Exploring the Effectiveness of Advanced Machine Learning Models in Speech Emotion Recognition

Kanika Jangra
Department of Electronics and Communication Enginnering
Lovely Professional University, Phagwara
Punjab, India
17.kanikajangra@gmail.com

Deepika Ghai
Department of Electronics and Communication Enginnering
Lovely Professional University, Phagwara
Punjab, India
money.ghai25@gmail.com

Sandeep Kumar
Department of Computer Science and Engineering
Koneru Lakshmaiah Educational Foundation
Vaddeswaram, India
er.sandeepsahratia@gmail.com

2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE) | 979-8-3503-6684-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/IC3SE62002.2024.10593399

Abstract— **The importance of recognizing emotion from voice stems from the basic human need to understand and communicate emotional states, which is vital in enhancing security, health care, etc. This study compares several advanced machine learning models to assess their effectiveness in recognizing emotions from speech, using the widely accepted RAVDESS, i.e. Ryerson Audiovisual Database of Emotional Speech Song. Our research focuses on the study of depth models of Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) versus conventional machine learning algorithms, like Support Vector Machines (SVMs), Random Forests (RFs), and Long-Range Machines (GBM). Through careful preprocessing, feature extraction using Mel-Frequency Cepstral Coefficients (MFCCs). The research concludes that LSTM performs better at 91% than the other implemented models. Thus, in future, voice-based emotion recognition can help with diagnosis with ongoing monitoring of mental health conditions like depression, anxiety and stress by detecting emotional distress or mood changes.**

Keywords— Machine Learning, emotion detection, voice detection, CNN, LSTM, SVM, GBM, RF, MFCC.

I. Introduction

Human speech includes numerous features that the listener examines to understand the complicated information supplied by the speaker. Inadvertently, the speaker conveys tone, intensity, tempo, and other auditory properties, which help to capture both the subtext or meaning and the precise words. Emotion detection has many applications in medical treatment, security, forensic sciences, and other fields. Models such as LSTM do computations in a timestep sequence. Numeric features are fed into a network of neural networks, which outputs the logit vector. LSTMs, the decoder, was built to be an attention-based machine that trained on the encoder's learnt representation to produce an output chance for the following character sequence. When examining MFCCs as time-series information, LSTMs or their more complicated counterparts are used to address the issue of the speech emotion recognition problem of classification. CNNs work with MFCCs in a single dimension or acquire to recognize Mel spectrograms applying 2D filters.

SET, which stands for speech emotion identification, has two steps: extracting features and categorizing features. Speech-processing researchers have developed several features, such as source-based excitement features, prosodic characteristics, vocal sliding factors, and mixed features. In the second step, nonlinear and linear algorithms are used to sort the features into groups. Bayesian networks (BN), which are sometimes called the Minimum Likelihood Principle (MLP), and Support Vector Machines (SVM) are the linear models that are most often used to recognize emotions. The speech sound is not usually thought of as being fixed. Since this is the case, nonlinear models should do well in SER. SER can be used with several different nonlinear classification methods. These are frequently employed to put data into groups based on basic-level traits.

A lot of the time, energy-based traits like Perceptual Linear Prediction cepstrum coefficients (PLP), Mel-Frequency Cepstrum Coefficients (MFCC), Linear Predictor Coefficients (LPC), and Mel Energy-spectrum Dynamic Coefficients (MEDC) are used to pick out feelings in speech accurately. Deep learning techniques for SER have multiple advantages over traditional methods. For example, they can find complex structures and includes without requiring human feature extraction and tuning. They also prefer to extract features at a low level from raw data and can work with data that has yet to be labelled. Deep Neural Networks (DNNs) with Convolutional Neural Networks (also known as CNN) are good at handling images and videos. Speech-based classification tasks like natural language processing (NLP) and speech recognition (SER) should use recurrent designs, such as recurrent neural networks (RNNs) with Long Short-Term Memory (LSTM). In other words, the study focuses on the efficiency of machine learning and deep-learning algorithms to detect feelings.

II. Literature Work

Speech emotion recognition enhances human-machine interaction through emotional classification [1]. Fusion of spatial and temporal feature depictions for speech emotion recognition [2] achieves higher accuracy on RAVDESS and IEMOCAP datasets and outperforms state-of-the-art models. Emotion recognition based on speech and audio features using MFCC and CNN+LSTM algorithms [3]. Anger and neutral emotion performed best, yielding an accuracy of 61.07%. Deep learning techniques are critical solutions for SER [4]. Speech emotion recognition (SER) is a method for extracting emotions from human speech. The analysis used the RAVDESS data set and achieved an accuracy of 80.64% with CNN LSTM [5]. Hybrid MFCCT features with CNN performed better than MFCC and domain features [6].

