Journal of Mines, Metals and Fuels, *73(9): 2933-2940; 2025. DOI: 10.18311/jmmf/2025/49511*

Print ISSN : 0022-2755

Journal of Mines, Metals and Fuels

Contents available at: *www.informaticsjournals.co.in/index.php/jmmf*

Gammatone Cepstral Coefficient-based Speech Emotion Recognition using Machine Learning Techniques

Pagidirayi Anil Kumar[1]*, Aketi Parvathi[1] and S. Sruthi[2]

[1]Department of ECE, G. Pulla Reddy Engineering College, Kurnool - 518002, Andhra Pradesh, India; anilkumar.ece@gprec.ac.in

[2]Department of ECE, Sree Rama Engineering College, Tirupati - 517507, Andhra Pradesh, India

Abstract

This research investigates the effectiveness of different feature extraction techniques for Speech Emotion Recognition (SER) and explores the potential of machine learning to improve accuracy. While Gamma Tone Cepstral Coefficients (GTCC) are designed to capture auditory features that align with human perception, they may not adequately capture subtle emotional cues. Mel Frequency Cepstral Coefficients (MFCC), on the other hand, have shown promise in effectively representing speech signals for emotion recognition. This work compares GTCC-based feature extraction with MFCC and employs a Cubic Support Vector Machine (SVM) classifier to enhance the system's ability to learn and distinguish between emotional states. Using datasets like CREMA-D and SAVEE, this research aims to advance the development of SER systems with improved accuracy and sensitivity for applications in human-computer interaction.

Major Findings: *This study focuses on enhancing GTCC features for better emotion recognition in speech. While MFCC showed higher accuracy, improvements to GTCC narrowed the performance gap. The results indicate that refined GTCC, combined with Cubic SVM, holds promise for effective SER.*

Keywords: *Cubic SVM, GTCC, MFCC, SER*

1.0 Introduction

Speech recognition can be achieved through various data sources, and speech signals offer practical advantages over biological signals. Speech is easily acquired and analysed, making it a cost-effective choice for emotion recognition systems. Furthermore, speech is a natural and efficient mode of human communication, crucial in time-sensitive human-computer interactions. Although significant progress has been made in speech recognition, enabling computers to understand the semantic content of speech, recognising the emotional nuances embedded within speech remains a challenge. This underscores the growing importance of SER research, which aims to bridge this gap and enable more authentic and emotionally intelligent interactions between humans and machines[1,2].

SER aims to assess a speaker's emotional state based on their voice. Expressing emotions through speech is complex, as it depends on various factors and situations. One of the main challenges in SER is identifying the most effective features that can reliably distinguish different emotions. Variations in speech—such as differences

**Author for correspondence*

in speaking styles, sentence structures, and speaker characteristics like speech rate—directly influence key attributes like pitch and energy, which are essential in SER. Extracting these elements from speech signals is crucial for accurately interpreting emotions and improving speech recognition systems. These features are generally categorised into three types: Continuous, Spectral, and Voice Quality, each playing a distinct role in analysing emotional cues within speech[3-6].

SER technology analyses vocal cues like pitch and tone to identify human emotions. This has broad applications across diverse fields: improving customer service interactions, monitoring patient well-being in mental health, creating more responsive human-computer interactions, assessing student engagement in education, enhancing immersion in entertainment, gauging consumer reactions in market research, improving driver safety in automotive, and bolstering security through vocal threat detection. As SER technology continues to evolve, it promises to revolutionise how we interact with machines and understand human emotions[7,8].

SER has a wide range of real-world applications. In a call centre, it helps determine the best way to respond to customers. In vehicles, it can assess a driver's mental state to prevent accidents. SER is also useful in healthcare for detecting emotional cues related to medical conditions. Additionally, it enhances e-learning and storytelling by adapting content based on the audience's mood. Human speech carries various emotions influenced by different situations, with key categories including anger, sad, happiness, fear, surprise and disgust, concerning neutral.

