Hindawi
Scientific Programming
Volume 2019, Article ID 7213717, 12 pages
https://doi.org/10.1155/2019/7213717

Research Article

DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network

Rami S. Alkhawaldeh

Department of Computer Information Systems, The University of Jordan, Aqaba 77110, Jordan

Correspondence should be addressed to Rami S. Alkhawaldeh; r.alkhawaldeh@ju.edu.jo

Received 14 March 2019; Revised 30 July 2019; Accepted 19 August 2019; Published 9 September 2019

Academic Editor: Autilia Vitiello

The speech entailed in human voice comprises essentially paralinguistic information used in many voice-recognition applications. Gender voice is considered one of the pivotal parts to be detected from a given voice, a task that involves certain complications. In order to distinguish gender from a voice signal, a set of techniques have been employed to determine relevant features to be utilized for building a model from a training set. This model is useful for determining the gender (i.e., male or female) from a voice signal. The contributions are three-fold including (i) providing analysis information about well-known voice signal features using a prominent dataset, (ii) studying various machine learning models of different theoretical families to classify the voice gender, and (iii) using three prominent feature selection algorithms to find promisingly optimal features for improving classification models. The experimental results show the importance of subfeatures over others, which are vital for enhancing the efficiency of classification models' performance. Experimentation reveals that the best recall value is equal to 99.97%; the best recall value is 99.7% for two models of deep learning (DL) and support vector machine (SVM), and with feature selection, the best recall value is 100% for SVM techniques.

1. Introduction

The voice of human speech is an effective communication method consisting of unique semantic linguistic and paralinguistic features such as gender, age, language, accent, and emotional state. The sound waves consisting of human voice are unique among all creatures producing sound since every single wave carries a different frequency. Identifying human gender based on voice has been a challenging task for voice and sound analysts who deploy numerous applications including (i) effective advertising and marketing strategies in customer relationship management (CRM) systems which depend on gender interoperability such as the user interface style as well as preferences of words and colours; (ii) investigating criminal voice in crime scenarios; and (iii) enhancing human-computer interaction (HCI) systems especially dialogue systems by customizing services that rely on gender voice and also improving the level of user satisfaction. Because of the importance of identifying gender through voice recognition, the human voice should be converted from the analogue to the digital form to extract useful features and then to construct classification models. The robustness and effectiveness of classifiers are determined by the quality of features that depend on a training set employing machine learning (ML) techniques. Therefore, eliciting voice features plays a vital role in improving the efficiency of classifiers since the human voice is liable for nonuseful features. Research on improving the efficiency of voice classifiers is copious, particularly studying the process of extracting efficient features from voice including identifying the linguistic content of speech signal components and disposing of nonuseful contents such as background noise.

There are a set of features used for recognizing the voice gender. Among the most common features utilized for voice gender recognition are mel-scaled power spectrogram (Mel), mel-frequency cepstral coefficients (MFCCs), power spectrogram chroma (Chroma), spectral contrast (Contrast), and

tonal centroid features (Tonnetz). By getting the extracted features combined with the gender label as a form of a training set, ML techniques are used to build a high-quality model for recognizing the voice gender, as shown in Figure 1. In particular, each classification technique is used to build a set of hypothesis models and selects the most optimal one. This model classifies the unknown voice label by receiving the voice features and categorizing the voice gender.

A multitude corpus of research has been conducted to address the efficiency of voice classifiers aiming to enhance the accuracy of programs being used. Hu et al. [1] used two level classifiers (pitch frequency and GMM classifier) to recognize the speaker's gender on the TIDIGITS dataset achieving a success rate of 98.65%. Djemili et al. [2] used four classifiers including GMM, multilayer perceptron (MLP), vector quantization (VQ), and learning vector quantization (LVQ) to analyze voices taken from IViE corpus. They managed to achieve a 96.4% success rate. Li et al. [3] combined the estimated voice acoustic level of five different methods into one score level. The results were obtained on using the aGender dataset for the gender category with a 81.7% success rate. Yücesoy and Nabiyev [4] proposed a system for identifying speakers using a fusion score of seven subsystems where the feature vectors are the MFCC, PLP, and prosodic on three different classifiers that are GMM, SVM, and GMM-SV-based SVM combined at the score level. The classification success rate on gender identification using the aGender database is 90.4%. Lee and Kwak [5] used two classifiers: SVM and decision tree (DT) with the MFCC feature, on a private corpus identifying gender voice. The overall accuracies using MFCC-SVM and MFCC-DT for gender classification were 93.16% and 91.45%, respectively.

The most efficient classifiers and feature extractors of superior accuracy on voice gender recognition include deep neural networks (DNNs) and convolutional neural networks. Qawaqneh et al. [6] proposed an adequate technique to enhance the MFCC features and then adjust the weights between DNN layers. These improved MFCC features are evaluated on DNN and I-Vector classifiers where the overall accuracies are 58.98% and 56.13%, respectively. Sharan and Moir [7] compared two classification techniques (DNN and SVM) using single and combined feature vectors for robust sound classifications. The results showed a better performance for the DNN technique as it is robust and has low sensing to the noise approach.

In this work, a novel approach is presented for characterizing the voice gender using a different set of features along with different ML algorithms from various families. These features showed their effectiveness in extracting the voice patterns, hence categorizing the gender. The contributions are demonstrated as follows:

(i) Studying a set of voice features and examining their effects as possible suitable features for gender classification techniques

(a) *RQ1*. To what extent is selecting voice signal features useful on building machine learning classifiers?

(ii) Using different ML techniques of various families for recognizing the speech gender from the extracted and efficient features

(b) *RQ2*. What is the performance of various ML models in gender voice-recognition applications?

(iii) Evaluating well-known natural feature selection techniques for choosing the most optimal features

(c) *RQ3*. To what extent using natural feature selection evaluators is useful for enhancing the performance of ML techniques?

The structure of this paper is organized as follows: The related works are presented in Introduction. Section 2 discusses the proposed approach in classifying a given voice gender. It provides a detailed discussion of the classifiers of voice recognition including the phases of preprocessing, extracting features, ML techniques, evaluation metrics, and feature extraction methods. Section 3 presents the experimental settings and answers the research questions in each section thereafter. The conclusions and future work are discussed and summarized in Section 4.

