International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Generative Adversarial Network with Convolutional Wavelet Packet Transforms for Automated Speaker Recognition and Classification

Venkata Subba Reddy Gade[1]
Research Scholar
Dept.of ECE, Sathyabama Institute of Engineering and Technology
Chennai, India
Associate Prof, Dept. of ECE, Gokaraju Rangaraju Institute of Engineering and Technology ,Hyderabad, India
Email:gvsreddy2005@gmail.com

Dr. M Sumathi[2]
Prof, Dept. of ECE, Sathyabama Institute of Science and Technology
Chennai, India
Email:sumagopi206@gmail.com

Abstract—Speech is an effective mode of communication that always conveys abundant and pertinent information, such as the gender, accent, and other distinguishing characteristics of the speaker. These distinctive characteristics allow researchers to identify human voices using artificial intelligence (AI) techniques, which are useful for forensic voice verification, security and surveillance, electronic voice eavesdropping, mobile banking, and mobile purchasing. Deep learning (DL) and other advances in hardware have piqued the interest of researchers studying automatic speaker identification (SI). In recent years, Generative Adversarial Networks (GANs) have demonstrated exceptional ability in producing synthetic data and improving the performance of several machine learning tasks. The capacity of Convolutional Wavelet Packet Transform (CWPT) and Generative Adversarial Networks are combined in this paper to propose a novel way of enhancing the accuracy and robustness of Speaker Recognition and Classification systems. Audio signals are dissected using the Convolutional Wavelet Packet Transform into a multi-resolution, time-frequency representation that faithfully preserves local and global characteristics. The improved audio features better precisely describe speech traits and handle pitch, tone, and pronunciation variations that are frequent in speaker recognition tasks. Using GANs to create synthetic speech samples, our suggested method GAN-CWPT enriches the training data and broadens the dataset's diversity. The generator and discriminator components of the GAN architecture have been tweaked to produce realistic speech samples with attributes quite similar to genuine speaker utterances. The new dataset enhances the Speaker Recognition and Classification system's robustness and generalization, even in environments with little training data. We conduct extensive tests on standard speaker recognition datasets to determine how well our method works. The findings demonstrate that, compared to conventional methods, the GAN-CWPTs combination significantly improves speaker recognition, classification accuracy, and efficiency. Additionally, the suggested model GAN-CWPT exhibits stronger generalization on unknown speakers and excels even with loud and poor audio inputs.

Keywords-Automated Speaker Recognition, Deep learning, Artificial intelligence Generative Adversarial Networks (GANs), Convolutional Wavelet Packet Transform (CWPT).

I. INTRODUCTION

Speech is the primary and inborn mechanism of human contact, efficiently and quickly transmitting important information. People devote time and effort to learning to use voice commands to interact with smart devices. In all, 7097 live languages have been shown to exhibit distinctive speech patterns [1]. At least one person uses a living language as their primary communication. Less than 23 languages, including Spanish, English, Chinese, and Hindi, are spoken by more than 50% of the world's population. Given that more languages are spoken on Earth than people, this number continues to be astounding. However, these few languages have a significant influence. They are essential to the development of AI disciplines, including Text-To-Speech (TTS), Computational Linguistics (CL), Natural Language Processing (NLP), and Automatic Speech Recognition (ASR).In contrast, it might be difficult for less widely used languages to get financing for specialized technological research and development [2]. Creating similar solutions for languages with limited resources is a difficult but necessary task.

ASR, also known as automatic speech recognition, has many applications in security, education [3], smart healthcare [4], and smart cities [5], making it a hotly debated and researched topic. Multiple algorithms are used in this process to correlate textual patterns with recognized speech signals, translating spoken language into written text [6]. The primary objective of automatic speech recognition (ASR) is to transform audio signals into text using a robust framework that integrates state-of-the-art semantic learning techniques. ASR integrates

several fields, including computer science, linguistics, AI, digital signal processing, acoustics, and statistics.

Recent advancements in ASR have been made possible by implementing different DL methods. Successful automatic speech recognition (ASR) engines have been built for European, English, and Asian languages by major computer companies like Microsoft, Apple, Amazon, Facebook, Google, and IBM. The advancement of automatic speech recognition (ASR) systems for Central Asian languages like Uzbek remains in its infancy. dialectal differences and the lack of a standardized Uzbek speech corpus mainly cause this. One of the key challenges is the lack of speech corpora needed to create an ASR system for Central Asian languages. Test results may be significantly improved with an appropriate classification methodology for the ASR system, and computing complexity may increase. One of the challenges in creating an ASR system is finding an appropriate speech corpus and acquiring the necessary data for training and testing the system. These training datasets are now only available for about 7,000 of the world's most widely spoken languages [7].

Generative Adversarial Networks (GANs) have developed as a game-changing way to learn complex data distributions and create realistic samples. The two neural networks that make up a GAN, the discriminator and the generator, engage in a strategic game of rivalry to enhance the quality of the data they generate. GANs, first introduced for picture synthesis, have been effectively extended to various domains, including speech and audio processing, powering rapid advances in automated speaker detection and categorization. The key concept behind employing GANs for automatic speaker detection and classification is to use their ability to capture underlying data distributions and generate synthetic but highly representative speech samples. In this technique, Generative Adversarial Networks (GANs) play an important role in augmenting existing training data, effectively solving the obstacle of insufficient labeled datasets—a prevalent problem in speech recognition systems. Furthermore, GAN-based algorithms provide greater generalization, increased resilience in noisy situations, and adaptability to different speech accents and languages.

The Continuous Wavelet Packet Transform (CWPT) is at the forefront of innovative techniques for extracting features in voice recognition and classification problems. CWPT, which serves as an improved iteration of traditional wavelet transformation methods, combines the strengths of both wavelets and convolutional neural networks (CNNs), resulting in a powerful tool for interpreting and representing complex audio data. By integrating wavelets' time-frequency analysis capabilities with CNNs' hierarchical learning capabilities, CWPT aims to improve feature discrimination, resulting in more precise and robust speaker identification.

This study provides a ground-breaking mechanism for speaker classification and automatic recognition. The technique dramatically improves the accuracy and efficiency of the process by combining Generative Adversarial Networks (GANs) and CWPT.Our innovative framework addresses the drawbacks of conventional speaker identification algorithms, which commonly fail to distinguish speakers accurately in challenging situations, including background noise, varying emotional states, and insufficient training data.

This study's primary objective is to show how the recommended approach may teach resilient and discriminative skills for speaker identification, even in challenging circumstances. Our strategy intends to dramatically improve the performance and generalization of automated speaker recognition systems by combining the GAN's capacity to produce realistic samples with the CWPT's powerful feature extraction capabilities.

The following is a summary of this paper's significant contributions:

- To improve the precision and robustness of Speaker Recognition and Classification systems, GAN - CWPT are merged in a novel way.
- Audio signals are dissected using the CWPT into a multi-resolution, time-frequency representation that faithfully preserves local and global characteristics.
- To improve audio features, better precisely describe speech traits and handle pitch, tone, and pronunciation variations that are frequent in speaker recognition tasks. Using GANs to create synthetic speech samples, our suggested method GAN-CWPT enriches the training data and broadens the dataset's diversity.
- The new dataset enhances the Speaker Recognition and Classification system's robustness and generalization, even in environments with little training data.
- Conduct extensive tests on standard speaker recognition datasets to determine how well our method works.
- The findings show that incorporating GAN-CWPTs significantly improves speaker recognition and classification efficiency and accuracy compared to standard methods.

The remainder of this essay is organized as follows: The literature survey of speech processing is examined in Section 2. Section 3 expands on the proposed methodology by delving into the architecture of the GAN-CWPT. Section 4 explains the experimental setup and assessment measures used to assess the performance of our model. Section 5 concludes by summarizing our contributions and outlining potential future research possibilities.

II. LITERATURE SURVEY

Ismail et al. [8] produced speech datasets in English and Urdu with five distinct regional accents used in GB, a region in northern Pakistan. These datasets are meant to promote and advance speaker recognition system research. The voice datasets comprise 7200 speech samples from 180 speakers, with identifying information, specific words, and 10- to 16-digit numeric strings. Four machine learning methods were trained on pre-processed datasets to extract speech attributes: RF, ANN, SVM, and KNN. With accuracy improvements of 88.53% and 86.58%, respectively, the ANN classifier model outperformed RF, SVM, and KNN, according to the English and Urdu datasets analysis.