A proposed self-concept-based deep learning model for speech perception recognition. The optimized data set obtained an experimental accuracy rate of 90% [7]. Bilingual Arabic English Speech Emotion Recognition System. High performance with low computing costs". Speech emotion recognition using audio features achieves an accuracy of 85%[8]. Detection of sarcasm was done with a score of 75%.

979-8-3503-6684-6/24/$31.00 ©2024 IEEE

"Human speech emotion recognition using CNN[9]. The model outperformed the other models and achieved an accuracy of 94.38%. A variety of audio and machine learning algorithms are used for emotion recognition. A proposed method for speech perception detection using masked sliding windows[10]. A deep neural network-based classifier achieves high accuracy with sentiment data sets. Emotion recognition uses speech signals in the intelligence system[11]. Deep learning techniques for feature extraction and model building. This paper describes a set of sound structures means built on Match Frequency Cepstral Coefficient (MFCC)[12], Wavelet Packet Transformation (WPT), Linear Predictive Cepstral Coefficient (LPCC), Zero Crossing Rate (ZCR), Spectrum Center, Spectral Rolloff. Spectral Kurtosis[13], Root surface square (RMS), pitch, jitter, and shimmer to improve a particular feature[14]. This paper explains acoustic text features in hidden space are used to select a perceptual class with minimum generalized reconstruction error as an SER result, which can be used as an indicator to decide whether the class is neutral or not and thus can be applied to it other classes of perception[15]. Voice is a powerful emotional state; loudness and tone often betray underlying emotional states. Advances in SER systems have been characterized by the inherently language-driven nature of consumer engagement to enhance user experience through responsive and sensitive technology[16].

Early approaches to SER, as described in the literature, included the development of unique classifiers based on extraction methods from speech signals. These classifiers were trained on tone, pitch, and strength to distinguish between emotional states[17]. One study stated that linear discriminant analysis (LDA) and support vector machine (SVM) were used to detect four primary emotions: happy, sad, angry, and neutral. Deep learning, especially 2D Convolutional Neural Networks (CNNs), shows essential progress in the field. CNNs have shown promise in classifying emotions, with a reported accuracy of around 70% when analyzing data sets[18]. Including CNNs highlights the shift towards architectures that can extract and learn the most suitable features for SER tasks with little domain knowledge. A notable case in this area is the RAVDESS, which contains linguistic content with various sensory properties. This data set has contributed to developing SER systems that are more nuanced and capable of understanding complex human emotions[19]. Research suggests that gender-based training can help develop more accurate SER models, emphasizing the importance of individualized programs. One of the recent studies proposed a less complex SER algorithm that showed good performance using only Mel-frequency cepstral coefficients (MFCC)[20-21]. Thus, the survey describes that extensive research in advanced machine learning techniques and deep learning techniques are used for speech emotion detection (SER) [22-23]. However, real-world applications often have environments with variable noise and sound, which can reduce the performance of SER models that are not explicitly designed to handle such situations; our proposed model performs augmentation by noise pitching to check the efficiency of the mode. Thus, the research is also efficient in giving excellent analysis by comparing the outcomes of multiple machine learning and deep learning algorithms.

III. PROPOSED WORK

The proposed model describes the efficient method of detecting emotions using machine learning algorithms, as shown in Figure 1. The model involves various steps for the detection of 6 classes of emotions.

