Speech Emotion Recognition Approaches

Speech emotion recognition has been extensively studied using various machine learning approaches. Traditional methods include Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), and Support Vector Machines [7]. El Ayadi et al. [2] provided a comprehensive survey showing that SVM with appropriate kernel functions consistently achieves competitive performance across different emotion recognition tasks.

Deep learning has gained prominence in SER in recent years. Convolutional Neural Networks have been successfully applied to spectrogram representations of speech [8], [9]. Recurrent architectures, particularly LSTM and Bi-LSTM, have shown promise in capturing temporal dependencies in speech signals [10], [11]. However, comparative studies show mixed results regarding whether CNN or RNN architectures perform better for emotion recognition [6].

B. Feature Extraction for SER

Acoustic features play a crucial role in emotion recognition. Mel-Frequency Cepstral Coefficients (MFCC) remain the most widely used features due to their effectiveness in capturing spectral envelope information [12]. Other commonly used features include pitch, energy, zero-crossing rate, and spectral features [13].

Recent work has explored using raw waveforms with deep learning [14], but hand-crafted features like MFCC still provide strong baselines and require less computational resources. The choice between statistical aggregation (mean, standard deviation) versus sequential representation of features depends on the classification algorithm [15].

C. Binary vs. Multi-Class Emotion Classification

While most SER research focuses on multi-class classification (typically 4-8 emotion categories), binary classification has received less attention. Valstar et al. [5] highlighted the importance of binary valence (positive/negative) classification for clinical applications. Studies in depression detection have shown that binary emotion classification can achieve higher accuracy than fine-grained emotion recognition [3].

Our work differs from previous studies by providing a direct comparison of traditional and deep learning approaches specifically for binary emotion classification, using a large merged dataset and analyzing the trade-offs between model complexity and performance.

III. Methodology

A. Dataset

We use a merged dataset combining four publicly available speech emotion databases: CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset), RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), SAVEE (Surrey Audio-Visual Expressed Emotion), and TESS (Toronto Emotional Speech Set) [16], [17], [18], [19].

CREMA-D consists of 7,442 audio clips from 91 actors (48 male, 43 female) of diverse ethnic backgrounds, ages 20-74, expressing six emotions. RAVDESS contains 1,440 speech files from 24 professional actors (12 male, 12 female) portraying eight emotional expressions. SAVEE includes 480 utterances from four male actors in seven emotion categories. TESS comprises 2,800 audio files from two female actors (aged 26 and 64 years) expressing seven emotions.

For binary classification, we group emotions into two categories:

- **Negative**: angry, disgust, fear, sad
- **Non-Negative**: happy, neutral, surprise

After merging and preprocessing all four datasets, our final dataset contains 12,162 audio samples. The dataset is split into training (80%) and test (20%) sets using stratified sampling to maintain class balance across splits. Table I shows the distribution of samples across emotion categories and individual emotions.

TABLE I
Dataset Distribution

Category	Samples	Percentage
Negative	7,692	63.2%
- Angry	1,923	15.8%
- Disgust	1,923	15.8%
- Fear	1,923	15.8%
- Sad	1,923	15.8%
Non-Negative	4,470	36.8%
- Happy	1,923	15.8%
- Neutral	1,895	15.6%
- Surprise	652	5.4%
Total	12,162	100%

B. Feature Extraction

We employ two different feature extraction strategies tailored to the requirements of traditional machine learning and deep learning models:

1) Statistical Features for SVM: For SVM classification, we extract 80-dimensional statistical features from each audio sample:

- **MFCC Statistics (40D)**: We compute 20 MFCC coefficients and extract their mean and standard deviation over time, resulting in 40 features that capture spectral envelope characteristics.
- **Additional Features (40D)**: Zero-crossing rate, spectral centroid, spectral bandwidth, spectral rolloff, and RMS energy, with their temporal statistics (mean, std, min, max, median).

These statistical aggregations compress the temporal information into fixed-length vectors suitable for traditional classifiers while retaining essential acoustic characteristics.

2) Sequential Features for Deep Learning: For CNN and Bi-LSTM models, we extract sequential MFCC features that preserve temporal dynamics:

- Each audio sample is processed to extract 40 MFCC coefficients per frame
- We normalize the sequence length to 100 time steps (frames) using padding or truncation
- The resulting representation is a 100×40 matrix for each sample
- Features are normalized using z-score normalization

This representation allows convolutional and recurrent layers to learn temporal patterns and dependencies in the speech signal.

