IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 7, 2020, accepted February 8, 2020, date of publication February 12, 2020, date of current version February 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2973541

Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network

RASHID JAHANGIR[1,2], YING WAH TEH[1], NISAR AHMED MEMON[3], GHULAM MUJTABA[4], MAHDI ZAREEI[5], (Member, IEEE), UZAIR ISHTIAQ[1,2], MUHAMMAD ZAHEER AKHTAR[2], AND IHSAN ALI[1]

[1]Department of Information Systems, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia
[2]Department of Computer Science, COMSATS University Islamabad, Vehari Campus, Vehari 61100, Pakistan
[3]College of Computer Sciences and Information Technology (CCSIT), King Faisal University, Al Ahsa 31982, Saudi Arabia
[4]Center of Excellence for Robotics, Artificial Intelligence and Blockchain, Department of Computer Science, Sukkur IBA University, Sukkur 65200, Pakistan
[5]Escuela de Ingeniería y Ciencias, Zapopan, Tecnológico de Monterrey, Zapopan 45138, Mexico

Corresponding authors: Rashid Jahangir (rashid@siswa.um.edu.my), Ying Wah Teh (tehyw@um.edu.my), and Ihsan Ali (ihsanalichd@siswa.um.edu.my)

This work was supported in part by the Ministry of Higher Education Malaysia under Grant FRGS-FP111-2018A, and in part by the University of Malaya under Postgraduate Research Grant under Grant PG035-2016A.

ABSTRACT Speaker identification refers to the process of recognizing human voice using artificial intelligence techniques. Speaker identification technologies are widely applied in voice authentication, security and surveillance, electronic voice eavesdropping, and identity verification. In the speaker identification process, extracting discriminative and salient features from speaker utterances is an important task to accurately identify speakers. Various features for speaker identification have been recently proposed by researchers. Most studies on speaker identification have utilized short-time features, such as perceptual linear predictive (PLP) coefficients and Mel frequency cepstral coefficients (MFCC), due to their capability to capture the repetitive nature and efficiency of signals. Various studies have shown the effectiveness of MFCC features in correctly identifying speakers. However, the performances of these features degrade on complex speech datasets, and therefore, these features fail to accurately identify speaker characteristics. To address this problem, this study proposes a novel fusion of MFCC and time-based features (MFCCT), which combines the effectiveness of MFCC and time-domain features to improve the accuracy of text-independent speaker identification (SI) systems. The extracted MFCCT features were fed as input to a deep neural network (DNN) to construct the speaker identification model. Results showed that the proposed MFCCT features coupled with DNN outperformed existing baseline MFCC and time-domain features on the LibriSpeech dataset. In addition, DNN obtained better classification results compared with five machine learning algorithms that were recently utilized in speaker recognition. Moreover, this study evaluated the effectiveness of one-level and two-level classification methods for speaker identification. The experimental results showed that two-level classification presented better results than one-level classification. The proposed features and classification model for identifying a speaker can be widely applied to different types of speaker datasets.

INDEX TERMS Speaker identification, MFCCT, pattern recognition, LibriSpeech, hierarchical classification, deep neural networks.

I. INTRODUCTION

Automatic speaker identification (ASI) is the process of extracting the identity of a speaker by using a machine from a group of familiar speech signals. Speech signals are powerful media of communication that always convey rich and useful information, such as emotion, gender, accent, and other unique characteristics of a speaker. These unique characteristics enable researchers to distinguish among speakers when calls are conducted over phones although the speakers are not physically present. Through such characteristics, machines can become familiar with the utterances of speakers, similar to humans. Speaker utterances are trained with machine learning algorithms from the collected dataset, and then speakers are identified using the test utterances.

In general, speakers can be identified using two different approaches: text-independent and text-dependent. For the

The associate editor coordinating the review of this manuscript and approving it for publication was Ahmed Farouk.

text-dependent speaker identification system, the text being spoken during testing must be exactly the same as that spoken during the training of the system. By contrast, for the text-independent speaker identification system, the speaker identification process does not depend on the text being spoken by the speaker. Furthermore, speaker recognition is divided into two processes: speaker identification and speaker verification. Speaker identification involves the identification of a speaker utterance from a group of trained speaker utterances. Then, the speaker with a high probability of test utterance is identified as the speaker. Alternatively, speaker verification involves the process of determining whether a speaker of a test utterance belongs to a group of speakers through binary classification. In this study, text-independent speaker identification task is considered due to its applications in current technological speech advancement.

Speaker recognition has become an area of intense research due to its wide range of applications, including forensic voice verification to detect suspects by government law enforcement agencies [1], [2], access control to different services, such as telephone network services [3], voice dialing, computer access control [4], mobile banking, and mobile shopping [5]. Furthermore, speaker identification systems are extensively used to improve security [6], automatic speaker labeling of recorded meetings [7], and personalized caller identification using intelligent answering machines [8]. Various studies have been conducted in the area of speaker identification. These studies utilize Mel frequency cepstral coefficients (MFCC)-based features [9], Gaussian mixture models (GMM) [9]–[11], and vector quantization [12] to identify speakers. Then, these features are fed to simple machine learning classifiers [13] to construct speaker identification models.

The major challenge in speaker identification is the extraction of discriminative features from speech signals that can elicit improved performance from classification algorithms. In this regard, many studies have proposed different feature-engineering techniques, such as MFCC, linear prediction cepstral coefficient (LPCC), power-normalized cepstral coefficient, spectral features, and time-domain features. However, the aforementioned features are inefficient for speaker recognition in complex and noisy datasets, such as LibriSpeech [14], and exhibit low classification performance. The classification performance of MFCC and LPCC degrades as a result of channel variations caused by environmental noise and magnetic interference in handsets or microphones [15]. To overcome the limitations of the aforementioned features, this study proposed a novel fusion of MFCC and time-based features (MFCCT) from speech signals for the speaker identification task. In addition, a deep neural network (DNN) [16], [17] was used to construct an artificial neural network (ANN) to identify speakers based on unique voice patterns [18]. Moreover, the proposed MFCCT and constructed deep neural network-based speaker identification system was evaluated on the publicly available LibriSpeech [14] corpus database. The main contributions of this paper are: 1. Propose efficient MFCCT based features and deep neural network (DNN) for speaker identification in large speech data to improve recognition accuracy. 2. Propose two-level hierarchical classification model to identify speakers' gender and identity. The first level identifies the gender of the speaker (i.e., male or female), whereas the second level identifies the specific identity of the speaker. 3. Evaluate the performance proposed features on well-standardized, complex, and publicly available LibriSpeech corpus database. 4. Rigorously evaluate the performance of the proposed deep neural network (DNN) model and MFCCT features by comparing their performance with baseline techniques and features; 5. Compare the suitability of the proposed hierarchical classification model of two-level with one-level classification models. 6. To the best of our knowledge, this study is the first to evaluate efficient Mel frequency cepstral coefficients (MFCC)-based features and time domain feature fusion (MFCCT) for speaker identification. Moreover, existing techniques were evaluated on a small corpus that contained speaker utterances with minimal length and varying sample rate while the proposed technique is compared with large speech dataset.

The rest of this paper is organized as follows. Section II describes existing works on speaker identification. Section III presents the corpus used for the experiments, feature extraction process, classification process, evaluation metrics, and different experimental settings. Section IV reports the results of different experimental setting. Section V discusses the significance of the observed findings. Finally, Section VI concludes this paper.

II. LITERATURE REVIEW

The field of artificial intelligence combined with cognitive science is rapidly growing. It includes design and development of various real-time applications such as speech recognition, decision making, face recognition, and DNA analysis. Recently, voice biometrics have been utilized to authenticate individual identification.

Human voice is the most useful medium of communication due to its features of simplicity, uniqueness, and universality. In comparison with other biometric verification systems, the benefits of speaker identification are as follows: 1. Voice is easily accessible, easy to use and costs are low. 2. Voice is very easy to obtain and comparatively simpler for users to recognize people.

As speech recognition systems need to operate under a wide variety of conditions, therefore, such systems should be robust to extrinsic variations induced by a number of acoustic factors such as transmission channel, speaker differences and background noise. In order to enhance classification performance, most of the speech applications perform digital filter, where the clean utterance estimation is learnt by passing noisy utterance through a linear filter. With such concept, the subject of noise reduction becomes how to design a best filter that can considerably remove noise without noticeable loss of useful information. Therefore, several researchers are investigating how to minimize the effects of environmental noise in order to correctly classify speech signals. For instance, Lim, *et al.* [19] proposed spectral subtraction, which overlays a slight background noise over the speech signals. In the speech signals, those components equivalent to the noise will be hidden. However, spectral subtraction can also destroy several spectral features in the original speech signal [20], leads to the loss of some valuable features. In order to overcome this issue, support vector machine (SVM) [21] classifies speech features into various classes, aiming to minimize the difference among speech features of same class to enhance classification accuracy. Nonetheless, this approach often needs a large number of training utterances and is not beneficial for timely response applications.

