World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

Check for updates

Voice recognition by deep transfer learning and vision transformers to secure voice authentication

Nayem Uddin Prince [1] *, Abdullah Al Masum [2], Salman Mohammad Abdullah [3] and Touhid Bhuiyan [4]

[1] Information Technology (2022), Washington University of Science and Technology, USA.
[2] Information Technology (2024), Westcliff University Irvine, USA.
[3] Information Technology (2023), Washington University of science and technology, USA.
[4] Cyber Security School of Information Technology Washington University of Science and Technology Virginia, USA.

World Journal of Advanced Research and Reviews, 2024, 23(03), 1365–1377

Publication history: Received on 02 August 2024; revised on 10 September 2024; accepted on 12 September 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.23.3.2781

Abstract

Speech recognition is crucial for ensuring the security of personal devices and financial transactions. Attaining high accuracy and robustness in voice authentication is challenging due to the presence of voice and environmental variability. Recent advancements in the field of deep learning, particularly in transfer learning and visual transformers, have the potential to enhance voice recognition systems. This study employs advanced deep transfer learning techniques, including Vision Transformers (ViT), VGG16, and a customized Convolutional Neural Network (CNN), to enhance the accuracy and security of speech authentication. The objective is to evaluate and contrast various solutions' voice recognition and authentication accuracy. The experiment included 3000 voice samples, with an equal distribution of 1500 samples from male participants and 1500 from female participants. The dataset was used to train Vision Transformers, VGG16 with transfer learning, and a custom CNN. The models were assessed based on their accuracy in identifying and authenticating voice samples. The VGG16 model achieved the highest level of accuracy in speech recognition, with a precision rate of 95%. The Vision Transformer and custom CNN exhibited satisfactory performance. However, VGG16 demonstrated higher accuracy. The most accurate voice authentication model studied is the VGG16 model based on transfer learning. This study suggests that the security and reliability of voice recognition systems can be enhanced through the use of deep learning techniques.

Keywords: Voice recognition; VGG16; CustomCNN; Vit; honey trap; webform; Cybercrime; Vision Transform; MFCCs.

1. Introduction

In the past several years, an alarming rise in spam and fraudulent calls has been observed in Bangladesh. Many of these calls originate from intricate schemes designed to exploit individuals who aren't vigilant enough. A disturbing trend is the increase of so-called "honey trap" scams, where victims are lured into vulnerable situations by deceptive phone calls or text messages, leading to financial losses, identity theft, or other forms of exploitation. One of the most prevalent types of cybercrime in Bangladesh is spam calls and phishing emails. Scammers now have more ways than ever to trick individuals all throughout the country, thanks to the widespread availability of cell phones and the internet. Unwanted phone calls often take the form of seemingly legitimate job offers, loan arrangements, or other tempting promotions. The con artists who make these calls use deceptive methods to trick people into falling for their schemes. Since 2023, the number of spam and fraudulent phone reports received by the authorities in Bangladesh has increased dramatically. [1] Not only do these calls put people at risk, but they also interrupt them without warning. In addition to severe money losses, several victims of these scams have reported feeling threatened for their safety. The "work-from-home" fraud that emerged during and after the COVID-19 epidemic is one example of these techniques. Scammers prey on people looking for remote jobs by promising them easy money. When victims part with their details or pay for the proposals in

* Corresponding author: Nayem Uddin Prince

advance, they put themselves at risk of financial fraud or identity theft. Many of these scams are worldwide in scope, making it difficult for local law enforcement to tackle the issue. In a "honey trap" situation, the perpetrator lures the victim into an intimate connection with the express purpose of financially or otherwise taking advantage of them. The perpetrators of these scams in Bangladesh are increasingly using technology, such as social media, messaging apps like WhatsApp, and direct phone calls, to contact their victims. Scammers con individuals into talking to them to get them to divulge personal information or money by making them believe they are talking to someone charming. Reportedly, numerous individuals fell prey to honey trap scams in the year 2023. [1][2] In the beginning, scammers often mean well, hoping to win over the victim's confidence and form an emotional bond. Subtly claiming to be in a difficult situation, they may begin to solicit financial assistance. However, they could use potentially humiliating photos or information from their interactions as a kind of blackmail. Victims of these scams often suffer severe psychological suffering as a consequence of their humiliation and betrayal, which prevent them from seeking help or reporting the occurrence. Victims of honeytrap scams often end up losing a substantial amount of money, if not their entire life savings. Many people are impacted by these dishonest acts. An investigation from 2023 found that numerous individuals in Bangladesh have lost millions of Takas as a result of these scams. Even though they may not be as used to dealing with cyber threats, people living in rural areas are just as susceptible as those living in cities. Law enforcement groups have recorded a 30% increase in reports of spam calls and honey traps in the past two years. [2] This growth is due to two factors: the public's growing reliance on digital communication channels and the increased sophistication of scammers. To combat these growing threats, both individuals and authorities must establish more robust security measures. This means that individuals should exercise caution when communicating via phone or text message with unknown numbers, especially if the caller seems to be trying to collect personal information or sends an unsolicited offer.[3] Lawmakers must strengthen cyber regulations and strengthen ties with international groups if they are serious about apprehending those responsible for these scams. Furthermore, public awareness efforts must be launched to inform the general population of the risks associated with unsolicited messages and the techniques used by con artists. Raising awareness of these crimes and providing victims with the support they need may help reduce their occurrence and the harm they cause to society. Spam, phishing, and honey traps are major concerns for the people of Bangladesh. Since these scams are constantly evolving to become more sophisticated, individuals and authorities must remain vigilant. The public may be better protected from being victims of these crimes if public education, more rules, and improved security measures were put into place.