Mokgonyane et al. [9] developed a text-independent speech detection system based on machine learning. The research involved many procedures, including testing and training. These procedures included feature extraction, model training, evaluation, speech activity detection, and a graphical user interface. The Sepedi speech dataset was made available by the NCHLT project. The Long-Term Spectral Divergence approach was used to identify speech activity, and the pyAudio Analysis program was used for feature extraction. Weka's SVM, KNN, RF, and MLP implementations were used to train the models.

Kamiski et al. [10] developed the ASR System to address concerns with speaker identification in open set settings and speaker verification in difficult recording circumstances such as telephone communications. According to their research based on a validated voice dataset, the developed speaker recognition system outperformed rival systems in both speaker identification and verification tasks. The internal settings and features of the ASR System are optimized via genetic algorithms. Gaussian mixture model feature creation and categorization had an impact on this as well.

Dhakal et al. [11] developed a new high-speed pipelined architecture for real-time speaker detection. To address this issue, the proposed intelligent system accurately recognizes approved users. GF, CNN, and statistical considerations all aid in extracting features. According to our investigation, the feature extraction methods already described and RF proved to be the most effective at obtaining and classifying speaker recognition features. Success in these feature extraction and categorization processes will determine how accurate our suggested design is. RF outperformed the other two voice recognition systems in rigorous testing with many datasets.

Ye et al. [12] developed the topic of speaker identification research. We created a deep RNN model with a 2-D CNN layer that incorporates rich voiceprint information from speech spectrograms and the efficiency of 2-D CNN in extracting features from 2-D structures. This method creates a robust identification system by fusing the 2-D CNN's feature extraction capabilities with the GRU cell units' temporal dynamics. Using GRU cell units and cyclic memory learning, we integrated time series into deep RNN networks to capture the unique characteristics of each speaker hierarchically. A softmax classifier layer that learns and distinguishes speaker traits is present in the top layer.

Ibrahim et al. [13] investigated the application and significance of voice recognition systems. The report looked at research on automated speech recognition and voice recognition systems. According to an assessment of the literature, the HTK system, developed by a Cambridge University team led by Steve Young, is the most well-known software program for automatic speech recognition. Sarma et al. [14] recommended utilizing emotion-invariant speaker embedding to transform i-vectors holding speaker-specific information into an emotion-invariant space. Regarding accuracy, the suggested strategy fared better than a framework that used a regular speaker model and a spectrum of emotions.

Shafik et al. [15] used radon transforms and spectrograms of audio signals to create a CNN-based model for speech recognition. The model maintains high precision even when the signal is corrupted by external factors such as musical interference or the speech of another speaker

Costantini et al.'s study [16] explores high-level AI approaches for speaker detection without using speaker-specific models. A shallow, custom CNN architecture outperforms AlexNet trained on the ImageNet dataset (90.15% accuracy) among the several CNN architectures studied. The most accurate are grayscale spectrograms, which even exceed MFCC graphs. Despite being a significantly lighter model, its accuracy falls short of a Nave Bayes trained on chosen acoustic features at 87.09%. Pitch/F0, MFCC, and voicing likelihood have been determined to be the most useful acoustic variables for categorizing various domains.

Li et al. [17] of Microsoft Speech and Language Group developed an RNN-T model encoder that uses CTC or Cross-Entropy (CE) training. The WER was decreased by 11.6% when the RNN-T encoder was initialized with CE, but it was increased by 12.8% when the future context model was compared to the zero-lookahead model.Transcripted Microsoft data totaling 65,000 hours made up the model's training data. However, Khassanov et al. [18] produced a Kazakh language voice corpus with more than 332 hours of audio transcribed and more than 153,000 utterances from people of all ages, genders, and regions. They used the text in Kazakh from various sources, including Wikipedia, online journals, blogs, and laws, to construct this corpus. A web-based speech-recording program that can be used from desktop computers and mobile devices was then used to narrate these sentences. Earlier research focused on creating speech recognition for the Uzbek language.

2.1 Limitations for Existing system

- Speaker identification algorithms rely substantially on vast amounts of high-quality labeled data for training. However, getting such data can be difficult and costly, particularly for narrow or specific speaker categories. The model's accuracy and generalization may need to be improved due to limited and biased training data.
- Speaker identification models trained on a single language may need help generalizing to different languages or dialects. Phonetic structures, accents, and speaking styles can all substantially impact the system's performance.
- Background noise, reverberations, and other acoustic fluctuations in real-world contexts can reduce the accuracy of speaker recognition systems. The existence of such noise might result in false positives or false negatives, lowering system reliability.
- Individuals' voices can alter owing to various circumstances, such as age, health, emotions, or the situation in which they speak. The system must account for intra-speaker variability for successful classification while maintaining inter-speaker distinctiveness.
- The distribution of speakers across classes may need to be more balanced in many speaker recognition programs. This can result in biased models that

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

perform better on majority speakers while failing on minority ones.

- Spoofing attacks on automated speaker identification systems occur when malicious individuals attempt to mimic or trick the system using voice recordings or speech synthesis techniques. Creating effective anti-spoofing techniques is a continuing issue.

2.2 Problem Identification for Existing system

- One of the most challenging tasks in automated speaker recognition and classification is effectively identifying and distinguishing distinct speakers based on their speech characteristics. To give consistent results among varied speakers, the system must be able to handle changes in speech patterns, accents, and emotional states.
- Real-world circumstances frequently bring environmental elements such as background noise, reverberation, and channel distortions, all of which can substantially impact the quality of audio samples. The system must be strong enough to withstand these conditions and accurately recognize speakers even in difficult acoustic situations.
- As the number of speakers and audio samples increases, managing a large-scale speaker database efficiently becomes increasingly important. The system should be able to handle massive data storage, retrieval, and indexing operations while also ensuring efficient and precise search and categorization processes.
- In some applications, such security systems and call center operations, real-time speech detection and classification are essential. In order to meet the demands of such time-sensitive scenarios, the system must have low latency and quick response times.
- Automated speaker identification systems commonly deal with sensitive data, such as voiceprints or biometric information. It is essential to protect the security and privacy of this data in order to avoid misuse, unwanted access, and potential data breaches.
- The system should be able to recognize and classify speakers of different languages in a worldwide context. While taking into consideration differences in pronunciation, intonation, and other language-specific characteristics, it must handle the difficulties of cross-linguistic and multilingual speaker recognition.

III. PROPOSED SYSTEM

This section explains GAN-CWPT, a revolutionary method for improving Speaker Recognition and Classification systems' precision and robustness. The CWPT is used to segment audio data into a multi-resolution, time-frequency representation that correctly preserves both local and global properties. As a result of the improved audio features, speaker recognition tasks can more precisely account for variations in pitch, tone, pronunciation, and other speech characteristics. Our suggested method, GAN-CWPT, uses GANs to create artificial speech samples, which enhances the training data and broadens the dataset. The generator and discriminator components of the GAN architecture have been altered to produce realistic voice samples with traits very similar to genuine speaker utterances. Thus, the new dataset improves the Speaker Recognition and Classification system's robustness and generalization even in contexts with minimal training data. We extensively test well-known speech recognition datasets to gauge how effectively our approach performs.

The GAN-CWPT technique's block diagram is shown in Fig. 1. First noticed from the speech input signal. Pre-processing is essential in systems when background noise or calm is undesirable. Effective feature extraction techniques from speech signals, the majority of which contain speaker-related information, are required by methods such as SI and Speech Recognition (SR). It has been demonstrated that Generative Adversarial Networks (GANs) are extremely proficient at producing fictional data and improving the efficiency of various machine-learning tasks. Convolutional Wavelet Packet Transform (CWPT) and Generative Adversarial Networks (GANs) produce Speaker Recognition and Classification systems with increased precision and durability. In actual use, feature extraction reduces the data dimensions of spoken signals while maintaining the integrity of essential information. Front-end processing, including feature extraction from speech signals, is carried out during the training and recognition phases. In feature extraction, sets of numerical descriptors or feature vectors that contain essential aspects of the speaker's voice are extracted from digital speech signals. Speech signals include various information, not all required for speaker identification. Finally, the decision of voice is accepted or rejected.