A. RELATED STUDIES

Human voice is used in the universal practice to exchange information with one another. Speaker recognition refers to the identification of speakers based on the vocal features of the human voice. Speaker recognition has become an area of intense research due to its wide range of applications, such as forensic voice verification to identify suspects by government law enforcement agencies [1], [2]. Feature extraction in the speaker recognition process plays a key role because it significantly affects the performance of a speaker recognition classification model. In recent years, various researchers in the area of speaker recognition have proposed novel features that have been proven useful in effectively classifying human voices. Murty, *et al.* [22] extracted residual phase and MFCC features from 149 male speaker utterances from the NIST 2003 dataset to form a master feature vector. The authors fed the extracted MFCC features as input to an auto-associative neural network classifier and obtained approximately 90% classification accuracy. Nonetheless, the proposed features and classifier may be ineffective for complex datasets, such as LibriSpeech. Fong, *et al.* [23] performed a comparative study to classify speakers using various time-domain statistical features and machine learning classifiers; they obtained the highest accuracy of approximately 94% by using the multilayer perceptron classifier. Although, the experimental results of the study achieved good classification accuracy, the results cannot be generalized to a wider scale because the authors used only 16 speaker voices from the PDA speech dataset in the experiment. In addition, the study used a small amount of speaker utterances in the training and testing sets. Ali, *et al.* [24] recently proposed a speaker identification model for identifying 10 different speakers using the Urdu language dataset. The study fused deep learning-based and MFCC features to classify speakers using a support vector machine (SVM) algorithm. The experimental results achieved 92% classification accuracy. Hence, the results are promising. However, the dataset used in the experiments suffers from several weaknesses. First, only 10 speaker utterances were used in the experiments. Second, each utterance comprised only one word. Thus, the fusion-based features proposed by the authors may be inefficient and ineffective for complex human voices. Soleymanpour, *et al.* [25] investigated clustering-based MFCC features coupled with an ANN classifier to categorize 22 speakers from the ELDSR dataset. The experimental results of the study achieved 93% classification accuracy. Laptik, *et al.* [26] and Prasad, *et al.* [27] evaluated MFFC features and a GMM classifier to classify 50 and 138 speakers from the CMU and YOHO datasets, respectively. The results of the experiment that used the proposed feature extraction methods exhibited 86% and 88% classification accuracy. Nidhyananthan, *et al.* [28] proposed a set of discriminative features to classify 50 speaker utterances from the MEPCO speech dataset. These authors extracted RASTA-MFCC features to classify speaker utterances. The extracted features were inputted into a GMM-universal background model classifier to learn the classification rules. The results achieved 97% classification accuracy. Although the results demonstrated reasonable classification accuracy, they cannot be applied to a wider scale because the study only utilized six utterances, with one utterance lasting for only 3 s. Therefore, for speaker utterances that are over 3 s long, RASTA-MFCC features may prove to be insignificant. To address the issues in the existing literature, Panayotov, *et al.* [14] provided a standard and complex speaker utterance dataset, called "LibriSpeech," for the speaker identification problem. The well-known MFCC features did not exhibit promising results when they were extracted from the LibriSpeech dataset and fed to a classifier. To improve the classification accuracy of the LibriSpeech dataset for speaker identification, the present study proposed novel MFFCT features to classify speaker utterances. Furthermore, a DNN was applied to the extracted MFCCT features to construct a speaker identification model. The details of the proposed features and model are discussed in the subsequent sections.

III. PROPOSED METHODOLOGY

This section describes in detail the methodology (Figure 1) used to identify the speakers. First, several speaker utterances were collected for the experiments. Second, various useful features were extracted from the collected speaker utterances to form a master feature vector. This master feature vector was then fed as an input to a feed forward deep neural network architecture to construct the speaker identification model. To investigate the classification performance of the proposed speaker identification model, two performance metrics, namely, overall accuracy and AUROC (Area Under the Receiver Operating Characteristics), were used. Finally, the performance of the constructed model was evaluated using a separate test set and existing speaker identification baseline techniques. The details of these methods are discussed in subsequent sections.

A. DATASET

The LibriSpeech [14] corpus was used for the experiments conducted in this study. This corpus is publicly available and is prepared from audiobooks of the LibriVox with careful

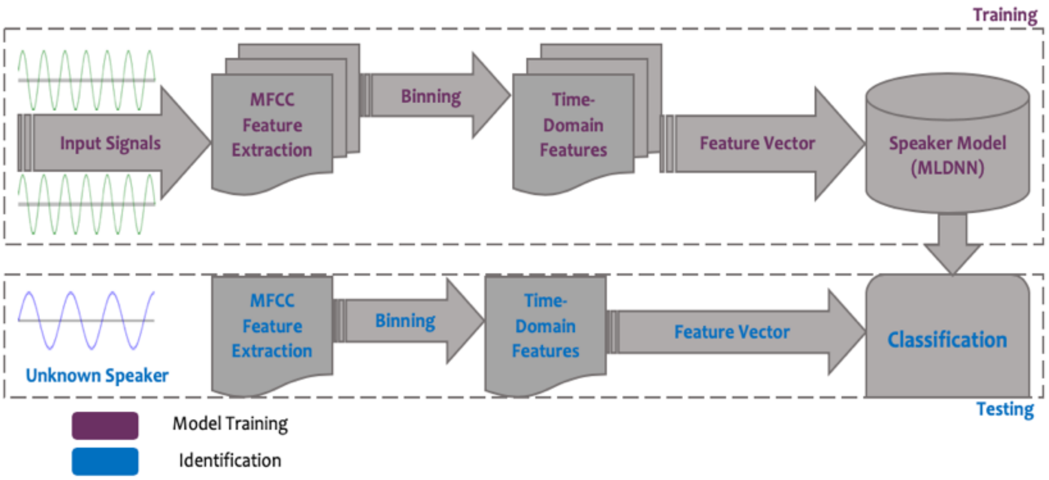


FIGURE 1. Proposed methodology for speaker identification system.

TABLE 1. Selected LibriSpeech dataset for the experiment.

Gender	Speaker	Utterances			Length
		Total	Training	Testing	
Male	50	5609	4505	1104	19h 47m
Female	50	5764	4634	1130	20h 03m

segmentation and alignment to develop automated speech recognition and speaker identification models using machine learning and deep learning techniques. LibriSpeech includes English speeches related audio files that belong to male and female of various accents but majority are USA English. All the utterances in this dataset are sampled at 16 kHz frequency and sample size of 16 bits. This corpus includes five different training and testing sets for developing an automatic speaker identification model. In this study, one subset of the corpus, i.e., *train-clean-100*, was considered for the experiments because it includes 100 h and 25 min of speeches of male and female speakers with several utterances. In addition, 50 male and 50 female speakers were selected from this dataset for the experiment (Table 3). Moreover, 80% and 20% of the utterances of each male and female speaker were used for training and testing, respectively. Each speaker served as a class label in the selected corpus to identify the speaker through MFCCT features and DNN [16] architecture.

B. SPEECH PRE-PROCESSING

Speech signals pre-processing is very critical phase in the systems where background-noise or silence is completely undesirable. Systems like automatic speaker identification and speech recognition requires efficient feature extraction approaches from speech signals where most of the spoken portion includes speaker-related attributes. Therefore, in this study pre-emphasis and silence removal techniques were employed.

The pre-emphasis method increases the strength of high frequencies of speech signal, while the low frequencies remain in their original condition in order to improve the signal-to-noise ratio. Pre-emphasis works by enhancing the high-frequency energy through applying high-pass filter (FIR) which is equivalent to

$$H(z) = 1 - \alpha z^{-1}, \quad \alpha = [1, -0.97] \tag{1}$$

where α is pre-emphasis coefficient.

FIR inevitably changes distribution of energy across frequencies along with overall energy level. This could have critical impact on the acoustic features related to energy [29]. On the other hand, signal normalization makes the speech signals comparable irrespective of variations in magnitude by using Eq 2.

$$S_{Ni} = \frac{S_i - \mu}{\sigma} \tag{2}$$

where $\mathrm{S_i}$ is the $\mathrm{i^{th}}$ part of signal S, σ and μ are are the standard deviation and mean of S respectively, $\mathrm{S_{Ni}}$ is the normalized $\mathrm{i^{th}}$ part of signal S.

C. FEATURE ENGINEERING

In general, classification performance relies on the quality of a feature set. Thus, irrelevant features may produce less accurate classification results. In deep learning and machine learning, extracting discriminative feature sets is an important task to obtain reasonable classification performance [30]. Moreover, the authors of [30] concluded that the feature engineering step is a key step in machine learning and deep learning because the success or failure of any speaker identification model heavily depends on the quality of the features used in the classification task. If the extracted features correlate well with the class, then classification will be easy and accurate. By contrast, if the extracted features do not correlate well with the class, then the classification task will be difficult and inaccurate. Furthermore, the collected speaker utterances are frequently unavailable in a proper form

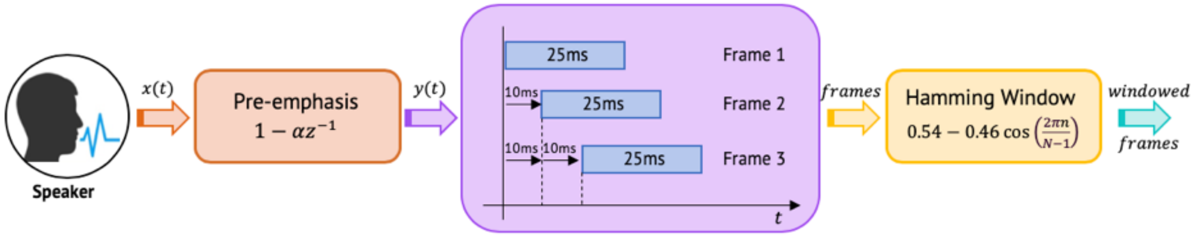


FIGURE 2. Framing of speaker utterance.

to learn the classification rules. Thus, to make these utterances useful for the speaker identification task, various useful features are extracted from collected utterances, and the extracted features are appropriate for learning classification rules. In general, most of the effort in speaker identification is required in the feature engineering step. It is an interesting step in the speaker identification process, where perception, innovation, intuition, creativity, and "black art" are equally important as technical and subject knowledge. The construction of a classification model is frequently the fastest step in the speaker identification task because feature engineering is responsible for extracting discriminative features from speaker utterances and transforming these features into a numeric master feature vector. This vector is then used by a machine learning or deep learning classifier to quickly learn the classification rules and develop a classification model. Feature extraction is more challenging than feature classification because of its domain-specific nature compared with the general-purpose nature of the classification task. Thus, in the present study, an innovative feature extraction process was adopted to extract useful and effective features, known as MFCCT features, from speaker utterances to construct an accurate classification model for speaker identification. The detailed functionality of the proposed MFCCT features is discussed in the subsequent subsection.

