Journal of Telecommunication, Electronic and Computer Engineering
ISSN: 2180 – 1843 e-ISSN: 2289-8131 Vol. 15 No. 1
jtec.utem.edu.my

Wavelet-based Parametric Feature Subset Selection for Speaker and Accent Recognition using Genetic Algorithm

Rokiah Abdullah[1], Vikneswaran Vijean[1], Hariharan Muthusamy[2], Farah Nazlia Che Kassim[1], Zulkapli Abdullah[1], Mohammad Nazri Md Noor[1] and Jamaludin A. R. Rawi[3]
[1]*Faculty of Electronic Engineering Technology, Universiti Malaysia Perlis (UniMAP), Kampus Pauh Putra, 02600 Arau, Perlis, Malaysia*
[2]*Department of Electronic Engineering, National Institute of Technology, Srinagal (Garhwal), Uttarakhand, India*
[3]*Kolej Komuniti Bandar Darulaman, No. 17, Bandar Darulaman Jaya, 06000 Jitra, Kedah, Malaysia*
rokiah@unimap.edu.my

Article Info

Article history:
Received Dec 28th, 2022
Revised Feb 17th, 2023
Accepted Mar 2nd, 2023

Index Terms:
Feature selection
Genetic Algorithm
Speaker and accent recognition
Wavelet packet Transform

Abstract

Research on speaker and accent recognition studies using the Malay language in the field of Automatic Speech Recognition (ASR) is limited, with most studies focusing on speech recognition. This study proposes to increase the performance Malaysian speakers and accent recognition using wavelets transform, namely Wavelet Packet Transform (WPT) and Dual-Tree Complex Wavelet Packet Transform (DT-CWPT). A variety of feature extraction combinations, including conventional Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC) and wavelets transform, were implemented to compare the effectiveness of the proposed method. Although the proposed approach resulted in improved detection rate, it faced challenges in terms of high feature dimensionality and increased computation time. To address these issues, the Genetic Algorithm (GA) approach has been adopted to reduce the number of irrelevant features, accelerate the learning system and achieve better performance. The extracted features were trained using various classifiers, including k-Nearest Neighbors (k-NN), Support Vector Machine (SVM) and Extreme Learning Machine (ELM). The experimental results showed that the best speaker recognition accuracy was 97.33% for English numbers using SVM classifier and 96.02% for Malay words using the ELM classifier with a combination of wavelets, LPC and MFCC features. For accent recognition, the ELM classifier yielded the best performance, achieving 95.28% accuracy for English numbers with a combination of wavelets and MFCC features and 96.72% for Malay words using combined feature extraction of wavelets, LPC and MFCC feature extraction. It can be concluded that Malay words yielded better recognition rates than English numbers. Furthermore, use of GA effectively reduced the overall number of features while maintaining high accuracy level.

I. INTRODUCTION

Automatic Speech Recognition (ASR) is a process that stores converted voice signal or continuous audio speech data in computational format. This storage of different voices is important to represent good pronounciation examples. The important part of ASR, is feature extraction, which entails it extracting relevant signal data from a particular voice or speech. Various feature extraction techniques have been developed to achieve higher recognition accuracy. Conventionally, Fourier Transform has been widely used for ASR analysis. However, this method is less effective due to the lost of time information of the signal [1] and [2]. Therefore, wavelet-based approaches have been recognised as valuable tools for analyzing non-stationary signals in both time and frequency scale. Previous studies have explored various feature extractions techniques. In relation to this, Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC) and Linear Predictive Cepstral Coefficients (LPCC) are among the popular feature extraction techniques used and yield good results. Recently, an alternative approach employing wavelet transform has become increasingly popular and extensively employed in numerous studies [3]-[5]. With the rapid development of variety methods in extracting the features of ASR, the often-cited problems faced by the researchers are dimensionality issues [6], [7]. To address this issue, Genetic Algorithm (GA) has been identified as one of the feature subset selection methods that was found to be effective and extensively used in ASR. The following works will describe the features and machine learning techniques implemented in this work, focusing on the application of GA optimization.

In 2018, optimization technique using GA for speaker and speech recognition using MFCC as well as Deep Neural Network (DNN) was reported by Kaur et al. [8]. The performance of the proposed method was compared with the Perceptual Linear Prediction (PLP), Relative Spectral

Transform (RASTA) PLP and LPCC methods. Experiments were conducted in clean environment with added White Gaussian Noise (WGN). The results showed that MFCC outperforms other techniques in both clean or noisy environments. The combination of DNN with MFCC and GA yielded the best performance with an average accuracy of 96.51%, while 94.08% accuracy was obtained using the combination of MFCC and DNN without GA.

Kawase et al. [9] proposed a speech enhancement parameter in an ASR system, applying GA to enhance the speech recognition effectiveness and roboustness under various acoustic conditions. This work investigated noisy speech signal and used GA to improve recognition accuracy. The initial population in GA was generated both randomly and manually, reducing the minimum error rate by 27% and 7% respectively. The proposed method showed improvements in the parameter-set values.

In a study to reduce the size feature vectors in phonemes recognition of Arabic speech, Ibrahim et al. [10] proposed a feature selection method using GA. The Feed Forward Neural Network (FFNN) was incorporated with the proposed GA based feature selection method. The King Abdulaziz City for Science and Technology (KACST) Arabic Phonetic Database was used in their study. The phoneme recognition accuracy results were slightly higher than those obtained using the full-fledge features vector recognition. A 90% recognition accuracy was achieved with a 50% reduction in the dimension of the input vector.

He et al. [11] proposed a speech recognition system design using Field Programmable Gate Array (FPGA)-based embedded real-time English. This integrated software and hardware system collected English speech samples, processed the data and outputs control commands. This work employed the Genetic Algorithm Continuous Hidden Markov Model (GA_CHMM) to obtain better recognition accuracy. The performance of GA_CHMM was compared with the CHMM algorithm and Viterbi algorithm. English speech samples of isolated words from 10 individuals were collected and the vocabulary size selected for the test was 100. The recognition rate was greatly improved with over 90% after using English speech by employing GA_CHMM, CHMM and Viterbi algorithm. GA_CHMM results were the highest with over 93% followed by CHMM and Viterbi algorithm.

In a recent study, Albadr et al. [12] presented Optimized GA with Extreme Learning Machine (ELM) machine learning in Speech Emotion Recognition (SER) system. The features were extracted using conventional MFCC features. This study tested seven emotions consisting of neutral, happiness, boredom, anxiety, anger, sadness and disgust using the Berlin Emotional Speech (BES) dataset. Four different experiments have been performed namely Subject Dependent (SD), SI (Speaker Independent), Gender Dependent Female (GD-Female), and Gender Dependent Male (GD-Male). It was observed that the optimized GA-ELM resulted in the highest performance accuracy of 100% (SD), 93.26% (SI), 96.14% (GD-Male) and 97.10% (GD-female).