2. Deep Gender Recognition (DGR)

The proposed methodology for speech gender classification includes a set of stages as briefly discussed. The stages start by converting the voice, from its abstract representation, into a consistent form in order to extract the relevant features. Then, the relevant features are selected as inputs for building a classifier model for recognizing the gender of a human voice. In addition, a DL model is being built to automatically extract useful features and feed them into a fully connected artificial neural network (ANN) for classification. However, here, a set of process for extracting features for other models rather than DL and classification techniques are summarized as follows.

2.1. Voice Preprocessing. A transmitted voice is inevitably vulnerable to noise interference and voice attenuation that needs a preprocessing process to purify it for feature extraction. This phase shows a set of steps as follows.

2.1.1. A/D Signal Conversion. A/D signal conversion is used to convert the given voice from the analogue to the digital signal by common sampling and quantization techniques [8]. The A/D conversion formulates the signal in an understandable form by machine for easy manipulation.

2.1.2. Preemphasis Process. Because of attenuation at high-frequency segments of the voice signal, there is a necessary need to use a preemphasis filter. The preemphasis filter flattens the signal (or speech) waveforms. The process filters low-frequency interference, especially power frequency interference at low-frequency segments, and emphasizes the high-frequency portions in order to produce a high-pass filter to carry out spectral analysis interference. This process

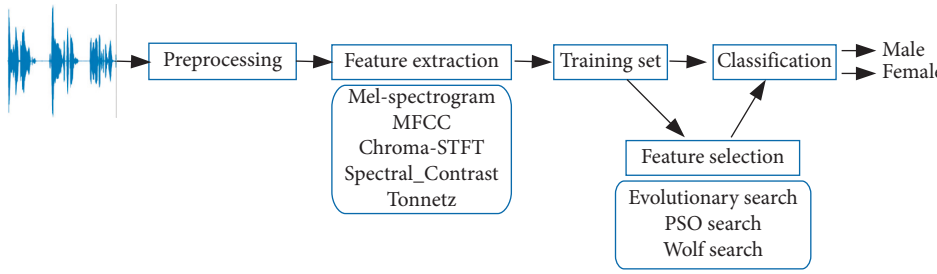


Figure 1: General gender recognition framework.

occurs after A/D conversion by the first-order digital pre-emphasis filter equation [9]:

$$H(z) = 1 - \mu z^{-1}, \quad (1)$$

where z represents the filter and μ is the preemphasis filter coefficient with the value ranging commonly within [0.9, 1].

2.1.3. Frame Blocking and Hamming Window. The frame blocking is a process of handling the filtered digital signal into a number of N small frame segments with adjacent frames separated by M ($M < N$). The process of the Hamming window minimizes speech signal discontinuities before and after each frame within the window frame. This method is popularly used in the MFCC before the mel-frequency warping step where mel scales are calculated. The analytical representation of the Hamming window is given by

$$w(n) = 0.54 - 4.4\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1, \quad (2)$$

where $w(n)$ is the window operation, n is the number of individual samples, and N is the total number of speech samples [9].

2.1.4. Fast Fourier Transform (FFT). The FFT algorithm is in general used for estimating the discrete Fourier transform (DFT) of any sequence, or its inverse form. In the speech voice signal, the FFT converts each frame of those N samples from the time-domain signal into a form of frequency domain [10]. The FFT is considered a computationally efficient implementation of the DFT method, which is defined on the set of N samples $\{x_n\}$ as follows:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-(j2\pi kn/N)}, \quad k = 0, 1, 2, \ldots, N-1, \quad (3)$$

where X_k is a complex number considered as an absolute value (frequency magnitude or modulus). $\{x_k\}$ is the resulting sequence interpreted as follows: the positive frequencies $0 \le f < (1/2)F_s$ correspond to the values $0 \le n \le (1/2)N - 1$, while the negative frequencies $-(1/2)F_s < f < 0$ correspond to $(1/2)N + 1 \le n \le N - 1$. F_s represents the sampling rate. The results obtained are called the frequency spectrum of the voice signal.

2.2. Extracting Features from Digital Voice Signals. There are a set of relevant features that could be inferred from the voice signal. Hence, a preprocessing phase is needed to prepare the speech signals as an input for a set of feature extraction techniques. These sets of features and a voice gender as a label represent the training set for building a classifier model in order to recognize the voice speech gender. For visualization, Figure 2 shows a voice sample, which is a British English female's voice and its features. The features used in this paper are as follows.

2.2.1. Mel-Spectrogram. Mel-spectrogram computes a mel-scaled power spectrogram coefficient. An object of mel-spectrogram type represents an acoustic time-frequency representation of sound, as shown in Figure 2(b). The power spectral density $P(f, t)$ is sampled into a number of points around equally spaced times t_i and frequencies f_j (on a mel-frequency scale). The mel-frequency scale is defined as

$$\text{mel} = 2595 * \log_{10}((1 + \text{hertz})/700). \quad (4)$$

2.2.2. MFCC. MFCC represents accurately the vocal tract that is a filtered shape of a human voice and also manifests itself in the envelope of a short-time power spectrum, as shown in Figure 2(c). In order to compute MFCCs, a set of sequential steps should be followed:

(1) Framing the Signal into Short Frames. The audio signal is framed into 20–40 ms (25 ms is standard) frames to overcome changes in the sample in a short time period as it is constantly changed in a long period of time.

(2) Periodogram of Power Spectrum. This calculates for each frame the periodogram estimation of the power spectrum, which identifies the frequencies in the frame.

(3) Applying the Mel Filterbank to the Power Spectra (or Summing the Energy in Each Filter). A filter is required for estimating the energies in various frequency regions that appear in a group of aggregated periodogram bins because of unnecessary information in periodogram spectral estimation. Hence, the mel filterbank estimates the energy near 0 Hz and then for higher frequencies as there is less concern for variations.

