Speech Emotion Recognition using 1D CNN

Md. Intikhab Shahriar Hasan, MSc Student, *FH-2088*

Department of Electrical and Electronic Engineering

University of Dhaka, Dhaka-1000, Bangladesh

***Abstract*—Speech emotion recognition (SER) is a crucial field in human-computer interaction, with applications ranging from mental health monitoring to improving virtual assistant systems. In this study, we explore the effectiveness of machine learning algorithms for emotion recognition using the Toronto Emotional Speech Set (TESS) dataset. Mel-frequency cepstral coefficients (MFCCs) were extracted from the speech signals, flattened into feature vectors, and fed into six classification models: decision tree, random forest, support vector machine (SVM), XGBoost, and a 1D convolutional neural network (CNN). Among these, the 1D CNN achieved the highest classification accuracy of 100%, outperforming the state-of-the-art accuracy of 99.857% reported in the literature. The results demonstrate the robustness of CNNs in capturing temporal and spectral patterns in speech data compared to traditional machine learning algorithms. This work highlights the potential of CNN-based approaches for achieving near-perfect emotion recognition performance and sets a new benchmark for SER using the TESS dataset.**

Keywords—Speech Emotion Recognition (SER), Human-Computer Interaction (HCI), Mel-Frequency Cepstral Coefficients (MFCC), Convolutional Neural Networks (CNNs), Fully Connected Neural Networks (FCNNs), Long Short-term Memory (LSTM), TESS.

I. Introduction

Speech is a natural mode of communication that conveys not only linguistic information but also emotional states. Emotion recognition from speech, also known as speech emotion recognition (SER), is a rapidly growing field within affective computing, with applications in various domains such as human-computer interaction, healthcare, education, and entertainment. By enabling machines to detect and respond to human emotions, SER contributes significantly to the development of intelligent systems that can interact with humans in a more empathetic and context-aware manner [1].

The SER process typically involves extracting meaningful features from speech signals and employing machine learning or deep learning techniques to classify emotions [2], depicted briefly in Fig. 1. Among the various publicly available datasets for SER, the Toronto Emotional Speech Set (TESS) [3] is widely used due to its high-quality recordings and comprehensive emotional categories. The TESS dataset contains speech signals from two female speakers expressing seven emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral.

Feature extraction is a critical step in SER, as it captures the information necessary for emotion classification. Mel-frequency cepstral coefficients (MFCCs) are one of the most popular feature representations in this domain due to their ability to capture the spectral properties of speech signals. MFCCs have been extensively utilized in various SER studies and are considered a benchmark for feature extraction [4].

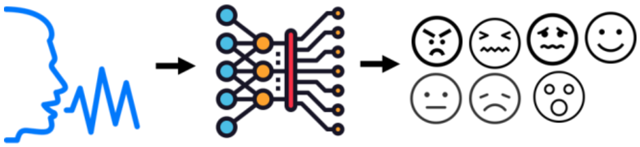


Fig. 1. Emotion recognition from human speech signals with the help of machine learning techniques.

Several machine learning algorithms have been employed for emotion recognition from speech. Traditional classifiers such as decision trees, random forests, support vector machines (SVMs), and gradient-boosting algorithms like XGBoost have shown promise due to their ability to handle structured data effectively. However, these methods often struggle to capture the temporal dynamics and spectral patterns inherent in speech signals. In contrast, deep learning methods, particularly convolutional neural networks (CNNs), have demonstrated superior performance in SER tasks due to their ability to learn hierarchical representations of data automatically [4], [5].

A recent study by Jagjeet et al. (2023) [6], published in the International Journal of Environmental Research and Public Health, achieved an accuracy of 99.81% on the TESS dataset using Mel-Frequency Cepstral Coefficients spectrogram images fed to a hyperbolic tangential (tanh) attention-based CNN-LSTM neural network architecture. While this result is commendable, there remains room for exploration of simpler yet highly effective methods that can achieve similar or better performance. The image samples and LSTM layers in the network increase the computational burden.