1) PROPOSED MFCCT FEATURES

This section discusses the functionality of MFCCT features which comprises of three distinct steps: (1) MFCC feature extraction, (2) time-domain feature extraction from MFCC features, and (3) appending target SIDs using the extracted features of each speaker utterance. These steps are discussed in the subsequent paragraphs.

a: EXTRACTING MFCC FEATURES

MFCC features were initially extracted from speaker utterances using Algorithm 1. MFCC-based features have been proven useful in speaker identification tasks [31]. These features represent the vocal tract information of a speaker. The MFCC feature extraction process comprises framing, windowing, discrete Fourier transform (DFT), logarithm of magnitude, warping frequencies on Mel scale and applying discrete cosine transform (DCT). Each speaker utterance was divided into a frame length of 25 ms. Moreover, 10 ms overlapping was used in successive frames to avoid information loss, as shown in Figure 2. Thus, the total number of frames for each speaker can be determined using Eq 3. In addition, the total number of samples per frame (N) can be computed using Eq 4. In the dataset used in this study, speaker utterances were recorded at a sample rate of 16 kHz and a frame step of 10 ms was used.

$$\mathit{Total\ frames} = \frac{\mathit{Number\ of\ Samples}}{\mathit{Frame}_{step} \times \mathit{Sample\ Rate}} \tag{3}$$

$$N = \mathit{Frame\ length} \times \mathit{Sample\ Rate} \tag{4}$$

Algorithm 1 MFCC Features of Speaker Utterance

```
Input : path to speaker utterances
Procedure: GetMFCCFeatures (path)
N ←total number of utterances
Sum ←0
Count ← 1
J ← 1
while J <=N do
    A ← MFCC matrix
    A ← matrixToColumn(A)
    M(:, Count) ←A
    Sum ←Sum + 1
    Sount ←Count + 1
end
X ← ceil (0.80 * Sum)
Y ←Sum - X
trainingMatrix ← M(:,1:X)
testingMatrix ← M(:,X+1:end)
trainingSamples ←X
testingSamples ←Y
end
```

After the framing steps, hamming windowing was performed on each individual frame to smooth the edge of each individual frame using Eq 5.

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}), \quad 0 \leq n \leq N \tag{5}$$

N is the number of samples in each frame.

Thereafter, the magnitude spectrum of each frame of N samples was computed by using DFT in the third step. Each magnitude spectrum was passed through a series of Mel-filter bank. Mel is a measuring unit based on the perceived frequency of human ears. The estimation of Mel can be written as

$$Mel(f) = 2595 \times \log_{10}(1 + \frac{f}{700}) \tag{6}$$

where *f* represents the physical frequency and *Mel(f)* represents the perceived frequency.

In order to imitate the perception of human ears, the warped axis was implemented using Eq 6. The most widely employed triangular filter bank with the Mel-frequency warping is the Mel-filter bank. Afterwards, the Mel-spectrum was calculated by multiplying the each of the triangular filters magnitude spectrum X(k) using Eq 7.

$$s(m) = \sum_{k=0}^{N-1} [|X(k)|^2 \times H_m(k)]; \quad 0 \leq m \leq M-1 \tag{7}$$

where M is the number of triangular filters. $H_m(k)$ is the weight assign to the kth bin of energy spectrum that contribute to the mth output band and is written as:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) < k < f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) < k < f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{8}$$

with m varies from *0* to *M-1*.

Finally, MFCC features were computed by taking DCT of each log Mel spectrum by using Eq 9.

$$c(n) = \sum_{m=0}^{M-1} log_{10}(s(m)) \cos(\frac{\pi n(m-0.5)}{M}) \tag{9}$$

n=0,1,2,..., C-1 where *C* is the total number of MFCCs.

b: EXTRACTING MFCCT FEATURES

After extracting the MFCC features from speaker utterances, MFCCT features were extracted from the extracted MFCC features. The detailed functionality is depicted in Algorithm 2. MFCCT features were extracted using three distinct steps. First, binning was performed on the extracted MFCC features for every 1500 rows of each column. A binning size of 1500 was used because it achieved better accuracy (for details, refer to Figure 9). In the second step, 12 different time-domain features (Table 2) were extracted from each bin of extracted MFCC features. The 12 features were used because they obtained the highest classification accuracy (for details, refer to Table 4). As shown in Algorithm 2, the variable *matrix* represents the extracted MFCC features in matrix format, and the *size* variable represents bin size (1500 in this case). The GetFeatureVector is a method that returns to the final master feature vector (*MFV*) for classification. The variable *rows* represent the number of rows that contain the MFCC feature matrix values. The variable *cols* is used for speaker utterances that has been used in columns. The variable *bins* is used to contain the total number of bins, and *n* represents the number of MFFCT features (12 in this case). Thus, for each speaker, the total number of rows will be the number of bins (bins) multiplied by n (MFCCT features), and the columns will be the number of utterances for each speaker.

TABLE 2. List of time-domain features.

Label	Statistical time-domain features
MIN	Minimum value of each bin
MAX	Maximum value of each bin
M$_n$	Mean value of each bin
M$_d$	Median of each bin
M$_o$	Mode of each bin
STD	Standard deviation of each bin
VAR	Variance of each bin
COV	Covariance of each bin
RMS	Root mean square of each bin
Q1	25th percentile of each bin
Q2	50th percentile of each bin
Q3	75th percentile of each bin

Algorithm 2 Master Features Vector

```
Input  : Matrix of MFCC Features
Input  : Binning Size
Output: Master Feature Vector. MFV
Procedure: GetFeatureVector ($matrix$, $size$)
rows ←number of rows of Matrix
cols ←number of columns of Matrix
bins ← (rows/size)
$i \leftarrow 1$
$n$ ←Number of Time Features
while $i$ <=cols do
    Initial ← 1
    $J \leftarrow 1$
    $K \leftarrow 1$
    while $J$ <=bins do
        $M$ ←matrix(Initial:J*size,i)
        $MFV(K, i)$ ←min(M)
        ....................
        $MFV(K + n - 1, i)$ ←std(M)
        $K$ ←K + n
        Initial ←Initial + size
    end
end
end
```

$$S_1 = \begin{bmatrix} V_{11} & V_{12} & \dots & V_{1n} \\ V_{21} & V_{22} & \dots & V_{2n} \\ \vdots & \ddots & & \vdots \\ V_{m1} & V_{m2} & \dots & V_{mn} \end{bmatrix} \tag{10}$$

$$S_2 = \begin{bmatrix} V_{11} & V_{12} & \dots & V_{1n} \\ V_{21} & V_{22} & \dots & V_{2n} \\ \vdots & \ddots & & \vdots \\ V_{m1} & V_{m2} & \dots & V_{mn} \end{bmatrix} \tag{11}$$

$$\vdots$$

$$S_{100} = \begin{bmatrix} V_{11} & V_{12} & \dots & V_{1n} \\ V_{21} & V_{22} & \dots & V_{2n} \\ \vdots & \ddots & & \vdots \\ V_{m1} & V_{m2} & \dots & V_{mn} \end{bmatrix} \tag{12}$$

The generalized form of above equations can be written as

$$S_t = [V_{ij}]_{m \times n} \tag{13}$$

where m = number of features

n = number of utterances for single speaker

i = 1,2,3,....,m

j = 1,2,3,.....n

t = 1,2,3,.....100

To the get the class label Eq 13 can be written as

$$S_t = [V_{ij}]_{m \times n}, V_{ij} = \begin{cases} 1, & t = i \\ 0, & t \neq i \end{cases} \tag{14}$$

Two feature vectors were prepared in the third step. In the first feature vector, each row represents one MFFCT feature, and columns represent speaker utterances (Eq 13). In the second feature vector, each row represents SID, and columns represent the number of each speaker utterances (Eq 14 and Algorithm 3). Finally, both feature vectors are fed as input to DNN [16] to construct a classification model for speaker identification.

Algorithm 3 Class Labels Vector

```
Input  : Vector V of N by 1
Output: Vector of Class Labels
Procedure: GetClassLabelsVector ($V$)
$N$ ←number of columns of V
$Total$ ←0
$J$ ← 1
while $J$ <=$N$ do
    $Total$ ←Total + V(J)
end
$Initial$ ← 1
$J$ ← 1
$classVector$ ← zeros(N,Total)
while $J$ <=$N$ do
    $Max$ ←V(J)+ Initial - 1
    $J$ ← $Initial$
    while $J$ <=$Max$ do
        $classVector(J, K)$ ← 1
    end
    $Initial$ ←Initial + V(J)
end
end
```

D. SPEAKER IDENTIFICATION MODEL

The hierarchical classification approach was used to identify the speaker. In this approach, the top-level classification layer identifies whether the speaker is male or female. Then, the second-level classification model is used to identify the specific SID. Thus, three classification models, namely, gender identification, male SID, and female SID models, were constructed. The detailed functionality is depicted in Figure 3. The two-level hierarchical classification approach was used because it obtained better results than the one-level classification model (Section IV-E). Furthermore, several recent studies in various domains have applied the hierarchical classification model and reported that it outperformed the one-level classification model [32]. In all three classification models, a feed forward deep neural network was used to construct the classification model. This classification algorithm was selected because it achieved promising results in several pattern recognition applications. Moreover, the performance of a feed forward deep neural network was compared with various traditional classification algorithms in Section IV-A. In the subsequent paragraph, a brief description of a deep neural network is presented.

