Chapter 9
A Smart System for the Assessment of Genuineness or Trustworthiness of the Tip-Off Using Audio Signals: An Explainable AI Approach

Sirshendu Hore and Tanmay Bhattacharya

Abstract Assessment of the genuineness or trustworthiness of a Tip-off is a challenging research area as it depends on the mental state and perception of the Tip-off providers. Thus, in the proposed work an attempt has been made to help the Law Enforcement (LE) personnel to assess the legitimacy of a Tip-off from a voice call. For the aforesaid objective, four widely used mental states such as 'Anger', 'Happy', 'Sadness', and 'Neutral' have been considered. To placate our goal, a few classical Machine Learning (ML) models, as well as a few latest ML models, have been employed. Regional, international, and a combination of both audio sets have been engaged for an in-depth study. The novelty of this work is to, select a set of Important 26 or 13 Mel-Frequency Cepstral Coefficients (MFCCs) using Explainable AI (XAI) approaches (Mean Decreased Impurity based Gini and Permutation), whereas most of the researchers had employed either the First 26 or 13 MFCCs in their works. The proposed model shows the supremacy over the conventional approach of using sequential MFCCs feature vector result analysis shows the supremacy of XAI-based features over conventional approaches thereby making our system better and smarter. Among the employed models, 1D CNN has shown its supremacy over other employed models for this study. Hence, the 1D-CNN-based Machine learning approach has been proposed.

Keywords Mental awareness · Machine learning · Smart system · Tip-off · Explainable AI approach

S. Hore (✉)
Department of CSE, Hooghly Engineering and Technology College, Pipulpati, Hooghly, West Bengal, India
e-mail: shirshendu.hore@hetc.ac.in

T. Bhattacharya
Department of IT, Techno Main, Salt Lake, Kolkata, India

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_9

9.1 Introduction

The purpose of a Tip-off is to send confidential information or to give an early warning to an individual or organization so that preventive measures can be initiated (https://dictionary.cambridge.org/dictionary/english/tip-off. xxxx). A timely Tip-off may prevent blood shade, and violence may bring down communal tension on the other hand it helps the LE personnel to capture a criminal who is involved in a heinous act such as murder or rape. In today's society, every citizen has the right to feel safe and stay safe. It is the collective responsibility of the citizens to become the ear and eyes of the local law enforcement official. If they observed any suspicious activity or crime then it's their basic responsibility to bring such incidents before the law enforcement officials. However, it has been observed that sometimes prank calls and intentional miss reporting for vengeance may distract the law enforcement officials, which may lead to a waste of time and effort (https://www.criminallawyersandiego.com//crimes-police-government/false-report/). By doing so these criminals and their associates can keep the police personnel busy in an area where no crime has been observed or taken place, at the same time allowing them to carry out criminal activity in other parts of the locality. In this way, they succeed to mislead the law enforcement agencies. Therefore, before initiating any action based on a Tip-off received, probably it will be better if these agencies tried to measure the genuineness or trustworthiness of the Tip-off. One way to measure the genuineness or trustworthiness of the Tip-off is to check the mental state of the Tip-off provider from their voice or speech. It has been reported in some literature that we can get people's mental state from their voices or speeches. Understanding people's mental state from their voices is one of the important areas of research. Consequently, lots of research work have been carried out in this direction (Basharirad and Moradhaseli 1891; Ayaida et al. 2011; Poria et al. 2017). The use of ML models has been used widely in SER based approach (Pinto et al. 2020; Akçay and Oguz 2020). It has been observed that RAVDESS, an English audio corpus (Livingstone et al. 2018), and EMODB, a German audio corpus (EMO-DB) have been used mostly in the SER-based system (Yang et al. 2020; Chatterjee et al. 2021; Lalitha et al. 2014; Iqbal and Barua 2019). Between the two methods, used popularly in SER based approach, isolated or static label-based systems are most commonly used in the SER because it is easier to implement. It has been found in various studies that MFCC is the most trusted audio feature among researchers (Shashidhar et al. 2018; Boles and Rad 2017; Panwar et al. 2017). Nowadays, several modern methods have been used to make ML-based AI models more explanatory. Collectively, these approaches are called XAI approaches. (McDermid 2021).One of the methods used in XAI (Saarela and Jauhiainen 2021; Fisher et al. 2019) is to look for important features from feature vectors. Thus, in the proposed work, we have generated 26 and 13 MFCCs using Mean Decrease in Impurity (MDI) and employing Permutation Importance with Correlated Features. RF-based XAI has been employed to achieve the stated objectives.

Motivation: As the legitimacy of the Tip-off in the form of a voice call depends highly on the mental state of the providers. It is our responsibility to help the LE

personnel by suggesting a smart XAI-based model, to assess the Genuineness of a voice call. This motivates us to pursue this work.

Objective: The objective is to build a smart system to assess the genuineness or trustworthiness of the Tip-off from a voice call based on Important MFCC(s) features employing XAI based approach.

Highlights: Following are some of the takeaways of the proposed work:

- Successfully able to assess the genuineness or trustworthiness of Tip-off by analyzing the provider's mental state from their voice sample.
- Employed RF-based XAI as a tool to find important MFCCs (26, or 13) from a set of 40 MFCCs
- Use of the Regional, international, and a combination of both audio sets
- Employed Conventional and Latest ML Models.
- For comparative analysis mean decreased based Gini and Permutation based approaches have been employed to find important MFCCs
- Execution Time for all the experimental works have been compared

The ***abbreviations*** used are Ar = Anger, Bm = Boredom, Cm = Clam, Dt = Disgust, Fr = Fear, Hy = Happy, Nl = Neutral, Sd = Sad, Su = Surprise; Perm = Permutation, CSL = Classical, LT = Latest, ML = Machine Learning, IMP = Important, CoD = EMODB + RAVDESS.

The rest of the paper is organized as follows: The background of the study has been done in Sect. 9.2, The methodology adopted has been discussed in Sect. 9.3 which is followed by the results and discussion section, Finally, the conclusion, limitation, and future scope of the study have been given.

9.2 Background

Nowadays Artificial Intelligence (AI) and Machine Learning are being used extensively, to assist the human decision-making process. These include some conventional ML models such as SVM, RF, kNN, and some advanced state of art approach-based ML models like CNN, DNN, LSTM, etc. Researchers are making use of these models because these models show good performance under different circumstances. With the advancement of cloud computing researchers have easier access to high-performance machine instances having higher throughput. Therefore the presence of ML-based solutions has been observed in various sectors of our life, such as audio and speech processing (Basharirad and Moradhaseli 1891; Ayadia et al. 2011; Poria et al. 2017; Pinto et al. 2020; Akçay and Oguz 2020; Yang et al. 2020; Chatterjee et al. 2021; Lalitha et al. 2014; Iqbal and Barua 2019; Shashidhar et al. 2018; Boles and Rad 2017; Panwar et al. 2017; Sinith et al. 2016), disorders during the growing time (Silva et al. 2020), image classification (Bendre et al. 2020), as well as in cyber-security (Parra et al. 2020). Pinto et al. (2020) proposed one 1D CNN model to govern the human emotional state using MFCCs as a feature set. The model has attained reasonable accuracy. Yang et al. (2020) suggested smart home assistance

