DHERF: A Deep Learning Ensemble Feature Extraction Framework for Emotion Recognition Using Enhanced-CNN

Shaik Abdul Khalandar Basha and P. M. Durai Raj Vincent *

School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, India
Email: khalandar.basha2016@vitstudent.ac.in (S.A.K.B.); pmvincent@vit.ac.in (P.M.D.R.V.)
*Corresponding author

***Abstract*—Artificial Intelligence (AI) based solutions are inevitable for real-time issues in any field where voluminous historical data is to be analyzed for accurate prediction analysis. Voice-operated smart AI devices like Alexa, Siri, etc., are a commercial success which are now part of most smart households. Voice-based acoustic datasets can also be leveraged to function like biomarkers in identifying the emotion of the speech signal. Existing deep learning models using Convolutional Neural Networks (CNN) have already been employed for emotion detection, but mediocre performance was reported when prediction was extracted from multimedia content analysis. To enhance the performance of CNN-based deep learning algorithms on multi-media content-based datasets, a novel configuration framework known as the Deep Human Emotion Recognition Framework (DHERF) has been proposed in this work. DHERF exploits multiple selective features from the training dataset with a learning-based phenomenon for enhancing prediction accuracy. The experimental study revealed that optimized feature selection in training the DHERF model resulted in better prediction performances of up to 85.70% accuracy as compared to conventional CNN baseline and Long Short-Term Memory (LSTM) models which attained a maximum prediction accuracy of 71.64% and 81.11% respectively, for the same experimental conditions.**

***Keywords*—deep learning, human emotion recognition, Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), baseline CNN, audio-based human emotion recognition**

I. Introduction

Human beings express different kinds of emotions from time to time based on certain incidents and situations. Such emotions provide sufficient cues to ascertain their state of mind and behavior. The speech of humans is one of the effective means of communication. If their speech can be automatically processed and the kind of emotion it carries be detected through technology-based AI applications, it can provide break through results [1]. With the emergence of Artificial Intelligence (AI) enabled approaches, deep learning has attracted researchers due to its capability to learn from audio samples. Many deep learning methods are found to be effective as investigated in research [2−4]. Sezgin *et al.* [4] proposed a hybrid deep learning model for the classification of human emotions that come from body sensors in the healthcare domain and verified that neural network-based deep learning methods are more effective in emotion recognition. Muhammad [5] proposed an audio-visual approach using CNNs which includes a two-phase fusion of features along with a big data concept toward automatic detection of human emotions. Lalitha *et al.* [3] explored the perpetual features of audio signals towards building a system for enhancing the emotion detection process. They exploited a multi-feature approach along with DNN towards emotion classification.

Some researchers contributed towards the detection of mood that could lead to finding mental states and possible depression in humans. Lalitha *et al.* [3] explored deep learning architectures for automatic mood detection of mood from audio and lyrics. They intend to improve it in the future with new feature extraction combined with unsupervised learning towards finding the depression mood of humans. Ma *et al.* [6] focused on mood detection and tracking to identify possible moods such as depression leading to making decisions. There are audio-visual approaches also found in the literature for comprehensive emotion detection as explored by Datta *et al* [8]. From the review of the literature, it is evident that pure Convolutional Neural Network (CNN) is best suited deep learning model for purely voice based samples, but fails to deliver satisfactory emotion prediction when extracted from multimedia content analysis. In this work we propose a novel selective feature based framework, which enhances the prediction accuracy of CNN based learning models on multimedia Content

Our contributions are as follows.

1. We propose a framework known as Deep Human Emotion Recognition Framework (DHERF) which exploits the CNN-based prediction model.
2. We proposed an algorithm named Deep Learning-based Human Emotion Recognition (DL-HER) where advanced CNN configurations are exploited for leveraging prediction performance.
3. We implemented the DHERF framework to evaluate the proposed DHLER algorithm and

Manuscript received January 3, 2024; revised January 17, 2024; accepted February 21, 2024; published July 18, 2024.

verified its enhanced prediction accuracy over conventional CNN baseline and Long Short-Term Memory (LSTM) models.

The remainder of the paper is structured as follows. Section II reviews different deep learning models employed in emotion recognition. Section III presents the methodology proposed for effective human emotion detection. Section IV presents the results of our empirical study while Section V concludes our work on CNN based model for emotion recognition besides bestowing future scope.

II. Related Work

This section reviews the literature on existing methods used for emotion recognition using deep learning. Pyrovolakis *et al.* [8] proposed a methodology that could detect human emotions live through the Internet of Things (IoT) healthcare use-case using the 2D-CNN technique. They also explored it further to recognize human sentiments. Sezgin *et al.* [4] proposed a method that could extract a novel set of features from audio, such as TEO (Teager Energy Operator)-based, spectral, qualitative, and continuous. It was a frame-based feature extraction approach for better performance in the recognition of human emotions. However, they intended to improve their work by reducing computational complexity. Patel *et al.* [9] investigated compact representation of audio with the help of a deep autoencoder and found that such representation helps improve recognition performance. They also found the need for using improved autoencoders for further enhancement in recognition capability. Bertero and Fung [10] studied the utility of CNN for emotion recognition by developing a human-machine interactive system. Pyrovolakis *et al.* [8] explored deep learning architectures for automatic mood detection from audio and lyrics. Further enhancement was possible by new selective feature extraction combined with unsupervised learning [11, 12].

