IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 11, No. 1, March 2022, pp. 161~172
ISSN: 2252-8938, DOI: 10.11591/ijai.v11.i1.pp161-172

Artificial speech detection using image-based features and random forest classifier

Choon Beng Tan[1], Mohd Hanafi Ahmad Hijazi[1], Frazier Kok[2], Mohd Saberi Mohamad[3], Puteri Nor Ellyza Nohuddin[4]
[1]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Kinabalu, Malaysia
[2]Bayurini Sdn Bhd, Penampang, Kinabalu, Malaysia
[3]College of Medicine and Health Sciences, United Arab Emirates University, Abu Dhabi, United Arab Emirates
[4]Institute of IR4.0, Universiti Kebangsaan Malaysia, Bangi, Malaysia

Article Info

Article history:

Received Jul 15, 2021
Revised Nov 26, 2021
Accepted Dec 11, 2021

Keywords:

Anti-spoofing voice recognition
Artificial speech detection
Speaker recognition
Speaker verification
Voice presentation attack detection

ABSTRACT

The ASVspoof 2015 Challenge was one of the efforts of the research community in the field of speech processing to foster the development of generalized countermeasures against spoofing attacks. However, most countermeasures submitted to the ASVspoof 2015 Challenge failed to detect the S10 attack effectively, the only attack that was generated using the waveform concatenation approach. Hence, more informative features are needed to detect previously unseen spoofing attacks. This paper presents an approach that uses data transformation techniques to engineer image-based features together with random forest classifier to detect artificial speech. The objectives are two-fold: (i) to extract image-based features from the mel-frequency cepstral coefficients representation of the speech signal and (ii) to compare the performance of using the extracted features and Random Forest to determine the authenticity of voices with the existing approaches. An audio-to-image transformation technique was used to engineer new features in classifying genuine and spoof voices. An experiment was conducted to find the appropriate combination of the engineered features and classifier. Experimental results showed that the proposed approach was able to detect speech synthesis and voice conversion attacks effectively, with an equal error rate of 0.10% and accuracy of 99.93%.

Corresponding Author:

Mohd Hanafi Ahmad Hijazi
Faculty of Computing and Informatics, Universiti Malaysia Sabah
Jalan UMS, Sabah, Malaysia
Email: hanafi@ums.edu.my

1. INTRODUCTION

Voice recognition, often known as speaker recognition, is the act of identifying and verifying a speaking human. It is divided into two categories: speaker identification and speaker verification. Speaker identification is the process of determining a speaking individual's identity, whereas speaker verification is the act of verifying that individual's claimed identity. Figure 1 depicts the distinction between speaker identification and speaker verification. In recognizing and validating the identity of a person from voice, speaker recognition employs both physiological and behavioral components.

Automatic speaker verification (ASV) is the process of verifying the claimed identity of a speaking individual automatically. In most ASV systems, the speaker enrolment phase and the speaker verification phase are the two key phases. During speaker enrolment, the ASV system captures the speaker's voice and extracts attributes that are utilized to create a speaker model of the speaking individual. The speaker model is then registered with the ASV system. During speaker verification, the speaking individual's voice is recorded to create a speaker model for verification. After that, the speaker model is compared to the claimed identity's speaker model in the ASV system. Finally, the matching will generate a score, with the claim being accepted if the score is equal to or greater than the ASV system's threshold. Otherwise, the claim will be turned down. Numerous types of features have been deployed for ASV systems. Gaussian mixture models (GMM) were extensively used in the past for feature extraction to produce robust ASV systems. In speaker verification, the universal background model (UBM) is a speaker model that represents broad attributes and characteristics that can be compared to the specific person being verified [1]. Later, i-vector and x-vector [2] based ASV systems were introduced to replace the gaussian mixture model-universal background model (GMM-UBM) based ASV systems. Deep learning approaches [3] such as recurrent neural network (RNN) [4] as a backend classifier was shown the capability in speaker verification with a low equal error rate (EER).

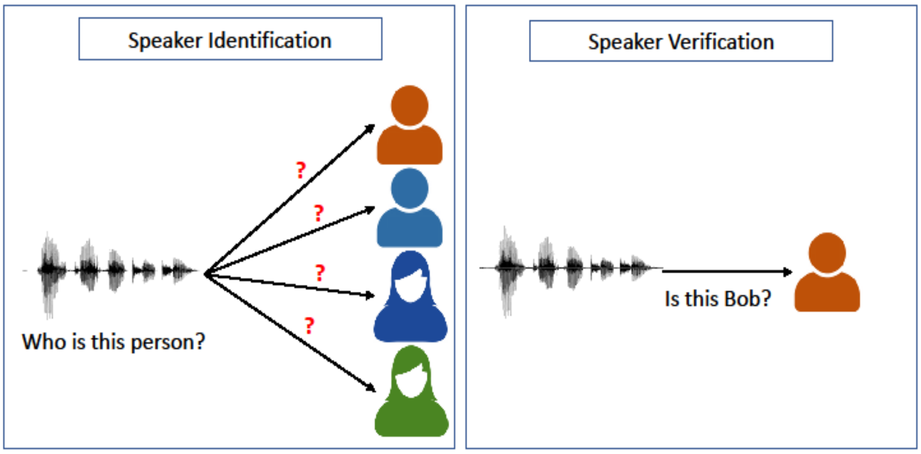

Figure 1. The illustration of speaker identification versus speaker verification

To mitigate the issue of security threats to the ASV systems, voice presentation attack detection (PAD) was introduced. Voice PAD can be categorized into two major types, namely artificial and replayed speech detection. The artificial speech was generated by speech synthesis and voice conversion, whereas replayed speech was generated by replaying the recordings of human speech. Several efforts can be seen to foster the development of countermeasures against spoofing attacks on ASV systems. First, the building of public datasets such as the dataset, which is made up of a collection of genuine and replayed speech [5]. In the ReMASC dataset, the human speech captured by the microphone array was labeled as genuine speech, whereas the playback of the replay source recordings generated in different replay settings was labeled as replayed speech. In particular, the ReMASC dataset made up of 9,240 genuine and 45,472 replayed recordings. The speech corpus was collected from a total of 50 speakers, in particular, 22 female and 28 male speakers with ages range from 18 to 36. Among the 50 speakers, there were 36 English native speakers, 12 Chinese native speakers, and 2 Indian native speakers. About 132 voice commands made up of 273 unique words were used as recording materials to provide reasonable phonetic diversity. Four different recording environments with different noise levels were used, namely one outdoor environment, two indoor environments (quiet and noisy), and one vehicle environment. The building of a public dataset allowed the community of spoofing and anti-spoofing for ASV to develop robust PAD for ASV systems.

