An Efficient Voice Authentication System using Enhanced Inceptionv3 Algorithm

[1]Kaladharan N and [2]Arunkumar R
[1,2]Department of Computer Science and Engineering, FEAT, Annamalai University, Tamil Nadu, India.
[1]Department of Computer Engineering, Government Polytechnic College, Theni, Tamil Nadu, India.
[1]nkeeeau@gmail.com, [2]arunkumar_an@yahoo.com

Correspondence should be addressed to Kaladharan N : nkeeeau@gmail.com.

Article Info
Journal of Machine and Comp uting (http://anapub.co.ke/journals/jmc/jmc.html)
Doi: https://doi.org/10.53759/7669/jmc202303032
Received 04 February2023; Revised from 18 May 2023; Accepted 18 June 2023.
Available online 05 October 2023.

Abstract – Automatic voice authentication based on deep learning is a promising technology that has received much attention from academia and industry. It has proven to be effective in a variety of applications, including biometric access control systems. Using biometric data in such systems is difficult, particularly in a centralized setting. It introduces numerous risks, such as information disclosure, unreliability, security, privacy, etc. Voice authentication systems are becoming increasingly important in solving these issues. This is especially true if the device relies on voice commands from the user. This work investigates the development of a text-independent voice authentication system. The spatial features of the voiceprint (corresponding to the speech spectrum) are present in the speech signal as a result of the spectrogram, and the weighted wavelet packet cepstral coefficients (W-WPCC) are effective for spatial feature extraction (corresponding to the speech spectrum). W-WPCC characteristics are calculated by combining sub-band energies with sub-band spectral centroids using a weighting scheme to generate noise-resistant acoustic characteristics. In addition, this work proposes an enhanced inception v3 model for voice authentication. The proposed InceptionV3 system extracts feature from input data from the convolutional and pooling layers. By employing fewer parameters, this architecture reduces the complexity of the convolution process while increasing learning speed. Following model training, the enhanced Inception v3 model classifies audio samples as authenticated or not based on extracted features. Experiments were carried out on the speech of five English speakers whose voices were collected from YouTube. The results reveal that the suggested improved method, based on enhanced Inception v3 and trained on speech spectrogram pictures, outperforms the existing methods. The approach generates tests with an average categorization accuracy of 99%. Compared to the performance of these network models on the given dataset, the proposed enhanced Inception v3 network model achieves the best results regarding model training time, recognition accuracy, and stability.

Keywords – Voice Authentication, Short-Time Fourier Transform, Weighted Wavelet Packet Cepstral Coefficient, Inception V3.

I. INTRODUCTION

People use internet services such as social networking sites, which provide several features for establishing interactions between people. Many users utilize social networking services not just to connect with others but also to obtain a plethora of details, and social networking services are conveniently accessible from any Internet-connected device [1]. Users conversant with information technology are familiar with social networking sites and use social media daily. Although social media content is generally open to the public, it constitutes personal information and must be safeguarded against attackers or unauthorized parties. Various assaults against social networking services have expanded dramatically as the number of SNS users has increased. At the same time, there are various risks against smart devices that can be utilized for social networking. Attackers exploit poor password hygiene, frequently snatching accounts and keeping them for ransom [2, 3].

Many security procedures have been researched, including wireless network security and malware detection [4, 5], where user authentication is regarded as a fundamental notion to ensure user system security. Various self-authentication solutions for user authentication formerly relied on resident registration numbers or public certificates; however, current approaches provide a high threat of personal data leakage accidents shortly. Personal data leakage instances are increasing yearly, and

existing identification methods need help with the difficulty of continually updating ways. Biometrics is one of the authentication methods used to address these issues [6, 7].

Facial recognition, speaker recognition, iris recognition, and fingerprint recognition are examples of biometric technologies. In recent years, biometric technology has started to involve numerous devices, including smartphones and laptops, and the biometrics industry is rising as the usage of biometrics by security-critical entities, including corporations, financial institutions, and government agencies, grows. It is also utilized in mobile apps, and telecoms utilize it to ensure consumers can access their accounts [8].

Many areas of today's world rely on Speech Recognition (SR) and Voice Recognition or Identification (VI) technologies, such as authenticating individuals over the phone for other security services. Speaker recognition difficulties are classified into two types: verification [9] and recognition [10]. Speaker verification aims to ascertain whether the speaker is who he claims to be. This necessitates the system's ability to identify the speaker among many possible deceivers. Alternatively, the purpose of speaker recognition is to be able to recognize a speaker from a set of previously registered speakers. When the system receives an unknown statement from a person, it makes an attempt to identify the speaker as one of the registrants.

With the advancement of technology in recent years, the speaker's voice evolved into a requirement for speaker verification and recognition techniques, including recognizing illegal suspects, increasing human-computer connection, coping with music while waiting in line, and so on. Despite several studies on feature extraction and the development of classifiers, classification accuracy still needs improvement.

The standard automatic speech recognition paradigm is sophisticated and demands a lot of computer and storage resources, pushing it hard for automated voice recognition techniques to infiltrate regular people's lives. However, deep learning (DL) development in voice recognition has substantially reduced the complexity of training models. This paper aims to investigate the application and development pattern of DL algorithms in the area of speech recognition and reveal the relationship between DL and conventional automatic speech recognition technology and explore the main directions and research approach of DL. The implementation of the DL algorithm in the field of speech recognition, as well as a comparison of the benefits and drawbacks of each direction.

In this paper, we offer a method for identifying and verifying individual speech patterns. Speaker recognition is the process of determining a voice's owner based on that voice's features. The Inceptionv3 neural network performs the learning process of the recorded speech data in speaker recognition, and every time speech-based authentication is done, the speech pattern of the particular individual under test is discriminated against using the taught DL method. The logged-in user's data is empty. Furthermore, we conduct tests to highlight the threat of synthesized speech in speaker recognition algorithms, and ways to avoid this problem are implemented in the suggested method.

Our research's key contribution can be stated as follows:

- A deep learning system based on speech data is proposed for user authentication.
- Framing, windowing, and short-time Fourier transform (STFT) generate a spectrogram. Furthermore, the Savitsky-Golay filter is used for a series of digital information points for softening information where data accuracy is improved without signal distortion.
- W-WPCC works well in extracting spatial features from speech spectrograms, and the speech spectrogram provides spatial features of the voiceprint and picked W-WPCC for obtaining features from the spectrogram.
- This study intends to use enhanced inception v3 to ascertain an audio transmission.
- Since the parameters were carefully tuned, the enhanced InceptionV3 learns around 40 seconds faster than the original model.
- MATLAB was used to test the performance of the suggested augmented inception v3-based model.
- The validation performance was compared to the reported technique.
- This paper demonstrates a masquerading exploit for avoiding voice-based identification.

The remainder of the paper is structured as follows: A current literature review is described in Section 2. Section 3 thoroughly describes the suggested method, including everything from audio signal segmentation to audio feature extraction and SVM network training. Section 4 describes the trials that were carried out, demonstrating the accuracy of the speech recognition system and analyzing the audio data used. Finally, Section 5 contains the paper's conclusions.

II. LITERATURE SURVEY

Tandel et al. (2020) suggested thoroughly reviewing the literature on classical and DL-based voice comparison and speaker detection approaches. In addition, the report presents publicly available datasets used by academics for speaker recognition and voice matching. This simple document will be helpful for both novices and researchers interested in voice recognition and speech comparison. Furthermore, the publication describes in sufficient depth the entire voice comparison method. This work

investigates and analyzes standard and DL-based speaker recognition and speech matching methods in the survey and suggests DL-based techniques to master speech processing [11].

