Automatic Gender Authentication from Arabic Speech Using Hybrid Learning

Amjad Rehman Khan

Artificial Intelligence and Data Analytics Lab (AIDA), College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia
E-mail: arkhan@psu.edu.sa

***Abstract*—Speech recognition is progressively being utilized in practical applications with time. Automatic gender identification is one of the most intriguing applications since it distinguishes female and male speeches from briefly spoken communication records. This is advantageous in various applications, including automated conversation systems, system verification, demographic attribute prediction and assessing speaker's expressions. Speech is a natural mode of communication, and pitch variation of a gender-specific speech signal is often used to identify a person as male or female. This paper presents a model to identify gender from Arabic speech by integrating audio preprocessing, Mel-Frequency Cepstral Coefficients (MFCC), Delta MFCC, and Log Filter bank feature extraction. Pre-processing involves testing pre-emphasis, framing, windowing, and Fast Fourier Transform. Finally, features are extracted using three feature extraction methods from the processed audios. Feed Forward Neural Networks and Keras-based Neural Networks are employed as classifier models. Regarding accuracy and simplicity, the proposed hybrid method surpasses most previous approaches discussed in the literature for gender categorization from Arabic speech. The proposed model achieved an average classification accuracy of 93.09%.**

***Keywords*—speech recognition, Arabic language, gender classification, hybrid learning, technological development**

I. Introduction

Automatic voice recognition is a field of study that enables models to take vocal input from people and interpret it with higher accuracy. Numerous methods exist for implementing voice recognition models. Gender recognition from Arabic speech was less explored in the literature, and some researchers recognized gender from the speeches with low accuracy. One of the developing methods for voice recognition is using neural networks with deep learning. Arabic is a widely spoken natural language that received less attention regarding voice recognition.

During the past decade, Deep Learning (DL) has emerged as an innovative and exciting field of Machine Learning (ML). It has subsequently been studied and applied to various research reported in the state of the art using deep and traditional learning. Typically, these are composed of non-direct processes of several layers. Machine learning techniques try to evaluate the extracted particular characteristics, features, and context from these Deep Neural Network (DNNs) that are mostly utilized to improve computers' skills further to comprehend what humans can accomplish, like voice recognition. As a result of the constant and fast growth of applications relating to human lives and well-being, automatic speaker age and gender categorization are thriving research areas [1, 2]. According to United Nations (UN), Arabic is the 6th authorized Semitic and globally 5th spoken language with estimated speakers of 422 million in the Arab world. According to region, Dialectal Arabic (DA) is often split into five major categories: Gulf, Iraqi, Egyptian, Levantine and Maghrebi [3, 4].

However, automatically recognizing a speaker's gender and identifying them based on their voice is a critical problem in audio signal processing. Gender recognition determines if a man or a female delivers a speech. This job is vital for speaker-specific Automated Speech Recognition (ASR) since it enables the ASR approach to be extra precise than speaker-autonomous techniques. Gender identification is the technique of spontaneously identifying and verifying genders based on the unique data included in the audio records [5, 6]. Gender categorization through voice is critical for speech recognition, which applies to various applications. It is often created via usual machine learning and deep learning techniques. Many businesses, including Amazon, Microsoft, Apple and Google, have created a voice recognition algorithm to achieve high accuracy and low error rates [7, 8]. It is not difficult to infer a human's gender from other people's audio recordings. By and large, individuals can readily determine the gender of the speaker in a discussion based on their learned experience. In general, gender identity may be accomplished via an audio and speaker recognition technique. The individual ear is a tremendous technique for determining sex through intensity and frequency features of audio and communication. Similarly, a computer may be trained to accomplish the same by selecting and integrating the appropriate characteristics from speech data into a machine-learning algorithm [9–11].

However, computer systems are not always accurate in determining whether a person is male or female. As a result, many articles and ideas have been submitted to address this

Manuscript received April 1, 2023; revised July 16, 2023, accepted August 14, 2023; published April 24, 2024.

issue via computer systems. Gender categorization is a challenging issue in speech recognition. While numerous research has been performed to enhance the accuracy of feature extraction and classifiers, classification accuracy has yet to reach the required level. The key problem in identifying the speaker's sex is creating strong characteristics and building an effective classification system. In addition, Arabic is a challenging semantic language compared to English and other languages [12, 13]. When many speakers with distinct dialects must be identified, significant problems arise. Because of the absence of standardization and norms, spoken Arabic varies from area to region. The Arabic used in everyday casual conversation differs from those used in books, magazines, newspapers, and television news broadcasts. The pronunciation of the Arabic alphabet in isolation is distinct from the pronunciation of the same letter in context. One of the primary problems Arabic ASR researchers' faces is a need for spoken and written training data. These issues may be mitigated by limiting the number of speakers and words and working in an acoustically sound environment. Additionally, various dialects may be overcome by ignoring the intricacies of fluent speaking and focusing on contemporary standard Arabic [14–16].

Various features have been created and researched for speech recognition in the related work. For example, MFCCs are a commonly utilized spectral feature set capable of modelling the vocal tract filter in a short-time power spectrum [8]. Since neural networks are simplified representations of organic nerve systems, they were inspired by the kind of computation done by the human brain. Neural Networks are high-performance information processing systems that combine parallelism, adaptability, and fault tolerance with real-time operation. These networks fall under the category of approximation computing methods. In the last decade, neural networks have grown in popularity and are promising to resolve the challenges presented by existing information processing systems [17–20]. Neural Network is extensively utilized in various areas and applications, including natural language, image processing, recognition and classification and computer vision. The fundamental strength of neural networks is derived from their deep architecture, which can transform raw input data into a rich and robust internal representation.

Arabic is one of the world's most broadly vocal dialects, with about 295 million native speakers. Additionally, it is the official language of the Arab world, which comprises 22 nations. It is exceptionally morphologically complex and ambiguous. Many colloquial dialects of Arabic vary by location and are utilized in everyday conversation. These dialects are distinct from Modern Standard Arabic (MSA), used in newspapers and other forms of formal communication. As a result, we selected Arabic as the language of this study. The reason for using Neural Network (NN) is to capitalize on its accomplishments in ASR. Using NNs effectively handles individual changes in the voice stream and improves the acoustic model's speaker invariance. This research aimed to explore the usage of neural networks for automated Arabic speech recognition by using various feature extraction methods to boost the performance of isolated word AASR. The primary objective of developing a voice recognition system is to create a robust, high-quality model.

The key contributions of this research can be described as follows.

- A hybrid gender classification model using Arabic speech is proposed.
- MFCCs, Delta MFCCs and Log Filter Bank features are extracted from pre-processed audio signals and fused to generate a feature vector.
- Proposed hybrid model exhibited significantly better performance than existing techniques.

Further, this research paper is arranged as follows: Section II presents the literature review based on various Arabic and other audio dialects databases. Section III extensively provided the proposed technique for gender classification. Finally, in Section IV, comprehensive experiments have been carried out, while Section V delivers conclusions and future work.

