Recognizing Semi-Natural and Spontaneous Speech Emotions Using Deep Neural Networks

AMMAR AMJAD, LAL KHAN, NOMAN ASHRAF, MUHAMMAD BILAL MAHMOOD, AND HSIENTSUNG CHANG

1Department of Computer Science and Information Engineering, College of Engineering, Chang Gung University, Taoyuan 33302, Taiwan
2CIC, Instituto Politécnico Nacional, Mexico City 7738, Mexico
3Department of Software Engineering, Dalian University of Technology, Dalian 116024, China
4Artificial Intelligence Research Center, Chang Gung University, Taoyuan 33302, Taiwan
5Department of Physical Medicine and Rehabilitation, Chang Gung Memorial Hospital, Taoyuan 333, Taiwan
6Bachelor Program in Artificial Intelligence, Chang Gung University, Taoyuan 33302, Taiwan

Corresponding author: Hsien-Tsung Chang (smallpig@widelab.org)

This work was supported by the Chang Gung Memorial Hospital under Grant CMRPD2J0023.

ABSTRACT We needed to find deep emotional features to identify emotions from audio signals. Identifying emotions in spontaneous speech is a novel and challenging subject of research. Several convolutional neural network (CNN) models were used to learn deep segment-level auditory representations of augmented Mel spectrograms. The proposed study introduces a novel technique for recognizing semi-natural and spontaneous speech emotions based on 1D (Model A) and 2D (Model B) deep convolutional neural networks (DCNNs) with two layers of long-short-term memory (LSTM). Both models used raw speech data and augmented (mid, left, right, and side) segment level Mel spectrograms to learn local and global features. The architecture of both models consists of five local feature learning blocks (LFLBs), two LSTM layers, and a fully connected layer (FCL). In addition to learning local correlations and extracting hierarchical correlations, LFLB comprises two convolutional layers and a max-pooling layer. The LSTM layer learns long-term correlations from local features. The experiments illustrated that the proposed systems perform better than conventional methods. Model A achieved an average identification accuracy of 94.78% for speaker-dependent (SD) with a raw SAVEE dataset. With the IEMOCAP database, Model A achieved an average accuracy of an SD experiment with raw audio of 73.15%. In addition, Model A obtained identification accuracies of 97.19%, 94.09%, and 53.98% on SAVEE, IEMOCAP, and BAUM-1s, the databases for speaker-dependent (SD) experiments with an augmented Mel spectrogram, respectively. In contrast, Model B achieved identification accuracy of 96.85%, 88.80%, and 48.67% on SAVEE, IEMOCAP, and the BAUM-1s database for SI experiments with augmented reality Mel spectrogram, respectively.

INDEX TERMS Speech emotion recognition, convolutional neural network, data augmentation, long-short-term memory, spontaneous speech database.

I. INTRODUCTION

Speech is an efficient, quick, and fundamental way of human communication. Speech signals are one of the most natural ways humans express their emotions. Speech emotion recognition (SER) is a challenging task in artificial intelligence, pattern recognition, signal processing, and other fields [1], [2]. The existing studies [3], [4] have been on SER problems using data gathered in laboratory-controlled conditions, such as the acted and simulated databases [5], [6] to identify emotions. Whereas semi-natural emotions are linked with high identification accuracy, these emotions can be easily exaggerated. However, acted emotions fail to accurately represent the features of human emotional expression in natural situations. Spontaneous emotions are more demanding and harder to describe than acted or semi-natural emotions in the wild. Therefore, emotion recognition in the wild has attracted much attention. Although, extracting the most discriminative features for speech expression feature extraction is essential for the SER frameworks. The fundamental emotional characteristics [7], [8] of speech are low-level descriptors (LLDs). The commonly used LLDs...
consist of prosody, voice quality, and spectral characteristics [7]–[9]. Recently, many large feature sets based on LLDs, including INTERSPEECH 2010 [10], ComParE [11], AVEC 2013 [12], and GeMAPS [13] have been proposed for SER. Handcrafted audio features are utilized as an input of neural networks when utilizing transfer learning and deep learning approaches for SER [3], [14]–[16]. The overall outcome of handcrafted features extraction from specific emotions is very high; however, the extraction of handcrafted features generally requires human effort [17], [18]. Accordingly, improved features extraction methods are required to effectively identify the most discriminative emotional features.

The newly developed deep learning methods [19], [20], which have attained significant attention in the SER, provide suitable solutions [14], [21], [22] to the problems mentioned above. Multiple deep neural frameworks have been utilized for high-level feature learning tasks, such as deep neural networks (DNNs) [23]–[25], deep convolutional neural networks (CNNs) [26], and Long short-term memory recurrent neural networks (LSTM-RNNs) [27]. DNNs comprise one or more underlying hidden layers between inputs and outputs based on feed-forward architectures. Automated feature learning approaches were developed to acquire high-level features that accurately recognize human emotions [10], [12], [13], [28], [29], and all have given better results. The DNN method was the first deep learning method deployed, correlated with handcrafted acoustic characteristics. In [30], introduced a DNN-based (Gerda) approach, following the Mahalanobis minimum-distance classifier (MDC) for SER, to learn the discriminative features of 6552 low-level acoustic descriptors (LLDs). The DNN [31] was utilized to learn high-level features from handcrafted features. Emotion classification was fostered by using an extreme learning machine (ELM) [32]–[34]. MFCCs are used as an input of DNN for obtaining high-level features. An extreme learning machine (ELM) is utilized to classify speech emotions [35]. The author in [36], utilized a DNN to compress an utterance into a fixed-length matrix by pooling the last hidden layer actions across time. Then, the encoded matrices are used to learn an ELM kernel for the utterance-level classification of speech emotions. DNNs cannot successfully acquire discriminative features for SER because DNNs typically require handcrafted characteristics as SER inputs.

CNN’s are composed of multi-level convolutional and pooling layers, which allows obtaining mid-level feature representations using data input and train models. To take advantage of CNNs’ outstanding performance in computer vision applications [26], 2D time-frequency representations generated from acoustic spectra are often input into CNNs for SER. Specifically, the researchers of [32] used spectrograms as input data of a hybrid network that comprises a sparse auto-encoder and a one-layer CNN to train salient features for SER. The suggested approach [37] used segment-level spectrograms as CNN input to extract discriminative characteristics. In [38], segment-level features are extracted using an image spectrogram as inputs to a deep network such as AlexNet [26]. Implementation of CNNs with LSTM-RNNs has recently been a new study topic in SER. The studies [39], [40] proposed an attention-based bidirectional LSTM with a spatial CNN for deep spectrum feature extraction by utilizing segment-level spectrograms. In [14], the proposed method used a segment-level spectrogram with a deep convolutional LSTM architecture.

