UDC 004.934.032.26 **doi: 10.32620/reks.2022.4.13**

**Sumon Kumar HAZRA, Romana Rahman EMA*, Syed Md. GALIB, Shalauddin KABIR, Nasim ADNAN**

***Department of Computer Science and Engineering, Jashore University of Science and Technology, Bangladesh***

# EMOTION RECOGNITION OF HUMAN SPEECH USING DEEP LEARNING METHOD AND MFCC FEATURES

***Subject matter:*** *Speech emotion recognition (SER) is an ongoing interesting research topic. Its purpose is to establish interactions between humans and computers through speech and emotion. To recognize speech emotions, five deep learning models: Convolution Neural Network, Long-Short Term Memory, Artificial Neural Network, Multi-Layer Perceptron, Merged CNN, and LSTM Network (CNN-LSTM) are used in this paper. The Toronto Emotional Speech Set (TESS), Surrey Audio-Visual Expressed Emotion (SAVEE) and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) datasets were used for this system. They were trained by merging 3 ways TESS+SAVEE, TESS+RAVDESS, and TESS+SAVEE+RAVDESS. These datasets are numerous audios spoken by both male and female speakers of the English language. This paper classifies seven emotions (sadness, happiness, anger, fear, disgust, neutral, and surprise) that is a challenge to identify seven emotions for both male and female data. Whereas most have worked with male-only or female-only speech and both male-female datasets have found low accuracy in emotion detection tasks. Features need to be extracted by a feature extraction technique to train a deep-learning model on audio data. Mel Frequency Cepstral Coefficients (MFCCs) extract all the necessary features from the audio data for speech emotion classification. After training five models with three datasets, the best accuracy of 84.35 % is achieved by CNN-LSTM with the TESS+SAVEE dataset.*

***Keywords:*** *speech emotion recognition (SER); deep learning method; advanced AI; mel frequency cepstral coefficients (MFCCs); audio data.*

## Introduction

Speech is the main means of communication used by people in the world and people express their emotions through speech. Emotions have long been studied in physiology and psychology [1]. Speech can express people's feelings. A human command will be better understood by an agent if it can explain a human's mental state [2]. The emotional speech reflects the mood and intention of the speaker [3]. It plays a very important role in real-life communication [3] Currently, more and more people are communicating through virtual voice assistants like Siri, Cortana, Alexa and Google Assistant. Interface performance is unreliable because it does not accurately understand human emotions [4]. Emotions are of different types: happiness, sadness, fear, surprise, anger, annoyance, neutral, etc. There are all kinds of people, men and women. Emotional language is also different for people of different languages. With them in mind, emotion recognition in speech is a challenging task for both men and women. And the more emotions that are identified, the more difficult the task becomes.

Interaction between humans and computers can improve the understanding of emotional states in human speech and distinguish different emotions from the same speech [5]. It will be more interactive [5]. Speech signals are also usually variable due to speaker characteristics, uncertain environmental noise, speaking style, etc. Therefore, speech emotion recognition is a challenging task [6].

If the machine can recognize emotions, emotional tasks in the real world will be easier and less time-consuming. Such as business marketing, personalized movie player or music player, suicide prevention, etc. [7]. For example, in business marketing, voice identifies product reviews as positive (customer happiness) or negative (customer sadness). This task is difficult and time-consuming for many products. So doing it by machine will be easy and save time.


Fig. 1. Speech emotion recognition

(Fig.1) simply shows the speech emotion recognition process. Where speech will be an input and it is classified by a deep learning model (for this system) and predicts an emotion.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Deep learning algorithms are used with audio speech data and trained on it. Here, five deep learning models (CNN, ANN, MLP, LSTM, CNN-LSTM) are applied to three datasets (TESS + SAVEE, TESS + RAVDESS, TESS + SAVEE + RAVDESS). MFCC technique is used for feature extraction and softmax is used as an activation function. The best model is recommended for the SER system.

SER systems can also be implemented for real-world use. The most common emotions are classified here. The seven emotions classified in the SER system are happiness, sadness, fear, disgust, surprise, anger, and neutrality. These seven emotions were identified for both male and female English language data in this work, which was a challenge.

The following sections of this paper are: section 1 describes related work (where previous works on SER systems are described), section 2 describes preliminaries (where speech emotion fundamentals, feature extraction and models are described), section 3 describes the methodology (how to build the speech emotion recognition system), results and analysis (model training results and their comparison for different models) are described in section 4 and conclusion (final comments about the SER system) describes. Its future scope is also discussed here. These sections comprise the entire SER system.

## 1. Related Work

Speech Emotion Recognition is the most wanted research topic nowadays. Researchers have tried various methods for speech emotion recognition but recently emotion recognition by analyzing speech and language with the help of deep neural networks has become the most popular [8]. Some of the previous tasks are:

Harshit Dolka et al. [9] showed the SER system on 4 datasets by training the ANN model. There TESS, SAVEE, CREMA-D, and RAVDESS datasets were trained with ANN models separately. ANN is used instead of CNN due to CNN training taking more time. RAVDESS and CREMA-D were trained with 6 emotions for the dataset and 7 emotions for the other datasets. An accuracy of 99.52 % was obtained for the TESS dataset, 88.72 % for the RAVDESS dataset, 71.69 % for the CREMA-D dataset, and 86.80 % for the SAVEE dataset. MFCC was used for feature extraction. As an activation function, ReLu is used. The highest accuracy on the TESS dataset was obtained with the proposed model (ANN+ReLu). Where 7 emotions are classified from 2800 women's audio speech.

Bagus Tris Atmaja et al. [10] trained two unidirectional LSTMs. Classified 3 emotions using both audio and text data from the IEMOCAP dataset. The model trained with the audio dataset gave an accuracy of 58.29 %. Combining audio and text datasets yielded 75.49 % accuracy. Used LSTM+Dense for classification. Bidirectional LSTM with ReLu and softmax activation functions were used in this model.

Zheng Lian et al. [11] used the Domain Adversarial Neural Network (DANN) model with the IEMOCAP dataset. classified 4 emotions in this dataset. They obtained 82.49% accuracy by training this model. This multimodal emotion recognition is mainly context-dependent. This method has been done with both audio and text data. The two types of data form a multimodal system. Multiple approaches were used MFCC, spectral, power, etc. to extract features. Acoustic and lexical features were used to represent the data.

Amiya Kumar Dash et al. [12] proposed a method for speech emotion recognition with the LSTM model and obtained 84.50% accuracy. The system worked by inputting audio and converting the audio to text. The transformed emotional words were run through feature extraction. The text was then divided into positive, negative and object. Finally, the text converted from the speech was trained by the LSTM model. 3000 audios were used for testing and converted to text. Considering SentiWordNet, the data was divided into positive, negative and neutral.