Kanjo *et al.* [2] proposed a hybrid deep learning model for the classification of human emotions that come from body sensors in the healthcare domain. They found that neural network-based deep learning methods are more effective in emotion recognition. Their methodology includes a feature fusion approach that could improve performance. They intend to explore different kinds of data such as EEG (electroencephalogram) signals that are sensor-driven in the future. Kao *et al.* [2] investigated different emotion detection approaches that are based on text. It is a learning-based hybrid approach to emotion detection. Zamil *et al.* [13] proposed a hybrid deep-learning model for emotion detection. It combines both discriminative and generative models that resulted in a stronger ability to classify speech samples [14, 15]. Hossain and Zamil *et al* [13] proposed an audio-visual approach using CNNs which includes a two-phase fusion of features along with a big data concept towards automatic detection of human emotions. Their frame work is also compatible with an edge-cloud-based approach in They found different kinds of features a multi-model approach using CNNs for emotion detection. Their work is towards screening of depression cases in the healthcare domain.

Naveenkumar and Kaliappan [16] explored features such as the Cepstral coefficient and Mel frequency for making a tool for emotion recognition. Besides, they investigated different AI-enabled methods explored in similar research. Zamil *et al.* [14] explored sound features to classify frames and perform voting ensembles on them toward an effective emotion detection mechanism. Kimmatkar and Babu [13] proposed a methodology for not only emotion recognition but also considering mental health screening in the healthcare domain in assisting physicians in possible diagnosis. It is a CNN-based approach that works in frequency and time domains. Sailunaz *et al.* [17] studied different text-based and audio-based emotion detection methods. It includes lexicon-based methods for machine learning, and deep learning. Kansizoglou *et al.* [18] defined an audio-visual approach using CNN and reinforcement learning towards emotion recognition. It is an active learning-based approach for automatic emotion identification. They found that the emotional state of humans helps in solving certain real-world problems. Batziou *et al.* [19] focused on the audio-visual method with deep learning approaches for emotion detection and even finding the impact of emotions as well.

Lalitha *et al.* [3] explored the perpetual features of audio signals towards building a system for enhancing the emotion detection process. They exploited a multi-feature approach along with DNN towards emotion classification. They intend to improve their method to deal with imbalanced datasets in the future. Hazarika *et al.* [20] defined a method with feature fusion and also self-attention mechanism along with deep learning to ascertain human emotions extension of their work was implementation on multi-model systems in the future. Lu *et al.* [20] focused on mood detection and tracking to identify possible moods such as depression leading to making decisions, impact of music on work pressure was further to be analyzed. Popovic *et al.* [21] considered multi-party conversations to detect emotions. Their methodology includes vision-based models, speech-based models, and Natural Language Processing (NLP) models towards a multi-model emotion recognition system. Nandwani and Verma [21] proposed a system with a web-based interface to detect emotions from text and find sentiments from the same. Rachman *et al.* [21] explored lyrical and audio features from audio to recognize human emotions. Mande and Dani [22] proposed a framework with multiple models in the pipeline toward effective recognition of emotions. Zhang *et al.* [22] exploited a hybrid deep learning-based model for emotion detection. It also includes the notion of learning effective features [23, 24]

Ogihara *et al.* [22] explored a content-based approach for automatically detecting emotions. It exploited an improved similarity search mechanism. Other important contributions include emotion recognition for social robotics [25], deep learning fusion [26], cross-toward approach [13], DNN [13], supervised ML [13] rule-based approach [13], and EEG-based hybrid deep learning [27].

From the review of the literature, it is understood that a CNN equipped with novel configurations improves prediction accuracy of the deep learning model for multimedia content analysis [28, 29].

Therefore, we propose a novel selective feature based DHERF framework and evaluated its performance.

III. Materials and Methods

This section presents materials and methods in terms of dataset details, the proposed framework, algorithms, and evaluation methodology.

A Dataset Details

RAVDESS is a widely used dataset [15, 30, 31] with 1440 audio samples from 24 professional actors of both genders covering 7 emotions such as "calm, happy, sad, angry, fearful, surprise, and disgust" [21, 32]. Each emotion has two levels of intensity known as normal and strong [33].

As presented in Table I, there are 8 classes of emotions including neutral which does not reflect any emotion. In fact, neutral does mean the absence of emotion. Therefore, there is no strong intensity level associated with this class.

TABLE I. Emotion Classes and Their Intensity in the RAVDESS Dataset

Class	Emotion	Intensity Levels
1	neutral	normal
2	calm	normal, strong
3	happy	normal, strong
4	sand	normal, strong
5	angry	normal, strong
6	fearful	normal, strong
7	disgust	normal, strong
8	surprise	normal, strong

B. The Framework

We proposed a framework for the automatic detection of human emotions from given audio samples. The framework takes the RAVDESS dataset as input for both training and test audio samples. Exploratory data analysis is made on the given dataset in order to understand details of the dataset and ascertain more insights into it. This will enable us to effectively sort, data imbalance and check for data augmentation. We followed a multi-feature approach to bring effectiveness to the learning process and eventually, appropriate feature selection is made to improve the quality of training for the proposed enhanced CNN model. The different kinds of features include MFCC, spectral, chromatogram, spectral bandwidth, and Mel.

Audio-related datasets categorize a few sounds like speech recognition, environmental and sound classification. Image analysis is based on classification of objects [dog or cat], while text classification checks detecting ham and span. Extraction of features in text, and images is comparatively easier than feature extraction from audio data. Classification techniques for images dataset using deep learning extracts features in numerical format with the help of already stored pixels, extraction from text dataset requires a decoder and sequential encoder. However, it is quite difficult to extract features of audio data because of time and frequency dependency, from which frequency and pitch need to be extracted properly.

Before preprocessing, were visualized hidden patterns from the audio files by using a Python library called librosa. To perform EDA, the audio dataset contains two patterns: a dimensional array and a sample rate. EDA helps to know the depth of the dataset with sample rate and 2D array. EDA with the help of librosa is much more popular specifically for signal processing due to 3 strong reasons.