Although the results reported in the literature look promising, most of their research focus on speech, emotion speech and phoneme recognition. Besides, many studies employ Fourier transform techniques, such as MFCC and LPC features, which are less effective due to the lost of time information signal during the transformation from time to frequency, as mentioned earlier by [1], [2]. These studies also generally use GA without indicating the best and appropriate parameter settings. Thus, there is no guarantee of optimality of the obtained solution. In addition, lengthy computation times of running GA was mentioned during the training phase. Therefore, in this study, WPT, DT-CWPT, LPC and MFCC are employed using a new database extracted from original voice signals of Malaysian speakers. After consolidation, a larger number of features are obtained after the consolidation. The combination of these extracted features produces a larger feature set. GA is proposed to select optimal features, overcoming the curse of dimensionality and improving learning speed. The k-Nearest Neighbors (k-NN), Support Vector Machine (SVM) and Extreme Learning Machine (ELM) classifiers are used to measure the performance of the recognition rate.

II. METHOD

The speech corpus was extracted from 75 speakers (female and male) of different races, who were undergraduate students at the University Malaysia Perlis. The speakers pronunounced selected Malay words and English numbers (0 through 9) for 15 sessions, resulting in 23, 625 speech sample files. The predefined numbers and Malay words were arranged randomly in each session. The Malay words consist of vowels /a/, /i/, /o/, /u/ and /ɛ/ represent “front” and “back” vowels and one diphthong (au). These words are monosyllable and bi-syllable structures. Both structures were chosen to compare the effectiveness of each structure in the recognition rate. This study is an extension of a work proposed by [13] and details of the collected dataset are presented in Table 1.

Table 1
Collected dataset details

Item	Description
Speakers	75
Session /speaker	15 times
Age	19-24 years old
Utterances	1. Malay word: Aku, Basi, Cap, Jalan, Jam, Muka, Pas, Pulau, Rabu, Sen, Tol 2. English numbers (0 until 9)
Race	Malay, Chinese, Indian
Microphone	Stereo microphone
Audio file format	Wav
Sampling frequency	16kHz
Room type	Close room environment

The proposed block diagram for speaker and accent recognition is shown in Figure 1. The 44.1 kHz voice signal sampling frequency was down sampled to 16 kHz for speech processing purposes. Ai et al. [14] reported that most of the salient speech features were within 8Khz bandwidth, hence, the down sampled value was considered reasonable. Voice signals from the database were extracted using feature extraction algorithms, such as MFCC, LPC and several types of wavelets transform (DT-CWPT and WPT). By this stage, all of the necessary information to differentiate between speaker and accent has been retained. The combined features were then optimized using the GA to select the best optimal feature subset from its original features based on natural selection [15]. The selected optimized features were investigated using the k-NN, SVM and ELM classifiers.

Experiments were conducted by grouping the features into four different categories: combining features from wavelets only, wavelets and LPC, wavelets and MFCC, as well as wavelets with LPC and MFCC features. Four different experiments as listed in Table 2 were used to determine the accuracy of learned speeches.

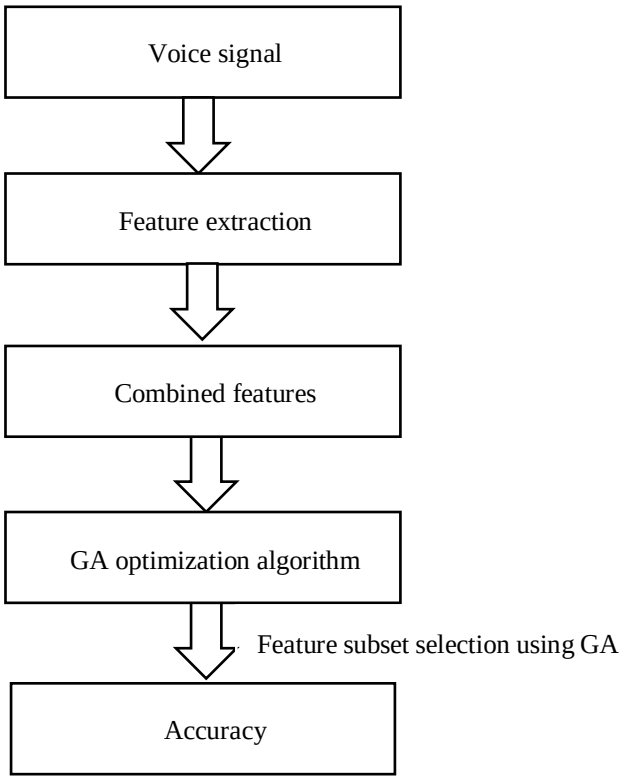

Figure 1. Block diagram of the proposed speaker and accent recognition

Table 2
Summary of experiments

Experiment	Feature extraction category	Classifier
1	DT-CWPT + WPT	k-NN, SVM and ELM
2	DT-CWPT + WPT + LPC	
3	DT-CWPT+WPT+ MFCC	
4	DT-CWPT + WPT + LPC + MFCC	

Many feature extraction methods were adopted in ASR. In this study, WPT, DT-CWPT, LPC and MFCC were employed and the explanation of these methods is described below.

A. Wavelet Packet Transform (WPT)

The WPT is an extension of the Discrete Wavelet Transform (DWT). The DWT analysis decomposed a signal into approximate coefficients and detail coefficients. A lower frequency band was used for further decomposition. In contrast, WPT decomposed both lower and higher frequency bands into two sub-bands, resulting in a balanced binary tree structure produced by a wavelet packet [16]. Therefore, this approach provides more information in terms of time-frequency resolution. Many studies on WPT were extensively reported by [17], [18] due to their high accuracy and encouraging performances. Fourth-order Daubechies wavelets were employed based on observations from previous studies as reported by [19], [20]. This particular wavelet family is best suited for the analysis of speech signals. Moreover, it had sharp filter transition bands, time-invariant and is computationally fast.

B. Dual-Tree Complex Wavelet Packet Transform (DT-CWPT)

The DT-CWPT technique contains two Discrete Wavelet Packet Transform (DWPTs) that operated in parallel. The voice signals were decomposed by DT-CWPT and divided into sub-bands, producing low and high frequencies. To construct DT-CWPT, both sub-bands were repeatedly decomposed using low and high pass perfect reconstruction (PR) filter banks (FB). The real part is the first FBs, while the imaginary part is the second FBs of the DT-CWPT. Since the DT-CWPT is made up of two FB wavelet packets operated in parallel, the filter for the second FB wavelet packet is identical to that of the first FB wavelet packet. Figure 2 shows the real part (first wavelet packet FB) at level five for DT-CWPT [21]. Replacing the first stage filters by $h_i^{(1)}(n)$ by $h_i^{(1)}(n-1)$ and replacing $h_i(n)$ by $h_i'(n)$ for i ∈ {0,1} will result in the second wavelet FB. In this study, 5 levels of DT-CWPT decomposition were employed, with non-linear entropy features extracted from each sub-band. The total number of feature vectors produced was 124, generated from both sub-bands of DT-CWPT.