Seunghyun Yoon et al. [13] used an Audio recurrent neural network (RNN) for a speech emotion recognition system. Used the IEMOCAP dataset for model training and classified 4 emotions. Used MFCC to extract features from audio data. This model achieved accuracy ranging from 68.8% to 71.8% through training. Classified happy, sad, angry and neutral emotions. also used text data in experiments and developed a multimodal dual recurrent encoder (MDRE). Textual data is also taken from the IEMOCAP dataset. An overall accuracy of 68.8 % to 71.8 % was obtained.

Kun Zhou et al. [3] trained one of many EST (DeepEST) models with the Emotional Speech Dataset (ESD). The system was trained with 4 emotions and gave 90% accuracy. The dataset is essentially a subset of the IEMOCAP dataset. The model was trained with the optimizer RMSProp and a learning rate of 1e-5. EST or Emotional Style Transfer is basically about learning how to transform an emotional style. Deep features were extracted. One-to-many emotion transfer was performed such as VAW-GAN, and DeepEST. These were then compared and found to have the highest accuracy in DeepEST. N2H, N2S, and N2A techniques for speech quality were also used with the model. Overall, the model got efficient.

Marta Zielonka et al. [14] proposed a custom CNN model. Mainly ResNet18 model was trained with four combined datasets (SAVEE, TESS, CREMA-D, RAVDESS). Six emotions were detected with 57.42 % accuracy. The IEMOCAP dataset was also used by the authors. Some experiments were done with this dataset.

A modern optimization framework today is Bayesian optimization, which is used worldwide for those black-box functions that are expensive [15]. It tunes artificial neural network models to tasks such as speech recognition, image classification and language modeling [15]. Two parts of Bayesian optimization: the surrogate model is used to model the objective function and the acquisition function that outputs a value that is determined by the objective function at a new point [15]. Bayesian is a probabilistic model [15]. Many people use it for good results [15].

From the description of previous work, it is clear that speech-emotion recognition systems need to be further improved. In previous work, some tasks have found low accuracy, some have given high accuracy but trained on data from only male or only female speakers that are efficient at identifying emotion in the speech of either male or female speakers. Some previous work showed high accuracy but a low number of emotions detected. Thus, speech emotion recognition with more emotions (7 emotions) is a very interesting topic of research, including both male and female speaker data. If more accuracy is achieved in this task, it will be great for speech emotion recognition. In this manuscript, this interesting work is proposed with satisfactory accuracy.

## 2. Preliminaries

This section describes the basics of the speech-emotion recognition system. This system inputs speech (audio) of various emotions. And gives a specific emotion based on the input data. Some terms are used for the system.

### 2.1 Speech (Audio)

Audio is a sound that has a frequency of 20 Hz to 20 K Hz. For the system, only human speeches are used. There have two types of audio: mono and stereo. Mono audio is recorded with a single channel (microphone) and stereo is recorded with multiple channels [16]. This system is used mono audio data. Audio data are in many formats such as .wav, .mp3, .mp4, .flac, .aac, .wma, etc. The datasets used here are in .wav format. WAV audio stores the waveform data. It is an uncompressed audio file. It is the strongest and has high-quality audio data.

### 2.2 Emotion

Human emotions are the mental states which are bought by the neurophysiological changes of humans. It is based on thoughts, behavioral responses, feelings, etc. Various types of emotion are seen in human behavior such as happiness, surprise, sadness, fear, anger, disgust, eutrality, etc. This system classified these emotions.

### 2.3 Feature Extraction

Feature extraction is extracting necessary features from the input data for the target classification. For audio data, some techniques can be used for feature extraction. The most used technique is MFCC.

MFCCs are the Mel Frequency Cepstral Coefficients. It is one of the best spectral features [17]. It converts conventional frequency into Mel Scale and for this MFCC is used [9]. Mel scale is mainly a non-linear scale [18]. For calculating mels, the equation [9] is:

$$f_{(mel)}=\{1000\log_{10}(1+f/1000)\}/\log_{10}2. \quad (1)$$

By calculating mel following equation (1), MFCC extracted the necessary features. In this system, the MFCC technique is used for the task of feature extraction. Which provides the best result for speech emotion recognition.

### 2.4 Models

Deep learning models are well for audio data classification. Also, machine learning algorithms can classify audio data. But deep learning models are easy to use, implement and provide the best performance. The Convolutional Neural Network (CNN), the Artificial Neural Network (ANN), Multi-Layer Perceptron (MLP), Long-Short Term Memory (LSTM), etc. are some deep learning models. These models and the merged CNN-LSTM model are used in this system.

#### 2.4.1 CNN

CNN is made with some convolutional layers, some pooling layers, some fully connected layers, and flatten layers. These layers are associated with datasets.

It convolves by the kernel over the feature. The feature can add padding in need. The pooling layer reduces input data size by taking the feature's max value (called Maxpooling) or the min value of the feature. Flatten layer changes into a single dimension from the input feature with the softmax activation function. (Fig. 2) is a sample CNN model of MFCC based. A CNN works better on noisy channel data due to its high tolerance and hence performs better than DNN [20]. In speech emotion recognition, the voice signal is converted into a spectral image, then CNN or other neural networks are run on it [21].