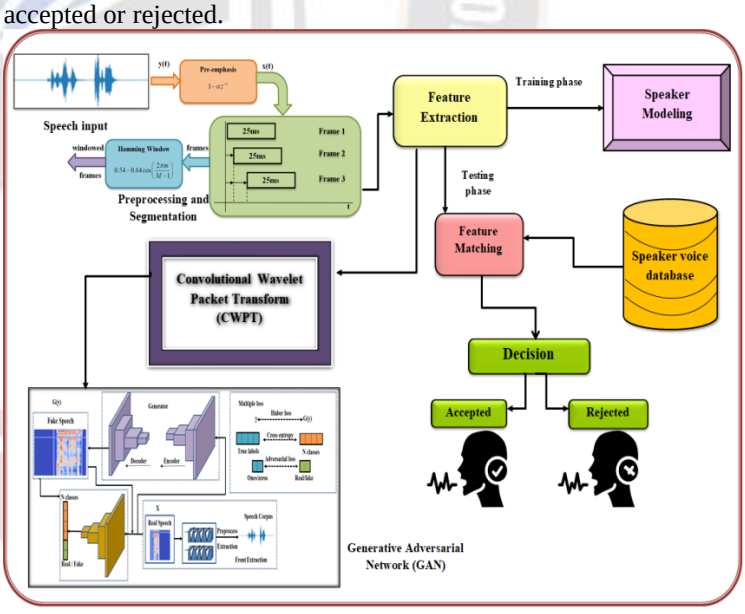

Figure 1: block diagramof GAN-CWPT method

3.1. Dataset

The National Centre provided the dataset for this investigation for the Human Language Technology (NCHLT) project of the Language Resource Management Agency [19]. The collection includes audio recordings of Sepedi voices made by various

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

speakers. Apiece of the 50 speakers that participated in our data collection provided 100 audio samples with 3-5 words apiece. As can be seen in Table 1, a total of 5000 audio files were used, totalling 294.6 minutes in length.

Table 1: Summary of the 100 Speaker Dataset

Unit	Value
Size	548 MB
Duration	294.6 minutes
Instances	5000

3.2 Speech pre-processing and segmentation

We provide an overview of the many pre-processing techniques researchers employ in various speaker identification domains in this section. In systems where background noise or stillness is not required, efficient feature extraction methods are essential. This is crucial for voice identification and speech recognition (SR) systems, mainly since a large amount of the uttered segment carries properties specific to the speaker [20]. The various pre-processing techniques used in different speaker identification research projects are described in depth in the next section.

3.2.1. Silence removal

There may be silence at several moments during the speech signal, such as at the beginning, in between the syllables of the sentence, and at the conclusion. Voice signals' unspoken portions are cut out to reduce processing time and complexity. Unspoken components must therefore be eliminated before continuing with the process. The unknown parts of a voice signal are successfully eliminated by utilizing statistical background-noise features to categorize each sample as either uttered or unuttered. Each voice sample's mean and standard deviation are computed as surveys:

$$\mu = \left(\frac{1}{M}\right)\sum_{k=1}^{N} y(t) \tag{1}$$

$$\sigma = \sqrt{\left(\frac{1}{M}\right)\sum_{k=1}^{N}\left(y(t)-\mu\right)^2} \tag{2}$$

Where the speech signal, mean, and standard deviation are represented by y (t), background noise is described by Equations 1 and 2. The sample is regarded as spoken if the one-dimensional Mahalanobis distance function, or $(|y-\mu|/\sigma) \geq 3$ for each sample, is greater than zero.

3.2.2. Pre-emphasis

Pre-emphasis is a filtering technique that emphasizes higher frequencies in spoken input. This technique attempts to equalize the frequency spectrum of spoken sounds, which often exhibits a steep fall in the upper-frequency area. The glottal source has an octave slope of roughly 12 dB for spoken sounds, but acoustic energy from the lips generates a +6 dB/octave increase in the spectrum. As a result, when a voice is recorded by a microphone that is situated at a distance, the recorded spectrum differs from the genuine vocal tract spectrum by roughly -6dB/octave. The use of pre-emphasis helps to lessen some of the overt glottal effects that are present in spoken speech. The pre-emphasis filter's most well-known substitute is.

$$H(z) = 1-\alpha z^{-1} \tag{3}$$

The range from -0.97 to 1 determines the pre-emphasis filter's slope. Both the overall energy level and the distribution of energy across various frequencies are altered by this filter. This may significantly impact the acoustic qualities related to energy.

3.2.3. Framing

As depicted in Fig. 2, signal framing is a technique for splitting a continuous speech signal into fixed-length chunks. Because the signal is nonstationary, the speaker's characteristics can change while speaking. In comparison, speech signals are projected to be steady for only 20 to 30 milliseconds. By dividing the signal into frames, it is possible to identify this stability and retrieve the pertinent acoustic properties. Furthermore, overlapping two successive frames makes it feasible to keep the information between them. There is a ten-millisecond gap between each succeeding frame, with each frame typically lasting 25 milliseconds.

3.2.4. Windowing

Due to their non-stationary nature and the changing statistical properties they display over time, voice signals cannot be subjected to DFT. These traits consist of random alterations in the vocal tract, prosody changes, and spectral patterns. However, the statistical characteristics of speech signals remain constant for most phonemes within brief periods of 10–20 ms. As a result, it is possible to apply standard signal processing techniques inside these time intervals, known as frames made up of N samples. Speech processing systems frequently split signals into overlapping frames and apply a Hamming window to each frame using Eq. 4 to extract information. Any potential spectral aberrations are lessened as a result.

$$w(m) = 0.54 - 0.64\cos\left(\frac{2\pi m}{M-1}\right), 0 \leq m \leq M \tag{4}$$

The amount of examples in each frame is indicated by the "M." Below is a description of the windowed speech signal's output:

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

$$y(m) = y(m)w(m) \tag{5}$$

The most widely used windowing functions are hanging half parallelograms and triangles.

3.2.5. Endpoint detection

This technique entails separating speech signal fragments from an ambient noise background. This background noise has a normal distribution because it is classified as white noise. A mathematical viewpoint is that.

$$p(y) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right) \tag{6}$$

The initial 3200 voice samples of a spoken phrase designated as background noise are used to compute the parameters σ and μ in the given equation. As a result, word Y is classified as a component of the white-noise distribution, and since $(|y-\mu|/\sigma) \geq 3$ is still valid for each speech sample x, it can be indifferent from the spoken phrase.

3.2.6. Speech signal normalization

Using Eq. 7, normalization makes voice signals equivalent regardless of magnitude variances.

$$s_{Mi} = \frac{s_i - s}{\sigma} \tag{7}$$

Let s_i be the ith component of signal s, σ and s indicate its mean and standard deviation, respectively, and s_{Mi} denotes the i^{th} element of the movement s that has been normalized.

3.2.7. Spectrogram

Using the short-term Fourier transform, speech samples are transformed into spectrograms to extract discriminative properties automatically. In two dimensions, an energy amplitude is shown visually in a spectrogram. The x and y axes indicate the time and frequency domains. The color of each point in the visualization, shown in Figure 2, denotes the energy amplitude at a particular instant. A spoken signal is initially separated into frames before being analyzed. Each frame is given a Hamming window, which produces a spectrogram. The fast Fourier transform (FFT) is then used to transform these windowed frames from the time domain to the frequency domain. In the frequency-domain representation, band-pass filters are used. The Mel-scale is used to distribute these filters, and the center frequency of each filter has been adjusted accordingly. Each band-pass filter's output is given a logarithmic function to decrease the dynamic range. The voice signal spectrogram is created by merging the results of these operations frame by frame.