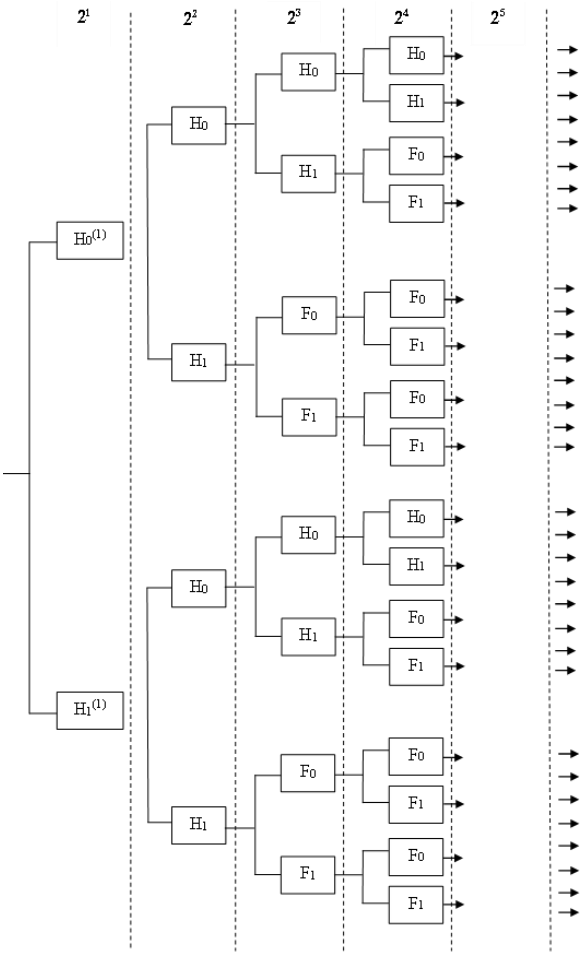

Figure 2. The forth-level of DT-CWPTs on the first wavelet packet FB

C. Linear Predictive Coding (LPC)

LPC methods are widely used in ASR. This method is based on a mathematical approximation of the vocal tract. The linear prediction is based on the concept that the present

sample was based on a linear combination of previous samples [22]. It can be described in Equation 1.

$$\hat{s}(n) = \sum_{k=1}^{p} a_k \; s(n-k) \quad (1)$$

According to Equation 1, ŝ(n) represents the estimated sample, a_k signifies the linear prediction coefficients, p denotes the order of the model, and s (n-k) refers to the previous speech sample. A suitable value for the p order was found between 8-20 in the recognition system, as reported by Prasana et al. [23]. However, Yusnita et al. [24] observed favorable result at 16 orders. Based on the experimental observation, the order of p was fixed as 16, as it generated a good performance.

D. Mel Frequency Cepstral Coefficients (MFCC)

The MFCC was introduced by Davis and Mermelstein based on the human auditory system. The Mel-frequency scale below 1 kHz is a linear frequency spacing and frequencies above 1 kHz is a logarithmic spacing. This can be expressed mathematically in Equation 2 as:

$$M(f) = 1125 * \log_e(1 + \frac{f}{700}) \quad (2)$$

In the MFCC approach, the voice signal undergoes several steps, including frame blocking, windowing, Fast Fourier Transform (FFT), Mel filter bank and computing Discrete Cosine Transform (DCT). First, the voice signal was segmented into small blocks for framing purposes. As reported Jain and Sharma [25], speech signals exhibit quasi-stationary over short periods(20-40ms). The short time frames of 32ms with a 50% overlap and a Hamming window were applied to the frames, making them more adaptable for spectral analysis and smoothing the signal. Then, FFT was applied, converting each sampling frame from the time domain into the frequency domain. The Mel scale was based on pitch perception. To obtain a smooth magnitude spectrum, a triangular-shaped filter was used. Finally, DCT was used to convert the log Mel-spectrum into the time domain, and MFCCs were generated after applying DCT. 13 MFCCs features were used to investigate the voice characteristic of the speaker.

E. Feature Subset Selection using Genetic Algorithm

In this study, GA was employed to select the best subset of eigen-features to classify the speaker and accent recognition. Feature subset selection is the process of choosing important/relevant features from the original data feature domain by removing redundant, noisy or irrelevant features without changing them. Therefore, the recognition rate increases when a small number of features, and the computation time decreases. According to a previous study reported by [8]-[12], GA is recognized as one of the best algorithms and feature selection methods due to its effectiveness and promising accuracy. Table 3 summarizes the parametric settings of GA employed for the speaker and accent recognition.

GA is a method for solving an optimization problem inspired by biological evolution based on natural selection. A basic form of GA relies on genetic operators: selection, crossover and mutation. The elements of GAs used in this study are population size, fitness function, selection method, crossover, mutation and elitism. The basics of GA started with the population, which is a subset of solutions to produce a new or current generation. Population size refers to the number of chromosomes (individuals) in the population. Previous studies have reported that a good population size is between 20–30, while others [26]-[28] suggest a size of 50–100. In this study, a population size of 100 was employed.

Selection is the process of selecting chromosomes from the population to contribute to the next generation. There are many methods used for selection, namely elitist, truncation, tournament, roulette wheel and more. For this study, tournament selection was employed. Tournament selection is a method that randomly selects an individual from a population, involving several "tournaments" among a few individuals. The winner from each tournament will be chosen for the crossover. This method is commonly used in studies as it can accommodate negative fitness values. Due to its efficiency, simplicity. and speed, as mentioned by [29]-[31], a tournament size of 2 was used. The crossover operator roughly mimics biological recombination, merging two individuals to form a crossover offspring for a new generation. The arithmetic crossover was applied, as this operator linearly combined the two parent chromosomes. Based on research by [32], [33] cross-values ranging from 0.5 to 1.0 were reported as optimal for GA employment. A mutation operator takes place after a crossover was performed. This operator was used for the exploration of search space to avoid local minima. The mutation probability (p_m) was set to a low value to prevent GA from being reduced to a random search when the probability was high. Thus, the mutation probability was set to 0.05 as proposed by the previous studies reported by [34]-[36].

Table 3
The Parameter Settings of GA

Parameters	Setting
Population size	100
Selection method	Tournament, size 2
Mutation type	Random, 0.05
Crossover type	Arithmetic, 0.05

F. Statistical Analysis

The features were exmined using one-way Analysis of Variance (ANOVA) and box plot analysis. ANOVA tests were performed before and after the application of GA for speaker and accent recognition, using the F score technique with a significance level of 0.05. Table 4 shows the p-values and F scores for accent recognition using four different feature sets. These features were derived from utterances using English numbers before and after the GA implementation.