using scaled MFCC as features to acquire the consumer's psychological state. In their work author(s) have engaged classical ML Models such as SVM, BPNN, ELM, etc. The proposed model has attained 92.4% accuracy. In the year 2021 Chatterjee et al. (2021) suggested one smart assistant system using 1D CNN. The suggested system used MFCC as an input to obtain higher accuracy. In their work 1D, CNN Model had been engaged. Lalitha et al. (2014) and Iqbal and Barua (2019) employed SER to determine human psychology in real-time, using both forms of the ML Model. In both works, they have employed MFCCs as a feature vector. Research works of Akçay and Oguz (2020), Shashidhar et al. (2018) suggested in detail the psychological models, dataset, classifier, pre-processing, and feature to be used in the SER system. At the same time, these models have given birth to too many queries because of a lack of interpretability (Gunning et al. 1973). XAI approaches have been used widely to mitigate some queries (McDermid 2021; Saarela and Jauhiainen 2021; Fisher et al. 2019). According to Bellotti (Bellotti 2009), there can be two levels of explainability. It can be either local or global or it could be Time based. The time-based interpretability can be further divided into three parts, Prior (what had been done), Contemporary (what is going on), and Post (what it has planned to do in the next). XAI methods help us to remove blind faith and bring transparency to the system (Zarsky 2016). It has been reported in various studies that transparency improves user awareness (Ananny and Crawford 2018), minimizes bias (Diakopoulos 2014), detects discrimination (Sweeney 2013), makes the user more accountable (Diakopoulos 2017), and helps to understand the functionalities of the intelligent system more in details (Lim and Dey 2009). Table 9.1 briefly describes some of the works carried out by the researchers.

9.3 Proposed Methodology

Figure 9.1 depicted the overall system in a nutshell.

9.3.1 Dataset Used

In the proposed work one regional language-based (EMO-DB), one international or mostly spoken language-based dataset (Livingstone et al. 2018), as well as a combined dataset combined these two datasets have been considered.

9.3.1.1 Regional Language-Based Audio Dataset

Berlin EMODB, a popularly used audio dataset used by the researchers to find participants' mental states. 5 Germans males and 5 Germans females have contributed to building the dataset. The dataset has seven emotion labels with a total size is 535. In

Table 9.1 A few studies to assess the mental state based on different parameters

Ref No (year)	Dataset language type	Employed features	Emotions labels used and count	Classifier used	Results
Lalitha et al. (2014)	Berlin EMODB Regional (German)	Frequency MFCCs, Scaled MFCCs,	Ar, Bm, Dt, Fr, Hy, Nl, Sd (7)	ANN	85.7%
Sinith et al. (2016)	Berlin EMODB Regional (German) SAVEE (English)	MFCCs, Pitch, Intensity	Ar, Hy, Nl, Sd (4)	SVM	Males: 67.5%, Females: 70%, Both: 75%,
Iqbal and Barua (2019)	RAVDESS SAVEE (English)	MFCCs, energy, spectral entropy	Ar, Dt, Fr, Hy, Nl, Sd, Sur (7)	GBM, SVM, KNN	Satisfactory
Yang et al. (2020)	Berlin EMODB Regional (German)	Scaled MFCCs	Ar, Hy, Nl, Sd (4)	SVM, BPNN, ELM, PNN	92.4%,77.8%, 7881%
Pinto et al. (2020)	RAVDESS (English)	MFCCs	Ar, Dt, Fr, Hy, Nl, Sd, Sur (7)	1D CNN	91%
Chatterjee et al. (2021)	RAVDESS, TESS (English)	MFCCs	Ar, Dt, Fr, Hy, Nl, Sd, Cm, Sur (8)	1D CNN	90.48%, 95.79%

the proposed work we have considered only four emotion labels with a total size of 337, Table 9.2.

9.3.1.2 International or Mostly Spoken Audio Dataset

It has been found in the literature that most of the researchers are using RAVDESS as their preferred dataset. In the dataset, there are eight emotion labels with a total size of 1440. In our proposed work we have only considered four emotion labels with a total size of 672. Twelve men and an equal number of women have contributed to building the dataset, Table 9.3.

9.3.1.3 Hybridization

This dataset has been obtained by combing regional audio datasets i.e. EMODB with international audio datasets i.e. RAVDESS, Table 9.4.

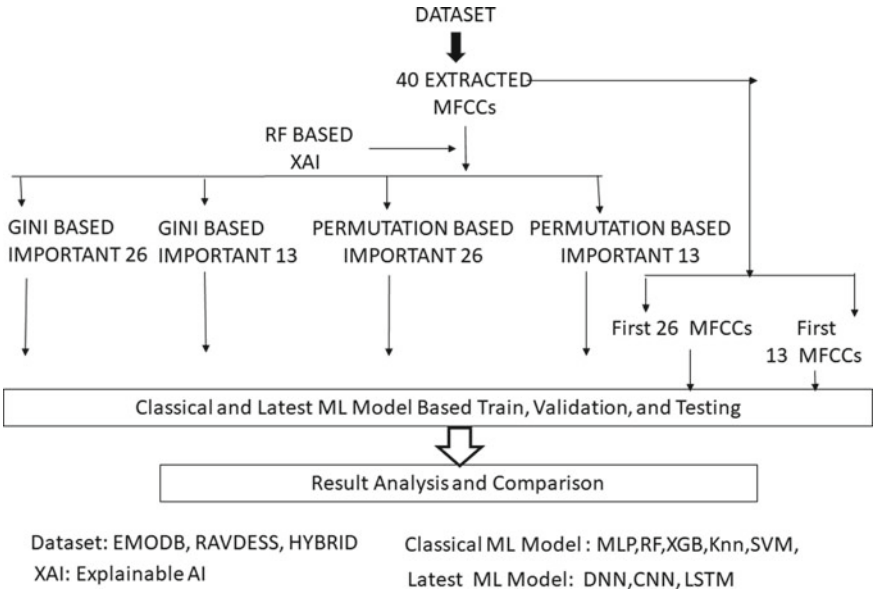

Fig. 9.1 Block diagram of proposed System

Table 9.2 Employed dataset and considered emotion labels. Here triple dots (…) de notes the emotion labels that have not been considered

Dataset	Emotion labels							
EMODB	*Ar*	Dt	Fr	*Hy*	*Nl*	*Sd*	Bm	Total
	127	48	69	69	79	62	61	535
	127	–	–	69	79	62	–	337

Table 9.3 Employed Dataset and considered emotion labels. Here triple dots (…) de notes the emotion labels that have not been considered

Dataset	Emotion labels								
RAVDESS	*Ar*	Dt	Fr	*Hy*	*Nl*	*Sd*	Su	Cm	Total
	192	192	192	192	96	192	192	192	1440
	192	–	–	192	96	192	–	–	672

Table 9.4 Hybridization of EMODB and RAVDESS

Dataset	Emotion labels				
CoD	Ar	Hy	Nl	Sd	Total
	319	261	175	254	1009

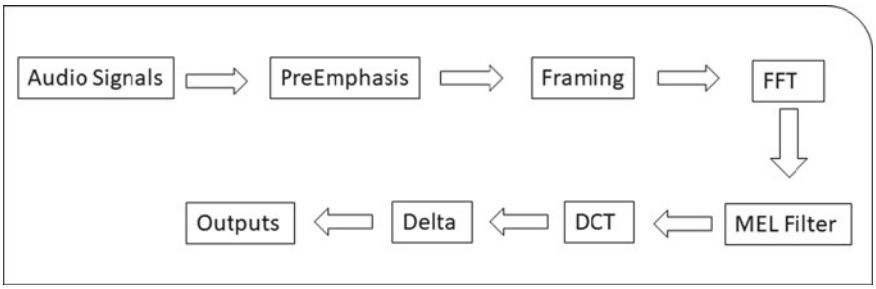

Fig. 9.2 MFCCS extraction process

9.3.2 Pre-processing

To make the system better and more robust pre-processing has been introduced. Through this process, we tried to increase the standard of voice or speech. Normalization, noise removal, trimming, etc. are some widely used pre-processing techniques.