Second, the ASVspoof challenges were held to encourage the development of voice spoofing countermeasures. Standard datasets, techniques, and evaluation criteria were utilized in the ASVspoof Challenge series. The first ASVspoof Challenge, which covered speech synthesis and voice conversion attacks, was held in 2015. In the ASVspoof 2015 Challenge, the rating was based on 16 primary submissions. The best system in the ASVspoof 2015 Challenge had an EER of 1.21% on average [6]. Then, to emphasize replay attacks, the ASVspoof 2017 Challenge was launched. There was a much higher number of submissions received in the ASVspoof 2017 Challenge, recorded 49 submissions compared to the previous challenge. The best performing system in the ASVspoof 2017 challenge has achieved 6.73% EER [7]. The ASVspoof 2019 challenge was later organized to include speech synthesis, voice conversion, and replay attacks. The ASVspoof 2019 dataset can be divided into two types of attacks: logical attacks (LA) and physical attacks (PA). Genuine speech, speech synthesis, and voice conversion attacks were included in the LA dataset, whereas genuine speech and replay attacks were included in the PA dataset. For the LA and PA scenarios, the best submissions achieved 0.22% and 0.39% EER, respectively [8]. However, only 56.25% and 64% submissions for the LA and PA scenarios had outperformed the baseline system, respectively. Nonetheless, in both the LA and PA scenarios, the EER of the majority of the submissions had not been less than 5%. Due to the ease of obtaining biometric data, especially through social media, the security threat to the ASV systems is significant. Publicly available biometric data in social media can be used by security adversaries to launch presentation attacks such as speech synthesis and voice conversion to spoof the ASV systems. Furthermore, a large amount of artificial speech can be generated using state-of-the-art speech synthesis and voice conversion algorithms to spoof ASV systems. Whereby in this paper, the focus is on artificial speech due to it is the common spoofing attack as it can be generated in a short time to spoof the ASV system. Several voice PAD systems were introduced to detect artificial speech. As most of the artificial speech was produced using parametric vocoders, phase information was an effective feature to detect speech synthesis attacks. As a result, phase-based voice PAD for detecting artificial speech has become state-of-the-art [9]. However, ASV systems are still prone to attacks from artificial speech as most of the phase-based voice PAD introduced were only effective against artificial speech generated using minimum-phase filters based parametric vocoders [10].

There were numerous works found in the literature whereby the application of image classification in the signal domain was shown to be effective. To apply an image classification approach, audio data were pre-processed and transformed into image data. For example, features extracted from the Spectrogram image were shown to improve the performance of acoustic event classifications [11]. Besides, Spectrogram images were also being used for rapid speaker recognition and artificial speech detection [12]. The recent work [12], which used raw Spectrogram image as input for an end-to-end Light-ResNet-34 model, has outperformed the conventional approach of using constant q cepstral coefficients (CQCC) and GMM in artificial speech detection. Another recent work that applied deep neural network (DNN) architecture as a backend classifier with constant-q equal subband transform (CQ-EST) features [13] was shown to outperform most of the state-of-the-art approaches with an EER of 0.06%. Other than backend classifiers, deep learning architecture such as convolutional neural network (CNN) was also used as a feature extractor in recent works. In other work, a light gated CNN was used as a feature extractor to extract features from spectrogram image and probabilistic linear discriminant analysis (PLDA) as backend classifier to achieve an EER of 0.16% in artificial speech detection [14].

A fused system using short time fourier transform (STFT) and modified group delay (MGD) features were introduced recently and produced a 0.02% EER in detecting artificial speech. The advantage of this kind of fused system [15] is that both magnitude and phase spectral features were used together. This method yielded better performance than a fusion of independent systems with one feature for each system. Although most of the recent works achieved good EER, however, most of them did not perform well in detecting an S10 attack, one of the attack scenarios of the ASVspoof 2015 dataset. This indicates that more generalized models of artificial speech attacks are needed. In the context of artificial speech detection, the most recent works were using CNN as a feature extractor to extract image-based features from the spectrogram. Nonetheless, CNN usually requires a large number of training samples, computing time, and resources for better performance and generalization. However, similar performance can be achieved by utilizing handcrafted features for image classification, dealing with the abovementioned drawbacks. Moreover, work that utilized handcrafted image-based features in detecting artificial speech was limited in the literature. It is conjectured that using similar approaches to extract image-based features (color, texture, or edges) could be useful to generate more generalized features for artificial speech detection.

Despite the advancements made possible by speech identification technology, spoofing attacks by security adversaries to evade ASV systems is always a problem. To spoof ASV systems, state-of-the-art speech synthesis and voice conversion algorithms could easily generate artificial speech in massive quantities. Furthermore, because it is so easy to get biometric data via social media, spoofing attacks on ASV systems are becoming more common. As a result, robust spoofing countermeasures are required. These countermeasures are commonly known as voice PAD. Voice PAD has been the subject of various research studies, which may be found in the literature. Recent voice PADs, on the other hand, were vulnerable to unknown spoofing techniques [6]. The voice PADs submitted in the ASVspoof 2015 competition demonstrate this. System A, the best system in the ASVspoof 2015, proposed using two features for artificial speech detection: mel-frequency cepstral coefficients (MFCC) and cochlear filter cepstral coefficients plus instantaneous frequency (CFCCIF) using GMM classifier. Although system A performed with an average of 1.21% EER, the average EER for known and unknown attacks were 0.41% and 2.01%, respectively. Similarly, most systems submitted to the ASVspoof 2015 challenge encountered a similar circumstance in which they were unable to identify the S10 attack effectively, which was the sole attack produced using the waveform concatenation method. This pattern can be interpreted as possible overfitting in the proposed voice

PADs. Hence, more informative features are needed to generalize voice PAD against unseen spoofing attacks [16]. One solution is to produce new features using feature engineering, in which new descriptive features are constructed to be used to train a predictive model. This paper is written to propose a new feature engineering approach using data transformation techniques for artificial speech detection. In this work, rather than using conventional signal processing to extract features from speech, we proposed to use data transformation to apply the techniques from other domains such as image processing for artificial speech detection. The proposed approach is motivated by the success of deploying image classification techniques to sounds classification and speaker recognition [11]. An ensemble classifier in the form of random forest (RF) was used to generate the artificial speech model. The performance of the proposed approach is detailed, along with the results and discussion. Then, issues and future work to mitigate the limitation of the proposed approaches are described in this paper.