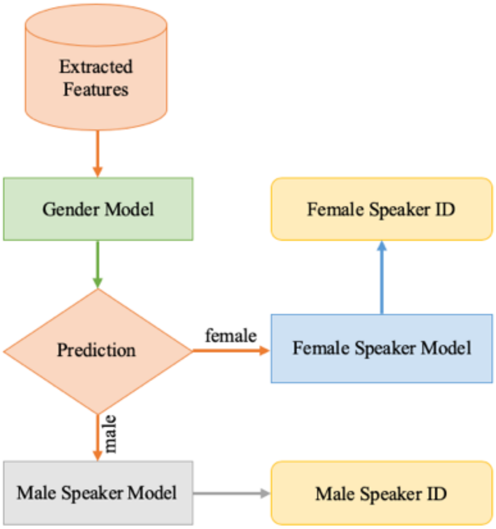


FIGURE 3. Proposed hierarchical model for speaker identification.

1) DEEP NEURAL NETWORK

In recent years, many ANNs have been proposed for speech recognition, speaker identification, image processing, sensor data processing and other application areas [17], [33]. The feedforward neural network (FFNN) as shown in Figure 5 has one input layer, one output layer, and one or more hidden layers. The input layer feeds the features to the hidden layer. The output layer computes the prediction of each class, and the results are applied to the input data through the series of functions of the hidden layers. Each layer consists of neuron-like information processing units, which are the basic building blocks of ANNs. Each neuron performs a simple weighted sum of the information it received and then applies the transfer function to normalize the weighted sum, as shown in Figure 4 [34]. Neural transfer functions are used to compute the output of a hidden layer from the input and to return the matrix of n elements. However, the softmax

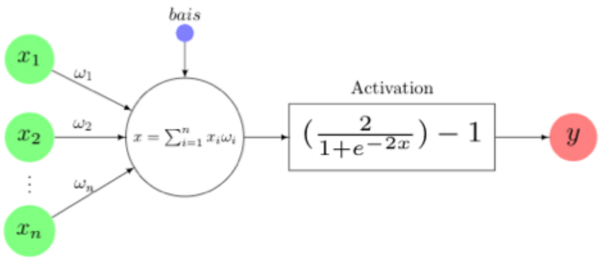

bais
x_1
x_2
x_n
ω_1
ω_2
ω_n
$x = \sum_{i=1}^{n} x_i \omega_i$
Activation
$\left(\frac{2}{1+e^{-2x}}\right) - 1$
y

FIGURE 4. Neuron with tansig transfer function.

neural transfer function is used differently in the output layer compared with that in the hidden layers to compute the predictions of each class. Figure 4 shows the weights w that are connected to each input x of a neuron and bias b. The two parameters are updated by a neural network during the training phase through the training function. Other details of ANN and various types of training and transfer functions can be found in [34].

In the current study, the customized FFNN was used as a classifier to identify the speaker. Several configurational changes were used in FFNN to identify the speakers and reduce the overall misclassification results [35]. The default FFNN architecture consists of one input layer, one hidden layer, and one output layer. The customized DNN architecture used in this study to classify speakers consist of 1 input layer, 5 hidden layers, and 1 output layer, as shown in Figure 5. Input layer used 48 neurons, which are equal to the number of features of each speaker utterance. Each hidden layer used 200 neurons, because the performance of neural networks depends on the number of neurons. A minimal number of neurons can contribute to underfitting, whereas a large number of neurons can lead to overfitting [34], [35]. Each hidden layer used the hyperbolic tangent-sigmoid (tansig) transfer function to compute output from the input within the range of -1 and 1.

However, the output layer used the softmax transfer function to compute the output values for multiclass classification (Table 3). Moreover, the trainscg function, which is the widely utilized function for pattern recognition-related problems, was used to train DNN [34]. Furthermore, to achieve a generalized performance of the training model and to avoid overfitting, different training functions, namely, trainscg, trainrp, traincgb, and traincgp, were used to train DNN [36]. Later MFCCT features were fed to the trained DNN to identify the speaker based on the unique patterns of the speaker's utterances.

E. EVALUATION METRICS

Overall accuracy and AUCROC were used to measure classification performance in all the experiments. These metrics are briefly discussed in the next paragraphs.

1) OVERALL ACCURACY

Overall accuracy is the ratio of the number of accurately predicted utterances to the total number of utterances for prediction. Eq 15 presents the mathematical definition of overall accuracy.

$$Accuracy = (\sum_{i=1}^{N} \frac{TP_i + TN_i}{TP_i + FN_i + TN_i + FP_i})/N \qquad (15)$$

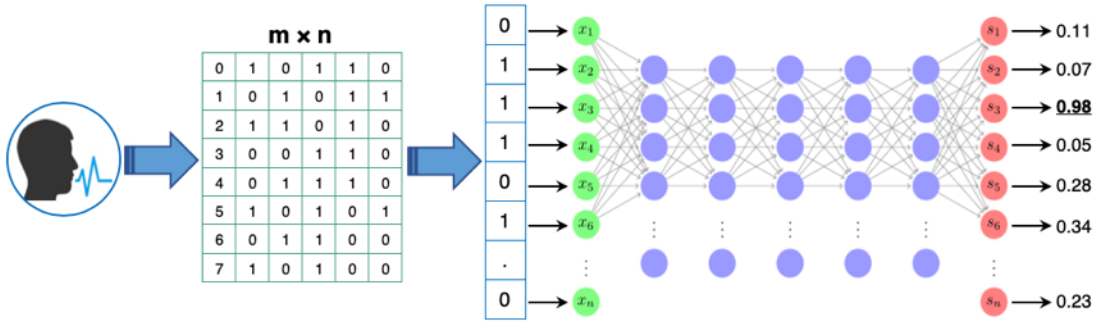
where N is the total number of instances.

2) AUROC

Area Under the Receiver Operating Characteristics (AUROC) is a useful measure and is extensively used in machine learning tasks that involve imbalanced datasets [39], [40]. Moreover, this measure analyzes the performance of a classifier with respect to each class and provides a good performance summary of ROC curves to compute the performance of a classifier by plotting a curve and computing the area under it. If the value of the area under the curve (AUC) is close to 1, then the performance of that classifier is good; by contrast, a value that is less than 0.5 indicates that the performance is poor [40], [41].

3) EQUAL ERROR RATE

Equal error rate (EER) is used to find the common value for its false acceptance rate (FAR) and its false rejection rate (FRR). The lower EER value indicates the higher accuracy of the system. FAR and FRR can be calculated using Eq 16


m x n

0	1	0	1	1	0
1	0	1	0	1	1
2	1	1	0	1	0
3	0	0	1	1	0
4	0	1	1	1	0
5	1	0	1	0	1
6	0	1	1	0	0
7	1	0	1	0	0

FIGURE 5. Deep neural network architecture.

TABLE 3. The detail on deep neural network.

Functions		Layers	Type	Neurons	Functions	
Training	Performance				Processing	Transfer
scaled conjugate gradient [37]	cross entropy [38]	1	input	48 (4 $bins$ $\times 12 features$)	mapminmax	
		2	hidden	200		tansig
		3	hidden	200		tansig
		4	hidden	200		tansig
		5	hidden	200		tansig
		6	hidden	200		tansig
		7	output	100 ($total$ $speakers$)		softMax

and 17 [42] while equal error rate (EER) can be calculated using Eq 18.

$$FAR = FPR = \frac{FP}{FP + TN} \tag{16}$$

$$FRR = FNR = \frac{FN}{FN + TP} \tag{17}$$

$$EER = \frac{FAR + FRR}{2} \tag{18}$$

F. EXPERIMENTAL SETUP

This section presents the experimental setup of the construction of the speaker identification model using the proposed MFCCT features and DNN algorithm. An extensive set of experiments was performed to measure the performance of the constructed model and to compare its performance with baseline speaker identification models. To evaluate the performance of the constructed speaker identification model through the proposed MFCCT features, experiments were performed systematically in four different settings, as follows:

1. Proposed MFCCT features and classification algorithms: In this setting, the proposed MFCCT features were extracted from human voices. The extracted MFCCT features were then fed to six different classification algorithms, namely, DNN, random forest (RF), k-nearest neighbor (k-NN), SVM, naïve Bayes (NB), and J48, to construct the speaker identification models. In this setting, six analyses (one feature engineering technique (MFCCT) × six classification algorithms) were run to evaluate the performance of the classification algorithms coupled with the proposed MFCCT features.

2. Performance comparison of the proposed MFCCT features with baseline features: In this setting, the performance of the proposed MFCCT features was compared with those of MFCC features and time-domain features.

3. Comparison of different binning sizes for MFCCT features: In this setting, the performance of several binning sizes (such as 500, 1000, 1500, 2000, 2500, and 3000) was evaluated to obtain the optimal learning curve for the DNN algorithm. In addition, these binning sizes were used because of their implementation feasibility, which allows the evaluation of the performance of the classification algorithms within a suitable operating range.

4. Selection of various time-domain features to compute effective MFCCT features: In this setting, the performance of several time-domain features (shown in Table 3) was evaluated to obtain the best set of time-domain features to compute the MFCCT features and determine the optimal learning curve for the DNN algorithm.

5. One-level versus two-level classification models: The hierarchical classification method was designed to improve the accuracy of speaker identification. To ascertain the efficacy of the hierarchical classification method, experiments were performed to compare the results of one-level classification with two-level classification. In one-level classification, all 40 speakers were labeled using their respective SID numbers. The proposed MFCCT features were used with the DNN algorithm to construct a classification model. 6. Evaluation of proposed method on different databases: In this setting, the performance of proposed MFCCT features coupled with DNN and other four classification algorithms were evaluated on three different speaker identification datasets to observe the effectiveness of proposed method. For this setting, 30 analyses (3 datasets 5 machine learning algorithms 2 classification models i.e. male and female) were performed. The EER evaluation metric was used to measure the effectiveness of all these 30 analyses. For all the experiments, speaker voice preprocessing, feature extraction, and classification were performed in MATLAB R2017a. The matplotlib Python library was used to generate accuracy graphs, AUC graphs, and utterance pattern.

