INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

EVALUATING THE EFFICACY OF VOICE RECOGNITION ALGORITHMS USING CUTTING-EDGE DEEP LEARNING TECHNIQUES

Garaga Srilakshmi[1], Dr. Alok Agarwal[2], Dr.Dola Sanjay S[3]
Research scholar ECE, JJTU, Rajasthan.
Guide, Department of Electronics and Communication Engineering, JJTU, Rajasthan.
Co-Guide, Principal Aditya College of Engineering And Technology, Surampalem. A.P

ABSTRACT

A speech enhancer is a piece of software designed to improve the quality of voice recordings. Thanks to the speech enhancement technique, the input voice is more easily recognized from the background noise. Restoration of the original quality of speech that has been damaged by background noise is the primary focus of speech enhancement research. This development has a wide range of potential applications, including but not limited to teleconferencing, speech recognition, mobile phones, and assistive listening devices. The potential for speech enhancement technology has grown significantly with the introduction of smart phones, smart watches, and other wearable and augmented reality devices. In order to improve the audibility of remote control and command, speech augmentation technology may be used. Modern smart devices with voice augmentation are designed to comprehend people more accurately. There may be room for improvement in speech augmentation for usage with voice-controlled devices. But current speech enhancers can only get rid of quasi-stationary noise because of computational limitations. Once the interference is gone, the strength of the remaining signal may start to fade.

Improvements in language detection and customization might increase hearing aids' use. The phrase "speech enhancement" is often used to refer to a broad variety of approaches intended to improve the quality of transmitted sound in a variety of ways, such as by boosting speech intelligibility or decreasing listener fatigue. It's possible that situations like these might benefit from the use of artificial speech augmentation to counteract background noise.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

I. INTRODUCTION

Communication between individuals of our species was a primary evolutionary goal. It's possible that a person's voice might be used as a kind of identification. Identifying a person just from their voice is what's known as "automatic speaker recognition" (ASR). Speech recognition and other forms of biometric identification are gaining popularity. Science concerned with determining a person's identity through physical traits. A person's voice may be utilized as a kind of biometric authentication due to its distinctive tone. Listening to someone talk may tell you a lot about their character and health. Iris, face, retina, fingerprints, ear, DNA, etc. are all examples of physiological biometrics, whereas voice, signature, keystrokes, typing, etc. are examples of behavioral biometrics. There has been an increase in studies including voice biometrics [1, 2].

Communicating verbally with another individual facilitates the exchange of ideas and emotions. Human nature makes verbal communication the default mode of engagement. Voice signals [3, 4] include not just auditory and articulation data, but also meaning and context data about the words being said. When communicating with others, humans have a natural talent for picking up on nuances in tone, body language, and emotional condition. The methods by which humans generate and decode speech signals have been the subject of several studies. Applications that can analyze and interpret voice signals automatically may benefit from this concept. Every person's voice is unique in some way. A person's voice will sound different from another's owing to both genetic and environmental factors [3-5].

Speaker recognition is the technique of identifying the speaker from information in an audio signal or waves. Speaker recognition is the most practical biometric identifying method in the modern digital environment [5]. Financial organizations, factories, and police departments all support this technology because to its increased security for massive datasets [5-6]. Separate from verification (1:1), identification (1: N matching) is another subset of speaker recognition. Identification, as most would agree, is more difficult than verification [7]. As the number of users grows, there is a greater possibility that a misidentified speaker may be picked up. Since just two speakers are compared, the performance of the speaker verification system is independent of the size of the voice database.

With the recent growth of digital information, verification is more important than ever. There is mounting evidence that biometric identification is the most secure method currently in use. Verifying a speaker's identity means either accepting or rejecting their claimed identity, as opposed to just

ISSN: 2057-5688

confirming that they made a certain statement, as is the case with speaker identification. Perhaps the future of biometric identification can be seen in speaker verification technologies. Both text-dependent and text-independent forms exist [8]. However, text-independent methods [7] don't need the speaker to repeat the same sentence/words throughout training and assessment. A text-based system is more trustworthy since it evaluates not just the content but also the speaker's voice. The same phrases may or may not be used by text-independent recognition systems for verifying a user's identity.