- EDA helps to interset the wave signal into one channel (Mono).
- Data normalization is represented in the range of −1 to +1 in audio signals.
- Able to see default sample rate (22 kHz).

As presented in Fig. 1, the proposed enhanced CNN model (as discussed later in this section) is trained with the selected features to gain emotion detection knowledge. This knowledge model is used for the detection of emotions in the given test samples. The proposed model is evaluated, and emotion recognition is facilitated with user-friendly client application. Once the model is trained with training samples, the client application reuses the saved model to predict emotions in the given sample.

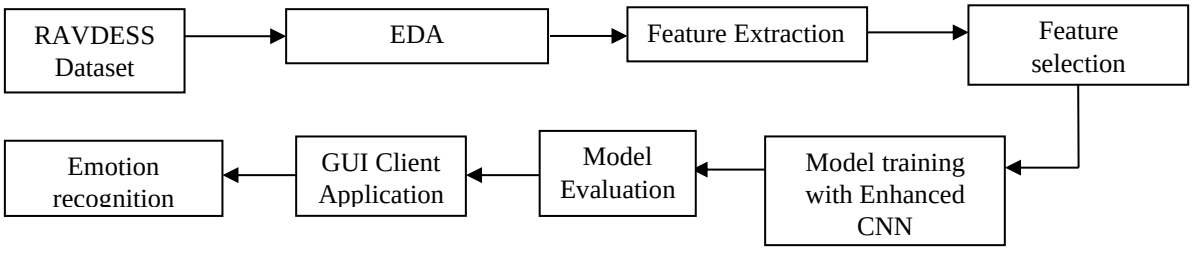

Fig. 1. Overview of the proposed deep human emotion recognition framework.

As presented in Fig. 2, after much empirical study and investigation, the enhanced CNN model is designed to leverage emotion recognition performance. It is made up of customized layers toward effectiveness in emotion identification more accurately. It has several trainable parameters. The configuration of layers has the potential to improve detection performance.

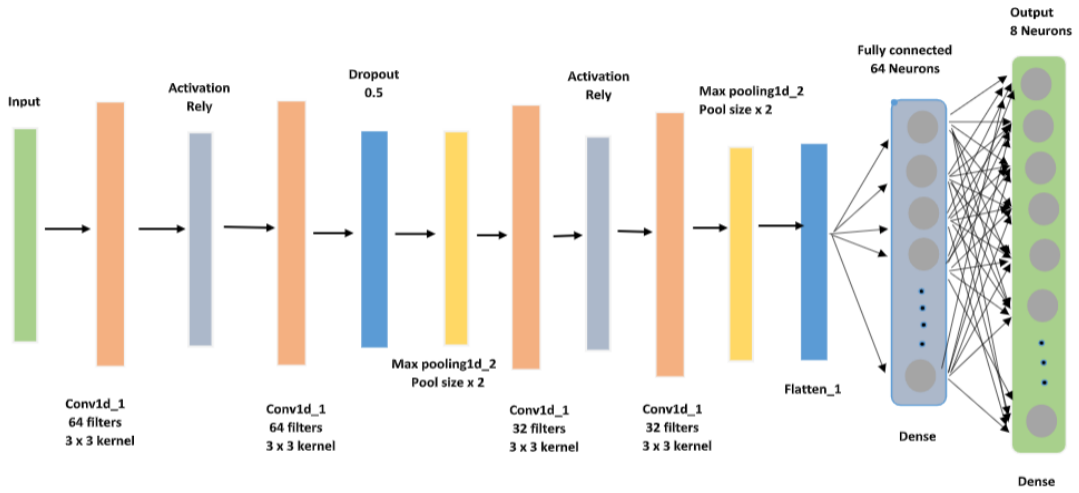

Fig. 2. Enhanced CNN model used for emotion recognition from audio.

As presented in Table II, details of layer type, output shape, and number of parameters used are provided. This deep learning model is designed based on CNN to overcome the limitations of the baseline CNN model with optimal configurations and usage of weights appropriately.

TABLE II. LAYERS AND PARAMETERS INVOLVED IN THE PROPOSED ENHANCED CNN MODEL

Layers (Type)	Output Shape
Conv1D (Conv1D)	(None. 188, 64)
Activation (Activation)	(None. 188, 64)
Conv1D_1 (Conv1D)	(None. 186, 64)
Dropout (Dropout)	(None. 186, 64)
Max_pooling1D (Maxpooling1D)	(None. 93, 64)
Conv1D_2 (Conv1D)	(None. 91, 32)
Activation_1 (Activation)	(None. 91, 32)
Conv1D_3 (Conv1D)	(None. 89, 32)
Max_pooling1D_1 (Maxpooling1D)	(None. 44, 32)
Flatten (Flatten)	(None. 1408)
Dense (Dense)	(None. 64)
Dropout_1 (Dropout)	(None. 64)
Dense_1 (Dense)	(None. 32)
Dense_2 (Dense)	(None. 10)
Total Params: 114.474 Trainable Params: 114,474 Non-Trainable Params: 0	

C. Ensemble Feature Extraction

1) Feature extraction

The most commonly used feature extraction is Mel Frequency Cepstral Coefficients (MFCC). MFCC is used to identify Mono syllables in sound datasets without classification of the speaker. In MFCC feature extraction, the initial process amplifies the sound signal's outrageous frequency, followed by windowing and framing phases. To restrict the influence of disturbances at the beginning and end of audio data windowing method is applied and the framing phases will break the audio data into multiple time slots in the range of 25−35 ms.