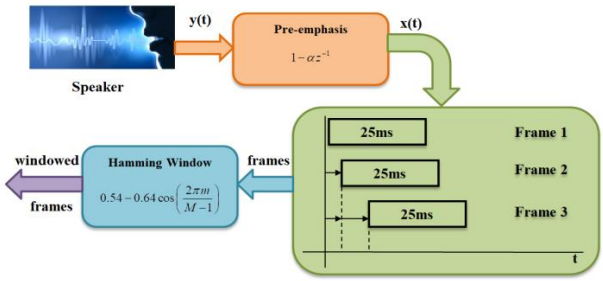

Figure 2: Speech Signal Framing and Windowing

3.3. Convolutional wavelet packet transforms

The CWPT, a powerful signal processing tool, combines the principles of convolution and wavelet packet transforms to analyze and represent data in both the temporal and frequency domains. Data compression, image analysis, and signal processing applications benefit most from this disruptive approach. The Continuous Wavelet Packet Transform (CWPT) employs a well-built array of convolutional filters to separate unique properties or motifs within a signal. The CWPT excels in detecting even the smallest variations and subtleties in the signal via these filters, allowing for specialized analysis. The CWPT develops a hierarchical arrangement of sub-bands having a tree-like structure by merging the convolution outputs with appropriate weightings that include varied scales and orientations. Unlike standard wavelet transformations, the CWPT, like traditional wavelet packet transformations, allows for signal investigation at many resolutions. However, its distinguishing feature is the use of convolutional filters, which augment the captured spatial information. This enhancement is especially useful for jobs requiring a more detailed spatial comprehension, such as image processing. In this way, the CWPT provides a more comprehensive representation of the signal, outperforming standard wavelet transformations. Notably, the CWPT distinguishes itself by its ability to capture signals with localized and anisotropic properties, which is a hurdle for many other transformation approaches. This adaptability extends across multiple data kinds, from pictures and audio signals to time-series data, making the CWPT a versatile instrument suitable for a wide range of applications. In conclusion, the CWPT cleverly combines convolution and wavelet packet transforms to create a robust and versatile approach to signal processing.It is a crucial tool in contemporary signal processing and data analysis because of its ability to capture fine details and applicability across numerous domains.

3.3.1. Sub-band based wavelet parameters

The inner product of the signal y (t) and mother wavelet $\psi(t)$ is computed to produce the wavelet transform.

$$\psi_{c,d}(t) = \psi\left(\frac{t-d}{c} \right) \tag{8}$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

$$W_{\psi} y(c,d) = \frac{1}{\sqrt{c}} \int_{-\infty}^{+\infty} y(t)\psi^{*}\left(\frac{t-d}{c}\right)dt \quad (9)$$

The variables c and d stand in for the scale and shift parameters, respectively. Users can either shift or expand the mother wavelet by changing these variables.

The wavelet, or the discrete wavelet transform (DWT), divides the signal into discrete domains to perform dyadic multi-resolution analysis (MRA). The discrete wavelet family's scale and translation parameters are identified in the DWT architecture by the integer's j and k, respectively. The family's collection of discretized parameter functions receives modifications.

$$c = c_0^j \quad (10)$$

$$d = kd_0c_0^j \quad (11)$$

$$\psi_{i,k}(k) = c_0^{-j/2}\psi(c^{-j}t\_kd_0) \quad (12)$$

$\psi_{j,k}(t)$ is referred to as the DWT base in the equation. The transform's time variable is still continuous despite the name. The DWT coefficients of a continuous time function are similarly defined as

$$e_{j,k}\left\langle f_v(t)\psi_{j,k}(t)\right\rangle = \frac{1}{c_0^{j/2}}\int f_v(t)\psi(a_0^{-j}t - kb_0)dt \quad (13)$$

Once the transformation is finished, it becomes clear how a function's $f_v(t)$ wavelet representation is expressed.

Discrete Wavelet Transform (DWT) signals alter data using high-pass and low-pass filters. After passing through a high-pass filter, the high-frequency components of vocal signals are kept as "details." Like the high-frequency components, the low-frequency components, which are kept as "approximations," can only be gradually dissected by repeated processes.

$$f_v(t) = \sum_j \sum_k \left\langle f_v(t)\psi_{j,k}(t)\right\rangle \psi_{j,k}(t) \quad (14)$$

Their most important feature is the low-frequency content of voice transmissions, which allows for signal identification. Frequency content can provide depth offlavour. The voice will change if the high-frequency speech signal components are removed, but speech can still be understood. The wavelet packet transform (WPT) uses an iterative binary tree approach to dissect the audio stream. The sole difference between the WPT and DWT is that the WPT decomposes approximations and details rather than just approximations. The core idea behind wavelet packets (WP) is that they use a pair of dual filters with low-pass and high-pass characteristics when given a signal. These filters separate the initial signal's sub-band frequency components into two sequences. The two orthogonal wavelet bases from the previous node has the subsequent names.

$$\psi_{j+1}^{2p}(t)\sum_{n=-\infty}^{\infty} h[n]\psi_j^p(k-2^j n) \quad (15)$$

$$\psi_{j+1}^{2p}(t)\sum_{n=-\infty}^{\infty} g[n]\psi_j^p(k-2^j n) \quad (16)$$

The wavelet function is denoted by v[n] in Eqs. (15 and 16). The number of decomposition levels and prior node nodes are represented by the symbols j and p, respectively. The Wavelet Packet Transform (WPT) and the Discrete Wavelet Transform (DWT) are used in this study's feature extraction process. However, due to the data's extreme length, classifiers cannot effectively use it. We must therefore look for a more accurate way to capture the speech features.

3.3.2. Energy index of the sub-band signals

It is common practice to estimate voice energy to enhance the depiction of sub-band signals. According to an earlier study, an energy index can be a helpful element in recognition tasks. Previous research has shown the value of using the energy index of a particular sub-band signal as a characteristic to recognize digital modulation in a biological setting. This investigation will partition the signal's energy into several resolutions, followed by an assessment of these indices. Mathematically

$$P_j = \frac{1}{N}\sum_k |v_{j,k}|^2 = \frac{\|v_j\|^2}{N_j} \quad (17)$$

Where $\|v_j\|$ denotes the norm of the expansion coefficient v_j

.3.4. Generative Adversarial Network (GAN)

While the basic GAN and convolutional GAN both strive to produce high-quality output samples, the discriminator's primary responsibility is to assess whether the inputs are valid. GANs are an obvious candidate for classification jobs due to their high content generation performance. When there is a lack of training data, the model's generalizability can be improved by using samples generated by the $p_g(y)$ that closely approximate the real data distribution. This study tests the CGAN's classification abilities by giving it access to more unlabelled examples. Figure 3 presents the suggested GAN framework.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

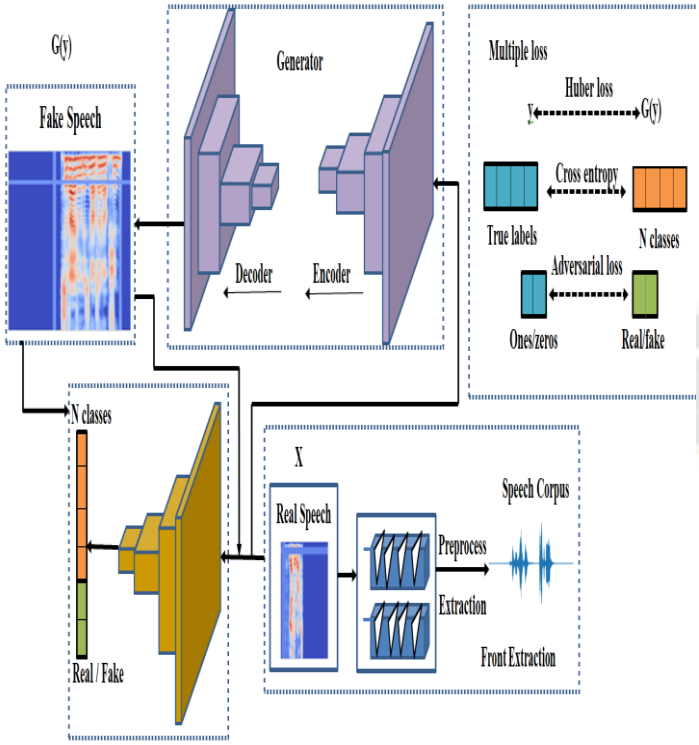

Figure 3: Framework of GAN

The "real/fake" output of a GAN discriminator is produced. A vector $l = \{l_1, ...l_N\}$ with N dimensions created by the standard classifier section, and l_k represents the probability $p_{model}(x = k \mid y)$ that the input y belongs to class k. The GAN enables the classifier to accept samples from the generator as input and produce N+1 units as output, with l_{N+1} reflecting the probability $p_{model}(x = N = 1 \mid y)$ that the inputs are authentic, this integrates the discriminator and classifier. As a result, the loss function is divided into two parts: an adversarial loss L_{adv} and a classification loss L_{class}.

$$L_{class}(Dis) = -E_{yx \sim p_{data}(y,x)} \sum_{k=1}^{N} x_k \log l_k \quad (18)$$

$$L_{class}(Gen, Dis) = -E_{y \sim p_g(y)}[\log p_{model}(x = k+1 \mid y)] - E_{y \sim p_g(y)}[\log[1 - p_{model}(x = k+1 \mid y)]] \quad (19)$$