In speaker modeling for idling applications, the Gaussian mixture model (GMM) is often used. Parameter estimate in GMM for identification use the EM approach when the principle of the training database is maximum likelihood (ML). Maximum a posteriori (MAP) training databases employ the EM method for parameter estimation. To train a speaker database, a GMM (universal background model) is built using the UBM (with particular data from registered speakers) [9, 11].

Speaker recognition and authenticity checking are the two main components of voice recognition. When a user's identification has to be validated in a voice-based application, speaker verification is utilized. Speaker identification is the process of attributing a particular speaker to a particular speech act. Confirming someone's identity is always preferable than just identifying them. Voice recognition software is very useful for forensic investigators. In this application, it is utilized to locate a database speaker who most closely resembles the suspect. Accurate speaker identification gets more challenging as audio files grow in size. For speaker confirmation, just a yes/no verdict is needed, hence a large speech database is unnecessary. Recognizability is heavily influenced by the audio signal spectrum and by session-to-session variation [1, 12].

II REVIEW OF LITERATURE

Similar to feature extraction methods, ML and DL classifiers can be used to evaluate the speaker identification model's overall efficacy. Many studies have been conducted to identify the best classifier for accurately identifying the speaker's social status. Classical speaker models include both template and stochastic models, also known as nonparametric and parametric, respectively. Parametric approaches typically make assumptions about the development of extracted features, while nonparametric approaches do not.

A hybrid approach, involving GMM at the front end and SVM at the back end, was proposed by Fine et al. (2001). In this method, the GMM output was used as input to the SVM classifier. The GMM's non-

ISSN: 2057-5688

linear probabilistic mean values are used to train a support vector machine to minimize the gap between the hyper planes. This method was developed on the premise that calculating the log-likelihood ratio is inefficient due to the inherent imprecision of probabilities.

Tolba (2011) used a Hidden Markov Model (HMM) to select a speech signal description (MFCCs) that could distinguish between different speakers of Arabic. The obtained feature chain is segmented and the likelihood function is generated under the assumption of vector independence during the analysis phase. The selected presenter is determined by contrasting likelihood functions from various models. Text-independent technologies have a relatively low level of accuracy when making predictions.

Zhao et al. (2014) claimed that the topic of speech recognition in a noisy environment had received little attention until they focused on it. To determine the similarity between two time sequences that vary in speed or duration, we can use a technique called dynamic time warping (DTW). It was more flexible and productive, but also more difficult to use.

Daqrouqa & Tutunjib (2015) trained the back-end feed-forward NN for speaker detection using audio formants and wavelet packet entropy-based 12 features. The proposed method is effective at capturing vowel features. One advantage of using vowels is that it's possible to distinguish between people even when only partial recordings of words are available. This may prove helpful if the recordings are erased or if the user is deaf or mute.

For the TIMIT 8K database, Ge et al. (2017) proposed a neural network architecture for text-free speaker verification and identification. The 39 MFCCs were generated using the pre-processed audio. In addition to ANN, SVM is also used in ASR. Pattern classification and regression using a NN's weight update becomes more difficult when external noise is present simultaneously with the recorded speech signal.

Several varieties of deep features for text-dependent SV were provided by Liu et al. (2015). All these novel components are provided in the GMM-UBM and identifying vector frameworks. Discrimination based on linear models and the law of probabilities is used in the ASR classifier's back end. Extraction of deep features by hand becomes more difficult as the number of speakers grows.

For ASI systems, Tirumala and Shahamiri (2017) suggested using Deep Auto Encoders (DAEs). The research confirms the relevance of scale, illuminating the differences in precision between standard back-propagation and layer-wise implementation. Due to the small size of the bottleneck layer in DAE, the output may be missing or incorrectly interpreting critical aspects of the input speech signal.