The key contributions of this paper are:

- Application of data transformation techniques to engineer image-based features to detect artificial speech
- Application of RF to be used with the new features engineered to detect artificial speech
- Empirical evaluation of the proposed approach with the existing work found in the literature

2. THE PROPOSED METHOD

In this paper, data transformation is considered to generate potential generalized features for voice PAD. In the conventional approach, features such as MFCC and CQCC are extracted directly from the speech signal to determine the genuineness of the speech. Recently, deep learning approaches, including CNN, were frequently being used to automatically extract features from image representation of speech signals. Unlike conventional and deep learning feature extraction approaches, the work presented in this paper proposed to use handcrafted features extracted from the image and hexadecimal frequency representation of the speech signal. In this paper, audio recordings were first transformed into images. Then, the image-based features are extracted from the transformed data to form the feature vectors. Figure 2 shows the differences between the conventional approach and the proposed feature engineering for artificial speech detection. The proposed feature engineering allows new features to be extracted from the speech data. Subsection 2.1 describes the generation of image-based features considered in this paper. Subsection 2.2 presents the RF classifier used for artificial speech detection in this work.

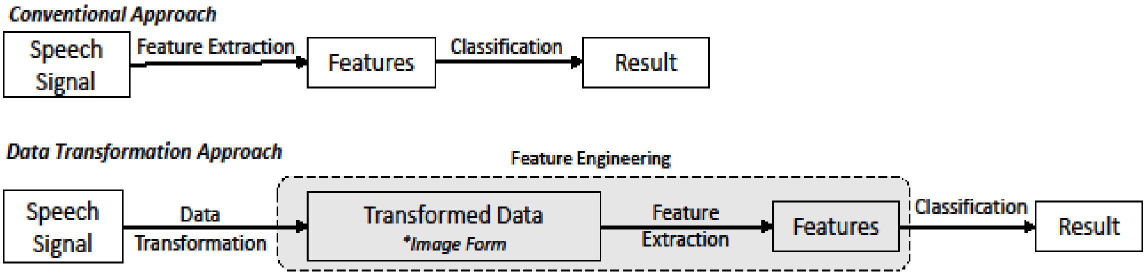

Figure 2. The comparison of the conventional approach and the proposed data transformation approach for artificial speech detection

2.1. Transformation of audio data into image representation and extraction of image-based features

Spectrogram and MFCCs are two common forms used to represent audio data [17], [18]. However, little attention has been paid to use both as images, whereby image-based features could be extracted for voice PAD. In work presented in this paper, the audio signals are represented as Spectrogram and MFCCs images. Different features were then extracted from each of the generated images. Figure 3 shows the process of the feature extraction from Spectrogram and MFCC images proposed in this paper. The speech signal is first transformed into spectrogram and MFCC images. Then, the color layout filter (CLF) and local binary patterns (LBP) features are extracted from the spectrogram to form the spectrogram-based features. concerning MFCC images, the CLF features are extracted to form the MFCC-based features.

A spectrogram is a representation of a signal that shows the signal's spectral information as frequency over time in the form of visual. Figure 4 shows how spatial differences between genuine and spoof voices using Spectrogram image representation could be observed. In this example, a genuine voice contains less background noise in a certain region of the spectrogram, while the spoof voice contains more background noise. To detect more differences between genuine and spoof voices in the voice-transformed images such as spectrogram, image classification techniques could be used as suggested in [17]. MFCC is an audio feature commonly used for signal processing, especially speech recognition [18], [19]. Figure 5 shows the generated MFCC images of genuine and spoof voices. From Figure 5, a slightly different color intensity in the region of a non-speech segment can be observed when comparing the MFCC images of genuine and spoof speech.

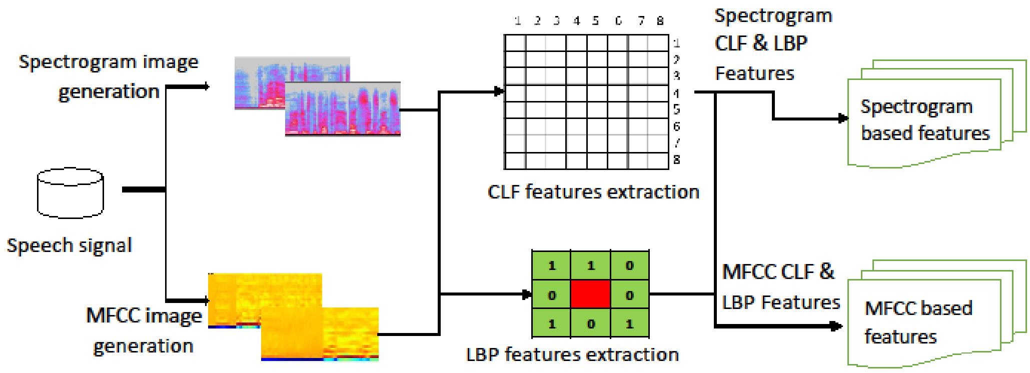

Figure 3. The feature extraction process

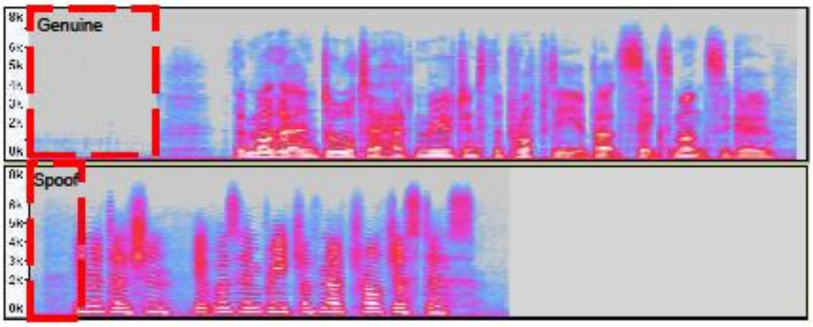

Figure 4. The observable spatial differences between genuine and spoof voices using spectrogram generated from the audacity tool