9.3.3 Feature Extracted

It has been reported in various studies that MFCC is the most trusted audio feature among researchers (Yang et al. 2020; Chatterjee et al. 2021; Lalitha et al. 2014; Iqbal and Barua 2019). MFCC 40, 26, and 13 have been used widely to determine human mental state from speech or voice. Figure 9.2 shows the general process followed to extract MFCCs from an audio signal.

9.3.4 Feature Selected

It is quite obvious that to make the system smart we need to reduce the size of the number of features. There have been various ways we can reduce the size of the feature vector (Velliangiria et al. 2019). With the advancement of XAI, people are making use of the XAI approach to reduce the number of features and included those features having more impact on the classification process.

The Mean Decrease in Impurity based on important features can be obtained by calculating the node probability using the following equation:

$$n_{ij} = w_j C_j - wl(j)Cl(j) - wr(j)Cr(j) \tag{9.1}$$

Here, n_{ij} is the importance of node j, wj is the weighted number of samples reaching node j, Cj represents the impurity value of node j, l(j) is the child node on the left of node j, r(j) is the child node on right of node *j*.

The Permutation Importance with Correlated Features can be obtained using the following equation below

$$i_j = s - \frac{1}{K}\sum_{k=1}^{k} s_{kj} \tag{9.2}$$

Where i_j the permutation of importance, S is the reference score of the employed RF model, J is the feature to be evaluated, K number of iterations, s_{kj} is the computed score based on the employed model.

9.3.5 Machine Learning in SER

It has been observed that Machine learning algorithms have been used widely to determine a human mental state from his/her voice or speech. It can broadly be categorized into the following two categories.

9.3.5.1 Classical ML Models

In the proposed work some of the widely used ML models such as Multilayer perceptron with the feed-forward network, Random forest, XGBoosing, K nearest Neighbour, and support vector machine have been engaged.

9.3.5.2 Latest ML Models

With the advisement of new technology, storage capacity, and advanced processing systems people have shifted their focus from the Conventional ML Approach to these new approaches. Alex Net, RNN, CNN, AutoML, etc. have been used extensively to solve the real-time problem.

9.3.6 Performance Index

Judging the performance of the employed model is one of the required parameters in all classification processes. In the proposed work following yardsticks have been engaged.

$$Precision = \frac{tp}{tp + fp} \tag{9.3}$$

$$Recall = \frac{tp}{tp + fn} \tag{9.4}$$

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \tag{9.5}$$

Where tp = True Positive, fp = False Positive, tn = True Negative, fn = False Negative.

9.4 Results and Discussion

Detail experiments have been conducted to test the impact of important features to determine the Tip-provider's mental state. Doing so will help the LE personnel to protect citizens from possible threats or untoward incidents To satisfy the objective, two popularly used audio/speech datasets as well as one dataset, combining these two datasets have been produced and employed. Scikit-learn Python library has been employed for all ML-based experimental processes. It is also used to find the important features. Google Colaboratory (Colab) has been used as an environment. For CSL-based ML, the underline is used to indicate the best accuracy while bold font represents the same for LT-based ML. Table 9.5 illustrates the parameter of the RF-based approach to find the important features.

9.4.1 Figure 9.3 shows the outline view of the employed DNN model while Figs. 9.4 and 9.5 show the architecture of the employed CNN and LSTM models. Table 9.6 tried to analyze the performance of CSL and LT models employing 40 MFCCs as a feature vector. Tables 9.7, 9.8, 9.9, 9.10, 9.11 and 9.12, tabulated the findings of CSL and LT ML models. In all such experimental findings, different MFCCs feature vector sizes such as 26 and 13 in sequence; MDI, and permutation-based important 26 and 13

Table 9.5 Parameters used to find important MFCCs using the XAI approach

Parameter name	Value	Parameter name	Value
Bootstrap	True	class_weight	None
Criterion	Gini	max_depth	None
max_features	Auto	max_leaf_nodes	None
min_impurity_decrease	0.0	min_impurity_split	None
min_samples_leaf	1	min_samples_split	2
min_weight_fraction_leaf	0.0	n_estimators	100
n_jobs	None	oob_score	False
random_state	0	Warm_start	False

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	896
dense_1 (Dense)	(None, 128)	8320
dense_2 (Dense)	(None, 256)	33024
dense_3 (Dense)	(None, 4)	1028

Total params: 43,268
Trainable params: 43,268
Non-trainable params: 0

Fig. 9.3 The architecture of the employed DNN model for the proposed study

have been engaged. These MFCCs have been extracted from EMODB, RAVDESS, and combined datasets. Tables 9.13, 9.14 and 9.15 shows the comparative analysis of the findings. Tables 9.16, 9.17 and 9.18 demonstrate the time taken to execute different models for different MFCCs in the employed datasets. Figures 9.6, 9.7, 9.8, 9.9, 9.10, 9.11, 9.12, 9.13, 9.14, 9.15, 9.16 and 9.17 show the selected 26 and 13 MFCCs, obtained using MDI, and Permutation based approach from the employed dataset. Figures 9.18, 9.19 and 9.20 display the confusion Matrix of adopted 1D CNN while Fig. 9.21 shows the accuracy and loss for the adopted 1D CNN models. Table 9.19 shows the performance of the adopted 1D CNN model based on the benchmark score.

9.4.2 Based on Table 9.6, it can be recorded that in the case of EMODB, MLP has shown its supremacy over other CSL models employed for this study. For the other two datasets, RF is comparatively better. In the case of LT-based ML models, CNN has a clear mandate over the other two. The result of Table 9.7 shows that for CSL models, the important 26 and 13 MFCCs obtained using MDI based approach are the most appropriate since both of them have achieved 95.59% accuracy. The kNN and MLP are the most suitable CSL ML models. CNN is the best LT-based ML model for the proposed work, Table 9.8. MFCCs 13 in sequence and MDI-based important 13 MFCCs have shown their supremacy over the rest. Both have achieved 98.88% accuracy using CNN.

9.4.3 Based on Table 9.9 it can be recorded that among the CSL models, MLP is the most suitable ML model. The model has achieved the best accuracy (71.53%) and MDI-based Important 26 MFCCs are the most appropriate feature set compared to other MFCCs involved. The MLP-based model also shows its supremacy. It has achieved 72.99% accuracy also MDI based MFCCs 13 is the most suitable feature vector. In the case of LT-based ML models, Table 9.10, CNN has a clear mandate

```
Model: "CNN"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 Conv1 (Conv1D)               (None, 26, 32)            192

 Activation1 (Activation)     (None, 26, 32)            0

 Dropout1 (Dropout)           (None, 26, 32)            0

 Conv2 (Conv1D)               (None, 26, 32)            5152

 Activation2 (Activation)     (None, 26, 32)            0

 Dropout2 (Dropout)           (None, 26, 32)            0

 Conv3 (Conv1D)               (None, 26, 64)            10304

 Activation3 (Activation)     (None, 26, 64)            0

 Dropout3 (Dropout)           (None, 26, 64)            0

 Flat1 (Flatten)              (None, 1664)              0

 Dense1 (Dense)               (None, 4)                 6660

 Activation4 (Activation)     (None, 4)                 0

=================================================================
Total params: 22,308
Trainable params: 22,308
Non-trainable params: 0
```

Fig. 9.4 The architecture of the employed CNN model for the proposed study

over the other two. The result of Table 9.9 also shows that MDI-based important 26 MFCCs (accuracy 78.11%) and MDI-based important 13 MFCCs (accuracy 74.72) are the most suitable feature sets among the feature sets employed for this study.