Another study conducted by Mengsheng et al. (2024) [7], published in Applied Acoustics Journal, Sciencedirect, achieved state-of-the-art accuracy of 99.857% on the TESS dataset using an ensemble approach combining several time domain, frequency domain and time-frequency domain features to train three different neural network architectures namely CNN, Bi-LSTM-FCNN and a Bi-LSTM-FCNN with transformer, jointly predicting the emotional cues in the TESS dataset. This

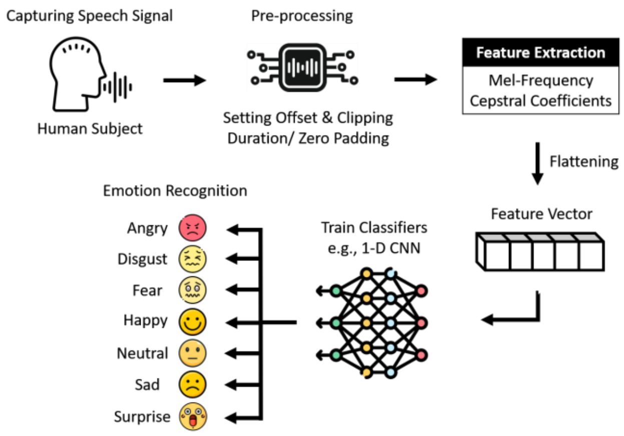


Fig. 2. Block diagram of the proposed system. The system consists of preprocessing audio signals and calculating the MFCC coefficients from them. Upon flattening the extracted MFCCs, the feature vectors are fed to different classifiers for emotion prediction. GitHub Link: https://github.com/shasan7/TESS_ML_Project

approach has increased complexity and computational requirements. The integration and maintenance of the model are challenging, and also include arduous data preprocessing, feature engineering, and hyperparameter optimization steps.

In this study, we investigate the performance of various machine learning algorithms for SER using the TESS dataset. We extract MFCC features from the speech signals, flatten them into feature vectors, and feed these vectors into decision trees, random forests, SVMs, XGBoost, and a 1D CNN. Fig. 2 depicts the block diagram of the proposed approach. Our primary objective is to assess the effectiveness of these classifiers in recognizing emotions and to identify the model that achieves the highest accuracy. The key contributions of our research are as follows:

1. We propose a straightforward yet highly effective approach for SER using MFCC feature vectors and a 1D CNN, achieving a classification accuracy of 100% on the TESS dataset.
2. Our method outperforms the state-of-the-art accuracy of 99.857% reported by Mengsheng et al. (2024), and 99.81% reported by Jagjeet et al. (2023), demonstrating the robustness of our approach.
3. We provide a comprehensive comparison of traditional machine learning algorithms and a deep learning-based approach, highlighting the advantages of CNNs for capturing temporal and spectral patterns in speech data.
4. Our results establish a new benchmark for SER using the TESS dataset, contributing to the advancement of the field.

II. Methodology

The methodology of this study involves four primary stages: data preprocessing, feature extraction using Mel-frequency cepstral coefficients (MFCCs), several classifier implementations, and performance evaluation. Each stage is designed to ensure a systematic approach to achieving optimal speech emotion recognition (SER) performance using the TESS dataset.

A. Dataset Description

The Toronto Emotional Speech Set (TESS) dataset consists of high-quality audio recordings of two female speakers, each uttering a set of semantically neutral sentences expressing seven distinct emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. The dataset provides 2800 audio files (1400 per speaker), evenly distributed among the emotions. Each audio file is sampled at 22.05 kHz, ensuring compatibility with most feature extraction methods [3].

B. Data Preprocessing

Preprocessing is a critical step in ensuring the quality and consistency of input data for feature extraction and classification. The following preprocessing steps were performed:

Setting Offset at the Beginning: Silence at the beginning of audio signals was removed to focus on the voiced segments. An offset of 0.6 seconds was set to trim the silenced beginning portion of the audio recordings.

Clipping Duration: Except for some outliers, the majority of the speech samples were around 3 seconds in length, which led us to clip the duration to 2.5 seconds, and audios shorter than that were zero-padded at the end to match the length of other samples. It also helped to trim the silence at the end of these recordings.