Notice that the two-dimension mentioned above (2D) CNN techniques such as CNN, CNN+LSTM, and CNN+RNN, effectively extract energy modulation features. As a result, these time and frequency features, which extract 2D time and frequency spectrograms of audio data, performed very well in SER applications. Moreover, various 1D CNN models have been employed in recent years for features extraction in SER, and other applications [41]–[45]. For example, the researchers in [46] investigated the efficacy of the different 1D CNN architectures for extracting features from the 1D original waveforms on SER challenges. However, these utilized 1D CNN models with one or two convolution layers are shallow, making their learned 1D CNN features inappropriate for SER. On the other hand, 1D CNNs trained on sample-level 1D audio waveforms have been effectively used for music categorization, using feature extraction learned from the original 1D raw audio waveforms.

In [32], a sparse auto-encoder was used to learn essential features from spectrogram for SER with one-layer CNN. The end-to-end SER framework with a combination of two-layers CNN with an LSTM is presented in [33], [34]. A recurrent neural network (RNN) is designed to handle long-range dependencies [47]. In suggested techniques [3], [32]–[34], researchers used limited data from publically available emotional speech databases to develop deep CNNs with one or two convolutional layers (CL). In [15], [48] found that varying lengths of spectrogram yield different affective cues for recognizing specific emotions because other emotions dominated separate segment-level features in an utterance. Therefore, when various segment features are used for speech emotion recognition, the discriminating strength of the retrieved utterance-level features changes for an utterance. This study proposed a new spontaneous and semi-natural SER approach based on the deep CNN+LSTM architecture. Similar to [14], [49], the proposed approach used multiple image scales as inputs to a single CNN. This approach is different from other approaches, so-called multiscale systems [50] in which CNN’s use with subnetwork is based on specific information. To further increase SER performance, we integrate deep LFLB with LSTM at different lengths of RGB spectrograms. The presented approach learns local features from raw audio spectrograms using deep CNNs on target publicly available databases. Then, to achieve utterance-level feature extraction for SER, the temporal dynamic information is modeled using an LSTM. Experiments were performed on two semi-natural datasets and one spontaneous emotional speech dataset (BAUM-1s) [51].
The main contributions of this research are as follows:

- Considering the augmented Mel spectrogram presents various emotional signals for recognizing certain emotions, this proposed study used a multiscale system for semi-natural and spontaneous datasets. We believe that this is the first work in which a multiscale framework with a local feature learning block has been used for spontaneous and semi-natural datasets for SER.

- The following layers are used to extract local level features from raw and augmented data: convolutional, batch normalization (BN), exponential linear unit, and max-pooling layers in the local feature learning block (LFLB).

- Two LSTM layers are added to build networks that connect to the LFLB to learn long-term dependencies from a series of obtained features.

- For the first time, a 1D CNN+LSTM (Model A) model can learn many emotional characteristics from raw speech datasets. However, the two-dimensional CNN+LSTM models (Model B) outperform Model A in the proposed study. The Model B was designed to learn local correlations and global contextual information from an augmented Mel spectrogram. The LFLB or LSTM layers may handle the augmented Mel spectrogram as a series or grid.

The rest of this paper is structured as follows.

The related works is given in Section 2. Details of our proposed method are given in Section 3. Results are provided in Section 4. Conclusion and future work are presented in Section 5.

II. RELATED WORKS

Distinguishing features are essential for recognizing speech emotions from audio signals. Spectrum features are among the different prosody features that are utilized in SER. AB Kandali et al. used the Gaussian mixture model (GMM) with MFCC to recognize emotions from the Assamese speech database [52]. VB Waghmare et al. used MFCCs as the main feature to identify emotions in the Marathi speech dataset [53]. After extracting MFCCs from EmoDB, Demircan, S., utilized k-NN to identify emotions [54]. Chenchah et al. employed HMM and SVM [55] to determine the spectral features obtained from raw speech data. In [56] proposed an SER approach with an auto-associative neural network (AANN) by fusing residual phase and MFCC features. AANN, SVM, and RBFNN are used to identify emotions in a music database using two acoustic features [57]. However, handcrafted features are highly effective for distinguishing emotions in audio data but are primarily low-level features. A generalized discriminant analysis (Gerda) deep neural network (DNN) layered with multiple limited Boltzmann machines (RBMs) was used in [30] to identify emotions. The results were significantly better than traditional baseline methods. In [58], the suggested approach used a regression-based DBN with three hidden layers to extract features and identify emotions from a music database.

The proposed technique [59] investigated a hybrid approach and obtained the best outcomes on FAU Aibo. The suggested research [60] used a deep neural network to detect utterance level emotions. It achieved a 20% relative accuracy increase over conventional state-of-the-art methods. In [32], developed a semi CNN approach with a linear SVM to identify emotional classes. In [61], the suggested CNN architecture was used to identify emotions from labeled datasets, and preliminary experimental outcomes noticed that this approach performed better than SVM-based classification. In [61] proposed a systematic method for developing an effective emotion detection system utilizing deep DCNNs and annotated training data. Furthermore, [15] compressed the extracted audio features using PCA. The proposed technique is different from the work mentioned above. The developed models learn local and global features to identify emotions from audio data. In general, models are only capable of identifying low-level features. Furthermore, existing methods based on CNN can only extract a single kind of emotion-related information, which is insufficient for recognizing emotions.