Table 4
ANOVA test for accent recognition (English numbers)

No	Consolidation of features	Before GA		After GA	
		p-value	F score	p-value	F score
1	Wavelets only	p<0.05	28.97	p<0.05	86.01
2	Wavelets + LPC	p<0.05	5.33	p<0.05	73.15
3	Wavelets + MFCC	p<0.05	5.29	p<0.05	74.02
4	Wavelets+ LPC + MFCC	p<0.05	5.37	p<0.05	64.43

From Table 4, all the p-value are less than 0.05, leading to the rejection of the null hypothesis, which states no differences in means exist within the group. The analysis results showed that there were significant differences in the combined feature extraction methods proposed before and after the implementation of the GA.

In addition, the F-score value increased after the application of GA. This result suggests that the mean of the interclass variation between samples was greater compared to the intra-class variation within samples.

Furthermore, these observations are in line with the findings from the box plot analysis, which is further explained below. Figure 3 and Figure 4 present the box-plot analysis using Malay words for accent recognition. The angled lines represent the confidence interval around the median, while the centreline of the box in the plot indicates the median. The ends of the boxes inside the figure refer to Inter-Quartile Range (IQR). The '+' sign in the plot refers to the outliers, while the whiskers below and above the box indicate the extent of the remaining data. In this box plot analysis, the mean between groups before the GA showed similarities, whereas, there were significant differences after using the GA between groups of different accents. For accent recognition using Malays words as illustrated in Figure 3, the IQR for group 1,2 and 3 ranged from 2.9-3.0. In contrast, Figure 4 showed the IQR for groups 1, 2 and 3 was around 0.7- 0.8. Hence, it showed a significant difference before and after the GA was applied.

G. Classification

The recognition results were obtained using 10-fold cross-validation to ensure the reliability of the classifiers. The method offers the advantage of avoiding overlap between training and validation data. In this method, nine sets of data were used for training and one set was used for validation. The cross-validation process was executed ten times without data repetition. For speaker recognition, 75 classes were employed to represent the speech of 75 individuals, while three classes of accent recognition were used to represent Malay, Chinese and Indian accents, with 25 samples per class accent. The total number of samples for the speaker is 23625. Since 10-fold cross-validation was used, the number of data for one class per fold was calculated by dividing 23,625 by10 and then dividing the result by 75 subjects, yielding 32. Therefore, for one fold, there are 32 data points for one class. Meanwhile for accent recognition, the total samples (23,625) were divided by 10 and then divided by three accents, resulting in approximately 788. Hence, for one fold, there are approximately 788 data for one class. The k-NN, SVM and ELM supervised classifiers were used for the speaker and accent recognition. The fundamentals of these classifiers are described in the following paragraphs.

1) K-Nearest Neighbor

The k-Nearest Neighbor (k-NN) is a classification approach based on lazy learning [37]. It is a simple and supervised algorithm that classifies new instance queries based on the majority voting of k-Nearest Neighbors. A class label is assigned according to the majority count of k-nearest neighbors from the training dataset. To locate the nearest neighbors, the k-NN algorithm is implemented using Euclidean distance. Euclidean distance is defined in Equation 3 below.

$$d_E[x,y] = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \quad (3)$$

From this k-NN category, the class label of the test sample can be determined by applying majority voting among the k-nearest training speech samples. Generally, higher values of k will reduce the effect of noise on the classification but may lead to less distinct boundaries between classes [38].

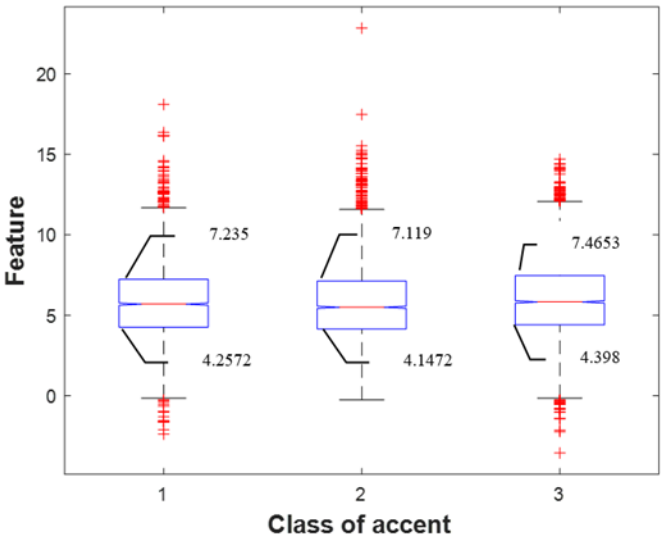

Figure 3. Boxplot before using GA for Wavelet and MFCC based features for accent recognition (Malay words)

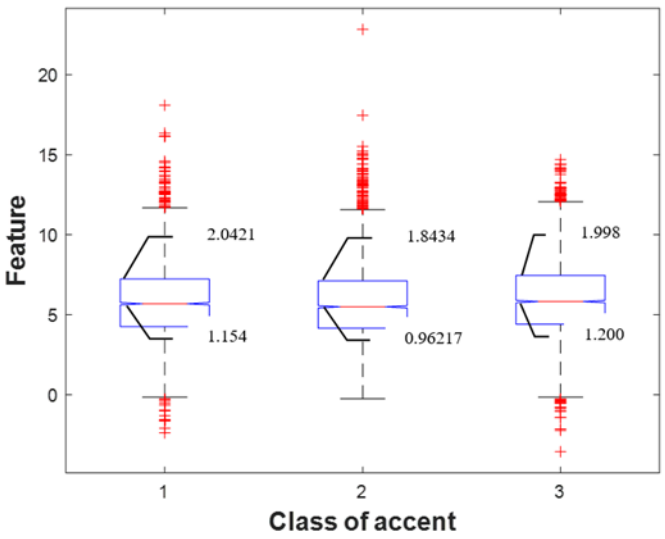

Figure 4. Boxplot after using GA for Wavelet and MFCC based feature for accent recognition (Malay words)

In this study, the value of k ranging between 1 and 10 was used and the best value of k was presented in Table 5 until Table 8.

2) Support Vector Machine

The Support Vector Machine (SVM) is a supervised algorithm designed to handle binary and multi-class classification problems. In cases of multi-class classification, multiple binary classifications are used, which can be either one-against-all or one-against-one. In this study, multi-class classification was employed since the data of this work contained non-linear samples. To distinguish between different classes of input data samples, SVM constructed a

hyperplane by searching for an optimal hyperplane. Since SVM employs linear algorithms on data in high dimensional space, it can achieve excellent performance [39]. The best parameters of the regularization parameter (Cost, C) and (gamma,γ) were optimized using the Lib SVM Tool [40]. Tables 5 through Table 8 show all parameters found empirically through simulation.

