Revue d'Intelligence Artificielle
Vol. 38, No. 2, April, 2024, pp. 397-405
Journal homepage: http://iieta.org/journals/ria

HERF: A Machine Learning Framework for Automatic Emotion Recognition from Audio

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

School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore 632014, India

Corresponding Author Email: pmvincent@vit.ac.in

https://doi.org/10.18280/ria.380203

Received: 21 July 2023
Revised: 12 December 2023
Accepted: 15 January 2024
Available online: 24 April 2024

Keywords:
human emotion recognition, machine learning, feature engineering, multilayer perceptron, audio based emotion recognition

ABSTRACT

Human emotion recognition from audio has potential applications such as healthcare, feedback assessment, gaming and advertisement to mention few. Human emotion detection often helps in assessing the feelings of person automatically. It could lead to making well informed decisions. Advancements in machine learning (ML) has paved way for unprecedented possibilities in emotion recognition from audio automatically. The methods identified in existing literature exhibit limitations, notably the absence of feature engineering to enhance predictive performance. A framework is proposed based on ML for automatic recognition of human emotions from a given voice content. The framework is called the Human Emotion Recognition Framework (HERF), designed to receive audio datasets as input and employs supervised learning for the automated identification of human emotions based on audio signals. We proposed two algorithms for realizing the framework. A Hybrid Feature Selection (HFS) algorithm is introduced to enhance the efficiency of identifying features that could have discriminative power. Additionally, the Neural Network-based Automatic Emotion Recognition (NN-AER) algorithm, utilizing Multilayer Perceptron (MLP) and HFS, is proposed for automatic emotion recognition. The feature selection provided by the HFS algorithm improves the training quality of NN-AER. RAVDESS is dataset used for empirical study. This dataset supports emotions such as neutral, happy, sad, disgust, angry, fearful and surprised. We designed a web application used to recognise emotion for given audio sample based on saved MLP model. Results of our study revealed that NN-AER outperforms many states of the art methods.

1. INTRODUCTION

The human voice reflects different kinds of emotions. Recognition of such emotions is useful in different applications. In the recent past, researchers started working on automatic emotion recognition with different techniques. The emergence of machine learning (ML), has brought unprecedented possibilities to exploit Artificial Intelligence (AI) to leverage emotion recognition performance using audio samples. However, it is challenging due to practical issues in extracting emotional content from audio samples [1]. Recognizing emotions from speech has practical applications in the real world. Such applications need interaction between humans and machines. Based on human emotion recognition, in the motor field, it is possible to know the mental state of the driver and take safety precautions. It can be used in the healthcare domain by therapists as a diagnostic tool. It is widely used in automatic translation systems where a speaker's emotion recognition plays a crucial role. It is of more use in aircraft cockpits due to the ability of the system to understand the stress level of pilots. It is used in call center applications and communications over mobiles. Many systems in the real world adapt responses to the recognition of human emotions.

The research found in the literature has revealed that there are many contributions to emotion recognition using ML models as explored in the study [1-3]. SVM is the ML technique used for emotion recognition. Feature selection and optimization using the evolutionary method are employed in the study [4]. Multi-model data and an ensemble of ML models are used in the study [5] for improving prediction performance. Another ensemble model with bagged SVMs is explored in the research [6] for exploiting many models. Deep learning models are also found useful in human emotion recognition as studied [7-10]. Pre-trained deep CNN model is used in the research [11] besides an attention model. In the study [12], Multilayer Perceptron (MLP) and DL models are employed. A weighted approach using deep learning is explored in the research [1]. From the literature, it is ascertained that ML models have the potential to acquire intelligence from historical data or training samples for the automatic detection of emotion from audio samples. However, there is a need for optimal feature selection and neural network combination to improve performance further. Towards this end, our contributions are given below.

1. A framework based on ML is proposed to recognize human emotions from a given voice content. The framework is named as Human Emotion Recognition Framework (HERF).
2. The Hybrid Feature Selection (HFS) algorithm is proposed to find features that have discriminative power. The system employs an iterative process to extract MFCC,

STFT, and Mel Spectrogram features from each audio sample. These features are fused to create a feature map, which is subsequently utilized in the NN-AER algorithm.

3. Another algorithm known as Neural Network-based Automatic Emotion Recognition (NN-AER) which exploits Multilayer Perceptron (MLP) and HFS for automatic recognition of human emotions from audio samples. During the training phase, the HFS algorithm is executed to acquire the feature map corresponding ground truth values to train the MLP classifier. Once the model is trained, it is stored for subsequent use, facilitating the automatic prediction of emotion recognition classes for a given audio sample.
4. Our framework is evaluated with web web-based interface to test new audio samples from the knowledge model saved after training a based classifier.

The following sections lighten different aspects of the paper. Section 2 reviews the literature of prior works on human emotion recognition from audio files. Section 3 presents the materials and methods associated with the research. Section 4 shows empirical results. Section 5 delves into the proposed framework and outlines the limitations of the underlying model. Section 6 concludes our work and provides future scope.

2. RELATED WORK

This section covers an overview of the prior works concerning existing approaches to emotion recognition from audio. Seng et al. [2] highlighted the societal shift towards spiritual aspects and proposed an emotion communication system to address non-line-of-sight challenges, facilitating real-time multimedia emotion transmission. Livingstone et al. [13] focused on speech emotion classification, focusing on feature selection, classification methods, and emotional speech database preparation. The discussion addressed performance limitations, emphasizing the need for improved classification accuracy and database quality. Proposed extensions included the integration of speaker-dependent and independent systems, along with exploring temporal structure modelling and multiple classifier systems. Kumar et al. [14] introduced a video-based abnormal human activity recognition system for elderly care, ensuring privacy through binary silhouettes. Using R-transform and KDA, the system achieved high recognition accuracy. Future enhancements involve incorporating depth silhouettes for heightened discrimination. Livingstone et al. [13] proposed a novel architecture to discriminate emotional physiological signals from multichannel bio signals, achieving a recognition rate of 77.68% for four emotional states. The approach holds promise for effective healthcare applications, particularly in monitoring elderly or chronically ill individuals. Kumar et al. [14] used EEG and ML to categorize human emotions, attaining a good level of accuracy. The identified features suggest practical applications in non-invasive emotional assessment, with future work aiming to deepen the understanding of brain responses to music at various stages.