ISSN: 2057-5688

Qayyum et al. (2018) stressed the importance of Bidirectional Long Short-Term Memory (BLSTM) for the success of Recurrent Neural Networks (RNNs). Using these features, a hybrid RNN-BLSTM system is built to improve the ongoing speaker recognition task. Weight sharing, the appropriate number of hidden units, and the best possible pooling strategy are just a few of the many techniques used in RNN that contribute to its high recognition rate.

Mutual Information (MI) or similar evaluation models were proposed by Ravanelli and Bengio (2019) as practical metrics for unsupervised learning of drawings. It determines the speaker by selecting sentences at random and then maximizing the mutual information between them. The SincNet architecture is used for feature extraction in the proposed encoder. During the show

III RESEARCH METHODOLOGY AND APPROACH

Pathological speech recognition has been extensively researched throughout the last decade. When it comes to the identification and classification of various vocal problems, speech processing is a valuable tool. Parkinson's disease (PD) and multiple sclerosis (MS) are two of the most studied neurodegenerative diseases due to their devastating effects on patients' capacity to communicate, comprehend, and move.

Speech analysis has been largely disregarded for decades because of how hard it is. Some of the numerous things that may go wrong with your voice include overuse, stress, smoking, acid reflux, and hormone abnormalities. The diagnostic procedure of choice for problems with the vocal folds is called direct laryngoscope, in which the larynx is seen directly via a camera. Local anesthesia is generally utilized since it is less intrusive and more pleasant for the patient. An inaccurate diagnosis might be made by a Pathologist who is overworked, stressed, or lacking in sleep. Consequently, incorrect diagnoses are a major issue. Indirect laryngoscope using a mirror requires less local anesthesia but produces similar results as direct laryngoscope. However, the continuous costs of maintaining such machinery are rather expensive. Pathologists are able to better identify voice problems, choose effective treatments, and predict the outcomes of their patients' treatments by carefully analyzing patient complaints. A rapid, painless, simple, and affordable approach of illness identification might be useful for the preliminary evaluation and diagnosis of voice abnormalities.

This research endeavors to solve these problems by developing a piece of software that can

ISSN: 2057-5688

automatically detect and label anomalies in human speech. The purpose of this research is to construct an algorithm that will be able to automatically detect potentially dangerous abnormalities in speech.

3.1 Research Design

There are several recommendations for improving the reliability of the study's results sprinkled throughout the research plan. We provide more practical methods for speech recognition in this study. When examining voice data, acoustic analysis is performed to help isolate the most relevant bits of information. In this study, we examined several classifiers and established a procedure for picking the most efficient one. Figure 3.1 is a research roadmap that details future upgrades and modifications to the Voice Pathological Identification System. Here are the steps that must be taken to complete this probe:

In the first stage, called "preprocessing," a parametric representation is created.

The first step is to offer a filter set that focuses on reducing noise and restoring silence by using a hybrid wiener and discrete wavelet transformation (HWFDWT).

Combining Cat Swarm Optimization with MFCC coefficients, the proposed technique CSOMFCC seeks to decrease the process's dimensionality and execution time while increasing feature selection and extraction.

Group the information in a meaningful way. The goal is to sort the data set into "normal" and "abnormal" categories based on the characteristics of the voices.

In this last stage, we dig further into the study to determine whether the suggested Modified BPNN algorithm can boost classification accuracy and speed.

Instructions for the Final Phases of Growth Pathology Disorders Classification Distinction the ROC Curve was created as a result of the success of Pathological Voices in terms of accuracy.

Some metrics that may be used to evaluate the performance of the suggested model include signal-to-noise ratio, accuracy, specificity, sensitivity, duration, and the Roc Curve.