In [62], authors used text and audio data from the IEMOCAP database to demonstrate a double recurrent encoder framework technique that uses MFCC with text tokens as input characteristics. The suggested multimodal approach resulted in a 71.8% accuracy on the testing dataset. In [63], the suggested model is used to train two classifiers, RNN and SVM, using the CREMA-D database for accounting for voice level variation. Three intensity levels were used to train the classifiers: low, medium, and high. The emotions labeled “happy” and “neutral” have the highest categorization accuracy, whereas the emotion labeled “disgust” has the lowest. Furthermore, they did not produce any epoch-by-epoch accuracy curves or a class-by-class confusion matrix to support their claims. Therefore, in [64] proposed the remaining block and memory attention methods and 3D LMS-based, dilated CNNs. They employed a combination of the static LMS feature, the $\Delta$ (delta), and the $\Delta\Delta$ (double delta) feature to create the feature vector from raw speech. When using the speaker-dependent collection, IEMOCAP obtained 74.96% accuracy, and when using the speaker-independent raw data, it achieved 69.32% accuracy. The model obtained the best accuracy of 90.37 percent with the IEMOCAP database.

High-level information extracted from speech spectrograms is used to build an SER model proposed in [65]. It was determined how well the model performed using two different data sets. The IEMOCAP and EMO-DB datasets have an accuracy of 77.1% and precision of 92.2%, respectively. Zhang et al. [66] offer a novel multi-task learning approach. The RA VDESS database contains both voice and music samples with four emotional states, and the model acquired a 57.14% accuracy rate for selecting group multi-task feature sets. In [67], scientists collected and computed mean values for the 20-MFCC, the twenty delta, and the twenty double delta characteristics. Input for the artificial neural network algorithm was these mean values. Using the EMO-DB and RA VDESS datasets, they obtained 82.3 percent and 87.8%
accuracy with their approach. Badshah et al. proposed an SER design for a 2D CNN-based model based on the EMO-DB database [68]. A pre-trained AlexNet [69] design was also used, but the results were dismal. They also looked into transfer learning. The first suggested model had an accuracy rate of 84.3% on the test set. Previous research has shown that ensemble approaches may improve speech recognition accuracy [70]–[72] in SER tasks.

III. PROPOSED METHOD

One of the primary objectives of SER is to extract more discriminative emotional features from raw audio and augmented data. Emotional features are divided into two types: handcrafted features and deep learning features. Many handcrafted feature extraction approaches are developed carefully with ingenious strategies. Most deep learning approaches [30], [48], [73]–[75] are used to extract deep learning features from speech emotional datasets and perform well for SER. Therefore, identifying emotions using deep learning features is becoming more popular. Raw databases, environments, and discriminative features are the main concerns for SER. The proposed approach is divided into three sections: (1) preparation of the raw data, (2) learning local and global features, and (3) the architecture of Model A and Model B.

A. PREPARATION OF THE RAW DATA

The initial step is to generate a suitable input for LFLB. For this purpose, we created different lengths of Mel spectrogram segments. The generated Mel spectrogram segments have a fixed input size (227 * 227 * 3). As stated in [4], we generate three channels of spectrogram segments. These spectrogram segments are identical to the RGB format of the original 1D audio signals. We created 2D Mel spectrogram segments of size (Mel_spectrogram = W * S) for an utterance. S represents the total number of Mel filter banks (MFB), and W is the context window size. We used (S = 64) to calculate the Mel spectrogram with a 25ms window size and a 10 ms overlap [14]. Then we use a contextual window of W frames to divide the Mel spectrogram into (64 * W) segments. In [76], the used 250 ms audio clip may provide information about emotions. This result indicates that W is greater than and equal to 23 and that its segment length is 245 milliseconds. Delta (Δ) and Double Delta (ΔΔ) coefficients are generally obtained from MFCCs in speaker recognition to represent the spatio-temporal in auditory segments. Also, first and second-order spectrogram regression coefficients of Mel_spectrogram calculated using their corresponding spectrogram slice’ (Δ) coefficients and (ΔΔ) coefficients. As a result, we can get three channels of Mel spectrogram slices with a scale of (64 * W * 3). The obtained Mel spectrogram is identical to the RGB image. The bilinear interpolation is used to resize various Mel spectrogram slices into a suitable size (227 * 227 * 3) as input for LFLB to recognize features of the obtained channels of the Mel spectrogram slices. The process for generating three channels of Mel spectrogram segment slices (static, delta, and double delta) used as inputs to LFLB is shown in Figure 1. Semantic information was not included in the auditory Mel spectrogram segments. As a result, obtained features are inputted into deep learning models to generate segment features. (W = 64) in [4], [14] was utilized as the primary SER function. This research aims to determine the effect of various subgroup length W inputs on frameworks. The detailed description of datasets is illustrated in Table 1, and Table 2 shows the structure of databases.

B. DEEP FEATURE LEARNING

A novel approach for learning local and global level discriminative features from raw audio databases and augmented Mel spectrograms is presented in our study. We combined an LSTM and LFLB to learn local level and global level features. The CL of LFLB processed a grid of values G [77]–[79]. The LFLB is used to learn a feature sequence and each feature in a sequence is a function of a limited number of input features. Whereas LSTM is used for processing the G-series of numbers [47], every element of the learning features is a function of the preceding output elements. The high-level features can be learned using a combination of the CNN+LSTM approach. The CNN+LSTM method contains both long-term contextual dependencies and local information.

C. LOCAL FEATURE LEARNING

LFLB extracts emotional features from the input signal. As shown in Figure 2, LFLB consists of five layers. 1) two convolutional layers (CL), 2) batch normalization layer (BNL) [80], 3) exponential linear unit (ELU), and 4) max-pooling layer (MPL). The first, second, and fourth are the core layers of LFLB. The main advantages of the CL are spatial locality and shared weights [77]–[79]. However, spatial locality and shared weights enable CL’s learning kernel capability. At each batch, BN normalized the activation function of the CL and enhanced the stability and performance of the deep neural network. Deep neural networks...
perform better and are more stable when using BN layers. It is possible to keep the mean activation near zero and the activation standard deviation close to one by using batch normalization [81]. The ELU determines the output of the BN layer. Although ELU has negative values in contrast to other activation functions, in the suggested approach, ELU speeds up the learning process and leads to better identification accuracy [82]. By using a PL, the extracted features can be more robust against distortion and noise. The most common non-linear function is max-pooling, which divides the input on-overlapping groups and returns the highest value for each sub-group [83].