Bhavan et al. [15] Introduced a Hindi speech recognition system utilizing the Discrete Wavelet Transform and the K Means Algorithm, where the Daubechies8 with 5-level decomposition demonstrated optimal outcomes. Speaker independence was achieved through Hidden Markov Models (HMMs). Tarantino et al. [16] presented a context-sensitive multimodal emotion recognition approach, incorporating BLSTM networks for capturing context-related information BLSTM exhibited superior performance compared to standard techniques, achieving notable discrimination in emotional space clusters. Prospective investigations may delve into dynamic modelling for low-level features and integrate linguistic information into the system. Li et al. [17] developed a SER system extracting features like MFCC and MEDC, utilizing SVM for emotional state recognition. The system demonstrated high accuracy, indicating independence from speaker and text variations. Batziou et al. [3] explored Speech Emotion Recognition, focusing on enhanced performance through various features. Feature selection using Fast Correlation-Based Filter (FCBF) identified 25 key features and the Fusion of Artificial Neural Networks (FAMNN) with Genetic Algorithm (GA) optimization. Xu et al. [18] proposed a modified supervised manifold learning algorithm (MSLLE) for spoken emotion recognition, emphasizing improved interclass distance and generalization. Experimental results on two databases showcased the superior performance of MSLLE over other methods.

Ma et al. [19] used a novel deep CNN architecture, PCNSE-SADRN-CTC, designed for discrete speech emotion recognition, demonstrating efficiency on various datasets. Future research aims to explore its applications in diverse speech-related tasks. Shah et al. [20] introduced Fusion-ConvBERT, a pioneering fusion network model for Speech Emotion Recognition (SER), showcasing superior performance across datasets. Imani et al. [21] proposed a novel Speech Emotion Recognition (SER) method, amalgamating DCNN and BLSTMwA models, surpassing popular approaches on the EMO-DB and IEMOCAP datasets. Zhang et al. [22] introduced the RAVDESS, a validated multimodal emotional database featuring 24 actors expressing diverse emotions in speech and song, freely accessible to researchers. Jiang et al. [23] discussed Chroma feature extraction using STFT and. CQT, highlighting the advantages of STFT for chord recognition projects.

Koduru et al. [24] examined various ANN models for forecasting based on temporal data. The study concludes that RBF yields improved accuracy, followed by RNN and MLP, while GRNN exhibits the lowest efficiency. Emphasizing the versatility of ANN models in psychology, the research advocates for effective pre-processing of time series data and proposes an enhanced performance index. Despite the flexibility demonstrated, researchers are encouraged to address limitations and explore the generalization of these models to other databases in future studies. Lee et al. [25] concentrated on voice-based emotion detection utilizing MLP and CNN classifiers, with CNN outperforming MLP in a web application. Plans involve expanding training data, experimenting with different architectures, and extending emotion detection to video, image, and text inputs. Lim et al. [26] tackled Speech Emotion Recognition (SER) challenges, introducing Head Fusion with multi-head attention, resulting in improved accuracy (76.18% WA, 76.36% UA) and robustness against added noise. Zhang et al. [27] used an audio-visual system integrating rule-based and machine-learning techniques. The system employs BDPCA+LSLDA and OKL-RBF neural classifiers for visual optimization. Future refinements include optimizing window length and overlap in the audio path and exploring applications such as customer satisfaction assessment. Shah et al. [20] conducted a review of EEG-based methods, encompassing feature

extraction, reduction, ML classifiers, and the correlation of EEG rhythms with emotions. The review compares ML and deep learning algorithms, identifying open problems certain issues

Raghu Vamsi et al. [28] Present an attention mechanism that aligns speech frames with recognized text, the proposed model achieves better results on the IEMOCAP dataset. Zhao et al. [29] contribute to speech emotion recognition by proposing a bagged ensemble with Gaussian-kernel SVMs, demonstrating superior performance across three datasets compared to state-of-the-art approaches. Xu et al. [30] introduce a multitask learning method, incorporating self-attention and gender classification as auxiliary tasks, resulting in a 7.7% improvement over existing methods on the IEMOCAP dataset. Focusing on enhancing speech emotion recognition, Lee et al. [25] utilize feature extraction methods like MFCC, DWT, pitch, energy, and ZCR, leading to improved accuracy and efficiency in experimental results. Addressing multi-modal emotion recognition challenges, Li et al. [17] propose a deep-weighted fusion method that incorporates cross-modal noise modelling and effective feature extraction.

Li et al. [31] Examining emotion recognition with eye-tracking technology, the study outlines pertinent features, and challenges, and acknowledges the limited existing literature in this domain. Zhang et al. [27] delve into enhancing Speech Emotion Recognition (SER) by employing a novel windowing system and self-attention, showcasing improved performance on the IEMOCAP dataset. Introducing the Probability and Integrated Learning (PIL) algorithm, Raghu Vamsi et al. [28] tackle complex human emotion recognition, specifically addressing emotional uncertainty through classification probability. While showing promise for affective computing in videos and artificial emotion for robots, the method warrants further exploration and refinement. Zhao et al. [29] contribute a cost-effective navigation and face recognition system for the visually impaired utilizing IoT, smartphones, GPS, and ultrasonic sensors. Achieving 90% face recognition accuracy and 95% obstacle detection, potential future work involves broader object recognition and dynamic face reactions with IoT integration, considering the current limitation of static face recognition. Recognizing the significance of learner emotions in e-learning, Xu et al. [30] explore methods such as voice recognition, facial expressions, and gestures. Their findings indicate that multimodal systems, combining various aspects prove more effective than single-modal approaches However, there are certain limitations in existing methods, particularly the lack of feature engineering to enhance prediction performance.