II. Related Work

Existing methods and models have been exhaustively documented in the published literature. Automatic speaker recognition models are constructed using characterization, pattern analysis, and engineering models. The article focuses on categorization, feature extraction, and selection techniques for voice recognition. The complexity reduction of the technique is critical in the precise feature selection in conjunction with the model. This procedure is critical for techniques that are deployed in real-time scenarios. The authors propose a novel speech recognition network based on deep learning for automated spoken word recognition. The supremacy of an input wave sample, in this case, voice sound, directly affects the process by which a classifier achieves accuracy. The Berlin database contains about 500 protests against male and female media personnel. The proposed model obtains a maximum accuracy of 94.21%, 83.54%, 83.65 %, and 78.13 % on the applied dataset using Linear Prediction Coefficients (LPC), LSP, prosodic and Mel Frequency Spectrum Coefficients (MFCC) features. The proposed network outperformed the previous approaches regarding recognition performance [21, 22].

An Arabic speaker identification system has been created to convert Arabic speech immediately. The speech samples were captured, the pre-processing activity was identified to distinguish vocal from unvoiced portions, and the Arabic Speech signals were segmented using frame and rectangle window sliding methods, followed by MFCC features extraction. The feature vectors for each spoken sample are aggregated using the Vector Quantization via Linde-Buzo-Gray (VQLBG) procedure and the Gaussian Mixer Model (GMM) to classify and recognize an unknown speaker based on his spoken phrases belonging to a distinct cluster from other Arabic speakers. This method was claimed to have a recognition rate of 95.5%. VQLBG has a reported accuracy of 75.6 %. Limitation: The words spoken by each Arabic speaker will

be classified as belonging to a distinct cluster distinct from other clusters associated with other Arabic speakers. This foundation method has been selected to be validated and acknowledged as being associated with the recognized speaker [23, 24].

A novel model for Arabic speech recognition was developed that automatically determines the speaker's gender. The spoken emphasized letter. The author's utterances information is assumed from these emphasized letters diacritical. The dataset utilized has 720 sounds. This group included 12 individuals: four ladies and eight males. During the testing, four letters with three diacritics were explored. Each participant made five recordings of each sound. To begin, we evaluated the efficacy of the Hidden Markov Models (HMM) technique, which, when utilized to input records from dataset obtained 63.42% gender identification. Then, a suggested hybrid technique based on K-Nearest Neighbors and Decision Trees (KNN-DT) is tested, which obtained 66.67% gender recognition in the case of DT and K-Nearest Neighbors (KNN), which achieved 52.78% gender recognition. This research aimed to continue investigating methods for directly mapping the raw audio wave shape to the appropriate languages.

Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNN) were also investigated to enhance the hybrid technique. Alternate area of study worth investigating is the impact of social data integration throughout the training phase. This may aid in discovering new correlations between recordings and identifying unique critical characteristics for speech recognition [25, 26]. A novel feature extraction and classification approach based on DNNs for speaker age and gender categorization has been presented. Each speaker is represented by a model created by the suggested approach. The comparison between the gender class models and test model is calculated for each test speech utterance. Shifting Delta Cepstral (SDC) and MFCCs coefficients were utilized as feature sets. The suggested model, which used the SDC feature set, outperformed the MFCCs in classification accuracy. The testing findings demonstrated that the suggested shifting delta cepstral gender model and shifting delta cepstral class model improved efficacy than other techniques by attaining a classification accuracy of 57.21% overall [27]. Bidirectional LSTM has been utilized to create speech recognition technique to distinguish gender categorization. To extract the features for training the Bidirectional LSTM, the MFCC is used.

A five-run evaluation utilizing a small database of thousand audio samples containing five hundred men and five hundred women demonstrates that the technique can correctly identify the speakers' gender. With an 80:20 train-test split, the method achieves minimum and maximum accuracy of 81.0% and 90.5%, respectively and overall recognition rate of 86.7%. Reduce the train-test split percentage and its efficacy will diminish. It remains constant when the splitting percentage is 50:50 [28]. A convolutional-recurrent neural network is used in the suggested method. A public dataset is Mozilla's "Common Voice" database, which was used to train and test the suggested algorithms.

A Voice Activity Detector, Short-Time Fourier Transform, and Mel scale are utilized for audio preprocessing. Gender is identified with an average mistake of less than 1.55%, while the likelihood of properly classifying speakers' ages into three age groups is more than 80 % on average. Considering the information in the datasets, these findings are very promising in justifying their application in the systems under consideration. The developed approach is independent of the input signal's length [29]. The topology of Deeper LSTM Systems was utilized to determine speaker gender from an acoustic records collection. Three major stages comprise the suggested method. To begin, the ten highly efficient information characteristics were chosen. Subsequently, using the double-layer LSTM structure, a DL technique was constructed. Finally, the classification performance comparison, specificity, sensitivity, and accuracy measures were computed [30]. Simultaneously, the suggested method's recognition rate was studied and compared with the performance produced using standard ML techniques. The research has a 98.4% success rate in predicting gender.

While voice gender identification is a difficult job that has been widely researched in the literature, the task becomes more difficult when sounds and uncontrolled surroundings surround the voice. Two Self-Attention-based approaches were given for implementing speech gender identification techniques in uncontrolled settings. The initial approach is composed of self-attention six levels and a dense level. The next approach augments the first model with convolutional layers and six inception-residual blocks before the self-attention layers. These approaches use MFCC to describe the speech input and Logistic Regression to classify it [31]. The tests were conducted in unrestricted settings, including background noise and speakers with various expressions, ages, accents and languages. The findings indicate that the suggested models were capable of 95.11% and 96.23% accuracy, respectively. These models outperformed all other models in every criterion and are considered literature work for Speech Gender Identification in unrestricted settings [32].

Gender identification was accomplished using supervised machine learning on voice signals. To begin, a database of Polish language speech signals was created. The following database contains male and female audio records of 6,797 s and 6,825 s, respectively with overall 3 s audio records of 13,649. The R language was used with the warbleR and specprop libraries to extract characteristics from these signals, including the mean, median, standard deviation, 1st quartile, 3rd quartile of frequency, and up to 20 other features. Additionally, the characteristics were utilized for NN training. The acoustic speech recognition and NN construction have being carried out in Python, while the feature computation was carried out in the R language. The training of neural networks was conducted through just the CPU, followed by CPU and GPU. The computation complexity of the algorithms was calculated. 92.4% gender recognition rate

was achieved. The usage of GPUs significantly expedited the system learning procedure. During training, the model was exposed to 10,927 s samples. Male and female audio records of 1,307 s and 1,415 s, respectively, were included in the testing set. Of the 2,722-test data, 2,514 instances had the gender properly recognized, whereas 208 cases had the gender erroneously identified. This results in an overall accuracy of 92.4% for gender recognition [33].