The LFLB can be customized in various ways, depending on the task. The modification in the LFLB is generally reflected in the various convolution and max-pooling parameters. A local feature extractor is performed by the convolution layer. The data convolved over the height and width of the input value using the kernels. When convolved features pass into the CL, we can obtain a feature map by computing the dot product of the input and kernel elements. Suppose a signal
s(n) is fed into a 1D CL. Convolution of the signal s(n) with the kernel k(n) and the size of the kernel z yields the output r(n). We randomly initialized the proposed approach to the 1D convolution kernel k(n).

\[ r(n) = s(n) \ast k(n) = \sum_{m=-z}^{z}s(m).k(n-m) \]  

(1)

If s(x, y) is the input of the 2D CL, the result r(x, y) is achieved by convolving a signal s(x, y) with the convolution kernel k(x, y), and size is i \ast j. In the proposed approach, we randomly initialized the 2D convolution kernel k.

\[ r(x \ast y) = s(x \ast y) \ast k(x \ast y) = \sum_{i=-a}^{a} \sum_{j=-b}^{b}s(i,j).k(x-i, y-j) \]  

(2)

The convolved features fed into BN normalized the activation function of preceding layers in every batch. The second layer of LFLB uses a transformation to maintain the mean constant. In the case of the convolved features, the variance is 1. When correlated features are input into the third layer of LFLB, output features are explained as below:

\[ r^z_x = \sigma(BN(b^z_x + \sum_{y} r^y_x \ast k^z_{xy})) \]  

(3)

where \( r^z_x \) and \( r^{y-1}_x \) defines the x output and y input features at the z layer and z-1 layer, in the above equation, \( k^z_{xy} \) represents the convolution kernel between the x and y features. The BN (\( \sigma \)) correlated with the learning features of the CL. In the proposed study, the ELU activation function is represented by the function \( \sigma(h) \), which can be written as:

\[ \sigma(h) = \begin{cases} h & h \geq 0 \\ \alpha(e^h - 1), & h < 0 \end{cases} \]  

(4)

Suppose the value of \( \alpha \) is greater than zero and \( ev \) is Euler’s value. In that case, output features are inputted into the MPL. The PL is used for non-linear down-sampling functionality, decreasing feature resolution. The extracted features obtained by the max-pooling layer are as follows:

\[ r^z_q = \max_{v \in \Omega_q} r^z_v \]  

(5)

In eq. (5), \( \Omega_q \) indicates the pooling value with the value of the index \( q \), and \( r^z_q \) and \( r^z_w \) defines the input and output features of the MPL with index \( q \) and \( w \).

**D. GLOBAL FEATURES LEARNING**

The architecture of LSTM is the same as a recurrent neural network (RNN) [47], [84]. LSTM specifically understands long-term dependencies from series of segments. So it is stacked upon the LFLB used to learn contextual dependencies from the sequences of extracted local level features. Because LSTM uses four components to modify the block state: an input gate, an output gate, a forget gate, and a self-recurrent unit at each time step. Let \( zp \) be the input volume and \( zq \) be the output volume of an LSTM network. The correlation between \( zp \) and \( zq \) can be written as:

\[ d^t = \sigma_p(G_a e^{z-1} + H_a e^{z-1} + k_a) \]  

(6)

\[ e^t = \sigma_p(G_e z_{t-1} + H_e z_{t-1} + k_e) \]  

(7)

\[ r^o = \sigma_p(G_r z_{t-1} + H_r z_{t-1} + k_r) \]  

(8)

\[ o_t = f_t + i_t \cdot \sigma_o(G_o e^{z-1} + H_o e^{z-1} + k_o) \]  

(9)

\[ m^t_f = o_t \cdot \sigma_m z_t \]  

(10)

where \( o_t \) is the LSTM unit state; \( G, H, \) and \( k \) denote parameter matrices and vectors; \( a_t, e_t, \) and \( r_t \) denote gate vectors; \( p \) denotes a sigmoid function; \( \sigma_m \) and \( \sigma_o \) is the hyperbolic tangent; The (\( \cdot \)) operator represented the Hadamard product. In equations (6)-(10), the input and output feature indices are superscripts \( x = 1 \) and \( l \). The variables \( a, e, \) and \( r \) in the equations above represent the forget, input, and output gate. In eq. (9), \( o \) represent a cell value. The variable \( p \) in the above equations represents the gate.
TABLE 1. Detailed description of datasets.

| Datasets   | Speakers       | Emotions                                                                 | Languages                                           | Size                          | Avg. samples of each emotion per person |
|------------|----------------|---------------------------------------------------------------------------|-----------------------------------------------------|-------------------------------|----------------------------------------|
| SAVEE      | 4 (male)       | Seven emotions (sadness, neutrality, frustration, happiness, disgust, anger, surprise) | 480 utterances (120 utterances per speaker)/British English | 480 utterances (120 utterances per speaker) | 15                                     |
| IEMOCAP    | 10 (5 Male, 5 Female) | Nine emotions (surprise, happiness, sadness, anger, fear, excitement, neutral, frustration and others) | English | 12 Hours Recording | 100.3                                  |
| BAUM-1s    | 31 (13Male, 18 Female) | Six emotions (anger, joy, sadness, disgust, fear, surprise) | unscripted and unguided way in Turkish | 1222 video sample |

TABLE 2. Original SAVEE, IEMOCAP and BAUM-1s databases structure.

| Datasets   | Sad | Angry | Happy | Disgust | Neutral | Surprise | Frustration | Excited | Joy | Fear |
|------------|-----|-------|-------|---------|---------|----------|-------------|---------|-----|------|
| SAVEE      | 60  | 60    | 60    | 60      | 120     | 60       | –           | –       | –   | –    |
| IEMOCAP    | 1182| 1229  | 495   | 4       | 575     | 24       | 3830        | 2505    | –   | 135  |
| BAUM-1s    | 139 | 56    | 179   | 36      | 187     | 43       | 221(Boredom) | –       | 64  | 38   |

TABLE 3. The Model A parameters are as follows: The output is determined by the length and number. In Learning Block 1, the CL is represented by C while the max-pooling layer is represented by P.