(4) Logarithm of All Filterbank Energies. Large variations of energies are scaled using a logarithmic scale as there are no different sounds in large energies. The logarithmic scale is a

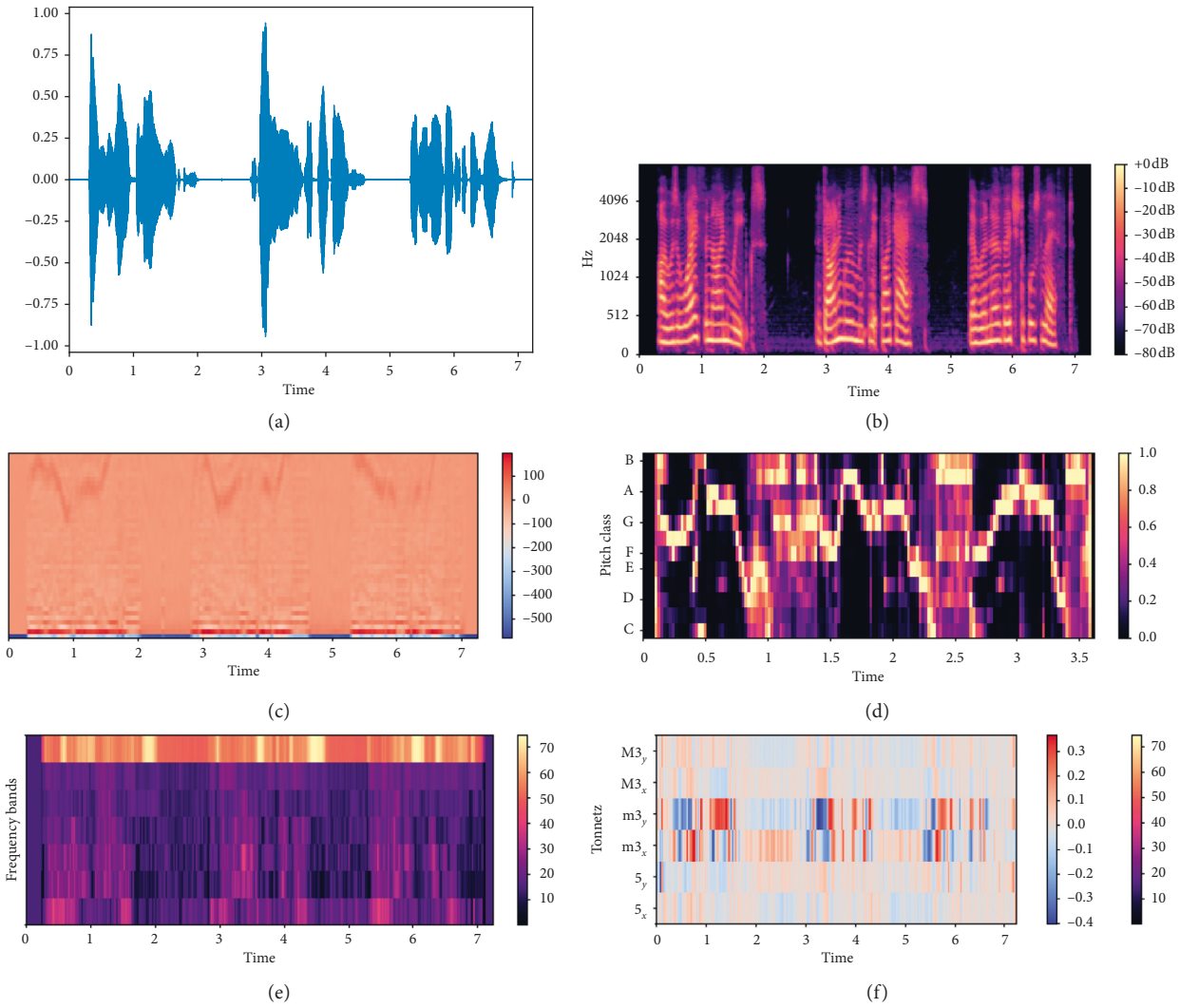


FIGURE 2: A British English female's voice sample and its features: (a) voice sample (B_eng_f1.wav); (b) mel-spectrogram; (c) MFCC; (d) chromagram; (e) spectral contrast; (f) tonal centroids (Tonnetz).

channel normalization technique that is also exploited for cepstral mean subtraction.

(5) DCT of the Log Filterbank Energies. Because of the correlation in filterbank energies that lead to overlapping, the DCT is used to decorrelate the energies. This generates diagonal covariance matrices as features.

(6) 2-13 DCT Coefficients. Higher DCT coefficients are chosen to reduce the fast changes in the filterbank energies and discard the rest.

2.2.3. Chorma-STFT (Short-Time Fourier Transform). Chorma-STFT computes a chromagram from a waveform or power spectrogram, as shown in Figure 2(d). Chroma features are powerful representatives for a music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or Chroma) of the musical octave.

2.2.4. Spectral_Contrast. Spectral_Contrast computes spectral contrast, using the method defined in [11]. It represents the relative spectral distribution instead of the average spectral envelope.

2.2.5. Tonnetz. Tonnetz computes the tonal centroid features (or Tonnetz), following the method in [1] that detects changes in the harmonic content of musical audio signals.

2.3. Classification Learning Techniques. Classification learning algorithms aim to find an optimal classifier model for recognizing test samples of provided features and unknown labels. Several learning techniques fundamentally reveal a philosophical theory in modelling knowledge as a mathematical form. In order to cover the diversity in using different forms, a set of classification learning algorithms of various families are used. In particular, the selected classifiers in terms of the family include the following [12].

2.3.1. Bayes. Bayes is a direct approach that finds the best hypothesis by using Bayes' theorem as a probability theorem for building rule- or graph-based classification models. Two well-known methods are used, which are the Bayesian network (BN) and naive Bayes (NB) models.

2.3.2. Functions. In this family, the classifier builds a function (or hypothesis) of the input domain (i.e., features) and maps it into a range of outputs (i.e., labels) to form a function for classification. A set of models are used which are multilayer perceptron (MLP), SMO (sequential minimal optimization for SVM), logistic (L), support vector machine (SVM linear (S_L), SVM polynomial (S_P), and SVM radial (S_R)), and latent Dirichlet allocation (LDA).