A proposed methodology uses deep neural networks in the expressive area of age and speaker identification through a person's speaking. The suggested model is trained and tested using the German speech corpus Gender. Tests were performed using a variety of network topologies, including FC and CNNs. The proposed network achieved an effectiveness of 48.41% in unweighted accuracy in a combination of individual age and gender identification. The performance achieved in a distinct mechanism of age and gender identification was 88.80% and 57.53%, correspondingly. In contrast to current conventional classification techniques, applied DNNs deliver the greatest outcomes for individual age identification [34].

Specifically for the airline industry, Jasuja *et al.* [35] designed a method to predict consumers' travel ancd tourism referral behavior. A multi-label classification strategy was used in the development of their forecasting methods. They relied on K-Nearest Neighbors, Support Vector Machine, Multi-layer Perceptron, Logistic Regression, Random Forest, and Ensemble Learning as core classification techniques to train their model. Additionally, they improved their predictive model by using label space partitioning techniques such as RAkELo and Louvain as a transformation technique to convert each label set into a multi-class classification problem. Using a variety of metrics, the suggested model outperformed previous binary classifications in terms of accuracy. Consumer recommendation plans were online systems that made it easier for customers as well as advertisers to deal with the amount of information available online by learning about user preferences and making suggestions based on those preferences. The travel business created machine learning and consumer recommendation engines to help customers have better experiences based on smart, data-driven decisions. ML algorithms were used by airline marketing to get a better idea of what their services were like [36].

In the study by Wahyuni [37], a Sparse Self-Attention Network (SpSAN) model was introduced to predict CRDs. The SpSAN model improved the fine-tuning performance of the Bidirectional Encoder Representations from Transformers (BERT) by incorporating sparsity into the self-attention procedure. The research achieved this by replacing the softmax function with a controllable sparse transformation during the fine-tuning phase with BERT. In article [38], a Self-Attentive Convolution Neural Network (SACNN) was introduced to solve natural language inference and sentence pair modelling tasks. The model employed CNNs to incorporate the interactions between sentences, considering each sentence's representation formulation alongside its corresponding counterpart. Table I presents state of art comparisons of different techniques reported with their pros and cons.

TABLE I. COMPARISON OF RELATED WORK AND ITS PROS, CONS

Techniques	Dataset	Pros	Cons
Deep Learning, MFFC, LPC, LSP [16]	Berlin voice sounds	The paper focuses on categorization, feature extraction, and selection techniques for voice recognition.	The procedure is critical for techniques that are deployed by the authors in real-time scenarios.
MFCC, VQLBG and GMM [24]	Arabic speech dataset	The feature vectors for each spoken sample are aggregated using the VQLBG procedure and the GMM (Gaussian Mixer Model) to classify and recognize an unknown speaker based on his spoken phrases	The words spoken by each Arabic speaker will be classified as belonging to a distinct cluster distinct from other clusters associated with other Arabic speakers.
KNN, DT and HMM [35]	720 sounds	A novel model for Arabic speech recognition was developed that automatically determines the speaker's gender. This research aimed to continue investigating methods for directly mapping the raw audio wave shape to the appropriate languages.	The accuracy of the speech recognition based on the gender was quite poor, and the dataset was limited.
SDC, DNN, MFCC [27]	Audio sounds	For each test speech word, a comparison is made between the gender class model and the test model. The values of the SDC (shifting delta cepstral) and MFCCs were used as feature sets.	The accuracy of 57.21% that was attained may be regarded as quite poor, indicating that more improvements are required to get better precision. The performance of the models may also differ based on the dataset employed, and they can have trouble extrapolating to new data.
Bidirectional LSTM, MFCC [30]	Audio sounds	The technique is capable of correctly identifying the speakers' gender utilizing a small database of thousand audio samples containing five hundred men and five hundred women	Despite the technique's encouraging accuracy findings, which range from 81.0% to 90.5% and have an overall identification rate of 86.7%, there is a clear difference between the minimum and maximum accuracy attained.
CNN, DNN, FC [16]	German speech	The proposed model is trained and tested using the German speech corpus Gender.	The comparatively low accuracy of gender recognition shows that there could be difficulties in accurately capturing the differentiating factors required for gender categorization.

III. Materials and Methods

This research aims to develop the best technique for a voice-based gender identification system. As a result, we optimized the pre-processing phase, especially the audio signal filtering, to attain the better possible technique performance. The proposed investigative data assessment and appraisal technique consists of three phases: audio dataset pre-processing, feature extraction, selection and gender classification.

A. Audio Preprocessing

In pre-processing phase, investigation involves evaluating various methods, including pre-emphasis, framing, windowing, and Fast Fourier Transform.

B. Pre-emphasis

This technique increases the signal's high-frequency component, flattens its frequency spectrum, and maintains it within the full frequency range from low to high. The frequency spectrum may be determined using the same signal-to-noise ratio [21]. Simultaneously, the situation is to remove the influences of the lips and vocal cords during the occurrence procedure, which make up for the suppressed maximum frequency component of the voice signal caused by the pronunciation scheme, towards emphasizing the high-frequency setup. The pre-emphasis technique is a high-pass filter applied to the speech signal [19] as shown in Eq. (1). In Eq. (1), the filter is signified as x. In contrast, filter coefficients of pre-emphasis are denoted via μ values between 0.9 and 1.

$$H(x) = 1 - \mu x1 \quad (1)$$

Figs. 1 and 2 show original female and male audio samples.

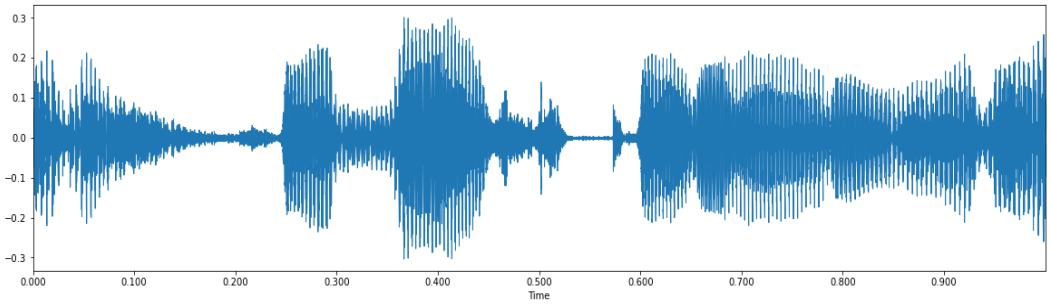

Fig. 1. Original female audio sample.

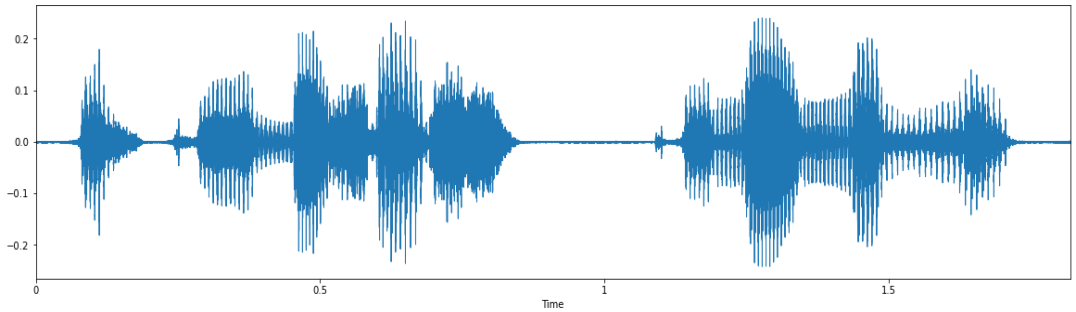

Fig. 2. Original male audio sample.