Khedir et al. (2021) offered an applied attempt to use convolutional neural networks (CNNs) to construct a system for detecting human speaker identities. This study employs two methods: RW-CNN and MFCC-CNN. The first strategy is the standard one employed by MFCC, in which audio functions are input into a CNN for processing. Begin the training process with the suggested CNN to take the information as a picture after training it. Then the approach, RW-CNN, follows the identical procedures as the first method but skips the MFCC stage and goes directly into the CNN. The system demonstrated great accuracy and low mean square error for both methods. The system was tested using audio and microphone recording files, and the results were primarily positive [12].

Kydyrbekova Aizat et al. (2020) created a block diagram for identifying system users based on individual speech features using a deep neural network (DNN) technique and the i-vector in the fundamental phonetic unit method. Security against many sorts of assaults on biometric systems, allowing users to be identified with first and second mistake probabilities are 0.025 and 0.005, respectively. Incorporating fundamental speech units in developed recognition algorithms can enhance computational metrics, reduce subjective decision-making in biometric techniques, and improve security against assaults on speech biometric techniques [13].

Zeng Taiyao et al. (2022) provided an overview of deep learning's use in speech recognition. Introduced contemporary deep-learning research results in speech recognition, studied the association between classic speech recognition methods and existing DL models, and examined the deep-learning development trajectory. The deep learning model should absorb the standard voice recognition model's principles to develop better a deep learning-based speech recognition system [14].

Ayad Alsobhani et al. (2021) developed a word-tracking model for audio recognition using deep convolutional neural learning. People of all ages use six control words (start, stop, forward, reverse, right, and left). Our voice dataset is divided into two equal sections, male and female, and is utilized to train and test the suggested DNN. The proposed deep neural network achieves 97.06% accuracy in word classification on entirely unknown speech samples. CNNs are employed in the training and testing of our data. Our findings differ from other research, which frequently employs reasonably consistent out-of-the-box data for isolated word kinds [15].

Zheng Ruxin et al. (2022) suggested a DL-based speech recognition system using DL theory. The research work includes the following aspects: providing a convolutional network to extract features; generating a "speech function template library" through colossal training. Study on matching and recognition algorithms; study on a voiceprint recognition-based sign-in system. The system interface will provide the voiceprint database's data and registration history, the existing recognition outcomes, the accuracy rate, and registration time. The system's average recognition rate is around 95%, sufficient for practical applications [16].

Feng Ye et al. (2021) presented a recurrent neural network (RNN) or CNN for speaker recognition. Convolutional layers are used in network model design to extract voiceprint features and lower the complexity of frequency and time domains, allowing for quicker calculation of GRU layers. The experimental findings demonstrate that our suggested DNN model (Deep GRU) obtains a most outstanding accuracy of 98.96%. Meanwhile, the results illustrate the efficacy of the presented GRU deep network framework compared to existing speaker identification models [17].

Bella et al. (2020) suggested a speech authentication system for one-time password (OTP) systems, including a speaker verification and voice recognition approach. Models for recognizing and verifying human voices expressed by MFCC feature vectors include a short-term memory network and a Siamese network with a CNN. Experiments revealed that the speech recognition model's verification accuracy was trustworthy, but the speaker verification technique did not produce better results. At the same time, the precision of the speaker verification model needs to be improved to develop a reliable security system, and further research is warranted [18].

Muruganantham et al. (2020) suggested a CNN (Convolutional Neural Network) model for speaker recognition. After collecting an individual's voice as an information source, MFCC (Mel-Frequency Cepstral Coefficients) calculations are used to calculate coefficients specific to a particular example. Then, use CNN to practice for the speech test. The testing procedure begins once the preparatory process is completed. If a prepared speech test is delivered as feedback during the test, the correct person can be identified, and the correct performance will occur [19].

Salahuddin-Duleby et al. (2020) devised a novel method incorporating CQCC and MFCC as input features, a convolution neural network-based front-end feature extractor, and an SVM back-end classifier. This paper greatly enhances accuracy and attains next-generation performance by applying improved training strategies such as batch normalization. The experiments are carried out on the ASVspoof dataset to show that the approach is promising and outperforms the traditional method presented in the 2017 Replay Attack Challenge. This paper was able to train CNN for feature extraction and SVM for feature classification in experiments. The results reveal promising progress [20].

Amira Shafiq et al. (2021) used CNNs to enhance speaker recognition accuracy in the presence of distractors for robotic control applications. Initially, the speaker's speech signal is segmented, and each segment is turned into a spectrogram, calculating the spectrogram's Radon transform. The suggested model has six max-pooling layers and six convolutional layers,

whereas the reference method has three max-pooling layers and three convolutional layers. Experiment findings show that the suggested RDLM model has a classification accuracy of up to 97.5%, over twice the level of specific established methods for speaker recognition [21].

Abdusalomov et al. (2022) proposed a machine learning (ML)-based algorithm for extracting feature metrics from voice signals to increase voice recognition applications' performance in real-time smart city surroundings. Furthermore, the notion of allocating main memory blocks to caches is efficiently used to cut computation time. Cache block size is a significant factor in cache performance. The issue of overclocking in digital voice signal processing has yet to be overcome. Compared to traditional voice recognition algorithms, experimental results reveal that the suggested method accurately recovers signal features and performs excellent classification [22] in **Table 1.**

Table 1. Summary of the Existing Voice Authentication Process

Reference	Techniques used	Advantages	Limitations
[12]	MFCC-CNN and RW-CNN	excellent accuracy with minimal mean square error	Utilized only a few datasets.
[15]	CNN	improved and more precise categorization accuracy	Over-fitting is possible due to the intricacy of the model structure.
[17]	Gated recurrent unit (GRU) with two-dimensional CNN (2-D CNN)	The suggested approaches combine the benefits of 2-D CNN feature extraction with the temporal dependence of the GRU cell block.	Need to consider the training phase's time cost, as well as the model's stability in complex noise settings
[18]	LSTM and siamese network with CNN	The voice recognition model's validation accuracy is dependable.	Meanwhile, the accuracy of the speaker verification model is insufficient to develop a reliable security system and has to be investigated.
[22]	MFCC		As a result of the diverse noise conditions, faults occur.
[23]	Connectionist Temporal Classification (CTC)	It is not necessary to align annotations and data one by one.	The CTC approach needs language modeling power for integrating linguistic models for combined optimization and a failure to predict output dependencies.
[24]	Listen, Attend, and Spell (LAS)	No independent assumption are necessary.	However, because the LAS method must recognize the entire input sequence thereafter, real-time performance is weak.
[25]	LSTM	improved word accuracy	Metrics used for LSTM training networks is four times that of typical RNN, making over fitting a possibility.

III. PROPOSED METHODOLOGY

Voice is most likely the most significant kind of human contact. Human speech or voice is an informational signal that provides various information, including linguistic material, speaker mood, and tone of voice. Based on speech features, VPR is intended to separate, identify, and identify speakers. There are various methods for making the speaker identification procedure easier. Typically, such systems require two stages: feature extraction and feature matching or classification, with the classification part consisting of two components: decision-making and pattern matching. A general-purpose voiceprint recognition system is depicted in **Fig 1.** The feature extraction module calculates a set of speech signal features that highlight speaker-specific data, after which each speaker's speech is recorded and utilized to develop a corresponding speaker model.

The voiceprint recognition technique's operation has been split into training and testing. During training, the device will gather the original speech signal through a series of preprocessing procedures based on the system's metrics. The extract method retrieves the feature parameters associated with each voice. For training, speech features are fed into the recognition network. Obtain speaker models and save them in a sample library for future evaluation. The voice samples in the test sample set are preprocessed, features are extracted during the testing phase, and the features of the voice samples to be examined are compared with the registered speaker model to identify the speaker's identity. **Fig 1** depicts the system framework 1.