This piece of writing touches on the following topics. You will discover synopses of significant articles in Section II. Section III outlined the steps to take. Section V evaluates our model, while Section IV delves more into the experimental findings. The following procedures are examined in Section VI.

2. Literature review

Deep learning and ML techniques can handle image processing problems. Efforts to raise the standard of living for individuals involved are crucial. Although we introduced a fresh perspective, some studies have previously employed methods that are very similar to ours. The differences are brought to light in two studies that compare:

Lee el. At. [4] To enhance voice recognition, they recommend using Adaptive MFCC in conjunction with Deep Learning. Voice recognition is made better by removing audio data from the initial stream. Nevertheless, current methods that lessen band noise diminish the quality of audio transmissions. In contrast to the current MFCC, the proposed filter compactly accumulates in the data density area to lessen data loss and impose the weighted value on the data area. The recognition rate is enhanced by avoiding data loss. Also, voice recognition independent of DB is now possible using Deep Learning.

Tandel el. At. [5] Speaker detection and voice comparison methods, both conventional and deep learning-based methods, are the focus of this research review. This study also uses public datasets for speaker detection and voice comparison. This little effort may aid speech detection and comparison researchers and beginners.

Suparatpinyo el. At. [6] They want voice-data self-assessment software. This study manipulated psychiatric hospital statistics to cheer up gloomy individuals. University student interviews and non-depressive speech show negative class statistics. To confirm that DCT does not lose spectrum information, researchers compared applied DCT and non-DCT spectrographs. Which impacts depression diagnosis ResNet-based deep learning categorization testing. Count ResNet-34, ResNet-50, and ResNet-101 convolution layers. Both ResNet-50 models trained on different spectrographs had F1 scores above 70% in experiments.

Ibrahim el. At. [7] Assessing the dependability of voice recognition security is done using the CNN method, which is a deep learning technique. Learning is enhanced by the CNN algorithm, which is safer, faster, and more accurate. By contrasting the expected and actual values, the results display the CNN algorithm's performance (the confusion matrix). At 15000 iterations, the best accuracy is 96.87% for sound files, 96.30% for 12000, and 95.77% for 6000. After 15000 repetitions of the voice file, CNN discovered a high level of accuracy.

Lowit el. At. [8] An innovative convolutional neural network (CNN)-based method for identifying suspicious voice patterns is shown here. The network is fed spectrograms of both normal and disordered speech by the one-of-a-kind technology. Pre-trained CNN weights are made using CDBN. This generative model investigates the structure of the input data using statistical methods. A CNN's weights are fine-tuned by supervised back-propagation learning. Using this deep learning method, we can see that even with very little data, we can get decent classification results.

Buyukyilmaz el. At. [9] A deep learning model called Multilayer Perceptron can identify voice gender. The collection contains 3,168 male and female voice samples. Acoustic analysis creates samples. An MLP deep learning system detected gender-specific characteristics. Our model is 96.74 percent accurate on test data. The gender of voice recognition interactive website was created.

Hamdani et al. [10] proposed an algorithm for classifying accents based on speech rhythm. This method may enhance language learning and voice recognition. The precise categorization of the three regional dialects by the method demonstrated that measurements of speech rhythm can be used to determine accents. In addition, they developed a neural network classifier with an accuracy rate of 88%.

Identification analyses are employed to study Bangladesh's linguistic variety, according to Mamun et al. [11]. The MFCC and repeated neural system method explore Bangladeshi voice linguistic variation. Registered nurse. People in Noakhali, Barisal, Mymensingh, Chittagong, Sylhet, Rajshahi, Khulna, and Dhaka (old) are speaking out. Intelligent voice recorders can record at 16 kHz with 16-bit quantization. Barisal led all regions with 83% accuracy.

A group of Nigerian scholars surveyed the three most widely spoken native languages in Nigeria, Yoruba [12]. This bolstered the researcher's faith in the data. Following the necessary preparations, the recording equipment is now prepared. The proposed model, which combines a 1D CNN and LSTM architecture, accurately categorized speakers of Hausa, Igbo, and Yoruba languages with average accuracy rates of 97.7%, 90.6%, and 96.4%, respectively. Categorizing speakers into three distinct groups is beneficial.

Weninger et al. [13] use wavelets with DWPT, DT-CWPT, and WPT-based non-direct highlights to identify speakers and highlights. Results are assessed using MFCC and LPC standards. People and emphasis are identified by k-NN, SVM, and ELM classifiers. ELM classifiers were best at speaker recognition using English numbers (92.16 percent) and Malay words (93.54 percent).