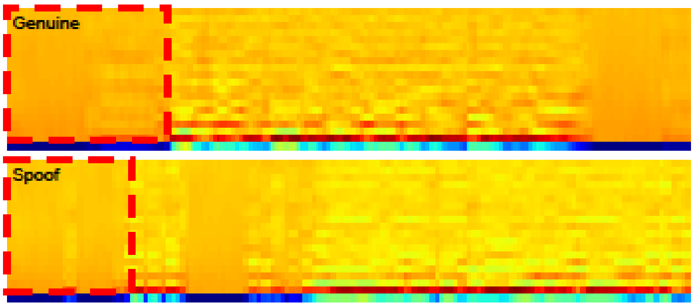

Figure 5. An example of MFCC images generated for genuine and spoof voices of a speaker

CLF was selected for feature extraction as it describes the spatial distribution of colors in an image and it works well in image classification when applied on color spectrogram [20]. In the CLF algorithm, the input image was divided into 64 blocks. Then, the values of all pixels within each block were averaged to obtain a representative color, resulting in three 8×8 arrays, collectively representing *YCbCr* color space.

Then, discrete cosine transform (DCT) was applied to the three 8×8 arrays, resulting in three DCT matrices, one for each *YCbCr* component. The CLF descriptor was formed by reading the coefficients from the matrices in zigzag order. The CLF descriptor contains a total of 33 features generated. As shown in Figures 4 and 5, genuine and artificial voices may be distinguished by the differences in the spatial distribution of color in certain regions of the generated Spectrogram and MFCC images. Figure 6 shows the process of CLF features extraction.

LBP is chosen in this paper as it is commonly used and produced good descriptors of texture in image classification. Figure 7 shows the process of LBP feature extraction. To extract LBP features from a Spectrogram image, the 3D color pixels were converted into 2D grayscale values. For each pixel in the converted grayscale image, a neighborhood radius *r* surrounding the center pixel was selected. Then, the LBP value was calculated for this center pixel and stored as a 2D array with the same height and width as the converted grayscale image. Then, the center pixel was compared to the surrounding neighborhood pixels, whether the neighbor pixels were greater-than-or-equal-to the center pixel. If the neighbor pixel was greater than or equal to the center pixel, then the value will be set as 1; otherwise, 0 will be set. The possible number of combinations of LBP codes was 2^p, where p is the number of neighborhood pixels. In original LBP, with neighborhood radius, =1 and p=8, there were 2^8=256 possible number combinations of the LBP codes, ranged 0-255. A frequency histogram of LBP codes was computed as LBP features. Details of LBP can be found in [21].

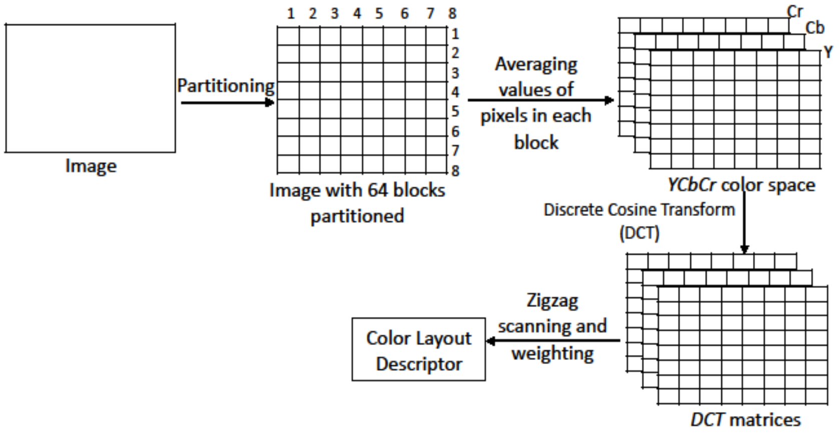
Image
Partitioning
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8
Image with 64 blocks partitioned
Averaging values of pixels in each block
Cr
Cb
Y
YCbCr color space
Discrete Cosine Transform (DCT)
DCT matrices
Zigzag scanning and weighting
Color Layout Descriptor

Figure 6. The process of CLF features extraction

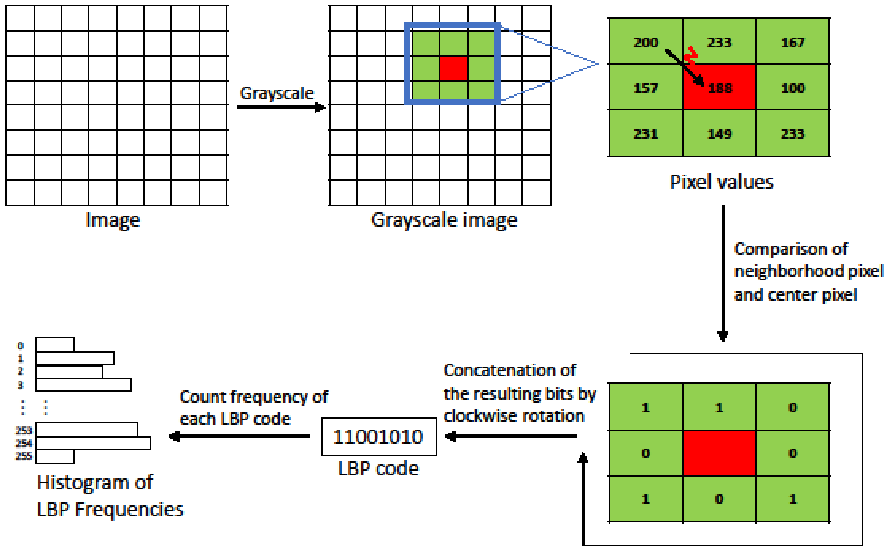
Image
Grayscale
Grayscale image
200
233
167
157
188
100
231
149
233
Pixel values
Comparison of neighborhood pixel and center pixel
1
1
0
0
0
1
0
1
Concatenation of the resulting bits by clockwise rotation
11001010
LBP code
Count frequency of each LBP code
0
1
2
3
253
254
255
Histogram of LBP Frequencies

Figure 7. The process of LBP features extraction

In this paper, the Spectrogram and MFCC images were generated from audio recordings using Python. Spectrogram and MFCC images were plotted using *pyplot* and *librosa* libraries, respectively. The generated images were then saved in PNG format with a size of 640×480 pixels. The CLF implementation in Weka [22] was used to extract the features. The *cvtColor()* algorithm of the *OpenCV* library (*cv2*) was employed for grayscale conversion in Python. Concerning LBP, a neighborhood radius r=1 was chosen as it was the setting used in the original LBP [21] and most used in the literature, in which there were eight neighboring pixels in a 3×3 pixels window. The frequency histogram computed from LBP was used as LBP features in this work. In total, 322 features were extracted from the spectrogram and MFCC represented audio data. A total of 289 features were generated from the spectrogram; 33 were CLF and 256 were LBP features. Concerning MFCC, a total of 33 CLF features were generated.