This work is structured as follows: Section II covers related research on SER, while Section III explains the methodology and proposed approach. Section IV presents the simulation results along with relevant tables, and Section V concludes the study.

2.0 Related Work

Hanilci *et al.* demonstrated the effectiveness of Linear Prediction Residual Cepstrum Coefficients (LPRC) for speaker verification, showing comparable performance to MFCC and LPCC features. Using the NIST 2001 Speaker Recognition Evaluation database and a GMM-UBM model, they achieved promising results, particularly when combining LPRC with LPCC and MFCC[9].

Maesa *et al.* developed a text-independent Automatic Speech Recognition (ASR) system using Mel-Frequency Cepstral Coefficients (MFCC) and Gaussian Mixture Models (GMM), achieving 96% accuracy on the Voxforge.org database. This real-time capable system, implemented in Matlab, demonstrates the effectiveness of MFCC and GMM for security access control applications[10].

Langari *et al.* proposed a speech feature extraction method using Principal Component Analysis (PCA) with dynamic time-frequency coefficients for SER. However, evaluations on the EMO-DB and SAVEE databases showed this approach to be ineffective[11].

Lanjewar *et al.* proposed a Speech Emotion Recognition (SER) system using Gaussian Mixture Models (GMM) and k-Nearest Neighbours (kNN) on the Berlin Emotion Database, analysing six emotions. While demonstrating reasonable accuracy for anger, the system struggled to identify other emotions[12].

Sharma explored a SER system using Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN) with the RAVDESS dataset. While achieving 60% accuracy, the system's performance was limited by the front-end feature extraction algorithms[13].

Puri proposed a hybrid approach using 2D Convolutional Neural Networks (CNNs) for Speech Emotion Recognition (SER) on the RAVDESS dataset, focusing on depressive emotions. However, the method's reliance on manual speech segmentation limits its practicality and effectiveness[14].

Huijuan developed a Speech Emotion Recognition (SER) system using a hierarchical multi-task learning architecture. This approach first identifies general categorical features, then leverages this information to enhance emotion recognition accuracy, ultimately improving performance on the IEMOCAP dataset[15].

Dimitri Bitouk *et al.*[16] used the Berlin EMO-DB to achieve 83% accuracy in emotion recognition by extracting spectral features like LPCC and MFCC. Their approach, which applied rank-based subset selection and SVM classification, proved effective for analysing emotional content in speech.

3.0 Proposed Work

Gamma Tone Cepstral Coefficients (GTCC) utilise gammatone filters, which are designed to mimic the human ear's frequency response. This biologically-inspired method enhances the system's ability to extract relevant speech features, especially in noisy conditions, and improves the capture of emotional nuances in speech.

MFCCs are widely used for SER due to their ability to effectively capture essential speech features, making them a reliable choice for feature extraction. In contrast, GTCC face limitations in low-frequency signals, which can reduce their effectiveness in accurately representing emotional variations. To overcome this drawback and improve classification accuracy, machine learning techniques can be applied. By leveraging advanced algorithms, feature selection, and deep learning models, an effort is made to enhance the performance of SER systems, ensuring better differentiation of emotions and improved overall recognition accuracy[17].

Figure 1 represents the block diagram process begins with acquiring speech signals, which then undergo preprocessing to enhance quality. GTCC-based feature extraction is applied to capture essential speech characteristics. These extracted features are then classified using a Cubic SVM model, which helps distinguish between different emotions. The detected emotions are evaluated, leading to the final output, where the classification accuracy is assessed before concluding the process.