| Name          | OutputDim | Size of Kernel | Stride |
|---------------|-----------|----------------|--------|
| Learning Block 1 | C^1*d and C^2*d | L^*128 | 4 | 1 |
| Learning Block 2 | P 1*d | L^*4*128 | 3 | 3 |
| Learning Block 3 | C 3*d and C4*h | L^*4*128 | 4 | 1 |
| Learning Block 4 | P 2*d | L^*16*128 | 3 | 3 |
| Learning Block 5 | C 5*h | L^*16*256 | 4 | 1 |
| Learning Block 6 | P 3*d | L^*64*256 | 3 | 3 |
| Learning Block 7 | C 6*h | L^*64*256 | 4 | 1 |
| Learning Block 8 | P 4*h | L^*256*256 | 3 | 3 |
| Learning Block 9 | C 7*h | L^*256*512 | 4 | 1 |
| Learning Block 10 | P 5*h | L^*512*512 | 3 | 3 |

E. ARCHITECTURE OF MODEL A

As shown in Figure 2, the architecture of Model A consists of five LFLBs, and each LFLB block consists of five layers. We apply the following rules to distinguish between different layers: 1) The number before the label specifies the network in which the building block or layer is located. 2) The number after the label specifies the index number of the layer in a network. Figure 2 depicts the general architecture of the proposed model. Model A is based on a deep learning method that learns from raw datasets. Consequently, in each LFLB, the convolution and pooling layers are one-dimensional. The kernel size for the first and second blocks is 128; for the third and fourth blocks, it is 256. The kernel size and stride of all the max-pooling layers are three. Model A’s parameters are illustrated in Table 3, and softmax is the last layer of Model A used to recognize emotions.

Next, a one-dimensional vector representing an audio clip is inputted into Model A, where LFLBs learn local features. After being reshaped, the output features from the five one-dimensional learning blocks are given to the LSTM layers. Finally, the contextual dependencies are identified from the inputted local hierarchical properties. Figure 3 illustrates the local level extracted features and contextual dependencies. So, the output of the LSTM layers comprises local and long-term contextual dependencies. Next, output features are inputted into FCL, followed by two LSTM layers. Below is the equation for the FCL:

\[ s^l = b^l + s^{l-1} \cdot kl \]  \hspace{1cm} (11)

The softmax layer is used as a classifier. The softmax layer makes predictions based on the input features. The Softmax function is described as follows:

\[ z_i = \sum_j h_j W_{ji} \] \hspace{1cm} (12)

\[ \text{softmax}(z_i) = p_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \] \hspace{1cm} (13)

where softmax input is \( Z_i \), \( W_{ij} \) is the weight, and \( h_j \) is the activation function. So, the predicted class label \( \hat{u} \) represented as:

\[ \hat{u} = \text{argmax} p_i \] \hspace{1cm} (14)

F. ARCHITECTURE OF MODEL B

The architecture of Model B is similar to that of Model A. As shown in Figure 2, Model B consists of five LFLBs,
TABLE 4. The parameters for Model B are shown in the table below, and output is specified by height * width * digit, where M indicates the low-level features. In 2D LFLB, convolutional and maximum pooling layers are denoted as C1^{st} and P1^{st}, respectively.

| Name | OutputDim | Size of Kernel | Stride |
|------|-----------|----------------|--------|
| Learning Block 1 | C1^{st} and C2^{nd} | M^*N^*64 | 2*2 | 1*1 |
| Learning Block 1 | P1^{st} | M/2^*N/2^*64 | 2*2 | 2*2 |
| Learning Block 2 | C3^{rd} and C4^{th} | M/2^*N/2^*64 | 2*2 | 1*1 |
| Learning Block 2 | P2^{nd} | M/8^*N/8^*64 | 2*2 | 2*2 |
| Learning Block 3 | C5^{th} | M/8^*N/8^*128 | 2*2 | 1*1 |
| Learning Block 3 | P3^{rd} | M/32^*N/32^*128 | 3*3 | 3*3 |
| Learning Block 4 | C6^{th} | M/32^*N/32^*128 | 2*2 | 1*1 |
| Learning Block 5 | P4^{th} | M/128^*N/128^*128 | 3*3 | 3*3 |
| Learning Block 5 | C7^{th} | M/128^*N/128^*128 | 2*2 | 1*1 |
| Learning Block 5 | P5^{th} | M/128^*N/128^*128 | 3*3 | 3*3 |

TABLE 5. The comparison of classification accuracies of Model A and Model B.

| Database              | Type       | Raw audio data | Augmented Mel spectrogram |
|-----------------------|------------|----------------|----------------------------|
| Accuracy of SD Experiments |           |                |                            |
| SAVEE dataset         | Semi-Natural | 94.76          | 97.19                      |
| IEMOCAP dataset       | Semi-Natural | 73.13          | 84.09                      |
| BAUM-1s dataset       | Spontaneous  | 31.61          | 53.98                      |

| Database              | Type       | Raw audio data | Augmented Mel spectrogram |
|-----------------------|------------|----------------|----------------------------|
| Accuracy of SI Experiments |       |                |                            |
| SAVEE dataset         | Semi-Natural | 87.54          | 96.85                      |
| IEMOCAP dataset       | Semi-Natural | 69.33          | 85.80                      |
| BAUM-1s dataset       | Spontaneous  | 29.10          | 48.67                      |

TABLE 6. The average classification accuracy of Model B was compared to that of other feature extraction and classification models using the SAVEE database.

| Research work | Accuracy of SD | Accuracy of SI |
|---------------|----------------|----------------|
| [3]           | 82.10          | 66.90          |
| [74]          | 88.3           | 85.2           |
| [75]          | –              | 92.9           |
| [85]          | 75.5           | –              |
| [86]          | 91.6           | 85.8           |
| Our Work      | 97.19          | 96.85          |

TABLE 7. Comparison of Model B recognition accuracy with various feature extraction approaches and classification methods using the IEMOCAP dataset.

| Research work | Accuracy of SD | Accuracy of SI |
|---------------|----------------|----------------|
| [3]           | 83.80          | 76.60          |
| [61]          | 91.6           | 85.8           |
| [73]          | 88.3           | 85.2           |
| [64]          | 79.96          | 69.32          |
| Our Work      | 94.09          | 88.80          |