2.2. Random forest (RF) classifier for artificial speech detection

Features were extracted from data samples and automatically learned using a deep learning process, which was then used to predict the data samples' class labels in end-to-end learning. Unlike the end-to-end approach, a backend classifier is needed to differentiate between genuine and spoof speech using the proposed handcrafted features. In this work, an ensemble classifier is selected as it shows good classification results when applied with handcrafted features [23]–[25].

RF is a supervised, ensemble learning model where decision trees are bagged for classification and regression. In an RF model, multiple decision trees based on randomly selected training subsets were trained and merged to get a more accurate and stable prediction via votes aggregation.

The use of the greedy algorithm to select the best split point at each step in the tree building process will lead to similar resulting trees for bagged decision trees. This resulting in the reduction in the variance of the predictions of the bagged decision trees. To mitigate this issue, RF is an improved version of bagged decision trees that disrupt the greedy splitting algorithm during tree creation. When the greedy splitting algorithm is disrupted during tree creation in RF, the split points of decision trees can only be chosen from a subset of the input features at random. As a result, the similarity between the bagged decision trees decreased and led to lower bias and higher variance of the predictions. Due to its simplicity and predictive performance, RF was chosen as a backend classifier for artificial speech detection in this work. Details of RF can be found in [26]. The Weka implementation of the identified classifiers was used in work presented in this paper.

3. RESEARCH METHOD

An experiment was executed to test the proposed approach's generalization capability in identifying artificial speech as an independent system rather than as part of an integrated ASV system. The ASVspoof 2015 dataset, the largest and most used public dataset for artificial speech detection, was used to measure the performance of the proposed approach. The recent ASVspoof 2019 dataset was not included as it was designed to evaluate the impact of the countermeasures on the reliability of an ASV system when subjected to spoofing attacks [27], which is out of the scope of the work presented in this paper.

The ASVspoof 2015 dataset used in the experiment was made up of speech synthesis and voice conversion attacks in addition to genuine speeches. The ASVspoof 2015 dataset was collected and generated from a total of 106 speakers, specifically 45 male and 61 female speakers. The genuine speeches of the ASVspoof 2015 were recorded in a semi-anechoic chamber having a solid floor, whereas the spoof speeches were generated using ten different common speech synthesis and voice conversion algorithms. These algorithms produced ten different categories of attacks (S1-S10). The known attacks in the ASVspoof 2015 dataset were made up of S1-S5 attacks, which used common voice conversion and speech synthesis algorithms. The unknown attacks in the ASVspoof 2015 dataset were made up of S6-S10 attacks. S1, S2, and S6-S9 attacks were generated using voice conversion algorithms, whereas S3, S4, and S10 attacks were generated using speech synthesis algorithms. Details on each of the ten spoofing algorithms used in the production of spoof speeches in the ASVspoof 2015 dataset are available in [6].

Four types of features were extracted from the audio recordings of the ASVspoof 2015 dataset, whereas the classifications were conducted using the weka tool. Most of the parameters set in the Weka tool were empirically found to be working well in most cases; hence this work uses the suggested parameters by Weka. There were training, development, and evaluation sets in the ASVspoof 2015 datasets. As described in [6], the training set is to train and build a PAD model, whereas the development set is for model tuning and refinement, and the evaluation set is for model evaluation. An experiment was conducted to evaluate the performances of the different combinations of the extracted features and classifiers. The experiment was conducted for each combination of features and classifiers, where both training and development sets were used to train the model, whereas the evaluation set was used for validation. The experiment was conducted using a machine with specifications: Intel i5-3210M processor, 2.50 GHz, 8 GB of RAM, Windows 10 (64-bit) OS.

4. RESULTS AND DISCUSSION

In this work, we showed the accuracy and F1-score of each model as a supporting metric in addition to the EER for a better comparison of the model performances. This is because a lower EER may not necessarily indicate that a model predicted more instances correctly. F1-score is often used in binary classification to evaluate how good the classifier is in detecting positive cases. Table 1 shows the results of experiment 1, whereby the detection was performed to identify genuine or artificial speech, with the best-performed combination of features and classifiers using the four features proposed. Several recent works that used both training and development sets for model training, namely Model$_1$-Model$_3$ [28]–[30], were used for comparison to present the competitiveness of the proposed approach to the state-of-the-art. In the remainder of this section, the combination of features and classifiers is represented using the model number, as shown in Table 1. Figure 8 shows detection error tradeoff (DET) curves for Model$_4$-Model$_7$ in experiment 1.

Table 1. The performances of the models trained with both train set and development set and tested with the evaluation set of the ASVspoof 2015 dataset

Model	Experiment 1 (Eval Set)		
	EER (%)	Accuracy (%)	F1-Score (%)
Model$_1$: Scattering cepstral coefficients (SCC)+GMM-UBM [28]	0.18	-	-
Model$_2$: Compressed sensing for high dimensional features (CS-HD)+i-vector [29]	0.24	-	-
Model$_3$: CQCC+SCC+GMM-UBM [30]	0.10	-	-
Model$_4$: Spectrogram image CLF+RF	17.61	93.02	96.42
Model$_5$: Spectrogram image LBP+RF	17.09	94.35	97.11
Model$_6$: MFCC image CLF+RF	0.10	99.93	99.96
Model$_7$: MFCC image LBP+RF	30.01	95.01	97.45