3. MATERIALS AND METHODS

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

3.1 Dataset details

RAVDESS is widely used dataset [31] with 1440 audio samples from 24 professional actors of both genders covering different emotions. Each emotion has two levels of intensity known as normal and strong.

As presented in Table 1, there are 8 classes of emotions including neutral which actually does not reflect any emotion. In fact, neutral does mean absence of emotion. Therefore, there is not strong intensity level associated with this class.

Table 1. Emotion classes and their intensity in 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

3.2 The framework

We proposed a ML based framework known as Proposed Human Emotion Recognition Framework (HERF). It has provision to take audio dataset as input and performs supervised learning for automatic detection of human emotions from audio. Given inputs are subjected to audio normalization followed by feature extraction and selection of features. Afterwards, a neural network based classifier is trained. The training results in knowledge model creation. This knowledge model is used for prediction of new audio samples and recognize 8 categories of emotions. The model is saved for further usage. A web based application is developed and deployed. This application enables users to choose an audio sample for testing and the saved model is reused to predict class label. Normalizing audio is critical as it involves adjusting the volume of an audio signal, ensuring a uniform level, and enhancing audibility for ease of listening.

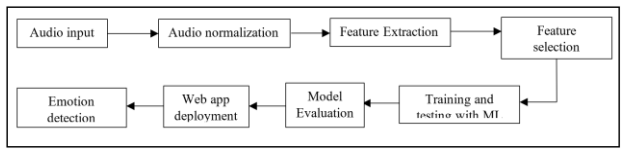

Figure 1. Proposed Human Emotion Recognition Framework (HERF)

As presented in Figure 1, the framework has provision for automatically recognizing human emotions. In the process feature selection plays crucial role as it can influence the quality of training a classifier. Feature selection the process in which different features from the audio sample are extracted and used. Three kinds of features such as MFCC, STFT and Mel spectrogram features are used in the empirical study. MFCC features effectively represent human voice, employing a logarithmic power spectrum's linear cosine transform on the Mel-frequency scale, which exhibits nonlinearity. These coefficients, derived from an audio clip, offer a cepstral representation of the audio. The advantage of MFCC over standard cepstral concepts lies in its utilization of evenly spaced frequency bands on the Mel scale, approximating human voice. Another method for audio feature extraction is Short Time Fourier Transform (STFT), which computes a Fourier transform from the audio signal, primarily designed for pitch shifting, pitch detection, and noise reduction. Additionally, this paper employs Mel spectrogram as a feature extraction technique, providing a visual representation of frequency spectra over time in the voice. The Mel spectrogram creates a 2D modelling of audio clip reflecting the amplitude

of energy associated with each frequency component in the temporal domain. These diverse features are sequentially combined to generate a single feature. Then the combined feature vector is used for training classifier. Figure 2 illustrates model training and testing associated with the HERF framework.

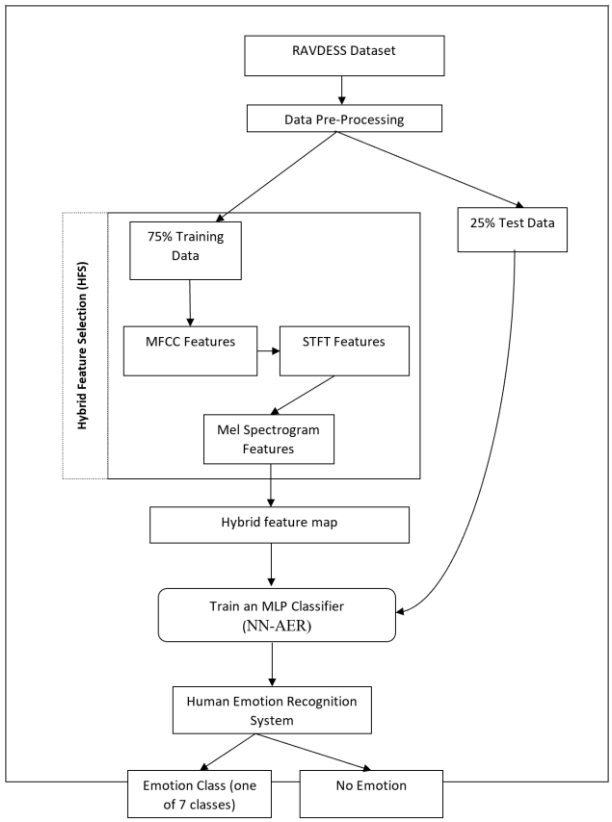

Figure 2. Illustrates model training and testing of HERF framework

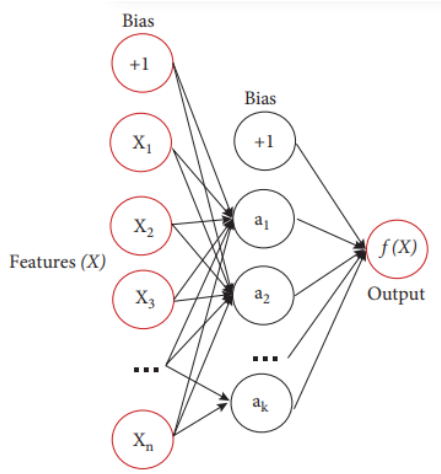

Figure 3. Architecture of MLP classifier

As presented in Figure 2, RAVDESS dataset is given as input. It is subjected to dividing 75% training and 25% testing. On the given training data MFCC features, STFT features and Mel spectrogram features extracted from training samples and combined into a single feature vector. An algorithm known as HFS is defined to select features. The hybrid feature map is used for training MLP classifier. In order to exploit MLP classifier and HFS algorithm we proposed an algorithm known as Neural Network based Automatic Emotion Recognition (NN-AER). The training results in a knowledge model which is used to classify test samples into different classes.