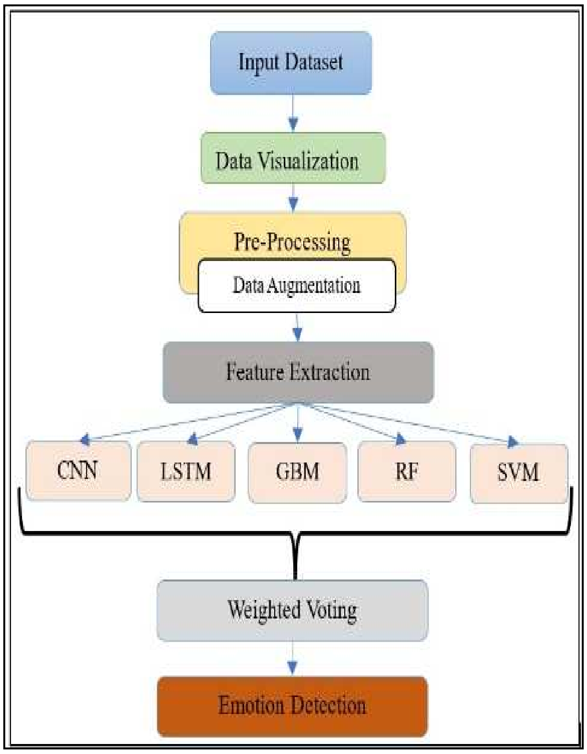


Fig. 1. Flow Chart of Proposed Work

A. Dataset Used

We used the RAVDESS dataset for this study because it is an open-source collection that scientists can use to find out how people feel when they talk. Research the Ryerson Audio-Visual Database of Emotional Speech and Songs, also known as RAVDESS, has 7356 recordings showing emotions. These files have three types: full AV, video-only, and audio-only. There are also two voice channels, one for spoken text and one for song. One character in each file plays one of the eight feelings below: neutral, happy, sad, angry, scared, shocked, or sickened..

B. Data Visualization

Figure 2 shows the count of emotions in the dataset; it describes the colour of each bar indicates a specific emotion, while its height indicates the frequency of that emotion.

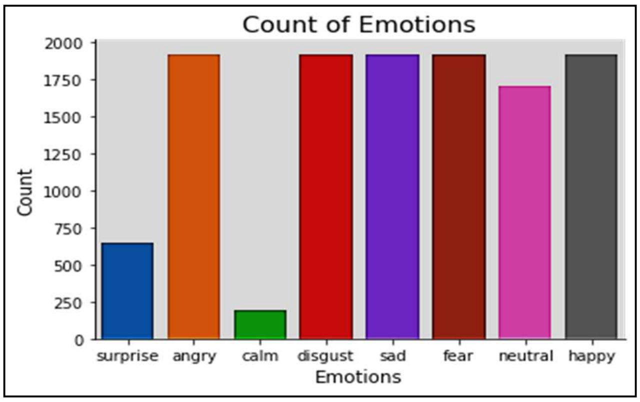


Fig. 2. Count of Emotions in RAVDESS Dataset

The research analysed the highest count of emotions, such as anger, sadness, fear, happiness, and disgust; each visual shows a waveform and a spectrogram with frequency. They explain waves: the figure on the left shows a wave that is the visual magnitude of an acoustic signal. The x-axis indicates time, and the y-axis represents amplitude. Peaks in the waveform indicate where the sound is loudest (highest amplitude) and troughs quietest (lowest amplitude). Spectrogram with fundamental frequency: The graph on the right is a spectrogram, which explains the representation of the spectrum frequencies in sound or other signals as they vary with time. Here again, time at x-3. Axis: the y-axis represents frequency (in Hz) and colours at any given time. The brighter the colour, which indicates the intensity or signal on each frequency, the more energy there is. The cyan line that appears to be traced through the centre of the spectrogram indicates the dominant frequency at which the signal evolves, the lowest frequency of the sound perceived as the pitch of the tone. Figure 3 shows anger. While Figure 4 shows disgust, Figure 5 fear, Figure 6 is happy, and Figure 7 is sad.