IV. EXPERIMENTAL RESULTS

This section presents the results of all the experiments discussed in Section III-F. The results are presented based on four experimental settings. First, the results of the proposed MFCCT features and classification algorithms were obtained. Second, the results of the performance comparison of the proposed MFCCT features with baseline features were obtained. Third, the results of the comparison of different binning sizes for MFCCT features were obtained. Fourth, the results of the selection of various time-domain features to compute effective MFCCT features were obtained. All the results are reported in the subsequent subsections.

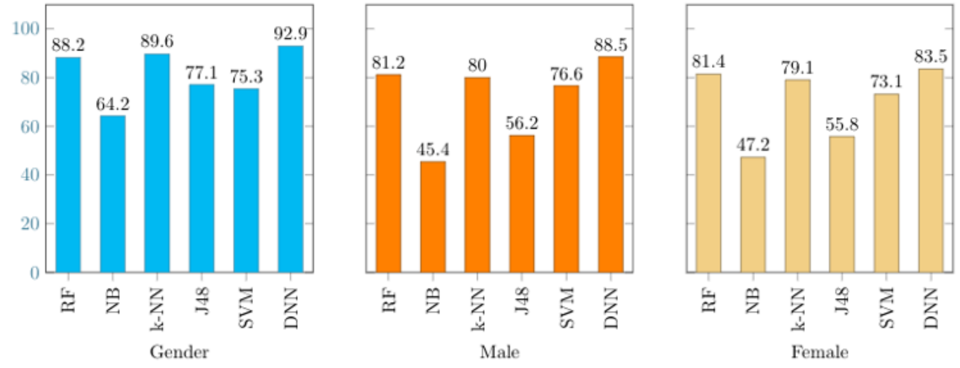


FIGURE 6. Overall accuracies of the first level and second level classification models.

A. RESULTS OF EXPERIMENTAL SETTING I

This section presents the results of Experimental setting I, in which the extracted MFCCT features were fed to five machine learning classification algorithms (i.e., RF, k-NN, NB, J48, and SVM) and DNN. The overall accuracies of all the aforementioned algorithms for the first level (gender-based classification model) and second level (male and female classification models) are shown in Figure 6. As shown in Figure 6, the DNN algorithm outperformed the other five machine learning-based algorithms by obtaining an overall accuracy of 92.9% for the gender identification classification model. In addition, this algorithm obtained 88.5% and 83.5% overall accuracy for the male and female speaker identification models, respectively. In the other five machine learning-based algorithms, an irregular trend can be observed in achieving overall accuracy. The k-NN and RF algorithms obtained the highest accuracy (89.6% and 88.2%) for the gender-based speaker identification model compared with the other three machine learning-based algorithms. In addition, in the male and female speaker identification models, the RF algorithm obtained the highest accuracy (81.2% and 80.4%, respectively) compared with the other four machine learning-based algorithms. In all the experiments, the NB algorithm obtained the lowest accuracy, followed by J48 and SVM.

In summary, the DNN algorithm outperformed the other five classification algorithms in obtaining good classification accuracy in the first-level and second-level classifications for speaker identification. In addition, the ROC diagrams [43] for all three classification models obtained through the highest-performing DNN algorithm are presented. Figure 8 (c) and (d) shows the ROC diagram and the confusion matrix for the first-level classification model. The performance of the male speaker class is marginally better than the performance of the female speaker class, because many analysis techniques, like pitch and formant are less accurate for high-pitched utterances (females) as compared to low-pitched utterances (males) [44]. Figure 8 (a) shows the ROC diagram for all 50 male speakers. The prediction accuracy of all male speakers is acceptable. Figure 8 (b) shows the ROC diagram for all 50 female speakers. The prediction accuracy of all female speakers is reasonable.

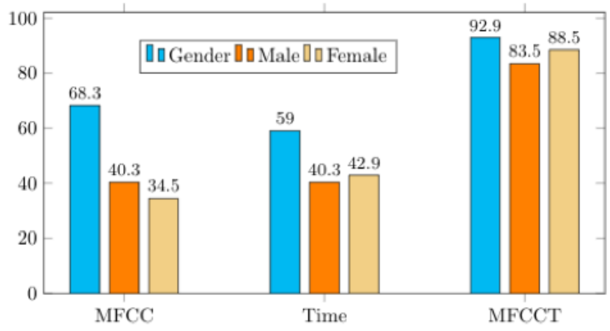


FIGURE 7. Performance comparison of different feature engineering techniques.

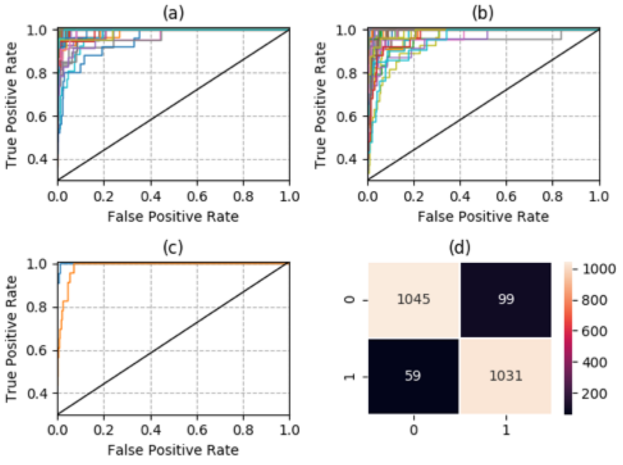


FIGURE 8. ROC Curves of Male (a). Female (b) and Gender (c) Classification Models along with confusion matrix of gender classification model (d).

B. RESULTS OF EXPERIMENTAL SETTING II

This section presents the results of Experimental setting II, in which the performance of the proposed MFCCT features were compared with other baseline features (i.e., MFCC and time-domain features) using the DNN algorithm. Thus, nine classification models [three different feature

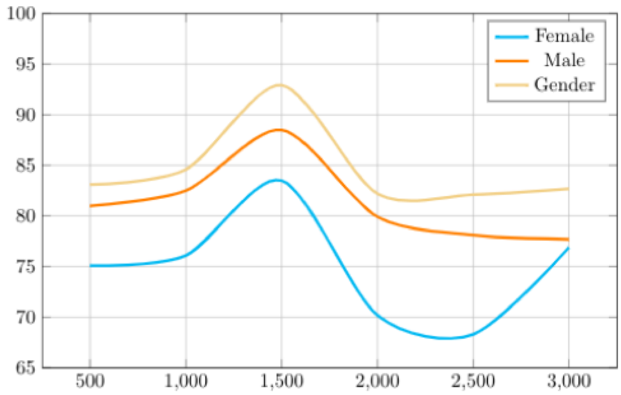


FIGURE 9. Accuracy comparison of bin sizes for mfcct features.

sets × (one gender-based model + one male model + one female model)] were constructed to evaluate the performances of the proposed MFCCT features. The results of all nine classification models are shown in Figure 7. The MFCCT features obtained better classification accuracy in all the experiments. Moreover, the lowest accuracy was observed when time-domain features were used. The proposed features attained approximately 50% more accuracy compared with existing baseline features.

C. RESULTS OF EXPERIMENTAL SETTING III

This section presents the results of Experimental setting III, in which the performances of different binning sizes for MFCCT features were compared. All three classification models were evaluated using different binning sizes (500, 1000, 1500, 2000, 2500, and 3000). The overall accuracies of these binning sizes are shown in Figure 9. The bin size of 1500 achieved the highest overall accuracy when MFCCT features were used. Overall accuracy gradually decreased when bin size reached more than 1500, and the lowest accuracy was observed at a bin size of 3000.

D. RESULTS OF EXPERIMENTAL SETTING IV

This section presents the results of Experimental setting IV. This setting evaluates the combination of various time-domain features (shown in Table 2) to compute effective MFCCT features. In this setting, the first two time-domain features of Table 2 were initially used to compute MFCCT features. The number of time-domain features was increased to 4, 6, 8, 10, and 12 to compute the MFCCT features. Thereafter, the resultant MFCCT features were computed from 2, 4, 6, 8, 10, and 12 time-domain features and then fed to the DNN algorithm to construct 18 different classification models (as shown in Table 4) to evaluate classification accuracy across all 18 models. Table 4 shows that an incremental trend is observed in classification accuracy. The highest accuracy of 92.9%, 88.5%, and 83.5% of the three models (i.e., gender of speaker model, male speaker model, and female speaker model) was observed when 12 different time-domain features were used to compute MFCCT features. In addition, the lowest classification accuracy was observed when MFCCT features were computed using 2 time-domain features.

TABLE 4. Time-domain features to compute effective MFCCT features.

Number of Features	Overall Accuracy		
	Gender	Male	Female
2	81.6	53.8	48.2
4	84.6	73.6	69.3
6	87.4	78.5	73.9
8	88.3	80.1	76.8
10	90.6	83.7	79.8
12	**92.9%**	**88.5%**	**83.5%**

TABLE 5. One-level and two-level accuracies.