G. HYPERPARAMETER OPTIMIZATION

It is crucial to select hyperparameters for a neural network before moving further. Hyperparameter optimization aims to maximize a deep neural network’s efficiency on a database independent of the deep neural network under evaluation. Confusion matrix, arbitrary search, and other search strategies were all effectively used in various deep learning models. They all help improve the training of the deep model. The Bayesian optimization technique produces higher performance with fewer testing datasets. [88]–[90]. The Bayesian optimization technique is used in our studies to choose hyperparameters for suggested DNNs. Bayesian optimization is a sequential design approach that effectively solves the optimization problem. In our studies, we utilized Hyperopt to maximize the hyperparameters [90]. Hyperopt provides a

input for the first LSTM layer. Local features are used to learn contextual dependencies. Figure 4 depicts the learning of local-level features and contextual dependencies. As a result, the LSTM layer’s output comprises spatial correlations and global contextual information. The FCL is used to categorize these features into the output space. Softmax is used for classification using learned features.
TABLE 9. Recognition accuracy of Mid Mel spectrogram techniques.

| Types | BAUM-1s | IEMOCAP | SAVEE |
|-------|---------|----------|--------|
|       | SD | SI | SD | SI | SD | SI |
| Mid  | 42.40 | 36.88 | 75.78 | 70.98 | 77.64 | 79.09 |

minimizer-friendly optimal solution and analyzes it like a randomized function. The goal function is also given prior knowledge. The probability distributions over the optimization problem are modified following the collected function evaluations. Using probability distributions, we may generate an acquired function. Recursively, hyperparameters were chosen. As a first step in selecting an optimization technique, a distribution over the variables is chosen (adagrad, Adam, SGD, and RMS). The proposed design is presented once the learning with optimal parameters has been performed.

IV. RESULTS

The proposed models’ performance was evaluated using semi-natural and spontaneous databases for SD and SI experiments. In addition, we evaluated our models using two experiments. The first experiment was performed using raw audio samples, and the second was performed using an augmented Mel spectrogram. Model A is used to extract features from raw audio samples in the presented study. In contrast, Model B extracts high-level features from augmented Mel spectrogram.

Additionally, the developed models are used for predictive capacity instead of limited explanatory power. Numerous methods are presented to reduce the possibility of overfitting in proposed studies. Overfitting is a factor in bad predictions for untrained datasets because overfit models memorize training sets instead of learning to predict better. Although overfitting occurs for numerous reasons, (1) when the model’s architecture is very complicated, (2) Overfitting occurs when a learning model becomes overtrained. (3) when the model degrees of freedom are too high [91]. Many techniques have been developed to minimize overfitting, including regularization [92], BN [81], cross-validation [93], early stopping [93], and model-selection [94].

A. SPEAKER DEPENDENT (SD) EXPERIMENTS

First, we performed SD experiments on both augmented Mel spectrogram and raw audio speech signals. The datasets were randomly divided into an 80:20 ratio for training and testing. Our experiment results imply that the proposed models efficiently identify speech emotions. The main goal of the proposed study is to identify emotions with high accuracy and generalization performance. Therefore, the best-predicted model is reported in our experiments. Figure 5 illustrates the results achieved from the SAVEE database for SD experiments. Model B recognized “neutral”, “angry”, “frustration”, and “sad” with the highest accuracies of 100%, 97.58%, 97.43%, and 97.12% respectively, with the SAVEE dataset with raw audio data. While Model B achieved the average accuracy of 97.19% with an augmented Mel spectrogram. As shown in Figure 6, the SAVEE database identified “anger” and “sad” with the highest accuracy of 100%. At the same time, “disgust,” “surprise,” and “neutral” were recognized with the highest accuracies of 98.66%, 98.13%, and 96.48% with the SAVEE datasets of an augmented Mel spectrogram, respectively. Model B recognized “frustration”, “happy”, and “anger” with the accuracies of 51.07%, 47.66%, and 45.20% with the IEMOCAP dataset of raw audio data, as illustrated in Figure 7. Figure 8 illustrates the IEMOCAP dataset contains four emotions, “sad”, “anger”, “happy”, and “neutral”, which are listed with accuracies of 97.28%, 94.66%, 56.66%, and 89.37%, with an augmented Mel spectrogram, respectively. Model B achieved 69.53% and 85.34% average accuracy with raw audio and augmented Mel spectrogram for the IEMOCAP database, respectively. As shown in Figure 9, the BAUM-1s database identified “joy” with the highest accuracy
of 75.99%. Model B achieved average accuracy with the BAUM-1s dataset of raw audio data is 40.57% for SD experiments. Model B achieved average accuracy with the BAUM-1s dataset of the augmented Mel spectrogram is 53.98% for SD experiments, as illustrated in Figure 10.

Validation accuracy is an important metric for assessing the generalization performance of a training model. The optimal model will be there when the validation accuracy of Model A and Model B achieves its maximum during the training process. The proposed study selects the best predictive model to minimize overfitting. The prediction performance of Model A and Model B improves as validation accuracy decreases. Overfitting occurs when the accuracy of validation decreases while training continuously increases. So, the training process will be early stopping. Avoiding overtraining and improving the model’s performance can be achieved by early stopping. To achieve the highest classification performance, we evaluated validation accuracy in our experiments. When validation accuracy in training no longer increases, the model will have higher prediction accuracy.

### B. SPEAKER-INDEPENDENT (SI) EXPERIMENTS

SI experiments were conducted using the same approach as SD experiments. However, the distribution of the dataset for IEMOCAP was different for SI experiments. The data were divided into two groups for the SI experiment depending on the subject. Since the utterance of emotions from the selected database was performed by twelve speakers, data from nine subjects was selected as the training sample, and data from the other three subjects was selected as the testing sample. The suggested model fitted the experimental data and has better predictive performance. The obtained results for SI experiments are shown in Figs. 11–15 to analyze the individual emotional groups’ identification accuracy. Model B achieved average accuracy with the SAVEE and IEMOCAP databases at 96.85% and 88.80%, respectively, with augmented Mel spectrograms for SI experiments. While with raw data, the obtained accuracy is 87.54% and 69.53%, respectively. The best-fitted and predictive models are recorded when the proposed model achieves the highest validation accuracy during training. Therefore, the suggested model is more accurate at fitting the experimental data and has higher classification accuracy. Figure 11 shows the average accuracy achieved by Model B with the SAVEE database is 87.54%.