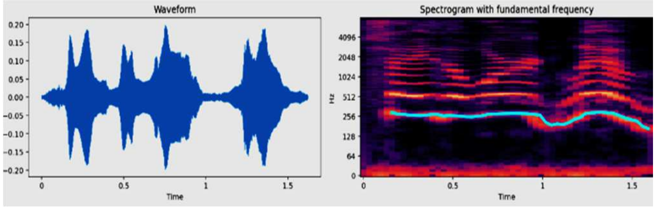


Fig. 3. Emotion: Angry on RAVDESS Dataset

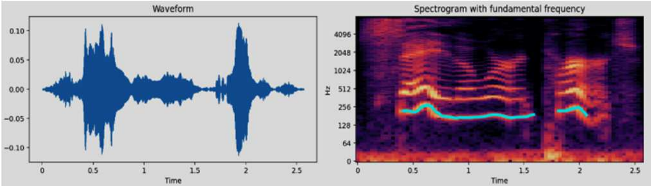


Fig. 4. Emotion: Disgust on RAVDESS Dataset

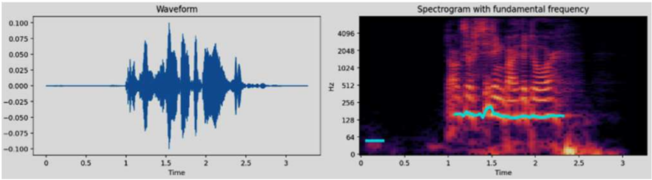


Fig. 5. Emotion: Fear on RAVDESS Dataset

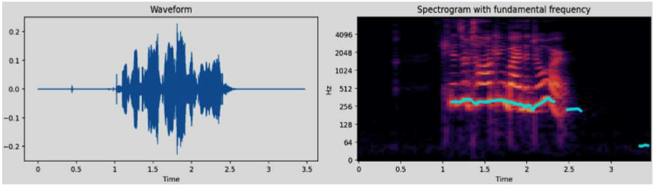


Fig. 6. Emotion: Happy on RAVDESS Dataset

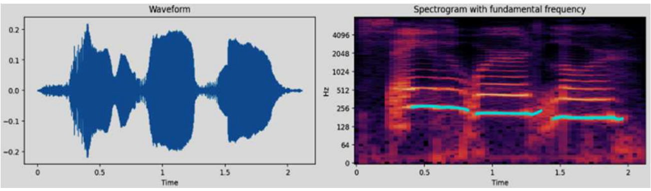


Fig. 7. Emotion: Sad on RAVDESS Dataset

C. *Pre-processing*

In the research, we have implemented the augmentation method for pre-processing in the context of voice recognition for emotional information; data augmentation describes the process of affectedly expanding the amount of data by manipulating existing data. Figure 8 shows the original voice; we have performed some standard data augmentation methods applied to voice emotion recognition, and Figure 8 shows the original voice.

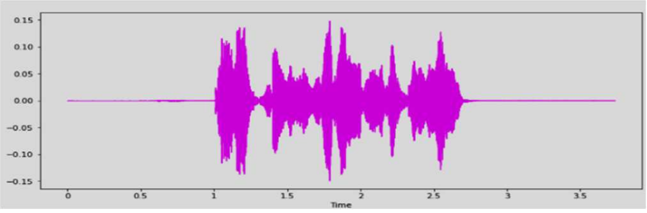


Fig. 8. Original Voice

Noise Injection: Adding background noises to clean audio samples helps to stabilize the image in real-world situations with background noise, the sample voice, by injecting the voice is as shown in Figure 9.

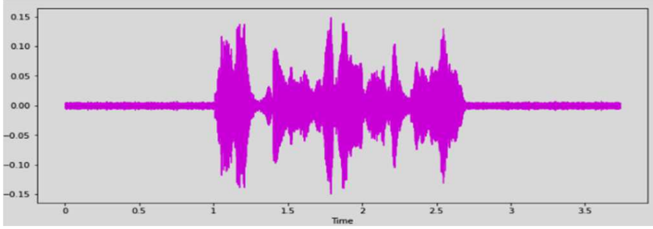


Fig. 9. Noised Voice

Time Stretch/Voice Modification: Using this model, we can recognize emotion even when the speech is different by varying the pronunciation speed without affecting the pronunciation rate (time dilation) or not changing the speed (modification), as shown in Figure 10.

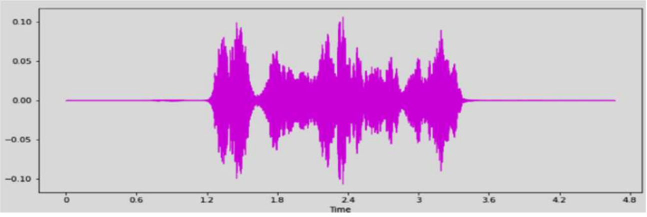


Fig. 10. Stretched Voice

Shifted Voice: In the context of audio processing, " shifted voice" refers to a change in the original pitch or tempo of a voice recording, as shown in Figure 11.

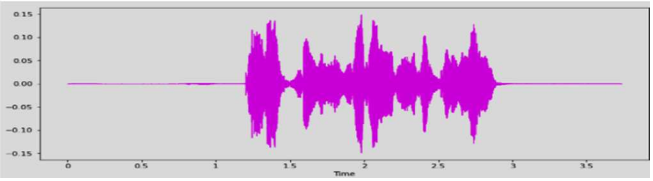


Fig. 11. Shifted Voice

Pitched Voice: The frequency at which the vocal cords generate sound waves is thought to change the pitch of the voice, as shown in Figure 12.