ISSN: 2057-5688

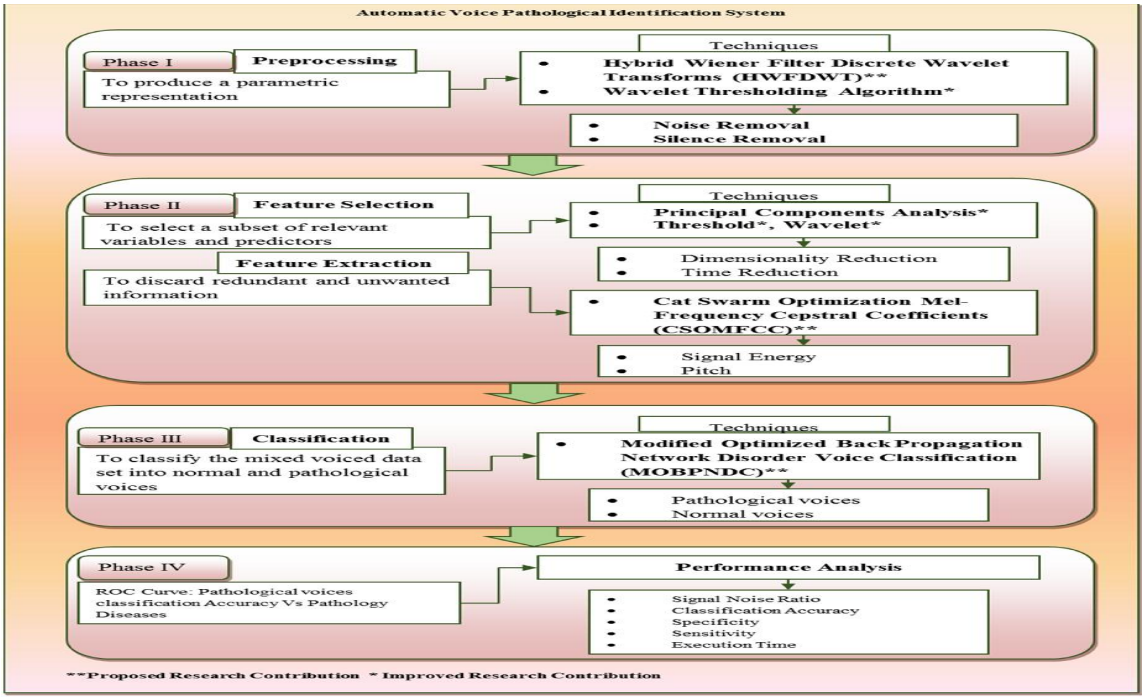

Figure 3.1 Proposed Research Design

Phase 1: Preprocessing

In order to investigate vocal pathology difficulties, we preprocess a wide variety of input speech signals. The preprocessing step is significantly impacted by noise cancellation and silence restoration methods. The Electro Grotto Graph (EGG) may then be subjected to acoustic analysis using the Wiener and DWT filters. An Electro Grotto Graph (EGG) is generated from each individual voice sample by first passing it through a Wiener filter and then a Discrete Wavelet Transform (DWT) filter. As part of this research, we created the Hybrid Wiener Filter Discrete Wavelet Transform (HWFDWT). Vocal pathology abnormalities are first diagnosed using a demising estimation. The proposed preprocessing approach has a higher computational cost than the present configuration. But where it truly comes into its own is in noise and silence restoration. Wiener filters may make use of linear time invariant (LTI) filtering to attempt to minimize the MSE between the estimated random process and the target process. The Discrete Wavelet Transform (DWT) was used to create the discrete wavelets that were collected by the different Fourier Transforms. Combining the "Wiener filter minimization" and "discrete wavelets sample" procedures yields a Hybrid Wiener Filter Discrete Wavelet Transform (HWFDWT).

ISSN: 2057-5688

Here, we separate the voices of the individuals speaking from the rest of the environment. Many studies have been conducted in the academic community on the topic of noise and silence cancellation. It has been suggested that we model both at once. In the presence of a lot of background noise, however, it was unable to correctly identify certain infected voices. The Hybrid Wiener Filter Discrete Wavelet Transform (HWFDWT) is a preprocessing technique that combines the Wiener Filter with the Discrete Wavelet Transform to solve this problem. Linear time-invariant (LTI) filtering in combination with a Voice Enhancing Threshold (VETh) may be used to accomplish Voice Demising in the Wiener filter. The input loud speech signal has a significant "Mean Square Error" all the time. Infrequently do we hear of this phenomenon being called a "voice noise spectrum?" A "Mean Square Error" threshold was applied to the affected person's voice in order to remove background noise. The MSE increases with signal noise level because noise acts as a signal amplifier. When the coefficient is exceedingly small, the Mean Square Error equals zero. Since the original input signal is kept intact, this technique is the gold standard for noise cancellation. Applying a Wiener filter to the noisy input speech signal may help us get closer to the target Demised speech and reduce the MSE.