3) Extreme Learning Machine

Extreme Learning Machine (ELM) is a new method introduced by Huang et al. [41] and is characterized as a Single Hidden Layer Feed-Forward Networks (SLFNs). Unlike conventional neural networks such as ANN, the hidden nodes or neurons in the ELM algorithm do not require tuning. Instead, the weights of the hidden nodes are randomly assigned, and the output nodes are calculated using a simple inverse operation. Therefore, ELM provides an extremely fast and efficient learning speeds, as well as strong generalization capabilities compared to other machine learning. The regularization coefficients of the ELM classifier, ranging between -10 and 10 are depicted in Table 5 until Table 8.

III. RESULTS AND DISCUSSION

The performance of different classification schemes was performed with the consolidation of features extracted from non-linear features using wavelets transform, LPC and MFCC. The recognition results are presented in Table 5 to Table 8 using k-NN, SVM and ELM classifier. These tables show the accuracy of the speaker and accent recognition employing English numbers and Malay words before and after applying GA.

Table 5 and Table 6 show the results of speaker recognition with a combination of wavelets only, wavelets + LPC, wavelets + MFCC and wavelets + LPC + MFCC for both English numbers and Malay words. As presented in Table 5 and Table 6, the performance of the recognition rate using wavelets transform for speaker recognition of English numbers and Malay words is promising when using the ELM classifier. For English numbers, the accuracy of 95.26% before GA and 93.14% after GA was achieved using wavelets transform as shown in Table 5. Moreover, for Malay words in Table 6, the accuracy of 96.06% before GA and 93.43% after GA was achieved using wavelets transform. It also can be observed that more than half the number of features was reduced after applying GA, with the best accuracy (97.33%) achieved through the combination of wavelets, LPC and MFCC using English numbers. Meanwhile, the best accuracy of 96.02% was obtained after applying GA from the combination of wavelets, LPC and MFCC using Malay words, as shown in Table 6. It can be concluded that the adoption of GA for feature combination resulted in a reduction of computation time using three different classifiers with the SVM classifier's computation time shrinking by more than 50%.

Table 5
Recognition rate (%) with standard deviation (SD) and computation time using different combined features for the speaker (English numbers) before and after GA applied.

ACCURACY BEFORE GA			
No. of features	**Classifiers**		
	k-NN (k)	SVM (C,γ)	ELM (λ)
Wavelets (558)	81.25±0.17 time=330.24 (3)	92.29±0.10 time=4734.35 (0.062, 1024)	95.26±0.06 time=0.53 (5)
Wavelets + LPC (574)	82.19±0.17 time=329.05 (3)	92.75±0.14 time=4769.01 (0.062, 32)	95.66±0.08 time=0.52 (6)
Wavelets + MFCC (571)	83.79±0.17 time=337.16 (3)	95.01±0.07 time=4518.17 (0.031, 64)	97.04±0.10 time=0.58 (6)
Wavelets + LPC+ MFCC (587)	84.32±0.11 time=341.50 (3)	95.06±0.08 time=4533.30 (0.031, 64)	**97.18±0.07 time=0.53 (6)**
ACCURACY AFTER GA			
No. of features	**Classifiers**		
	k-NN (k)	SVM (C,γ)	ELM (λ)
Wavelets (153)	81.61±0.13 time=90.67 (3)	91.41±0.13 time=1736.02 (0.5, 1024)	93.14±0.07 time=0.38 (3)
Wavelets + LPC (148)	81.76±0.12 time=88.71 (3)	91.74±0.08 time=1693.17 (0.5, 32)	93.41±0.07 time=0.37 (3)
Wavelets + MFCC (148)	83.95±0.15 time=96.94 (3)	94.09±0.1 time=1653.52 (0.25, 32)	94.88±0.06 time=0.38 (3)
Wavelets + LPC+ MFCC (155)	84.24±0.12 time=92.17 (3)	**97.33±0.92 time=1508.33 (0.25, 32)**	95.27±0.07 time=0.40 (3)

Table 6
Recognition rate (%) with standard deviation (SD) and computation time using different combined features for the speaker (Malay words) before and after GA applied

ACCURACY BEFORE GA			
No. of features	**Classifiers**		
	k-NN (k)	SVM (C,γ)	ELM (λ)
Wavelets (558)	82.60±0.15 time=407.10 (3)	93.82±0.06 time=5025.38 (0.062, 32)	96.06±0.04 time=0.67 (6)
Wavelets + LPC (574)	83.49±0.14 time=408.08 (3)	94.05±0.07 time=5108.43 (0.062, 32)	96.34±0.06 time=0.62 (6)
Wavelets + MFCC (571)	85.33±0.10 time=403.39 (3)	95.82±0.06 time=4469.67 (0.015, 128)	97.26±0.05 time=0.63 (5)
Wavelets + LPC+ MFCC (587)	85.85±0.13 time=405.80 (3)	95.96±0.09 time=4546.84 (0.015, 128)	**97.65±0.05 time=0.61 (6)**
ACCURACY AFTER GA			
No. of features	**Classifiers**		
	k-NN (k)	SVM (C,γ)	ELM (λ)
Wavelets (130)	81.39±0.13 time=101.90 (3)	92.67±0.10 time=1604.50 (0.5, 32)	93.43±0.10 time=0.45 (3)
Wavelets + LPC (148)	81.75±0.13 time=88.22 (3)	91.75±0.09 time=1700.37 (0.5, 32)	93.41±0.08 time=0.38 (3)
Wavelets + MFCC (147)	83.61±0.15 time=105.70 (3)	93.71±0.09 time=1550.12 (0.25, 32)	94.76±0.08 time=0.44 (3)
Wavelets + LPC+ MFCC (146)	86.57±0.11 time=104.37 (3)	95.23±0.07 time=1487.67 (0.25, 32)	**96.02±0.08 time=0.45 (3)**

Table 7 and Table 8 present the results of accent recognition using English numbers and Malay words. From

Table 7
Recognition rate (%) with standard deviation (SD) and computation time using different combined features for the accent (English numbers) before and after GA applied