Park et al. [14] ranked regional Mandarin dialects extensively. They analyze Mandarin dialogue data from fifteen Chinese cities to cover the country. The study advocated a bLSTM-emphasized classifier to simplify the transition between simple and complicated Mandarin ASR models and investigated Mandarin accent classifiers. The discourse database has 466 persons and 135k words (84.7 h). This study achieved relative CERRs of 13.2% for Audience A1, 15.3% for A2, and 14.6% for A3 using i-vector speaker correction.

2.1. Comparison with other work

Table 1 shows how our work compares to others. Our study outshines its competitors, according to the table analysis. [15] This study is more accurate than its predecessors. One more thing we didn't anticipate is that our model makes use of really sophisticated algorithms, which are among the most crucial values we've ever come across.

Table 1 Comparison Table

Other work			**Our work**
Author Name	Algorithm	Accuracy	Algorithm & Accuracy
Lee el. At.	MFCC	90.00%	VGG16, CustomCNN, Vision Transformer 95.00%
Suparatpinyo el. at.	ResNet50,	70.00%	
Ibrahim el. at.	CNN	95.00%	
Buyukyilmaz el. at.	MLP	96.74%	

3. Material and methods

Figure 1 depicts the diagram of the methods used in this project. In this section, six key methodologies are implemented to accomplish this job successfully. Initially, we have gathered recordings of both male and female voices. After obtaining this data, I meticulously gathered several types of sentences. The primary focus of our project is data preprocessing, where we employ three factors to convert speech into a format that can be understood by machines. Additionally, it employs classification techniques to create a well-balanced and trained dataset. In the fourth phase, we utilized custom CNN, Vgg16, and Vit models to make predictions on our data. The number five represents the outcome analysis, which is a crucial component of this project. The last component of this project entails evaluation. In this section, we employ a confusion matrix to provide justification for the primary objective of this research. [16]

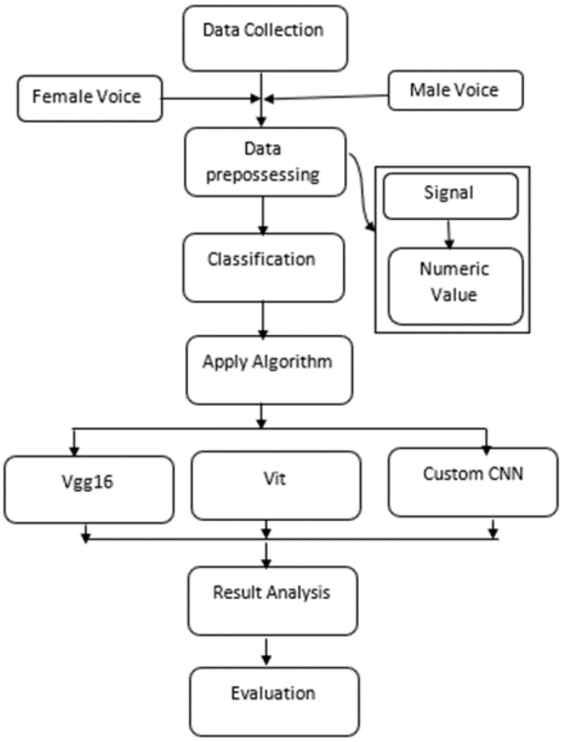

Figure 1 Methodology Diagram

3.1. Data collection and preprocessing

A total of 1500 male and 1500 female voices were recorded specifically for this project. As a result, we gathered a compilation of distinct statements from each individual. For this specific occasion, we utilized the recorder on our personal mobile device. Our research indicates that all data is in the form of voice. However, machines cannot understand voice. Therefore, all data must be made understandable to machines. For this specific case, we utilized three variables to convert our vocal input into numerical data. The parameters of Theo's system are pitch, resonance, and MFCCs. The novelty of this project lies in the utilization of three factors for the subsequent phase.

3.2. Dataset representation

Figure 2 depicts the graphical representation of the dataset. Our dataset demonstrates an equal distribution of male and female voices, with each gender accounting for 50% of the data. The term "project dataset" refers to a dataset that is well-balanced and suitable for use with algorithms.[17]

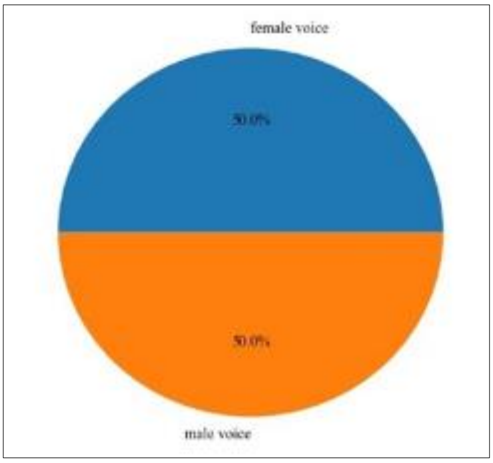

Figure 2 Dataset Representation

3.3. Data Classification

In this section, we have selected the parameters that were utilized in the training dataset. Figures 3 and 4 depict the acoustic waveforms of male and female voices, respectively. The vertical amplitude of the waveform represents the intensity of the sound wave. Greater amplitudes correspond to higher sound intensity, whilst lower amplitudes correspond to lower sound intensity. The waveform represents the sound produced by overlapping many sounds. We analyze the waveform of the identical sentence said by both male and female voices for comparison.