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

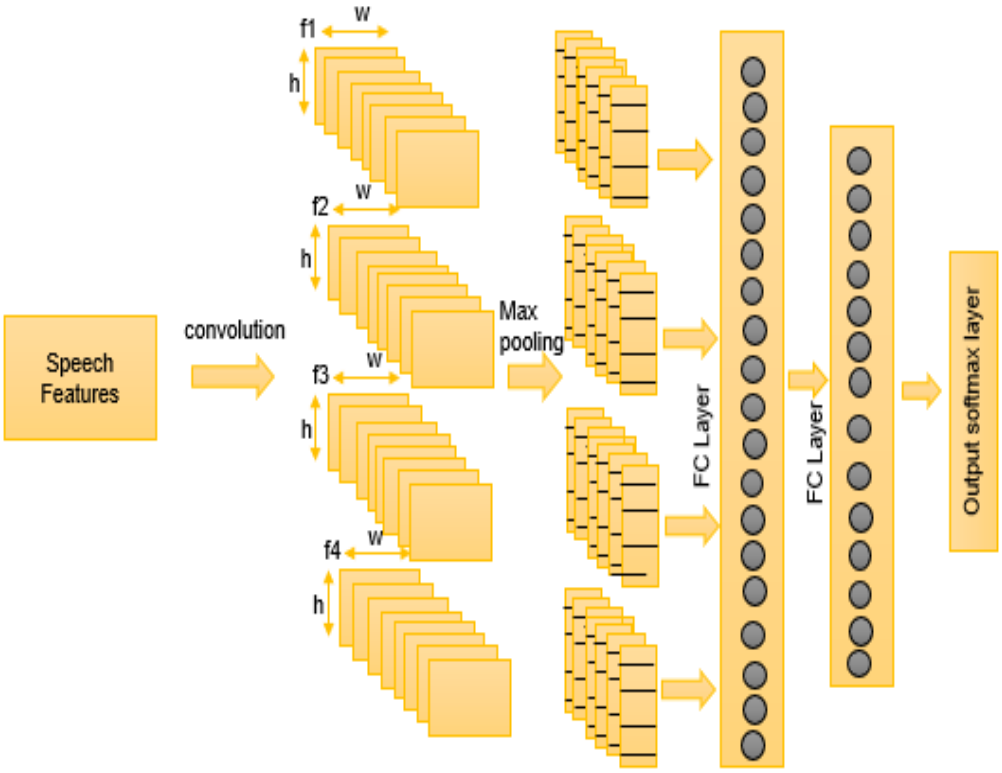


Fig.2. A sample CNN model (MFCC based) [5, 19]

### 2.4.2 LSTM

Long Short-Term Memory (LSTM) is made with both Long-Term Memory (LTM) and Short-Term Memory (STM) and used the gate concept for simple and effective calculation [22]. For a forget gate in an LSTM cell, forward pass equations are:

$$\mathbf{p_t} = \boldsymbol{\sigma_g}\left(\mathbf{W_p x_t} + \mathbf{U_p h_{q-1}} + \mathbf{b_p}\right), \quad (2)$$

$$\mathbf{q_t} = \boldsymbol{\sigma_g}\left(\mathbf{W_q x_t} + \mathbf{U_q h_{t-1}} + \mathbf{b_q}\right), \quad (3)$$

$$\mathbf{r_t} = \boldsymbol{\sigma_g}(\mathbf{W_r x_t} + \mathbf{U_r h_{t-1}} + \mathbf{b_r}), \quad (4)$$

$$\tilde{\mathbf{s}}_\mathbf{t} = \boldsymbol{\sigma_s}(\mathbf{W_s x_t} + \mathbf{U_s h_{t-1}} + \mathbf{b_s}), \quad (5)$$

$$\mathbf{s_t} = \mathbf{p_t} \odot \mathbf{s_{t-1}} + \mathbf{q_t} \odot \tilde{\mathbf{s}}_\mathbf{t}, \quad (6)$$

$$\mathbf{h_t} = \mathbf{r_t} \odot \boldsymbol{\sigma_h}(\mathbf{s_t}). \quad (7)$$

Where x_t is the LSTM cell's input vector, p_t is forget gate's activation vector, q_t is the input/update gate's activation vector, r_t is the output gate's activation vector, h_t is the hidden vector, s̃_t is cell input's activation vector, s_t is state vector of the cell [23].

(Fig. 3) shows an LSTM cell where the green rectangles are the layers, blue circle rectangle means the component-wise. Below to upper direction means the copy and upper to below direction means concatenate [23]. In this figure, σ is the sigmoid function whose range is (0,1). Another activation function named tanh is also used here. That range is (-1,1). The tanh activation function is connected to the output end. So, its output is in the range of (-1,1).

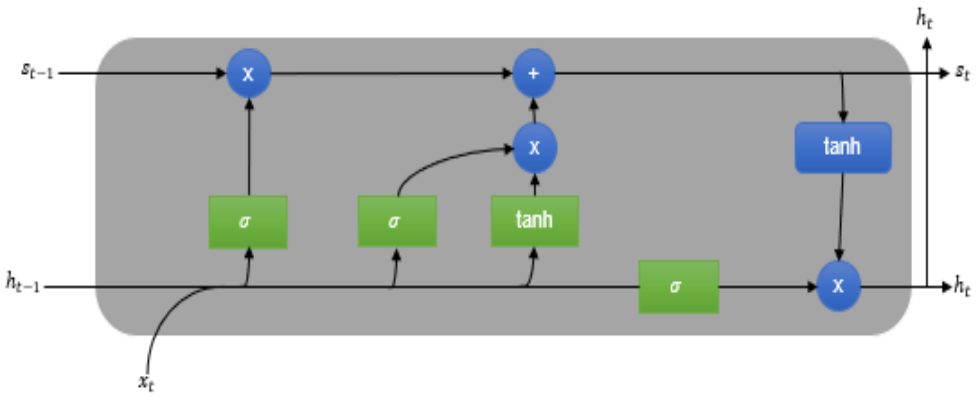


Fig. 3. LSTM cell [23]

### 2.4.3 ANN

Artificial Neural Network (ANN) is a Deep Learning (DL) algorithm that has artificial neurons like the human brain. It has three types of layers: the input layer, the hidden layer, and the output layer. There can have multiple hidden layers. Every input layer and output layer are connected to the hidden layer. The input layer takes inputs. It may have a weight that is also considerable for the output calculation. The hidden layer set up the relation between the input and the output layers. The hidden layer has a bias value for output calculation. After calculating a hidden layer with bias, weight, and input, the output is obtained. The output equation is:

$$y_j = \sum_{j=1}^{m} w_j * x_j + b, \quad (8)$$

where, x = input, w = weight, b = bias, y = output [24].

### 2.4.4 MLP

Multi-Layer Perceptron (MLP) is mainly a typical ANN network with a multi-layer, which means a series of layers. It also has input, hidden, and output layers. MLP is a feed-forward network that treats non-linear and distinguishes data that is not separable from linear [25].

### 2.4.5 CNN-LSTM

The Convolutional Neural Network (CNN) and the Long Short-Term Memory (LSTM) are merged and made the CNN-LSTM model. The LSTM layer is added to the CNN layer here. CNN and LSTM are described before in this section. This model is shown in fig. 4.

In (Fig. 4), it can be clear how the LSTM layer is added to the CNN layer. Here, input features are connected to the convolution layer and max pooling is added. Then LSTM layer is added. A fully connected layer and the activation function are added after the LSTM layer. The output layer is located at the end. It is similar to the CNN architecture (see Figure 2) as it is a merged network of CNN and LSTM. Just an LSTM layer is added after the CNN layer.

## 3. Methodology

The methodology section describes the method of the work. Here the sequence of operations of a system is described. A flow chart shows the working steps of this speech-emotion recognition system. The flowchart is given in (Fig. 5).