C. Framing and Windowing

Although the speech signals are non-stationary, it is typically stationary within a specific time variation between 20 ms to 40 ms, i.e., frames or small windows. The audio wave has been split into multiple frames and processed throughout the framing process. First, collect N sample points into a single observation unit called a frame. Normally, N is 256 or 512, about 20 to 30 ms. There will be an overlapping space between two consecutive frames to prevent too big changes between them. This overlapping region has M sampling points, typically equal to 1/2 or 1/3 of N. In most cases, the voice signal in speech recognition has a sample frequency of 8 KHz or 16 KHz. When the sampling point is 256 for the length of the frame, then 256/80001000 = 32 ms is the time duration for 8 KHz [19].

To improve the steadiness amongst frame's initial and last portions, the hamming window is increased frame by frame. Suppose that the resulting signal following framing is $F(x)$, x=0, 1, ..., X–1, X denotes size of the frame, therefore again increased the Hamming window [20].

$$F'(x) = F(x) \times G(x) \quad (2)$$

After that, $W(n)$ has the following form:

$$G(x, b) = (1b)\ b \times \cos(2\pi x / X - 1),\ 0 \le x \le X - 1 \quad (3)$$

D. Fast Fourier Transform (FFT)

The Fast Fourier Transform method calculates any order's Discrete Fourier Transform (DFT) or inverse form. This method transforms Z audio records signal frame by frame into the frequency domain from the time domain in

the speech voice signal. The Fast Fourier Transform method is an effective version of the discrete Fourier transform technique, which is described as follows for a collection of *Z* samples *Xz*:

$$X_{k=} \sum_{z=0}^{Z-1} x_z e^{-\left(\frac{j2\pi kz}{Z}\right)}, k = 0,1,2, \ldots, Z-1 \quad (4)$$

$$0 \leq f < (1/2)\, Fs \Rightarrow 0 \leq z \leq (1/2)\, Z - 1 \quad (5)$$

$$-(1/2)\, Fs < f < 0 \Rightarrow (1/2)\, Z + 1 \leq z \leq Z - 1 \quad (6)$$

In Eq. (4), composite values are denoted by X_k with a frequency magnitude or absolute output, which may be deduced through Eqs. (5) and (6) using positive and negative frequencies with the associated computations. Fs represent the sampling rate. The frequency spectrum of the voice signal [27] is the name given to the findings produced.

Feature extraction and fusion: Feature's extraction plays the main role for the image representation in machine and deep learning domain.

E. Mel-Frequency Cepstral Coefficients (MFCC)

In this research, we utilize MFCC to distinguish and recognize gender through speech signals. In practice, we use LibROSA, a python library for audio signal analysis. MFCC is a popular technique for extracting speech recognition features. Its visual and auditory systems are modelled like those of humans. As a result, it is important in a variety of speech-related applications for example speaker, gender identification and speech recognition and so on. First, the Fourier transform of the windowed signal is retrieved. The next stage is to use triangle filter banks to map the spectrum's power on the Mel-scale. Take the log of the triangle filter bank's resultant. Calculate the DCT of the log of the Mel power signal as well. The amplitudes of the DCT resultant result are used to calculate MFCCs. There is a total of 13 MFCC characteristics that have been retrieved.

MFCC feature descriptor is the quickest speech recognition method in case of computations cost with decent characteristics integrity. Higher DCT coefficients are used to minimize the filter bank energy' rapid changes and reject the remainder. Figs. 3 and 4 present male audio sample of MFCC and Delta MFCC features post-preprocessing. Similarly, Fig. 5 presents male audio sample post-preprocessing.

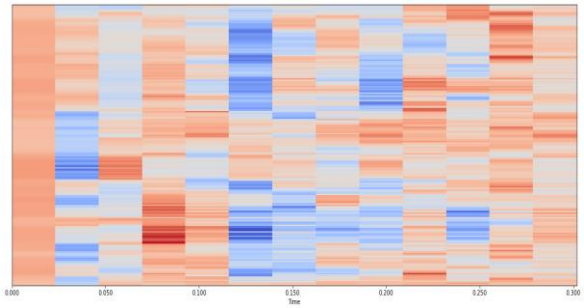

Fig. 3. Male audio sample Delta MFCC features after signal pre-processing.

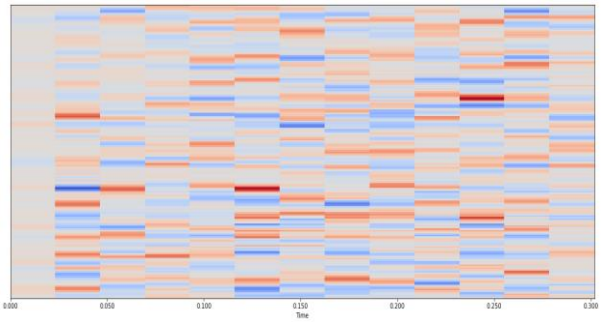

Fig. 4. Male audio sample MFCC features after signal pre-processing.

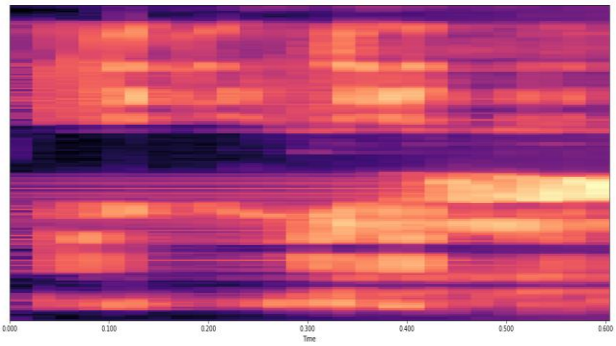

Fig. 5. Male audio sample log filter bank features after signal pre-processing.

F. Delta or Velocity of MFCC

Temporal features contain dynamic information that may be used to derive the audio signal's aggressive behaviour. One such temporal characteristic is velocity, which is derived from stationary MFCCs. Signal's temporal data could be unable to identify due to stationary MFCCs. Temporal information may be used to infer a speaker's gender. As a result, velocity characteristics have been utilized to augment MFCCs. After using velocity information, we observed adequate differentiation in the classification performance. Calculating first-order derivatives on the MFCCs yields the velocity characteristics. As a result, they are often referred to as Delta () coefficients. Where the velocity characteristics are expressed in terms of MFCCs. Eq. (7) provides the generic formula for computing the coefficients.

$$\Delta Cep_k(t) = \frac{\sum_{i=-x}^{x} y_i Cep_k(t+1)}{\sum_{i=-x}^{x} |i|} \quad (7)$$

where *t* is denoted is the frame time for the kth attribute of MFCCs $\mathrm{Cep_k(t)}$ features. *x* and y_i describe full consecutive and antecedent frames and extra ith weight, respectively. Generally, *x* is supposed as 2 in the experimental processes. Total 13 delta MFCCs features are extracted [1].