SID	One Level	Two Level	SID	One Level	Two Level
026	26.1	75.9	019	90.9	90.9
027	51.9	88.9	032	78.3	87.1
060	78.9	70.0	039	54.2	84.8
078	82.6	82.6	040	86.4	90.9
118	70.4	78.3	083	91.7	100.0
163	100.0	95.5	087	90.5	95.2
196	57.1	85.7	089	83.3	91.7
201	84.0	84.0	103	75.0	85.0
229	83.3	88.9	125	25.9	57.3
233	81.8	95.5	150	68.2	68.2
254	91.7	91.7	198	68.0	81.3
307	91.7	91.7	200	87.0	100.0
311	100.0	100.0	211	78.8	81.8
332	82.4	94.1	226	76.2	81.0
374	95.5	95.5	248	77.3	77.3
405	56.5	73.9	250	77.3	86.4
412	100.0	100.0	289	76.5	87.3
445	66.7	75.0	298	81.8	81.8
446	81.8	90.9	302	54.5	54.5
458	87.5	100.0	322	95.8	95.8
460	76.0	84.0	328	94.7	94.7
481	80.0	92.0	403	72.7	90.9
625	95.5	95.5	426	91.3	91.3
831	91.7	95.8	441	87.0	91.3
839	85.0	90.0	587	52.2	72.3
909	86.4	90.9	669	47.8	83.9
911	77.3	77.3	696	81.0	85.7
1034	73.7	81.2	730	90.6	96.9
1040	75.0	87.5	887	73.9	78.3
1081	50.0	80.0	1069	81.0	88.6
1235	40.0	76.1	1088	77.3	86.4
1334	85.7	90.0	1098	83.3	94.4
1355	89.5	100.0	1116	88.0	88.0
1455	65.4	80.8	1183	84.6	92.3
1594	56.0	80.0	1246	30.4	76.2
1624	90.0	95.0	1263	90.5	90.5
1723	90.9	95.5	1363	94.7	100.0
1743	50.0	80.0	1447	52.0	64.6
1867	88.9	92.6	1502	81.5	96.3
2002	73.9	91.4	1553	35.0	65.0
2136	84.0	84.0	1578	69.6	69.6
2159	95.5	95.5	1737	60.0	88.0
2289	88.5	96.2	1841	34.8	82.6
2384	72.2	83.3	1898	36.0	77.6
2436	73.9	84.2	1926	42.3	80.8
2514	100.0	100.0	1963	32.0	61.3
2518	77.3	81.8	1970	58.3	91.7
2843	95.7	100.0	1992	90.9	90.9
2893	76.2	95.2	2007	79.2	79.2
2911	91.7	91.7	2092	24.8	47.3
	78.9%	**88.5%**		**70.7%**	**83.5%**

E. RESULTS OF EXPERIMENTAL SETTING V

In this setting, experiments were performed to ascertain the effectiveness of the hierarchical classification model. Table 5 shows the classification results obtained from the one- and

TABLE 6. Classification results (EER %) of different speaker identification databases.

Databases	Males					Females				
	SVM	RF	k-NN	J48	DNN	SVM	RF	k-NN	J48	DNN
LibriSpeech	0.12	0.09	0.10	0.23	**0.06**	0.14	0.09	0.11	0.23	**0.08**
VCTK	0.34	0.28	0.36	0.38	**0.17**	0.36	0.29	0.37	0.39	**0.18**
ELSDSR	0.34	0.24	0.35	0.38	**0.18**	0.26	0.26	0.29	0.38	**0.22**

two-level classification models. The hierarchical classification model achieved better results than the one-level classification model. Nonetheless, 22 speakers exhibited accuracies that were the same in the one-level and hierarchical classification models. Moreover, one speaker (SID 060) achieved high accuracy in the one-level classification model. However, accuracy was reduced to 9% in the two-level classification model. To summarize, the two-level classification yielded better results than the one-level classification in most of the cases.

F. RESULTS OF EXPERIMENTAL SETTING VI

This section shows the experimental findings of 24 analyses which were performed to evaluate the effectiveness of proposed MFCCT features across three different randomly selected datasets. The detailed results are shown in Table 6. As can be seen here, across all three datasets, our proposed MFCCT features coupled with DNN shown the best classification performance in both male and female classification models. The performance of proposed approach showed lowest EER using LibriSpeech dataset followed by VCTK dataset across both male and female classification models. The highest EER was observed in ELSDSR dataset across both male and female classification models. The lowest EER was observed in male classification model compared to female classification model across all three datasets. The DNN classifier outperformed SVM, RF, k-NN, and J48 classifiers in both male and female classification models across all the datasets. It can be inferred from these 30 analyses that the proposed MFCCT features coupled with DNN is robust and it has merit to perform better across several speaker identification datasets.

G. COMPARISON OF PROPOSED MODEL WITH BASELINES

To show the effectiveness of our proposed MFCCT features coupled with DNN, we compare the performance of proposed method with three baseline methods. The results of these experiments are shown in Table 7. As shown here, the proposed MFCCT features coupled with deep neural network outperformed aforementioned three baseline methods [45], [46]. However, the performance of our proposed method is a hair less than that of baseline [47]. The possible reason behind this marginal performance difference may be because, in [47] the authors have classified only 10 speakers with 8 utterances. Conversely, our proposed model was developed using 100 speakers (50 male and 50 female) and achieved 89% accuracy which is a hair less than that of [47]. Therefore, the model proposed in [47] possibly shows less performance as the number of classes or speakers increase. To confirm this, we performed an experimental evaluation where we initially employed our proposed model on 10 speakers and we then gradually increase the number of speakers by 10 till 50. The comparative analysis of these experiments can be seen in Figure 11. As can be seen here, as we increase the number of speakers, the accuracy of classification model is decreasing. Nevertheless, the accuracy of proposed model is much better on 10 speakers from LibriSpeech dataset (Figure 10) compared to the model proposed in [47]. Hence, it can be concluded that our proposed model is much more accurate and generic than the model proposed in [47].

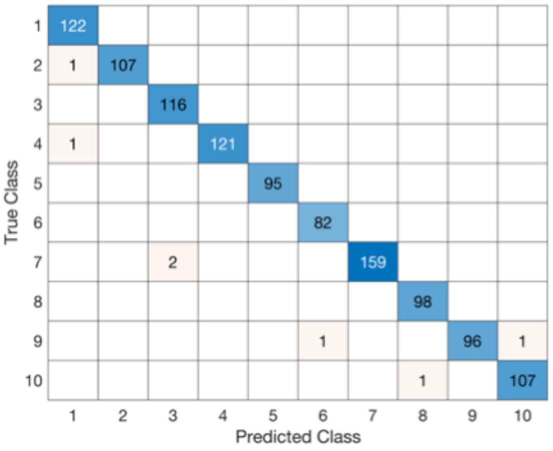


FIGURE 10. Confusion matrix of proposed model (10 Speakers).

TABLE 7. Comparison of baselines methods and proposed method.

Studies	Speakers	Database	Classifier	Acc.%	ERR%
Baseline 1 [45]	100	LibriSpeech	MMSE s-vector	-	0.158
Baseline 2 [47]	10	LibriSpeech	DNN	90	-
Baseline 3 [46]	40	LibriSpeech	GMM-UBM	86	-
Proposed Method	100	LibriSpeech	DNN	89	0.11

V. DISCUSSION

This sectionprovides the theoretical analysis of the speaker identification techniques used in this study. The experimental

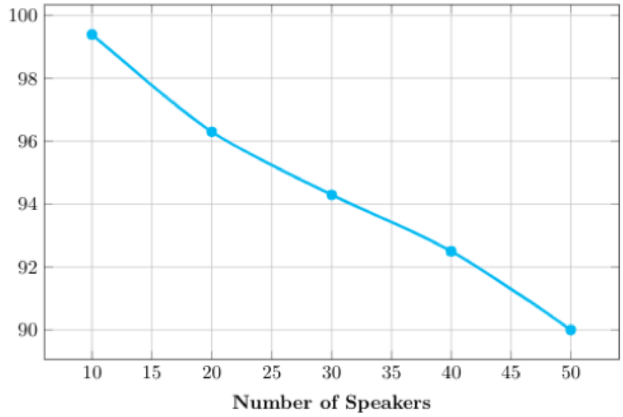


FIGURE 11. Performance comparison on different number of speakers.

results of this study show that the proposed MFCCT features and DNN can classify speaker utterances with an overall accuracy between 83.5% and 92.9%. As indicated in the experimental results (Section IV-B), the proposed MFCCT features presented the highest accuracy and outperformed MFCC and time-domain features. The possible reasons for the poor performance of MFCC features may be their production of short-time Fourier transform, which has an extremely weak time-frequency resolution, and their inherent pre-assumption that a signal is stationary [48]. Meanwhile, the possible reason for the poor performance of time-domain features is their inability to produce representative and discriminative visual patterns for different speaker utterances. To support this assertion, the visual patterns of three different speaker utterances are presented in Figure 12. Each column in Figure 12 shows the patterns generated through the proposed MFCCT, MFCC and time-domain features respectively. The utterance patterns of visual speakers generated by MFCCT features are discriminative across all three different speakers. The utterance patterns of visual speakers generated by MFCC and time-domain features are not sufficiently discriminative across all three different speakers. Thus, the classifier can effectively classify the patterns generated by MFCCT features and produce less classification error. By contrast, the classifier may encounter difficulties in classifying the patterns generated by MFCC and time-domain features, and thus, can cause high misclassification rates. Therefore, MFCCT features are recommended to accurately identify speaker utterances with a minimal misclassification rate instead of MFCC and time-domain features. The proposed MFCCT features consume less computational time during speaker identification model training and classification because these features are extracted from MFCC features by computing various descriptive statistical functions using a specific binning size. Thus, an enormous number of MFCC features can be transformed into few powerful and discriminative MFCCT features. To transform MFCC features into MFCCT features, 12 different descriptive statistical functions shown in Table 3.4 were used for each binning size. In our experiments, the combination of different descriptive statistical functions was evaluated to effectively transform them into MFCC features. Our experimental results showed high correlation between descriptive statistical and classification accuracy improvements. Moreover, classification performances were evaluated using various binning sizes to transform MFCC features into MFCCT features and to obtain the optimal learning curve for the classifier. The obtained results showed that the learning curve increased in performance with an increasing order of binning size from 1 to 1500. The learning curve decreased in performance when binning size crossed 1500. Thus, the binning size of 1500 should be used when transforming MFCC features into MFCCT features.