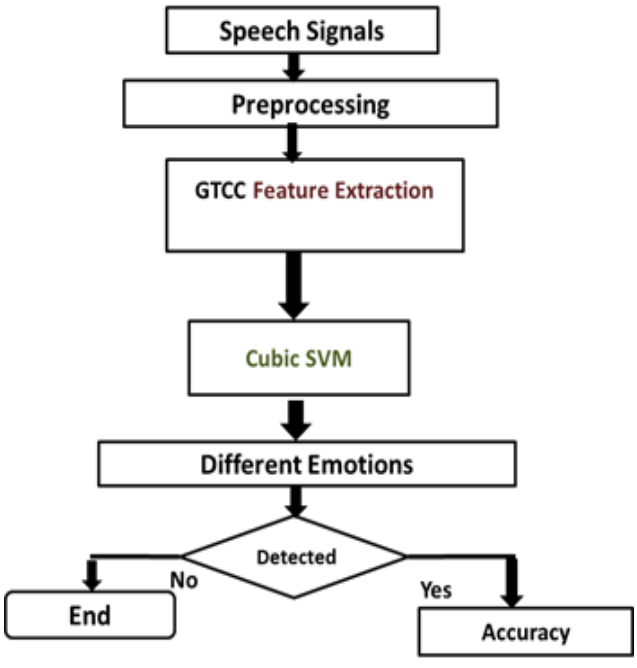


Figure 1. Block diagram of the proposed method.

In an SER system, preprocessing is the first step in speech recognition, helping to differentiate between voiced and unvoiced signals while forming feature vectors. Filtering is used to minimise noise from recordings or the environment, with a pre-emphasis filter boosting higher frequencies for better analysis. This filter improves the Signal-To-Noise Ratio (SNR), making the speech signal more suitable for feature extraction. Additionally, preprocessing ensures that background noise and other disturbances are accounted for in speech signal processing. In the early stages of audio analysis, it is crucial to extract relevant features from a continuous speech signal. This helps in isolating essential components for further processing. The pre-emphasis filter is mathematically represented using the z-transform.

$$H(x) = 1 - ax^{-1} \quad (1)$$

The 'a value ranges between 0.9 and 1.0 in Equation (1). It is well-known that speech signals are not stationary, making them difficult to analyse. After filtering, the speech signal is divided into short overlapping frames, usually 20–40 ms long, to capture its short-term stationary characteristics. This segmentation enhances the accuracy of feature extraction. A window function is then applied to each frame to minimise abrupt transitions that could introduce spectral distortions. The Hamming window is widely used as it smooths frame edges and reduces spectral leakage. The Hamming window w(m) is

$$w(m) = 0.54 - 0.46\cos\left(2\pi\frac{m}{M}\right), \text{where } 0 \leq m \leq M \quad (2)$$

These preprocessing steps refine the speech signal, making it more suitable for feature extraction and enhancing the accuracy and efficiency of further analysis.

4.0 GTCC Feature Extraction

This research utilises the GTCC feature extraction technique, as shown in Figure 2, commonly used in SER systems. The term "Gammatone" originates from the gamma function in statistics. It is preferred as an auditory filter due to its alignment with human auditory perception. The Gammatone function's impulse response is defined by parameters where 'a controls duration and 'm determines filter order. This makes it an effective choice for speech signal analysis.

$$G(t) = a * t^{m-1} * e^{(-2\pi at)} * \cos(2\pi ft + \varphi) \quad (2)$$

Patterson and Moore demonstrated that the Gammatone function closely mimics the human auditory system. It is used to define the Equivalent Rectangular Bandwidth (ERB) of auditory filters. They suggested that a fourth-order filter (n=4) effectively models the human auditory filter. While the exact phase characteristics of the auditory filter are uncertain, it is assumed to be near minimum phase. This assumption is based on the Gammatone filter's minimum phase properties[18].

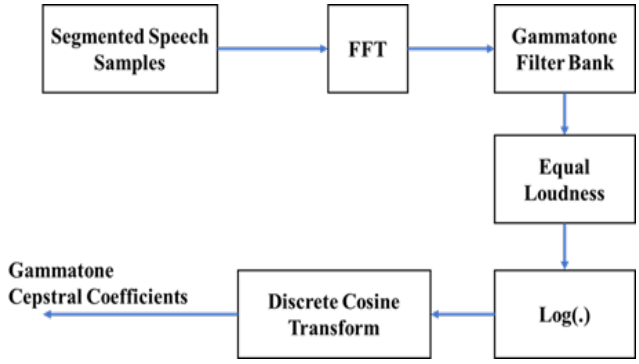


Figure 2. Block diagram for GTCC features extraction process.