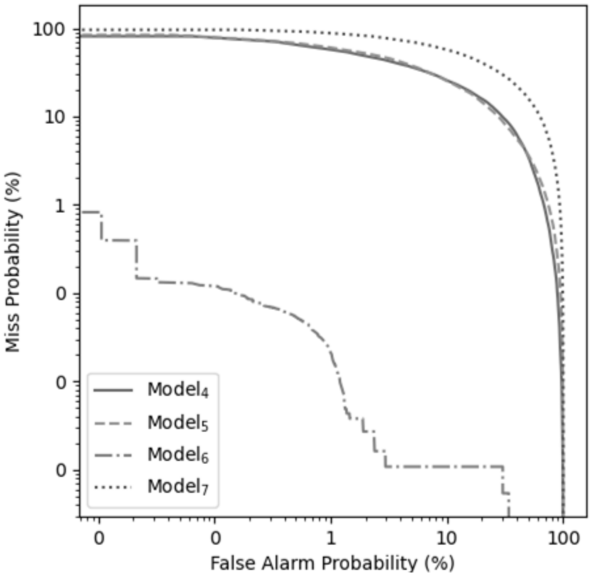

Figure 8. DET curves of Model$_4$-Model$_7$ in experiment 1

As the primary metric used in the ASVspoof 2015 was EER only, hence the accuracy of the Model$_1$-Model$_3$ are not shown in Table 1. From Table 1, most of the proposed models performed with over 17% EER in experiment 1. All the proposed models (Model$_4$-Model$_7$) achieved accuracy over 90% in experiment 1. In experiment 1, Model$_6$ achieved the lowest EER and the highest accuracy among the proposed models. It can be observed clearly in Figure 8 that the performance of Model$_6$ was far better than other models. On the other hand, all the proposed models (Model$_4$-Model$_7$) achieved over 96% F1-score. This indicates that the proposed models were very good in detecting the spoof voices while at the same time has low misclassification of genuine instances as a spoof.

The combination of the MFCC image with the CLF feature extractor has produced a robust feature that enabled the Model$_6$ to perform the best in experiment 1. MFCC uses a Mel scaling that produces a series of coefficients resembling the resolution of the human auditory system, which is different from spectrogram that uses a linear frequency scaling. In addition, the differences in the spatial distribution of color in the

MFCC images between genuine and spoof could be detected by the CLF feature extractor. Therefore, the MFCC based features performed the best when using the CLF feature extractor for artificial speech detection. Other than the robustness of the MFCC CLF feature, the use of RF may be one of the factors that lower EER was achieved by Model$_6$ in experiment 1. Due to the nature of RF, decision trees with more variation were built when the number of instances in training data to be randomly selected increases. Eventually, this produces a more generalized predictive model as the similarity between the bagged trees decreased. Therefore, having more data in model training may improvise the detection rate, though it may not always be the case.

In terms of detection error trade-off, a DET curve was presented in Figure 8 using the results obtained in experiment 1. A DET curve shows the detection error trade-off between the false-negative rate (miss probability) and false positive rate (false alarm probability) of a binary classification model. From Figure 8, Model$_6$ has a significantly lower detection error trade-off than other models in experiment 1. This indicates Model$_6$ performed significantly better than other models in experiment 1.

Besides, the robustness of Model$_6$ can also be seen by looking at the ISO/IEC standard metrics, namely, attack presentation classification error rate (APCER) and bonafide presentation classification error rate (BPCER) of the model. In total, only 130 out of 193,404 instances (0.07%) in the ASVspoof 2015 evaluation set were misclassified by Model$_6$. The APCER of Model$_6$ was 0.02%, given 29 out of 184,000 spoof instances were misclassified as genuine. The BPCER of Model$_6$ was 1.07%, given 101 out of 9,404 genuine instances were misclassified as a spoof. Nonetheless, the difference between APCER and BPCER was about 1%. To further compare our best model, Model$_{6,}$ with recent works, the comparison of artificial speech detection by category of attacks (S1-S10) is presented in Table 2.

Table 2. The comparison of the performance of our best model with recent works on the evaluation set of the ASVspoof 2015 dataset by category of attacks (S1-S10)

Model					EER (%)					
			Known Attack					Unknown Attack		
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Model$_1$		0.02					0.33			
Model$_2$	0.02	0.03	0.01	0.01	0.02	0.01	0.00	0.01	0.00	26.28
Model$_3$	0.00	0.01	0.00	0.00	0.02	0.01	0.01	0.00	0.00	0.95
Model$_6$	0.14	0.10	0.02	0.02	0.21	0.21	0.14	0.02	0.12	0.03

From Table 2, it can be observed that our model, Model$_{6,}$ significantly outperformed Model$_1$ - Model$_3$ for the S10 attacks scenario, the most difficult spoofing attack. Model$_6$ also produced comparable performances on other categories of attack. Besides, it can be observed that Model$_6$ recorded 0.02-0.21% EER across S1-S10 attacks. Model$_6$ that performed with an overall EER of 0.10%, recorded a significantly higher EER of 0.95% on the S10 attack despite the model achieved below 0.02% EER in other attacks (S1-S9). This indicates that Model$_6$ is more generalized than the others.

There were 9,404 genuine instances and 18,400 spoof instances of each attack type (S1-S10) in the ASVspoof 2015 evaluation set. The misclassified instances for genuine, known, and unknown attacks by Model$_6$ were 101, 11, and 18 instances, respectively. An interesting observation is that there were no instances from S10 attacks being misclassified as genuine by Model$_6$, in which the 0.03% EER for S10 attacks were incurred by the false alarm. Unlike known attacks (S1-S5), which were generated using a vocoder, S10 attacks were generated without a vocoder. This was the factor that has caused most of the state-of-the-art voice PAD systems to suffer from significantly higher EER on S10 attacks. Comparably, Model$_6$ has successfully identified all S10 attacks as a spoof; hence it was more generalized and effective in detecting artificial speech regardless of the use of the vocoder.

To prevent unreliable performance evaluation, EER, accuracy, and F1-score were used in this work. From the results shown in Table 1, the performance of Model$_6$ is reliable as the EER, accuracy, and F1-score were good. Very high accuracy and F1-score but high EER can be obtained if the proportion of either class of the test set was overwhelming and the model biased toward one of the classes with overwhelming proportion. For example, a test set was made up of 100 instances with 90 spoof instances and ten genuine instances. If the model was bias and overfitting, it might predict all instances of the test set as a spoof to achieve high accuracy and F1-score. In this case, the accuracy of the model would be 90%, whereas the EER would be 50%. However, Model$_6$ was not the case. From the results, as shown in Tables 1 and 2, the low EER achieved by Model$_6$ indicated that the high accuracy achieved was neither due to bias nor overfitting. The combination of RF and MFCC image-based CLF features was shown to be effective in detecting artificial speech as Model$_6$ produced a low EER of 0.10% while achieving high accuracy and F1-score of 99.93% and 99.96%, respectively. It is also shown to be able to produce similar detection performance on all categories of attacks (S1-S10).