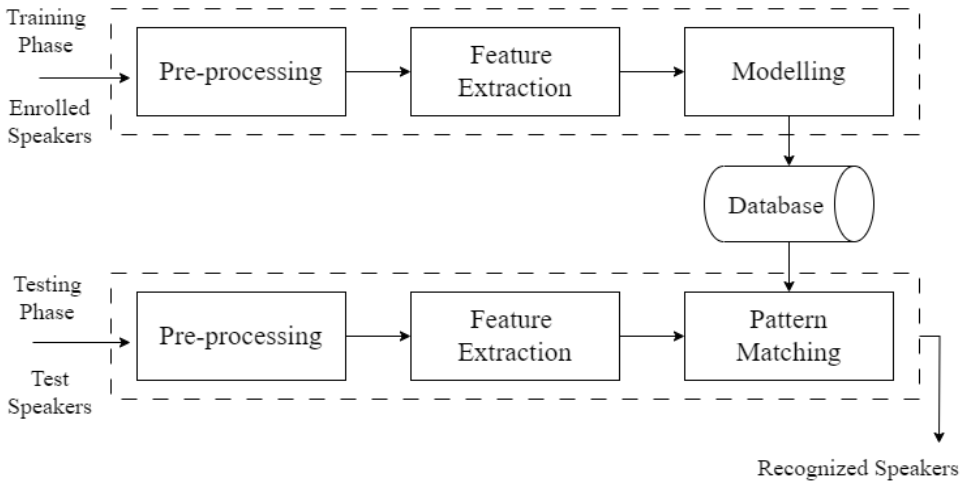

Fig 1. Frame of Voice Authentication System

Preprocessing

In the context of automated voice authentication, voice authentication consists of two steps: speech simulation and voice recognition. In the speaker representation job, features extracted from known persons' signals are registered in a database for eventual usage in test situations with various classifiers. Acoustic interference degrades the performance of voice authentication systems. Deconvolution and noise reduction techniques can be utilized as pretreatment stages in the testing stage to achieve improved recognition results. Before entering the system, the noisy interfering voice signal initially travels through a speech improvement stage, where standard methods are utilized to eliminate noise from the speech signal.

Our objective in data preparation is to acquire spectrograms that contain rich audio characteristics of loudspeakers. Framing, windowing, the short-time Fourier transform (STFT), and other techniques are used to create spectrograms [26]. STFT is the process of computing a signal's discrete Fourier transform (DFT) in short, overlapping windows. Short overlapping windows in speech signal processing helps keep the voice signal stable in the short term. The voice sample frequency in this experiment is 16 kHz, and the number of FFT points is 512. The framed speech signal's discrete STFT is computed using the equ (1)

$$X[K] = \sum_{n=0}^{N-1} x\, x[n] e^{-j\frac{2\pi nK}{n}} \quad K = 0,1,2,\ldots,N-1 \tag{1}$$

Where x[n] represents the framed signal, n denotes the frame number, and N represents the frame size. Formula (2) is used to define the power spectrum.

$$P(K) = \left| X(K) \right|^2 \tag{2}$$

In our studies, a float-point series of times is returned after loading an audio file. Convert its energy spectrum to the Mel scale after determining its energy spectrum. Furthermore, performance can be analyzed by converting the power spectrum (width squared) to dB units. The frame length was fixed to 32ms throughout this operation, the window between subsequent frames was reduced to 16ms, and the window function was set to Hamming window. Lastly, as seen in Figure 2, we receive the spectrogram.

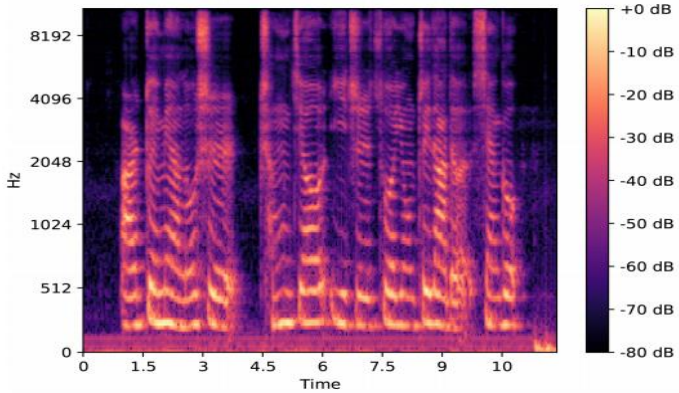

Fig 2. Spectrogram of Speech

The Savitsky-Golay filter is an example of a digital filter. This filter is applied to a series of numeric data points to smooth the data and improve its accuracy while not distorting the signal. Precision is maximized by a procedure known as convolution, which involves picking a contiguous subset of nearby data points utilizing a low-degree polynomial and the least squares method. Suppose all data points are equally spaced; then, a mathematical solution to a least-squares formula can be obtained in a set of "convolution coefficients" that can be applied to all subsets of the data to provide a smooth estimate. Signal (or smoothed signal derivative) at each subset center point.

Feature Extraction using W-WPCC

A spectral feature is the essence of wavelet packet cepstral coefficient (WPCC) [27]. Given the complementary features between WPCC and noise-resistant sub-band spectral centroids, the technique integrates them using specific methodologies to provide new noise-resistant speech signal characteristics. W-WPCC is formed by combining the sub-band spectrum's centroid with the wavelet packet's cepstral coefficient using a weighted technique. On the other side, because of the white noise condition, distinct sub-band energies of the voice signal mixed with white noise are easily confused. The sub-band energy is less impacted by noise than the sub-band spectral centroid. As a result, the cepstral coefficients of the weighted wavelet packet are incredibly resistant to speech signal white noise. On the other hand, the sub-band energy weighted by the sub-band spectrum centroid is a new definition of the frequency domain disbandment of the speech signal that contains essential information about the speech signal, can distinguish the speech category, and can construct the characteristics of the speech signal. The algorithm for obtaining weighted wavelet packet cepstral coefficient (W-WPCC) characteristics from sub-band spectral centroids is depicted in **Fig 3.**