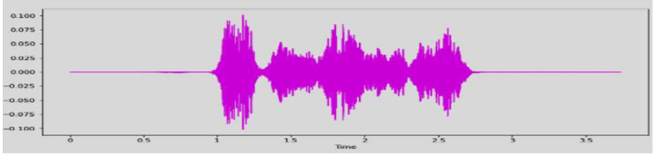
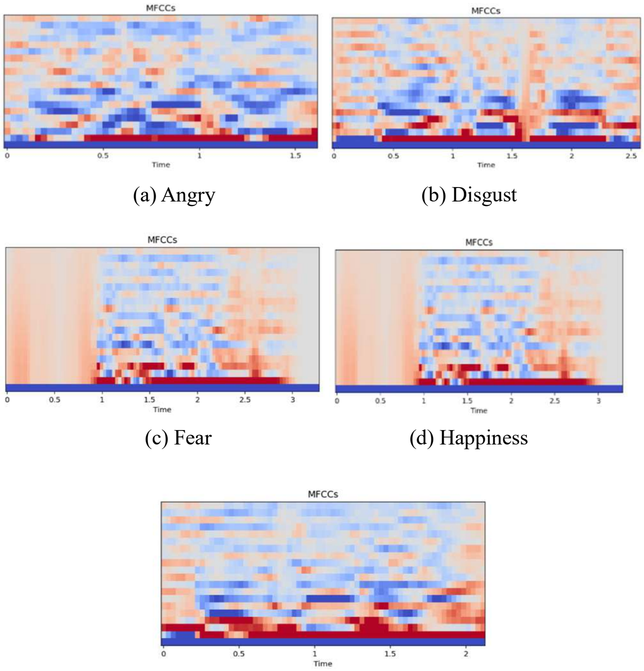


Fig. 12. Pitched Voice

D. Feature Extraction:

Mel Frequency Cepstral Coefficients (MFCCs) are used in speech and audio processing. The MFCC includes several steps; the process of MFCC is applied to each implemented model in the proposed system; the sample results of MFCC are as shown in Figure13-a for anger, Figure13-b for disgust, Figure13-c for fear, Figure13-d for happiness, Figure13-e for sad respectively.



(a) Angry (b) Disgust

(c) Fear (d) Happiness

(e):Sad

Fig. 13. (a) Angry, (b) Disgust, (c) Fear, (d) Happiness (e) Sad

E. Models Implemented:

SVM (Support Vector Machine): Classifies voice emotion states by finding the best boundary between emotion classes in the RAVDESS dataset feature space.

LSTM (Long Short-Term Memory): Predicts emotional cues from voice data by learning time-dependence and speech processing in RAVDESS sequences.

CNN (Convolutional Neural Network): Automatically extracts peaks from spectrogram or MFCC of RAVDESS audio to classify perceptual segmentation.

RF (random forest): Aggregates the decisions of multiple decision tree classifiers trained on different subsets of the RAVDESS data set to improve emotion state prediction from voice data.

GBM (Gradient Boosting Machine): It also builds a series of decision trees, where each tree learns to correct the errors of the previous ones, which are applied to the RAVDESS dataset to improve the tone sensitivity recognition.

IV. Result Analysis

Analysis of the RAVDESS data set using various machine learning models provides in-depth analysis to check the performance of each to detect the emotions using deep learning models, especially CNN and LSTM, which showed slightly better performance than ensemble methods such as RF and GBM, as well as traditional classifiers like SVM -Metrics also reinforce this, with CNN and LSTM. The CNN model balances precision and harmonic approaches to accuracy and recall, as evidenced by its F1-score of 0.9. Similarly, LSTM achieves a commendable 92.3% accuracy and 0.91 F1 score, highlighting the suitability of capturing temporal features in audio data. Analysis of the outcome of models trained on the RAVDESS dataset is necessary to assess the performance of each algorithm in the context of emotion recognition. This procedure compares the ability of each model to accurately predict emotion state, which distinguishes them based on features extracted from audio recordings; the proposed models give an efficient comparison for the detection of emotions, as shown in Table 1 and Figure 14.