L_{class} Incorporates all labeled data and indicates the cross-entropy loss using the label vector y from a training sample. The labels of the genuine original samples that served as their foundation are passed down to the freshly created samples. L_{adv} stands for the adversarial loss when attempting to grasp the target distribution. The strategy uses the least square GAN (LSGAN) technique to improve training stability and handle the problem of gradient vanishing [21]. The adversarial loss L_{adv} is then broken down into separate generator and discriminator components.

$$L_{adv}(Dis) = E_{y \sim p_{data}(y)}[(D(y)-1)^2] E_{y \sim p_{data}(y)}[(D(G(y))^2] \quad (20)$$

$$L_{adv}(Gen) = E_{y \sim p_{data}(y)}[D(G(y))-1)^2] \quad (21)$$

The conventional GAN can provide outputs of varying quality. The learned data distribution when more than one mapping complies with the GAN framework and the adversarial loss is insufficient to preserve context gained from genuine inputs, errors may occur. The Huber loss is used to guarantee an accurate depiction of the actual data. The Huber loss functions as a strong regression loss function that can tolerate outliers in the data better than the squared error loss. The Huber loss can be defined as follows, given an input of dimensions H*W and its corresponding generated output G(y) of the same size:

$$L_{Huber}(Gen) = \frac{1}{H.W} \sum_{i=1}^{H} \sum_{j=1}^{W} \delta(i, j) \quad (22)$$

$$\delta(i, j) = \begin{cases} \frac{1}{2}[y_{ij} - G(y)_{ij}]^2, & x_{ij} - G(y)_{ij} \le 1 \\ |\, y_{ij} - G(y)_{ij}\, |, & otherwise \end{cases} \quad (23)$$

Where the residual is known as $y_{ij} - G(y)_{ij}$ and the residual threshold is modifiable.

The following is a list of the overarching goals that the generator and discriminator must minimize:

$$L(Gen) = L_{adv}(Gen) + L_{Huber}(Gen) \quad (24)$$

$$L(Dis) = \lambda_1 L_{adv}(Dis) + \lambda_2 L_{class}(Dis) \quad (25)$$

Where λ_1 and λ_2 are trade-off variables that support categorization and unlabeled learning, respectively.

3.4.1. Network architecture

GAN employs a variety of generator and discriminator designs. The generator is always built in an encoder-decoder configuration, as illustrated in Fig. 4. The generator should be able to encode both temporal and spatial data and provide excellent feature sequences upon decoding. The discriminator must handle enormous amounts of training data to approximate speaker sequencing. The acquisition of the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

feature sequences corresponding to the speaker's voice follows. The generator creates synthetic samples with lengths that match these real sequences. Convolutional and shuffler layers are used in a series of iterative stages that involve downsampling and upsampling to achieve this. The discriminator then distinguishes between real and synthetic samples, classifying them into separate groups, using the created samples and actual acoustic properties from the dataset.

3.4.1.1. Generator design

The generator's objectives are the reproduction of speaker feature sequences and the identification of correlations from inputs. Some GANs use generators made up of interconnected layers or essential convolutional layers for specific speech-related tasks, including voice conversion and language detection. The samples produced by these generators are unreliable because they only capture relationships between feature dimensions. Although it takes time because of how sophisticated parallel computation is, the RNN is an excellent method for solving this issue. Considering these elements, we build the generator utilizing gated CNNs [22], allowing for the insertion of sequentialstructure and promoting rapid convergence.

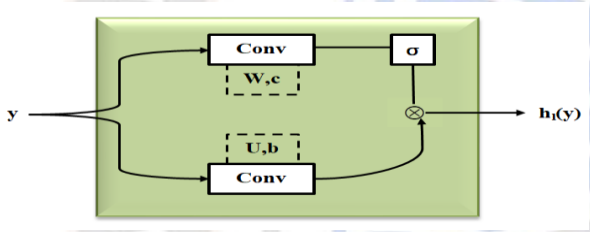

Figure 4: The architecture of the gated CNN block

The gated CNN block, which introduces a change from the traditional rectified linear unit (ReLU), is shown in Figure 4. Convolutional operations and the gated linear unit (GLU) are combined in this block. The GLU functions as both activation and a computational component to create the hidden layer's output, denoted by the symbol $h_l(x)$.

$$h_l(y) = (y*U + b) \otimes \sigma(y*W + c) \quad (26)$$

The y_x values represent the hidden layer's (h_l) input. U, b, W, and c are designated as parameter elements for convolutional layers within linear projection layers. The symbol represents the sigmoid function and denotes element-wise matrix multiplication. The information going through the hierarchy can be regulated using this gating mechanism dependent on the states of the preceding layers.
The feature map dimension is increased by using pixel shuffler up sampling layers after extracting patterns from the inputs. Pixel shufflers are used in computer vision processing to recreate high-resolution images.

3.4.1.2. Discriminator design

The Discriminator Network's job is to determine the input image's legitimacy to increase the denoised outcome's aesthetic appeal. The goal is to retain the value for created samples near zero while assigning a probability value for actual picture data as close to one as practical. The Discriminator Network's smooth operation is made possible by this interaction.
The following three components are located between input and output:
1. Convolutional layer and leaky ReLU activation function integration
2. In seven blocks, A batch normalization (BN) layer, a recurrent convolutional (Conv) layer, and a leaky ReLU layer were all present. Each block's kernel size rose gradually from 64 to 512 while the strides alternated between 2 and 1 cyclically.

IV. RESULT AND DISCUSSION

This section evaluates the suggested near real-time speaker recognition system's perceived efficacy. All the experiments are performed on a Windows operating system with an i7 processor, Nvidia GT 640, 32 GB RAM clocked at 3.4 GHz using Matlab software. We used the MSR Identity toolbox to implement the speaker recognition system. To validate the improvement in the system's performance with our proposed approach, we conducted experiments for speaker verification and identification. The research makes use of many well-known techniques, including Artificial Neural Networks (ANN), Support Vector Machines (SVM), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN).

4.1. Classification Evaluation

There are several differences between the performance metrics employed by classification systems. The domain of the categorization system must be appropriate for the critical study performance metrics. Using test data, a classification task's performance can be evaluated using a confusion matrix (Tab. 2). Predicting both positive and negative events is a component of its application. The terms "true positive" and "true negative" refer to instances when both the actual and projected classifications are incorrect (i.e., negative), respectively. TP refers to situations where both the primary and forecasted classifications are accurate. False negatives, or FNs, occur when the anticipated class is negative, but the actual class is positive

.

Table 2: Confusion matrix

	Actual Instance	
	Yes	**No**
Predicted Instances		
Yes	TP	FN
No	FP	TN

False positives occur when the expected class is positive, but the actual class is negative. Reduce FP and FN to achieve the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

optimal ASR system performance. A variety of performance measures were utilized to evaluate the classifier's performance. The key steps used in automatic speaker identification are precision, recall, accuracy, false acceptance rate (FAR), false rejection rate (FRR), equal error rate (EER), and execution time. A summary of these performance benchmarks is provided in the section that follows.

4.1.1 Precision Analysis

It measures the proportion of predicted unfavourable events. In most cases, exactness is calculated using the precision measure. The precision values increase as the FP rate decreases.

$$\Pr esision = \frac{TP}{FP + TP} \tag{27}$$

Table 3: Precision Analysis for GAN-CWPT method with existing systems

Number of data from dataset	ANN	SVM	RNN	CNN	GAN-CWPT
200	85.039	91.324	89.034	87.234	94.983
400	86.324	91.928	89.617	87.928	95.425
600	85.653	92.617	90.123	88.192	95.827
800	86.435	92.928	90.425	88.562	96.039
1000	87.094	93.627	90.928	88.928	96.824

In Fig. 5 and Tab. 3, the precision of the GAN-CWPT methodology is compared to that of other frequently used methods. The graph demonstrates how the deep learning strategy outperforms the different alternatives regarding precision. For instance, the precision values for the ANN, SVM, RNN, and CNN models are 85.039%, 91.324%, 89.034%, and 87.234%, respectively, while the precision value for the GAN-CWPT model is 94.983% for 200 data. The suggested GAN-CWPT model has a precision value of 96.824% under 1000 data, which is higher than the ANN, SVM, RNN, and CNN models, with precision values of 87.094%, 93.627%, 90.928%, and 88.928%, respectively.

