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

Cyber threat detection using voice and speech analysis

Kaosar Hossain [1,*], Sufia Zareen [2] and Sahadat Khandakar [3]

[1] *Student, BSc in Computer Science, American International University-Bangladesh.*
[2] *Student, Masters in Genetics, Osmania University, Hyderabad, India.*
[3] *Student, BSc in Electrical Engineering, BRAC University, Bangladesh.*

World Journal of Advanced Research and Reviews, 2021, 10(03), 508-517

Publication history: Received on 26 April 2021; revised on 19 June 2021; accepted on 28 June 2021

Article DOI: https://doi.org/10.30574/wjarr.2021.10.3.0254

Abstract

Using voice interactions for fraud detection has become an active field of research as the number of fraudulent activities through this channel has increased. Current systems fail to identify threatening voices because usual and malignant voice features are not detected with the conventional methods of analysis. However, in particular, we would examine how Visions Transformer (ViT) models could be explored and their capability of complex pattern recognition in voice signals can be utilised for speech fraud detection. Voiceprints are used not just in the voice-based ID process; they also squeeze more samples out of a system that can otherwise be squirrelly at detecting when users are not who they say they are, something significant given how voiceprint use is on the ascendency for user verification in things like customer service and telecommunications settings. We address this by using ViT, specially trained on male and female voices, to detect light abnormalities in the speech pattern caused by fraud. Figure 18 shows the comparative analysis, which proves the effectiveness of the proposed method, where ViT achieved an accuracy of 95% and also outperforms classic models like CNN in terms of precision, recall, and F1-score. However, they also show signs of overfitting, which is an issue where the model performs very well on training data but fails to generalize to the validation set. In short, ViT seems to hold potential for voice fraud detection; however, additional efforts are necessary in the area of regularization, early stopping, and possibly also data augmentation to prevent overfitting and improve performance when it comes to out-of-sample accuracy.

Keywords: Voice-Based Fraud; Detection Vision Transformer (Vit); Machine Learning; Anomaly Detection; Overfitting; Precision and Recall; Data Augmentation; Model Generalization; Speech Signal Processing

1. Introduction

The recent increase in cyber threats, with particular attention to voice-based deception — like phone scams and impersonations coming up the ranks, knows no limits. A 2021 report by the Federal Trade Commission (FTC) revealed that phone-based fraud complaints were responsible for 32% of total fraud, illustrating the growth of voice-based crime (Federal Trade Commission, 2021). Also, the cybersecurity statistics 2020-2021 point to a 20% rise since last year in identity theft cases committed via voice calls. These numbers underscore the critical necessity of possessing more sophisticated technologies to discover and neutralize these unmistakable threats. Since voice and speech signals are unique (carrying properties that can isolate outliers in authentic versus fraudulent interactions), applying data science tools, specifically machine learning and deep learning for voice analysis, is one prospective way to detect a transaction as suspicious or genuine (Hussain and Rizvi, 2020). Executive summary: Phone fraud represents a potentially ruinous double threat to businesses and consumers around the world. The impact of fraud runs into billions of dollars every year. As such, especially in areas like finance, telecoms, and e-commerce, it is important to have processes in place that drastically reduce the probability of fraud occurring. AI voice and speech analysis combined with most of this fraud detection system, makes for a solid solution. AI models are capable of detecting subtle inconsistencies or patterns that

* Corresponding author: Kaosar Hossain

suggest the authenticity of the speaker's voice, relying also on vocal features peculiar to male and female voices, showing that this type of heuristic can be a powerful tool in cybersecurity (Chen and Xu, 2021).Machine learning and deep learning models such as Vision Transformers (Vitt), VGG16, and Custom CNNs work very well in extracting the patterns out of complex, unstructured data like audio signals. The results of this project, which are presented further, led to excellent performance compared to those obtained while using other approaches. Results. For example, a convolutional neural network (CNN) model, VGG16, achieved the highest precision of 0.96, indicating that this method can ultimately detect fraudulent calls with high accuracy. Credit: Zhang and Li (2020). The other models, Custom CNN and ViT, are also obtaining excellent results here with high F1-scores to make a balance between precision and recall for our fraud detection. By analyzing male and female voice samples, the project intends to solve the fraud detection problem, considering that changing network security protocols are no longer confined to being virtual or real but also encompass who you call and how you answer your phone. The system learns to classify genuine and fraudulent calls by training three models (Vitt, VGG16, Custom CNN) on speech spectrograms. Gender-specific voice analysis can isolate such inconsistencies to stop potentially fraudulent activity. Thus, the deployment of these state-of-the-art ML algorithms provides a cost-effective solution for fraud detection in real-time voice interaction, making it an efficient solution for the telecommunication and customer service industries (Vasilenko and Andreev, 2021). In voice and speech analysis, machine learning can lead to fraud detection and security of a new era in several sectors (Hussain and Rizvi, 2020; Chen and Xu, 2021).