The FFT stage is essential for converting speech signals from the time domain to the frequency domain, allowing detailed spectral analysis. This transformation helps extract key speech features, reduce noise, and improve speech recognition and speaker identification.

The Gammatone filter mimics human cochlear frequency analysis, making it valuable in speech and audio processing. Its design, based on the gamma function, is widely used in feature extraction, hearing aids, and noise reduction. Equivalent Rectangular Bandwidth (ERB), the Gammatone filter closely matches the auditory filter bandwidths found in human hearing, given in Equation (3).

$$RB = 24.7 * (\frac{4.37}{1000} * f + 1) \quad (3)$$

In GFCC computation, the logarithm function compresses the dynamic range of filter energy outputs, making feature extraction more robust. This transformation aligns with human auditory perception and enhances resilience to loudness and noise variations. The Discrete Cosine Transform (DCT) further processes these log-compressed values, reducing redundancy and improving feature independence. By compacting key information into fewer coefficients, DCT enhances efficiency in speech recognition and speaker identification.

5.0 Classifiers

Cubic SVM is an advanced variant of the Support Vector Machine (SVM) classifier that utilises a cubic kernel function to map input data into a higher-dimensional space, enabling better separation of complex patterns. This classifier is particularly effective in handling nonlinear data distributions, making it well-suited for applications

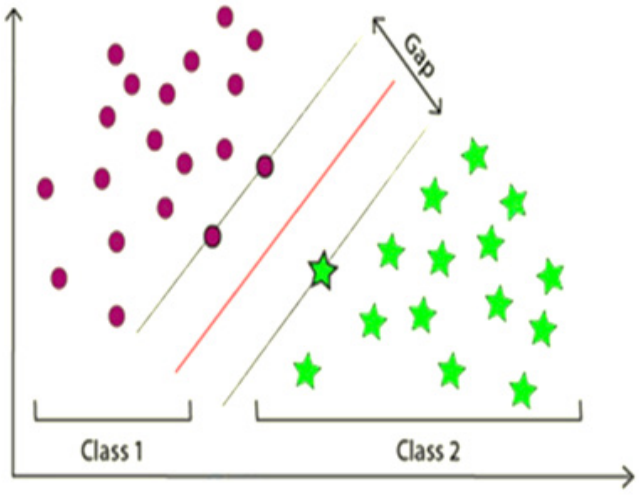


Figure 3. Cubic SVM.

where linear boundaries are insufficient. The cubic kernel enhances flexibility by capturing intricate relationships between data points, improving classification accuracy[19]. Additionally, SVMs, including cubic SVM, are memory-efficient as they rely on support vectors rather than storing the entire dataset, making them suitable for large-scale machine learning tasks shown in Figure 3. The cubic function is given in Equation (4).

$$K(x_i, x_j) = (x_i^T, x_j)^3 \quad (4)$$

Wide Neural Network classifiers are designed with an increased number of neurons per layer, allowing them to capture intricate patterns and relationships in data effectively. They are particularly useful for structured datasets where feature interactions play a crucial role, offering improved generalisation without requiring deep architectures. Compared to deep networks, wide models train faster due to their reduced layer depth, making them suitable for real-time applications and computationally constrained environments. They excel in handling tabular and categorical data by learning meaningful feature transformations, often enhancing tasks in finance, healthcare, and recommendation systems. Additionally, they are frequently integrated with deep networks in hybrid models, such as Wide & Deep Learning, combining memorisation and generalisation capabilities to improve performance in applications like speech recognition, natural language processing, and personalised recommendations.

Cross-validation improves model training by splitting data into multiple training and testing sets, ensuring better generalisation and reducing overfitting. It selects the most frequently appearing features across different subsets for optimal performance. The process involves defining the number of training iterations (n train) and selected features (n select). For instance, if n train = 10 and n select = 2, ten distinct datasets are used, and only features appearing more than twice are finalised. This approach enhances feature selection reliability and ensures robust model performance.