As presented in Figure 3, Multilayer Perceptron (MLP) is a kinf of ANN model. It has 3 significant layers and the functioning is reflected in Eq. (1) and Eq. (2).

$$h^1 = step(z^1) = step(w^1.x + b^1) \quad (1)$$

$$y = step(z^2) = step(w^2.h^1 + b^2) \quad (2)$$

ANN variants need batch based training where X is the input vector. Eq. (3) shows how k instances are obtained from available ones.

$$x_1 = \begin{pmatrix} x_{1,1} \\ \dots \\ x_{1,n} \end{pmatrix}, \dots\dots, x_k = \begin{pmatrix} x_{k,1} \\ \dots \\ x_{k,n} \end{pmatrix} \quad (3)$$

Afterwards combining the instance is done as in Eq. (4).

$$X = \begin{pmatrix} x_1^T \\ \dots \\ x_k^T \end{pmatrix} = \begin{pmatrix} x_{1,1} \dots\dots x_{1,n} \\ \dots \\ x_{k,1,} \dots\dots\dots, x_{k,n} \end{pmatrix} \quad (4)$$

Having understood it, y is computed as in Eq. (5).

$$y=step(z)=step(X.W+b) \quad (5)$$

where, (k, n) is shape of vector X, n denotes number of values in input while k denotes number of instances. W is a matrix.

3.3 Proposed algorithms

We proposed two algorithms for realizing the framework. HFS is proposed to find discriminating features (HFS) is proposed to find features that contribute to class label prediction. Another algorithm known as Neural Network based Automatic Emotion Recognition (NN-AER) which exploits Multilayer Perceptron (MLP) and HFS for automatic emotion recognition.

Algorithm 1. Hybrid Feature Selection Algorithm

Algorithm: Hybrid Feature Selection (HFS)
Input: RAVDESS dataset ***D***
Output: Features ***M***

1. Start
2. Initialize MFCC feature vector ***F1***
3. Initialize STFT feature vector ***F2***
4. Initialize Mel feature vector ***F3***
5. Initialise feature vector ***F***
6. Initialize feature map ***M***
7. For each *d* in *D*
8. ***F1***←ExtractMFCCFeatures(***d***)
9. ***F2***←ExtractSTFTFeatures(***d***)
10. ***F3***←ExtractMelFeatures(***d***)
11. ***F***←***F1***+***F2***+***F3***
12. Add d and ***F*** to ***M***
13. End For
14. Return ***M***
15. End

Algorithm 2 takes RAVDESS as input and generate feature map. It has an iterative process for extracting MFCC, STFT and Mel Spectrogram features from each audio sample and combine the features to realize a Then the feature map is used

to predict emotion classes for given audio test samples. It generates confusion matrix for knowing efficiency.

As presented in Algorithm 2, it takes RAVDESS training set and testing set. In the training process HFS algorithm is run to obtain features map and ground truth values and use them for training MLP classifier. After training the model is saved for further usage to have automatic prediction of emotion recognition classes for given audio sample. Once training is completed, test samples are used to obtain feature map using HFS algorithm. Then the feature map is used to predict emotion classes for given audio test samples. It generates confusion matrix for knowing efficiency.

Algorithm 2. Neural Network based Automatic Emotion Recognition (NN-AER) algorithm

Algorithm: Neural Network based Automatic Emotion Recognition (NN-AER)
Inputs:
RAVDESS training set ***T1***
RAVDESS test set ***T2***
Output: Emotion recognition results ***R***

1. Start
2. Initialize feature map ***M***
3. Initialize recognition results ***R***
4. ***M***←Run HFS(***T1***)
5. ***m***←Train MLP model using ***M*** and ground truth values
6. Save model ***m***
7. M←Run HFS(***T2***)
8. For each map entry in M
9. class←PredictEmotion(***m***)
10. Update ***R***
11. End For
12. Display ***R***
13. Generate confusion matrix
14. Evaluation
15. Display evaluation results
16. End

3.4 Evaluation methodology

Figure 4 shows confusion matrix which helps in deriving values that show difference between ground truth and predictions.

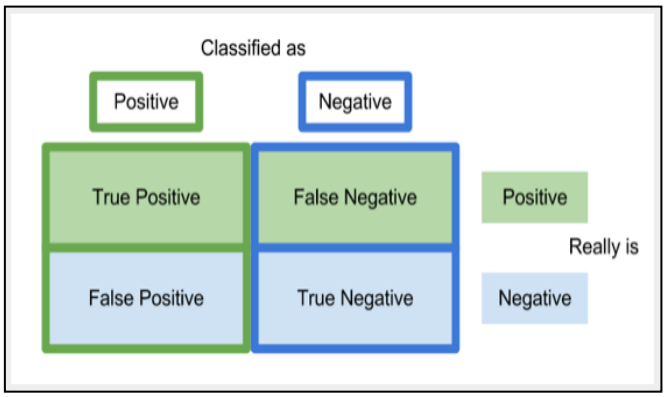

Figure 4. Confusion matrix

Computation of the performance metrics are based on correct and wrong predictions of a ML model. Precision and recall are computed as in Eq. (6) and Eq. (7).

$$\text{Precision}=\frac{TP}{TP+FP} \quad (6)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (7)$$

Other metrics such as F1-score and accuracy are computed as in Eq. (8) and Eq. (9).

$$\text{F1-score}=2*\frac{(p*r)}{(p+r)} \quad (8)$$

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

In F1-score computation, p denoted precision while r denotes recall values. All these metrics when evaluated result in a range of value from 0 to 1 reflecting least and highest performance.