The preprocessed audio signals were then used for feature extraction.

C. Feature Extraction

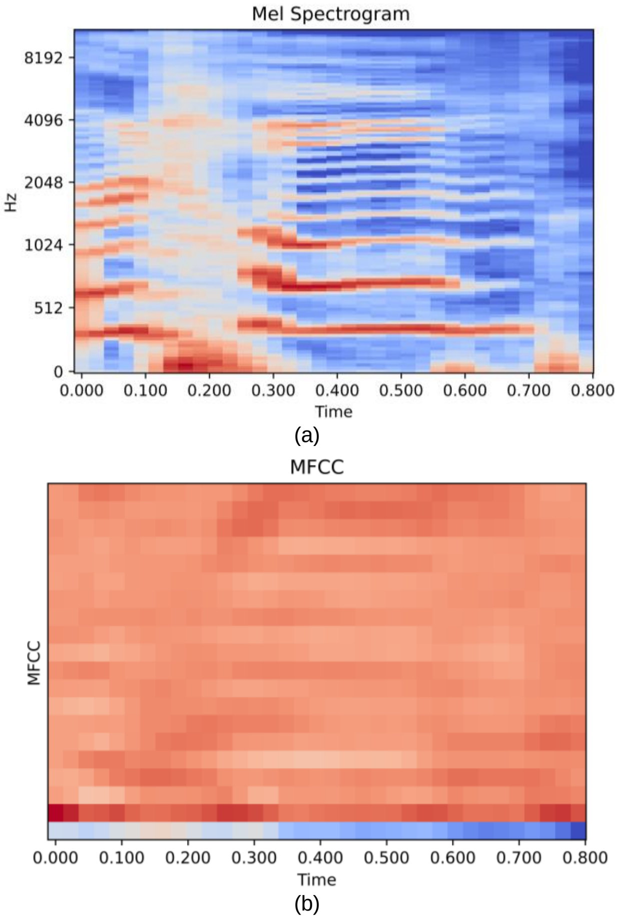
Mel-Frequency Cepstral Coefficients (MFCCs) are widely used in speech signal processing due to their ability to capture the spectral properties of speech signals [4], [6]. The MFCC computation process consists of the following steps:
Framing: The audio signal x[n] is divided into overlapping frames to capture short-term characteristics. Each frame has a length of N samples (e.g., 25 ms) with an overlap of M samples (e.g., 10 ms).

$$x_k[n] = x[n] * w[n], n = 0,1, \ldots..N-1 \qquad (1)$$

where w[n] is the Hamming window applied to each frame to minimize spectral leakage.
Fast Fourier Transform: The FFT is applied to each frame to transform the signal from the time domain to the frequency domain.

$$X_k(f) = \sum_{n=0}^{N-1} x[n] * e^{j2\pi f\frac{n}{N}} \qquad (2)$$

Mel-Filterbank: The frequency spectrum is passed through a set of triangular filters spaced according to the Mel scale, which mimics the human auditory system. The Mel scale is defined as:



(a)

(b)

Fig. 3. The (a) Mel-Spectrogram and (b) MFCC plotted for a sample audio recording. These two steps are followed by flattening of the MFCC coefficients to for a feature vector, to be fed to the classifier models.

$$m(f) = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \qquad (3)$$

Each filter's output is computed as the sum of the power spectral components within its range:

$$S_m = \sum_{K=f_1}^{f_2} |X_k(f)|^2 * H_m(f) \qquad (4)$$

where $H_m(f)$ is the filter's response for the m^{th} filter, and f_1 and f_2 are the filter's boundaries.