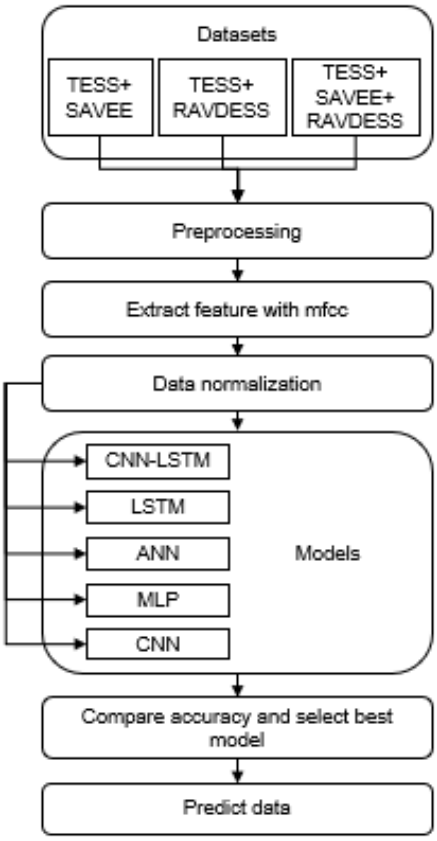


Fig. 5. Flowchart of the SER system [19]

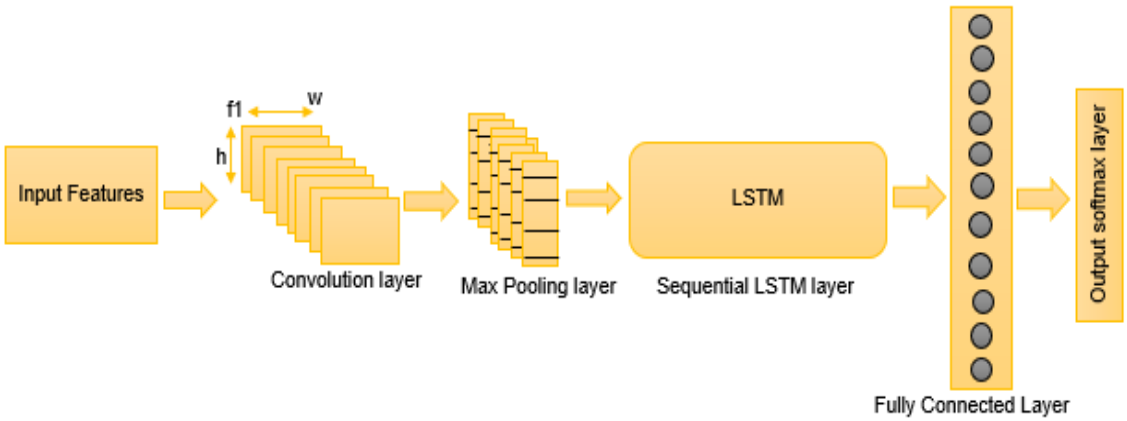


Fig. 4. A sample CNN-LSTM model [26]

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

It can be seen in Figure 5 that first the audio dataset is taken as input and it is preprocessed. Then features are extracted from there using the MFCC technique. After normalizing the data, the model is trained with the help of different algorithms of deep learning. Then the best model is selected by comparing the accuracy of different algorithms and finally the emotional data is predicted. This Speech Emotion Recognition (SER) system can be described by the following steps:

### 3.1 Datasets

Three types of speech datasets are used in this SER system. TESS, SAVEE, and RAVDESS are the most common audio dataset for speech emotion classification.

#### 3.1.1 TESS Dataset

The Toronto Emotional Speech Set (TESS) provides us with high-quality speeches by female speakers [27]. Most of the datasets are male only [27]. But this female-only dataset gives a balance in audio classification [27]. So, this dataset can train well and provides a good model (without overfitting) [27]. It contains at most 2800 audio files in WAV format [27]. Two female speakers speak 200 target audio words [27]. That means a total of 400 audio data is spoken for a single emotion [27]. Seven emotions are in this dataset: fear, happiness, anger, disgust, neutral, surprise, and sadness [27].

#### 3.1.2 SAVEE Dataset

The Surrey Audio-Visual Expressed Emotion (SAVEE) is a high-quality audio dataset that speaks by only male speakers [28]. A total of 480 audio files are in the SAVEE dataset in WAV format. 120 audios are for neutral emotions and 60 files are for other emotions [28]. Seven emotions are in this dataset: fear, happiness, anger, neutral, surprise, disgust, and sadness [28]. It would be a good interaction with other male-only datasets [28].

#### 3.1.3 RAVDESS Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a well-known dataset for audio classification [29]. 24 speakers speak in a total of 1440 audio speeches [29]. 12 male and 12 female speakers speak of eight emotions here: happy, sad, calm, angry, surprised, disgusted, fearful, and neutral [29]. Every speaker speaks 60 speeches and all are in WAV format [29]. It also provides audio files with strong and normal emotional intensity [29].

These three datasets are concatenated with each other as TESS+SAVEE, TESS+RAVDESS, and TESS+SAVEE+RAVDESS for increasing the number of training datasets and classifying speech emotions more accurately. So, these three datasets TESS + SAVEE, TESS + RAVDESS, and TESS + SAVEE + RAVDESS datasets are used for this experiments and training model. In this experiment, 7 emotions are used: fear, happiness, neutral, anger, disgust, sadness, and surprise.

### 3.2 Preprocessing

The dataset is preprocessed with the resample type of kaiser_fast, reducing the dataset's load time. It sampled with a sample rate of 44100. Data is split into training and testing parts. Here, 75% of the data is used for model training, and 25% is used for model testing. Data split is done by shuffling the dataset with the random state 42. This increases the performance of the model train. We used the 80:20 data split and the 75:25 data split. Where we got the best results for a 75:25 split of the data. To get a good model, it needs more data and tests with less data than train This work provided good results for this segmentation of 75% train data and 25% test data.

### 3.3 Feature Extraction

Speech input needs to convert into digital signals to train the model [9]. After converting into digital signals, the audio signal is processed, and extract the related suitable features for training the model [9]. For extracting feature, MFCCs is used in this system.

### 3.4 Data Normalization

Primarily the audio data are converted to the array. Then mean and the standard deviation (SD) are found from the dataset. Then normalize the data with subtract the mean and dividing the subtraction by the standard deviation (SD). Data are again converted to a numpy array and expanded dims for train CNN, ANN, CNN-LSTM, and LSTM models. It needn't expand dims to train the MLP model.

### 3.5 Models

In this system, 5 deep learning models are used. All 5 models are trained with all three datasets (TESS + SAVEE, TESS + RAVDESS, TESS + SAVEE + RAVDESS). The models are CNN, LSTM, ANN, MLP, and CNN - LSTM. In details:

#### 3.5.1 CNN

The Convolutional Neural Network (CNN) is model of a deep learning for classification. This model uses the sequential method. In this model, the softmax activation function is used, which reduces overfitting.