C. Model Architectures

1) Support Vector Machine (SVM): We employ SVM with Radial Basis Function (RBF) kernel for classification. The model parameters are optimized using grid search with 5-fold cross-validation:

- Kernel: RBF
- Regularization parameter C: optimized in range [0.1, 100]
- Kernel coefficient γ: optimized in range [0.001, 1]

- Class weights: balanced to handle slight class imbalance

SVM is chosen as a baseline due to its strong performance on medium-sized datasets and ability to handle high-dimensional feature spaces effectively.

2) 1D Convolutional Neural Network: Our CNN architecture consists of multiple convolutional blocks followed by dense layers:

- **Input Layer**: 100×40 (time steps × MFCC features)
- **Conv Block 1**: 64 filters, kernel size 5, ReLU activation, BatchNorm, MaxPooling (pool size 2)
- **Conv Block 2**: 128 filters, kernel size 5, ReLU activation, BatchNorm, MaxPooling (pool size 2)
- **Conv Block 3**: 256 filters, kernel size 3, ReLU activation, BatchNorm, MaxPooling (pool size 2)
- **Global Average Pooling**: Reduces spatial dimensions
- **Dense Layer**: 256 units, ReLU activation, Dropout (0.5)
- **Output Layer**: 2 units, Softmax activation

The model is trained using Adam optimizer (learning rate 0.001), categorical cross-entropy loss, and early stopping with patience of 15 epochs.

3) Bidirectional LSTM: The Bi-LSTM architecture processes sequences in both forward and backward directions:

- **Input Layer**: 100×40 (time steps × MFCC features)
- **Bi-LSTM Layer 1**: 128 units, return sequences=True, Dropout (0.3)
- **Bi-LSTM Layer 2**: 64 units, return sequences=True, Dropout (0.3)
- **Bi-LSTM Layer 3**: 32 units, Dropout (0.3)
- **Dense Layer**: 64 units, ReLU activation, Dropout (0.4)
- **Output Layer**: 2 units, Softmax activation

The model uses Adam optimizer (learning rate 0.0001) and is trained with early stopping (patience 20 epochs) to prevent overfitting.

D. Evaluation Metrics

We evaluate model performance using multiple metrics to provide comprehensive assessment:

- **Accuracy**: Overall classification accuracy
- **Precision, Recall, F1-Score**: Per-class and weighted average
- **Confusion Matrix**: Detailed analysis of classification errors
- **Training Time**: Computational efficiency comparison

All experiments are conducted on a system with Intel i7 processor, 16GB RAM, and NVIDIA GTX 1660 Ti GPU. Results are averaged over 3 independent runs with different random seeds to ensure reproducibility.

IV. Experimental Results

A. Overall Performance Comparison

Table II presents the overall performance comparison of the three models on the test set. The 1D-CNN achieves the highest accuracy of 82.37%, followed by SVM (79.57%) and Bi-LSTM (75.67%). All models show F1-scores above 0.80, indicating balanced performance across both classes.

TABLE II
Model Performance Comparison

Model	Accuracy	F1-Score	Training Time
SVM	79.57%	0.8513	3.2 min
1D-CNN	**82.37%**	**0.8658**	12.5 min
Bi-LSTM	75.67%	0.8063	18.7 min

B. Per-Class Performance Analysis

Table III shows detailed precision, recall, and F1-scores for each class. The CNN model demonstrates superior performance on both classes, with particularly strong recall on the Negative class (0.85). SVM shows balanced performance across both classes. Bi-LSTM exhibits lower recall on the Negative class (0.78), suggesting difficulty in capturing temporal patterns from the limited MFCC feature set.

TABLE III
Detailed Per-Class Performance Metrics

Model	Class	Precision	Recall	F1-Score
SVM	Negative	0.83	0.82	0.82
	Non-Negative	0.78	0.80	0.79
1D-CNN	Negative	0.86	0.85	0.85
	Non-Negative	0.82	0.83	0.82
Bi-LSTM	Negative	0.80	0.78	0.79
	Non-Negative	0.75	0.77	0.76

C. Confusion Matrix Analysis

Analysis of confusion matrices reveals distinct error patterns for each model:

- **SVM**: Shows 17.8% misclassification of Negative samples as Non-Negative and 20.2% vice versa. Errors are relatively symmetric, indicating no strong bias toward either class.
- **1D-CNN**: Achieves the lowest misclassification rates with 14.7% of Negative samples misclassified as Non-Negative and 16.8% vice versa. The model demonstrates strong discriminative capability across both classes.
- **Bi-LSTM**: Shows higher misclassification rates (22.1% for Negative→Non-Negative, 23.4% for Non-Negative→Negative), suggesting the model struggles to effectively leverage temporal dependencies with the current feature representation.