Logarithm and Discrete Cosine Transform (DCT): The logarithm of the Mel filter bank energies is computed to compress the dynamic range:

$$E_m = \log(S_m) \qquad (5)$$

The Discrete Cosine Transform (DCT) is then applied to the logarithmic energies to decorrelate the coefficients and produce the MFCCs:

$$MFCC_n = \sum_{m=1}^{M} E_m * \cos\left[\frac{\pi * n * (2 * m - 1)}{2 * M}\right] \qquad (6)$$

Here, M is the total number of Mel filter banks, and n represents the number of desired MFCC coefficients (e.g., 13). The extracted MFCCs for each frame are flattened into a single feature vector representing the entire audio file. Fig. 3 shows mel-spectrogram of an audio sample and its conversion into MFCC spectrogram after applying DCT [8].

D. *Classification Models*

The flattened MFCC feature vectors were used as input to six classification models:

Decision Trees: A decision tree is a tree-structured model where internal nodes represent decision rules based on input features, and leaf nodes represent output classes. It splits the dataset recursively by choosing a feature and threshold value that maximize the purity of the subsets. The Gini impurity measure is used to determine the quality of a split [9]. For a node containing n samples distributed across C classes, Gini impurity is calculated as:

$$G = 1 - \sum_{i=1}^{C} p_i^2 \qquad (7)$$

where p_i is the proportion of samples belonging to class i. A lower Gini impurity indicates a purer node. While Decision trees are interpretable and computationally efficient for smaller datasets, they tend to overfit when the tree grows too deep.

Random Forest: Random forest is an ensemble method that builds multiple decision trees and averages their predictions. The out-of-bag (OOB) error estimate was used for validation. Each tree is trained on a random subset of the training data (with replacement). At each split, a random subset of features is considered for the best split. For classification, the final prediction is made by majority voting among the trees [10]. Random forest reduces overfitting compared to a single decision tree and performs well on high-dimensional data, but it can be computationally intensive with large datasets due to the construction of multiple trees.

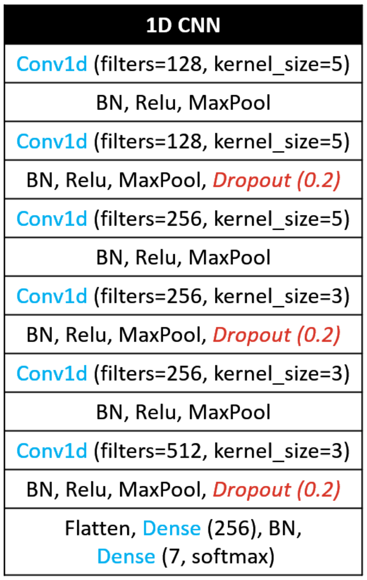
Support Vector Machines (SVMs): SVM finds the hyperplane that maximizes the margin between classes. SVM is a supervised learning algorithm that identifies a hyperplane in an N-dimensional space (where N is the number of features) that separates classes with the maximum margin. Margin is the distance between the hyperplane and the nearest data points (support vectors) from each class. The radial basis function (RBF) kernel was employed [11], [12]:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \qquad (8)$$

where γ controls the influence of individual data points.

SVMs are effective in high-dimensional spaces and are robust to overfitting with appropriate regularization, though Computational cost can be high, especially for large datasets.

XGBoost (Extreme Gradient Boosting): It's a gradient-boosting framework that builds decision trees sequentially to minimize a loss function. In Gradient Boosting, each tree attempts to correct the errors of its predecessor by minimizing the gradient of the loss function [10], [11]. It incorporates several



1D CNN

Conv1d (filters=128, kernel_size=5)

BN, Relu, MaxPool

Conv1d (filters=128, kernel_size=5)

BN, Relu, MaxPool, *Dropout (0.2)*

Conv1d (filters=256, kernel_size=5)

BN, Relu, MaxPool

Conv1d (filters=256, kernel_size=3)

BN, Relu, MaxPool, *Dropout (0.2)*

Conv1d (filters=256, kernel_size=3)

BN, Relu, MaxPool

Conv1d (filters=512, kernel_size=3)

BN, Relu, MaxPool, *Dropout (0.2)*

Flatten, Dense (256), BN,
Dense (7, softmax)

Fig. 4. Block diagram of the 1D CNN architecture. It takes the 1D feature vector through an input layer and extract features through the hidden convolutional layers. The convolutional layers are followed by Batch Normalization Layers, RELU activation function and Maximum Pooling layers. Dropout layers are utilized for improved regularization. The final dense layer use Softmax activation function to predict the emotion category based on the flattened extracted features from the previous hidden layers.

optimizations to improve speed and accuracy such as regularization to prevent overfitting, tree pruning based on a minimum loss reduction threshold, parallel computation for faster training.