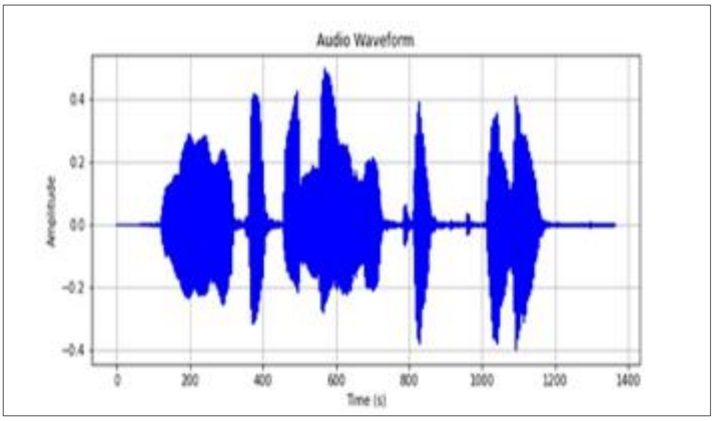

Figure 3 Male voice waveform.

Graphic 3 displays the male waveform. This figure exhibits reduced amplitude distortion, and the majority of the sound waveform exhibits a consistent increase in volume.

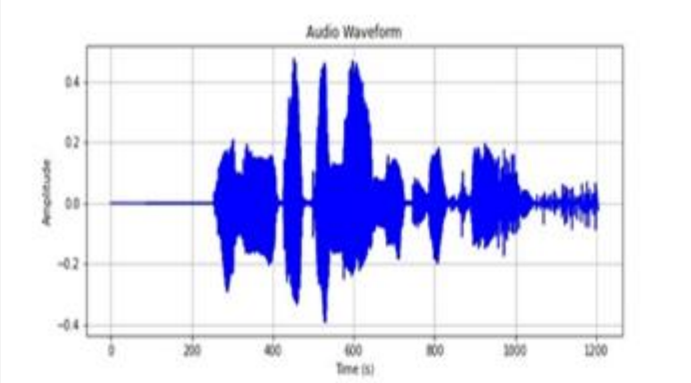

Figure 4 Female voice waveform

The image labeled 4 displays the female waveform. This image exhibits a notable degree of amplitude distortion. The predominant portion of the sound waveform does not indicate whether it should be amplified or attenuated. Amplitude and shape are often distinct from each other.

Both of these graphs exhibit linear behavior. However, in this particular circumstance, we are unable to derive significant information from the data. Therefore, it is necessary to employ another technique.

There is a scale called the Mel scale that utilizes MFCC information. The graphs display the male and female MFCC information, namely in graphs 5 and 6. The MFCC graph consists of a sequence of lines or bars, each representing a specific frequency band spaced on a mel scale. It is used to gather a large amount of voice data.

Figure 5 shows the yellow color in this picture, which represents a high decibel (dB) value of the voice. That indicates a voice that is both high-pitched and kind. In this project, we have gathered more data to enhance its uniqueness and accuracy.

The image depicted in Figure 6 illustrates that the yellow hue corresponds to a higher decibel (dB) value of the voice. This suggests a voice that is characterized by both a higher tone and a gentle demeanor. For this project, we have collected more data to improve its distinctiveness and precision.

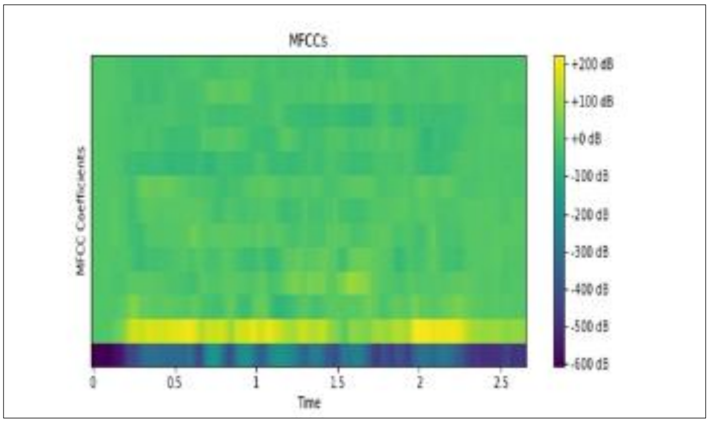

Figure 5 MFCC for male voice

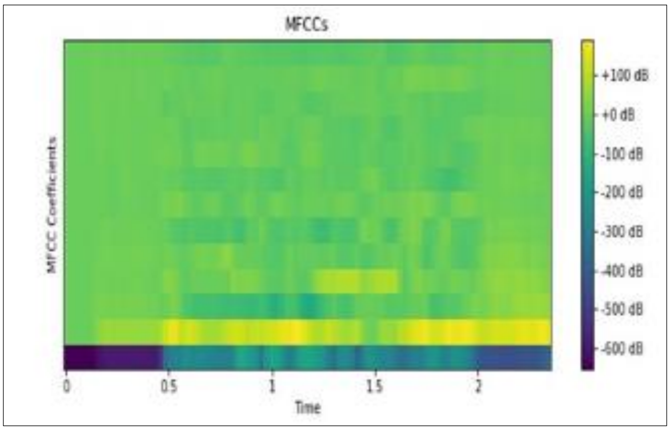

Figure 6 MFCC for female voice

Another piece of information is pitch control, which indicates the range of high and low tones that are perceptible to human ears. Figures 7 and 8 depict the pitch control graphs of female and male voices, respectively. The two graphs are significantly different.