4. RESULTS AND DISCUSSION

Experiments are made to evaluate the emotion recognition efficiency of the proposed MLP-based ML framework with underlying algorithms. RAVDESS dataset [1] is used for experiments. Data is split into 70% (training) and 25% (testing). MLP classifier is configured using the values presented in Table 2. After setting parameters, MLP is trained with a 70% training set. After training the model, it is tested with the 25% test samples that were not known to the model. This section presents exploratory data analysis and performance evaluation.

Table 2. Parameters configured with MLP classifier

Parameter	Value
max_iter	500
learning_rate	adaptive
hidden_layer_sizes	300
epsilon	1e-08
batch_size	256
alpha	0.01

The proposed algorithm named Neural Network-based Automatic Emotion Recognition (NN-AER) exploits the MLP classifier and also the HFS algorithm to improve prediction performance. On predicting the class of test samples, ground truth is used to generate a confusion matrix. Figure 8 shows the generated confusion matrix based on which performance evaluation is made.

4.1 Data analysis

Data analysis with the RAVDESS dataset is provided in this section. It has 8 classes of emotions in training and test samples.

As presented in Figure 5, wave plot is generated for an audio consisting of happy emotion in its audio content.

As presented in Figure 6, Spectrogram is generated for an audio consisting of happy emotion in its audio content.

As presented in Figure 7, wave plot is generated for an audio consisting of sad emotion in its audio content.

As presented in Figure 8, Spectrogram is generated for an audio consisting of sad emotion in its audio content.

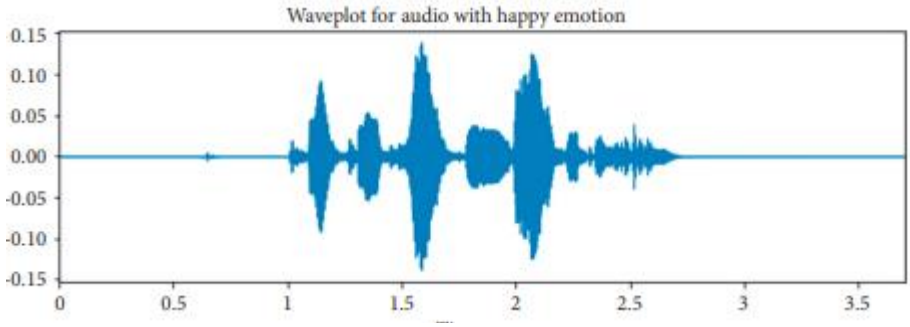

Figure 5. Sample audio with happy emotion

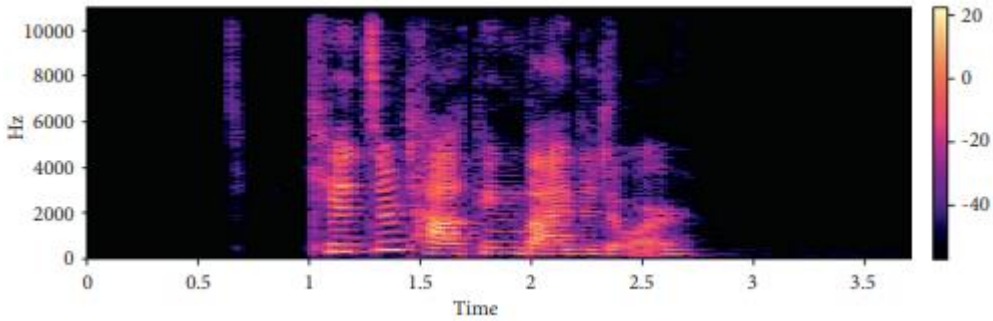

Figure 6. Spectrogram of sample audio with happy emotion

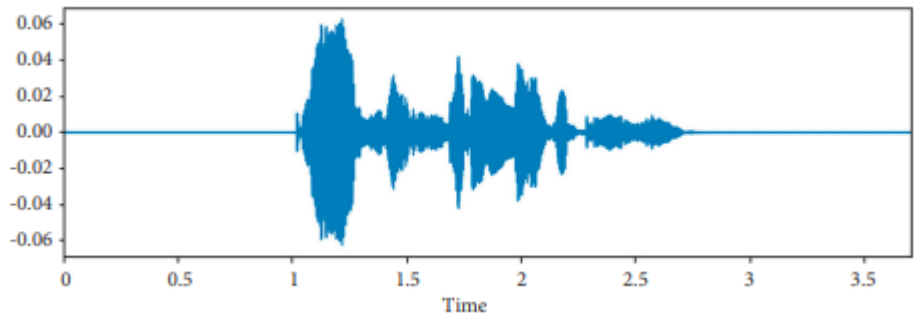

Figure 7. Sample audio with sad emotion

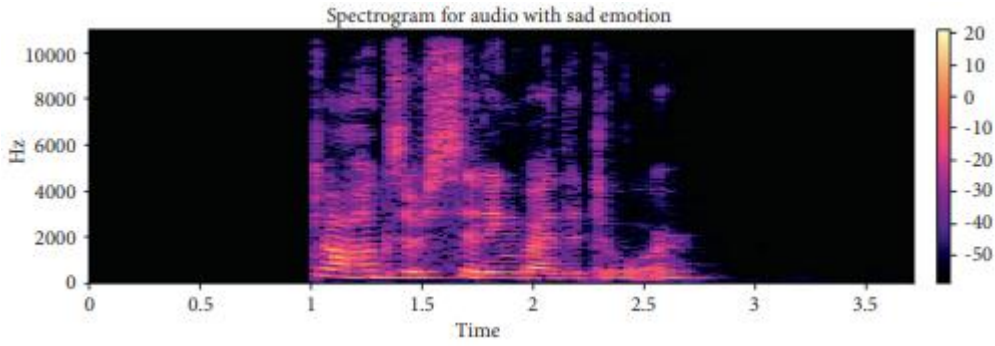

Figure 8. Spectrogram of sample audio with sad emotion

4.2 Experimental results

With the given 25% test data, the proposed algorithm with underlying MLP classifier and HFS algorithm could produce prediction results in the form of confusion matrix for all emotion classes.