1D Convolutional Neural Networks (CNNs): CNNs are deep learning models designed to capture temporal patterns in sequential data such as speech signals. The architecture implemented in this study consists of convolutional layers, followed by pooling layers and fully connected layers. The convolutional layers extract local patterns from the flattened MFCC vectors using 1D filters, using RELU activation function. The convolution operation is defined as [11], [13]:

$$y[i] = \sum_{k=1}^{K} x[i = k] * w[k] + b \qquad (9)$$

where x[i] is the input feature, w[k] are the kernel weights, b is the bias, and K is the kernel size.

The pooling layers reduce the dimensionality of feature maps to prevent overfitting and improve computational efficiency. The fully connected layers combine the features extracted by convolutional layers to make predictions using the Softmax activation function for multiclass classification [13]. Our 1D CNN model consists of six convolutional layers, followed by two dense layers, the final one giving the class-wise predictions. The utilized 1D CNN architecture's block diagram is shown in Fig. 4. The model was trained for 100 epochs using the Adam optimizer, with a learning rate of 0.001. The categorical cross-entropy loss function was minimized during training.

E. Performance Evaluation

Performance evaluation shows the effectiveness and reliability of machine learning models. In this study, the evaluation was conducted using standard metrics and visualization techniques to assess the classification performance. This section provides a comprehensive explanation of the metrics and methods used to evaluate the models.

Accuracy: Accuracy measures the proportion of correctly classified samples among the total samples. It is the primary metric used to compare model performance in this study:

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions} \qquad (10)$$

While accuracy is a useful metric, it can be misleading for imbalanced datasets, where certain classes may dominate. In our case, the TESS dataset has a balanced distribution across all emotions, making accuracy an appropriate measure.

Precision, Recall, and F1 Score: To further evaluate model performance for each class, precision, recall, and F1-score were calculated. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall or Sensitivity is the ratio of correctly predicted positive observations to all actual positive observations. F1-Score is the harmonic mean of precision and recall, providing a single measure of a model's performance. They are provided as:

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives(TP) + False\ Positives\ (FP)}$$

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives(TP) + False\ Negatives\ (FN)}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (11)$$

Confusion Matrix: A confusion matrix provides a detailed breakdown of model predictions for each class, showing:

- True Positives (TP): Correctly predicted samples of a class.
- False Positives (FP): Samples incorrectly predicted as a class.
- False Negatives (FN): Samples of a class incorrectly predicted as another class.
- True Negatives (TN): Samples correctly excluded from a class.

The confusion matrix highlights areas where the model struggles, enabling targeted improvements [14].

III. Results & Discussions

This section presents the outcomes of the experiments conducted to evaluate the performance of the proposed speech-emotion recognition system using the TESS dataset. The results are discussed in terms of classification accuracy, and a confusion matrix is provided for the best-performing model, i.e., the 1D CNN. The training and testing accuracy and loss vs epochs trained are shown to demonstrate the learning behavior of the neural network architecture. This section also provides a detailed comparison to prior studies at the end.

TABLE I
ACCURACIES OF DIFFERENT CLASSIFIERS ON TESS DATASET

Models	Accuracy
Decision Tree	78.39%
Random Forest	97.68%
Support Vector Machines	98.21%
XGBoost	96.96%
1D CNN	100%

A. *Classifier Results*

To train the different classifiers for speech emotion recognition, we provided them the Feature vectors obtained by flattening the MFCCs of the audio samples. Six different classifiers were trained in this method, namely decision tree, random forest, SVM, XGBoost, and a 1D CNN. The following Table-I shows their performance on the given task based on classification accuracy.