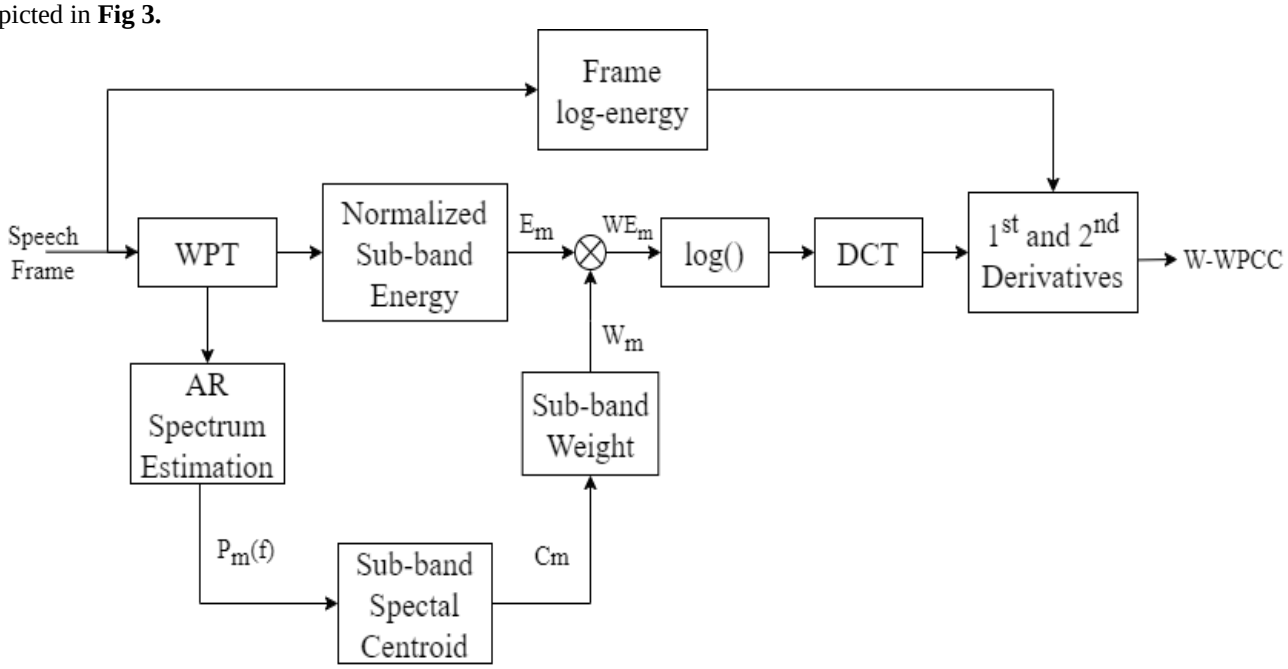

Fig 3. Block Diagram of the W-WPCC Feature Extraction

Voice Signal Authentication Model Using Enhanced Inception V3 Network

Inceptionv3 model

Szegedy et al. 2014 submitted the Inception model in the Large-Scale ImageNet Visual Recognition Challenge to lessen the impact of computational performance and low metrics in the application scenario [17]. Inception-v3's input picture size is 299 x 299. Despite being 78% larger than VGGNet (244 x 244), Inception-v3 outperforms VGGNet. The following are the primary reasons why Inception-v3 is so effective: Inception-v3 has fewer than half (60,000,000) of the parameters of AlexNet and less than a quarter of the parameters of AlexNet. (140,000,000); also, the total amount of floating-point calculations performed by the complete Inception-v3 network is around 5,000,000,000 times, significantly greater than that of Inception-v1 (approximately 1,500,000,000 times) [28-30].

These features enable Inception-v3 to be easier to implement, as it can be readily deployed on standard servers to deliver quick response services. Inception-v3 employs convolution kernels of varying sizes, resulting in various receptive fields. A modular architecture decreases the network's design area, and the ends are combined to realize the fusion of characteristics of different scales. Table 2 summarizes the Inception-V3 network parameters. Figure 4 depicts the Inception-v3 configuration.

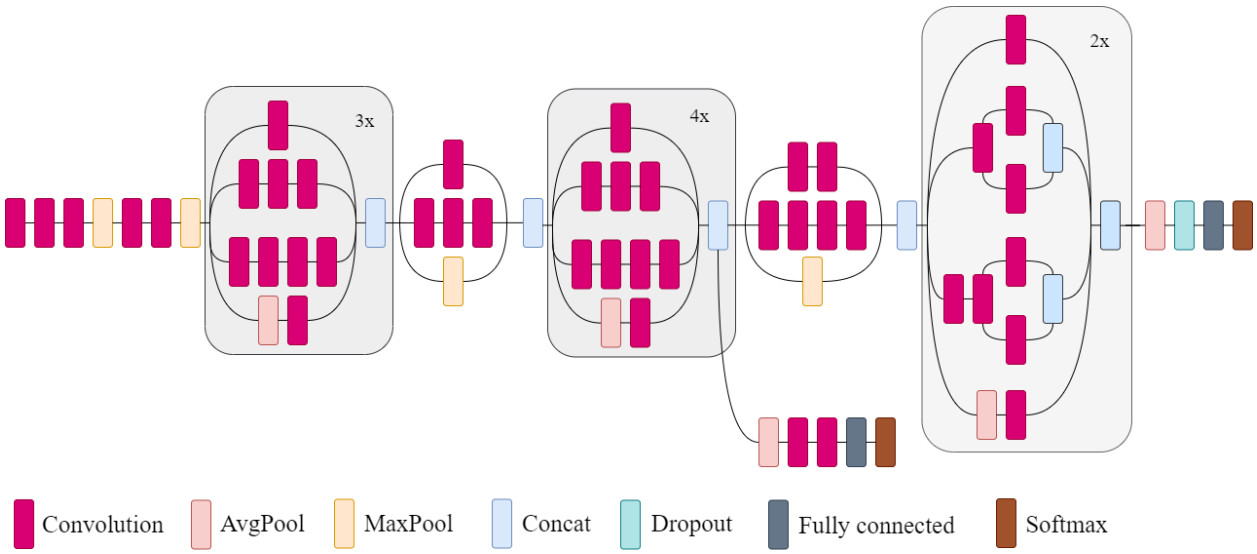

Fig 4. Structure Diagram of The Inception-V3 Model

Table 2. Network Structure of the Inception-V3 Model

Type	Input size	Patch stride/size
convolution	299 x 299 x 3	3 x 3/2
convolution	149 x 149 x 32	3 x 3/1
convolution	147 x 147 x 32	3 x 3/1
pooling	299 x 299 x 64	3 x 3/2
convolution	73 x 73 x 64	3 x 3/1
convolution	71 x 71 x 80	3 x 3/2
convolution	35 x 35 x 192	3 x 3/1
3 x Inception	35 x 35 x 288	---
5 x Inception	17 x 17 x 768	---
2 x Inception	8 x 8 x 1280	---
pooling	8 x 8 x 2048	8 x 8
linear	1 x 1 x 2048	logits
softmax	1 x 1 x 1000	classifier

Batch normalization (BN) layers are used as regularizers among auxiliary classifiers and fully connected (FC) layers in Inception-v3. The BN model can use the batch gradient descent approach to increase the deep neural network model's training and convergence speeds. The BN formula is written as follows:

$$B = \{X_{1....m}\}, \gamma, \beta \tag{3}$$

$$\{y_i = BN_{\gamma.\beta}(X_i)\} \tag{4}$$

$$\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} cv\, X_i \tag{5}$$

$$\sigma_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} v\,(X_i - \mu_B)^2 \tag{6}$$

$$\hat{X}_i \leftarrow \frac{X_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \tag{7}$$

$$y_i \leftarrow \gamma \hat{X}_i + \beta = BN_{\gamma.\beta(X_i)} \tag{8}$$

Proposed Enhanced Inceptionv3

Fig 5 depicts the construction of the suggested network model, an algorithm for identifying voice signals. By employing fewer parameters, this architecture decreases the intricate nature of the convolution process and increases learning speed. The model, depicted in Figure 5, uses the basic architecture of the InceptionV3 system, though the architecture of the fully connected layers differs.

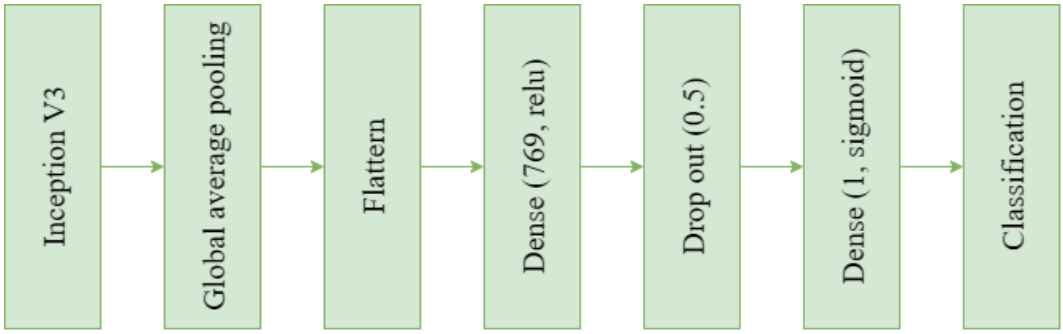

Fig 5. Structure of the Proposed Inceptionv3 Modification to Improve Accuracy

The suggested InceptionV3 system retrieves features from convolutional and pooling layer input data; however, it improves learning efficacy by changing the architecture of the layers that are fully linked. The dropout rate is set to 0.5, widely used to avoid over-tuning the neural network and increase efficiency. The primary difference between the suggested approach and InceptionV3 is that the activation function utilized in the final classification process is a sigmoid function suitable for speech signal classification rather than the Softmax function employed in the InceptionV3 method.