Figure 9 shows confusion matrix for 8 classes. By using the prediction results, efficiency of the models is measured in terms of accuracy, F1-score, precision and recall

As presented in Table 3, performance of the proposed model for human emotion recognition from audio is provided with several metrics such as precision, recall and F1-score.

As presented in Figure 10, precision performance of the proposed model is observed for different classes. Precision for neutral class is 88%, calm 91%, happy 71%, sad 70%, angry 84%, fearful 87%, disgust 78% and surprised 71%. Least

precision is exhibited for sad emotion with 70%. Highest precision is exhibited by 91% for detection of calm emotion.

Figure 11 shows precision of the model for different classes. Recall for neutral class is 88%, calm 95%, happy 67%, sad 81%, angry 78%, fearful 85%, disgust 69% and surprised 77%. Least precision is exhibited for happy emotion with 67%. Highest precision is exhibited by 95% for detection of calm emotion.

Figure 12 shows F1-score the model. F1-score for neutral class is 88%, calm 93%, happy 69%, sad 75%, angry 81%, fearful 86%, disgust 73% and surprised 74%.

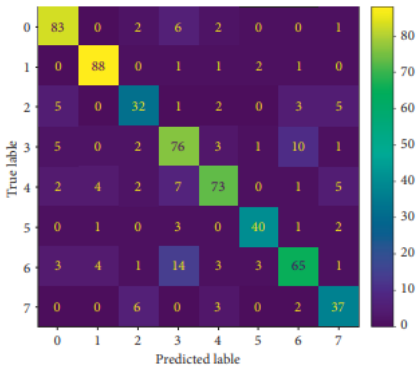

Figure 9. Prediction results reflecting 8 classes in the form of confusion matrix

Table 3. Performance of the proposed model for human emotion recognition from audio

Emotion	Performance (%)		
	Precision	Recall	F1-score
Angry	0.84	0.78	0.81
Calm	0.91	0.95	0.93
Disgust	0.78	0.69	0.73
Fearful	0.87	0.85	0.86
Happy	0.71	0.67	0.69
Neutral	0.88	0.88	0.88
Sad	0.7	0.81	0.75
Surprised	0.71	0.77	0.74

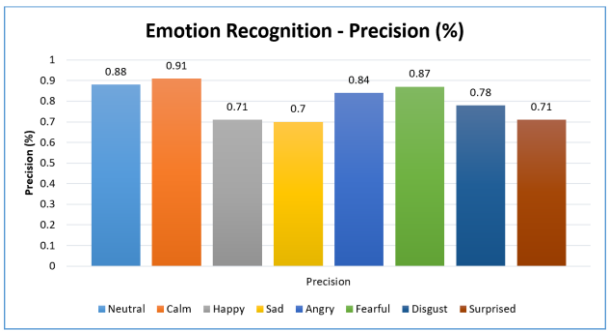

Figure 10. Precision performance of the model

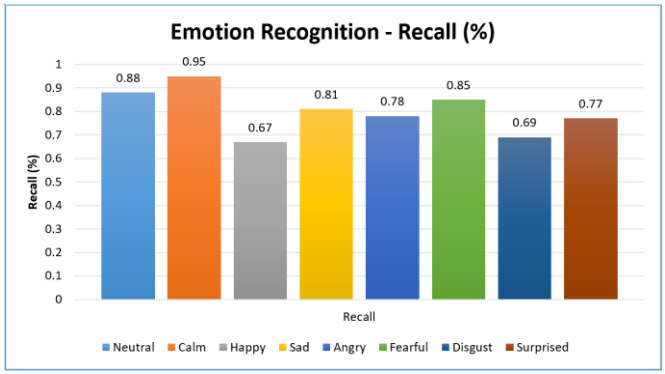

Figure 11. Recall performance of the proposed model for all classes

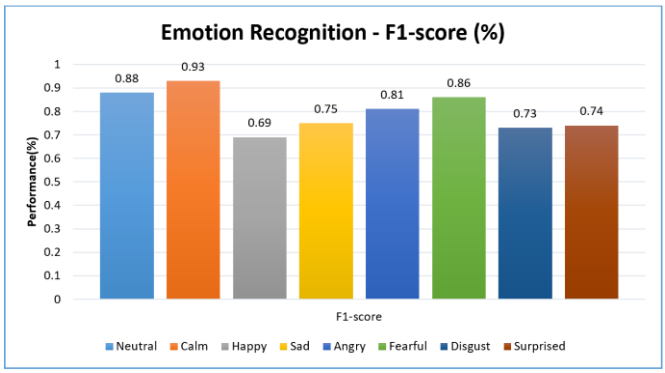

Figure 12. F1-score performance of the proposed model for all classes

4.3 Performance evaluation

This section presents compares our model with existing models found in Raghu Vamsi et al. [28] and Xu et al. [30].

As presented in Table 4, there are 8 classes of emotions for which the performance of each model is provided.

Figure 13 shows the performance of all models. For the neutral class, the highest performance is shown by 88.29% of the proposed model. For the neutral class, the highest performance is shown by 88.29% of the proposed model. For calm emotion highest performance exhibited by the proposed model with 90.72%. About happy class's highest performance is shown by the study [13] with 73.49%. When sad emotion is considered, the highest performance is observed at 81% [14]. The highest performance for angry emotion is exhibited by the study [14] with 88%. Concerning fearful emotion showed the highest performance with 89% [13]. Disgust is the emotion that exhibited the highest performance 7831% with the proposed model. The highest performance is exhibited by the study [14] for surprised emotion.