Algorithm 1: Algorithm of proposed methodology

Input: Speech Signals

Output: Gender classification

Start:

1. Original input speech signals
2. Preprocessing:

pre-emphasis,

Farming,
Windowing,
FFT techniques.

3. Feature Engineering:
   MFCC,
   Delta 2,
   MFCC and Log Filter Bank
4. Concatenate ← Extracted Features which are computed in step 3.
5. Neural network models ← Concatenated Features
6. Gender prediction

End

G. Log Filter Bank

The log filter bank is approximated using 26 vectors of different lengths based on the FFT computation. Each vector contains mainly zeros but contains values greater than zero for a portion of the spectrum. To get the energies of filter banks, every filter bank is increased by the power spectrum and followed by the coefficient's improvement. After this is completed, 26 numbers are remained which indicate the amount of energy contained in each filter bank. Maintain a record of the 26 energies. As a result, 26 energies of log filter bank are obtained.

IV. Proposed Hybrid Model

The proposed neural network model is built using the Keras library. Using stratified k-fold cross validation, scikit-learn is used to assess the model. This is accomplished by partitioning the data into k parts and training the model on all except one of them, which serves as a test set for evaluating the model's performance. This method is continuous k times, then the standard score obtained from all built models is used to get a robust assessment of performance. We must use the Keras classifier wrapper with scikit-learn in order to utilize Keras models with scikit-learn. This class accepts a method that is responsible for creating and returning our neural network model. Additionally, it accepts parameters that will be sent to fit (), such as the number of epochs and the batch size. With 60 neurons and 52 input variables, our model will have a single fully connected hidden layer. A new layer (one line) to the network is added, which adds a second hidden layer with 30 neurons following first. Initialize the weights using a tiny Gaussian random number. The activation function of the Rectifier is utilized. To make predictions, the output layer comprises a single neuron. crisp class values can be and consequentially translated from utilizing activation function (sigmoid) that provide possibility outcome amongst 0 and 1 in the output layer.

Conclusively, during training of binary classification tasks, binary cross entropy (logarithmic loss function) has been applied, which is the ideal loss function. Additionally, the network employs the effective Adam optimizer method for accuracy and gradient descent measurements are gathered during training. The accuracy of the network is used to determine its performance.

Fig. 6 presents proposed hybrid speech-based gender classification model.

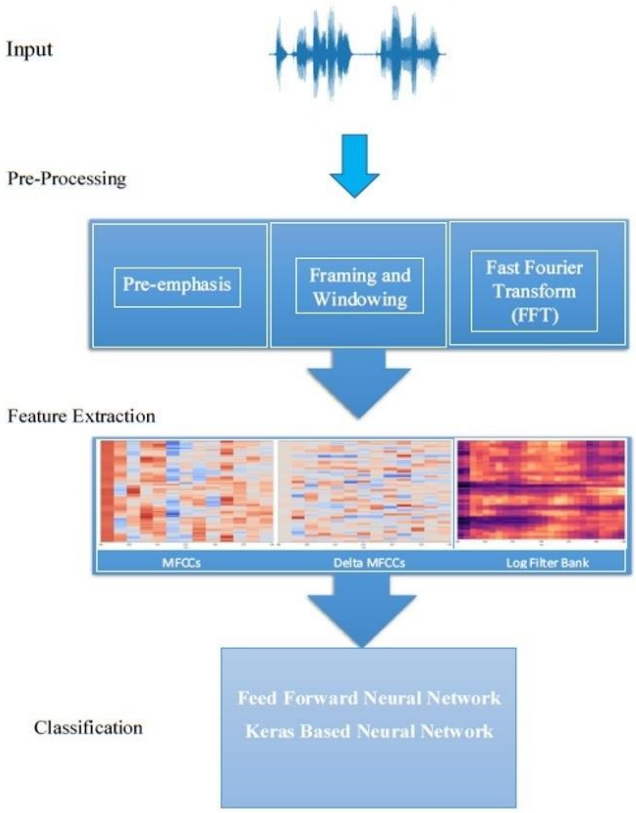

Fig. 6. Proposed research framework.

V. Results Analysis and Comparisons

Numerous open-access databases are available for training and testing systems for automated speech recognition and speaker verification/identification. The experimental procedure uses the Urban Jordanian Arabic (UJA) dataset. UJA is a corpus of Arabic recordings spoken by 2/3 of Jordan people in the country's main cities. Microphone N/D767 Electro-Voice and solid-state recorder PMD671 Marantz were used with a 22.05 kHz sampling rate in WAV audio format while capturing entire speakers at Hashemite University in Zarqa, Jordan.

This research analyzed a UJA speech dataset that included male and female speakers; there are six males and six women in these sound recordings.

This section summarizes the major findings of gender estimates and categorization analysis. The experiments were conducted by splitting the dataset into three subsets: first for training, which includes 70% of the available files; second for validation, which contains 15% of the files; and a third for evaluation, which contains 15%. Speaker independence is maintained between the training, validation, and assessment datasets.

As previously stated, we assess the proposed approach's performance by extracting feature sets such as MFCCs, delta MFCCs, and log filter banks. A speech utterance is split into 25 ms frames for the MFCCs. Each frame is retrieved using 52 features, 12 MFCCs, 12 delta MFFCs first derivatives, and 26 Log filter banks. As discussed in the preceding sections, feed-forward neural networks and Keras-based neural network models use these features for pattern recognition. The new technique's effectiveness was tested and compared to current methods using performance metrics such as cross-entropy (Table II), Receiver Operating Characteristic (ROC) confusion matrices, and accuracy. The performance of the feedforward neural network is presented in Figs. 7–9.

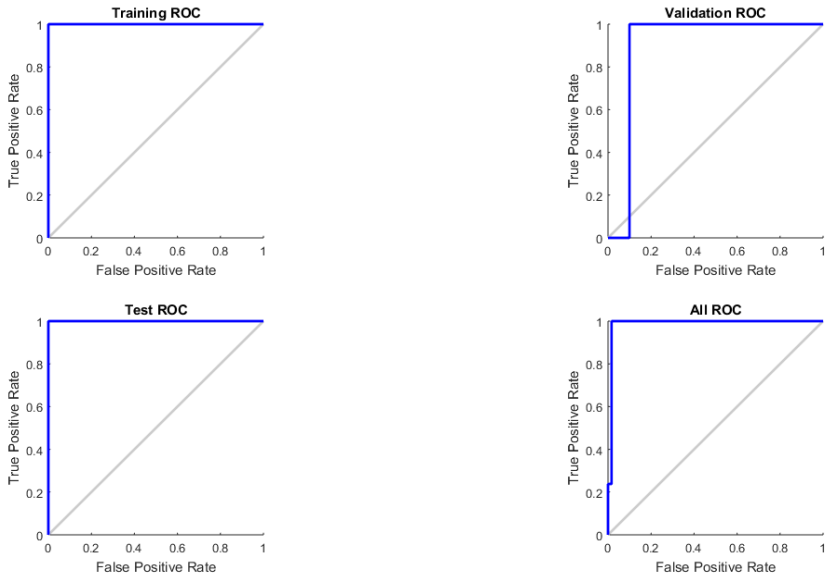

Fig. 7. ROC performance obtained using Feed forward neural network.