The SAVEE database contains seven emotion categories, four of which, “anger”, “sad”, “ surprise,” and frustration, were identified with accuracies of 100%, 99.56%, 98.32%, and 98.28%, respectively, by Model B with augmented
Mel spectrogram as shown in Figure 12. In contrast, the other three emotions were identified with less than 98.00% accuracy, as represented in Figure 12. Model B achieved an average accuracy with the IEMOCAP database of 88.80%. Figure 13 shows that the IEMOCAP database, “anger,” “sad,” and “excited” were recognized with accuracies of 96.73%, 99.52%, and 93.22%, respectively, by the Model B with an augmented spectrogram. As shown in Figure 14, Model B recognized “sad” and “neutral” with the highest accuracies of 99.02% and 96.38% with the IEMOCAP dataset. Model B achieved average accuracy with the BAUM-1s dataset of the augmented Mel spectrogram of 48.67% for SI experiments.

C. RECOGNITION ACCURACY COMPARISON

Table 5 shows that the Model-A can learn emotional characteristics from raw audio samples to identify speech emotions. Furthermore, compared to Model-A, Model-B has significant advantages. The obtained average recognition and validation accuracy with the augmented Mel spectrogram are higher than the obtained accuracy from raw data. Model B achieved the best validation accuracy with fewer epochs and converged quicker than Model A. The proposed Model B performed adequately compared to other feature extraction and techniques. Also, Model-B is quicker in convergence than Model-A.

D. DISCUSSION

Model A and Model B are comprised of five LFLBs and two LSTM layers to learn local level features and global level features. Because speech signals are time-varying signals and require complex evaluation to analyze time-varying features, the CNN+LSTM approach is proposed to recognize emotional states. Although this study has successfully acquired...
Suggested approach is used to generate suitable inputs for the simultaneous databases with an augmented Mel spectrogram. We proposed a new SER approach for semi-natural and spontaneous speech emotions. Tests are performed on numerous handcrafted parameters. These attempts have been effective in recognizing emotions; tests are performed every layer. Also, we discover whether handmade features are effective in recognizing emotions; tests are performed on numerous handcrafted parameters. These attempts have allowed us to disclose additional information about the DNN models employed to evaluate speech.

E. BLACK BOX

In the past few years, experts have started to investigate the “black box” issue to understand what is going on inside. Google researchers developed a new approach for the image classification model in 2015 to determine which features are utilized for classification. During the same year, researchers from the University of Wyoming identified how specific images might deceive a system by evaluating DNNs. A software engineer and a neurologist proposed the “information bottleneck” in 2017 [95]. Lehigh University developed DeepXplore to analyze neural networks by evaluating millions of neurons [96]. Stanford University [97] introduced ReLuPlex based on mathematical arguments to validate the features of DNNs. Although these methods have taken a significant step in image classification, it is still not a universal answer to the “black box” issue [98], [99]. Additionally, we have studied significantly to understand better the developed DNNs employed to evaluate the speech. In the proposed study, we determine the impact of fundamental parameters of the proposed networks on classification results, and multiple models with changing layers and filters are counted at every layer. Also, we discover whether handmade features are effective in recognizing emotions; tests are performed on numerous handcrafted parameters. These attempts have allowed us to disclose additional information about the DNN in the experiments.

V. CONCLUSION

We proposed a new SER approach for semi-natural and spontaneous databases with an augmented Mel spectrogram. The suggested approach is used to generate suitable inputs for the 1D (Model A) and 2D (Model B) CNN+LSTM framework from an original audio dataset and develop appropriate deep models for feature learning. The proposed method learns local and global features from raw data and augmented Mel spectrogram. We used five LFLB blocks to extract local-level features from inputted data. Local features are inputted into the LSTM layer to understand contextual correlations. Moreover, features extracted by proposed models consist of local and long-term contextual dependency.

The overall performance of the proposed approach was analyzed on spontaneous and semi-natural databases. We noted that Model A and B extract discriminative features and represent high-level abstractions of speech datasets. The proposed approach showed that the overall accuracies of Model B are higher than other feature extraction and state-of-the-art approaches. However, the DNNs discussed in this study have improved their performance in speech emotion detection. However, several areas still need to be addressed. First, the mechanism by which the proposed networks identify emotions cannot be fully described. The “Black box” of both models has not been investigated. However, most studies focused on the deep learning techniques employed in image processing. Speech is distinct from images; elucidating the “Black Box” of deep networks optimizing speech processing requires extensive research. Secondly, achieving better accuracy in SER is not the final goal. A novel approach capable of learning more specific features or training a more accurate prediction model must be investigated. Finally, to maximize the advantages of different extracted features, create a mechanism for combining different deep features acquired by different deep learning models.
A. Amjad, L. Khan, and H.-T. Chang, “Semi-natural and spontaneous speech emotion recognition using deep neural networks,” IEEE Access, vol. 10, 2022.
T. Anvarjon, Mustaqeem, and S. Kwon, “Deep-Net: A lightweight CNN-based speech emotion recognition system using deep frequency features,” Sensors, vol. 20, no. 18, p. 5212, Sep. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/18/5212

B. Zhang, E. M. Provost, and G. Essi, “Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2016, pp. 5805–5809.

P. Nantasri, E. Phaisangitsaagul, J. Karnjana, S. Boonkla, S. Keerativitanawat, A. Rugehataron, S. Usanavavasin, and T. Shinoraki, “A lightweight artificial neural network for speech emotion recognition using average values of MFCCs and their derivatives,” in Proc. 17th Int. Conf. Electr. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON), Jun. 2020, pp. 41–44.

A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, “Speech emotion recognition from spectrograms with deep convolutional neural network,” in Proc. Int. Conf. Platform Technol. Service (PleCon), Feb. 2017, pp. 1–5.

A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, vol. 25, F. Pereira, J. C. B. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/39998323dbd6b76c9a436e924a48c545b-Paper.pdf

A. Bhavan, P. Chauhan, Hiruk, and R. R. Shah, “Bagged support vector machines for emotion recognition from speech,” Knowl.-Based Syst., vol. 184, No. 10, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705119303533

K. Zvarevase and O. Olugbara, “Ensemble learning of hybrid acoustic features for speech emotion recognition,” Algorithms, vol. 13, no. 4, p. 70, Mar 2020. [Online]. Available: https://www.mdpi.com/1999-4893/13/3/70

C. Zheng, C. Wang, and N. Jia, “An ensemble model for multi-level speech emotion recognition,” Appl. Sci., vol. 10, no. 1, p. 205, Dec. 2019. [Online]. Available: https://www.mdpi.com/2076-3417/10/1/205

Y. Kim, H. Lee, and E. M. Provost, “Deep learning for robust feature generation in audiovisual emotional recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., May 2013, pp. 3687–3691.