The findings of Experimental setting I is discussed in Section IV-A. The DNN classifier outperformed five other machine learning classifiers. The DNN classifier is well-suited for identifying complicated and nonlinear patterns from high-dimensional datasets [49], thereby providing better discriminative power for speaker identification. Moreover, the visual patterns learned by DNN were similar in intra-speaker utterances and discriminative in

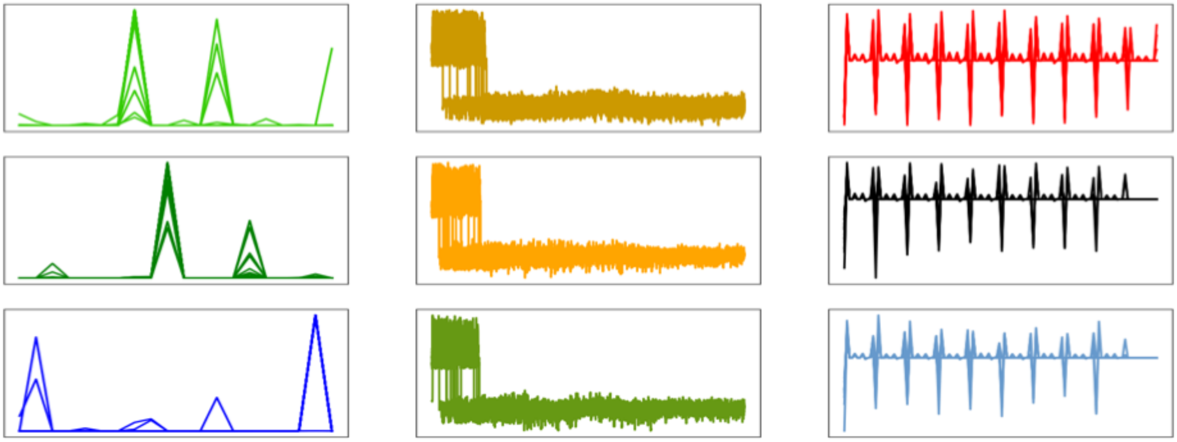


FIGURE 12. Visual patterns of different feature engineering techniques.

inter-speaker utterances. Thus, classification accuracy was better with the DNN classifier compared with the other classifiers. However, the experimental results showed that female speakers model yielded the low accuracy compared to male speakers' model. This is because in dataset there is one female (SID 2092) where the misclassification rate is higher. The possible reason behind this misclassification might be because the frequency of voices of this speaker is very similar to other female speakers and hence are challenging for classifier to classify accurately. To confirm this, we have performed experiments by replacing that female speaker (SID 2092) with female speaker (SID 2182) from the dataset. Our experimental results showed the improved accuracy (90% accuracy) on female model. In our future work, we will investigate the features that best classify the voices of speakers like SID 2092 will minimum misclassification rate.

The classification performance of RF was marginally lower than that of DNN. The fully grown trees in RF were not pruned [50], and the random split selection of features [51] led to computing better classification results. However, a large number of trees in RF can make the classification slow in real-time applications [52]. The classification performance of k-NN was marginally lower than that of RF. k-NN is highly effective when the amount of training data is large [53]. However, its computation cost is high due to the distance calculations for each cluster [53]. In many cases, J48 and SVM exhibit good performance; however, they demonstrated poor performance on the LibriSpeech dataset because the slight difference in training speaker utterances and single uncharacteristic features [54] can lead J48 to exhibit poor classification performance [55]. The default settings of several key parameters may cause the SVM classifier to present low classification performance [56]. The lowest overall accuracy was observed in the NB classification algorithm. The NB classifier assumes conditional independence among features that are probably invalid for the current dataset [57] and may result in poor performance. This conditional dependence on features becomes more complicated as the number of features increases, thereby negatively affecting the performance of the NB classifier. The results in Experimental Setting V are discussed in Section IV-E. The hierarchical classification approach outperformed the traditional one-level classification approach possibly because it split speaker training utterances to build two sub-classifiers [58] for effective prediction. Moreover, hierarchical classification takes less computational time than one-level classification [59]. The possible reason for the low performance of traditional one-level classification is because it considers 40 classes at a time; hence, differentiating among the utterances of 100 different speakers with different genders may be challenging for any classifier [60].

VI. CONCLUSION

In this study, effective MFCCT features were proposed for speaker identification through a hierarchical classification approach. The hierarchical classification approach was implemented in cascading style, where the first-level classification layer identifies the speaker gender and the second-level identifies the specific speaker identity. Moreover, five machine learning algorithms and one deep learning-based DNN were used to classify speaker gender and SID. The rigorous experimental results showed that the performance of the proposed MFCCT features in terms of overall accuracy was approximately 83.5%-93%. Moreover, DNN was found to be suitable for speaker identification through the proposed MFFCT features. The experimental results prove that the proposed speaker identification system is efficient, accurate, and robust in terms of number of speakers, testing utterances, and utterance length compared with other baseline speaker identification models. The promising results show that the proposed speaker identification system can be used in many application areas, including access control and security. In the future, we intend to improve classification accuracy by reducing the classification errors between speakers with similar voice patterns using deep learning with deeper architectures. Moreover, deep learning hyper-parameter tuning can be implemented to enhance the speaker recognition model. In addition, we are currently collecting a large corpus of speaker identification dataset to further improve the proposed model.

REFERENCES

[1] G. S. Morrison, F. H. Sahito, G. Jardine, D. Djokic, S. Clavet, S. Berghs, and C. G. Dorny, "INTERPOL survey of the use of speaker identification by law enforcement agencies," *Forensic Sci. Int.*, vol. 263, pp. 92–100, Jun. 2016.

[2] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 95–103, Mar. 2009.

[3] A. K. Hunt and T. B. Schalk, "Simultaneous voice recognition and verification to allow access to telephone network services," Google Patents 5 499 288, Mar. 12, 1996.

[4] J. Naik and G. Doddington, "Evaluation of a high performance speaker verification system for access control," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2005, pp. 2392–2395.

[5] M. G. Gomar, "System and method for speaker recognition on mobile devices," Google Patents 9 042 867, Mar. 26, 2015.

[6] M. Faundez-Zanuy, M. Hagmüller, and G. Kubin, "Speaker identification security improvement by means of speech watermarking," *Pattern Recognit.*, vol. 40, no. 11, pp. 3027–3034, Nov. 2007.

[7] B. M. Arons, *Interactively Skimming Recorded Speech*. Cambridge, MA, USA: Massachusetts Institute of Technology, 1994.

[8] C. Schmandt and B. Arons, "A conversational telephone messaging system," *IEEE Trans. Consum. Electron.*, vols. CE–30, no. 3, pp. 21–24, Aug. 1984.

[9] A. Maurya, D. Kumar, and R. K. Agarwal, "Speaker recognition for hindi speech signal using MFCC-GMM approach," *Procedia Comput. Sci.*, vol. 125, pp. 880–887, Jan. 2018.

[10] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture Models," *Digit. Signal Process.*, vol. 10, nos. 1–3, pp. 19–41, Jan. 2000.

[11] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, nos. 1–2, pp. 91–108, Aug. 1995.

[12] W.-C. Chen, C.-T. Hsieh, and C.-H. Hsu, "Robust speaker identification system based on two-stage vector quantization," *J. Sci. Eng.*, vol. 11, no. 4, pp. 357–366, 2008.

[13] D. B. A. Mezghani, S. Z. Boujelbene, and N. Ellouze, "Evaluation of SVM kernels and conventional machine learning algorithms for speaker identification," *Int. J. Hybrid Inf. Technol.*, vol. 3, pp. 23–34, Jul. 2010.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[15] M. A. Islam, W. A. Jassim, N. S. Cheok, and M. S. A. Zilany, "A robust speaker identification system using the responses from a model of the auditory periphery," *PLoS ONE*, vol. 11, no. 7, Jul. 2016, Art. no. e0158520.

[16] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*. Cambridge, MA, USA: MIT, 2016.

[17] H. F. Nweke, Y. W. Teh, M. A. Al-garadi, and U. R. Alo, "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges," *Expert Syst. Appl.*, vol. 105, pp. 233–261, Sep. 2018.

[18] R. Karthik, D. Tyagi, A. Raut, S. Saxena, and R. Kumar M, "Implementation of neural network and feature extraction to classify ECG signals," 2018, *arXiv:1802.06288*. [Online]. Available: http://arxiv.org/abs/1802.06288

[19] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[20] Z. Wu and Z. Cao, "Improved MFCC-based feature for robust speaker identification," *Tinshhua Sci. Technol.*, vol. 10, no. 2, pp. 158–161, Apr. 2005.

[21] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[22] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, no. 1, pp. 52–55, Jan. 2006.

[23] S. Fong, K. Lan, and R. Wong, "Classifying human voices by using hybrid SFX time-series preprocessing and ensemble feature selection," *BioMed Res. Int.*, vol. 2013, Oct. 2013, Art. no. 720834.

[24] H. Ali, S. N. Tran, E. Benetos, and A. S. D. A. Garcez, "Speaker recognition with hybrid features from a deep belief network," *Neural Comput. Appl.*, vol. 29, pp. 13–19, Mar. 2018.

[25] M. Soleymanpour and H. Marvi, "Text-independent speaker identification based on selection of the most similar feature vectors," *Int. J. Speech Technol.*, vol. 20, no. 1, pp. 99–108, Mar. 2017.

[26] R. Laptik and T. Sledevič, "Fast binary features for speaker recognition in embedded systems," in *Proc. Open Conf. Elect., Electron. Inf. Sci.*, New York, NY, USA, Apr. 2017, pp. 1–4.

[27] S. Prasad, Z.-H. Tan, and R. Prasad, "Frame selection for robust speaker identification: A hybrid approach," *Wireless Pers. Commun.*, vol. 97, no. 1, pp. 933–950, Nov. 2017.

[28] S. S. Nidhyananthan, R. S. S. Kumari, and T. S. Selvi, "Noise robust speaker identification using RASTA–MFCC feature with quadrilateral filter bank structure," *Wireless Pers. Commun.*, vol. 91, no. 3, pp. 1321–1333, Dec. 2016.

[29] X. Zhao and D. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7204–7208.

[30] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, Oct. 2012.

[31] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Syst. Appl.*, vol. 90, pp. 250–271, Dec. 2017.