In speech recognition to achieve the acoustic properties of a speaker MFCC technique is applied. In the next phase as shown in Fig. 3 the combination of Ensemble feature extraction using MFCC, Chroma, and Mel spectrogram is shown. Post windowing, discrete cosine transform, Fast Fourier Transform and Mel flitter bank techniques are applied on the dataset. To extract auditory perception of frequency, Mel-Scale method is used through MFCC. MFCC features are extracted by considering the logarithm of the power spectrum which is converted to cepstral, it helps in reducing processing complexity and feature dimensionality. MFCC can express temporal information in sound data between short durations by splitting the recording variation in sound data.

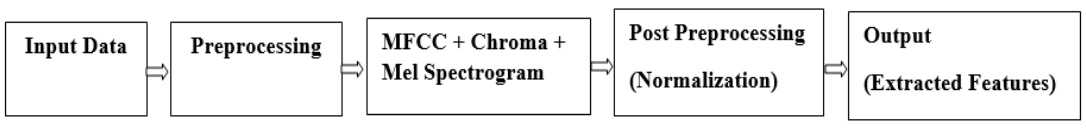

Fig. 3. Ensemble feature extraction pattern using MFCC, Chroma, and Mel spectrogram.

2) Chroma feature

In sound singles, the expression of the chromatogram is sharply relayed to 8 different pitch labels. Extraction of features using Chroma is mentioned as 8 pitch class classification. The Chroma technique is used to retrieve melodic and harmonic characteristics of sound singles. Harmonic features like keys and chords are extracted using chromo features for a short time window. Other features to extract magnitude are STFT, CENS, and CQT.

3) Mel spectrogram

In the third phase of the feature extraction mel spectrogram is applied upon MFCC and chroma. To extract features from signals Mel-spectrogram is used. Speech emotion recognition deals with low-frequency bands with preprocessing of audio data. Mal spectrogram

combined with CNN (Neural Nets) is used to carry forward investigation in detail. Mal-Spectrogram consists of a short-term Fourier transform from the signal window of the spectrum i.e. amplitude or energy from linear range to logarithmic scale. Mel-Scale and eigenvector is calculated through filter bank with N_meb = 128, N_fft = 1024 and Hot length = 512.

D. Proposed Algorithm

We proposed an algorithm known as Deep Learning-based Human Emotion Recognition (DL-HER) based on the architecture of enhanced CNN presented in Fig. 3. This algorithm is a novel contribution that exploits customized configurations of different layers keeping the emotion recognition problem in hand from given audio samples.

As presented in Algorithm 1, Consider inputs such as RAVDESS dataset D, number of epochs n, and batch size m. It has multiple convolutional and max-pooling layers besides other layers for optimal feature selection and learning from features. It performs multi-class classification of the given test samples. Imbalance audio dataset means fabricating conventional audio dataset using an argumentation framework specific to explore time stretching, pitch shifting and time shifting. A method called audio data augmenter is applied to perform an amalgam of deterministically or augmentation probabilistically in parallel or in series. The proposed Framework is divided into 3 parts feature selection, feature extraction, and classification of emotion. Feature extraction entirely depends upon multi-modal recognition. Feature selection focuses on input model information converted into features. Multi-emotion occupies different modalities and weights in classification.

Algorithm 1: Deep Learning based Human Emotion Recognition (DL-HER)

Inputs:
RAVDESS dataset *D*
number of epochs *n*
batch size *m*

Output:
Anomaly detection results *P*

1. Begin
2. D' PreProcess(D)
3. (T1, T2) SplitData(D')

Enhanced CNN Model Configuration

4. Create model m
5. Add Conv1D layer to m
6. Add activation ReLU layer to m
7. Add Conv1D layer to m
8. Add Dropout layer to m
9. Add MaxPool1D layer to m
10. Add Conv1D layer to m
11. Add ReLU activation layer to m
12. Add Conv1D layer to m
13. Add MaxPool1D layer to m
14. Add Flatten layer to m
15. Add Dense layer to m
16. Add Dropout layer to m
17. Add Dense layer to m (with ReLU activation)
18. Add Dense layer to m (softmax)
19. *m'* = TrainModel(*m, T1*)
20. **For each** Epoch *e* in *n*
21. **For each** batch *b* in *m*
22. Update m'
23. **End For**
24. **End For**

Fitting the Model

25. *m'* = FitModel(*m',T2*)

Prediction and Evaluation

26. *P* = PredictAnomalies(*m', T2*)
27. *R* = ModelEvaluation()
28. **Print** *R*

Return *P*

E. Evaluation Methodology

The proposed deep learning model is evaluated for its performance using a confusion matrix. Fig. 4 shows the confusion matrix which helps in deriving values for false negative, true positive, false positive, and true negative. These are used to know the performance of the given model in terms of precision, accuracy, F1-Score and recall.

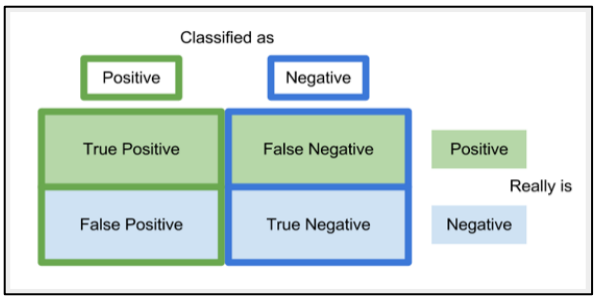

Fig. 4. Confusion matrix.

Computation of the performance metrics is based on correct and wrong predictions of an ML model. Precision and recall are computed as in Eqs. (1) and (2).