ACCURACY BEFORE GA			
No. of features	**Classifiers**		
	k-NN (k)	SVM (C,γ)	ELM (λ)
Wavelets (558)	90.11±0.12 time=318.56 (3)	93.55±0.14 time=8026.54 (2, 8)	95.61±0.07 time=0.51 (4)
Wavelets + LPC (574)	90.53±0.11 time=327.71 (3)	94.85±0.06 time=9263.43 (0.5, 8)	93.60±0.11 time=0.49 (3)
Wavelets + MFCC (571)	91.57±0.08 time=339.03 (3)	96.30±0.07 time=9239.27 (0.25, 16)	97.25±0.09 time=0.52 (4)
Wavelets + LPC+ MFCC (587)	91.83±0.09 time=347.26 (3)	96.37±0.08 time=9423.11 (0.25, 32)	**97.32±0.06 time=0.53 (4)**
ACCURACY AFTER GA			
No. of features	**Classifiers**		
	k-NN (k)	SVM (C,γ)	ELM (λ)
Wavelets (153)	89.66±0.12 time=89.67 (3)	93.09±0.10 time=5603.60 (2, 1024)	93.30±0.08 time=0.41 (2)
Wavelets + LPC (140)	89.05±0.11 time=81.95 (3)	93.54±0.12 time=3218.34 (2, 8)	93.43±0.08 time=0.35 (2)
Wavelets + MFCC (162)	91.35±0.09 time=103.56 (3)	94.92±0.07 time=3624.12 (2, 4)	**95.28±0.14 time=0.38 (2)**
Wavelets + LPC+ MFCC (149)	90.51±0.07 time=86.97 (3)	94.54±0.08 time=3457.27 (2, 16)	94.55±0.11 time=0.36 (2)

Table 8
Recognition rate (%) with standard deviation (SD) and computation time using different combined features for the accent (Malay words) before and after GA applied.

ACCURACY BEFORE GA			
No. of features	**Classifiers**		
	k-NN (k)	SVM (C,γ)	ELM (λ)
Wavelets (558)	90.97±0.10 time=383.52 (3)	95.54±0.08 time=9046.94 (0.5, 8)	96.52±0.09 time=0.57 (4)
Wavelets + LPC (574)	91.39±0.08 time=394.13 (3)	95.75±0.10 time=9103.65 (0.5, 8)	96.74±0.10 time=0.59 (4)
Wavelets + MFCC (571)	92.52±0.10 time=392.97 (3)	97.09±0.08 time=9147.73 (0.25, 32)	97.93±0.05 time=0.59 (4)
Wavelets + LPC+ MFCC (587)	92.87±0.11 time=419.96 (3)	97.19±0.07 time=9597.77 (0.25, 64)	**98.17±0.04 time=0.60 (5)**
ACCURACY AFTER GA			
No. of features	**Classifiers**		
	k-NN (k)	SVM (C,γ)	ELM (λ)
Wavelets (153)	91.55±0.12 time=111.99 (3)	95.10±0.13 time=3867.42 (2,8)	94.61±0.08 time=0.49 (3)
Wavelets + LPC (140)	91.30±0.10 time=118.01 (3)	95.01±0.08 time=4145.98 (2,16)	95.05±0.10 time=0.42 (2)
Wavelets + MFCC (162)	91.44±0.10 time=110.75 (3)	95.66±0.08 time=3992.25 (2,8)	95.73±0.10 time=0.42 (2)
Wavelets + LPC+ MFCC (149)	93.50±0.07 time=122.59 (3)	96.43±0.08 time=4429.35 (2,1024)	**96.72±0.07 time=0.43 (2)**

Table 7, the performance of three different classifiers showed that the recognition accuracy was above 89%. The best accuracy for accent recognition after applying GA using English numbers was 95.28%, which is achieved through the combined features of wavelets and MFCC using an ELM classifier. The k-NN classifier gave the lowest recognition accuracy at 89.05%. Table 8 presents the result of accent recognition using Malay words. As shown in Table 8, before applying GA, the best recognition rate of 98.17% was obtained from the combination of wavelets, LPC and MFCC and after applying GA, a 96.72% recognition rate was achieved using the same combination. It can be observed that almost all of these experiments achieved a feature reduction of over 70% from the original feature, along with a reduction in computation time.

The recognition results shown in Table 5 to Table 8, demonstrate that the performance using wavelets transform (WPT and DT-CWPT) features is promising. The wavelet features derived from WPT and DT-CWPT provide good accuracy due to the wavelets transform's ability to perform time-frequency analysis. The WPT algorithm offers more information on time-frequency resolution, while the DT-CWPT algorithm comprised of a real and imaginary tree structure (lowest and highest frequency sub-band), resulting in more efficient, detailed and appropriate information features. The implementation of GA successfully reduces a large number of features and speed up the computation time for the combince features employed. GA significantly reduces computation time while maintaining the recognition rate with only slight differences within 2%. GA's effectiveness is due to its ability to remove irrelevant and redundant features from the complete feature set, thereby, mitigating the impact of dimensionality, enhancing generalization capability, speeding up the learning process and improving model interpretability. Therefore, the performance of the learning models improves.

Based on the results depicted in Table 5 to Table 8, ELM classifier performs best , followed by SVM and k-NN classifiers. However, the results of SVM and ELM classifiers are comparable. The ELM algorithm' superior performance is attributed to its learning efficiency, fast learning speed and universal approximation capability. The SVM classifier achieves comparable results with ELM in the experiments due to its kernel function algorithm, which maps the training samples of classes into higher dimensional space.

The results indicate that the combination features using Wavelets, MFCC and LPC achieves the best recognition compared to other approaches. This is due to the time-frequency analysis capability of wavelets transform, the high accuracy and low complexity of MFCC features, and the reliability, accuracy and the robustness of LPC features, which contribute to improved recognition results. Therefore, the combined features of wavelets transform with MFCC and LPC features results in a better recognition rate. Most current ASR research using Malay words focuses solely on speech or accent recognition. For example, K. Anggraini et al. [42], examined speech recognition for English sentences with Riau Malay accent. They used Google Recognizer to isolate words from sentences, MFCC for feature extraction and the Hidden Markov Model (HMM) for classification. The method achieved 94.02% accuracy. In comparison, our method is able to predict accent and speaker identification with an accuracy of up to 95.28%.

S. Darshana et al. [43] investigated eight different accents of English using MFCC and Mel-Spectrogram features. They investigated a novel, well-structured database, which contains speech samples from six different states of non-native Indian English speaker accents to address the unbalanced dataset and speaker mismatch issues. The performance of the proposed models was evaluated on the novel database using metrics such as precision, accuracy, F1-score, and recall. Accent classification performance was tested using three pre-trained models: ResNet18, ResNet50, xResNet18. The xResNet18 achieved the best results, with an average of 100% accuracy, precision, recall, and F1-score. Trained on Mel-Spectrogram with zero padding at the end, the model shows its effectiveness in predicting the English accents of non-native speakers. However, its performance in terms of malay accent and speaker detection is not available.

In contrast, this work focuses on non-linear features using wavelets transform WPT and DT-CWPT, LPC and MFCC for speaker and accent recognition. A direct comparison with with the previous works that focus on speech and accent recognition in feasible since this work examines speaker and accent recognition.