2.3.3. Deep Neural Network (DNN). DNN is a framework of two phases, which are feature engineering and classification [13, 14]. The feature engineering process automatically extracts useful and nonlinear features from the raw data using convolutional and pooling layers by optimizing the weights W (or feature maps) between layers [15]. In the classification phase, the useful features are flattened as a vector to be fed into a fully connected ANN. In this work, the architecture of the DNN, as shown in Figure 3, receives the MFCC features of the input voice as one-dimensional (1D) data. These features are then fed into a convolutional layer that consists of three layers of 32, 48, and 120 neurons using ReLu as a nonlinear activation function. A pooling (or subsampling) layer follows the conventional layers using max function to reduce the size of resulted features. Finally, such features are flattened as the input vector to a fully connected ANN which is three dense layers of 128 neurons, 64 neurons using ReLu function, and 2 output neurons representing the gender of the input voice using softmax function (i.e., a normalized probability function). Two 1D DNN models are used which are the normalized deep convolutional neural network (DL_norm) and deep convolutional neural network (DL). The parameter settings of the DNN are 1000 epochs (or the number of iterations), 25% dropout (or regularization), Adam optimizer, and pooling and feature map size of 2×2.

2.3.4. Lazy. Lazy learners simply classify a new sample by estimating the vector similarity between the sample features and the vectors of samples in the training set and then assign the label of the most similar ones to that test sample. Lazy classifiers differ from other methods known as eager learners. Eager learners construct a machine learning model before the testing process as ready-to-use classifier models. The lazy learners used in this study are IBk and KStar (K*).

2.3.5. Meta. The idea is to learn an expert classifier of ensemble weak classifiers combined in a way to predict a label using averaging or voting methods. AdaBoost (Ada) and bagging (B) are well-known algorithms.

2.3.6. Trees. Each classifier is a form of a hierarchical tree where a node at each level represents the best attribute at that level, while the arcs represent the values of that attribute. The decision tree (J48) and random forest (RF) models are used.

2.3.7. Rules. Rules traverse each feature value and create a rule by finding the most frequent label. The criterion for selecting features depends on calculating the error rate of the rule. Three techniques are used, which are OneR (1R), Ridor (R), and rough set (RS) models.

2.4. Feature Selection Techniques. Building an optimal classifier model is affected by no-relevant features used for constructing such a model. These features drive the model to produce low accuracy for provided labels that leads to the underfitting or overfitting problem. Therefore, the necessity for selecting the relevant subset is needed. Three feature selection optimizers are used which are derived from the natural behaviour—evolutionary search, particle swarm optimization (PSO) search, and wolf search [16–18]. Each algorithm generates a set of individual solutions and then selects the optimal solution based on an evaluation metric and a learner optimizer (or evaluator). In this work, the evaluation metric used is area under the ROC curve (AUC) to validate whether a classifier can separate positive and negative samples and identify the best threshold for separation [19]. On the contrary, the RF classifier from the tree family is used as an evaluator to select the best subset features.

2.5. Evaluation. In this phase, particularly in the training phase, the 10-fold cross-validation method is used for each experiment by repeating it 10 times at each process of building a classifier. The evaluation metrics used are precision and recall [20]. Precision is the ratio of relevant samples to the retrieved ones, while recall is the ratio of retrieved and relevant samples to the total amount of relevant samples.

3. Experimental Results

A set of experiments are conducted for evaluating the contributions, which include studying the efficiency of extracted features, evaluating different learning techniques, and analyzing the three natural optimizers used for feature selection. This section also shows the datasets, experimental parameters, and settings and then presents the evaluation of the presented contributions.

3.1. Experimental Settings. A standard dataset of artificial voices from the study in [21] is used. The dataset consists of 20 languages. Each language has 16 voice samples of eight files for each gender. The artificial voice is a signal mathematically produced for regenerating the time and spectral characteristics of the human speech. These artificial voices have bandwidth between 100 Hz and 8 kHz, which significantly affects the performance of linear and nonlinear

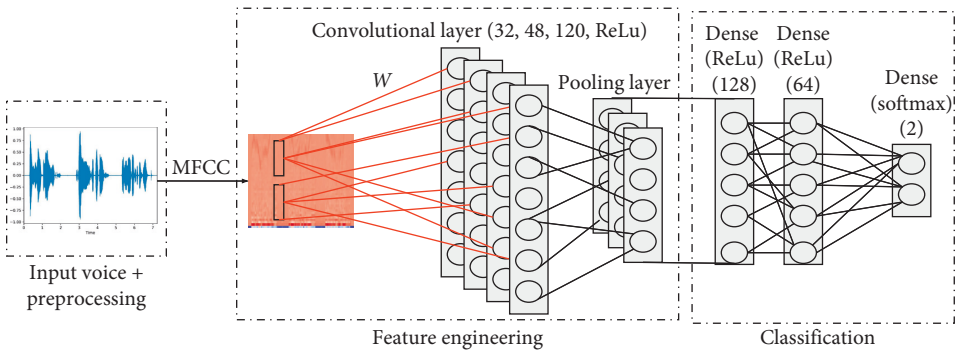


Figure 3: One-dimensional conventional neural network.

telecommunication systems. The artificial voice is mainly used for objective evaluation of speech processing systems and devices. A single channel with continuous activity (i.e., without pauses) is sufficient for measuring characteristics. The advantage of generating artificial voice is that it is more easily generated and has smaller variability than real voice.

The parameter settings of natural feature selection methods in Weka tool (https://www.cs.waikato.ac.nz/ml/weka/) are used as default settings. Although there are a set of natural methods, this study presents the most common ones, which are the evolutionary search, PSO search, and wolf search.