IV. RESULTS AND DISCUSSION

Scalability and real-time implementation depend on hardware and audio I/O library optimization. The suggested and current models' training data sets include mel-spectrogram pictures for stimulated and genuine speech and distinct speech kinds for realism and speech classification, respectively. In this experiment, we compare the enhanced inceptionv3 speech analysis method with conventional machine learning and cutting-edge deep learning models for identifying speech signals as real or stimulated. The normalized data is then divided into two sets: 75% training and 25% validation.

Database

Our research focuses on English speakers on social networks. As a result, we used the voices of five English speakers from YouTube in.wav format. Audacity software was used to convert the audio files to nomo files. Finally, we use MATLAB to successively slice every file into 100 samples of 2 seconds each. Eighty samples are used for training, and the remaining 20 for testing each speaker. Preprocessing yielded sound characteristics and descriptions. The influence of various dataset features on classification success may vary. Classification success rates can be increased by removing some of these aspects. W-WPCC coefficients are extracted as sample features for the W-WPCC method, and Enhanced Inception v3 is employed as the classifier. We display the waveform of each sample and store it as a JPEG image file for training and testing with the suggested model for enhancing Inception v3 with raw waveform images.

Performance Metrics

$$Accuracy = \frac{TN+TP}{TP+TN+FN+FP} \tag{9}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{10}$$

$$Specificity = \frac{TN}{TN+FP} \tag{11}$$

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

$$F-measure = 2\ x\ \frac{Precision\ x\ recall}{Precision+\ recall} \tag{13}$$

$$G-mean = \sqrt{Sensitivity.Specificity} \tag{14}$$

$$Error\ rate = 1 - [Accuracy] \quad (15)$$

Where TN-True negative, TP- True positive, FN-False negative, FP-False positive.

Table 3. Comparison of the Suggested Method

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)
SV	91	88	90
RF	92	91	92
ResNet	95	93	94
VGG-16	96	94	95
AlexNet	97	95	97
Proposed method	**99**	**98**	**98.5**

Table 3 compares the suggested enhanced start V3 model to the other models, revealing that it surpasses all models in all categories of performance ranking. Figures 6 and 7 compare the performance of the suggested models. Enhanced Inception v3's CNN transfer deep learning approach achieves 99% accuracy, 98% sensitivity, 98.5% specificity, and 98.2% accuracy.

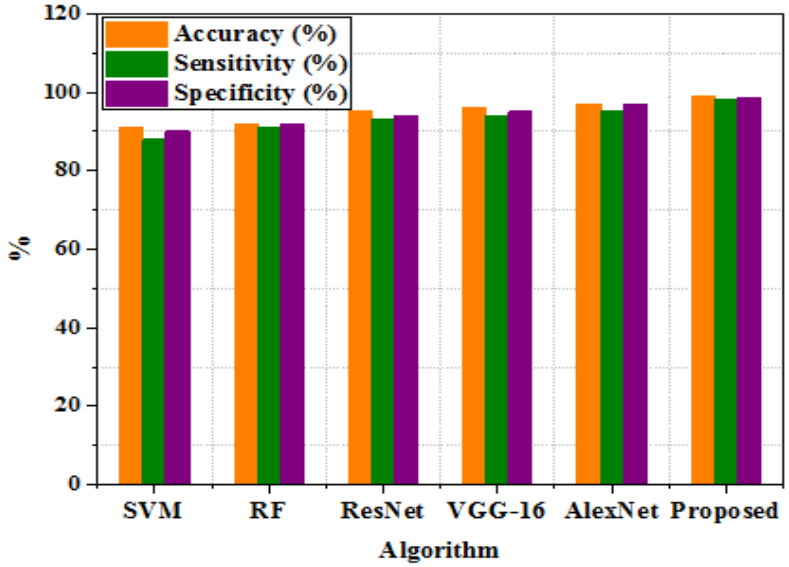

Fig 6. Performance Comparisons in Terms of Accuracy, Sensitivity and Specificity

Table 4. Comparison of the Proposed Method

Algorithm	Precision (%)	F1 score (%)	G-mean (%)
SVM	90.1	90.5	89.3
RF	93.4	92.6	91.7
ResNet	95.6	94.1	93.4
VGG-16	96.2	95	95.8
AlexNet	97.5	96.1	96.3
Proposed method	**98.2**	**98**	**98.9**

According to the study's findings, the proposed model beats existing methods in terms of evaluation metrics. The accuracy rates of the SVM, RF, ResNet, VGG-16, and AlexNet algorithms are 91%, 92%, 95%, 96%, and 97%, respectively**. Table 4** compares the proposed model to current models based on classification accuracy, F1 score, and mean g. The suggested Enhanced Inception V3 approach outperforms the other algorithms in terms of accuracy (98.2%), F1-score (98%), and G-means (98.9%). When compared to the original model, the approach suggested yields better results.

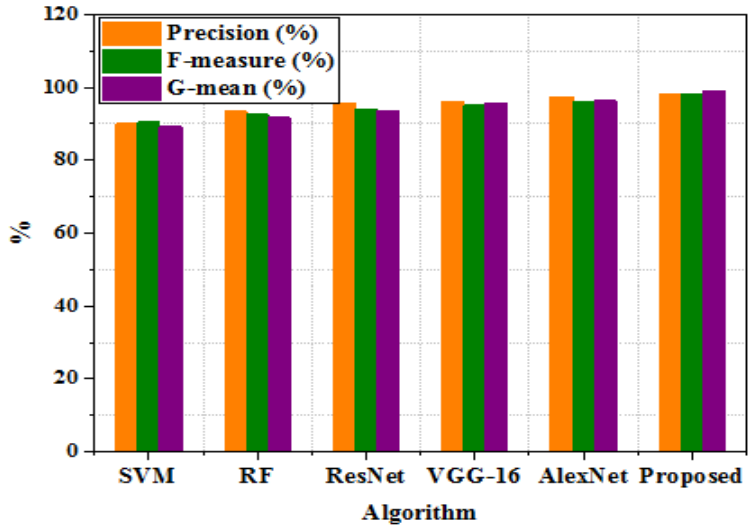

Fig 7. Performance Comparisons in Terms of Precision, F-Measure and G-Mean

$$\text{Kappa}=\frac{P_0-P_e}{1-P_e} \quad (16)$$

Where P_0 is the model's overall accuracy, and P_e is the index of the consistency of model predictions and actual class values.

Table 5. Comparison of the Proposed Method

Algorithm	Error rate (%)	Kappa
SVM	0.15	0.6
RF	0.13	0.8
ResNet	0.095	0.84
VGG-16	0.07	0.88
AlexNet	0.05	0.92
Proposed method	0.018	0.99

The error rates of the suggested approach and known algorithms are shown in **Fig 8 and Table 5**. Current algorithms, including SVM, RF, ResNet, VGG-16, and AlexNet, have error rates of 0.15%, 0.13%, 0.095%, 0.07%, and 0.05%, respectively. The graphic shows that the suggested approach substantially decreases the error rate compared to the current techniques.