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

Other metrics such as F1-Score and accuracy are computed in Eqs. (3) and (4).

$$F1\text{-}Score = 2 \times \frac{p \times r}{p+r} \quad (3)$$

In F1-Score computation, p denotes precision while r denotes recall values.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

All these metrics when evaluated result in a range of values from 0 to 1 reflecting the least and highest performance.

IV. Results and Discussion

This section presents experimental results. Our empirical study is made with an implementation of the

framework and the underlying algorithm. The results are provided in terms of exploratory data insights, detection of human emotions, and performance evaluation. Splitting of train and test data is considered to be a logical process because it identifies independent and dependent features. Our dataset consists of 1440 audio files with 8 classes. We employed the random split method to achieve division between test and train datasets and a stratified split of Train 0.7, Test 0.2, and cross-validation 0.1 is achieved.

A. Exploratory Data Analysis

This sub-section presents the dataset insights in order to understand the data distribution dynamics and features.

As presented in Fig. 5, emotions from the dataset are identified and emotion-wise data distribution dynamics is provided.

As presented in Fig. 5, different kinds of features are extracted from audio samples to realize a multi-feature-based model.

Fig. 6 shows a neural audio sample that has no emotions is visualized graphically.

In Fig. 7, a graphical representation of a given audio sample is provided to reflect the duration dynamics of emotion.

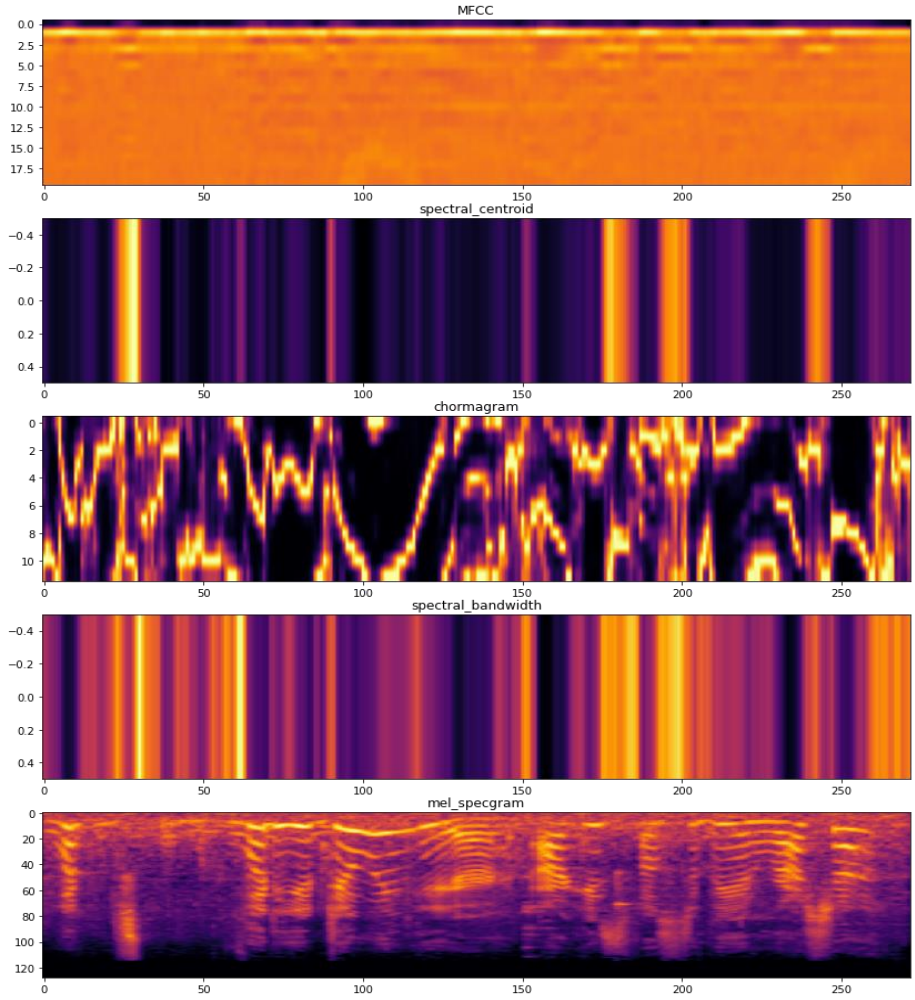

Fig. 5. Extraction of features of different kinds from audio.

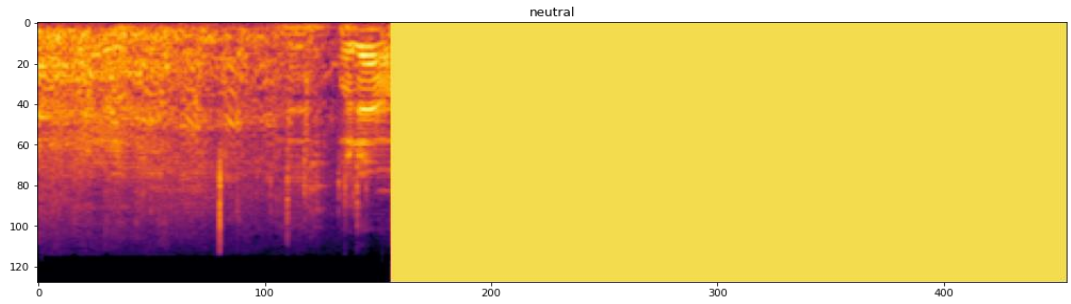

Fig. 6. An audio sample with neutral emotion.

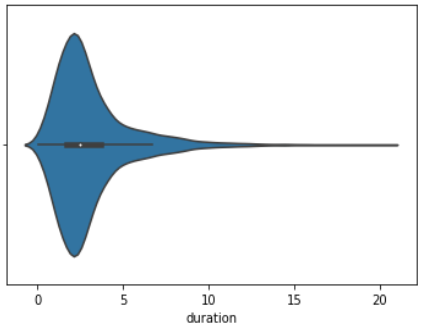

Fig. 7. Emotion pitch dynamics of an audio sample.