5. CONCLUSION

In this paper, a feature engineering approach to produce handcrafted features for artificial speech detection was proposed. The contribution of this paper is in the proposed combination of features engineered using data transformation approaches and RF classifier for artificial speech detection. Four types of image-based spectrogram and MFCC features were extracted to classify genuine and spoof speeches. The ASVspoof 2015 dataset was used in the experiment to determine the effectiveness of the proposed approach against artificial speech. An experiment was run to compare the performance of the new features with the RF classifier for artificial speech detection. From the experiment, the results showed that the proposed approach could produce a model (Model$_6$) which used an Image Filter called CLF to extract features from MFCC images and a Random Forest as the classifier. The combination of the MFCC CLF feature and RF classifier generated a well-performed model, which yields good EER, accuracy, and F1-score of 0.10%, 99.93%, and 99.96%, respectively, in detecting artificial speech. However, in a real-world scenario, speech data were always exposed to various noises that deteriorate the audio quality. As the ASVspoof 2015 dataset contained only clean audio recording, the proposed approach may not be able to achieve similar performance when tested on the noise added dataset. Hence, future work is directed to test the proposed approach on the noise added dataset. Then, the investigation of feature fusion and ensemble classifiers to improve the performance further and to expand the detection to other types of presentation attacks such as replay attacks. In addition, more datasets will be used to evaluate the generalization capability of feature engineered using a data transformation approach against previously unseen spoofing attacks. Lastly, the integration of the proposed approach with the ASV systems will be conducted and tested on the ASVspoof 2019 dataset.

REFERENCES

[1] A. A. Mallouh, Z. Qawaqneh, and B. D. Barkana, “New transformed features generated by deep bottleneck extractor and a GMM–UBM classifier for speaker age and gender classification,” *Neural Computing and Applications*, vol. 30, no. 8, pp. 2581–2593, Oct. 2018, doi: 10.1007/s00521-017-2848-4.

[2] A. I. Abdurrahman and A. Zahra, “Spoken language identification using i-vectors, x-vectors, PLDA and logistic regression,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2237–2244, Aug. 2021, doi: 10.11591/eei.v10i4.2893.

[3] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, “End-to-end text-dependent speaker verification,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 5115–5119, doi: 10.1109/ICASSP.2016.7472652.

[4] S.-H. Yoon and H.-J. Yu, “A simple distortion-free method to handle variable length sequences for recurrent neural networks in text dependent speaker verification,” *Applied Sciences*, vol. 10, no. 12, Jun. 2020, doi: 10.3390/app10124092.

[5] Y. Gong, J. Yang, J. Huber, M. MacKnight, and C. Poellabauer, “REMASC: realistic replay attack corpus for voice controlled systems,” in *Interspeech 2019*, Sep. 2019, pp. 2355–2359, doi: 10.21437/Interspeech.2019-1541.

[6] Z. Wu *et al.*, “ASVspoof: the automatic speaker verification spoofing and countermeasures challenge,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, Jun. 2017, doi: 10.1109/JSTSP.2017.2671435.

[7] T. Kinnunen *et al.*, “The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection,” in *Interspeech 2017*, Aug. 2017, pp. 2–6, doi: 10.21437/Interspeech.2017-1111.

[8] M. Todisco *et al.*, “ASVSpoof 2019: future horizons in spoofed and fake audio detection,” in *Interspeech 2019*, Sep. 2019, vol. 2019-Septe, pp. 1008–1012, doi: 10.21437/Interspeech.2019-2249.

[9] M. Pal, D. Paul, and G. Saha, “Synthetic speech detection using fundamental frequency variation and spectral features,” *Computer Speech and Language*, vol. 48, pp. 31–50, Mar. 2018, doi: 10.1016/j.csl.2017.10.001.

[10] C. Demiroglu, O. Buyuk, A. Khodabakhsh, and R. Maia, “Postprocessing synthetic speech with a complex cepstrum vocoder for spoofing phase-based synthetic speech detectors,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 671–683, Jun. 2017, doi: 10.1109/JSTSP.2017.2673807.

[11] I. Ozer, Z. Ozer, and O. Findik, “Lanczos kernel based spectrogram image features for sound classification,” *Procedia Computer Science*, vol. 111, no. 2015, pp. 137–144, 2017, doi: 10.1016/j.procs.2017.06.020.

[12] P. Parasu, J. Epps, K. Sriskandaraja, and G. Suthokumar, “Investigating light-ResNet architecture for spoofing detection under mismatched conditions,” in *Interspeech 2020*, Oct. 2020, pp. 1111–1115, doi: 10.21437/Interspeech.2020-2039.

[13] J. Yang, R. K. Das, and H. Li, “Significance of subband features for synthetic speech detection,” *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2160–2170, 2020, doi: 10.1109/TIFS.2019.2956589.

[14] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, “A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection,” in *Interspeech 2019*, Sep. 2019, vol. 2019-Septe, pp. 1068–1072, doi: 10.21437/Interspeech.2019-2212.

[15] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, “A gated recurrent convolutional neural network for robust spoofing detectionn,” *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 27, no. 12, pp. 1985–1999, Dec. 2019, doi: 10.1109/TASLP.2019.2937413.

[16] C. B. Tan *et al.*, “A survey on presentation attack detection for automatic speaker verification systems: state-of-the-art, taxonomy, issues and future direction,” *Multimedia Tools and Applications*, vol. 80, no. 21–23, pp. 32725–32762, Sep. 2021, doi: 10.1007/s11042-021-11235-x.

[17] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, “Speech emotion recognition from spectrograms with deep convolutional neural network,” in *2017 International Conference on Platform Technology and Service, PlatCon 2017 - Proceedings*, Feb. 2017, vol. 24, no. 6, pp. 1–5, doi: 10.1109/PlatCon.2017.7883728.

[18] C. K. On, P. M. Pandiyan, S. Yaacob, and A. Saudi, “Mel-frequency cepstral coefficient analysis in speech recognition,” in *2006 International Conference on Computing and Informatics*, Jun. 2006, pp. 1–5, doi: 10.1109/ICOCI.2006.5276486.

[19] D. Y. Mohammed, K. Al-Karawi, and A. Aljuboori, “Robust speaker verification by combining MFCC and entrocy in noisy conditions,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2310–2319, Aug. 2021, doi: 10.11591/eei.v10i4.2957.