Dropout is also used for reduced overfitting. Adam optimizer is used for reduced loss. Loss is calculated by categorical_crossentrophy. The pseudocode of the CNN model is [30]:

Pseudocode 1: CNN

Notations: In = Input, N = Number of neurons/hidden layers, E = Repeat/Epochs, M = Number of MaxPooling layers

```
1. PROCEDURE CNN (IN, N, E)
        Import library and datasets
2.  Input ←——— Dataset with variable combinations.
        Training CNN
3.     for In = 1 to End of input do
4.        for N = 1 to 256 do
5.           for p=1 to 2 do
6.              for E = 1 to 200 do
7.                    Add Conv1D
8.                    Add MaxPooling
9.                    Add Softmax
10.                  Train CNN
11.               end for
12.            end for
13.         end for
14.    end for
15.    return CNN-metrics
17. end procedure
```

### 3.5.2 LSTM

The Long-Short Term Memory (LSTM) is another model for classifying speech data. This is also defined as sequential. Adam optimizer and softmax activation function is used for reduced overfitting. Loss is calculated by categorical_crossentrophy. The batch size is 16. The pseudo-code of the LSTM model is [30]:

Pseudocode 2: LSTM

Notations: In = Input, N = Number of neurons for this model, E = Repeat/Epochs

```
1. procedure LSTM (In, N, E)
        Import library and datasets
2.  Input ←——— Dataset with variable combinations.
        Training LSTM
3.     for In = 1 to End of input do
4.        for N = 1 to 256 do
5.           for E = 1 to 200 do
6.                 Add Softmax
7.                 Train LSTM
8.              end for
9.           end for
10.    end for
11.    return LSTM-metrics
12. end procedure
```

### 3.5.3 ANN

Artificial Neural Network (ANN) is a very fast-running model. Softmax activation function and the Stochastic Gradient Descent (SGD) optimizer are used here. Every dataset is trained with the batch size 64. For loss calculation, categorical_crossentrophy is used. The pseudocode for ANN is [30]:

Pseudocode 3: ANN

Notations: In = Input, N = Number of Neurons,
E = Repeat/Epochs

```
1. procedure ANN (In, N, E)
        Import library and datasets
2.  Input ←——— Dataset with variable combinations.
        Training ANNs
3.     for In = 1 to End of input do
4.        for N = 1 to 4 do
5.           for E = 1 to 200 do
6.                 Add Softmax
7.                 Train ANN
8.              end for
9.           end for
10.    end for
11.  return ANN-metrics
12. end procedure
```

### 3.5.4 MLP

Multilayer Perceptron (MLP) is a deep neural network. That is often used for sound classification. The pseudocode of MLP model is [30]:

Pseudocode 4: MLP

Notations: In = Input, H = Hidden layer size,
I = Iterations

```
1. procedure MLP (In, Neurons, N)
        Import library and datasets
2.  Input ←——— Dataset with variable combinations.
        Training MLP
3.     for In = 1 to End of input do
4.        for H= 1 to 300 do
5.           for I = 1 to 500 do
6.                 Set alpha = 0.01
7.                 Set epsilon = 1e-08
8.                 Add adaptive learning rate
9.                 Train MLP
10.             end for
11.          end for
12.            Calculate accuracy
13.    end for
14.     return MLP-accuracy
15. end procedure
```

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

### 3.5.5 CNN-LSTM

CNN and LSTM models are merged and the CNN_LSTM model is built. Here, CNN's convolution layer sequentially stays and the LSTM layer is included in the CNN-LSTM model. This network is very much useful for audio data classification. In this model, the softmax activation function and the Adam optimizer are used. The use of the dropout function also reduces overfitting. With a batch size of 16. Pseudocode of CNN-LSTM is [30]:

Pseudocode 5: CNN-LSTM

Notations: In = Input, N = Number of neurons/hidden layers, E = Repeat/Epochs

```
1. procedure CNN-LSTM (In, N, E)
        Import library and datasets
2. Input ←——— Dataset with variable combinations.
        Training CNN-LSTM
3.    for In = 1 to End of input do
4.       for N = 1 to 512 do
5.          for E = 1 to 200 do
6.             Add Conv1D
7.                for Neurons = 1 to 256 do
8.                   for N = 1 to 200 do
9.                      Add LSTM layer
10.                     Add Softmax
11.                     Train CNN-LSTM
12.                  end for
13.               end for
14.         end for
15.      end for
16.   end for
17.   return CNN-LSTM-metrics
18. end procedure
```

### 3.6 Compare accuracy and Predict data

After training all the models with each dataset, the accuracy of the models is gotten. Not only train accuracy but also validation accuracy and test accuracy are targeted. A model is best fitted when high validation accuracy and test accuracy are available. This accuracy is compared for all models on all datasets. Chosen the best model with the corresponding dataset, where the validity and test accuracy are highest. After obtaining the best model, some audio data is predicted and the desired class (emotion) is obtained.

## 4. Result and Analysis

Now, the results of the model are discussed. The models have already been described in the Methods section of this paper. CNN, LSTM, CNN-LSTM, MLP, and ANN are applied to the TESS+SAVEE, TESS+RAVDESS, and TESS+RAVDESS+SAVEE datasets. By training these models, the CNN-LSTM model obtained the best accuracy for the TESS+SAVEE dataset. Therefore, this is considered the proposed model. Further explanation is:

Table 1

Accuracy table

| Dataset | Model | Validation accuracy | Test accuracy |
|---|---|---|---|
| TESS+RAVDESS (8 emotion) | CNN | 72.86 % | 72.86 % |
| | LSTM | 60.38 % | 60.38 % |
| | CNN+LSTM | 69.34 % | 69.34 % |
| | MLP | 53 % | 53 % |
| | ANN | 63.02 % | 63.02 % |
| TESS+RAVDESS +SAVEE (8 emotion) | CNN | 72.54 % | 71.86 % |
| | LSTM | 57.63 % | 57.63 % |
| | CNN+LSTM | 65.34 % | 62.46 % |
| | MLP | 44.83 % | 44.83 % |
| | ANN | 57.88 % | 57.63 % |
| TESS+SAVEE (7emotions) | CNN | 83.97 % | 83.97 % |
| | LSTM | 27.95 % | 27.95 % |
| | MLP | 54 % | 54 % |
| | ANN | 77.44 % | 77.44 % |
| | **CNN+LSTM (Proposed)** | **84.35 %** | **84.35 %** |

The TESS+RAVDESS dataset provides the highest accuracy of 72.86 % in CNN. ANN is one of the most common models in deep learning. It lasts a short time. But complex audio datasets are difficult to classify more accurately by this model. LSTM models are a bit more complicated. This provides more accuracy with fewer datasets, such as the TESS or SAVEE datasets. But it provides less accuracy in combining datasets. CNN-LSTM provides accuracy close to CNN. The MLP model gives high accuracy when the number of impulses is low. But it gives less accuracy in increasing the number of impulses.