The primary goal of the HWFDWT is to estimate a silence signal from a noisy speech signal by exploiting the silence signal's link to the speech signal. The Mean Square Error is calculated for each input signal. By removing noise from incoming signals, the MSE may be successfully mitigated. This statistical approximation is used in all filtering processes. We calculated the systematic pitch period and the approximate pitch period using the acoustic indices. In an Electro Glotto Graph (EGG), vocal fold vibrations are plotted versus perceived loudness. The strength of a voice signal varies greatly during the course of a single glottal cycle, as seen in this animated GIF. The creation of the Electro Glotto Graph (EGG) is impeded by the presence of noise in the voice input.

IV A PREPROCESSED REPRESENTATION OF THE PARAMETERS

Safekeeping of Private Information for the Time Being

The encrypted, real-time data set contains eighty voice samples of varied pitches. This study was painstakingly compiled by pathologists at Karpagam Nursing College and Karpagam Academy of Higher Education (KAHE) in Coimbatore, Tamil Nadu, and India. Karpagam Academy of Higher

ISSN: 2057-5688

Education is located in Coimbatore, Tamil Nadu, and India. It was established with approval from the Indian Ministry of Human Resource and Development under Section 3 of the UGC Act 1956. The data is quite helpful since it is separated by age and gender. To conduct our experiments, we gathered a total of 80 voice recordings, 20 each from men and women with typical and abnormal speech patterns. People with this disease produce vowels (/a/, /e/, /i/, /o/, and /u/) with a wide range of pitch, from low to high and back down again. The input is two seconds of 50 kHz, 16-bit resolution, vowel samples. People between the ages of 30 and 35 make up the bulk of the market. The total number of samples from the Real-time Dataset is shown in Table 4.2.

Table 4.2- Private Real-Time Database

Real-time Dataset	**Department of Pathology, Karpagam Faculty of Medical Sciences and Research, Coimbatore.**
No. of Samples	**80 samples; Male-20; Female-30;**
Recording Collections	**vowels /a/,/e/, /i/ /o/, /u/ and Phrases**
Minimum Duration	**2 seconds**
Frequency	**50 kHz**
Resolution	**16-bit**
Voice Types	**Normal – 20 pathology - 30**
Age	**30-35 years**

4.6 Acoustic Parameters

Some of the auditory characteristics that have been studied in relation to noise reduction include Mel filtering, jitter, shimmer, and pitch. An acoustic analysis of the vibration of the vocal folds demonstrates the aberrant voice quality. Breathiness, hoarseness, and roughness in the voice are investigated, and their possible causes are explored. The study takes into account much

ISSN: 2057-5688

interference to determine the robustness of a spoken signal. It's preferred that the voice have a few distinct styles. It's possible that Mel filtering, Jitter, Shimmer, and Pitch readings are all off. This exhibits the capacity to accurately estimate the quality of a voice. It is challenging to extend from this work to acoustic analysis due to issues with sensitivity, selectivity, and accuracy. Sensitivity, Selectivity, and Accuracy are often used as the major metrics of prediction in Mel filtering, Jitter, Shimmer, and Pitch computations. These settings may be fine-tuned in accordance with the fundamental frequency of the input speech stream.