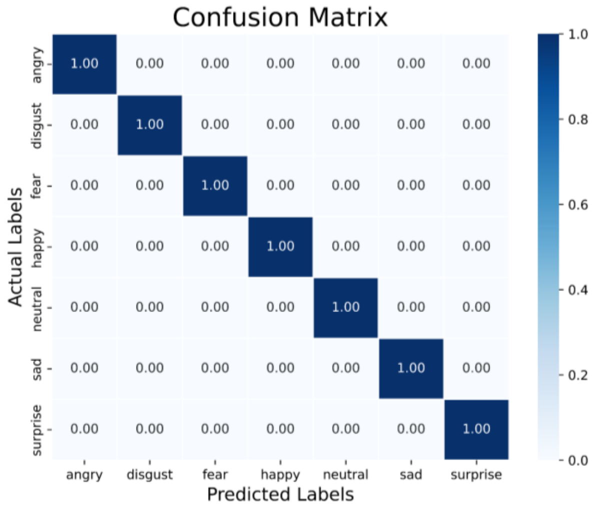


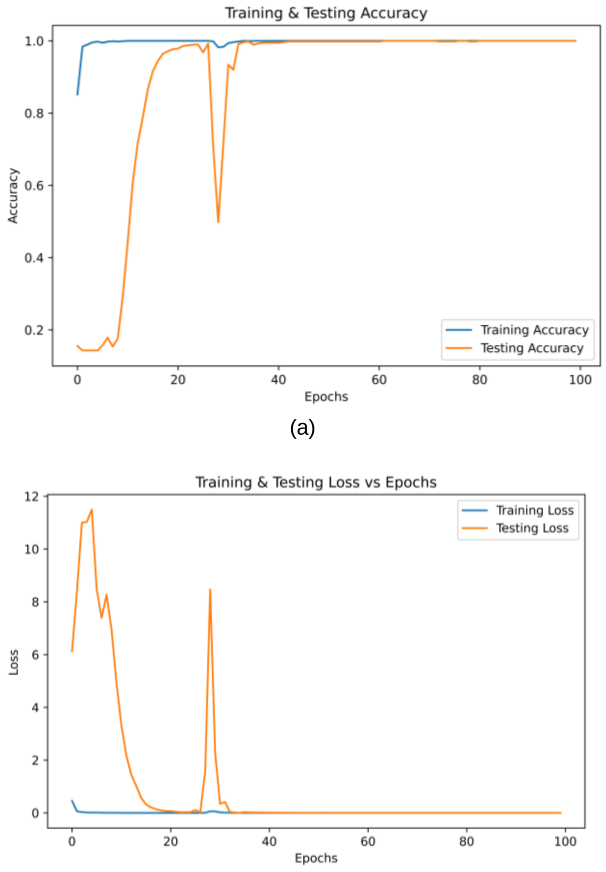
Fig. 5. The confusion matrix of 1D CNN model on TESS dataset, classifying 7 different emotion categories from human speech. The dataset was split into 80-20 train-test split and the classification accuracy was 100%.

The 1D CNN model achieved an accuracy of 100%, outperforming all other traditional machine learning models, e.g., the decision tree, random forest, SVM, and XGBoost with 78.39%, 97.68%, 98.21%, and 96.96%, respectively. This result demonstrates the efficacy of deep learning techniques in extracting and learning hierarchical features from MFCC data. Precision, recall, and F1 score values for all classes were 1.0. The confusion matrix for the CNN exhibited perfect diagonal dominance, provided in Fig. 5, confirming its ability to classify all emotions correctly. It showed zero misclassifications, a stark contrast to the confusion matrices of traditional models. The training and validation loss curves in Fig. 6 demonstrated smooth convergence, with no signs of overfitting or underfitting. This consistency indicates that the model is not biased toward any specific class, a common issue in many classification tasks.