6.0 Results

In SER analysis, speech samples are down-sampled to 16 kHz, and features are extracted for classification using datasets such as SAVEE, CREMA-D, RAVDESS, and TESS

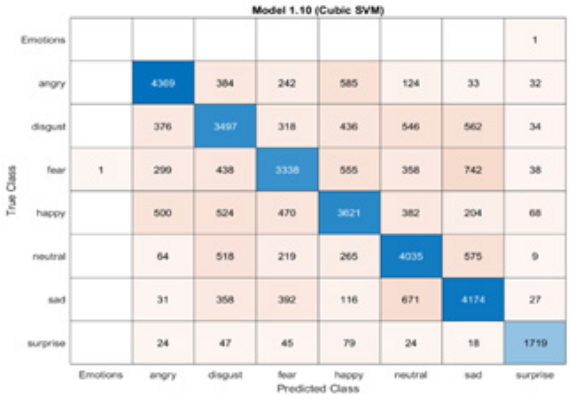


Figure 4. Number of observations of the cubic SVM classifier.

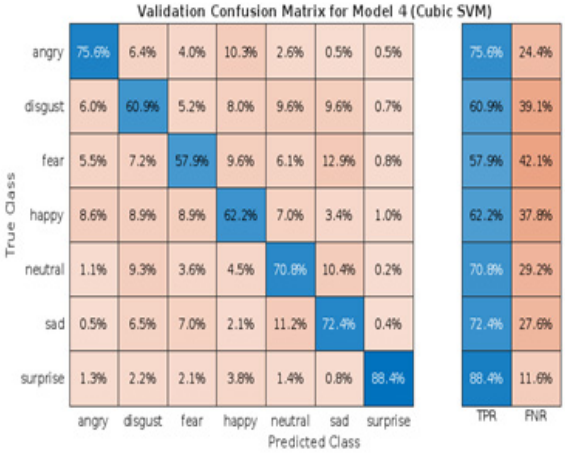


Figure 5. Accuracy of the cubic SVM classifier.

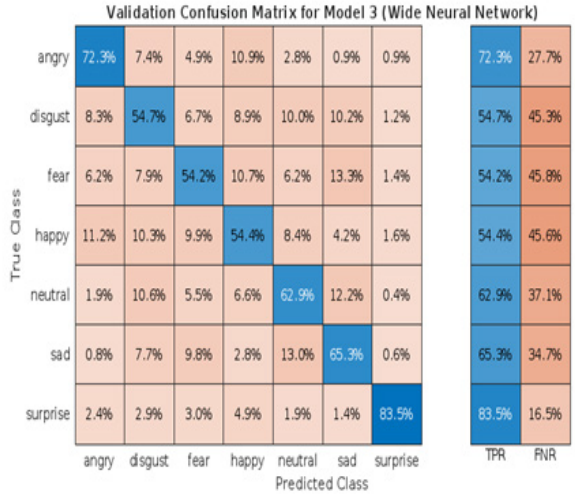


Figure 6. Accuracy of wide neural networks.