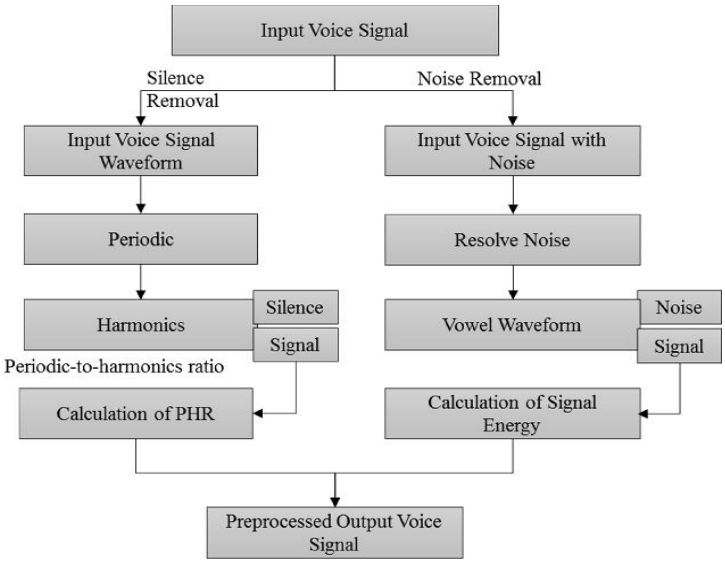

Figure 4.2. Preprocessed Voice Signal

The energy of the input speech signal and the periodic-to-harmonics ratio are two of the most important procedures in acoustic analysis, and both rely on this acoustic parameter. Any supplied voice warning signals will keep their original glottal waveform. It is common practice to exclude glottal noise when computing the periodic-to-harmonics ratio or the Signal Energy because of the peculiarities of the glottal waveform. Figure 4.2 provides a great example of this behavior. Both may be used to clean up audio samples of a person's voice that have been captured.

Input speech samples are analyzed using the Periodic-to-Harmonics Ratio (PHR). Equation 4.1 illustrates how the PHR is determined from the waveform of the input speech stream. This

ISSN: 2057-5688

forecast will go silent after about 25 glottal cycles. The equations we've already seen can be used to figure this out.

$$PHR = \frac{1}{Avg\left|\max_{Periodic} glottalnoise\right| - Avg\left|\max_{harmonics} glottalnoise\right|}$$

$$Glottal\ Noise = averagedwaveform - individualglottalwaveforms \quad (4.1)$$

The purpose of this PHR-based glottal noise estimate is to separate the signal from the background noise in the vowel waveform. We are measuring the intensity of this glottal-noise signal.

V CONCLUSION

The goal of this research is to provide a system for categorizing and treating vocal pathology issues based on their severity. Noise Removal and Silence Removal are performed after preliminary processing of various speech signals, such as the voice pathology disorder input. Together, the Wiener and DWT filters and the Electro Grotto Graph (EGG) were employed to get rid of them. The Hybrid Wiener Filter Discrete Wavelet Transforms (HWFDWT) technique was created for the goals of voice demising and pathologic voice prediction. It was recommended that the most important data be extracted from the presented pathologic speech sounds using Cat Swarm Optimization and Mel Frequency Cestrum Coefficients (CSOMFCC).

It has been shown that with cautious feature selection and extraction, it is possible to minimize both work and data size. The CSO algorithm improves the accuracy with which gender and abnormalities in voices may be identified. The features of the voice samples are captured and stored in a database after they have been used for training and testing purposes. When compared to MFCC and LPC analysis, Cat Swarm Optimization and Mel Frequency Cestrum Coefficients (CSOMFCC) provide more insightful results. Using the proposed approach CSOMFCC, the features were recovered in reduced dimensionality and in a shorter amount of time.

A Back Propagation Neural Network (BPNN) is used to classify the input audio signal as accurately as

ISSN: 2057-5688

possible. During the classification stage, speech samples are correctly labeled as normal or pathological using an Automatic speech Pathological Identification System. Making this call requires the usage of the MOD Optimized BPNC Disturbance Voice Classification (MOBPNDVC). MOBPNDVC accepts both healthy and sick voices as input, and then randomly generates male and female normal voices and male and female diseases.