9.4.4 Table 9.11 shows that among the employed Conventional ML models RF is the most appropriate. The RF-based models have achieved 76.60% and 73.68% accuracy for MFCCs 26 and MFCCs 13 respectively. The MDI-based important approach also shows its supremacy. Table 9.12 again shows the supremacy of 1D CNN among the employed LT models. It has achieved 83.46 and 81.18% of accuracy for MFCCs 26 and MFCCs 13 respectively. The MDI-based important approach also shows its supremacy.

9.4.5 Table 9.12 shows that for MFCCs 40, both CSL and LT-based ML models' best accuracy value achieved is 94.12% and 98.52% respectively, and the employed

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 lstm (LSTM)                  (None, 13, 64)            16896

 activation (Activation)      (None, 13, 64)            0

 dropout (Dropout)            (None, 13, 64)            0

 flatten (Flatten)            (None, 832)               0

 dense_4 (Dense)              (None, 4)                 3332

 activation_1 (Activation)    (None, 4)                 0

=================================================================
Total params: 20,228
Trainable params: 20,228
Non-trainable params: 0
```

Fig. 9.5 The architecture of the employed LSTM model for the proposed study

Table 9.6 Evaluation of results by employing CSL and LT algorithms in terms of **accuracy (%)** based on 40 MFCCs, datasets employed are EMODB, RAVDESS, and COMBINED

Employing 40 MFCCs Accuracy (%)								
Dataset	CSL ML models					LT ML models		
	MLP	RF	XGB	kNN	SVM	DNN	CNN	LSTM
EMODB	*94.12*	92.71	91.18	92.65	91.18	95.50	**98.52**	92.64
RAVDESS	69.34	*70.80*	68.61	68.61	66.42	69.89	**75.91**	74.45
CoD	73.66	*75.12*	73.66	75.61	71.71	78.61	**80.00**	78.07

Table 9.7 Evaluation of result by employing CSL algorithms in terms of accuracy (%) based on different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem) dataset engaged EMODB

CSL ML models	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
MLP	94.12	94.18	94.12	94.12	*95.59*	94.18
RF	89.71	89.24	89.71	89.66	92.65	89.24
XGB	92.42	92.65	92.65	92.02	94.12	92.58
kNN	92.65	*95.59*	94.12	91.18	92.65	91.18
SVM	91.18	92.65	94.12	86.76	89.71	89.71

Table 9.8 Evaluation of result by employing LT algorithms in terms of accuracy (%) based on different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem), dataset engaged EMODB

LT ML models	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
DNN	92.65	97.06	97.06	98.53	98.53	97.95
CNN	97.05	**98.52**	**98.52**	98.80	**98.88**	98.82
LSTM	96.05	**98.52**	97.05	95.58	95.58	95.58

Table 9.9 Evaluation of result by employing CSL algorithms in terms of accuracy, based on different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem), dataset engaged RAVDESS

CSL ML models	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
MLP	68.61	*71.53*	70.07	69.14	*72.99*	69.20
RF	69.34	70.80	68.61	70.45	70.51	*70.58*
XGB	66.15	66.42	66.17	62.04	65.69	64.96
kNN	65.69	65.77	65.75	64.03	64.20	64.13
SVM	65.69	66.15	66.02	66.40	66.42	65.98

Table 9.10 Evaluation of result by employing LT algorithms in terms of accuracy based on different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem), dataset engaged RAVDESS

LT ML Models	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
DNN	75.18	75.72	76.64	72.26	72.53	71.88
CNN	78.10	**78.11**	77.64	74.56	**74.72**	74.38
LSTM	73.72	75.91	73.72	72.86	73.99	72.96