B. Results and Evaluation

Performance of the proposed framework and its comparison with existing models is shown in Fig. 8. It is observed that the proposed system has a GUI client to take a new test audio sample as input and discover emotions besides providing a final prediction along with its probability. This client application exploits our saved enhanced CNN model after the due training process. In Table III, provides a qualitative comparison of selective features used in deep learning existing CNN models and proposed CNN model with DHERF framework configuration.

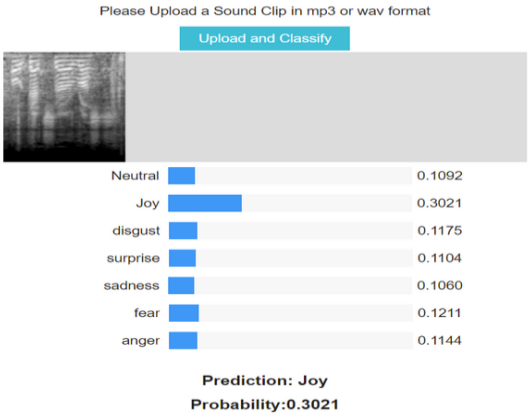

Fig. 8. Application exploiting trained deep learning model for emotion recognition from given audio sample.

TABLE III. COMPARATIVE TABLE OF PROPOSED RESEARCH AND THE EXISTING WORK

Authors	Feature section techniques (fst)	Audio features	Emotion Labels classes	ML and DL Algorithms	Evaluation metrics	Results
Dong *et al.* [34]	fst by Fisher Criterion and correlation analysis	spectrum features, prosodic and quality	sad, neutral, happy, fear, surprise, and angry	NN-BP, SVM, KNN	Average Recognition Rate	SVM:79%; NN-BP:80.1%
Sng *et al.* [35]	Extraction of Framing with 20 Ms	shimmer, pitch, Spectral, intensity, and jitter	sadness, anger, neutral, fear, and happiness	VQ; K, ANN and GMM	Accuracy	ANN: 72%: GMM: 79%, VQ: K 57%
Ma *et al.* [6]	Fourier-transform-based filter bank	LLDs	Valence (Positive and negative)	PCASS, MT-; KMM	Emo-DBy, Accuracy	MT: 62.52%; KMM: 66.36%; SHLA: 63.75%
Li *et al.* [36]	Segmenting signals	MFCCs and log-energy	neutral state, anger, sadness, boredom, and fear	Baseline: LSTM, RNN, CNN	Accuracy, F1-Score and precision	LSTM: 81.11% RNN: 78.83%, CNN: 78.31%
Lu *et al.* [26]	Not mentioned	energy augmented and MFCCs	Disgust, Neutral, fearful, calm, angry, happy, and sad	Baselines-CNN	Accuracy	CNN: 71.64%
Proposed Enhanced CNN	MFCC, Chroma, Mel-spectrogram	windowing and framing	Neutral state, anger, sadness, boredom, and fear, Neutral, Disgust, Surprised, Fearful, happy, calm, and sad	Enhanced CNN	Accuracy	Enhanced CNN: 85.7%

As presented in Table IV, the performance of different existing models in emotion detection accuracy along with the proposed enhanced CNN model is provided. RAVDEES dataset has been considered for an experiment that includes 1,441 speech audio single data and 1,010 emotional audio single data stored in wave audio file format. The proposed Ensemble feature extraction techniques are MFCC, Chroma, and Mel Spectrogram. Totally 190 features were extracted among these 45 MFCC features, 128 features from the Spectrogram and 17 from the Chroma technique.

TABLE IV. PERFORMANCE OF DIFFERENT DEEP LEARNING MODELS

Model	Calm	Angry	Sad	Happy	Fearful	Surprised	Disgust	Neutral
CNN Baseline	84.51045	81.1739	60.08895	73.857	63.6869	68.91285	51.2349	65.13405
LSTM Model	71.355	88.44	81.405	70.35	89.445	81.405	73.365	65.325
Proposed Enhanced CNN	91.1736	84.3195	70.6515	71.466	87.3848	71.50575	78.70155	88.73145

Fig. 9 shows the emotion recognition performance for each kind of emotion. The proposed model is compared against the CNN baseline and LSTM models. With many kinds of emotions [37], the proposed model showed the highest performance. CNN algorithm is a very popular method for image classification, in a recent study CNN shows a significant impact on audio file classification. Generally, audio data first converts singles into

spectrogram (2D representation) and applies 2D CNN architecture with a typical approach called sector-temporal feature section. But in our study, we applied DGT log mel-spectrogram and MFCC parameter selection entail. Instead of considering an average of 1D and 2D CNN features alone, we adopted an enhanced CNN algorithm (a combination of DGT log Mel-Spectrogram with MFCC) to achieve high accuracy compared with conventional 2D CNN.