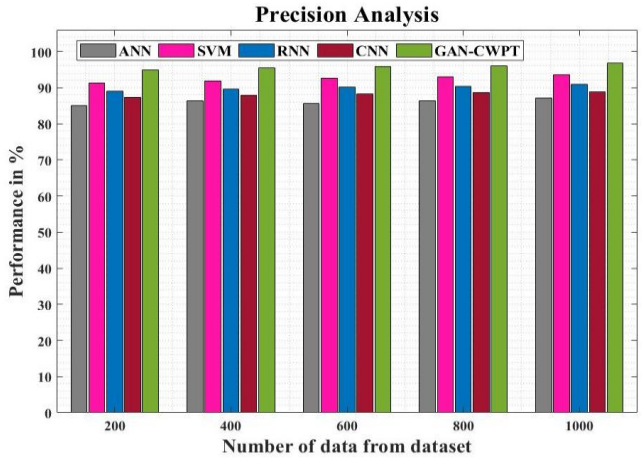

Figure 5: Precision Analysis for GAN-CWPT method with existing systems

4.1.2 Recall Analysis

The proportion of true positives (TPs) or accurately predicted positive outcomes to all positives is called recall. The true positive rate (TPR) is another term for recall.

$$\text{Re}\,call = \frac{TP}{FN + TP} \tag{28}$$

Table 4: Recall Analysis for GAN-CWPT method with existing systems

Number of data from dataset	ANN	SVM	RNN	CNN	GAN-CWPT
200	81.982	85.093	83.727	79.617	87.928
400	81.092	85.213	83.028	80.182	87.425
600	82.324	85.637	83.938	80.617	87.637
800	82.738	86.637	84.223	80.917	88.927
1000	82.083	86.435	84.627	81.425	89.435

In Fig. 6 and Tab. 4, the recall of the GAN-CWPT strategy is compared to that of other commonly utilized methods. The graph depicts how the deep learning technique outperforms the different strategies in terms of recall. For instance, the recall value of the GAN-CWPT model for 200 data is 87.928%, whereas for the ANN, SVM, RNN, and CNN models it is 81.982%, 85.093%, 83.727%, and 79.617%, respectively. With recall values of 82.083%, 86.435%, 84.627%, and 81.425%, for the ANN, SVM, RNN, and CNN models respectively, under 1000 data, the proposed GAN-CWPT model has recall value of 89.435%.

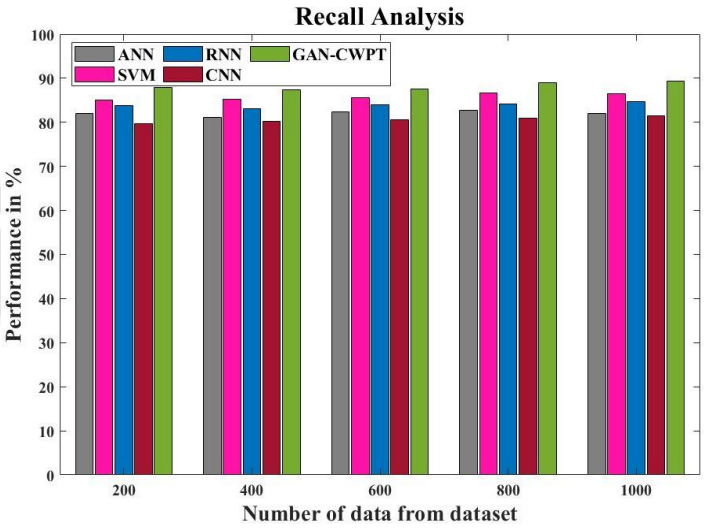

Figure 6: Recall Analysis for GAN-CWPT method with existing systems

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

4.1.3 Accuracy Analysis

The performance parameter known as accuracy measures how many examples a given classification system properly identified. It calculates the proportion of accurately anticipated occurrences to total instances. Eq. (29), which displays the mathematical expression for accuracy.

$$Accuracy = \frac{(TP + TN)}{(TN + TP + FN + FP)} \quad (29)$$

Table 5: Accuracy Analysis for GAN-CWPT method with existing systems

Number of data from dataset	ANN	SVM	RNN	CNN	GAN-CWPT
200	86.535	91.028	89.415	83.927	95.647
400	87.917	92.516	90.615	84.425	96.324
600	86.213	93.715	89.914	85.572	95.725
800	87.415	92.916	90.314	84.927	97.726
1000	88.118	94.553	91.018	85.332	98.627

In Fig. 7 and Tab. 5, the accuracy of the GAN-CWPT methodology is compared to that of other commonly used methodologies. The graph depicts how the deep learning method has an enhanced accuracy performance. For instance, the GAN-CWPT model has an accuracy of 95.647% for 200 data, compared to the accuracy values of 86.535%, 91.028%, 89.415%, and 83.927% for the ANN, SVM, RNN, and CNN models, respectively. In terms of accuracy under 1000 data, the suggested GAN-CWPT model outperforms with 98.627% of accuracy while the ANN, SVM, RNN, and CNN models have an accuracies of 88.118%, 94.553%, 91.018%, and 85.332%, respectively.

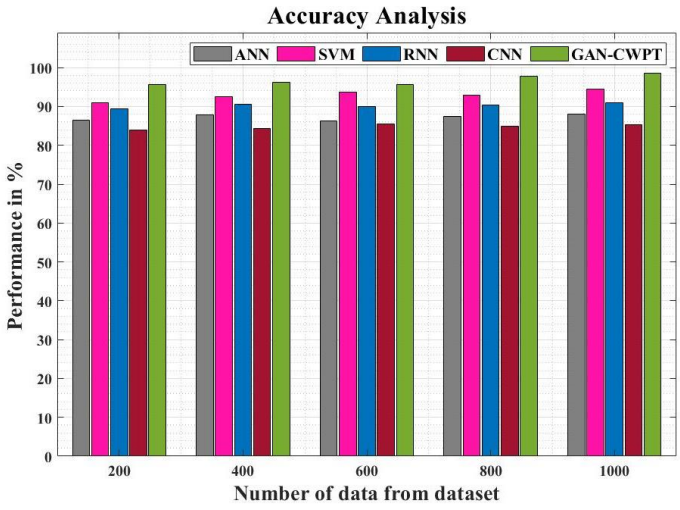

Figure 7: Accuracy Analysis for GAN-CWPT method with existing systems

4.1.4 EER Analysis

The mean of the false acceptance rate (FAR) and false rejection rate (FRR) is calculated using the Equal Error Rate (EER). A decrease in the EER value translates into a loss in system precision. Eqs 30 and 31 provide the equations for calculating FAR and FRR, respectively, while Eq 32 determines the EER.

$$FAR = \frac{FP}{FP + TN} \quad (30)$$

$$FRR = \frac{FN}{FN + TP} \quad (31)$$

$$EER = \frac{FAR + FPR}{2} \quad (32)$$

Table 6: EER Analysis for GAN-CWPT method with existing systems

Number of data from dataset	ANN	SVM	RNN	CNN	GAN-CWPT
200	22.516	20.415	18.324	17.134	15.424
400	22.817	20.928	18.927	17.322	15.827
600	22.019	20.516	19.028	17.625	16.029
800	23.656	21.412	19.425	18.028	16.324
1000	23.926	21.762	19.726	18.324	16.871

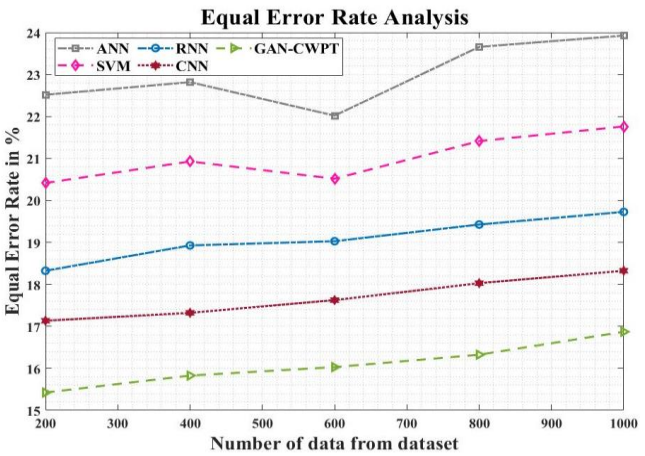

Figure 8: EER Analysis for GAN-CWPT method with existing systems

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Figure 8 and Tab. 6 display an EER comparison of the GAN-CWPT technique with various well-known techniques. The graph shows how the deep learning increases the output while decreasing EER. The ANN, SVM, RNN, and CNN models have EER values of 22.516%, 20.415%, 18.324%, and 17.134%, respectively, while the GAN-CWPT model's EER value for 200 data is 15.424%. The GAN-CWPT model, however, has demonstrated to function optimally over various data sizes with low EER values. Like this, for 1000 data, the EER value for the GAN-CWPT is 16.871%, compared to the values for the ANN, SVM, RNN, and CNN models, which are 23.926%, 21.762%, 19.726%, and 18.324%, respectively.