dataset is EMODB. Table 9.13 displays that the highest accuracy achieved is 95.59% using EMODB as a dataset. MFCCs involved are MDI-based IMP-26 and IMP-13 respectively. For RAVDESS best accuracy achieved is 72.99 and MFCCs involved are MDI-based IMP-13. Finally, for the combined dataset best accuracy achieved is 76.60 and MFCCs involved are MDI-based IMP-26. Table 9.14 demonstrates that for EMODB best accuracy achieved is 98.88 and the MFCCs involved are MDI-based

Table 9.11 Evaluation of result by employing CSL algorithms in terms of accuracy based on different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem), dataset engaged combined

CSL ML models	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
MLP	71.24	74.63	74.25	70.73	71.19	68.34
RF	76.59	*76.60*	76.58	73.66	*73.68*	70.73
XGB	71.22	71.42	71.32	66.83	66.83	69.27
kNN	72.68	73.66	72.68	69.27	69.76	66.34
SVM	70.24	70.26	70.24	72.68	72.82	70.24

Table 9.12 Evaluation of result by employing LT algorithms in terms of accuracy based on different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem), dataset engaged combined

LT ML models	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
DNN	78.05	79.02	78.21	79.89	80.08	77.10
CNN	83.41	**83.46**	81.51	81.12	**81.18**	79.53
LSTM	82.43	81.97	81.54	77.07	77.34	74.18

Table 9.13 Comparative Findings of 40 MFCCs, using CSL and LT Models based on the best accuracy (%) score achieved

Dataset	Accuracy (%) MFCCs 40	
	Classical ML models accuracy (%)	Latest ML models accuracy (%)
EMODB	*94.12*	**98.52**
RAVDESS	70.80	75.91
CoD	75.61	80.00

Table 9.14 Comparative findings of different MFCCs(First26 and 13 or Important 26 and 13 using MDI and Prem) based on the best accuracy(%) obtained, by employing CSL Model, engaged dataset EMODB, RAVDESS, and combined

Dataset	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
EMODB	94.12	*95.59*	94.12	94.12	*95.59*	91.18
RAVDESS	69.34	71.53	70.07	70.45	*72.99*	70.58
CoD	76.59	*76.60*	76.58	73.66	73.68	70.73

Table 9.15 Comparative findings of different MFCCs(First26 and 13 or Important 26 and 13 using MDI and Prem) based on the best accuracy(%) obtained, employed Model (LT), engaged dataset EMODB, RAVDESS, and combined

Dataset	Accuracy (%)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
EMODB	97.05	98.52	98.52	98.80	**98.88**	98.52
RAVDESS	78.10	**78.11**	77.64	74.72	74.72	74.38
CoD	83.41	**83.46**	81.54	81.12	81.18	79.53

Table 9.16 Comparative findings of the execution time of the CSL model based on the maximum time taken

MFCCs 40		
Dataset	CSL ML models Time in seconds	LT ML models Time in seconds
EMODB	0.0180	83.13
RAVDESS	0.0294	145.89
CoD	*0.0330*	**213.45**

Table 9.17 Comparative findings of the execution time using different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem) of the CSL Models based on the maximum time taken

Dataset	Execution time in seconds					
	IMP-26 MFCCs (MDI)	26 MFCCs	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
EMODB	0.0120	0.0120	0.0120	0.0040	0.0041	0.0040
RAVDESS	0.0173	0.0172	0.0170	0.0018	0.0018	0.0018
CoD	0.0290	0.0300	0.0280	0.0230	0.0230	0.0230

Table 9.18 Comparative findings of the execution time using different MFCCs (First26 and 13 or Important 26 and 13 using MDI and Prem) of the LT Model based on the maximum time taken

Dataset	Execution time (in seconds)					
	26 MFCCs	IMP-26 MFCCs (MDI)	IMP-26 MFCCs (Perm)	13 MFCCs	IMP-13 MFCCs (MDI)	IMP-13 MFCCs (Perm)