TESS+RAVDESS+SAVEE is the largest dataset in this system. The above-mentioned 5 models are applied to this combined dataset as well. This dataset gives the highest accuracy of 72.54 % for the CNN model. This dataset is very complex audio data. Hence overfitting problems arise and accuracy decreases.

The TESS+SAVEE datasets are two of the best, least complex audio datasets. Overfitting is minimized in this dataset.

So, for emotion audio classification, this dataset is more suitable. The proposed dataset provides the highest accuracy of 84.35 % in the CNN-LSTM (proposed) model. The accuracies for all models are given in table 1. The proposed model gives a validation accuracy of 84.35 % after 200 epochs. It also gave a test accuracy of 84.55 %, which is most important and valuable for this emotion classifier. Table 2 compares the accuracy of the

Table 2

Accuracy comparison

| Model | Dataset | Emotion Number | Accuracy |
|---|---|---|---|
| Multimodal Dual Recurrent Encoder (MDRE) | IEMOCAP | 4 | 68.8% to 71.8%. |
| MLP | IEMOCAP | 6 | 66.1% |
| LSTM | | | 64.2% |
| multi-hop attention model (MHA)_MHA-2 | IEMOCAP | 4 | 65-80% |
| MFCC and text model (Model 4C) | IEMOCAP | 4 | 76.1% |
| One-to-many EST(DeepEST) | ESD | 4 | 90% |
| wav2vec 2.0 | IEMOCAP+ RAVDESS | 8 | RAVDESS: 84.1% |
| Two unidirectional LSTM (Text+Speech) | IEMOCAP | 3 | 75.49% |
| Domain adversarial neural network (DANN) | IEMOCAP | 4 | 82.49% |
| ANN Model for SER using Mel Frequency Cepstral Coefficients (MFCCs) feature extraction | TESS(Female only), RAVDESS, CREMA, SAVEE(Male only) | 6(RAVDESS , CREMA) , 7(TESS, SAVEE) | TESS- 99.52%, RAVDESS- 88.72%, CREMA- 71.69%, SAVEE- 86.80% |
| Model-Agnostic Meta-Learning (MAML) algorithm | TESS, RAVDESS, EMODB, SAVEE, EMOVO, SHEMO, URDU | 4 | 50-72% |
| ResNet18 | Combined TESS, SAVEE, RAVDESS, CREMA-D | 6 | 57.42% |
| **CNN+LSTM (Proposed)** | **TESS+SAVEE** | **7** | **84.35%** |

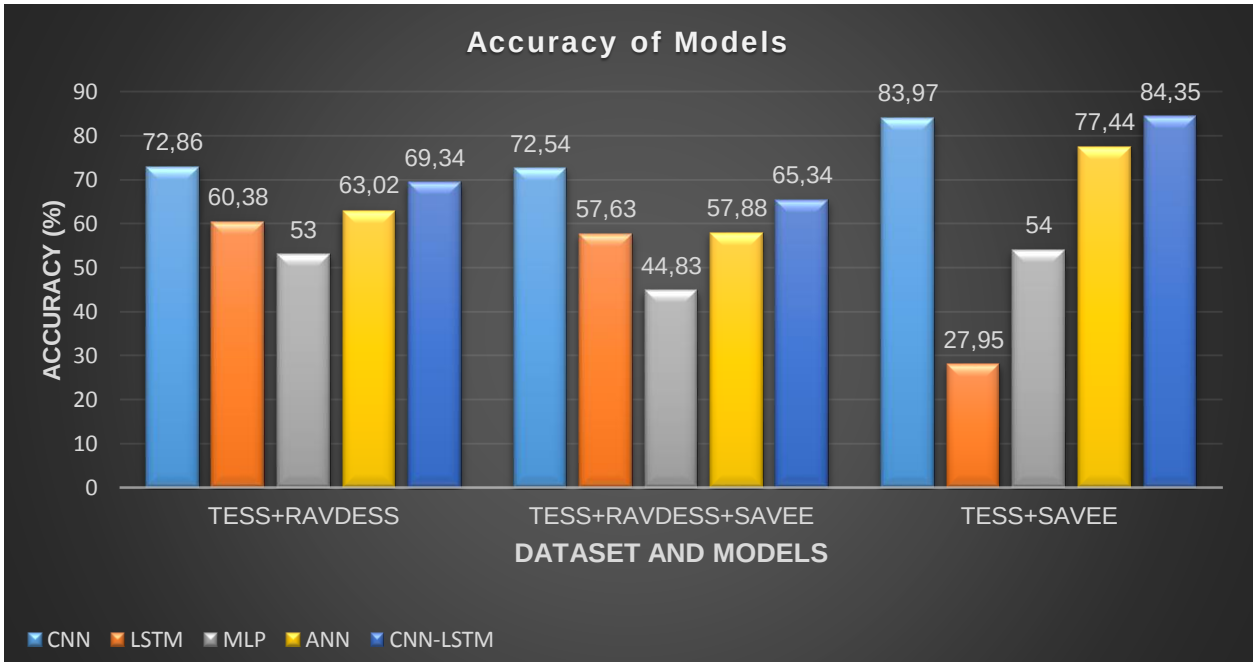


Fig.6. Bar diagram of accuracy table 1

proposed model with the accuracy of all previous models. It can be seen that the accuracy of the data which worked with both male and female speech and detected emotions like 6-7 is low. But the proposed model identified 7 emotions with male-female data. Its accuracy is more than in previous works.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

It can be easily understood that the comparison of accuracy of all models by the bar diagram (Fig. 6).

Fig. 6 clearly shows that the proposed model is the best in terms of accuracy. For the TESS+SAVEE dataset, CNN-LSTM is the recommended model. This model has 92.35 % training accuracy, 84.35 % validation accuracy, and 84.35 % testing accuracy. So, it is a good combination of the TESS+SAVEE dataset and the CNN-LSTM model for emotion classification.

| | actualvalues | predictedvalues |
|---|---|---|
| 210 | fear | surprise |
| 211 | angry | angry |
| 212 | happy | happy |
| 213 | neutral | neutral |
| 214 | sad | sad |
| 215 | fear | fear |
| 216 | fear | fear |
| 217 | angry | angry |
| 218 | neutral | neutral |
| 219 | neutral | neutral |

Fig. 7. Prediction

Fig. 7 shows the prediction results for 10 experimental data sets. where 9 out of 10 facts predicted the correct emotion. So, it is best for speech emotion classification of the audio dataset will that successfully predict the emotion of the voice data. It is a sample result of SER system implementation. Where the prediction results can be shown for the test data sample.