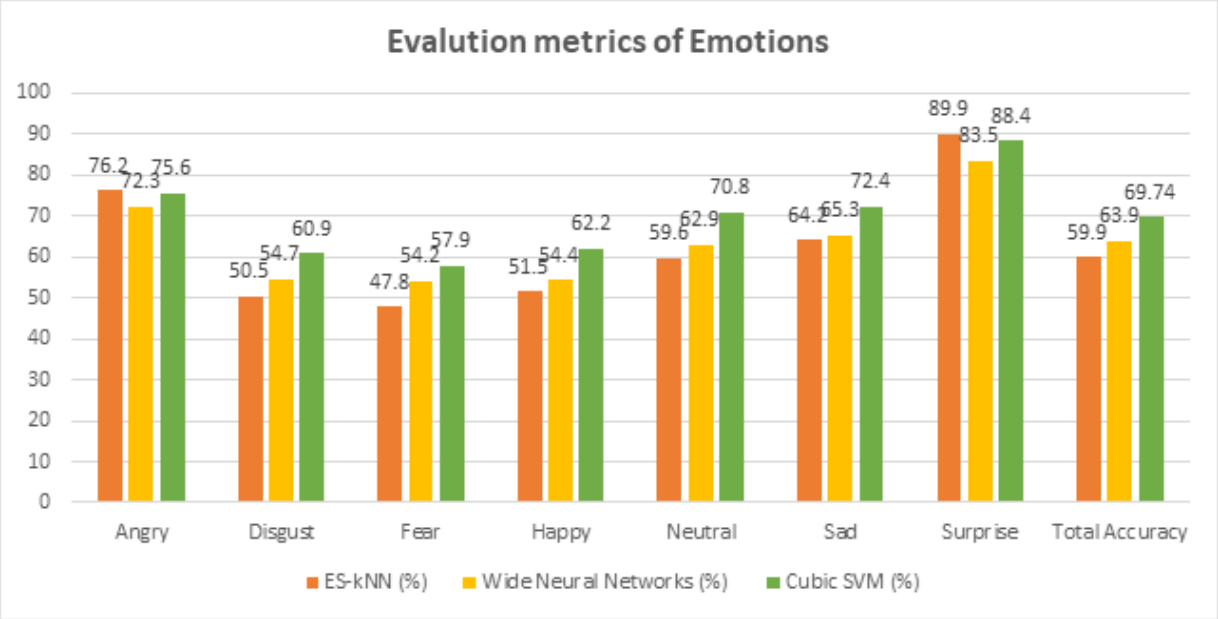


Figure 7. Evaluation metrics of emotions.

Table 1. Comparison table of emotions for ES-KNN, wide neural networks, and cubic SVM

S.No	Emotions	ES-kNN (%)[21]	Wide Neural Networks (%)	Cubic SVM (%)
1.	**Angry**	75.2	72.3	**75.6**
2.	**Disgust**	50.5	54.7	**60.9**
3.	**Fear**	47.8	54.2	**57.9**
4.	**Happy**	51.5	54.4	**62.2**
5.	**Neutral**	59.6	62.9	**70.8**
6.	**Sad**	64.2	65.3	**72.4**
7.	**Surprise**	89.9	83.5	**88.4**
8.	**Total Accuracy**	59.9	63.9	**69.74**

in CSV format. The study compares the performance of different classifiers, including Random ES-kNN, Cubic SVM, and Wide Neural Networks. The dataset consists of 12,162 files with 36,487 observations, divided into training and testing sets[20]. Performance evaluation is conducted using confusion matrices and comparative analysis with prior models. While GTCC-based feature extraction is employed, the Cubic SVM classifier effectively defines decision boundaries, whereas the Wide Neural Network captures complex emotional patterns. A comparative study highlights the strengths of each model in terms of accuracy, robustness, and computational efficiency, providing insights into their suitability for SER applications.

Figures 4 and 5 represent the confusion matrices of Cubic SVM with True positive and true negative, with the number of observations and accuracy table and wide neural networks in Figure 6.

Table 2 represents the comparison of the proposed method with the existing method, which gives a good accuracy compared to other methods and is represented graphically in Figure 7.