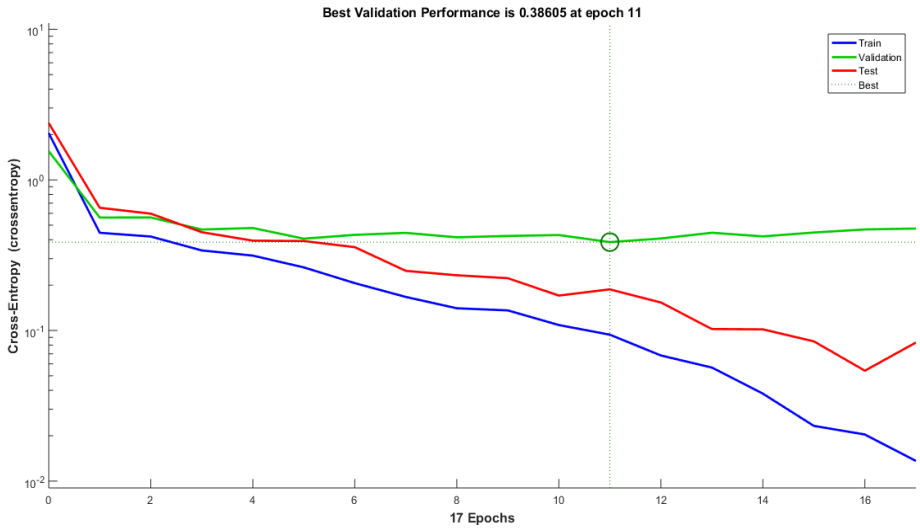

Fig. 8. Validation performance obtained using Feed forward neural network.

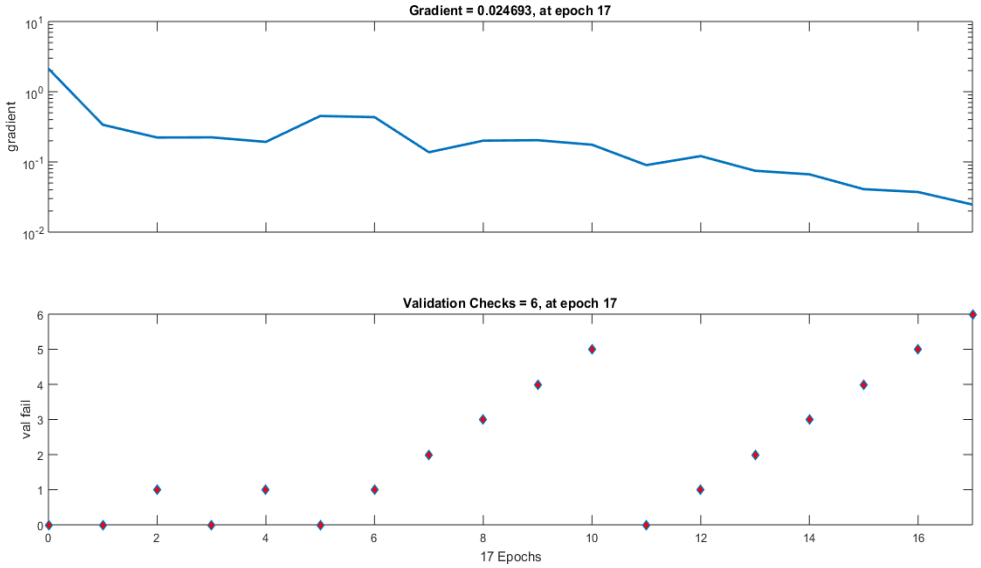

Fig. 9. Training state of feed forward neural network.

All tests were conducted using a Windows 10 operating system with 8GB RAM, an Intel (R) Core (TM) m3-7Y30 CPU running at 1.00 GHz and MATLAB 2016a. The experiments were executed in two stages:

A. Feed Forward Neural Network

The performance of the feedforward neural network is presented in Table II.

Table II described that better classification achieved through low Cross Entropy. 0 denotes no miscalculation. Percent Error reveals the misclassification of audio records by representing misclassifications portion by zero while highest misclassifications through 100.

TABLE II. PERFORMANCE EVALUATION RESULTS USING FEED FORWARD NEURAL NETWORK

Items	Samples	Cross Entropy	Percent Error
Training	72	7.403e–1	2.777e–0
Validation	15	2.042e–0	6.666e–0
Testing	15	2.045e–0	6.666e–0

The comprehensive analysis of feed-forward neural network is described in Figs. 10 and 11.

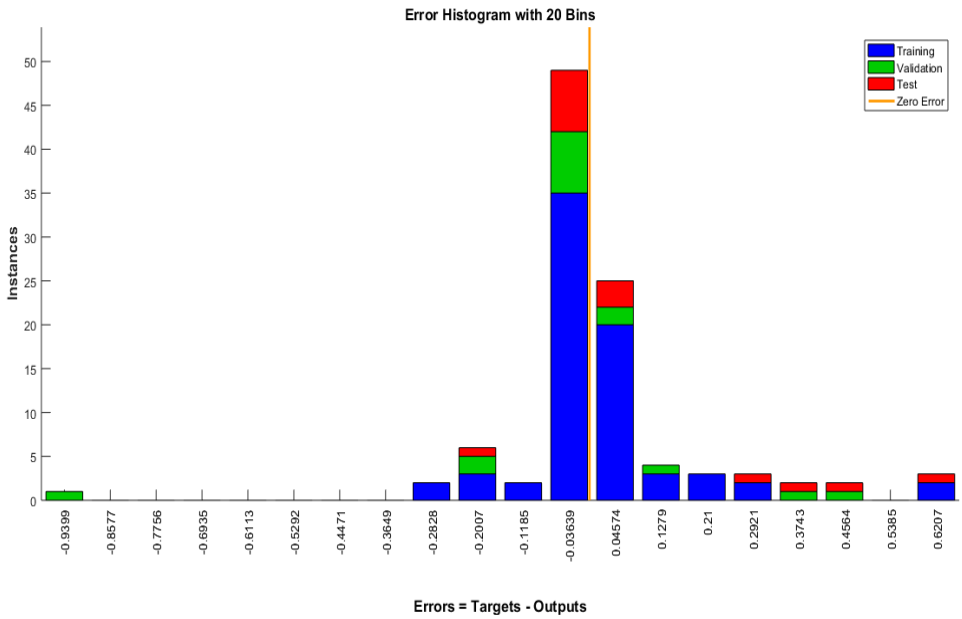

Fig. 10. Error histogram of feedforward neural network.