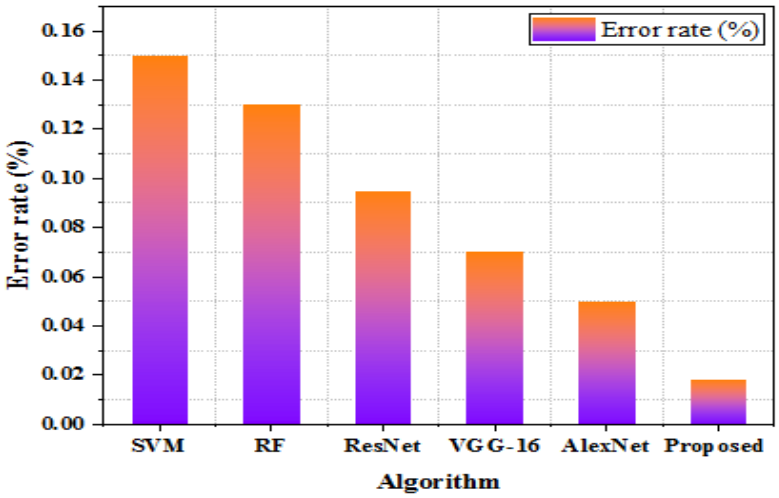

Fig 8. Performance Comparisons of Kappa

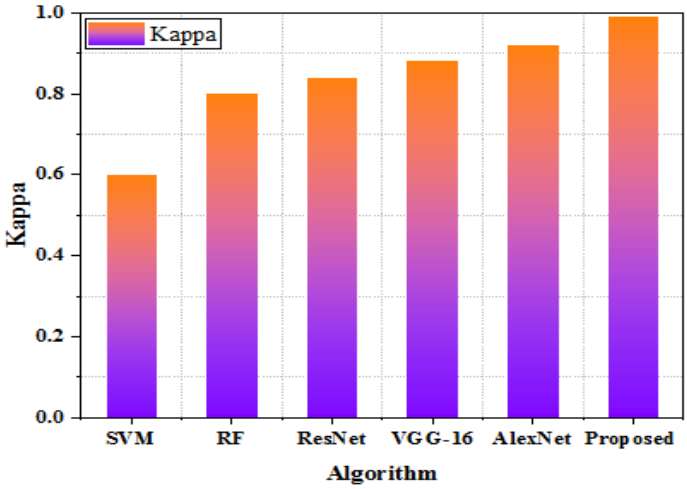

Fig 9. Performance Comparisons of Kappa

Cohen's kappa coefficient can be utilized to define the degree of agreement between two traditional nominal classifications. If Cohen's kappa is utilized to quantify category balance, the ranks of all nominal categories are regarded to be the same, making sense if all conceptual categories represent various sorts of "presence". The kappa values of the suggested approaches are compared in **Fig 9**. For the provided dataset, the suggested technique achieves a high kappa value (0.99), while current techniques, including SVM, RF, ResNet, VGG-16, and AlexNet, achieve 0.6, 0.8, 0.084, 0.088, and 0.92, respectively.

$$\text{Word error rate (WER)} = \frac{s+d+i}{n} \quad (17)$$

Where s represents the number of substitutions, i represents the number of iterations, d represents the number of deletions, and n represents the number of words in the reference.

Table 6 . Comparison of Proposed Method

Algorithm	Word error rate (%)	Simulation time (sec)
SVM	32.63	120.7
RF	33.91	80.5
ResNet	33.69	60.3
VGG-16	30.31	45.8
AlexNet	28.08	40.87
Proposed method	20.26	25.32

The word error rate (WER) is a typical metric of machine translation or speech recognition system performance**. Fig 10** depicts a performance comparison of the proposed approach with current techniques regarding word mistake rate. The graph illustrates that the suggested V3-inception model surpassed previous techniques in terms of word mistake rate. Existing algorithms with WERs of 32.63, 33.91, 33.69, 30.31, and 28.08 are SVM, RF, ResNet, VGG-16, and AlexNet, respectively.

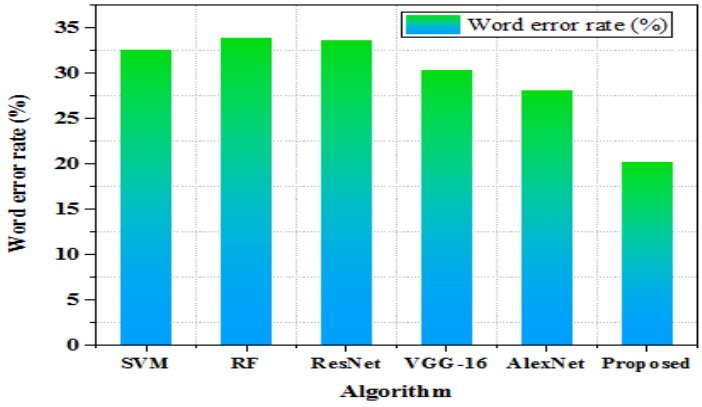

Fig 10. Performance Comparison of Word Error Rate

Fig 11 compares the suggested approach with existing methods in terms of simulation time. The time it takes to compete this operation is called simulation or execution time. The proposed v3 startup paradigm, as shown in the figure, takes substantially less time (25.32 seconds) to complete. Existing methods, including SVM, RF, ResNet, VGG-16, and AlexNet, have simulation times of 120.7 s, 80.5 s, 60.3 s, 458 s, and 40.87 s, respectively.

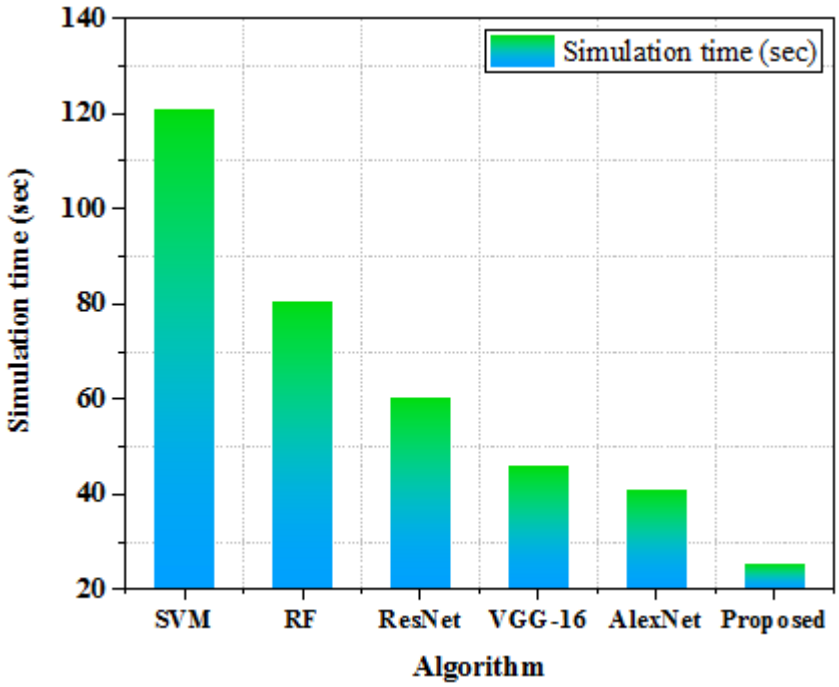

Fig 11. Performance Comparison of Simulation Time

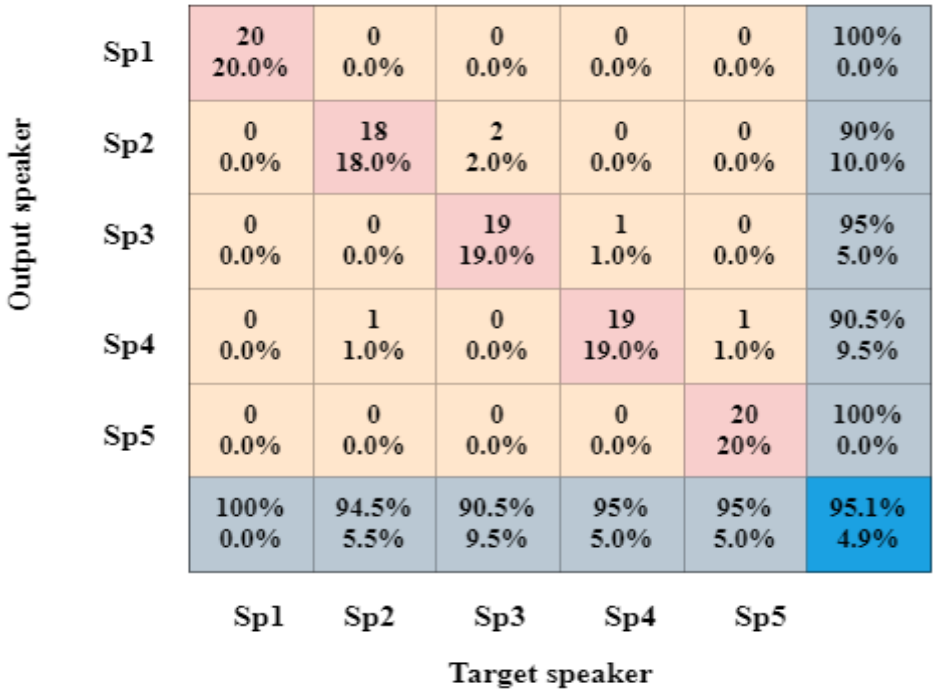

Fig 12. Confusion Matrix of Voice Authentication

The success rate of the classification model is one of many factors considered in its evaluation. This procedure necessitates the use of many factors. A confusion matrix is a table needed to determine these parameters. Displays the category to which

every statistic in the confusion matrix dataset belongs. Calculating this table's values yields several parameters and information about the classifier's performance.