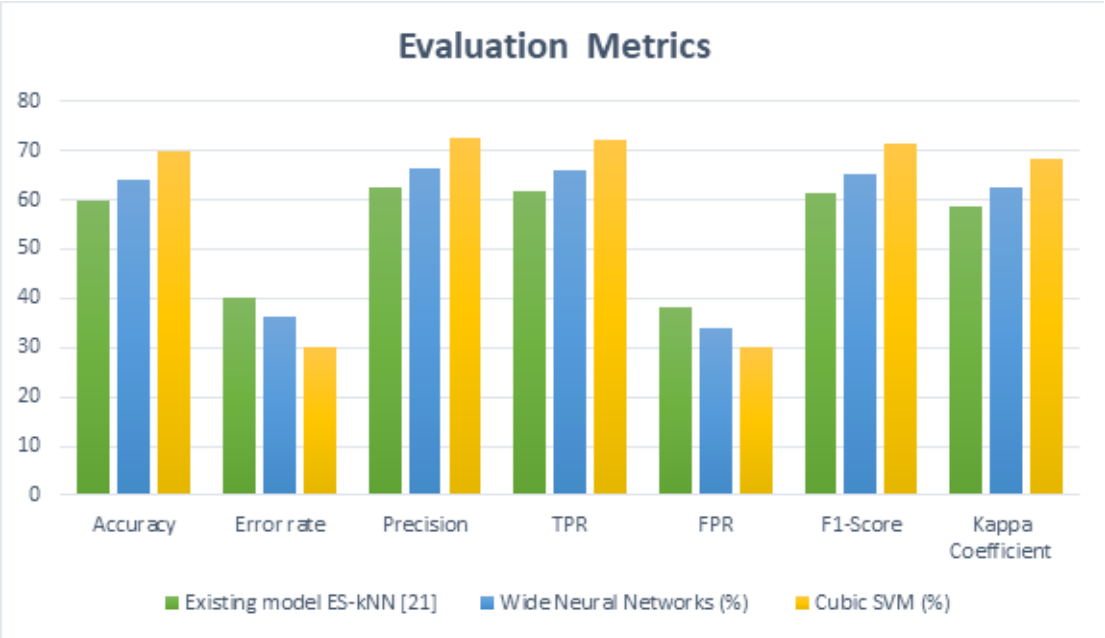


Figure 8. Evaluation Metrics.

Table 2. Comparison of performance metrics of the existing and proposed models

S.No	Metrics(%)	Existing model ES-kNN[21]	Wide Neural Networks (%)	Cubic SVM (%)
1.	**Accuracy**	59.9	63.9	**69.74**
2.	**Error rate**	40.1	36.1	**30.26**
3.	**Precision**	62.55	66.45	**72.52**
4.	**True positive Rate (TPR)**	61.92	66.07	**72.11**
5.	**False positive Rate(FPR)**	38.08	33.93	**30.26**
6.	**F1-Score**	61.29	65.36	**71.34**
7.	**Kappa Coefficient**	58.79	62.62	**68.34**

Table 2 represents the evaluation metrics that depict the accuracy, TPR, FPR, F1-score and Kappa coefficient and are represented graphically in Figure 8.

7.0 Conclusion

The proposed method utilising GTCC feature extraction demonstrates improved accuracy and efficiency in SER compared to existing models. With a cubic SVM classifier achieving the highest accuracy (69.74%) and lowest error rate (30.26%), it outperforms Wide Neural Networks and ES-kNN. The enhanced precision, F1-score, and kappa coefficient indicate the method's reliability in emotion classification. While GTCC shows competitive performance, ongoing improvements aim to further refine its effectiveness, making it a strong alternative to MFCC for speech emotion recognition.