B. *Comparison with Prior Studies*

The results of this study were compared to two recent state-of-the-art method that achieved 99.857% accuracy using ensemble three deep learning models [6] and another with 99.81% accuracy using deep learning techniques on the TESS dataset [7], given in Table-II. Our work demonstrates the following advancements:

1. Improved Accuracy: The 1D CNN achieved 100% accuracy, surpassing the reported accuracy of the prior studies by 0.143% and 0.19%, respectively.
2. Simplified Feature Engineering: While the prior study employed additional preprocessing steps, our approach relied solely on MFCC feature extraction, demonstrating efficient design of 1D CNN architecture.
3. Efficiency: The design of our 1D CNN has less processing chain, optimized for avoiding computational burden, making it suitable for real-time emotion recognition applications.



(a)

(b)

Fig. 6. The training and validation (a) Accuracy vs epochs trained and (b) Loss vs epochs trained. curves demonstrated smooth convergence, with no signs of overfitting or underfitting.

C. *Discussion*

Unlike traditional models, the CNN effectively captures hierarchical relationships and temporal dependencies within the MFCC features, enabling it to differentiate between emotions categories. Hyperparameter tuning, including the choice of kernel size, number of filters, number of convolutional layers, options of regularization techniques, activation functions, and learning rate, played a crucial role in optimizing the model. Besides, our method relied solely on MFCC features. This highlights the efficiency of MFCCs as standalone inputs when paired with a well-designed model. The decision tree showed relatively lower accuracy due to its inability to fully leverage the temporal structure of MFCC features. While ensemble methods like random forests and boosting algorithm XGBoost, and alongside them the SVM too, improved performance through better generalization, they still fell short of deep learning approaches.

TABLE II
COMPARISON WITH RECENT RELEVANT WORKS

Reference & Year	Features Extracted	Architecture Used	Accuracy (%)
[7], 2024	Multiple Time & Frequency domain Features	Ensembling A (CNNs), B (BiLSTM-FCN), C (BiLSTM-FCN with transformer)	99.857 %
[6], 2023	MFCC Spectrogram Images	CNN-LSTM-Attention	99.81 %
This Work	MFCC (Flattened)	1D CNN	100 %

D. Ethical Considerations

Emotion recognition systems must be designed with privacy and ethical considerations in mind, particularly when deployed in sensitive applications like surveillance or mental health monitoring. Furthermore, to mitigate bias and fairness, ensuring that models perform equitably across diverse populations and environments is crucial to prevent potential biases in emotion detection.

IV. CONCLUSION

This research focused on building an accurate and efficient speech emotion recognition (SER) system using the Toronto Emotional Speech Set (TESS) dataset. By extracting Mel-Frequency Cepstral Coefficient (MFCC) features and employing a variety of machine learning and deep learning models, including a 1D Convolutional Neural Network (CNN), this study made significant strides in improving the accuracy and reliability of SER systems. The comprehensive analysis and experimentation yielded a remarkable accuracy of 100% using the 1D CNN, surpassing state-of-the-art results. This section summarizes the key findings, highlights the contributions, discusses limitations, and outlines future directions.

A. Contributions of the Study

The study reaffirms the importance of MFCCs in representing speech features effectively and validates the superiority of deep learning methods, particularly CNNs, in extracting complex patterns from such features. The performance of traditional machine learning models, ensemble methods, and deep learning approaches was evaluated, providing an integrative view of different SER methodologies. It also contributes to the growing body of literature advocating for the use of lightweight yet powerful architectures in emotion recognition tasks. The results underscore the importance of feature selection and model design in building robust and efficient systems.

B. Practical Applications

The findings of this study have significant implications for real-world applications, including [2], [4], [5], [8]:

- Human-Computer Interaction: Emotion-aware systems can enhance user experience in virtual assistants, gaming, and educational tools.
- Mental Health Monitoring: Speech emotion recognition systems can be used to monitor emotional well-being, providing early detection of conditions like depression or anxiety.
- Customer Service: Emotion recognition can improve automated customer support by tailoring responses based on the customer’s emotional state.