EMODB	48.66	48.73	45.23	39.41	39.46	39.71
RAVDESS	52.23	52.23	52.11	49.08	49.08	49.90
CoD	147.59	147.85	147.12	103.45	104.22	104.00

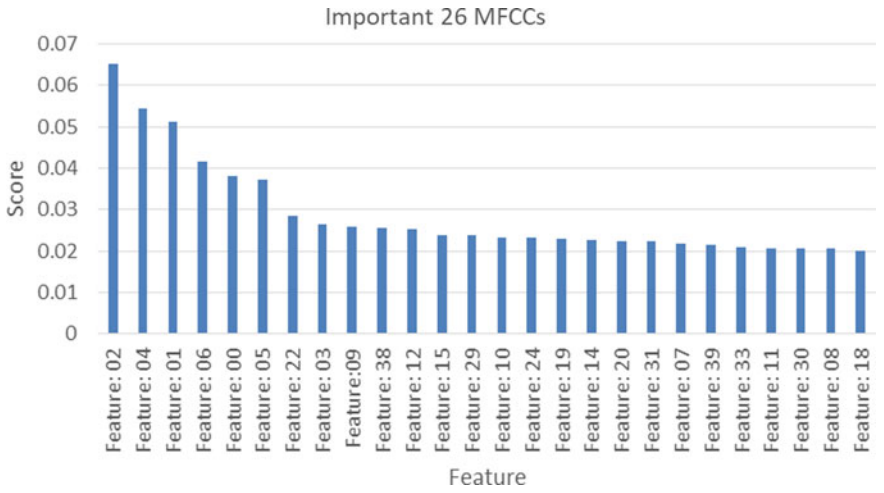

Fig. 9.6 Selected 26 MFCCs according to their importance using the MDI approach. Dataset employed EMODB

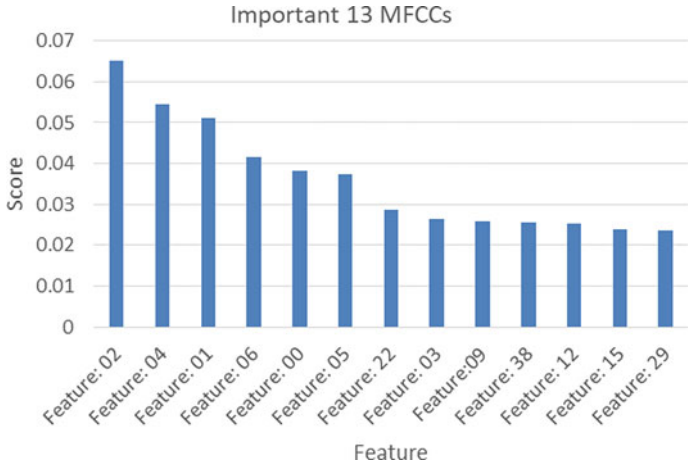

Fig. 9.7 Selected 13 MFCCs according to their importance using the MDI approach. Dataset employed EMODB

IMP-13. For RAVDESS and Combined best accuracy achieved is 78.11 and 83.46 respectively. In both cases, MFCCs involved are MDI-based IMP-26.

9.4.6 Tables 9.16, 9.17 and 9.18 shows the time taken to execute increases as the size of the dataset increases, as well as the number of MFCCs increases. It also shows that the execution time of LT-based ML is much higher compared to CSL ML models.

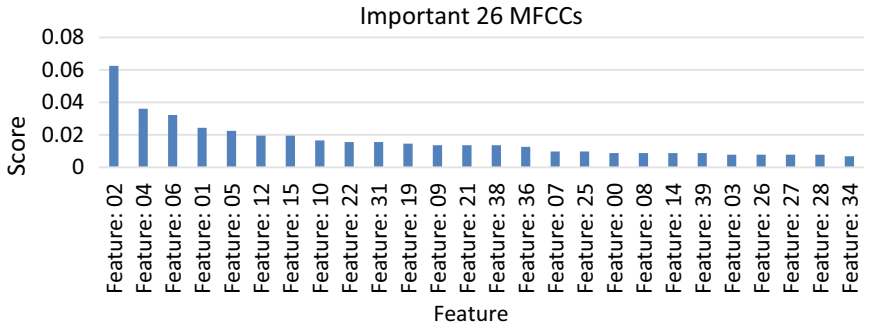

Fig. 9.8 Selected 26 MFCCs according to their importance using the Permutation approach. Dataset employed EMODB

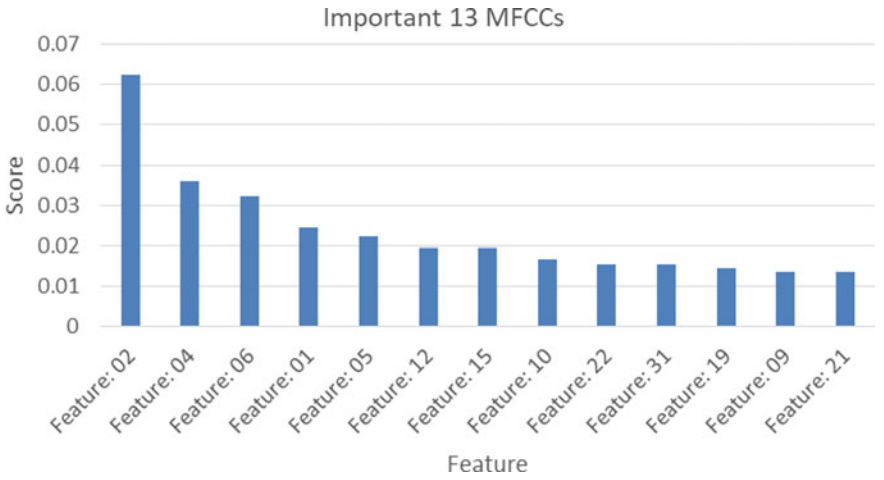

Fig. 9.9 Selected 13 MFCCs according to their importance using the Permutation approach. Dataset employed EMODB

9.5 Conclusion

The result analysis shows the supremacy of XAI based (MDI, Permutation) approach over the traditional approach by selecting the important MFCCs. This helps to make the system better and smarter. It also shows that among the employed ML models 1D CNN-based model achieved the best accuracy. For the regional dataset 1D, CNN achieved 98.88% accuracy, for the International dataset it has achieved 78.11% accuracy, and for the combined dataset 83.46%accuracy has been achieved. It further shows that MDI based approach outperforms the permutation-based approach. The LT-based ML models show better performance though they are a bit expensive in

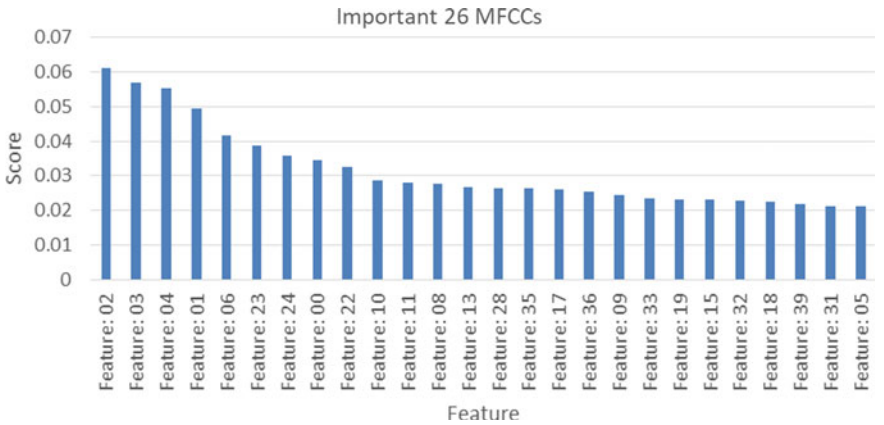

Fig. 9.10 Selected 26 MFCCs according to their importance using the MDI approach. Dataset employed RAVDESS

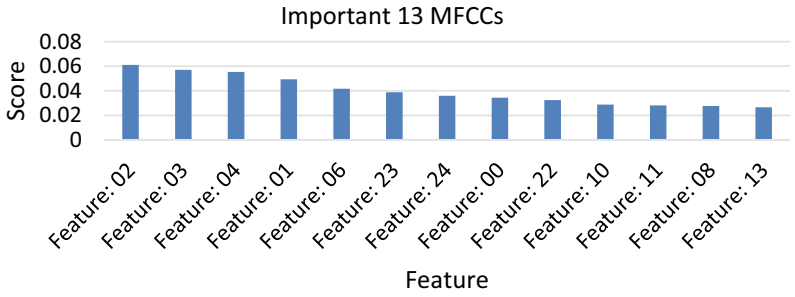

Fig. 9.11 Selected 13 MFCCs according to their importance using the MDI approach. Dataset employed RAVDESS

terms of execution times. Thus, a 1D CNN model has been proposed, where important features are extracted and used employing XAI based (MDI) approach.

Limitation and Future Scope: This study has a few limitations. Firstly as a feature, only MFCC(s) have been considered. Secondly, we have only considered two audio datasets and one combined dataset. The inclusion of a more diversified audio dataset may make the system more robust and better. Important MFCCs have been obtained using the RF algorithm. Thus for comparative analysis, other algorithms can be employed in the future for selecting the important MFCCs.

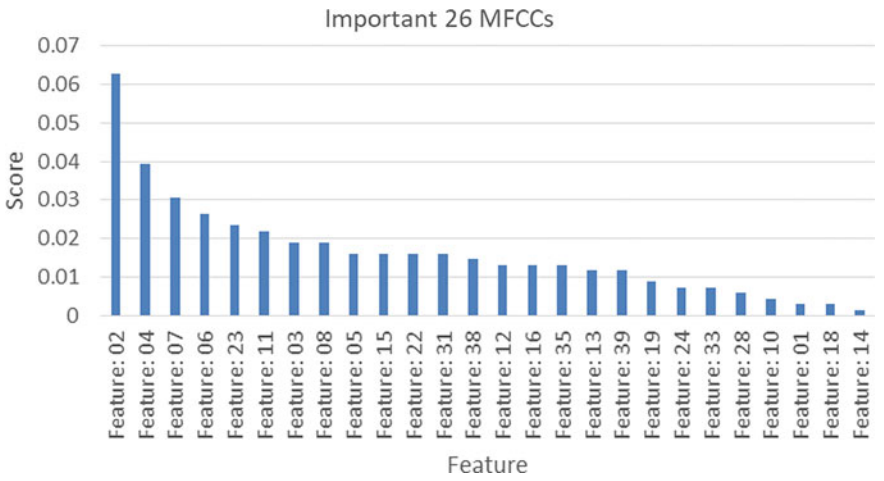

Fig. 9.12 Selected 26 MFCCs according to their importance using the Permutation approach. Dataset employed RAVDESS

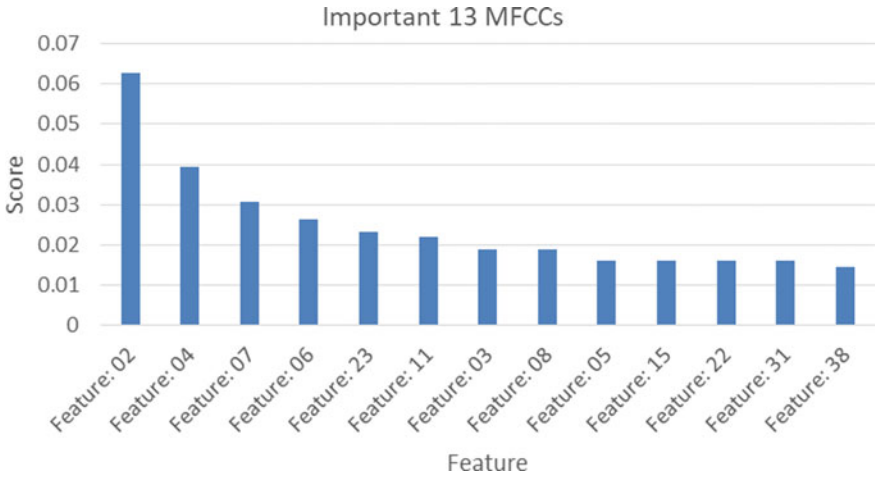

Fig. 9.13 Selected 13 MFCCs according to their importance using the Permutation approach. Dataset employed RAVDESS

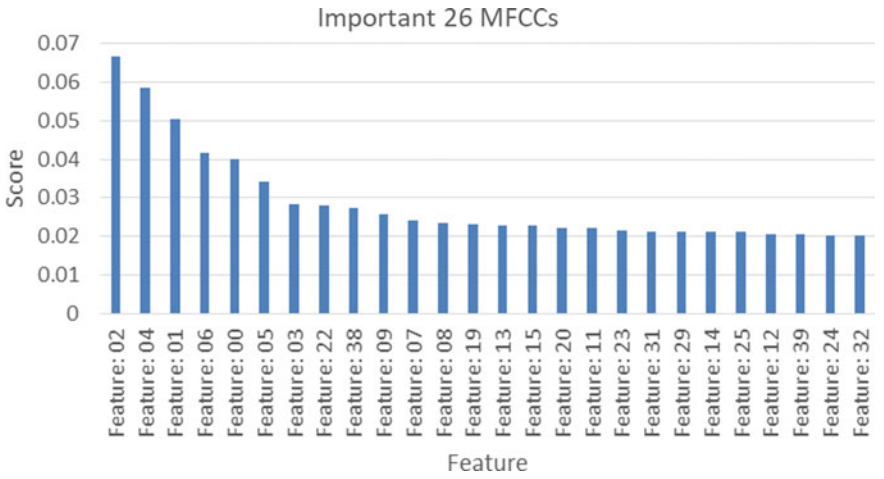

Fig. 9.14 Selected 26 MFCCs according to their importance using the MDI approach. Dataset employed combined

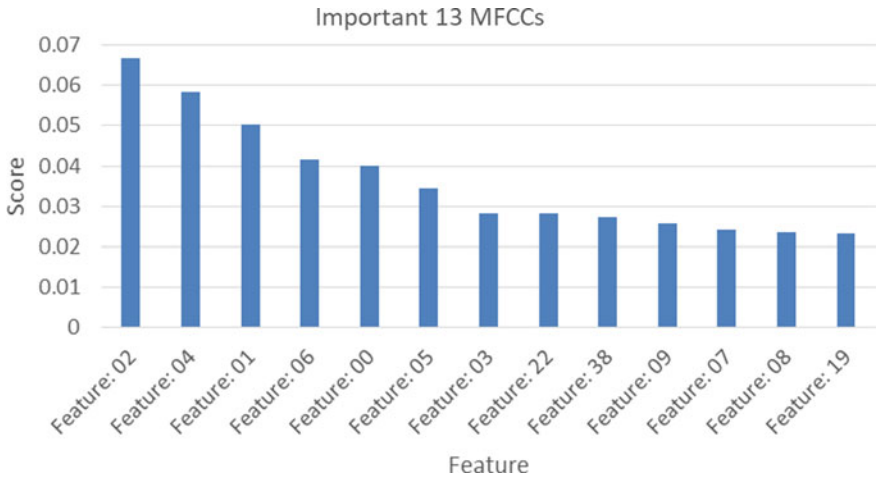

Fig. 9.15 Selected 13 MFCCs according to their importance using the MDI approach. Dataset employed combined

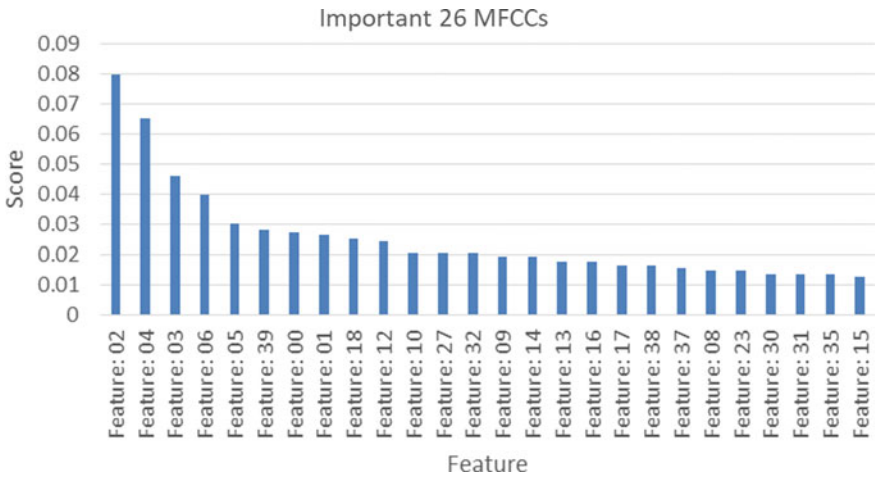