8.0 References

1. Nicholson J, Takahashi K, Nakatsu R. Emotion recognition in speech using neural networks. Neural Comput Appl. 2000; 9:290-6. https://doi.org/10.1007/s005210070006
2. Corive R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, *et al*. Emotion recognition in human-computer interaction. IEEE Signal Process Mag. 2001; 18(1):32-80.
3. Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. J Pers Soc Psychol. 1996; 70(3):614-36. https://doi.org/10.1037//0022-3514.70.3.614
4. Kumar AP, Anuradha B. Bi-LSTM based speech emotion recognition. In: Malhotra M, editor. Innovative emerging trends in computing and information technology. Cham: Springer; 2025. p. 199-213. (Communications in Computer and Information Science; vol. 2125). https://doi.org/10.1007/978-3-031-80839-5_14
5. Schuller B, Batliner A, Steidl S, Schiel F, Krajewski J. The INTERSPEECH 2011 speaker state challenge. In: Proceedings of Interspeech 2011; 27-31; Florence, Italy: ISCA; 2011. p. 3201-4. https://doi.org/10.21437/Interspeech.2011-801
6. Schuller B, Rigoll G, Lang M. Hidden Markov model-based speech emotion recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP); 2003; 6-10; Hong Kong: IEEE; 2003. p. 1-4. https://doi.org/10.1109/ICASSP.2003.1202279
7. Djamel A, Ahmed SS, Boudraa M, Boudraa B. An efficient front-end for distributed speech recognition over mobile. Int J Comput Commun Eng. 2012; 1(3):232-5. https://doi.org/10.7763/IJCCE.2012.V1.60
8. Ververidis D, Kotropoulos C. Emotional speech recognition: resources, features and methods. Speech Commun. 2006; 48(9):1162-81. https://doi.org/10.1016/j.specom.2006.04.003
9. Hu H, Xu MX, Wu W. GMM supervector-based SVM with spectral features for speech emotion recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP); 2007; 15-20; Honolulu, HI: IEEE; 2007. p. 413-6. https://doi.org/10.1109/ICASSP.2007.366937
10. Maesa A, Garzia F, Scarpiniti M, Cusani R. Text independent automatic speaker recognition system using Mel-frequency cepstrum coefficient and Gaussian mixture models. J Inf Secur. 2012; 3(4):335-40. https://doi.org/10.4236/jis.2012.34041
11. Lanjewar RB, Mathurkar S, Patel N. Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-nearest neighbour (K-NN) techniques. Procedia Comput Sci. 2015; 49:50-7. https://doi.org/10.1016/j.procs.2015.04.226
12. Sharma S. Emotion recognition from speech using artificial neural networks. In: Proceedings of the 11th International Conference on Cloud Computing, Data Science and Engineering (Confluence); 2021; 28-29; Noida, India: IEEE; 2021. p. 153-8. https://doi.org/10.1109/Confluence51648.2021.9377192
13. Puri T, Soni M, Dhiman G, Khalaf OI, Alazzam M, Khan IR. Detection of emotion of speech for RAVDESS audio using a hybrid convolutional neural network. J Healthc Eng. 2022; 2022:8472947. https://doi.org/10.1155/2022/8472947
14. Zhao H, Ning Y, Wang R. Coarse-to-fine speech emotion recognition based on multi-task learning. J Signal Process Syst. 2021; 93(2-3):299-308. https://doi.org/10.1007/s11265-020-01538-x
15. S SPBA. Speaker identification and spoken word recognition in a noisy environment using different techniques. Int J Res IT Comput Commun Eng. 2016; 4(6):590-2.
16. Bitouk D, Verma R, Nenkova A. Class-level spectral features for emotion recognition. Speech Commun. 2010; 52(7-8):613-25. https://doi.org/10.1016/j.specom.2010.02.010
17. Fernandes B, Mannepalli K. Speech emotion recognition using deep learning LSTM for Tamil language. Pertanika J Sci Technol. 2021; 29(3):2107-24. https://doi.org/10.47836/pjst.29.3.33
18. Jain U, Nathani K, Ruban N, Raj ANJ, Zhuang Z, Mahesh VGV. Cubic SVM classifier-based feature extraction and emotion detection from speech signals. In: Proceedings of the International Conference on Sensor Networks and Signal Processing (SNSP); 2018; 28-31; Xi'an, China: IEEE; 2018. p. 386-91. https://doi.org/10.1109/SNSP.2018.00081
19. Pagidirayi AK, Bhuma A. Speech emotion recognition using machine learning techniques. Rev Intell Artif. 2022; 36(2):271-8. https://doi.org/10.18280/ria.360211
20. RAVDESS, TESS, CREMA-D, SAVEE datasets. Kaggle. Available from: https://www.kaggle.com
21. Kumar PA, Anuradha B. Speech feature extraction and emotion recognition using deep learning techniques. I-manager's J Digit Signal Process. 2024; 12(2):1-12. https://doi.org/10.26634/jdp.12.2.21179