3.2. Voice Feature Effect and Correlation. In order to study what features are relevant to build an optimal classifier, the correlation between features has to be examined to demonstrate how they are related to each other. Four feature types, in the present work, are considered including MFCCs, Chroma, Mel, and Tonnetz. The correlation between features is presented in Figure 4 that shows a scattered plot representing the correlation relationship among different feature types. Each chart contains a linear regression equation that formulates the evaluated feature values. In addition, it clarifies the R^2 correlation coefficient. R^2 is a statistical measure that determines how close the real data points are fitted by the linear regression model. This means that if the R^2 value is close to 1, the data are highly fitted to the regression line, and there is no difference in their effects on tested labels, or there is, in contrast, a bad correlation to the labels.

In particular, as shown in Figure 4, the best R^2 of 0.332 is between the MFCC and Chroma features. In contrast, based on the chart, the worst correlation occurs between the Chroma and Contrast features with R^2 equal to 0.35. At each feature category, the MFCC feature has the best correlation with the Chroma feature with R^2 equal to 0.332 and worst correlation with the Tonnetz feature with R^2 equal to 0.0012. The Chroma feature has the worst correlation with the Mel feature with R^2 equal to 0.0004 compared to the other feature categories. The Mel feature has a worse correlation value in comparison with the Tonnetz feature, as shown in Figure 4(c), with R^2 equal to 0.0062. The Tonnetz features have the worst correlation value of $3.2e^{-6}$ compared to the Contrast feature category.

In summary, the MFCC, Chroma, and Mel features show an efficient performance as they are more related to each other. The reason is that these features extract high-energy coefficients from the signal where, in contrast, the other features concern about the tone of the musical signals. This answers the research question RQ1 that ensures the importance of selecting more suitable features for possibly building more accurate classifiers for voice paralinguistic information aspects.

3.3. Voice Gender Recognition and Classification. This work aims to build a classifier for recognizing the gender of a given human voice. The gender voices, whether they are male or female, are different from each other by the signal energy and tune. Therefore, it is necessary to construct a classifier model to differentiate the male human voice from female voice because it is important in many applications as discussed before.

In order to ensure diversity in constructing and using the classification models, different theoretical supervised classifiers are built, as shown in Figure 5. The figure shows a bar chart, where the x-axis represents the supervised learning methods and the y-axis represents the precision/recall evaluation metrics ranging from 0.86 to 1. As shown, the function family experimentally has overall superior performance results compared to the other families especially at the DL_norm and SMO techniques with precision/recall values of approximately 99.97% and 99.7%, respectively. However, in particular, the BN technique shows better performance than the NB technique in the Bayes family with a precision/recall value of approximately 10.2%. This means that the probabilities in the network graph of the BN method are more robust compared to the generated rules in the NB method. In the function families, the DL_norm and SMO techniques locally obtain significant performance as they have in general. The IBk technique gains leading performance compared to the K* method with 2.2% as the IBk method has a recall value of 99.1% and K* method has a recall value of 97%. The AdaBoost of the meta family gets high values of precision/recall compared to the B technique with percentage increasing up to 13.12%. In incremental rules family, the R technique gains consistent performance values compared to the 1R and RS methods. The methods in

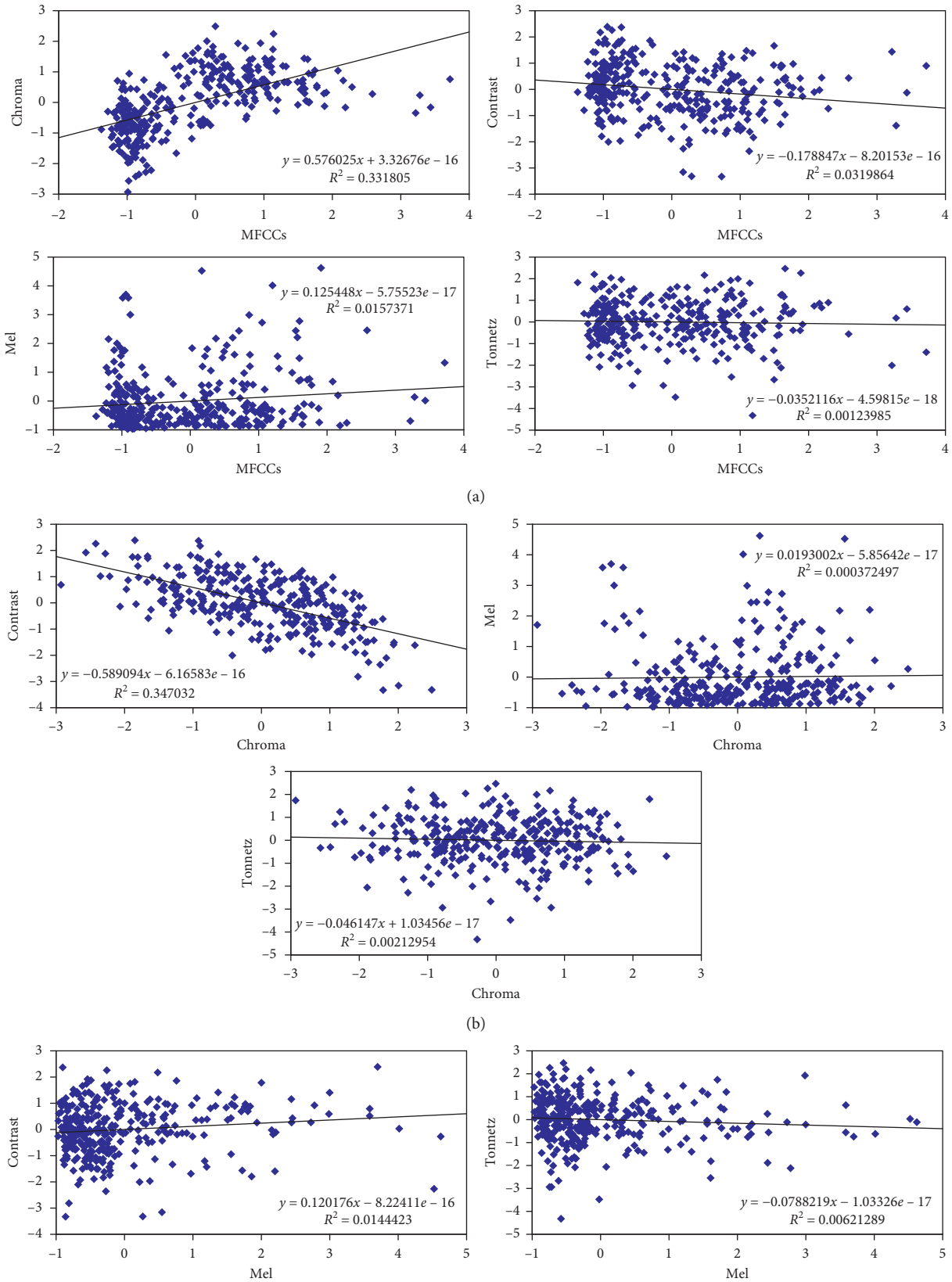
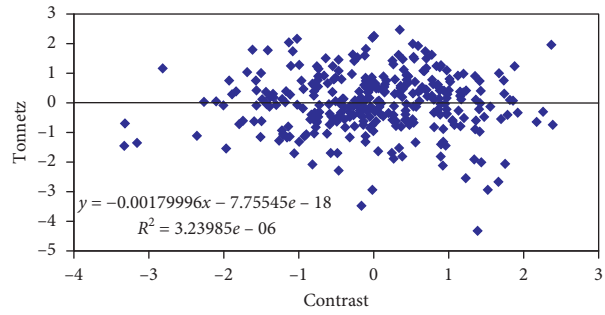