[20] M. Towsey *et al.*, “Long-duration, false-colour spectrograms for detecting species in large audio data-sets,” *Journal of Ecoacoustics*, vol. 2, no. 1, pp. 1–1, Apr. 2018, doi: 10.22261/JEA.IUSWUI.

[21] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, Jan. 1996, doi: 10.1016/0031-3203(95)00067-4.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, Nov. 2009, doi: 10.1145/1656274.1656278.

[23] J. Monteiro, J. Alam, and T. H. Falk, “An ensemble based approach for generalized detection of spoofing attacks to automatic speaker recognizers,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6599–6603, doi: 10.1109/ICASSP40776.2020.9054558.

[24] M. H. A. Hijazi, S. Kieu Tao Hwa, A. Bade, R. Yaakob, and M. Saffree Jeffree, “Ensemble deep learning for tuberculosis detection using chest X-ray and canny edge detected images,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 8, no. 4, pp. 429–435, Dec. 2019, doi: 10.11591/ijai.v8.i4.pp429-435.

[25] T. R. Jayanthi Kumari and H. S. Jayanna, “Limited data speaker verification: fusion of features,” *International Journal of Electrical and Computer Engineering*, vol. 7, no. 6, pp. 3344–3357, Dec. 2017, doi: 10.11591/ijece.v7i6.pp3344-3357.

[26] O. Sagi and L. Rokach, “Ensemble learning: a survey,” *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4, pp. 1–18, Jul. 2018, doi: 10.1002/widm.1249.

[27] X. Wang *et al.*, “ASVspoof 2019: a large-scale public database of synthetized, converted and replayed speech,” *Computer Speech and Language*, vol. 64, Nov. 2020, doi: 10.1016/j.csl.2020.101114.

[28] K. Sriskandaraja, V. Sethu, E. Ambikairajah, and H. Li, “Front-end for antispoofing countermeasures in speaker verification: scattering spectral decomposition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 1–1, 2017, doi: 10.1109/JSTSP.2016.2647202.

[29] Y. Zhao, R. Togneri, and V. Sreeram, “Compressed high dimensional features for speaker spoofing detection,” in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec. 2017, pp. 569–572, doi: 10.1109/APSIPA.2017.8282108.

[30] Y. Zhao, R. Togneri, and V. Sreeram, “Spoofing detection using adaptive weighting framework and clustering analysis,” in *Interspeech 2018*, Sep. 2018, vol. 2018-Septe, pp. 626–630, doi: 10.21437/Interspeech.2018-1042.

BIOGRAPHIES OF AUTHORS

Choon Beng Tan received his B.Comp.Sc. and M.Sc. degrees from Universiti Malaysia Sabah (UMS) in 2016 and 2018. He is now doing his PhD study in Computer Science in Universiti Malaysia Sabah (UMS). His recent work includes malware classification using machine learning and ensemble techniques, and cloud data integrity scheme; he is now working on voice presentation attack detection. His research interests include information security, cloud computing, and voice biometric security. He can be contacted at email: tanchoonbeng@ums.edu.my.

Mohd Hanafi Ahmad Hijazi is an Associate Professor of Computer Science at the Faculty of Computing and Informatics, Universiti Malaysia Sabah in Malaysia. His research work addresses the challenges in knowledge discovery and data mining to identify patterns for prediction on structured and/ or unstructured data; his particular application domains are medical image analysis and understanding and sentiment analysis on social media data. He has authored/ co-authored more than 50 journals/ book chapters and conference papers, most of which are indexed by Scopus and ISI Web of Science. He also served on the program and organizing committees of numerous national and international conferences. He is the leader of data technologies and applications research group at the faculty. He can be contacted at email: hanafi@ums.edu.my.

Frazier Kok holds a Degree in Computer Science from Universiti Teknologi Malaysia (UTM) 2001 and has 20 years’ experience in ICT industry. He has completed various projects ranging from being a programmer, designer, solution architect, project coordinator and project manager. Now as an Executive Director in Bayurini Sdn Bhd, he is also extensively involved in Major Product Deployment in various customer sites. He can be contacted at email: frazier@bayurini.com.

Dr. Mohd Saberi Mohamad is a Professor of Bioinformatics and Artificial Intelligence at the Department of Genetics and Genomics, the College of Medicine and Health Sciences, UAE University. His research areas are Bioinformatics, Artificial Intelligence, Data Science, and Computational Biology. He has received the 2018 Malaysia's Research Star Award in the category of young researchers by the Malaysia Ministry of Higher Education. He has been the project leader for 19 research grants and the project member for 20 research grants. He has also published 282 articles in the international refereed journal, international conferences, and book chapters; and produced 14 books (research books, edited books, and original book). Previously before joining UAUE, he had experience as a Director for the Institute For Artificial Intelligence and Big Data, a founder of the Department of Data Science, and a head of the Artificial Intelligence and Bioinformatics Research Group. Besides, he was also a head and member of 65 academic and research committees at the faculty, university, national, and international levels. He also has been appointed as an advisory board for Artificial Intelligence Research Institute and IoT Digital Innovation Hub in Europe. He served as seven chairman and 39 members of the committees for several international conferences in Europe, ASEAN, Japan, US, Australia, China, and Malaysia. He has become an accreditation panel to Malaysia Qualification Agency (MQA) to review academic programs in the field of Bioinformatics since 2013. He can be contacted at email: saberi@uaeu.ac.ae.

Dr. Puteri Nor Ellyza binti Nohuddin received her B.Sc. in Computer Science from University of Missouri-Columbia, USA and her M.Sc. IT from Universiti Teknologi MARA. In 2012, she was awarded her Ph.D. in Computer Science from the University of Liverpool, UK. Puteri joins Institute of IR4.0 (IIR4.0), Universiti Kebangsaan Malaysia as a Research Fellow in July 2015. Prior to coming to IIR4.0, she was lecturer at the Universiti Pertahanan Nasional Malaysia, Kuala Lumpur. Prior to her academic career, she worked with several conglomerates such as ExxonMobil, Sime Darby, Shell IT and Malaysian Resources Corporation Berhad as System Analyst. Puteri's teaching interests include Programming, Database systems and Data mining. Her primary research interests are in the field of Big Data, Data Mining and Knowledge Engineering. Specifically, she is interested in Time Series Clustering, Trend mining, Tacit Knowledge, and Social Network Analysis. She can be contacted at email: puteri.ivi@ukm.edu.my.