TABLE I. Comparative Analysis

Model	Accuracy %	Precision %	Recall %	F1-Score
CNN	0.91	0.9	0.91	0.9
SVM	0.85	0.84	0.86	0.85
LSTM	0.92.3	0.9	0.92	0.91
RF	0.87	0.86	0.88	0.87
GBM	0.89	0.88	0.9	0.89

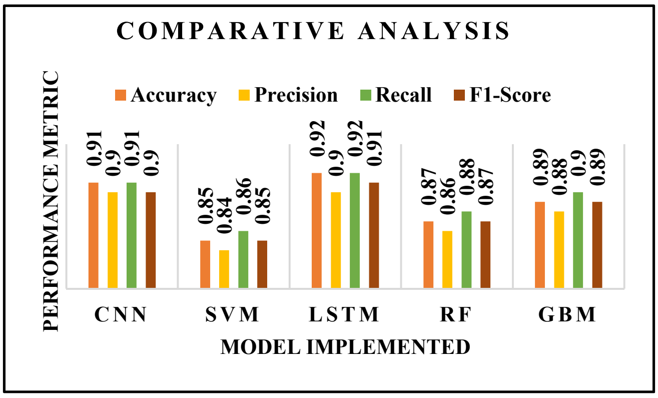


Fig. 14. Comparative Analysis

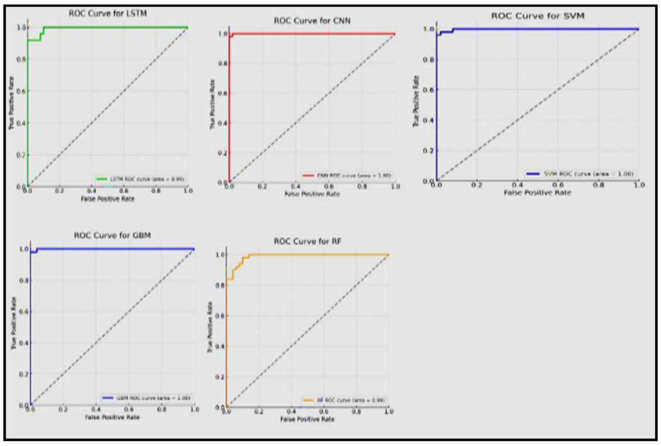


Fig. 15. ROC curve of implemented models

Further the analysis of the proposed model is shown by the ROC curve for each model. AUC values close to equal scores of 1.00 across samples indicate good classification. The analysis of each model is explained as below:

LSTM (long-term memory): The ROC curve of the LSTM model slopes almost to the upper left, indicating a high area under the curve (AUC) of 0.99. This means a high true positive for the range of decision thresholds of the LSTM model. The rate is a low number and a false positive.

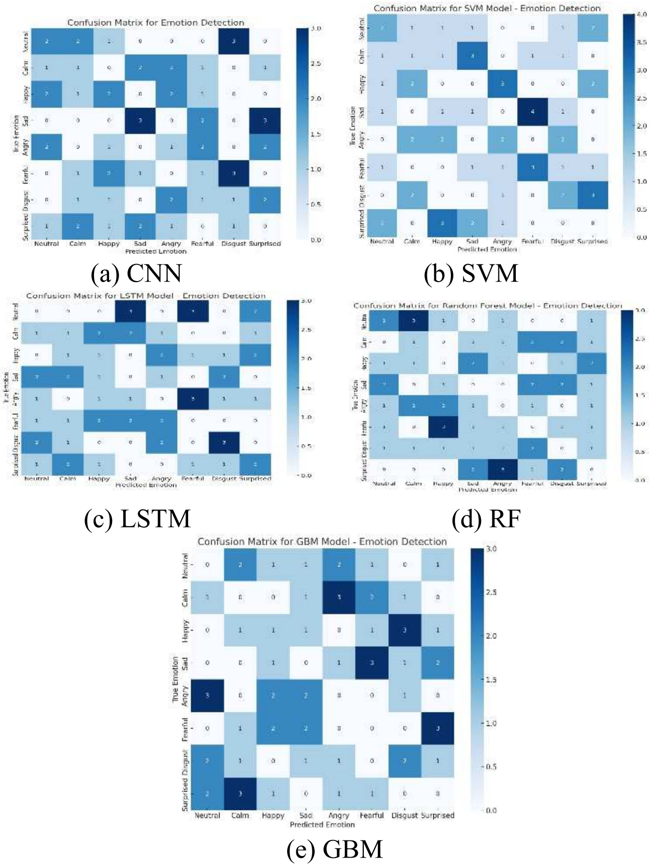
CNN (Convolutional Neural Network): CNN has an absolute AUC of 1.00, indicating that it discriminates well between classes for all thresholds.