As presented in Table 5 accuracy of different models is observed. Higher accuracy denotes better performance.

Table 4. Shows performance of all emotion recognition models

Model	Performance (%)							
	Calm	Angry	Sad	Happy	Fearful	Surprised	Disgust	Neutral
Raghu Vamsi et al. [28]	84.09	80.77	59.79	73.49	63.37	68.57	50.98	64.81
Xu et al. [30]	71	88	81	70	89	81	73	65
Proposed	90.72	83.9	70.3	71.11	86.95	71.15	78.31	88.29

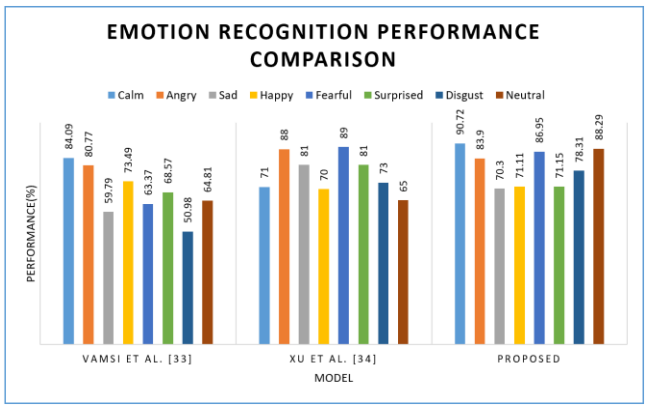

Figure 13. Emotion recognition performance comparison

Table 5. Shows accuracy comparison of different emotion recognition models

Models	Accuracy (%)
Raghu Vamsi et al. [28]	69.49
Xu et al. [30]	76.36
Proposed	81

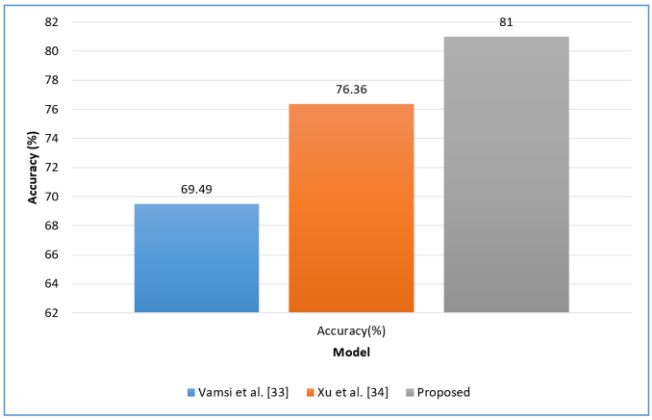

Figure 14. Accuracy of all models

Figure 14 shows models. The accuracy of the study [13] is 68.49% while the accuracy of the study [14] is 76.36%. The highest accuracy is exhibited by the proposed model with 81%. Therefore, it can be understood that, as far as human emotion recognition from audio is concerned, the proposed model outperforms existing ones.

5. LIMITATIONS OF THE PROPOSED FRAMEWORK

The presented framework relies on machine learning techniques for the recognition of emotions from audio content. It adopts an effective approach to feature engineering, enhancing the quality of training during the learning process. The Multilayer Perceptron (MLP) serves as the neural network model, demonstrating efficiency in classification tasks and is adept at categorizing eight distinct emotion classes. However, despite achieving a model accuracy of 81%, the highest among the evaluated models, there remains potential for further improvement in accuracy. Additionally, it's noteworthy that the study utilizes the RAVDESS dataset. To ensure the model's efficiency is generalizable, it is imperative to evaluate its performance across multiple datasets.

6. CONCLUSION AND FUTURE WORK

A framework is proposed based on ML for automatic recognition of human emotions from a given voice content. The framework is named as Human Emotion Recognition Framework (HERF). We proposed two algorithms for realizing the framework. HFS is proposed to find discriminating features. Another algorithm known as Neural Network based Automatic Emotion Recognition (NN-AER) which exploits Multilayer Perceptron (MLP) and HFS for automatic emotion recognition. RAVDESS is dataset used for empirical study. This dataset supports 8 categories of emotions. HERF introduces novelty by harnessing multiple feature selection methods to enhance the training process. We also designed a web application used to recognise emotion for given audio sample based on saved MLP model. Experimental results revealed that the proposed algorithm NN-AER shows better performance of prior methods with highest accuracy 81%. In future we propose DL based framework for improving accuracy further.

REFERENCES

[1] Chen, M., Zhou, P., Fortino, G. (2016). Emotion communication system. IEEE Access, 5: 326-337. https://doi.org/10.1109/ACCESS.2016.2641480

[2] Seng, K.P., Ang, L.M., Ooi, C.S. (2016). A combined rule-based & machine learning audio-visual emotion recognition approach. IEEE Transactions on Affective Computing, 9(1): 3-13. https://doi.org/10.1109/TAFFC.2016.2588488

[3] Batziou, E., Michail, E., Avgerinakis, K., Vrochidis, S., Patras, I., Kompatsiaris, I. (2018). Visual and audio analysis of movies video for emotion detection. Emotional Impact of Movies task MediaEval 2018.

[4] Lin, Y.P., Wang, C.H., Jung, T.P., Wu, T.L., Jeng, S.K., Duann, J.R., Chen, J.H. (2010). EEG-based emotion recognition in music listening. IEEE Transactions on Biomedical Engineering, 57(7): 1798-1806. https://doi.org/10.1109/TBME.2010.2048568

[5] Ranjan, S. (2010). Exploring the discrete wavelet transform as a tool for Hindi speech recognition. International Journal of Computer Theory and Engineering, 2(4): 642-646.