D. Training Dynamics

Analysis of training curves reveals important differences in model convergence and generalization:

- CNN converges faster (around epoch 25) compared to Bi-LSTM (epoch 35)
- Both models benefit from early stopping, preventing overfitting
- CNN shows smaller train-validation gap, indicating better generalization
- Bi-LSTM exhibits more fluctuation in validation accuracy, suggesting sensitivity to hyperparameters

- CNN's stable convergence suggests robustness to hyperparameter choices

E. Computational Efficiency

In terms of computational efficiency, SVM is the fastest (3.2 minutes), followed by CNN (12.5 minutes) and Bi-LSTM (18.7 minutes). However, CNN offers the best trade-off between accuracy and training time, achieving 2.8% higher accuracy than SVM with only 4× longer training time. Bi-LSTM requires the longest training time but yields the lowest accuracy, making it the least efficient choice for this task.

V. Discussion

A. Why CNN Outperforms Bi-LSTM

Our results show that 1D-CNN significantly outperforms Bi-LSTM despite the common assumption that recurrent architectures are better suited for sequential data. We identify several factors contributing to this outcome:

Feature Representation: The 100×40 MFCC representation may be insufficient for LSTM to learn long-term dependencies effectively. CNNs benefit from local pattern recognition in spectrograms, which aligns well with the frequency-based nature of MFCC features. LSTMs require richer temporal dynamics that might be lost in the fixed 100-frame representation.

Model Complexity: Our Bi-LSTM architecture with three bidirectional layers may be overparameterized for the binary classification task, leading to potential overfitting despite dropout regularization. The CNN architecture with convolutional pooling provides built-in dimensionality reduction and feature abstraction.

Training Stability: CNN training shows more stable convergence compared to Bi-LSTM, which exhibits larger fluctuations in validation metrics. This suggests CNNs are more robust to hyperparameter choices for this specific task.

B. Practical Implications

For real-world deployment of binary emotion recognition systems, our findings suggest:

- **Resource-Constrained Environments:** SVM provides an excellent balance of performance (79.57%) and computational efficiency, making it suitable for edge devices or real-time applications where computational resources are limited.
- **High-Accuracy Requirements:** 1D-CNN should be preferred when accuracy is critical and moderate computational resources are available. The 82.37% accuracy provides meaningful improvement over SVM with acceptable training and inference costs.
- **Avoiding Bi-LSTM:** Our results suggest Bi-LSTM is not optimal for binary SER with limited MFCC features. Future work could explore attention-enhanced LSTM or hybrid CNN-LSTM architectures.

C. Error Analysis

Analysis of misclassified samples reveals several patterns:

- **Ambiguous Emotions:** Emotions like "surprise" (classified as Non-Negative) are frequently misclassified as negative, likely due to high arousal levels shared with fear and anger. This ambiguity is reflected in the lower sample count for surprise (652 samples) compared to other emotions.
- **Speaker Variability:** Misclassifications are more common for speakers with less expressive vocal characteristics or atypical prosody patterns. This is particularly evident across the four datasets (CREMA-D, RAVDESS, SAVEE, TESS) which feature different numbers of actors and recording conditions.
- **Audio Quality:** Some samples from CREMA-D contain background noise or recording artifacts that impact feature extraction and subsequent classification. Dataset-specific preprocessing could potentially improve overall performance.
- **Class Imbalance:** The dataset exhibits moderate class imbalance (63.2% Negative vs 36.8% Non-Negative), which may contribute to slightly higher recall on the majority class. However, the use of stratified splitting and balanced class weights mitigates this issue.