Figure 4: Continued.



(d)

Figure 4: Feature correlations and line equations: (a) MFCCs vs rest; (b) Chroma vs rest; (c) Mel vs rest; (d) Tonnetz vs rest.

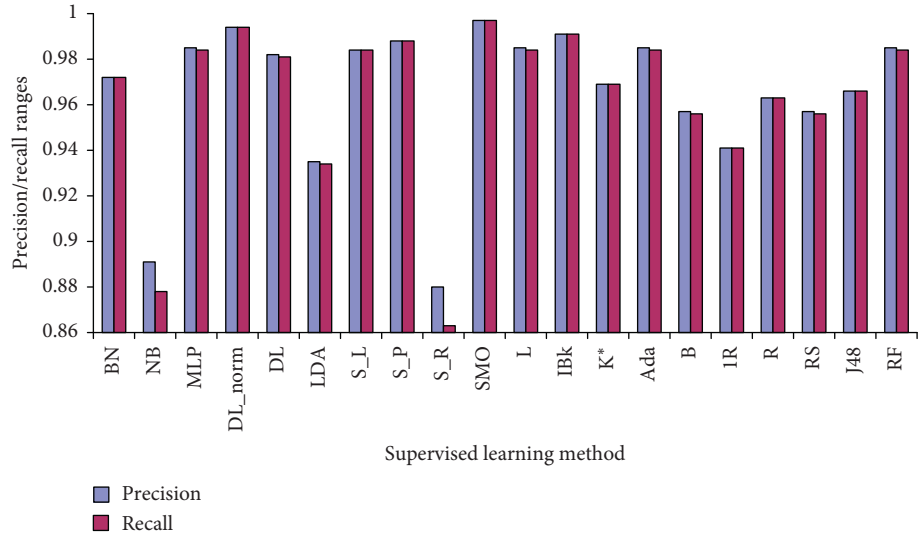


Figure 5: Precision/recall of classifier models.

the rules family show low performance compared to the other families, but the results are presented to explain such a conclusion. In the tree family, it is clear that the RF gains consistent results in comparison with the J48 technique with the performance value increasing up to 2.1%.

The DL_norm and SMO are used as significant techniques to evaluate their performance in comparison with the best techniques discussed in Introduction. The experiments were conducted for fairness on the same dataset. Figure 6 shows the accuracy values of the DL_norm and SMO techniques and the state-of-the-art methods. As shown, DL_norm and SMO still show better performance in comparison with the other state-of-the-art methods, which ensures the promising use of DL_norm and SMO techniques for gender recognition.

In brief, we discussed constructing classifier models for recognizing the voice gender by using various techniques of different theoretical families. The results showed that the function family gains consistent and significant performance values using the DL_norm and SMO techniques. This means that the theoretical methodology of building such models also has an effect on discriminating the human voice gender. Each ML technique algorithmically using good features could be a promising method for the voice gender recognition applications. Thus, these results lead us to answer the research question RQ2 affirmatively.

3.4. Feature Selection Techniques and Results. The performance of building classification techniques is affected by the quality of features in the training set. As such, it is necessary to select the most optimal features for enhancing the performance of voice gender recognition models. The three most common optimizers are exploited for feature selection inspired by natural behaviour, which are the EA, PSO, and wolf techniques as wrapper selection algorithms. These methods technically depend on searching a space of solutions and selecting the most optimal ones in an evaluated classifier such as the RF evaluator.

Five feature categories are discussed with a set of features for each category. There are 40, 12, 128, 7, and 6 subset features for MFCCs, Chroma, Mel, Contrast, and Tonnetz, respectively. However, Figure 7 shows a bar chart that explains the number of selected subfeatures at each category using the three selection methods. The x-axis represents the feature categories, while the y-axis represents the number of

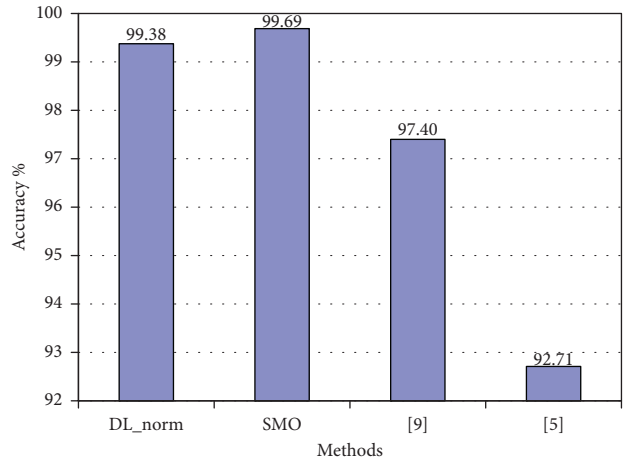


Figure 6: Comparison of best and state-of-the-art techniques.