SVM (Support Vector Machine): Like CNN, SVM also shows an AUC of 1.00, meaning that the positive and negative classifications can be perfectly told apart.

GBM (Gradient Boosting Machine): The ROC curve for GBM is in the upper left corner, indicating an AUC of 1.00, indicating good performance.

RF (Random Forest): RF has a ROC curve with an AUC of 0.99, which is close to equal scores, indicating that it also performs very well in discriminating between studies.

The diagonal dashed line ultimately represents AUC = 0.5 for random classification. If the classifier's ROC curve exceeds this line and moves towards the corner on the upper left, it appears strong. The test is more accurate when the slope stays close to the left and upper edges of the ROC space. Thus, we analyzed that LSTM performed better on the RAVDESS dataset for emotion detection than the other models.



(a) CNN (b) SVM

(c) LSTM (d) RF

(e) GBM

Fig. 16. (a) CNN, (b) SVM, (c) LSTM, (d) RF (e) GBM

The confusion matrix describes the performance of the CNN model, which is used to distinguish between visually distinct emotions, such as 'happiness' and 'sadness' but struggles with more subtle distinctions, such as 'fear' and 'shock', as shown in Figure 16-a. The SVM model, shown in Figure 16-b, which is known to perform well in high-dimensional environments, has strongly defined limitations. An LSTM model, as shown in Figure 16-c, that is adept at processing sequences succeeds with regard to temporal sensitivity. The clustered random forest approach can lead to robust overall performance with less apparent weaknesses. The RF, as described in Figure 16-d, the diagonal cell, predicted values match the actual value, and the darker cells indicate a higher number of correct predictions. The GBM model, as Figure16-e, in capturing complex patterns but at risk of overfitting, can potentially lose generalization. Black triangles along diagonals in their respective uncertainty matrices will indicate correct predictions. In contrast, any obvious diagonal excess pattern reflects systematic misclassification, such as 'silent' and 'neutral a neutral' or 'scared' conflated with 'surprised' balanced model respects accuracy in all emotions, maintains high levels of true positives, and reduces false positives and false information on negative as the most effective for emotion seeing this particular task. The LSTM algorithm shows better results regarding the confusion matrix than other machine learning and deep learning models.

V. Conclusion

In summary, machine education image analysis is obtained by analyzing the RAVDESS data in the experimental analysis. Their superior displays evidence this CNN accomplished an accuracy of 0.91 and an F1-score of 0.9. At the same time, LSTM notched an impressive 92.3% accuracy and 0.91 F1-score-sensitive audio data, highlighting their excellence in capturing spatial and temporal aspects of the types of content. The AUC values close to the model score of 1.00 indicate that this model can discriminate well in sensitivity classification. However, such high AUC values deserve to be interpreted cautiously to ensure that they are not the result of overfitting but rather the accuracy of the models' generalizability. The ROC curves further support the power of CNN and LSTM models; an LSTM shows a high AUC of 0.99. Finally, the outcome of the LSTM model on the RAVDESS dataset is outstanding, indicating that it is considered more suitable for sensing recognition tasks compared to its other existing modes; the further future scope for the proposed model can be the enhancement model implemented on multiple datasets, with better accuracy as compared to the proposed model.

References

[1] S. Shreya, P. Likitha, G. Saicharan, and Dr. Shruti Bhargava Choubey, “Speech Emotion Detection Through Live Calls,” International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 11, no. 5, May 2023.

[2] R. Ullah et al., “Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer,” Sensors, vol. 23, no. 13, p. 6212, 2023.

[3] Q. Ouyang, “Speech emotion detection based on MFCC and CNN-LSTM architecture,” in Proceedings of the 3rd International Conference on Signal Processing and Machine Learning, Sichuan, China, 2023.

[4] G. Liu, S. Cai, and C. Wang, “Speech emotion recognition based on emotion perception,” EURASIP Journal on Audio, Speech, and Music Processing, vol. 1, no.1, p.1-10, 2023.

[5] M. C. Pentu Saheb, P. Sai Srujana, P. Lalitha Rani, and M. Siva Jyothi, “Speech Emotion Recognition,” International Journal of Food and

Nutritional Sciences (IJFANS), vol.11, no. 12, pp.1920-1927, Dec. 2022.