Fig. 9.16 Selected 26 MFCCs according to their importance using the Permutation approach. Dataset employed combined

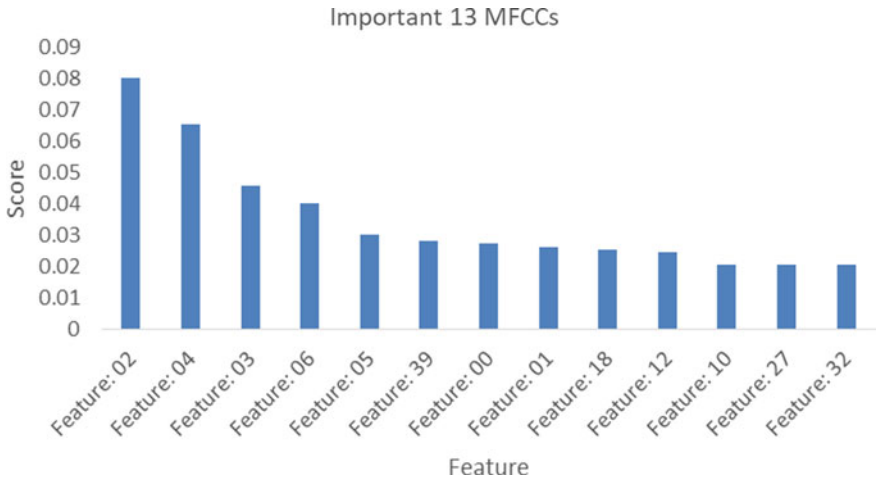

Fig. 9.17 Selected 13 MFCCs according to their importance using the Permutation approach. Dataset employed combined

Fig. 9.18 Confusion Matrix of adopted 1D CNN. 26 Important MFCCs selected using MDI based approach, Dataset employed RAVDESS

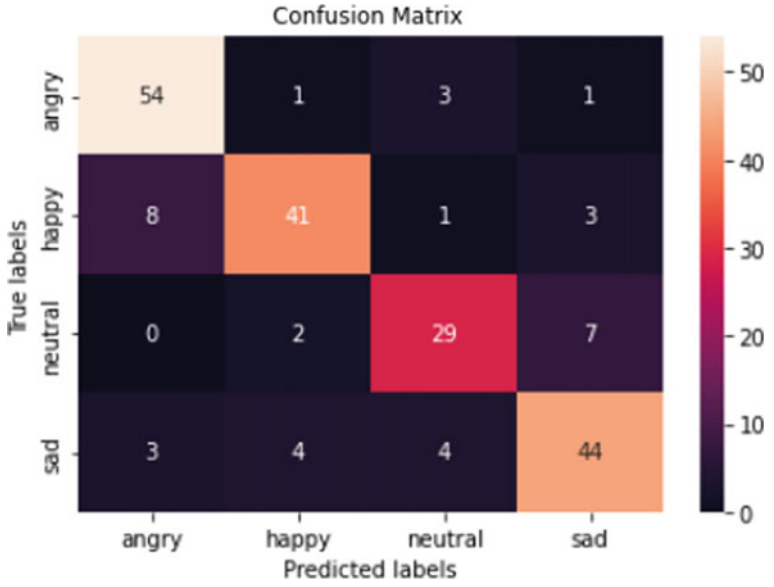

Fig. 9.19 Confusion Matrix of adopted 1D CNN 26 Important MFCCs selected using MDI based approach, Dataset employed combined

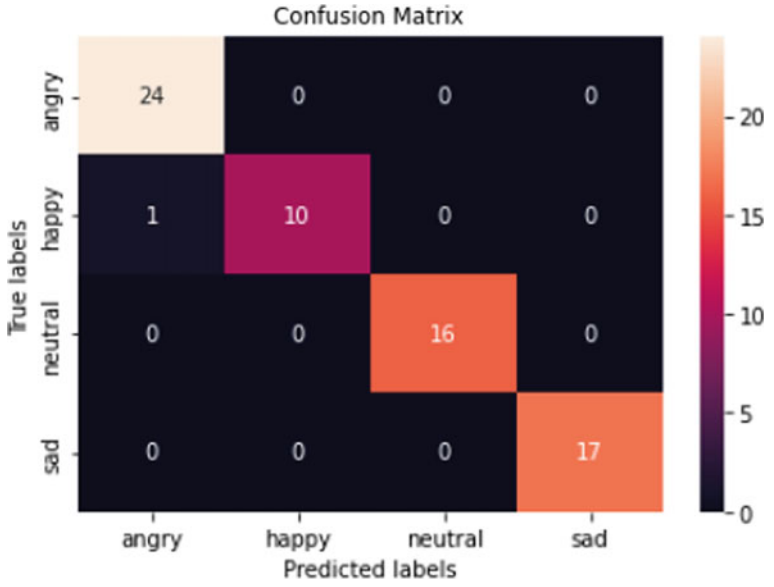

Fig. 9.20 Confusion Matrix of adopted 1D CNN. 13 Important MFCCs selected using MDI based approach, Dataset employed EMODB

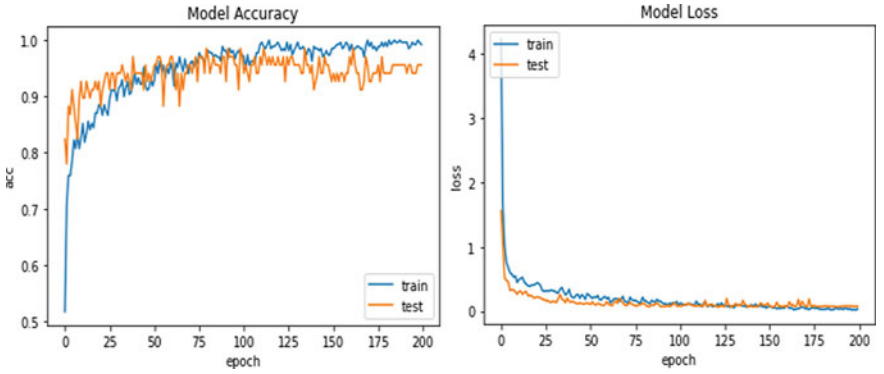

Fig. 9.21 Model accuracy and Model Loss of adopted 1D CNN. 13 Important MFCCs selected using MDI based approach, Dataset employed EMODB

Table 9.19 Evaluation of result based on employed 1D CNN using three benchmark scores against FOUR mental health states. Feature vector employed important 13 MFCCs MDI based

Mental state	Benchmark score			
	Precision	Recall	F1 Score	Support
Angry	0.96	1.00	0.98	24
Happy	1.00	0.91	0.95	11
Neutral	1.00	1.00	1.00	16
Sad	1.00	1.00	1.00	17
Accuracy	–	–	**0.99**	**68**
Macro avg	0.99	0.98	**0.98**	**68**
Weighted avg	0.99	0.99	**0.99**	**68**

References

Akçay, M.B., Oguz, K.: Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. **116**, 56–76 (2020). https://doi.org/10.1016/j.specom.2019.12.001

Ananny, M., Crawford, K.: Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media and Soc. **20**, 3 973–989 (2018)

Ayadia, E.M., Kamel, S., M, Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit. **44**, 572–587 (2011). https://doi.org/10.1016/j.patcog.2010.09.020

Basharirad, B., Moradhaseli, M.: Speech Emotion Recognition Methods: A Literature Review. In: AIP Conference Proceedings vol. 1891, pp. 020105. (2017).https://doi.org/10.1063/1.5005438

Bellotti, K.: Edwards: Intelligibility and accountability: human considerations in context-aware systems. Hum. Comput. Interact. **16**, 193–212 (2009)

Bendre, N., Ebadi, N. Prevost, J.J., Najafirad, P.: Human action performance using deep neuro-fuzzy recurrent attention model. IEEE Access **8**, 57 749–57 761 (2020)

Boles, A., Rad, P.: Voice biometrics: deep learning-based voiceprint authentication system. In: 12th System of Systems Engineering Conference (SoSE), pp. 1–6. IEEE. (2017).

Chatterjee, R., Majumder, S., Sherratt, R.S., Halder, R., Maitra, T., Giri, D.: Real-time speech emotion analysis for smart home assistants. IEEE Trans Consum Electronics **67**(1), 68–76 (2021). https://doi.org/10.1109/TCE.2021.3056421

Diakopoulos, N.: Algorithmic-accountability: the investigation of black boxes. Tow Cent. Digit. JIsm. (2014).

Diakopoulos, N.: Enabling accountability of algorithmic media: transparency as a constructive and critical lens. In: Transparent Data Mining for Big and Small Data, pp. 25–43. Springe. (2017)

EMO-DB: Berlin Database of Emotional Speech, [Online]. 671. http://emodb.bilderbar.info/start.html

Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn Res. **20**(177), 1–81 (2019)

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. ORCID: 0000-0001-6482-1973 and Yang, G-Z.: XAI-Explainable artificial intelligence. Sci. Robot **4**(37), eaay7120, (2019). https://doi.org/10.1126/scirobotics.aay7120V

https://dictionary.cambridge.org/dictionary/english/tip-off.