## Conclusions

This paper proposes a merged CNN-LSTM model with TESS+SAVEE (combined dataset) that provides the best results for speech-emotion recognition systems. Accuracy is derived from each dataset by training each model separately. The model with the highest accuracy is considered the best fit model. System performance depends on the model (model classification performance may vary) and dataset (clear and high-quality audio support to get a good result). Feature extraction techniques are also important. MFCC is used for this task. Any other feature extraction technique (such as ZCR) may give different results. Moreover, the speakers are very important for this system. It needs both male and female speakers to perform better in tests. This system uses both male and female speaker voices. Thus, this system is more efficient for speech emotion recognition (gives 84.35 % train and test accuracy). Never before has such accuracy been achieved in identifying 7 emotions with both male and female datasets. So, getting 84.35 % accuracy for both the male and female dataset is a big achievement. This is why it has emerged as the most preferred technique and method for emotional recognition.

The future scope of the system is that any other audio dataset can be used with any other model. Also, any other method can be used to train the model (such as machine learning methods). In this system (CNN-LSTM network) only CNN and LSTM are combined. Any other model can be combined to train the audio data. It would be great for speech-emotion recognition systems if some other method could increase the accuracy.

**Author contributions: c**oding analysis, comparison and drafting of the manuscript were done – **Sumon Kumar Hazra**. The research idea was conceived and led with overall supervision, revision, guidance, coding, analysis, comparison and drafting of the manuscript were also done – **Romana Rahman Ema**. Overall supervision, revision and guidance of the paper were done – **Syed Md. Galib**. Analysis, comparison and drafting of the manuscript were done – **Shalauddin Kabir**. Analysis, comparison and drafting of the manuscript were done – **Nasim Adnan**.

## References (BSI)

1. Lin, Y. L. and Wei, G. Speech emotion recognition based on HMM and SVM. *2005 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2005, vol. 8, pp. 4898-4901. DOI: 10.1109/ICMLC.2005.1527805.

2. Li, M., Yang, B., Levy, J., Stolcke, A., Rozgic, V., Matsoukas, S., Papayiannis, C., Bone, D. and Wang, C. Contrastive unsupervised learning for speech emotion recognition. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6329-6333. DOI: 10.1109/ICASSP39728.2021.9413910.

3. Zhou, K., Sisman, B., Liu, R. and Li, H. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 920-924). DOI: 10.1109/ICASSP39728.2021.9413391.

4. Pepino, L., Riera, P. and Ferrer, L. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv*, 2021, vol. abs.2104.03502. DOI: 10.48550/arXiv.2104.03502.

5. Tripathi, S., Kumar, A., Ramesh, A., Singh, C. and Yenigalla, P. Deep learning-based emotion recognition system using speech features and transcriptions. *arXiv*, 2019, vol. abs.1906.05681. DOI: 10.48550/arXiv.1906.05681.

6. Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G. and Yu, D. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 2014, vol. 22, iss. 10, pp. 1533-1545. DOI: 10.1109/TASLP.2014.2339736.

7. Yoon, S., Byun, S., Dey, S. and Jung, K. Speech emotion recognition using a multi-hop attention mechanism. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2822-2826. DOI: 10.1109/ICASSP.2019.8683483.

8. Lim, W., Jang, D. and Lee, T. Speech emotion recognition using convolutional and recurrent neural networks. *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*, 2016, pp. 1-4. DOI: 10.1109/APSIPA.2016.7820699.

9. Dolka, H., VM, A. X. and Juliet, S. Speech emotion recognition using ANN on MFCC features. *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, 2021, pp. 431-435. DOI: 10.1109/ICSPC51351.2021.9451810.

10. Atmaja, B. T., Shirai, K. and Akagi, M. Speech emotion recognition using speech features and word embedding. *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 519-523. DOI: 10.1109/APSIPAASC47483.2019.9023098.

11. Lian, Z., Tao, J., Liu, B., Huang, J., Yang, Z. and Li, R. Context-Dependent Domain Adversarial Neural Network for Multimodal Emotion Recognition. *Interspeech, 2020*, pp. 394-398. DOI: 10.21437/Interspeech.2020-1705.

12. Dash, A. K., Pradhan, R., Rout, J. K. and Ray, N. K. A constructive model for sentiment analysis of speech using deep learning. *2018 International Conference on Information Technology (ICIT)*, 2018, pp. 1-6. DOI: 10.1109/ICIT.2018.00013.

13. Yoon, S., Byun, S. and Jung, K. Multimodal speech emotion recognition using audio and text. *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 112-118. DOI: 10.1109/SLT.2018.8639583.

14. Zielonka, M., Piastowski, A., Czyżewski, A., Nadachowski, P., Operlejn, M. and Kaczor, K. Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets. *Electronics*, 2022, vol. 11, iss. 22, article no. 3831. DOI: 10.3390/electronics11223831.

15. Yaloveha, V., Podorozhniak, A. and Kuchuk, H. Convolutional neural network hyperparameter optimization applied to land cover classification. *Radioelectronic and Computer Systems*, 2022, no. 1, pp. 115-128. DOI: 10.32620/reks.2022.1.09.

16. *Mono vs Stereo: The Complete Guide*. Available at: https://www.hifireport.com/mono-vs-stereo-the-complete-guide/ (accessed Oct. 21, 2022).

17. Selvaraj, M., Bhuvana, R. and Padmaja, S. Human speech emotion recognition. *International Journal of Engineering & Technology*, 2016, vol. 8, no. 1, pp. 311-323. Available at: https://www.enggjournals.com/ijet/docs/IJET16-08-01-090.pdf. (accessed Oct. 01, 2022).

18. Chakroborty, S., Roy, A. and Saha, G. Fusion of a complementary feature set with MFCC for improved closed set text-independent speaker identification. *2006 IEEE International Conference on Industrial Technology*, 2006, pp. 387-390. DOI: 10.1109/ICIT.2006.372388.

19. Varshini, P., Soundarya, R. et al. Speech Emotion Analyzer. *International Journal of Research Publication and Reviews*, 2021, vol 2, Iss. 8, pp. 1026-1034. Available at: https://www.ijrpr.com/uploads/V2ISSUE8/IJRPR1095.pdf. (accessed Oct. 01, 2022).