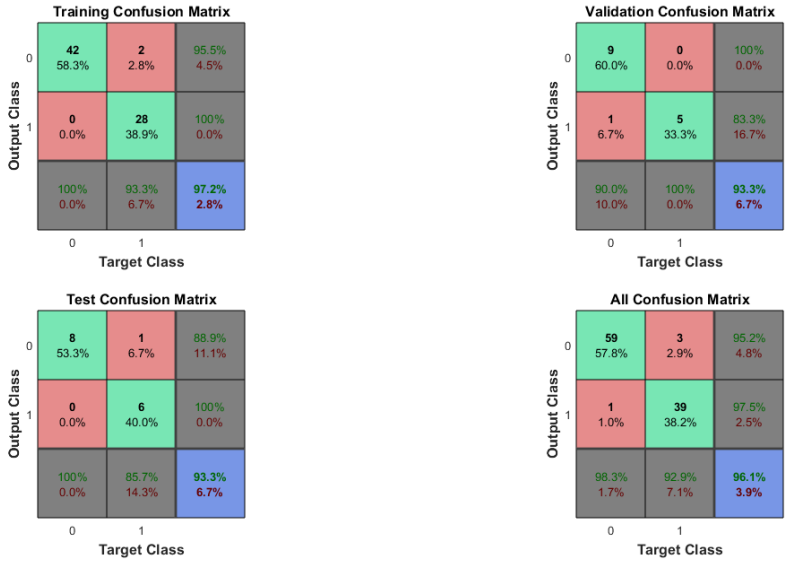

Fig. 11. Confusion matrices using feedforward neural network.

In confusion matrices, the number 0 indicates male-class speakers, whereas the number 1 denotes female-class speakers. Confusion matrices showed a classification rate of 96.1% and a misclassification rate of 3.9%. Male speakers are properly categorized 96.3% of the time, whereas female speakers are incorrectly classified 1.7%. 92.9% of female speakers are properly categorized, whereas 7.1% of male speakers must be correctly identified.

B. Keras-Based Neural Network

The proposed estimator is based on a neural network that was built in Python using Keras, the deep learning library. Keras is a Python-based high-level neural network API that plays as the leading role in as compared to that of Theano, CNTK or TensorFlow. It was designed with the goal of facilitating rapid experimentation [28].

The experiments performed without applying any preprocessing techniques using the proposed method. The proposed method achieved an 89.93% accuracy score before applying any preprocessing stages, and when we applied preprocessing techniques to the dataset and performed experiments, the proposed model achieved a 93.09% accuracy score in gender voice recognition. To get the best results from various machine learning or deep network jobs, preprocessing is quite helpful. Preprocessing with Fast Fourier Transform, windowing, frame, and pre-emphasis provides outclass results (see Table III).

TABLE III. PROPOSED METHOD ACCURACY BEFORE AND AFTER APPLYING PREPROCESSING TECHNIQUES

Proposed method	Accuracy (%)
Before preprocessing	89.93%
After preprocessing	93.09%

Keras functions were used to implement the estimator. The recognition accuracy was found to be 93.09 % when different combinations of retrieved features were used in conjunction with the Keras base Neural network model. We trained and tested the proposed model using tenfold cross validation in order to make predictions on all of our data. The dataset is grouped into ten equally portions (folds). Each fold serves as a validation set once, while the remaining nine folds serve as training data. The procedure is performed ten times, and the average error across the ten trials is calculated. The training uses a loss function (binary cross-entropy) of 6.37% and optimization across 250 epochs. Fig. 12 presents comparison of proposed model in state of art.

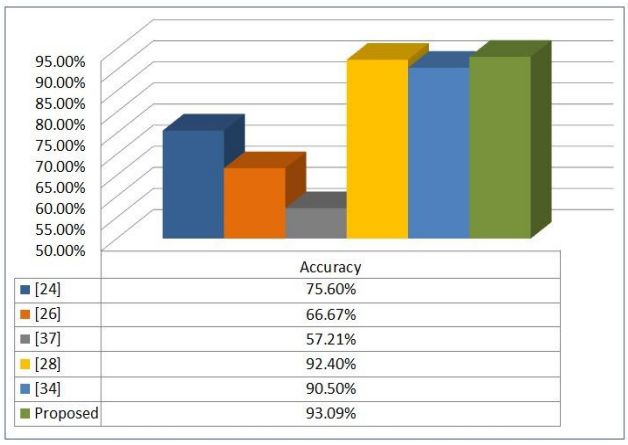

Fig. 12. Comparison with state-of-the-art techniques.

The comprehensive analysis demonstrates that the proposed approach outperforms state-of-the-art techniques. Physical features of speech are examined to determine gender [16]. Pitch and formant are gender-dependent characteristics; however, generic speech features such as fundamental frequency, autocorrelation coefficients, Linear Prediction Coefficients (LPC), and Mel-Frequency Cepstral Coefficients (MFCC) may also be utilized. Due to the limitation of the feature's number, the accuracy of gender identification decreased, as in the case of female

voice adoption, the model failed to perform well. The network often misidentified the speaker as a woman.

VI. Conclusion and Future Work

Recent improvements in deep learning have resulted in neural network-based Arabic speech recognition outperforming statistical pattern recognition methods by a wide margin. The greatest results are obtained using neural networks with RELU and sigmoid activations in the hidden layer. The primary objective was to create a gender recognition system based on speech signals. By integrating MFCCs, Delta MFCCs, and Log Filter Bank characteristics, this research offers a hybrid method for gender classification. Feed Forward Neural Networks and Keras-based Neural Networks are utilized as a classifier model.

Regarding accuracy and simplicity, the suggested hybrid model of gender categorization outperforms most previous approaches described in the literature. Before the preprocessing stage, the proposed method achieved an accuracy score of 89.93%. After applying preprocessing techniques to the dataset and conducting experiments, the proposed model achieved an accuracy score of 93.09% in gender voice recognition. Preprocessing is beneficial for maximizing the efficacy of machine learning and deep network tasks. The preprocessing results with Fast Fourier Transform, windowing, frame, and pre-emphasis are superior.

In the future, this work could be improved to handle large datasets of Arabic voice recognition with the aim of dimensionality reduction, speech feature extraction, and grouping using convolutional neural networks. This enhanced strategy will make processing Arabic voice data more effective and efficient. To improve the model's capacity for accurate speech recognition and comprehension even in difficult and noisy circumstances, noisy speech samples can test the model's resilience. By enabling better performance and application in real-world circumstances, this upcoming extension will aid in developing Arabic Speech Recognition technology.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgement

This research is supported by Artificial Intelligence & Data Analytics Lab (AIDA) CCIS Prince Sultan University, Riyadh, Saudi Arabia. The author also would like to acknowledge the support of Prince Sultan University for paying the APC of this publication.

References

[1] K. Alrajhi and M. A. Elaffendi, "Automatic Arabic part-of-speech tagging: Deep learning neural LSTM versus Word2Vec," *International Journal of Computing and Digital Systems*, vol. 8, no. 3, pp. 307–315, 2019.