This graph illustrates the pitch contour of a female voice, with a value of 7. The curve line indicates a quick increase, which signifies that the voice is loud and can potentially cause discomfort to our ears. Additionally, this increase in pitch suggests that the voice is becoming hoarser.

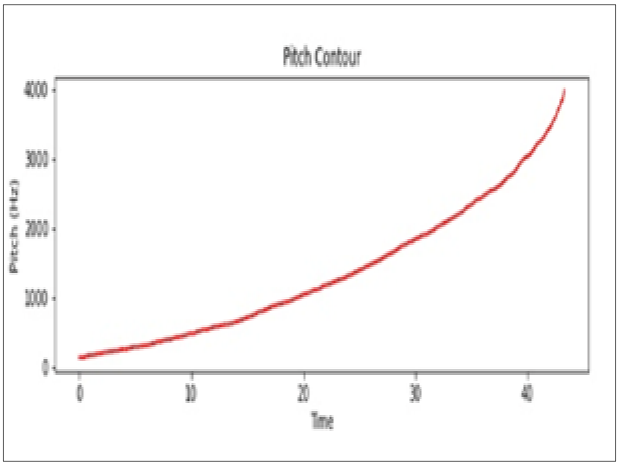

Figure 7 Pitch contour for female voice

The graph depicting the male voice pitch contour at 8 indicates that the curve will progressively steepen, indicating that the voice will gradually become audible and imply that it will be soft.

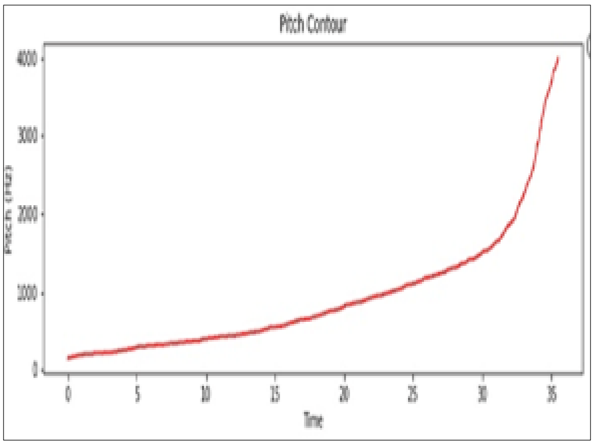

Figure 8 Pitch contour for male voice

3.4. Model selection and algorithms

The Given their architecture and picture classification capabilities, VGG16, CustomCNN, and Vision Transformer (ViT) were selected for this inquiry. An ideal choice for establishing a standard would be the VGG16 model because of its deep architecture and straightforward design. On the other hand, just for voice recognition, the CustomCNN model was developed in [18]. The ViT model, which is built on transformers, uses attention techniques to capture global image dependencies.

3.5. Evaluation

The three models were given 80% of the preprocessed dataset for training, 10% for validation, and 10% for test sets. We were able to fine-tune the performance of each model by adjusting the learning rates, batch sizes, and epoch counts. To improve the gradient descent approach during training, the Adam optimizer and categorical cross-entropy loss function were used. To measure the effectiveness of the model, several measures were employed, such as F1-score, recall, accuracy, and precision. The phenomenal achievement of 95.00% precision, recall, and F1 score in male/female speech detection and classification was achieved by the VGG16.

3.6. Implementation

A web-based interface and an AI-based platform are prerequisites for the VGG16 paradigm. Cybersecurity, scam, and bogus call detection can all benefit from this all-encompassing method, which entails meticulous data collecting, preprocessing, advanced model training, and actual implementation.

4. Results and discussion

Three deep-learning algorithms are utilized in this section. Examples are ViT, CustomCNN, and VGG16. We similarly employed three matrices to refine our algorithm for this method. Gains were achieved by the majority of the algorithms. Table 2 represents the accuracy of this project. In this table, we take three matrices: precision, recall, and f1-score. For this project, we applied three algorithms: ViT, Vgg16, and Custom CNN. Vit gained 0.91% for three matrixes. Custom CNN shows 0.92% for all three matrixes. VGG16 achieved precision, recall, and f1-score is 0.95%. The VGG16 algorithm gained the highest accuracy of all other algorithms.

Table 2 Accuracy table

Algorithms	precision	Recall	F1-score
Vit	0.91	0.91	0.91
VGG16	0.95	0.95	0.95
Custom CNN	0.92	0.92	0.92

4.1. Classification Report

The Vit classification for our two labels female voice and male voice is displayed in Table 3. The female voice gained a 0.92 recall rate, as shown in this table. A man's voice increased in accuracy by 0.93. With 303 pieces of supplementary data, the overall accuracy of the Vit algorithm was 0.92. As a result, macro and weight averages score similarly.