4.1.5 FRR Analysis

Table 7: FPR Analysis for GAN-CWPT method with existing systems

Number of data from dataset	ANN	SVM	RNN	CNN	GAN-CWPT
200	28.425	26.314	25.082	23.425	21.029
400	28.526	26.651	25.872	23.927	22.092
600	29.627	27.425	25.324	24.324	21.827
800	29.926	27.827	26.213	24.026	22.313
1000	29.213	28.029	26.029	24.725	23.029

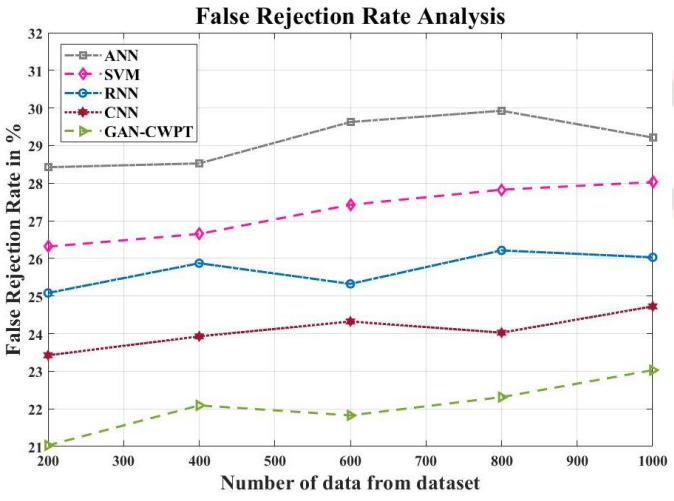

Figure 9: FPR Analysis for GAN-CWPT method with existing systems

Figure 9 and Tab. 7 display a FRR comparison of the GAN-CWPT strategy with various well-known techniques. The graph shows how the productivity increases while FRR decreases with deep learning. For example, the FRR value of GAN-CWPT model is 21.029%, for 200 data while it is 28.425%, 26.314%, 25.082%, and 23.425%, for ANN, SVM, RNN, and CNN models respectively. But the GAN-CWPT model has proven to perform most effectively with low FRR values across various data sizes. In a similar vein, with 1000 data, the FRR value for the GAN-CWPT is 23.029%, while it is 29.213%, 28.029%, 26.029%, and 24.725% for the ANN, SVM, RNN, and CNN models, respectively.

4.1.6 Receiver operating characteristics

The false acceptance rate (FAR) and false rejection rate (FRR) are displayed as functions of various values on the Receiver Operating Characteristic (ROC) curve. The classification algorithm's performance compared to the discriminating threshold is depicted graphically. FAR and FRR are affected by the training database's size and the decision threshold used to calculate the score. Examining the ROC curve as a function of the decision threshold, the erroneous rejection rate may be compared to the equivalent false acceptance rate. Changing the decision threshold can also affect the classifier's output.

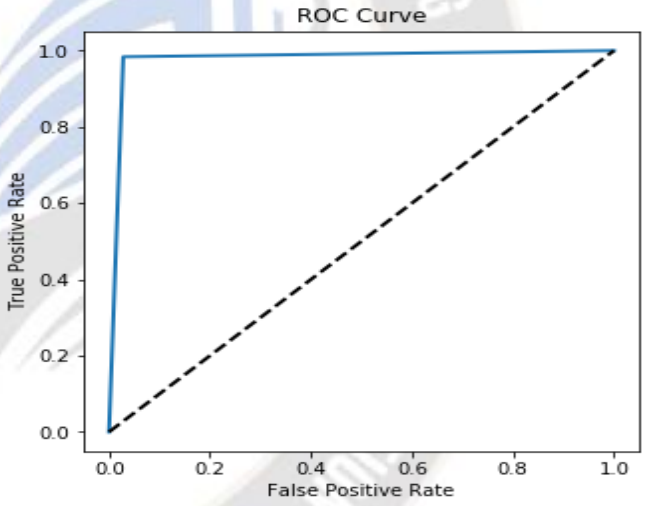

Figure 10: ROC curve analysis for GAN-CWPT method

4.1.7 Execution Time Analysis

The time it takes for a computer system or algorithm to analyze and process audio data to detect and classify the speakers in the input audio is called execution time. This time measurement is critical for evaluating the speaker recognition and classification system's efficiency and real-time capabilities.

Table 8: Execution Time Analysis for GAN-CWPT method with existing systems

Number of data from dataset	ANN	SVM	RNN	CNN	GAN-CWPT
200	0.132	0.173	0.182	0.159	0.109
400	0.138	0.179	0.192	0.152	0.115

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

600	0.140	0175	0.198	0.156	0.121
800	0.146	0.189	0.194	0.163	0.119
1000	0.142	0.185	0.197	0.168	0.125

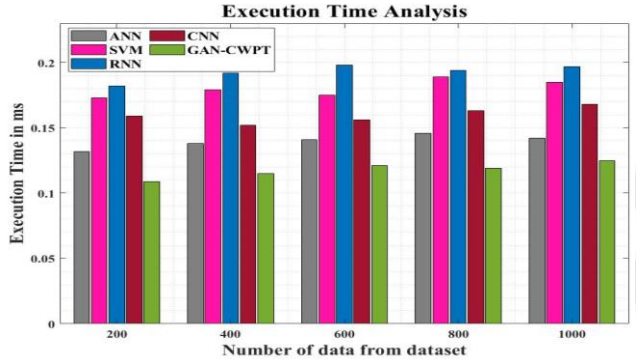

Figure 11: Execution Time Analysis for GAN-CWPT method with existing systems

In Tab.8 and Fig.11, the computational time of the proposed GAN-CWPT methodology is compared to that of existing techniques. The data clearly shows that the GAN-CWPT technique has outperformed all other strategies. The suggested GAN-CWPT approach, for example, took only 0.109ms to compute 200 data, whereas other current methods such as ANN, SVM, RNN, and CNN have taken 0.132ms, 0.173ms, 0.182ms, and 0.159ms, respectively. Similarly, the suggested GAN-CWPT approach takes 0.125ms to compute 1000 data, while existing techniques like ANN, SVM, RNN, and CNN have taken 0.142ms, 0.185ms, 0.197ms, and 0.168ms, respectively as their execution time.

4.1.8 Training and Testing Accuracy Analysis

Fig. 12 shows the training accuracy and testing accuracy of the GAN-CWPT system on 80:20 of the TR dataset/TS dataset. The evaluation of the GAN-CWPT approach on the TR dataset defines the training accuracy. In contrast, the testing accuracy is computed by assessing the performance on a separate testing dataset. The results show that training and testing accuracy increase with increase in epochs that increases the performance of the GAN-CWPT method on the TR and TS datasets.

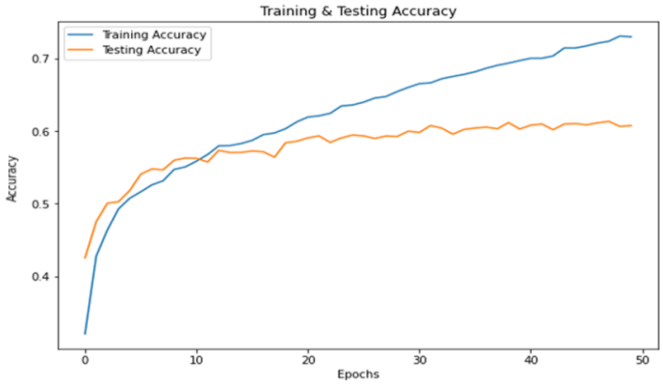

Figure 12: Training and Testing Accuracy Analysis for GAN-CWPT method

4.1.9 Training and Testing loss Analysis

Fig. 13 shows the training loss and testing loss outcome of the GAN-CWPT system on 80:20 of the TR dataset/TS dataset. The training loss defines the error between the predictive performance and original values on the TR data. The testing loss measures the performance of the GAN-CWPT approach on individual validation data. The findings show that the training loss and testing loss tend to decrease with increase in epochs. It depicted the superior performance of the GAN-CWPT method and its ability to generate precise classification. The reduced value of training loss and testing loss illustrates the superior performance of the GAN-CWPT technique in capturing patterns and relationships.