Z. Huang, M. Dong, Q. Mao, and Y. Zhan, “Speech emotion recognition using CNN,” in Proc. 22nd ACM Int. Conf. Multimedia (MM), New York, NY, USA: Association for Computing Machinery, Nov. 2014, pp. 801–804, doi: 10.1145/2647878.2654984.

S. Behnke, “Discovering hierarchical speech features using convolutional neural networks,” in Proc. 9th Int. Conf. Artif. Neural Netw., vol. 2, 1999, pp. 850–855.

Proc. 17th Int. Conf. Electr. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON), Jun. 2020, pp. 41–44.

A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, “Speech emotion recognition from spectrograms with deep convolutional neural network,” in Proc. Int. Conf. Platform Technol. Service (PleCon), Feb. 2017, pp. 1–5.

A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, vol. 25, F. Pereira, J. C. B. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012. [Online]. Available: https://proceedings.neurips.cc/paper/2012/file/39998323dbd6b76c9a436e924a48c545b-Paper.pdf

A. Bhavan, P. Chauhan, Hiruk, and R. R. Shah, “Bagged support vector machines for emotion recognition from speech,” Knowl.-Based Syst., vol. 184, No. 10, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705119303533

K. Zvarevase and O. Olugbara, “Ensemble learning of hybrid acoustic features for speech emotion recognition,” Algorithms, vol. 13, no. 4, p. 70, Mar 2020. [Online]. Available: https://www.mdpi.com/1999-4893/13/3/70

C. Zheng, C. Wang, and N. Jia, “An ensemble model for multi-level speech emotion recognition,” Appl. Sci., vol. 10, no. 1, p. 205, Dec. 2019. [Online]. Available: https://www.mdpi.com/2076-3417/10/1/205

Y. Kim, H. Lee, and E. M. Provost, “Deep learning for robust feature generation in audiovisual emotional recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., May 2013, pp. 3687–3691.

Z. Huang, M. Dong, Q. Mao, and Y. Zhan, “Speech emotion recognition using CNN,” in Proc. 22nd ACM Int. Conf. Multimedia (MM), New York, NY, USA: Association for Computing Machinery, Nov. 2014, pp. 801–804, doi: 10.1145/2647878.2654984.

S. Demircan and H. Kahramanli, “Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech,” Neural Comput. Appl., vol. 29, no. 8, pp. 59–66, Apr. 2018.

E. M. Provost, “Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., May 2013, pp. 3682–3686.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” Neural Comput., vol. 1, no. 4, pp. 541–551, 1989.

B. Boser, E. Sackinger, J. Bromley, Y. LeCun, R. Howard, and L. Jackel, “An analog neural network processor and its application to high-speed character recognition,” in Proc. Int. Joint Conf. Neural Netw. (IJCNN-Seattle), vol. 1, 1991, pp. 415–420.

S. Behnke, “Discovering hierarchical speech features using convolutional non-negative matrix factorization,” in Proc. Int. Joint Conf. Neural Netw., vol. 4, 2003, pp. 2758–2763.

S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. 32nd Int. Conf. Mach. Learn., vol. 37, F. Bach and D. Blei, Eds. Lille, France: PMLR, Jul. 2015, pp. 448–456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html

S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” CoRR, vol. abs/1502.03167, pp. 1–11, Feb. 2015.

D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (ELUs),” 2015, arXiv:1511.07289.

D. Ciresan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2012, pp. 3642–3649.

F. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: Continual prediction with LSTM,” in Proc. 9th Int. Conf. Artif. Neural Netw. (ICANN), vol. 2, 1999, pp. 850–855.

Y. Huang, A. Wu, G. Zhang, and Y. Li, “Speech emotion recognition based on Coiflet wavelet packet cepstral coefficients,” in Pattern Recognition, S. Li, C. Liu, and Y. Wang, Eds. Berlin, Germany: Springer, 2014, pp. 436–443.
AMMAR AMJAD received the master’s degree in computer science from the National College of Business Administration and Economics, in March 2017. He is currently pursuing the Ph.D. degree in electrical engineering with the Division of Computer Science and Information Engineering, Chang Gung University, Taiwan. His research interests include speech processing, language learning, speech analysis, speech synthesis, voice pathologies, auditory neuroscience, and machine learning.

LAL KHAN was born in D G Khan, Punjab. He received the M.S. degree in computer science from the Federal Urdu University of Arts, Science and Technology, Islamabad, in 2017. He is currently a Ph.D. Scholar with the Department of Computer Science and Information Engineering, Chang Gung University, Taiwan. He is also working in NLP task for resource-deprived languages. His research interests include machine learning, deep learning, natural language processing (NLP), and speech recognition.

NOMAN ASHRAF received the Ph.D. degree from the Centro de Investigación en Computación, Instituto Politécnico Nacional (IPN). His specialization lies in natural language processing (NLP). His research interests include hate speech detection, depression detection, and emotion detection. He joined Mayo Clinic as a Research Scholar and currently working on several projects related to Breast Cancer (BC) Prediction from Social Media (SM) and Clinical Text Datasets. Before joining Mayo Clinic, he was working as a Machine Learning Engineer at International Consulting Associates, Inc.

MUHAMMAD BILAL MAHMOOD received the master’s degree in computer engineering from the National University of Sciences and Technology, Islamabad, Pakistan, in 2015. He is currently a Ph.D. Scholar with the Department of Software Engineering, Dalia University of Technology, China. His research interests include natural language processing (NLP), the IoT, and cloud computing.

HSIEN-TSUNG CHANG received the M.S. and Ph.D. degrees from the Department of Computer Science and Information (CSIE), National Chung Cheng University, in July 2000 and July 2007, respectively. He joined the Faculty of Computer Science and the Information Engineering Department, Chang Gung University and worked as an Associate Professor. He is also a member of the Artificial Intelligence Research Center, Chang Gung University, and the Department of Physical Medicine and Rehabilitation, Chang Gung Memorial Hospital. His research interests include artificial intelligence, natural language processing, information retrieval, big data, web services, and search engines. He is the Director of the Web Information and Data Engineering Laboratory (WIDE Laboratory).