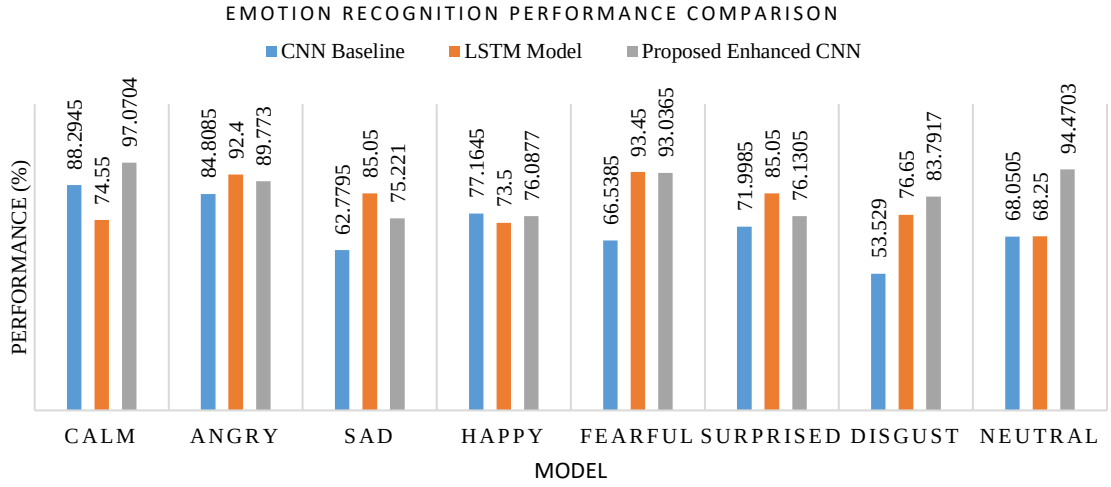

Fig. 9. Model accuracy comparison for each emotion.

As presented in Table V, the accuracy of different models in human emotion recognition is provided. The accuracy of the proposed enhanced CNN model is compared against existing models such as CNN baseline and LSTM. It is observed that the proposed model outperforms existing ones. The accuracy of the CNN baseline is 71.64% while LSTM showed 81.11% accuracy. The highest accuracy is exhibited by the proposed enhanced CNN model with 85.70%.

TABLE V. PERFORMANCE OF DIFFERENT DEEP LEARNING MODELS

Models	Accuracy (%)
CNN Baseline	71.64
LSTM Model	81.11
Proposed Enhanced CNN	85.7

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a framework known as Deep Human Emotion Recognition (DHERF) which exploits CNN-based prediction models. We proposed an algorithm named Deep Learning-based Human Emotion Recognition (DL-HER) where advanced CNN configurations are exploited for leveraging prediction performance. Our methodology exploits multiple features from training data with a learning-based phenomenon toward leveraging detection performance. The DL-HER algorithm is an important contribution to this work as it resulted from extensive research on multiple customized layers in the CNN model. Our experimental results showed better performance over existing models in terms of accuracy. The proposed enhanced CNN model achieves the highest accuracy with 85.7% accuracy while existing models such as CNN baseline and LSTM models showed 71.64% and 81.11%, respectively. From the results of our empirical study, it is understood that the proposed CNN-based enhanced model can be used in diversified applications including the healthcare domain. In the future, we intend to develop an integrated framework that exploits multiple best models with an ensemble approach for improving emotion recognition performance further.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Shaik Abdul Khalandar Basha conducted the research and wrote the paper; P. M. Durai Raj Vincent analyzed the data; all authors had approved the final version.

REFERENCES

[1] A. J. Datta, R. Taylor, G. Will, and G. Ledwich, "An investigation of earth grid performance using graphene-coated copper," *IEEE Access*, vol. 3, pp. 1042–1050, 2015. doi: 10.1109/ACCESS.2015.2454295

[2] E. Kanjo, E. M. G. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Inf. Fusion*, vol. 49, pp. 46–56, 2019. doi: 10.1016/j.inffus.2018.09.001

[3] S. Lalitha, S. Tripathi, and D. Gupta, "Enhanced speech emotion detection using deep neural networks," *Int. J. Speech Technol.*, vol. 22, no. 3, pp. 497–510, 2019. doi: 10.1007/s10772-018-09572-8

[4] M. C. Sezgin, B. Gunsel, and G. K. Kurt, "Perceptual audio features for emotion detection," *Eurasip J. Audio, Speech, Music Process.*, vol. 2012, no. 1, pp. 1–21, 2012. doi: 10.1186/1687-4722-2012-16

[5] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio—Visual emotional big data," *Inf. Fusion*, vol. 49, pp. 69–78, 2019. doi: 10.1016/j.inffus.2018.09.008

[6] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-Visual Emotion Fusion (AVEF): A deep efficient weighted approach," *Inf. Fusion*, vol. 46, pp. 184–192, 2019. doi: 10.1016/j.inffus.2018.06.003

[7] M. S. Hossain and G. Muhammad, "An audio-visual emotion recognition system using deep learning fusion for a cognitive