Fig 12 depicts the confusion matrix derived from the categorization results. Each speaker (labeled Sp1 through Sp5) gets 20 samples to test. Overall, the categorization rate was 95.1% .**Fig 13** depicts the suggested model's model accuracy and a plot of epochs vs accuracy. The training and validation graphs are depicted in the picture. **Fig 14** depicts the suggested model's loss curves, incorporating training and validation losses.

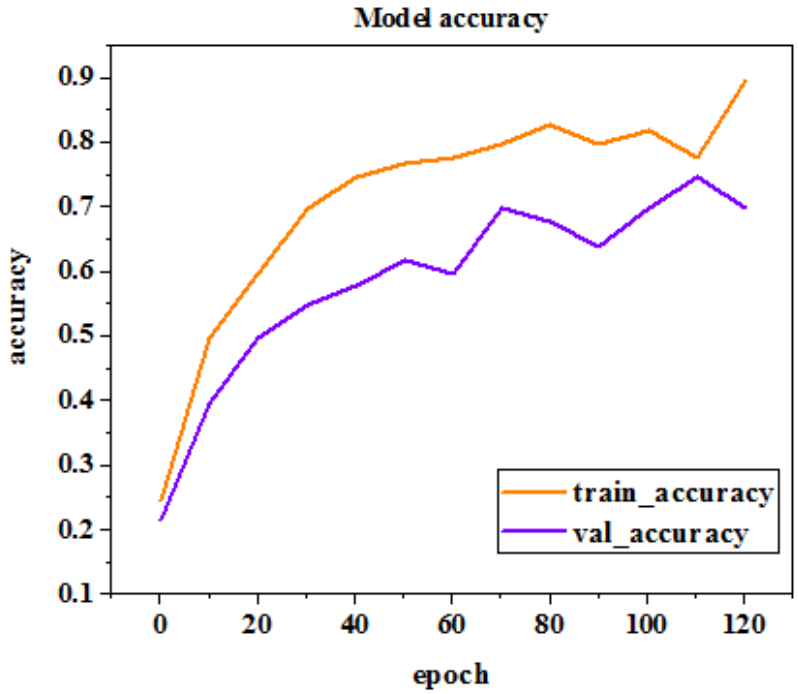

Fig 13. Accuracy Curve of Proposed Model

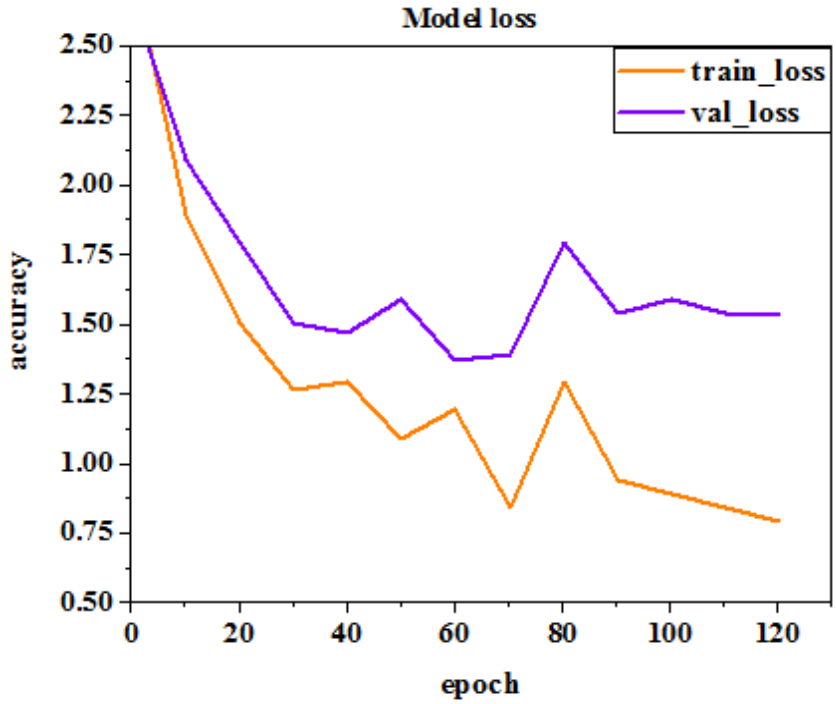

Fig 14. Loss Curve of Proposed Model

Table 7. False Positive Rate Comparison

Algorithm	FPR (%)
SVM	70.8
RF	68
ResNet	60
VGG-16	55
AlexNet	50
Proposed method	45

Furthermore, we assess the false positive findings of the chosen methodologies. **Fig 15** shows that the suggested strategy has the fewest mistakes. Furthermore, efficient parallel computing algorithms reduce mistakes in feature selection and extraction from sound data dramatically. Overfitting is a significant issue in training, and practically all machine learning approaches will suffer. We attempt to limit the danger of overfitting by employing feature selection approaches that prioritize the value of current attributes in the dataset and remove less important ones (without producing new ones). The FPR for current techniques, including SVM, RF, ResNet, VGG-16, and AlexNet, are respectively 70.8%, 68%, 60%, 55%, and 50%.

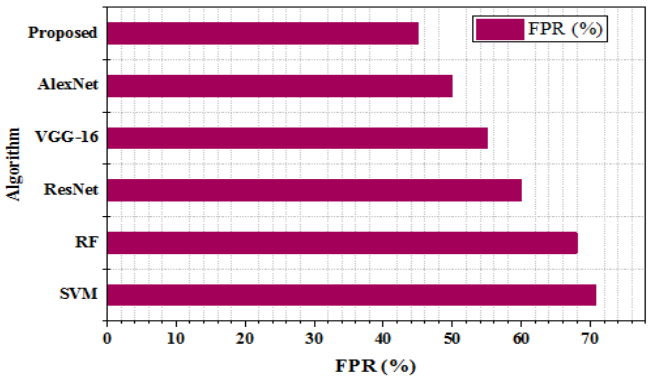

Fig 15. Evaluation of FPR (%)

V. CONCLUSION

In this paper, we offer Inception v3, a method for user authentication using a modified deep learning algorithm. The W-WPCC function is used in the user authentication model. The architecture of the fully connected layer is updated in this study to enhance the accuracy of image classification employing audio signals, similar to the fundamental framework of the InceptionV3 approach, which has outstanding classification capabilities. We tested the user authentication model using speech data from registered users in various situations and discovered that it accurately differentiates every registered user. The research was conducted with five English speakers whose voices had been retrieved from YouTube.

The results reveal that, compared to current methods, the suggested enhanced method based on Inception v3 (trained on speech spectrogram pictures) performs the best. This method's tests produced an average categorization result that was 99% accurate. The suggested approach is suitable for text-independent algorithms that require only short speech utterances as input.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding was received to assist with the preparation of this manuscript.

Ethics Approval and Consent to Participate

The research has consent for Ethical Approval and Consent to participate.