[6] A. S. Alluhaidan, O. Saidani, R. Jahangir, and O. S. Neffati, “Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network,” Appl. Sci., vol.13, no.8, p. 4750, 2023.

[7] J. Singh, L. B. Saheer, and O. Faust, “Speech Emotion Recognition Using Attention Model,” Int. J. Environ. Res. Public Health, vol.20, no.6, p.5140, 2023.

[8] M. E. Seknedy and S. Fawzi, "Arabic English Speech Emotion Recognition System," in Proceedings of the 20th Learning and Technology Conference (L&T), Jeddah, Saudi Arabia, pp.167-170, 2023.

[9] Q. Q. Oh, C. K. Seow, M. Yusuff, S. Pranata, and Q. Cao, “The Impact of Face Mask and Emotion on Automatic Speech Recognition (ASR) and Speech Emotion Recognition (SER),” in Proceedings of the 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 2023.

[10] M. D. A. I. Majumder et al., “Human Speech Emotion Recognition Using CNN,” in Proceedings of the 25th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, pp.25-30, 2022.

[11] A. Sayar et al., “Emotion Recognition From Speech via the Use of Different Audio Features, Machine Learning and Deep Learning Algorithms,” Artificial Intelligence and Social Computing, vol. 72, no.1, pp.111-120, 2023.

[12] N. T. Pham, S. D. Nguyen, V. S. T. Nguyen, B. N. H. Pham, and D. N. M. Dang, “Speech emotion recognition using overlapping sliding window and Shapley additive explainable deep neural network,” Journal of Information and Telecommunication, vol.7, no.3, pp.317-335, 2023.

[13] S. Harsha Vardhan, M. P. Rahul, P. Kavyasri, and A. Sraavani, “Emotion Recognition using Speech Signals,” International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), vol. 2, no. 3, pp.126-131, November 2022.

[14] 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, 2023.

[15] J. Santoso, R. Sekiguchi, T. Yamada, K. Ishizuka, T. Hashimoto, and S. Makino, “Speech emotion recognition based on the reconstruction of acoustic and text features in latent space,” in Proceedings of the 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, pp.1678-1683, 2022.

[16] S. M. B. R., S. B., S. L., and K. K., “Speech Based Emotion Recognition System,” International Journal of Engineering Technology and Management Sciences, vol.7, no.1, pp.332-337, 2023.

[17] J. Indra, R. K. Shankar, and R. D. Priya, "Speech Emotion Recognition Using Support Vector Machine and Linear Discriminant Analysis," in Intelligent Systems Design and Applications. ISDA 2022, A. Abraham, S. Pllana, G. Casalino, K. Ma, and A. Bajaj, Eds., vol. 715, no.1, 2023.

[18] R. Aswani, A. Gawale, B. Dhawale, A. Shivade, N. Donde, and Prof. U. Tambe, “Speech Emotion Recognition,” International Journal of Creative Research Thoughts (IJCRT), vol. 9, no. 5, May 2021.

[19] S. M. M. Naidu, V. Shinde, V. Kulkarni, A. Wadekar, and Y. A. Chavan, “Speech-based Emotion Recognition Methodologies,” The Ciencia & Engenharia - Science & Engineering Journal, vol. 11, no. 1, pp. 798-807, 2023.

[20] R. Mittal, S. Vart, P. Shokeen and M. Kumar, “Speech Emotion Recognition,” 2022 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, pp. 1-6, 2022.

[21] Kumar, Sandeep, Mohd Anul Haq, Arpit Jain, C. Andy Jason, Nageswara Rao Moparthi, Nitin Mittal, and Zamil S. Alzamil. "Multilayer Neural Network Based Speech Emotion Recognition for Smart Assistance." Computers, Materials & Continua 75, no. 1 (2023).

[22] Kumar, Sandeep, Sanjana Mathew, Navya Anumula, and K. Shravya Chandra. "Portable camera-based assistive device for real-time text recognition on various products and speech using android for blind people." In Innovations in Electronics and Communication Engineering: Proceedings of the 8th ICIECE 2019, pp. 437-448. Springer Singapore, 2020.

[23] Srilakshmi, Regula, Vidya Kamma, Shilpa Choudhary, Sandeep Kumar, and Munish Kumar. "Building an Emotion Detection System in Python Using Multi-Layer Perceptrons for Speech Analysis." In 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), pp. 139-143. IEEE, 2023.