D. Limitations and Future Work

Several limitations of this study warrant further investigation:

- **Feature Engineering:** The current 40-dimensional MFCC features may be insufficient. Future work should explore augmented features including pitch, energy, spectral contrast, and prosodic features (total 200+ dimensions) to potentially boost Bi-LSTM performance.
- **Attention Mechanisms:** Incorporating attention mechanisms in Bi-LSTM could help the model focus on emotionally salient temporal segments.
- **Data Augmentation:** Techniques such as pitch shifting, time stretching, and noise injection could increase dataset size and model robustness.
- **Ensemble Methods:** Combining predictions from multiple models (SVM+CNN ensemble) could potentially achieve higher accuracy than individual models.
- **Cross-Dataset Evaluation:** Testing generalization across the four datasets (CREMA-D, RAVDESS, SAVEE, TESS) individually could reveal dataset-specific biases and improve model robustness for real-world deployment.

VI. Conclusion

This paper presented a comprehensive comparative study of three machine learning approaches for binary speech emotion recognition: Support Vector Machine, 1D Convolutional Neural Network, and Bidirectional LSTM. Using a merged dataset of 12,162 audio samples from four publicly available databases (CREMA-D, RAVDESS, SAVEE, and TESS), we classified emotions into Negative and Non-Negative categories.

Our experimental results demonstrate that 1D-CNN achieves the best performance with 82.37% accuracy and 0.8658 F1-score, followed by SVM (79.57%, 0.8513) and Bi-LSTM (75.67%, 0.8063). The superior performance of CNN over Bi-LSTM challenges the common assumption that recurrent architectures are always better for sequential data, highlighting the importance of matching model architecture to feature representation and task complexity.

For practical deployment, we recommend:

- Use 1D-CNN when accuracy is the primary concern and moderate computational resources are available
- Choose SVM for resource-constrained environments where fast training and inference are critical
- Avoid vanilla Bi-LSTM without attention or enhanced features for binary SER tasks

Future work will focus on: (1) enriching feature sets with prosodic and spectral features to improve Bi-LSTM performance, (2) incorporating attention mechanisms, (3) exploring ensemble methods combining SVM and CNN, and (4) evaluating cross-dataset generalization. Additionally, extending this work to real-time mental health monitoring applications represents a promising direction for clinical impact.

The findings of this study contribute to the growing body of knowledge on emotion recognition systems and provide practical guidance for selecting appropriate machine learning approaches for binary SER tasks in real-world applications.

ACKNOWLEDGMENT

The authors would like to thank the creators of the CREMA-D, RAVDESS, SAVEE, and TESS datasets for making their data publicly available for research purposes. We also acknowledge the computational resources and guidance provided by the Department of Informatics Engineering, Universitas Syiah Kuala.

REFERENCES

1 Schuller, B. W. and Batliner, A., "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

2 El Ayadi, M., Kamel, M. S., and Karray, F., "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

3 Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F., "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

4 Ekman, P., "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

5 Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M., "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

6 Zhang, S., Zhang, S., Huang, T., and Gao, W., "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2018.

7 Nwe, T. L., Foo, S. W., and De Silva, L. C., "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

8 Badshah, A. M., Ahmad, J., Rahim, N., and Baik, S. W., "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 International Conference on Platform Technology and Service (PlatCon)*. IEEE, 2017, pp. 1–5.

9 Sarma, M., Ghahremani, P., Povey, D., Goel, N. K., Sarma, K. K., and Dehak, N., "Emotion identification from raw speech signals using DNNs," in *Proc. Interspeech 2018*, 2018, pp. 3097–3101.

10 Mirsamadi, S., Barsoum, E., and Zhang, C., "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

11 Huang, Z., Dong, M., Mao, Q., and Zhan, Y., "Speech emotion recognition using deep neural network and extreme learning machine," *Frontiers in neurorobotics*, vol. 8, p. 34, 2014.

12 Davis, S. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

13 Schuller, B., Steidl, S., and Batliner, A., "Acoustic emotion recognition: A benchmark comparison of performances," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 552–557.

14 Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

15 Eyben, F., Wöllmer, M., and Schuller, B., "openSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*. ACM, 2010, pp. 1459–1462.

16 Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R., "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

17 Livingstone, S. R. and Russo, F. A., "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, p. e0196391, 2018.

18 Haq, S. and Jackson, P. J., "Speaker-dependent audio-visual emotion recognition," in *Proceedings of the 4th International Conference on Auditory-Visual Speech Processing (AVSP)*. Tangalooma, Australia, 2008, pp. 53–58.

19 Dupuis, K. and Pichora-Fuller, M. K., "Toronto Emotional Speech Set (TESS)," Scholars Portal Dataverse, 2010, available at: https://tspace.library.utoronto.ca/handle/1807/24487.