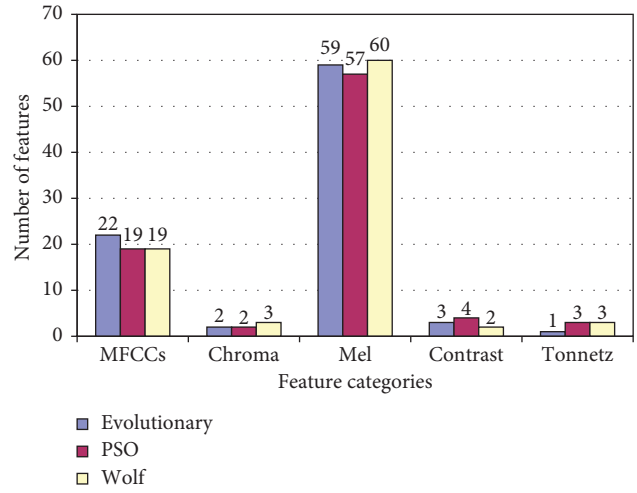


Figure 7: Number of features at each category.

selected subfeatures for the selection methods. In particular, the percentage of selected subfeatures using the EA, PSO, and wolf techniques is (55%, 47.5%, 47.5%), (16.7%, 16.7%, 25%), (41.1%, 44.5%, 46.88%), (42.86%, 57.1%, 28.6%), and (16.7%, 50%, 50%) on average of 34.5%, 43.2%, and 39.6% for MFCCs, Chroma, Mel, Contrast, and Tonnetz, respectively. The percentages show that the EA algorithm selects a small number of subfeatures at Chroma and Tonnetz categories. The PSO technique has a small number of subfeatures only at the Chroma category, while the wolf method selects a small number of subfeatures at the Chroma and Contrast categories. The question *how the effect would be if these subfeatures are used on ML techniques for voice gender recognition applications* needs to be answered.

In order to evaluate the selected subfeatures from the three feature selection techniques on the recognition of the human voice gender, similar experiments are conducted on the same families of ML techniques, as shown in Figure 8. The figure shows three bar charts of the three selection algorithms with x-axis and y-axis similar to those in Figure 5. The results manifest improvements in the performance of classifier models in overall families with approximately the same effect of three techniques on selecting the optimal features according to the performance of ML techniques. In particular, the best performance precision/recall values after subfeature selection for the evolutionary, PSO, and wolf methods are for the L method with 99.7%, the S_P method with 100%, and the RF method with 99.4%, respectively.

In summary, these results ensure that the EA selection algorithm gains a high evaluation performance with a 99.7% precision/recall value in the L method and also with a small number of subfeatures on percentage average of 34.5% compared to the PSO and wolf techniques. If there is tolerance in the percentage of selecting subfeatures, the PSO shows a superior result in classifying the human voice gender reaching 100% using the SVM_P ML technique. Hence, using natural feature selection algorithms is useful in enhancing the performance of ML techniques, resulting in a small number of relevant features. This accordingly answers the research question RQ3.