20. Dua, S., Kumar, S. S., Albagory, Y., Ramalingam, R., Dumka, A., Singh, R., Rashid, M., Gehlot, A., Alshamrani, S. S. and AlGhamdi, A. S. Developing a Speech Recognition System for Recognizing Tonal Speech Signals Using a Convolutional Neural Network. *Applied Sciences*, 2022, vol. 12, iss. 12, article no. 6223. DOI: 10.3390/app12126223.

21. Trinh Van, L., Dao Thi Le, T., Le Xuan, T. and Castelli, E. Emotional Speech Recognition Using Deep Neural Networks. *Sensors*, 2022, vol. 22, iss. 4, article no. 1414. DOI: 10.3390/s22041414.

22. *Understanding Architecture of LSTM*. Available at: https://www.analyticsvidhya.com/blog/2021/01/understanding-architecture-of-lstm/ (accessed Oct. 21, 2022).

23. *Long short-term memory*. Available at: https://en.wikipedia.org/wiki/Long_short-term_memory (accessed Oct. 21, 2022).

24. *Artificial Neural Network Tutorial.* Available at: https://www.javatpoint.com/artificial-neural-network (accessed Oct. 21, 2022).

25. Sahu, G. Multimodal speech emotion recognition and ambiguity resolution. *arXiv*, 2019, vol. abs.1904.06022. DOI: 10.48550/arXiv.1904.06022.

26. Livieris, I. E., Pintelas, E. and Pintelas, P. A CNN–LSTM model for gold price time-series forecasting. *Neural computing and applications*, 2020, vol. 32, iss. 23, pp.17351-17360. DOI: 10.1007/s00521-020-04867-x.

27. *Toronto emotional speech set (TESS)*. Available at: https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess (accessed Oct. 21, 2022).

28. *Surrey Audio-Visual Expressed Emotion (SAVEE)*. Available at: https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee (accessed Oct. 21, 2022).

29. *RAVDESS Emotional speech audio*. Available at: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio (accessed Oct. 21, 2022).

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

30. *Pseudocode for the automated artificial neural network algorithm to produce the trained network library*. Available at: https://www.researchgate.net/figure/Pseudocode-for-the-automated-artificial-neural-network-algorithm-to-produce-the-trained_fig1_325677157 (accessed Oct. 21, 2022).

*Надійшла до редакції 10.09.2022, розглянута на редколегії 20.11.2022*

## РОЗПІЗНАВАННЯ ЕМОЦІЙ ЛЮДСЬКОГО МОВЛЕННЯ ЗА ДОПОМОГОЮ МЕТОДУ ГЛИБОКОГО НАВЧАННЯ ТА ФУНКЦІЙ MFCC

***Сумон Кумар Хазра, Романа Рахман Ема*, Саїд М. Галіб, Шалауддін Кабір, Насім Аднан***

**Тема:** Розпізнавання мовних емоцій (SER) зараз є популярною цікавою темою дослідження. Його мета – налагодити взаємодію між людьми та комп'ютерами за допомогою мови та емоцій. Для розпізнавання мовних емоцій у цій статті використовуються п'ять моделей глибокого навчання: згорточна нейронна мережа, довгострокова пам'ять, штучна нейронна мережа, багаторівневий персептрон, об'єднана CNN і мережа LSTM (CNN-LSTM). Для цієї системи використовувалися набори даних Набір емоційних промов Торонто (TESS), Суррей Аудіовізуальні виражені емоції (SAVEE) і Аудіовізуальна база даних емоційної мови та пісні Райерсона (RAVDESS). Вони навчаються шляхом злиття 3 способів TESS+SAVEE, TESS+RAVDESS і TESS+SAVEE+RAVDESS. Ці набори даних представляють собою велику кількість аудіозаписів, які розмовляють як чоловіки, так і жінки, які розмовляють англійською мовою. У цьому документі класифіковано сім емоцій (сум, щастя, гнів, страх, огида, нейтральність і здивування), що є проблемою для визначення семи емоцій як для чоловіків, так і для жінок. У той час як більшість працювали з виявленням мовлення й емоцій лише для чоловіків або лише для жінок з обома даними, це дало низьку точність. Щоб навчити модель глибокого навчання на аудіоданих, функції потрібно виділити за допомогою техніки вилучення ознак. Мел-частотні кепстральні коефіцієнти (MFCC) виділяють усі необхідні характеристики з аудіоданих для класифікації мовних емоцій. Після навчання п'яти моделей із трьома наборами даних найкращу точність 84,35 % досягає CNN-LSTM із набором даних TESS+SAVEE.

**Ключові слова:** розпізнавання мовних емоцій (SER); метод глибокого навчання; просунутий ШІ; Мел-частотні кепстральні коефіцієнти (MFCC); аудіодані.

**Сумон Кумар Хазра –** студент інформатики та інженерії, Джашорський університет науки та технологій, Бангладеш.

**Романа Рахман Ема (* відповідний автор) –** доцент кафедри комп'ютерних наук та інженерії, Джашорський університет науки та технологій, Бангладеш.

**Сайєд Мд. Галіб –** д-р, проф. комп'ютерних наук та інженерії, Джашорський університет науки і технологій, Бангладеш.

**Шалауддін Кабір –** викл. інформатики та інженерії, Джашорський університет науки та технологій, Бангладеш.

**Насім Аднан –** д-р мед. наук, доц. каф. комп'ютерних наук та інженерії, Джашорський університет науки та технологій, Бангладеш.

**Sumon Kumar Hazra** – Student of Computer Science and Engineering, Jashore University of Science and Technology, Bangladesh,
e-mail: sumon.just.cse@gmail.com, ORCID: 0000-0001-8796-3456.

**Romana Rahman Ema (* corresponding Author)** – Assistant professor of Computer Science and Engineering, Jashore University of Science and Technology, Bangladesh,
e-mail: rr.ema@just.edu.bd, ORCID: 0000-0002-2384-9539.

**Syed Md. Galib** – Dr., Professor of Computer Science and Engineering, Jashore University of Science and Technology, Bangladesh,
e-mail: galib.cse@just.edu.bd, ORCID: 0000-0002-5708-727X.

**Shalauddin Kabir** – Lecturer of Computer Science and Engineering, Jashore University of Science and Technology, Bangladesh,
e-mail: sks.kabir@just.edu.bd, ORCID: 0000-0002-0031-8807.

**Nasim Adnan** – Dr. Md., Assistant professor of Computer Science and Engineering, Jashore University of Science and Technology, Bangladesh,
e-mail: nasim.adnan@just.edu.bd, ORCID: 0000-0001-9210-2896.