[2] L. Berriche, "Comparative study of fingerprint-based gender identification," *Security and Communication Networks*, 1626953, 2022.

[3] M. Alian, A. Awajan, and B. Ramadan, "Unsupervised learning blocking keys technique for indexing Arabic entity resolution," *Int. J Speech Technol*, vol. 22, pp. 621–628, 2019.

[4] S. L. M. Sainte *et al*., "A new framework for Arabic recitation using speech recognition and the Jaro Winkler algorithm," *Kuwait J. Sci*., vol. 49, 2022.

[5] E. Alsharhan and A. Ramsay, "Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition," *Language Resources and Evaluation*, vol. 54, no. 4, pp. 975–998, 2020.

[6] F. Afza *et al*., "A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection," *Image and Vision Computing*, vol. 106, 104090, 2021.

[7] O. Mamyrbayev *et al*., "Neural architectures for gender detection and speaker identification," *Cogent Engineering*, vol. 7, no. 1, 1727168, 2020.

[8] E. Sezgen, K. J. Mason, and R. Mayer, "Voice of airline passenger: A text mining approach to understand customer satisfaction," *Journal of Air Transport Management*, vol. 77, pp. 65–74, 2019.

[9] K. Nugroho, E. Noersasongko, and H. A. Santoso, "Javanese gender speech recognition using deep learning and singular value decomposition," in *Proc. 2019 International Seminar on Application for Technology of Information and Communication*, 2019, pp. 251–254.

[10] I. E. Livieris, E. Pintelas, and P. Pintelas, "Gender recognition by voice using an improved self-labeled algorithm," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 492–503, 2019.

[11] M. Alian, A. Awajan, and B. Ramadan, "Unsupervised learning blocking keys technique for indexing Arabic entity resolution," *International Journal of Speech Technology*, vol. 22, pp. 621–628, 2019.

[12] T. Saba, A. Rehman, and G. Sulong, "Cursive script segmentation with neural confidence," *Int. J. Innov Comput. Inf. Control (IJICIC)*, vol. 7, no. 7, pp. 1–10, 2011.

[13] F. Ertam, "An effective gender recognition approach using voice data via deeper LSTM networks," *Applied Acoustics*, vol. 156, pp. 351–358, 2019.

[14] I. E. Livieris, E. Pintelas, and P. Pintelas, "Gender recognition by voice using an improved self-labeled algorithm," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 492–503, 2019.

[15] S. Rami and D. G. R. Alkhawaldeh, "Gender recognition of human speech using one-dimensional conventional neural network," *Sci. Program*, vol. 3, 2019.

[16] M. O. A. Albaraq, "Arabic speaker recognition system using gaussian mixture model and em algorithm," *International Journal of Advanced Research in Computer Science*, vol. 11, no. 2, 2020.

[17] M. H. Al *et al*., "Harris Hawks sparse auto-encoder networks for automatic speech recognition system," *Applied Sciences*, vol. 12, no. 3, 1091, 2022.

[18] A. Rehman and T. Saba, "Performance analysis of character segmentation approach for cursive script recognition on benchmark database," *Digital Signal Processing*, vol. 21, no. 3, pp. 486–490, 2011.

[19] M. Raza *et al*., "Appearance based pedestrians' gender recognition by employing stacked auto encoders in deep learning," *Future Generation Computer Systems*, vol. 88, pp. 28–39, 2018.

[20] T. Saba *et al*., "Fundus image classification methods for the detection of glaucoma: A review," *Microscopy Research and Technique*, vol. 81, no. 10, pp. 1105–1121, 2018.

[21] T. Saba, A. Rehman, and G. Sulong, "An intelligent approach to image denoising," *Journal of Theoretical and Applied Information Technology*, vol. 17, no. 2, pp. 32–36, 2010.

[22] A. Rehman and T. Saba, "Off-line cursive script recognition: current advances, comparisons and remaining problems," *Artificial Intelligence Review*, vol. 37, pp. 261–288, 2012.

[23] P. K. Jain, E. A. Yekun, R. Pamula, and G. Srivastava, "Consumer recommendation prediction in online reviews using Cuckoo optimized machine learning models," *Computers and Electrical Engineering*, vol. 95, 107397, 2021.

[24] B. Mouaz *et al*., "A new framework based on KNN and DT for speech identification through emphatic letters in Moroccan dialect,"

Indonesian Journal of Electrical Engineering and Computer Science, vol. 21, no. 3, pp. 1417–1423, 2021.
[25] P. K. Jain *et al*., "SpSAN: Sparse self-attentive network-based aspect-aware model for sentiment analysis," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 4, pp. 3091–3108, 2023.
[26] S. Joudaki *et al*., "Vision-based sign language classification: A directional review," *IETE Technical Review*, vol. 31, no. 5, pp. 383–391, 2014.
[27] R. S. Alkhawaldeh, "DGR: Gender recognition of human speech using one-dimensional conventional neural network," *Scientific Programming*, 7213717, 2019.
[28] H. A. S. Hevia *et al*., "Convolutional-recurrent neural network for age and gender prediction from speech," in *Proc. 2019 Signal Processing Symposium*, 2019, pp. 242–245.
[29] M. M. Nasef, A. M. Sauber, and M. M. Nabil, "Voice gender recognition under unconstrained environments using self-attention," *Applied Acoustics*, vol. 175, 2021.
[30] S. ElSayed and M. Farouk, "Gender identification for Egyptian Arabic dialect in twitter using deep learning models," *Egyptian Informatics Journal*, vol. 21, no. 3, pp. 159–167, 2020.
[31] M. Harouni *et al*., "Online persian/Arabic script classification without contextual information," *The Imaging Science Journal*, vol. 62, no. 8, pp. 437–448, 2014.
[32] A. Majkowski *et al*., "Identification of GENDER based on speech signal," in *Proc. 2019 IEEE 20th International Conference on Computational Problems of Electrical Engineering*, 2019.
[33] W. Quamer *et al*., "SACNN: Self-attentive convolutional neural network model for natural language inference," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 3, pp. 1–16, 2021.
[34] M. Markitantov and O. Verkholyak, "Automatic recognition of speaker age and gender based on deep neural networks," *Speech and Computer*, pp. 327–336, 2021.
[35] L. Jasuja, A. Rasool, and G. Hajela, "Voice gender recognizer recognition of gender from voice using deep neural networks," in *Proc. 2020 International Conference on Smart Electronics and Communication (ICOSEC)*, 2020, pp. 319–324.
[36] S. Hamdi *et al*., "Gender identification from Arabic speech using machine learning," in *Proc. International Symposium on Modelling and Implementation of Complex Systems*, 2020, pp. 149–162.
[37] E. S. Wahyuni, "Arabic speech recognition using MFCC feature extraction and ANN classification," in *Proc. 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering*, 2017, pp. 22–25.
[38] G. L. Soon *et al*., "Evaluating the effect of multiple filters in automatic language identification without lexical knowledge," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, 2020.