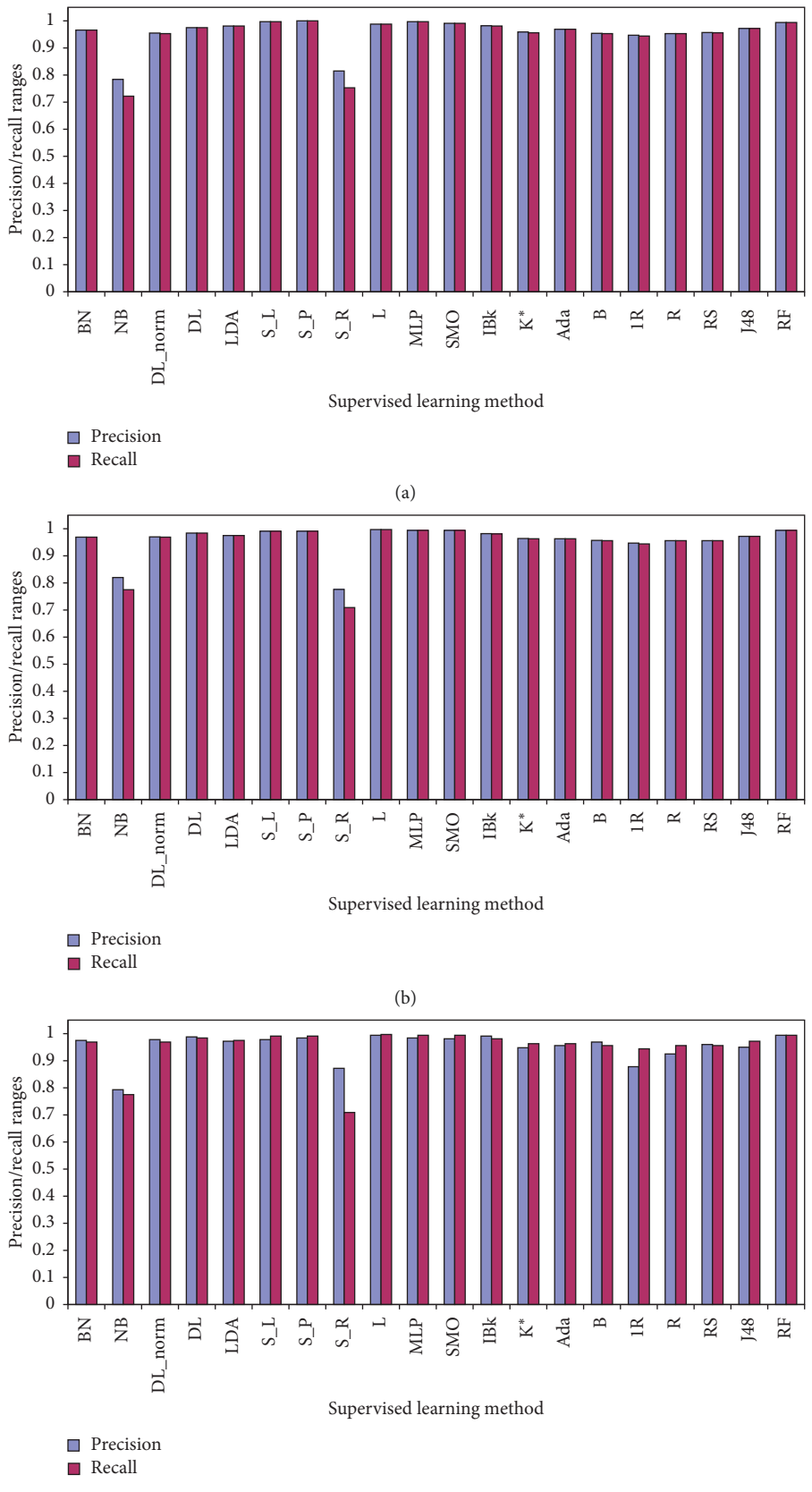


Figure 8: Swarm feature selection techniques: (a) PSO search; (b) evolutionary search; (c) wolf search.

4. Conclusion and Future Work

Recognizing the gender of human voice has been considered one of the challenging tasks because of its importance in various applications. The contributions are threefold including (i) studying the extracted features by examining the correlation between each other, (ii) building classification models using different ML techniques from distinct families, and (iii) evaluating the natural feature selection techniques in finding the optimal subset of relevant features on classification performance. In particular, three feature categories perform efficiently because of their theoretical methodology in extracting the relevant coefficient energies in the voice signal, which are the MFCCs, Chroma, and Mel, and this answers the research question RQ1. From the perspective of the performance of classifiers, the ML techniques behave in different ways. The results showed that the function family gained better performance compared to the other families. Although the function family has superior results, the other techniques have promising results, and this leads us to answer the research question RQ2. Finally, a set of experiments are conducted using three common feature selection techniques inspired by nature, which are EA, PSO, and wolf methods using the RF as an evaluator. These wrapper selection techniques select subfeatures from the feature categories which are on average approximately 39.1% on overall features. In spite of a small number of subfeatures, the performance of ML techniques was increased as there are some features that are not relevant in determining the gender of human voices. This also answers the research question RQ3.

In the future work, more experiments are being conducted to use many feature categories, ML techniques, and other natural feature selection techniques. Furthermore, the proposed techniques are being examined on different datasets since here only a standard artificial voice from the study in [21] is used. The reasons behind using it are that it contains many different languages (i.e., 20 languages), as well as the voice text is too long.

Data Availability

The standard dataset of artificial voices from the study in [21] used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

References

[1] Y. Hu, D. Wu, and A. Nucci, "Pitch-based gender identification with two-stage classification," *Security and Communication Networks*, vol. 5, no. 2, pp. 211–225, 2012.

[2] R. Djemili, H. Bourouba, and M. C. A. Korba, "A speech signal based gender identification system using four classifiers," in *Proceedings of the 2012 International Conference on Multimedia Computing and Systems*, pp. 184–187, Tangiers, Morocco, May 2012.

[3] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.

[4] E. Yücesoy and V. V. Nabiyev, "A new approach with score-level fusion for the classification of a speaker age and gender," *Computers & Electrical Engineering*, vol. 53, pp. 29–39, 2016.

[5] M.-W. Lee and K.-C. Kwak, "Performance comparison of gender and age group recognition for human-robot interaction," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 3, no. 12, 2012.

[6] Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, "Deep neural network framework and transformed MFCCs for speaker's age and gender classification," *Knowledge-Based Systems*, vol. 115, pp. 5–14, 2017.

[7] R. V. Sharan and T. J. Moir, "Robust acoustic event classification using deep neural networks," *Information Sciences*, vol. 396, pp. 24–32, 2017.

[8] J. G. Proakis and D. G. Manolakis, "Digital signal processing," in *Principles, Algorithms, and Applications*, Prentice-Hall, Upper Saddle River, NJ, USA, 3rd edition, 1996.

[9] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1097–1111, 2008.

[10] K.-I. Kanatani, "Fast fourier transform," in *Particle Characterization in Technology*, pp. 31–50, CRC Press, Boca Raton, FL, USA, 2018.

[11] W. Abdulla, N. Kasabov, and D.-N. Zealand, "Improving speech recognition performance through gender separation," *Changes*, vol. 9, p. 10, 2001.

[12] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011.

[13] W. Di, A. Bhardwaj, and J. Wei, *Deep Learning Essentials: Your Hands-On Guide to the Fundamentals of Deep Learning and Neural Network Modeling*, Packt Publishing, Birmingham, UK, 2018.

[14] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2018.

[15] S. Hinterstoisser, V. Lepetit, P. Wohlhart, and K. Konolige, "On pre-trained image features and synthetic images for deep learning," in *Proceedings of the European Conference on Computer Vision*, pp. 682–697, Springer, Munich, Germany, September 2018.

[16] Y. Chtioui, D. Bertrand, and D. Barba, "Feature selection by a genetic algorithm. Application to seed discrimination by artificial vision," *Journal of the Science of Food and Agriculture*, vol. 76, no. 1, pp. 77–86, 1998.

[17] E. Emary, H. M. Zawbaa, C. Grosan, and A. E. Hassenian, "Feature subset selection approach by gray-wolf optimization," in *Afro-European Conference for Industrial Advancement*, A. Abraham, P. Krömer, and V. Snasel, Eds., pp. 1–13, Springer International Publishing, Cham, Switzerland, 2015.

[18] B. Xue, M. Zhang, and W. N. Browne, "Particle Swarm Optimization for feature selection in classification: a multi-Objective approach," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1656–1671, 2013.

[19] L. B. Lusted, "Signal detectability and medical decision-making," *Science*, vol. 171, no. 3977, pp. 1217–1219, 1971.

[20] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*, pp. 233–240, ACM, New York, NY, USA, 2006.

[21] ITU-T Recommendation P.50, "Objective measuring apparatus," in *Proceedings of the International Telecommunication Union-Telecommunication Standardization Sector (ITU-T)*, Geneva, Switzerland, September 1999.