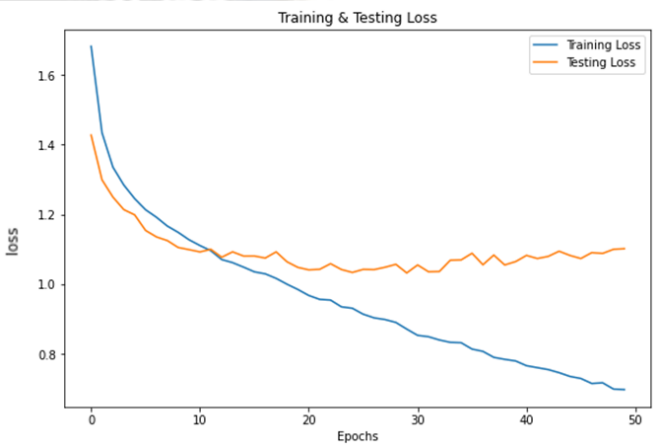

Figure 13: Training and Testing loss Analysis for GAN-CWPT method

V. CONCLUSION

In this study, the components that make up an autonomous speaker recognition system are identified and described in depth. Speech databases, including recordings made in various languages using recording techniques in both quiet and noisy environments, are crucial for training these systems. Creating a dependable and efficient Speaker Identification (SI) system is then made possible by preprocessing the speech signals to prepare them for feature extraction. The enhanced audio features can handle pitch, tone, and pronunciation fluctuations that are common in speaker recognition tasks and more accurately describe speech characteristics. Our proposed approach GAN-CWPT enhances the training data and increases the diversity of the dataset by employing GANs to generate synthetic speech samples. The generator and discriminator parts of the GAN architecture have been modified to create realistic voice samples with characteristics that are quite comparable to actual speaker utterances. The additional dataset improves the Speaker Recognition and Classification system's generalization and robustness, even in settings with little training data. We run extensive experiments on popular speaker identification datasets to assess how well our technique functions. The outcomes show that the suggested WMDCNN-RIO approach beats state-of-the-art

methods in terms of accuracy and effectiveness. The presented model had a total accuracy of 98.627%, making it the most accurate. Future research on speaker emotion recognition will center on using brief spoken instructions in voice interactive systems. We aim to improve command prioritizing in multi-user environments by utilizing this information. This will assist these systems in understanding the command and how an order is said. An order given in a stern or obnoxious tone will be given higher priority than one. The user's experience can be significantly improved by a voice interaction system that can recognize emotion.

REFERENCES

1. Mukhamadiyev, Abdinabi, IlyosKhujayarov, OybekDjuraev, and Jinsoo Cho. "Automatic speech recognition method based on deep learning approaches for Uzbek language." *Sensors* 22, no. 10 (2022): 3683.
2. De Lima, T.A.; Da Costa-Abreu, M. A survey on automatic speech recognition systems for Portuguese language and its variations. Comput. Speech Lang. 2020, 62, 101055. [CrossRef]
3. Chen, Y.; Zhang, J.; Yuan, X.; Zhang, S.; Chen, K.; Wang, X.; Guo, S. SoK: A Modularized Approach to Study the Security of Automatic Speech Recognition Systems. arXiv 2021, arXiv:2103.10651.
4. Xia, K.; Xie, X.; Fan, H.; Liu, H. An Intelligent Hybrid–Integrated System Using Speech Recognition and a 3D Display for Early Childhood Education. Electronics 2021, 10, 1862.
5. Sodhro, A.; Sennersten, C.; Ahmad, A. Towards Cognitive Authentication for Smart Healthcare Applications. Sensors 2022, 22, 2101.
6. Avazov, K.; Mukhriddin, M.; Fazliddin, M.; Young, I. Fire Detection Method in Smart City Environments Using a Deep-Learning Based Approach. Electronics 2021, 11, 73. [CrossRef]
7. Qian, Y.; Zhou, Z. Optimizing Data Usage for Low-Resource Speech Recognition. IEEE/ACM Trans. Audio Speech Lang. Processing 2022, 30, 394–403.
8. Ismail, Muhammad, Shahzad Memon, Lachhman Das Dhomeja, ShahidMunir Shah, Dostdar Hussain, Sabit Rahim, and Imran Ali. "Development of a regional voice dataset and speaker classification based on machine learning." *Journal of Big Data* 8 (2021): 1-18.
9. Mokgonyane, Tumisho Billson, Tshephisho Joseph Sefara, Thipe Isaiah Modipa, Mercy MosibudiMogale, Madimetja Jonas Manamela, and Phuti John Manamela. "Automatic speaker recognition system based on machine learning algorithms." In *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, pp. 141-146. IEEE, 2019.
10. Kamiński, K. A., & Dobrowolski, A. P. (2022). Automatic Speaker Recognition System Based on Gaussian Mixture Models, Cepstral Analysis, and Genetic Selection of Distinctive Features. *Sensors*, *22*(23), 9370.
11. Dhakal, Parashar, Praveen Damacharla, Ahmad Y. Javaid, and Vijay Devabhaktuni. "A near real-time automatic speaker recognition architecture for voice-based user interface." *Machine learning and knowledge extraction* 1, no. 1 (2019): 504-520.
12. Ye, Feng, and Jun Yang. "A deep neural network model for speaker identification." *Applied Sciences* 11, no. 8 (2021): 3603.
13. Ibrahim, H., & Varol, A. (2020, June). A study on automatic speech recognition systems. In *2020 8th International Symposium on Digital Forensics and Security (ISDFS)* (pp. 1-5). IEEE.
14. Sarma, B.D.; Das, R.K. Emotion invariant speaker embeddings for speaker identification with emotional speech. In Proceedings of the 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, 7–10 December 2020; pp. 610–615.
15. Shafik, A.; Sedik, A.; El-Rahiem, B.; El-Rabaie, E.-S.; El Banby, G.; El-Samie, F.; Khalaf, A.; Song, O.-Y.; Iliyasu, A. Speaker identification based on Radon transform and CNNs in the presence of different types of interference for Robotic Applications. Appl. Acoust. 2021, 177, 107665.
16. Costantini, Giovanni, Valerio Cesarini, and Emanuele Brenna. "High-Level CNN and Machine Learning Methods for Speaker Recognition." *Sensors* 23, no. 7 (2023): 3461.
17. Li, J.; Zhao, R.; Meng, Z.; Liu, Y.; Wei, W.; Parthasarathy, S.; Gong, Y. Developing RNN-T models surpassing high-performance hybrid models with customization capability. arXiv 2020, arXiv:2007.15188
18. Khassanov, Y.; Mussakhojayeva, S.; Mirzakhmetov, A.; Adiyev, A.; Nurpeiissov, M.; Varol, H.A. A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline. arXiv 2020, arXiv:2009.10334.
19. N. J. De Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. de Wet, E. Barnard, and A. De Waal, “A smartphone-based asr data collection tool for under-resourced languages,” Speech communication, vol. 56, pp. 119–131, 2014
20. Jahangir, Rashid, Ying WahTeh, Henry Friday Nweke, Ghulam Mujtaba, Mohammed Ali Al-Garadi, and Ihsan Ali. "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges." *Expert Systems with Applications* 171 (2021): 114591.
21. P. Ravi Prakash, D. Anuradha, Javid Iqbal, Mohammad GouseGalety, & Ruby Singh (2022) A novel convolutional neural network with gated recurrent unit for automated speech emotion recognition and classification, Journal of Control and Decision, DOI: 10.1080/23307706.2022.2085198
22. Chiai AI-Atroshi, J. Rene Beulah, Kranthi Kumar Singamaneni, C. Pretty Diana Cyril, & S. Velmurugan (2022) Automated speech based evaluation of mild cognitive impairment and Alzheimer’s disease

detection using with deep belief network model, International Journal of Healthcare Management, DOI: 10.1080/20479700.2022.2097764

23. X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2794–2802.
24. Y.N. Dauphin, A. Fan, M. Auli, D. Grangier, Language modeling with gated convolutional networks, in: Proceedings of the 34th International Conference on Machine Learning, ser. ICML'17, vol. 70, 2017, p. 933–941.