Then, separate ROC curves are created for each clinical condition, such as laryngitis, diplophonia, dysphonic, laryngoscopes, and chordates. A modified version of the Support Vector Machine (SVM) was created for voice classification, using the Modified Optimized Back Propagation Network Disorder Voice Classification (MOBPNDVC) as its basis. The proposed MOBPNDVC, which makes use of recognized and modified Back propagation neural network models, achieved 97.79% accuracy on the Saarbrucken data set test model, and 97.50% accuracy on the Real-time Dataset of the Department of Pathology, Karpagam Faculty of Medical Sciences and Research, Coimbatore. This section presents and discusses the findings from the investigation of the planned Automatic Voice Pathological Classification System.

REFERENCES:

1. *Al Amrani, Y, Lazaar, M & El Kadiri, KE 2018, 'Random forest and support vector machine based hybrid approach to sentiment analysis based hybrid approach to sentiment analysis the first international conference on intelligent computing in data sciences', Procedia Computer Science, vol. 127, pp. 511-520, Available from: <https://doi.org/10.1016/j.procs. 2018.01.150>.*
2. *Ali, T, Spreeuwers, L, Veldhuis, R & Meuwly, D 2014, 'Biometric evidence evaluation: An empirical assessment of the effect of different training data',IET Biometrics, vol. 3, no. 4, pp. 335-346, DOI: 10.1049/iet-bmt.2014.0009*
3. *Amami, RD, Ben Ayed & Ellouze, N 2013, 'Adaboost with SVM using GMM super vector for imbalanced phoneme data', 2013 6th International Conference on Human System Interactions (HSI), Sopot, pp. 328-333,DOI: 10.1109/HSI.2013.6577843.*
4. *Atal, BS, Chang, JJ, Mathews, MV & Tukey, JW 1978, 'Inversion of articulatory-to-acoustic transformation in the vocal tract by a computersorting technique', J. Acoust. Soc. Am. vol. 63, no. 5, pp. 1535-1555, Available from: <https://doi.org/10.1121/1.381848>.*

ISSN: 2057-5688

5. *Bargarai, F, Abdulazeez, A, Tiryaki, V & Zeebaree, D 2020, 'Management of wireless communication systems using artificial intelligence-based software defined radio', International Journal of Interactive Mobile Technologies, vol. 14, no. 13, pp. 107-133, Available from: <https://doi. org/10.3991/ijim.v14i13.14211>.*

6. *Ben-David, A 2007, 'A lot of randomness is hiding in accuracy', Eng. Appl.Artif.Intell. vol. 20, no. 7, pp. 875-885, Available from: <https://doi.org/10.1016/j.engappai.2007.01.001>.*

7. *Benesty, J, Chen, J, Huang, YA & Doclo, S 2005, 'Study of the wiener filterfor noise reduction, speech enhancement', Signals and CommunicationTechnology, Berlin, Heidelberg, pp. 9-41, Available from: <https://doi.org/10.1007/3-540-27489-8_2>*

8. *Boughorbel, S, Jarray, F & El-Anbari, M 2017, 'Optimal classifier for imbalanced data using Matthews correlation coefficient metric', PLoS One, vol. 12, no. 6, Available from: <https://doi.org/10.1371/journal.pone.0177 678.>*

9. *Breiman, L 2001, 'Random forests', Machine Learning, vol. 45, no. 1, pp. 5-32, DOI: 10.1023/A: 1010933404324.*

10. *Breiman, L 2003, Manual–Setting Up, Using, and Understanding Random Forests, v 4.0, Available from: <ftp://ftp.stat.berkeley.edu/pub/users/ breiman. Using_random_forests_V3. 0. Pdf>.*

11. *Breiman, L, Friedman, JH, Olshen, RA & Stone, CJ 1984, Classification And Regression Trees (1st ed.), Belmont, CA: Wadsworth, International Group, Available from: <https://doi.org/10.1201/9781315139470>*

12. *Browman, CP & Goldstein, L 1992, 'Articulatory phonology: An overview,Phonetica', vol. 49, no. 3-4, pp. 155-180, Available from: <https://doi.org/ 10.1159/000261913>.*