Table 3 Classification table for custom CNN

Label	precision	Recall	F1-score	Support
Female voice	0.90	0.92	0.91	142
Male voice	0.93	0.91	0.92	161
Accuracy	0.92			303
Macro avg	0.92	0.92	0.92	303
Weight avg	0.92	0.92	0.92	303

Table 4 displays the customized Convolutional Neural Network (CNN) classification results for our two voice labels: female and male. The table displays that the female voice achieved a score of 0.92 in all three matrices. The male voice increased by 0.93, resulting in three matrices. The unique CNN algorithm obtained an accuracy of 0.92, using a support data size of 303. The macro and weighted average scores are also reported as the outcome.

Table 4 Classification table for vit

Label	precision	Recall	F1-score	Support
Female voice	0.92	0.92	0.92	142
Male voice	0.93	0.93	0.93	161
Accuracy	0.92			303
Macro avg	0.92	0.92	0.92	303
Weight avg	0.92	0.92	0.92	303

The VGG16 classification for our two labels, female voice, and male voice is displayed in Table 5. The female voice gained a 1.00 precision rate, as shown in this table. A man's voice increased in accuracy by 1.00 as recall rate. With 303 pieces of supplementary data, the overall accuracy of the Vit algorithm was 0.96. As a result, macro and weight averages score similarly. This VGG16 algorithm also performed the highest accuracy for classification reports. So, we take this algorithm for final implementation

Table 5 Classification table for vgg16

Label	precision	Recall	F1-score	Support
Female voice	1.00	0.91	0.95	142
Male voice	0.92	1.00	0.96	161
Accuracy	0.96			303
Macro avg	0.96	0.95	0.96	303
Weight avg	0.96	0.96	0.96	303

4.2. Evaluation

Figure 9. shows the accuracy of training and validation over 21 epochs. As the training accuracy increases, the model learns from the data and gets closer and closer to 0.9. A sharp decline in validation accuracy is an indication of overfitting, a condition in which VGG16 does well on the training data but not well enough on new, unknown data. There seems to be an overfitting issue with the model as the gap between training and validation accuracy widens, particularly towards the end. This could be remedied by using early halting or regularisation in order to improve the model's performance on unknown data.

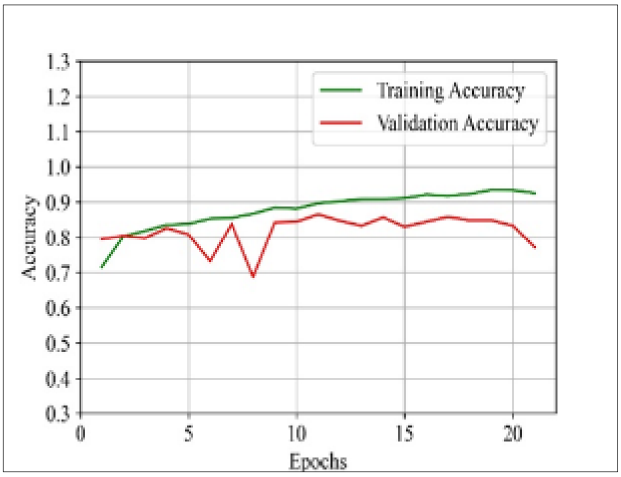

Figure 9 Training vs Validation Accuracy for VGG16.

Training and validation losses are shown in the graph for each of the 21 epochs. Losses are decreasing, indicating model improvement in prediction. Green and decreasing training loss indicate consistent improvement. In contrast, the validation loss (red) is more volatile, with huge changes between epochs 15 and 21, when it increases significantly. Even though it performed better on the training set, the model struggles to generalize to the validation set, perhaps due to overfitting. The large validation loss at the end suggests that the model started remembering the training data instead of learning generalizable patterns, which could explain why it performs badly on novel data. Regularization or early halting may solve this problem.

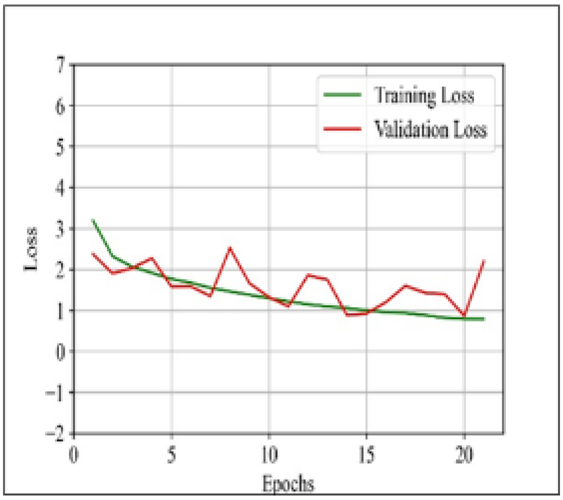

Figure 10 Training vs Validation loss for VGG16.