[32] G. Mujtaba, L. Shuib, R. G. Raj, M. A. Al-Garadi, R. Rajandram, and K. Shaikh, "Hierarchical text classification of autopsy reports to determine MoD and CoD through term-based and concepts-based features," in *Proc. Ind. Conf. Data Mining*. Cham, Switzerland: Springer, 2017, pp. 209–222.

[33] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[34] M. H. Beale, M. T. Hagan, and H. B. Demuth, *Neural Network Toolbox User's Guide*. MathWorks, Natick, MA, USA, 1992.

[35] J. M. Alvarez and M. Salzmann, "Learning the number of neurons in deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2270–2278.

[36] *Improve Shallow Neural Network Generalization and Avoid Overfitting—MATLAB & Simulink*. Accessed: 2013. [Online]. Available: https://www.mathworks.com/help/deeplearning/ug/improve-neural-network-generalization-and-avoid-overfitting.html

[37] M. F. Müller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Netw.*, vol. 6, no. 4, pp. 525–533, Jan. 1993.

[38] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013, p. 3.

[39] F. J. Provost and T. Fawcett, "Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions," in *Proc. KDD*, 1997, pp. 43–48.

[40] F. J. Provost, T. Fawcett, and R. Kohavi, "The case against accuracy estimation for comparing induction algorithms," in *Proc. ICML*, 1998, pp. 445–453.

[41] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[42] S. Marcel, M. Nixon, and S. Li, *Handbook of Biometric Anti-Spoofing-Trusted Biometrics Under Spoofing Attacks* (Advances in Computer Vision and Pattern Recognition). London, U.K.: Springer-Verlag, 2014.

[43] Scikit. *Feature Transformations With Ensembles of Trees*. [Online]. Available: https://scikit-learn.org/stable/auto_examples/ensemble/plot_feature_transformation.html

[44] R. Saeidi, P. Alku, and T. Backstrom, "Feature extraction using power-law adjusted linear prediction with application to speaker recognition under severe vocal effort mismatch," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 1, pp. 42–53, Jan. 2016.

[45] S. Ramoji and S. Ganapathy, "Supervised I-vector modeling for language and accent recognition," *Comput. Speech Lang.*, vol. 60, Mar. 2020, Art. no. 101030.

[46] S. Chakraborty and R. Parekh, "An improved approach to open set text-independent speaker identification (OSTI-SI)," in *Proc. 3rd Int. Conf. Res. Comput. Intell. Commun. Netw. (ICRCICN)*, Nov. 2017, pp. 51–56.

[47] T. Lin and Y. Zhang, "Speaker recognition based on long-term acoustic features with analysis sparse representation," *IEEE Access*, vol. 7, pp. 87439–87447, 2019.

[48] K. Daqrouq, A. Alkhateeb, M. Ajour, and A. Morfeq, "Network for noisy speaker identification system," *J. Appl. Sci.*, vol. 14, no. 19, pp. 2341–2349, 2014.

[49] F. Guenther, "Neural networks: Biological models and applications," in *International Encyclopedia of the Social & Behavioral Sciences*. 2001, pp. 10534–10537.

[50] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, Jan. 2005.

[51] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Comput.*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998.

[52] P. Latinne, O. Debeir, and C. Decaestecker, "Limiting the number of trees in random forests," *Int. Workshop Multiple Classifier Syst.* Berlin, Germany: Springer, 2001, pp. 178–187.

[53] S. B. Imandoust and M. Bolandraftar, "Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background," *Int. J. Eng. Res. Appl.*, vol. 3, pp. 605–610, Sep. 2013.

[54] P. Kapoor, R. Rani, and R. JMIT, "Efficient decision tree algorithm using j48 and reduced error pruning," *Int. J. Eng. Res. General Sci.*, vol. 3, no. 3, pp. 1613–1621, 2015.

[55] M. Quwaider and M. Alfaqeeh, "Social networks benchmark dataset for diseases classification," in *Proc. IEEE 4th Int. Conf. Future Internet Things Cloud Workshops (FiCloudW)*, Aug. 2016, pp. 234–239.

[56] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, 1998, pp. 137–142.

[57] D. D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, 1998, pp. 4–15.

[58] C. N. Silla and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Min Knowl Disc.*, vol. 22, nos. 1–2, pp. 31–72, Jan. 2011.

[59] S. Dumais and H. Chen, "Hierarchical classification of Web content," in *Proc. 23rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2000, pp. 256–263.

[60] R. Cerri, A. C. P. L. F. De Carvalho, and A. A. Freitas, "Adapting non-hierarchical multilabel classification methods for hierarchical multilabel classification," *Intell. Data Anal.*, vol. 15, no. 6, pp. 861–887, Nov. 2011.

RASHID JAHANGIR received the bachelor's degree in computer engineering from the University of Engineering and Technology Lahore, Pakistan, and the master's degree from the University of New South Wales (UNSW), Sydney, Australia. He is currently pursuing the Ph.D. degree with the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. He worked as a Software Engineer in software house in Lahore for two years. He has been a Lecturer with COMSATS University Islamabad, Vehari Campus, Pakistan, since 2014. He is working on digital signal processing and deep learning. The key research areas of his interest are deep learning, pattern recognition, machine learning, and data mining.

YING WAH TEH received the B.Sc. and M.Sc. degrees from Oklahoma City University, Oklahoma, USA, and the Ph.D. degree from the University of Malaya, Malaysia. He is currently a Professor with the Information Science Department, Faculty of Computer Science and Information Technology, University of Malaya. He has published several articles in academic journals indexed in well-reputed databases, such as ISI-indexed and Scopus-indexed. His research interests include data mining and text mining.

NISAR AHMED MEMON received the B.E. degree in computer systems engineering from Mehran University, Jamshoro, Pakistan, in 1989, the M.S. degree in computer engineering from the Ghulam Ishaq Khan Institute (GIKI), Pakistan, in 2005, and the Ph.D. degree in computer engineering from the GIK Institute, Pakistan, in 2010. He is currently working with King Faisal University, Saudi Arabia, as an Assistant Professor with the College of Computer Sciences and Information Technology. His current research interests include digital image processing, digital image watermarking, pattern recognition, data authentication, and cryptography. He has published more than 40 articles in international journals and conferences in the field of image processing, data authentication, medical image watermarking, data security, and biometrics.

GHULAM MUJTABA received the master's degree in computer science from FAST National University, Karachi, Pakistan, and the Ph.D. degree from the Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. He has received the gold medal for the master's degree. He has been an Associate Professor with Sukkur IBA University, Sukkur, Pakistan, since 2006. Prior to join Sukkur IBA University, he was with a well-known software house in Karachi for four years. He has vast experience in teaching and research. He has also published several articles in academic journals indexed in well-reputed databases, such as ISI and Scopus. His research interests include machine learning, online social networking, text mining, deep learning, and information retrieval.

MAHDI ZAREEI (Member, IEEE) received the M.Sc. degree in computer networks from the University of Science, Malaysia, in 2011, and the Ph.D. degree from the Communication Systems and Networks Research Group, Malaysia-Japan International Institute of Technology, University of Technology, Malaysia, in 2016. In 2017, he joined the School of Engineering and Sciences, Tecnologico de Monterrey, as a Postdoctoral Fellow, where he is currently a Research Professor. His research interests include wireless sensor and ad hoc networks, energy harvesting, cognitive radio networks, and performance optimization. He is a member of the Mexican National Researchers System (level I). He is also serving as an Associate Editor for IEEE Access.

UZAIR ISHTIAQ received the bachelor's degree in information technology from Bahauddin Zakariya University, Multan, Pakistan, and the master's degree in computer science from National University, FAST, Islamabad, Pakistan. He is currently pursuing the Ph.D. degree with the Faculty of Computer Science and Information Technology, University of Malaya, Malaysia. He has been a Lecturer with COMSATS University Islamabad, Vehari Campus, Pakistan, since 2014. His research interests include image processing, medical image analysis, and deep learning. He received a Gold Medal for the bachelor's degree.

MUHAMMAD ZAHEER AKHTAR received the bachelor's degree in computer science from Allama Iqbal Open University, Islamabad, Pakistan, and the master's degree from the University of Agriculture, Faisalabad, Pakistan. He is currently pursuing the Ph.D. degree with the Department of Computer Science, Allama Iqbal Open University. He has been a Lecturer with COMSATS University Islamabad, Vehari Campus, Pakistan, since 2014. His research interests include pattern recognition, time-series data analysis, data mining, and machine and deep learning.

IHSAN ALI received the M.S. degree in computer system engineering from the GIK Institute, in 2008. He is currently pursuing the Ph.D. degree with the Faculty of Computer Science and Information Technology, University of Malaya.

He is currently an Active Research Associate with the Centre for Mobile Cloud Computing Research (C4MCCR), Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia. He has published more than 40 high impact research journal articles including a highly reputable the *IEEE Communication Magazine*. He has been actively involved in research & teaching activities for the last ten years in different country including Saudi Arabia, USA, Pakistan, and Malaysia. His research interests include wireless sensor networks, robotics in WSNs, sensor cloud, fog computing, the IoT, and ML/DL in wireless sensor networks.

Mr. Ihsan has served as a Technical Program Committee Member for several well-known conferences including IWCMC 2017–2018, AINIS 2017, Future 5V 2017, ICACCI-2018, INAIT 2019, DiCES-N19, CCNC2020, ICCAIS2020, and CSNT2020, and also as an Organizer of the Special Session on Fog Computing in Future 5V 2017. He is also an Active Reviewer of *Computers & Electrical Engineering*, *KSII Transactions on Internet and Information Systems*, *Mobile Networks and Applications*, the *International Journal of Distributed Sensor Networks*, the *Journal of Advanced Transportation*, the IEEE Transactions on Intelligent Transportation Systems, *Computer Networks*, IEEE Access, *Wireless Communications and Mobile Computing*, and the *IEEE Communication Magazine*.