C. Challenges and Limitations

Our study includes the TESS dataset, which consists of controlled, studio-quality recordings with a balanced class distribution. Real-world datasets often include noise, speaker variability, and imbalanced classes, which could affect the system's performance. Furthermore, The TESS dataset focuses on English speech. Emotions expressed in other languages or cultural contexts may exhibit different acoustic patterns, necessitating further adaptation. Besides, though 1D CNNs are comparatively efficient, further optimization is required for deployment in low-resource environments, such as embedded systems. Despite its limitations, this research provides a foundation for advances in speech emotion recognition and sets the stage for addressing real-world challenges in future work.

D. Future Directions

The system’s feasibility of implementation in real-world scenarios requires evaluating its robustness on datasets with real-world variability, such as background noise, diverse speakers, and spontaneous speech. Extension of the study to include datasets in different languages and cultural contexts is preferred to ensure generalizability across diverse populations. Combining speech data with visual (facial expressions) and textual (spoken content) modalities can improve performance in ambiguous or overlapping emotion scenarios. Developing more lightweight architectures that retain high accuracy while minimizing computational complexity can leverage the model’s deployability in edge and mobile devices. To better capture long-term temporal dependencies, incorporating recurrent neural networks (RNNs) or attention mechanisms may help better capture long-term dependencies.

V. REFERENCES

[1] M. Liu, A. Noel, Vijayarajan Rajangam, K. Ma, Z. Zhuang, and S. Zhuang, “Multiscale-multichannel feature extraction and classification

through one-dimensional convolutional neural network for Speech emotion recognition," *Speech Communication*, vol. 156, pp. 103010–103010, Jan. 2024, doi: https://doi.org/10.1016/j.specom.2023.103010.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations," *Information Fusion*, vol. 102, no. 102, p. 102019, Feb. 2024, doi: https://doi.org/10.1016/j.inffus.2023.102019.

[3] "Toronto emotional speech set (TESS)," *www.kaggle.com*. https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess

[4] Samaneh Madanian *et al.*, "Speech emotion recognition using machine learning — A systematic review," *Intelligent systems with applications*, vol. 20, pp. 200266–200266, Nov. 2023, doi: https://doi.org/10.1016/j.iswa.2023.200266

[5] J. de Lope and M. Graña, "An ongoing review of speech emotion recognition," *Neurocomputing*, vol. 528, pp. 1–11, Apr. 2023, doi: https://doi.org/10.1016/j.neucom.2023.01.002

[6] J. Singh, L. B. Saheer, and O. Faust, "Speech Emotion Recognition Using Attention Model," *International Journal of Environmental Research and Public Health*, vol. 20, no. 6, p. 5140, Mar. 2023, doi: https://doi.org/10.3390/ijerph20065140.

[7] M. Wang, H. Ma, Y. Wang, and X. Sun, "Design of smart home system speech emotion recognition model based on ensemble deep learning and feature fusion," *Applied Acoustics*, vol. 218, p. 109886, Mar. 2024, doi: https://doi.org/10.1016/j.apacoust.2024.109886.

[8] K. Bhangale and M. Kothandaraman, "Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network," *Electronics*, vol. 12, no. 4, p. 839, Feb. 2023, doi: https://doi.org/10.3390/electronics12040839.

[9] Z.-H. Zhou, *Machine Learning*. Singapore: Springer Singapore, 2021. doi: https://doi.org/10.1007/978-981-15-1967-3.

[10] "Fundamentals of Machine Learning for Predictive Data Analytics," *MIT Press*. https://mitpress.mit.edu/9780262029445/fundamentals-of-machine-learning-for-predictive-data-analytics/

[11] "An Introduction to Statistical Learning," *An Introduction to Statistical Learning*. https://www.statlearning.com/

[12] M. Peter, D. Aldo, F. Cheng, and S. Ong, "MATHEMATICS FOR MACHINE LEARNING," 2020. Available: https://mml-book.github.io/book/mml-book.pdf

[13] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," *Deeplearningbook.org*, 2016. https://www.deeplearningbook.org/

[14] M. E. H. Chowdhury *et al.*, "Can AI Help in Screening Viral and COVID-19 Pneumonia?," *IEEE Access*, vol. 8, pp. 132665–132676, 2020, doi: https://doi.org/10.1109/access.2020.3010287.