IV. CONCLUSION

This study reports an experimental study aimed at improving the recognition of Malaysian speakers and accents using a combination of Wavelets, LPC and MFCC. To reduce the large number of features, a feature optimization technique, Genetic Algorithm (GA), was adopted. The combined features were then classified using three different classifiers: k-NN, SVM and ELM. The findings demonstrate that the GA-based feature subset selection has successfully produced a smaller subset of features from the combined features, yielding favorable speaker and accent recognition results. The highest recognition accuracy, 97.33%, was obtained using a combination of wavelets, LPC and MFCC features. For speaker recognition with Malay words, a 96.02% accuracy was achieved using the same combination. In terms of accent recognition, the combination of wavelet and MFCC produced the highest recognition of 95.28% for English numbers and 96.72% for Malay words. It is worth noting that some recognitions accuracies slightly decreased after adopting GA, with differences ranging from 45% to 2.04%, which is still acceptable. Furthermore, GA application led to a significant reduction (almost 50%) in computing time, particularly for the SVM classifier. It was observed that the combination of wavelets, LPC and MFCC features yielded the best recognition performance (the highest performance). Among the three classifiers used, the ELM algorithm outperformed both SVM and k-NN classifiers. Recognition rates using Malay words were found to be superior to those using English numbers. There is still room for improvement, particularly the use of other Malay words to evaluate speaker and accent recognition. In the future, we intend to extend this study using a larger and more diverse database and explore other feature optimization techniques, such as PSO, PCA and others.

ACKNOWLEDGMENT

The authors wish to acknowledge the encouragement and express gratitude to Dr Noriha Basir, Senior Lecturer of Centre for Liberal Sciences, Faculty of Applied and Human Sciences, Universiti Malaysia Perlis (UniMAP) as language consultant and expertise in suggesting the wordlist.

REFERENCES

[1] M. Hariharan, K. Polat and S. Yaacob, "A new feature constituting approach to detection of vocal fold pathology", International Journal of Systems Science, vol 45(8), 2014, doi: 10.1080/00207721.2013.794905.

[2] R Rasnayake, M.W.P Maduranga and J.P.D.M. Sithara, "Surface Electromyography signal acquisition and classification using Artificial Neural Networks (ANN)", International Journal Modern Education and Computer Science, vol 3, pp.64-75, 2022, doi: 10.5815/ijmecs.2022.03.04

[3] H.Ali, AFA Zaidi, WKW Ahmad, MSZ Azalan, TST Amran, M. R Ahmad and M Elshaikh. "A cascade hyperbolic recognition of buried objects using hybrid feature extraction in ground penetrating radar images," Journal of Physics: Conference Science vol 1997, 2021, doi: 10.1088/1742-6596/1997/1/012018.

[4] F He and Q Ye, " A Bearing fault diagnosis method based on Wavelet Packet Transform and Convolutional Neural Network optimized by simulated Annealing Algorithm", Sensors, vol 22 (4), 2022, doi:10.3390/s22041410.

[5] A Kamra, K Singh and SS Dhaliwal, " Speech Signal Analysis using Wavelet Domain", nternational Journal for Research in Applied Science & Engineering Technology (IJRASET), vol 8 (X), pp. 83-88, 2020, doi: 10.22214/ijraset.2020.31806.

[6] H Muthusamy, K Polat and S Yaacob, " Improved emotion recognition using Gaussian Mixture Model and Extreme Learning Machine in speech and glottal signals", Mathematical Problems in Engineering, vol 2015, 2015, doi: 10.1155/2015/394083.

[7] L. Lei and S. Kun, "Speaker recognition using Wavelet Packet Entropy, I-Vector, and Cosine Distance Scoring," Journal of Electrical and Computer Engineering, 2017, doi: 10.1155/2017/1735698.

[8] G. Kaur, M. Srivastava and A. Kumar, "Genetic Algorithm for combined speaker and speech recognition using Deep Neural Networks," Journal of Telecommunications and Information Technology, pp. 23-31, 2018, doi:10.26636/jtit.2018.119617.

[9] T. Kawase, M. Okamoto, T. Fukutomi, Y. Takahashi, R. Masuda and T. Ootake, "Self-adjustable speech enhancement and recognition system," International Conference on Consumer Electronics (ICCE). pp. 1-2, 2019, doi:10.1109/ICCE.2019.8661925.

[10] A. Ibrahim, Y. Mohammad Seddiq, A. Hamid Meftah, M. Alghamdi, S. Ahmed Selouani, M. A. Qamhan, "Optimizing Arabic speech distinctive phonetic features and phoneme recognition using Genetic Algorithm," IEEE Access vol.8, pp. 200395-200411, 2020, doi:10.1109/ACCESS.2020.3034762.

[11] L. He, G. Jin and S. Bing Tsai, "Design and implementation of embedded real-time English speech recognition system based on big data analysis," Mathematical Problems in Engineering, vol 2021, Sept 2021, doi:10.1155/2021/6561730.

[12] M. Abbass Abood Albadr, S. Tiun, M. Ayob, F. Taha Al-Dhief, K. Omar and M. Khaled Maen, "Speech emotion recognition using optimized Genetic Algorithm-Extreme Learning Machine", Multimedia Tools and Applications, pp. 1-27, 2022.

[13] R. Abdullah, H. Muthusamy, V. Vijean, Z. Abdullah and F. Nazlia Che Kassim, "Real and Complex Wavelet Transform approaches for Malaysian speaker and accent recognition," Pertanika Journal of Science & Technology, 27(2), pp. 737-752, 2019.

[14] O. Chia Ai, M. Hariharan, S. Yaacob and L. Sin Chee, "Classification of speech dysfluencies with MFCC and LPCC features," Expert Systems with Applications, vol 39(2), pp. 2157-2165, 2012, doi:10.1016/j.eswa.2011.07.065.

[15] RL. Haupt and SE. Haupt, " *Practical Genetic Algorithms*", 2nd Ed John Wiley & Sons, 2004.

[16] N. Aida Amira Johari, M. Hariharan, A. Saidatul and S. Yaacob, "Multistyle classification of speech under stress using Wavelet Packet Energy and entropy features," IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (STUDENT), pp.74-78,2011, doi:10.1109/STUDENT.2011.6089328.

[17] S. Z. Bong, K. Wong, M. Murugappan, N. Mohamed Ibrahim, Y. Rajamanickam and K. Mohamad, "Implementation of Wavelet Packet Transform and non-linear analysis for emotion classification in stroke patient using brain signals," Biomedical signal processing and control 36, pp. 102-112, 2017, doi:10.1016/j.bspc.2017.03.016.

[18] M. Hariharan, R. Sindhu, V. Vijean, H. Yazid, T. Nadarajaw, S. Yaacob and K. Polat, "Improved binary dragonfly optimization

algorithm and Wavelet Packet based non-linear features for infant cry classification," Computer Methods and Programs in Biomedicine,155, pp. 39-51, 2018,
doi: 10.1016/j.cmpb.2017.11.021.