[6] Wöllmer, M., Metallinou, A., Eyben, F., Schuller, B., Narayanan, S. (2010). Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. in Proceedings of the INTERSPEECH 2010: 2362-2365. https://doi.org/10.21437/Interspeech.2010-646

[7] Frantzidis, C.A., Bratsas, C., Klados, M. A., Konstantinidis, E., Lithari, C.D., Vivas, A.B., Bamidis, P.D. (2010). On the classification of emotional biosignals evoked while viewing affective pictures: An integrated data-mining-based approach for healthcare applications. IEEE Transactions on Information Technology in Biomedicine, 14(2): 309-318. https://doi.org/10.1109/TITB.2009.2038481

[8] El Ayadi, M., Kamel, M.S., Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3): 572-587. https://doi.org/10.1016/j.patcog.2010.09.020

[9] Khan, Z.A., Sohn, W. (2011). Abnormal human activity recognition system based on R-transform and kernel discriminant technique for elderly home care. IEEE Transactions on Consumer Electronics, 57(4): 1843-1850. https://doi.org/10.1109/TCE.2011.6131162

[10] Moreno, M., José, J., Pol, P., Gracia, A.F., del Pilar, M., (2011). Artificial neural networks applied to forecasting time series. LIAM - Laboratori de Modelització i Anàlisi de la Informació. http://hdl.handle.net/2117/12502.

[11] Gharavian, D., Sheikhan, M., Nazerieh, A., Garoucy, S. (2012). Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Computing and Applications, 21: 2115-2126. https://doi.org/10.1007/s00521-011-0643-1

[12] Zhang, S., Zhao, X. (2013). Dimensionality reduction-based spoken emotion recognition. Multimedia Tools and Applications, 63: 615-646. https://doi.org/10.1007/s11042-011-0887-x

[13] Livingstone, S.R., Russo, F.A. (2018). 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, 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391

[14] Kumar, P.M., Gandhi, U., Varatharajan, R., Manogaran, G., Jidhesh, R., Vadivel, T. (2022). Intelligent face recognition and navigation system using neural learning for smart security in Internet of Things (Retraction of Vol 22, Pg S7733, 2017).

[15] Bhavan, A., Chauhan, P., Shah, R.R. (2019). Bagged support vector machines for emotion recognition from speech. Knowledge-Based Systems, 184: 104886. https://doi.org/10.1016/j.knosys.2019.104886

[16] Tarantino, L., Garner, P.N., Lazaridis, A. (2019). Self-attention for speech emotion recognition. In Interspeech, pp. 2578-2582. http://doi.org/10.21437/Interspeech.2019-2822

[17] Li, Y., Zhao, T., Kawahara, T. (2019). Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In Interspeech, pp. 2803-2807. http://doi.org/10.21437/Interspeech.2019-2594

[18] Xu, H., Zhang, H., Han, K., Wang, Y., Peng, Y., Li, X. (2019). Learning alignment for multimodal emotion recognition from speech. arXiv preprint arXiv:1909.05645. https://doi.org/10.48550/arXiv.1909.05645

[19] Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., Košir, A. (2019). Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. Information Fusion, 46: 184-192. https://doi.org/10.1016/j.inffus.2018.06.003

[20] Shah, A.K., Kattel, M., Nepal, A., Shrestha, D. (2019). Chroma feature extraction. In Proceedings of the Conference: Chroma Feature Extraction Using Fourier Transform, pp. 1-14.

[21] Imani, M., Montazer, G.A. (2019). A survey of emotion recognition methods with emphasis on E-Learning environments. Journal of Network and Computer Applications, 147: 102423. https://doi.org/10.1016/j.jnca.2019.102423

[22] Zhang, J., Yin, Z., Chen, P., Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Information Fusion, 59: 103-126. https://doi.org/10.1016/j.inffus.2020.01.011

[23] Jiang, D., Wu, K., Chen, D., Tu, G., Zhou, T., Garg, A., Gao, L. (2020). A probability and integrated learning based classification algorithm for high-level human emotion recognition problems. Measurement, 150: 107049. https://doi.org/10.1016/j.measurement.2019.107049

[24] Koduru, A., Valiveti, H.B., Budati, A.K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. International Journal of Speech Technology, 23(1): 45-55. https://doi.org/10.1007/s10772-020-09672-4

[25] Lee, S., Han, D.K., Ko, H. (2020). Fusion-ConvBERT: Parallel convolution and BERT fusion for speech emotion recognition. Sensors, 20(22): 6688. https://doi.org/10.3390/s20226688

[26] Lim, J.Z., Mountstephens, J., Teo, J. (2020). Emotion recognition using eye-tracking: Taxonomy, review and current challenges. Sensors, 20(8): 2384. https://doi.org/10.3390/s20082384

[27] Zhang, H., Gou, R., Shang, J., Shen, F., Wu, Y., Dai, G. (2021). Pre-trained deep convolution neural network model with attention for speech emotion recognition. Frontiers in Physiology, 12: 643202. https://doi.org/10.3389/fphys.2021.643202

[28] Raghu Vamsi, U., Yuvraj Chowdhary, B., Harshitha, M., Ravi 'eja, S., Divya Ud. J. (2021). Speech emotion recognition(ser) using multilayer perceptron and deep learning techniques. IEEE Access, 27(5): 386-394.

[29] Zhao, Z., Li, Q., Zhang, Z., Cummins, N., Wang, H., Tao, J., Schuller, B.W. (2021). Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition. Neural Networks, 141: 52-60. https://doi.org/10.1016/j.neunet.2021.03.013

[30] Xu, M., Zhang, F., Zhang, W. (2021). Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset. IEEE Access, 9: 74539-74549. https://doi.org/10.1109/ACCESS.2021.3067460

[31] Li, T., Ogihara, M. (2004). Content-based music similarity search and emotion detection. 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada. https://doi.org/10.1109/ICASSP.2004.1327208