Competing Interests

There are no competing interests.

Reference

[1]. H. Park and T. Kim, "User Authentication Method via Speaker Recognition and Speech Synthesis Detection," Security and Communication Networks, vol. 2022, pp. 1–10, Jan. 2022, doi: 10.1155/2022/5755785.

[2]. S. K. Wong and S. M. Yiu, "Location Spoofing Attack Detection with Pre-Installed Sensors in Mobile Devices," Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), vol. 11, no. 4, pp. 16–30, Dec. 2020, doi: 10.22667/JOWUA.2020.12.31.016.

[3]. A. S. Kitana, T. Issa, and W. G. Isaac, "Towards an Epidemic SMS-based Cellular Botnet," Journal of Internet Services and Information Security (JISIS), vol. 10, no. 4, pp. 38–58, Nov. 2020, doi: 10.22667/JISIS.2020.11.30.038.

[4]. G. S. Kasturi, A. Jain, and J. D. Singh, "Detection and Classification of Radio Frequency Jamming Attacks using Machine learning," Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), vol. 11, no. 4, pp. 49–62, Dec. 2020, doi: 10.22667/JOWUA.2020.12.31.049.

[5]. A. L. Marra, F. Martinelli, F. Mercaldo, A. Saracino, and M. Sheikhalishahi, "A Distributed Framework for Collaborative and Dynamic Analysis of Android Malware," Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), vol. 11, no. 3, pp. 1–28, Sep. 2020, doi: 10.22667/JOWUA.2020.09.30.001.

[6]. D. Berbecaru, A. Lioy, and C. Cameroni, "Supporting Authorize-then-Authenticate for Wi-Fi access based on an Electronic Identity Infrastructure," Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), vol. 11, no. 2, pp. 34– 54, June. 2020, doi: 10.22667/JOWUA.2020.06.30.034.

[7]. S. H. K. Wong and S. M. Yiu, "Identification of device motion status via Bluetooth discovery," Journal of Internet Services and Information Security (JISIS), vol. 10, no. 4, pp. 59–69, Nov. 2020, doi: 10.22667/JISIS.2020.11.30.059.

[8]. J. A. Unar, W. C. Seng, and A. Abbasi, "A review of biometric technology along with trends and prospects," Pattern Recognition, vol. 47, no. 8, pp. 2673–2688, Aug. 2014, doi: 10.1016/j.patcog.2014.01.016.

[9]. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, no. 1–3, pp. 19–41, Jan. 2000, doi: 10.1006/dspr.1999.0361.

[10]. D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 72–83, 1995, doi: 10.1109/89.365379.

[11]. N. H. Tandel, H. B. Prajapati, and V. K. Dabhi, "Voice Recognition and Voice Comparison using Machine Learning Techniques: A Survey," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Mar. 2020, doi: 10.1109/icaccs48705.2020.9074184.

[12]. H. Y. Khdier, W. M. Jasim, and S. A. Aliesawi, "Deep Learning Algorithms based Voiceprint Recognition System in Noisy Environment," Journal of Physics: Conference Series, vol. 1804, no. 1, p. 012042, Feb. 2021, doi: 10.1088/1742-6596/1804/1/012042.

[13]. K. Aizat, O. Mohamed, M. Orken, A. Ainur, and B. Zhumazhanov, "Identification and authentication of user voice using DNN features and i-vector," Cogent Engineering, vol. 7, no. 1, p. 1751557, Jan. 2020, doi: 10.1080/23311916.2020.1751557.

[14]. T. Zeng, "Deep Learning in Automatic Speech Recognition (ASR): A Review," Proceedings of the 2022 7th International Conference on Modern Management and Education Technology (MMET 2022), pp. 173–179, Dec. 2022, doi: 10.2991/978-2-494069-51-0_23.

[15]. A. Alsobhani, H. M. A. ALabboodi, and H. Mahdi, "Speech Recognition using Convolution Deep Neural Networks," Journal of Physics: Conference Series, vol. 1973, no. 1, p. 012166, Aug. 2021, doi: 10.1088/1742-6596/1973/1/012166.

[16]. R. Zheng, Y. Fang, and J. Dong, "Voice Print Recognition Check-in System Based on Resnet," Highlights in Science, Engineering and Technology, vol. 16, pp. 98–108, Nov. 2022, doi: 10.54097/hset.v16i.2473.

[17]. F. Ye and J. Yang, "A Deep Neural Network Model for Speaker Identification," Applied Sciences, vol. 11, no. 8, p. 3603, Apr. 2021, doi: 10.3390/app11083603.

[18]. Bella, J. Hendryli, and D. E. Herwindiati, "Voice Authentication Model for One-time Password Using Deep Learning Models," Proceedings of the 2020 2nd International Conference on Big Data Engineering and Technology, Jan. 2020, doi: 10.1145/3378904.3378908.

[19]. T. Muruganantham, N. R. NAGARAJAN, and R. Balamurugan, "Biometric Of Speaker Authentication Using CNN,". 13. 1417-1423.

[20]. S. Duraibi, W. Alhamdani, and F. T. Sheldon, "Voice Feature Learning using Convolutional Neural Networks Designed to Avoid Replay Attacks," 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Dec. 2020, doi: 10.1109/ssci47803.2020.9308489.

[21]. A. Shafik et al., "Speaker identification based on Radon transform and CNNs in the presence of different types of interference for Robotic Applications," Applied Acoustics, vol. 177, p. 107665, Jun. 2021, doi: 10.1016/j.apacoust.2020.107665.

[22]. A. B. Abdusalomov, F. Safarov, M. Rakhimov, B. Turaev, and T. K. Whangbo, "Improved Feature Parameter Extraction from Speech Signals Using Machine Learning Algorithm," Sensors, vol. 22, no. 21, p. 8122, Oct. 2022, doi: 10.3390/s22218122.

[23]. W. Jia, L. Dongmei, "A review of deep learning applications in speech recognition," Computer Knowledge and Technology, 13(16): 191-197, 2020.

[24]. M. Han, T. Roubing, Z. Yi, et al. "Survey on Speech Recognition," Computer Systems & Applications, 31(1):1−10, 2022.

[25]. M. Wollmer, F. Eyben, B. Schuller, and G. Rigoll, "A multi-stream ASR framework for BLSTM modeling of conversational speech," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2011, doi: 10.1109/icassp.2011.5947444.

[26]. Wen-kai Lu and Qiang Zhang, "Deconvolutive Short-Time Fourier Transform Spectrogram," IEEE Signal Processing Letters, vol. 16, no. 7, pp. 576–579, Jul. 2009, doi: 10.1109/lsp.2009.2020887.

[27]. Y. Huang, K. Tian, A. Wu, and G. Zhang, "Feature fusion methods research based on deep belief networks for speech emotion recognition under noise condition," Journal of Ambient Intelligence and Humanized Computing, vol. 10, no. 5, pp. 1787–1798, Dec. 2017, doi: 10.1007/s12652-017-0644-8.

[28]. J. Cao, M. Yan, Y. Jia, X. Tian, and Z. Zhang, "Application of a modified Inception-v3 model in the dynasty-based classification of ancient murals," EURASIP Journal on Advances in Signal Processing, vol. 2021, no. 1, Jul. 2021, doi: 10.1186/s13634-021-00740-8.

[29]. Q. Zou, Y. Cao, Q. Li, C. Huang, and S. Wang, "Chronological classification of ancient paintings using appearance and shape features," Pattern Recognition Letters, vol. 49, pp. 146–154, Nov. 2014, doi: 10.1016/j.patrec.2014.07.002.

[30]. S. Raj, P. Prakasam, and S. Gupta, "Audio signal quality enhancement using multi-layered convolutional neural network based auto encoder–decoder," International Journal of Speech Technology, vol. 24, no. 2, pp. 425–437, Jan. 2021, doi: 10.1007/s10772-021-09809-z.