The confusion matrix shows in Figure 11 how well the voice recognition computer distinguishes men and women. The matrix has four corners: Twenty true positives correctly recognized female noises are in the top left corner. The upper right quadrant had two false negatives (female sounds misinterpreted as male). Zero false positives (male voices misidentified as female): good male voice classification (bottom-left quadrant). In the bottom-right quadrant, 23 male noises were accurately identified as true negatives. From this matrix, the model accurately reflects male and female voices, indicating success. Only two female voice misclassifications suggest a slight imbalance, but the model is durable.[19]

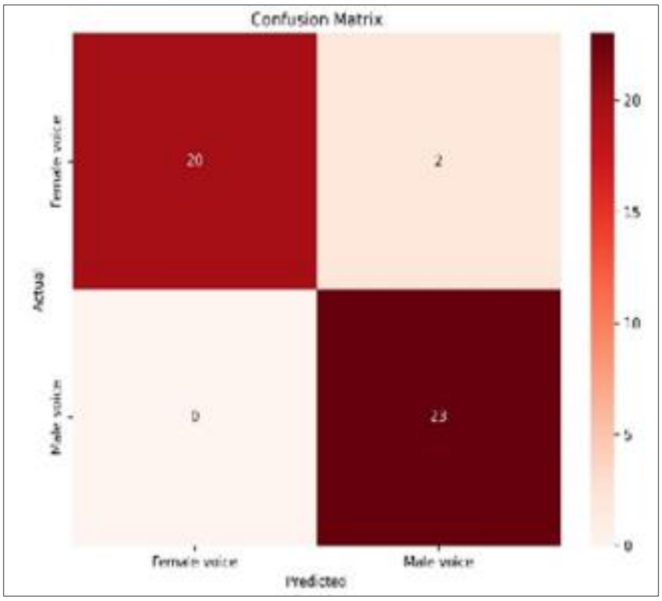

Figure 11 Confusion Matrix.

Figure 12 shows the evaluation graph in this figure. Our selected Algorithm's model, VGG16, performed very well. [20] The blue bar refers to the real data that has not been trained before, and the orange bar is predicted data. For the male voice, our model acts very well; it can detect all real data. For female voice, our model detects almost the same as real data; only two fewer data cannot be predicted. After analysis of this graph, it can be said that our model performed outstandingly.

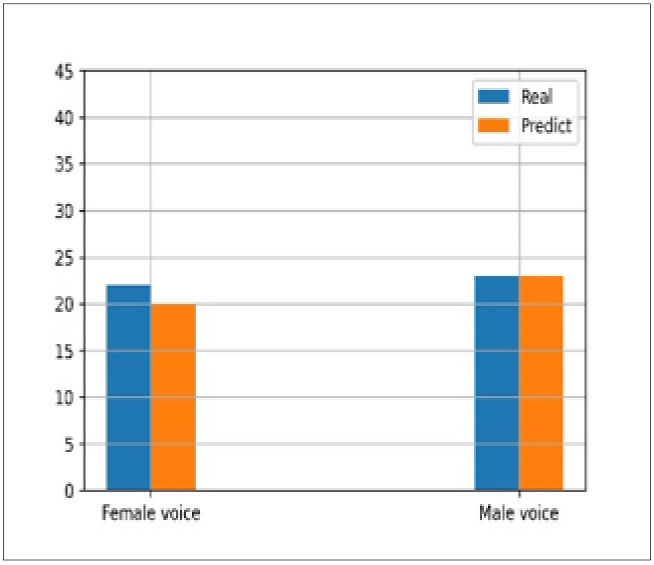

Figure 12 Evaluation Graph

4.3. Implementation

Our project has accomplished a tremendous deal. A good project should aim toward real-life applications. Here, we implement our label detection using OpenCV. [21] Our model readily detects all of our voice samples. Each label, along with its name, is detected by our vgg16 model in a green box.[22]

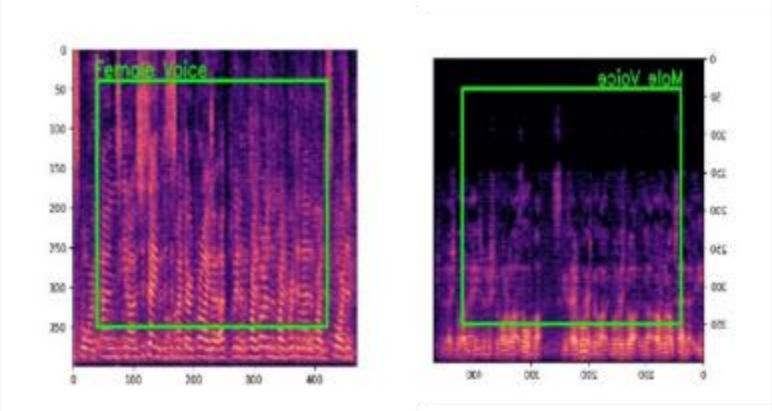

Figure 13 Implementation of VGG16

5. Conclusion

Using deep transfer learning and vision transformers (ViT), this study evaluated the security of voice authentication. An analysis was conducted on a dataset consisting of 1500 voices, both male and female, utilizing ViT, VGG16, and a custom CNN. The model with the best accuracy (0.95%) was VGG16. Overfitting was evident in the validation accuracy and loss graphs despite the models' strong performance on training data. Models, especially VGG16, may struggle with generalizing to unknown data, which is crucial for effective voice authentication, according to variations in validation metrics. It is possible to train more generalizable models with a dataset that includes voices from a wider range of ages, dialects, and environments. By using the spatial awareness of CNN and the global context comprehension of transformers, hybrid models that combine ViT and CNN strengths have the potential to enhance the accuracy of feature extraction and classification. This study's speech authentication dataset could be improved upon in future studies by using cross-domain transfer learning to models that have already been trained on big, diverse voice datasets.