https://www.criminallawyersandiego.com/crimes-police-government/false-report/.

Iqbal, A., Barua, K.: A real-time emotion recognition from speech using gradient boosting. In: International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 1–5. IEEE (2019)

Koolagudil, S.G., Srinivasa Murthy1, Y.V., Bhaskar1, S.P.: Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. Int. J. Speech Technol. (2018). https://doi.org/10.1007/s10772-018-9495-8

Lalitha, S., Madhavan, A., Bhushan, B., Saketh, S.: Speech emotion recognition. In: Proceedings of the International Conference on Advances in Electronics, Computers and Communications, ICAECC 2014, pp. 1–4. IEEE (2015b). http://doi.org/https://doi.org/10.1109/ICAECC.2014.7002390

Lim, B.Y., Dey, A.K.: Assessing demand for intelligibility in context-aware applications. In: Proceedings of the 11th International Conference on Ubiquitous Computing, pp. 195–204. ACM (2009)

Livingstone, S.R., Thompson, W.F., Wanderley, M.M., Palmer, C.: Common cues to emotion in the dynamic facial expressions of speech and song. Q. J. Exp. Psychol. 1–19 (2018). https://doi.org/10.1371/journal.pone.0196391

McDermid, J.A., Jia, Y., Porter, Z., Habli, I.: Artificial intelligence explainability: the technical and ethical dimensions. Phil. Trans. R. Soc. A **379**, 20200363 (2021). https://doi.org/10.1098/rsta.2020.0363

Panwar, S., Das, A., Roopaei, M., Rad, P.: A deep learning approach for mapping music genres. In: 12th System of Systems Engineering Conference (SoSE) , pp. 1–5. IEEE. (2017)

Parra, G.D.L.T., Rad, P., Choo, K.-K.R., Beebe, N.: Detecting internet of things attacks using distributed deep learning. J. Netw. Comput. Appl. 102662 (2020)

Pinto, MGD. Polignano, M., Lops, P., Semeraro, G.: Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients. In: EAIS, IEEE (2020). https://doi.org /978-1-7281-4384-22020

Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: from unimodal analysis to multimodal fusion. Inf. Fusion **37**, 98–125 (2017). https://doi.org/10.1016/j.inffus.2017.02.003

Saarela, M., Jauhiainen, S.: Comparison of feature importance measures as explanations for classification models. SN Appl. Sci. **3**, 272 (2021). https://doi.org/10.1007/s42452-021-04148-9

Silva, S.H., Alaeddini, A., Najafirad, P.: Temporal graph traversals using reinforcement learning with proximal policy optimization. IEEE Access, **8**, 63 910 (2020)

Sinith, M.S., Aswathi, E., Deepa, T.M., Shameema, C.P., Rajan, S., 2016. Emotion recognition from audio signals using support vector machine. In: Proceedings of the IEEE Recent Advances in Intelligent Computational Systems, RAICS, pp. 139–144. IEEE. (2015). http://doi.org/https://doi.org/10.1109/RAICS.2015.7488403

Sweeney, L.: Discrimination in online ad delivery. Commun. ACM **56**(5), 44–54 (2013).

Velliangiria, S., Alagumuthukrishnan, S., Iwin, S., Joseph, T.: A review of dimensionality reduction techniques for efficient computation. Procedia Comput. Sci. **165**, 104–111 (2019). https://doi.org/10.1016/j.procs.2020.01.079

Yang, N., Dey, N., Sherratt, S., Shi, F.: Emotional state recognition for AI smart home assistants using Mel-frequency Cepstral coefficient features. J. Intell. Fuzzy Syst. **39**(2), 1925–1936 (2020). ISSN 1875–8967 (E)

Zarsky, T.: The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making. Sci. Technol. Human Values **41**(1), 118–132 (2016)