[19] R. Abdullah, V. Vijean, H. Muthusamy, F. Nazlia Che Kassim and Z. Abdullah, "Real and Complex Wavelet Transform using Singular Value Decomposition for Malaysian speaker and accent recognition," Advances in Mechatronics, Manufacturing, and Mechanical Engineering, pp. 22-35, Springer, 2021,
doi:10.1007/978-981-15-7309-5_3.

[20] L. Lei and S. Kun, "Speaker recognition using Wavelet Packet Entropy, I-Vector, and Cosine Distance Scoring," Journal of Electrical and Computer Engineering, 2017,
doi: 10.1155/2017/1735698.

[21] F. Nazlia Che Kassim, H. Muthusamy, V. Vijean, Z. Abdullah and R. Abdullah, "Dual-Tree Complex Wavelet Packet Transform for voice pathology analysis," Pertanika Journal of Science & Technology, 28(3), pp. 839-858, 2020.

[22] M.P. Paulraj, S. Yaacob and S. A Mohd Yusof, "Vowel recognition based on frequency ranges determined by bandwidth approach", International Conference on Audio, Language and Image Processing (ICALIP), pp.75-79, 2008,
doi: 10.1109/ICALIP.2008.4590133

[23] S. R. Mahadeva Prasanna, C. S. Gupta and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," Speech Communication, 48(10), pp. 1243-1261, 2006, doi: 10.1016/j.specom.2006.06.002.

[24] M. A Yusnita, M. P Paulraj, S. Yaacob and A. B. Shariman, "Classification of speaker accent using hybrid DWT-LPC features and K-nearest neighbors in ethnically diverse Malaysian English," 2012 International Symposium on Computer Applications and Industrial Electronics (ISCAIE), pp. 179-184, 2012,
doi:10.1109/ISCAIE.2012.6482092.

[25] A. Jain and O. P. Sharma, "A Vector Quantization approach for voice recognition using Mel Frequency Cepstral Coefficient (MFCC): A Review 1," International Journal of Electronics & Communication Technology (IJECT),vol 4(4), pp. 26-29, 2013.

[26] M. Inal, "Feature extraction of speech signal by Genetic Algorithms-simulated annealing and comparison with Linear Predictive Coding based methods," International Conference on Adaptive and Natural Computing Algorithms, pp. 266-275, 2007, doi:10.1007/978-3-540-71618-1_30.

[27] T. Chen, K. Tang, G. Chen and X. Yao, "A large population size can be unhelpful in evolutionary algorithms," Theoretical Computer Science, 436, pp. 54-70, 2012, doi:10.1016/j.tcs.2011.02.016.

[28] O. Roeva, S. Fidanova and M. Paprzycki, "Population size influence on the genetic and ant algorithms performance in case of cultivation process modeling," Recent advances in computational optimization, pp.107-120, 2015, doi:10.1007/978-3-319-12631-9_7.

[29] A. E. Eiben and J. E. Smith, "Introduction to evolutionary computing," *Natural Computing Series*. Second Edition. Springer, 2003, pp.1-287.

[30] B. Baudry, F. Fluerey, J. M. Jezequel, Y. L. Traon, "Automatic test case optimization: A bacteriologic algorithm," IEEE software, vol 22(2), pp.76-82, 2005, doi:10.1109/MS.2005.30.

[31] P.Civicioglu, "Transforming geocentric cartesian coordinates to geodetic coordinates by using differential search algorithm," Computers & Geosciences, 46, pp. 229-247, 2012.
doi:10.1016/j.cageo.2011.12.011.

[32] M. Srivinas and L. M. Patnaik, "Adaptive probabilities of crossover and mutation in Genetic Algorithms," IEEE Transactions on Systems, Man, and Cybernetics, 24(4), pp. 656-667, 1994, doi:10.1109/21.286385.

[33] W. Y. Lin W. Y Lee and T. P. Hong, "Adapting crossover and mutation rates in Genetic Algorithms," Journal of Information Science and Engineering, vol 19(5), 889-903, 2003.

[34] M. Hariharan, L. Sin Chee, O. Chia Ai and S. Yaacob, "Classification of speech dysfluencies using LPC based parameterization techniques," Journal of medical systems, 36(3), pp. 1821-1830, 2012.

[35] H. Thanh Le, L.Van Tran, X. Hoai Nguyen and T. Hien Nguyen," Optimizing Genetic Algorithm in feature selection for named entity recognition," Proceeding of the Sixth International Symposium on Information and Communication Technology (SoICT),pp. 11-16, 2015, doi: 10.1145/2833258.2833262.

[36] H. Saleem Ibrahim Harba and E. Saleem Ebraham Harba, "Voice recognition with Genetic Algorithms two modules crossover and mutation," International Journal of Modern Trends in Engineering and Research (IJMTER), vol 2 (12), pp. 144–155, 2015.

[37] C. Liang Liu, C. Hoang Lee and P. Min Lin, "A fall detection system using k-Nearest Neighbor classifier," Expert systems with applications, vol 37(10), pp. 7174-7181, 2010,
doi:10.1016/j.eswa.2010.04.014.

[38] R. Amami, D. Ben Ayed and N. Ellouze, "Practical selection of SVM supervised parameters with different feature representations for vowel recognition," International Journal of Digital Content Technology and its Application (IJDCTA), vol 7(9), pp. 418-424, 2015, doi:10.48550/arXiv.1507.06020.

[39] S. Sangeetha and N. Radha, "A new framework for IRIS and fingerprint recognition using SVM classification and Extreme Learning Machine based on score level fusion," 2013 7th International Conference on Intelligent Systems and Control (ISCO), pp. 183-188, 2013, doi: 10.1109/ISCO.2013.6481145.

[40] C. Chung Chang and C. Jen Lin, "LIBSVM: a library for support vector machines," ACM transactions on intelligent systems and technology (TIST), vol *2*(3), pp. 1-27, 2011,
doi: 10.1145/1961189.1961199.

[41] G. Bin Huang, H. Zhou, X. Ding and R. Zhang, "Extreme Learning Machine for regression and multiclass classification," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol 42(2), pp.513-529, 2011,
doi:10.1109/TSMCB.2011.2168604.

[42] K. Anggraini, L. L. Van, and Y. Darmayunata , "Speech recognition for English sentences with Malay accent", Jurnal Teknologi Informasi dan Komunikasi, vol 13(2), 2022,
doi: 10.31849/digitalzone.v13i2.10759.

[43] S. Darshana, H. Theivaprakasham, G. J. Lal, B. Premjith,V. Sowmya and K.P Soman, "A Hybrid Deep CNN-based Multi-accent recognition system for English language", International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR), 2022.
doi: 10.1109/ICAITPR51569.2022.9844177.