References

[1] M. H. Marof, “Rise in work-from-home job scams in Bangladesh: What you need to know,” The Business Standard, Sep. 22, 2023. Available from https://www.tbsnews.net/bangladesh/crime/rise-work-home-job-scams-bangladesh-what-you-need-know-705162

[2] HT News Desk, “WhatsApp honey trap scam: What is it? How does it work and how to stay safe?,” Hindustan Times, Mar. 21, 2024. Available from https://www.hindustantimes.com/business/whatsapp-honey-trap-scam-what-is-it-how-does-it-work-and-how-to-stay-safe-101711015746506.html

[3] Tamal, M. A., Islam, M. K., Bhuiyan, T., Sattar, A., & Prince, N. U. (2024). Unveiling suspicious phishing attacks: enhancing detection with an optimal feature vectorization algorithm and supervised machine learning. Frontiers in Computer Science, 6, 1428013.

[4] H. -S. Bae, H. -J. Lee and S. -G. Lee, "Voice recognition based on adaptive MFCC and deep learning," 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), Hefei, China, 2016, pp. 1542-1546, doi: 10.1109/ICIEA.2016.7603830.

[5] 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), Coimbatore, India, 2020, pp. 459-465, doi: 10.1109/ICACCS48705.2020.9074184.

[6] Suparatpinyo, S., & Soonthornphisaj, N. (2023). Smart voice recognition based on deep learning for depression diagnosis. Artificial Life and Robotics, 28(2), 332-342.

[7] Ibrahim, W., Candra, H., & Isyanto, H. (2022). Voice recognition security reliability analysis using deep learning convolutional neural network algorithm. Journal of Electrical Technology UMY, 6(1), 1-11.

[8] Wu, H., Soraghan, J., Lowit, A., & Di Caterina, G. (2018, September). A deep learning method for pathological voice detection using convolutional deep belief networks. In Interspeech 2018.

[9] Buyukyilmaz, M., & Cibikdiken, A. O. (2016, December). Voice gender recognition using deep learning. In 2016 International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA2016) (pp. 409-411). Atlantis Press.

[10] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget and Z. Smekal, "Voice Pathology Detection Using Deep Learning: a Preliminary Study," 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI), Funchal, Portugal, 2017, pp. 1-4, doi: 10.1109/IWOBI.2017.7985525.

[11] Droua-Hamdani G: Classification of regional accent using speech rhythm metrics. In: Salah, A. A., et al. (eds.), SPECOM 2019, LNAI 11658, pp. 75–81 (2019)

[12] Mamun, R.K., Abujar, S., Islam, R., Badruzzaman, K.B.M., Hasan, M.: Bangla speaker accent variation detection by MFCC using recurrent neural network algorithm: a distinct approach. In: Saini, H., Sayal, R.,

[13] Salau, A.O., Olowoyoand, T.D., Akinola, S.O.: Accent Classification of the Three Major Nigerian Indigenous Languages Using 1DCNN LSTM Network Model, (2020). Springer Nature, Singapore Pte Ltd.

[14] Weninger, F., Sun, Y., Park, Y., Willett, D., Zhan, P.: Deep Learning based Mandarin Accent Identification for Accent Robust ASR (2019) ISCA.

[15] Jiao, Y., Tu, M., Berisha, V., Liss, J.: Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features (2019) ISCA.

[16] J. Park, F. Diehl, M. J. F. Gales, M. Tomalin and P. C. Woodland, "Training and adapting MLP features for Arabic speech recognition," 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 2009, pp. 4461-4464, doi: 10.1109/ICASSP.2009.4960620.

[17] Buyya, R., Aliseri, G. (eds.), Innovations in computer science and engineering. Lecture notes in networks and systems, vol. 103 (2020). Springer, Singapore.

[18] He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep Residual Learning for Image Recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770-778.

[19] Simonyan, K., & Zisserman, A. (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition." International Conference on Learning Representations.

[20] Hridoy, R. H., Arni, A. D., & Haque, A. (2023). Improved vision-based diagnosis of multi-plant disease using an ensemble of deep learning methods. International Journal of Electrical and Computer Engineering (IJECE), 13(5), 5109-5117.

[21] Chowdhury, R. H., Prince, N. U., Abdullah, S. M., & Mim, L. A. (2024). The role of predictive analytics in cybersecurity: Detecting and preventing threats. World Journal of Advanced Research and Reviews, 23(2), 1615-1623.

[22] M. A. Rahman, A. A. Khan, M. M. Hasan, M. S. Rahman and M. T. Habib, "Deep Learning Modeling for Potato Breed Recognition," in IEEE Transactions on AgriFood Electronics, doi: 10.1109/TAFE.2024.3406544.