wireless framework," *IEEE Wirel. Commun.*, vol. 26, no. 3, pp. 62–68, 2019. doi: 10.1109/MWC.2019.1800419
[8] K. Pyrovolakis, P. Tzouveli, and G. Stamou, "Mood detection analyzing lyrics and audio signal based on deep learning architectures," in *Proc. Int. Conf. Pattern Recognit.*, 2020, pp. 9363–9370. doi: 10.1109/ICPR48806.2021.9412361
[9] N. Patel, S. Patel, and S. H. Mankad, "Impact of autoencoder based compact representation on emotion detection from audio," *J. Ambient Intell. Humaniz. Comput.*, vol. 13, no. 2, pp. 867–885, 2022. doi: 10.1007/s12652-021-02979-3
[10] D. Bertero and P. Fung, "A First look into a convolutional neural network for speech emotion detection," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process*, 2017, pp. 5115–5119.
[11] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," *Int. J. Speech Technol.*, vol. 23, no. 1, pp. 45–55, 2020. doi: 10.1007/s10772-020-09672-4
[12] J. Z. Lim, J. Mountstephens, and J. Teo, "Emotion recognition using eye-tracking: Taxonomy, review and current challenges," *Sensors (Switzerland)*, vol. 20, no. 8, pp. 1–21, 2020. doi: 10.3390/s20082384
[13] A. A. A. Zamil, S. Hasan, S. M. Jannatul Baki, J. M. Adam, and I. Zaman, "Emotion detection from speech signals using voting mechanism on classified frames," in *Proc. 1st Int. Conf. Robot. Electr. Signal Process. Tech. ICREST 2019*, 2019, pp. 281–285. doi: 10.1109/ICREST.2019.8644168
[14] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011. doi: 10.1016/j.patcog.2010.09.020
[15] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5. 2018. doi: 10.1371/journal.pone.0196391
[16] M. Naveenkumar and V. K. Kaliappan, "Audio based emotion detection and recognizing tool using mel frequency based cepstral coefficient," *J. Phys. Conf. Ser.*, vol. 1362, no. 1, 2019. doi: 10.1088/1742-6596/1362/1/012063
[17] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, "Emotion detection from text and speech: A survey," *Soc. Netw. Anal. Min.*, vol. 8, no. 1, 2018. doi: 10.1007/s13278-018-0505-2
[18] I. Kansizoglou, L. Bampis, and A. Gasteratos, "An active learning paradigm for online audio-visual emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 756–768, 2022. doi: 10.1109/TAFFC.2019.2961089
[19] E. Batziou, E. Michail, K. Avgerinakis, S. Vrochidis, I. Patras, and I. Kompatsiaris, "Visual and audio analysis of movies video for emotion detection, " in *Proc. MediaEval'18*, 2018, no. 11, pp. 2018–2021.
[20] D. Hazarika, S. Gorantla, S. Poria, and R. Zimmermann, "Self-attentive feature-level fusion for multimodal emotion detection," in *Proc. IEEE 1st Conf. Multimed. Inf. Process. Retrieval, MIPR 2018*, 2018, pp. 196–201. doi: 10.1109/MIPR.2018.00043
[21] I. Popovic, D. Culibrk, M. Mirkovic, and S. Vukmirovic, "Automatic speech recognition and natural language understanding for emotion detection in multi-party conversations," in *Proc. MuCAI 2020, the 1st Int. Work. Multimodal Conversational AI*, 2020, pp. 31–38. doi: 10.1145/3423325.3423737
[22] A. A. Mande, "Emotion detection using audio data samples," *Int. J. Adv. Res. Comput. Sci.*, vol. 10, no. 6, pp. 13–20, 2019. doi: 10.26483/ijarcs.v10i6.6489
[23] Y. P. Lin *et al.*, "EEG-based emotion recognition in music listening," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1798–1806, 2010. doi: 10.1109/TBME.2010.2048568
[24] S. Ranjan, "Exploring the discrete wavelet transform as a tool for hindi speech recognition," *Int. J. Comput. Theory Eng.*, vol. 2, no. 4, pp. 642–646, 2010. doi: 10.7763/ijcte.2010.v2.216
[25] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio-visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, 2018. doi: 10.1109/TCSVT.2017.2719043
[26] L. Lu, D. Liu, and H. J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 1, pp. 5–18, 2006. doi: 10.1109/TSA.2005.860344
[27] I. M. R. Amer, B. Siddiquie, C. Richey, and A. Divakaran, "Emotion detection in speech using deep networks," in *Proc. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 3724-3728. doi: 10.1109/ICASSP.2014.6854297
[28] M. L. Dhore and I. P. Yesaware, "Speech emotion recognition using support vector machine," *Int. J. Comput. Appl.*, vol. 1, no. 20, pp. 975–8887, 2010.
[29] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling," in *Proc. 11th Annu. Conf. Int. Speech Commun.*, 2010, pp. 2362–2365. doi: 10.21437/interspeech.2010-646
[30] M. Xu, F. Zhang, and W. Zhang, "Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset," *IEEE Access*, vol. 9, pp. 74539–74549, 2021. doi: 10.1109/ACCESS.2021.3067460
[31] Z. Zhao *et al.*, "Combining a parallel 2D CNN with a self-attention dilated residual network for CTC-based discrete speech emotion recognition," *Neural Networks*, vol. 141, pp. 52–60, 2021. doi: 10.1016/j.neunet.2021.03.013
[32] R. Rekha and R. S. Tharani, "Speech emotion recognition using multilayerperceptron classifier on ravdess dataset," in *Proc. ICCAP 2021*, 2021. doi: 10.4108/eai.7-12-2021.2314726
[33] A. H. Wheeb, "Performance evaluation of UDP, DCCP, SCTP and TFRC for different traffic flow in wired networks," *Int. J. Electr. Comput. Eng.*, vol. 7, no. 6, pp. 3552–3557, 2017. doi: 10.11591/ijece.v7i6.pp3552-3557
[34] T. Dong *et al.*, "Deriving maximum light use efficiency from crop growth model and satellite data to improve crop biomass estimation," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 10, no. 1, pp. 104–117, 2017. doi: 10.1109/JSTARS.2016.2605303
[35] K. P. Sng, L. M. Ang, and C. S. Ooi, "A combined rule-based and machine learning audio-visual emotion recognition approach," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 3–13, 2018. doi: 10.1109/TAFFC.2016.2588488
[36] T. Li and M. Ogihara, "Content-based music similarity search and emotion detection," in *Proc. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process*, 2004, vol. 5. doi: 10.1109/icassp.2004.1327208
[37] N. V. Kimmatkar and B. Vijaya Babu, "Novel approach for emotion detection and stabilizing mental state by using machine learning techniques," *Computers*, vol. 10, no. 3, 2021. doi: 10.3390/computers10030037

Copyright © 2024 by the authors. This is an open access article distributed under the Creative Commons Attribution License (CC BY-NC-ND 4.0), which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.