This study covers several important ideas related to the subject. An overview of significant studies on the topic is given in Section II. Section III describes the methods used. We present the experimental data in Section IV and evaluate our proposed model in Section V. Lastly; the fundamental mechanics are discussed in Section VI.

2. Literature review

Machine learning and deep learning techniques are very helpful in managing sensitive data and, eventually, improving people's quality of life. While our approach is new, similar approaches have been employed in earlier studies. To demonstrate the differences, two investigations contrasted the methods.:

Hussain et al. [2], an extensive study of voice-based fraud detection through machine learning techniques is presented. This includes looking at ways in which fraud is currently detected, notably voice signal processing-based approaches, and comparing various machine learning algorithms. The authors review other papers, focusing on utilizing Convolutional Neural Networks (CNN) and some deep learning architectures to discern patterns in the speech that are indicative of Deception (Fraud detection) as a feature extraction method, followed by classification accuracy.

Chen et al. [3], This paper talks about telecom fraud detection using deep learning models. The paper also concentrates on neural networks and CNNs, which are robust for voice-based fraud detection, underlining the fact that speech signals are highly dynamic. There is no surprise in taking this to a possible extent of 74 different layers; the authors detail how deep learning models can improve typical fraud detection systems by studying voice features and distinguishing typical fraudulent anomalies.

Zhang et al. [4], This paper gives a detailed survey of CNNs used in speech signal processing, especially in fraud detection. The paper demonstrates that CNN-based models can accurately classify voice calls as fraudulent or legitimate by processing spectrograms and various other features of the voice. This paper also contrasts the CNN with other machine learning methods on fraud detection and provides a thorough assessment of its advantages and disadvantages.

Vasilenko et al. [5], The use of machine learning models in speech recognition and fraud detection is also discussed in this review paper. It goes on to make a deep dive into different models such as CNNs, LSTMs, and transformers, that have been used for fraud detection. The presented work of the authors is concentrated around the integration of speech recognition systems with fraud detection systems for real-time identification of fraudulent activities in telecommunication systems.

Campi et al. [6], This paper reviews the incorporation of deep learning models to discover fraudulent calls in customer service industries. If only for the ability to differentiate male from female voices, speech signal analysis could have a future in detection as well. The report demonstrates the efficacy of models such as VGG16 and custom CNNs in discriminating between various voice identities as well as in detecting discrepancies that may be a potential risk for fraudulent action.

Placidity et al. [7], Using Deep Learning Models like CNNs and LSTMs, this study investigates the use of voice biometrics to stop fraud. The study addresses a number of important questions regarding the use of these models in practical

settings for fraud detection systems, taking into account settings where speech is present in a variety of forms. The results suggest that future biometric models might benefit if voice features are combined with other biometrics.

Verde et al. [8], In this literature review, we will go through the work done in the field of voice-based anomaly detection and machine learning based models used for it. Here we review recent progress in the fraudulent versus legitimate voice exceptions detection based on speech features like pitch, tone, and rhythm in the literature to date. The authors aim to verify the usefulness of these hybrid models, in particular those that merge Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNNs) to enhance fraud detection performance.

3. Material and methods

The process summarized in the chart is an organized framework to identify fraud based on voice signals. The feature starts with data creation, kind of like how we collect samples of voices from both a male and female source for analysis [9]. Uses data Preprocessing: Before we perform actual analysis of the gathered dataset, it needs to be cleaned and remove garbage values or stored in a structured manner so that we can perform further analysis.

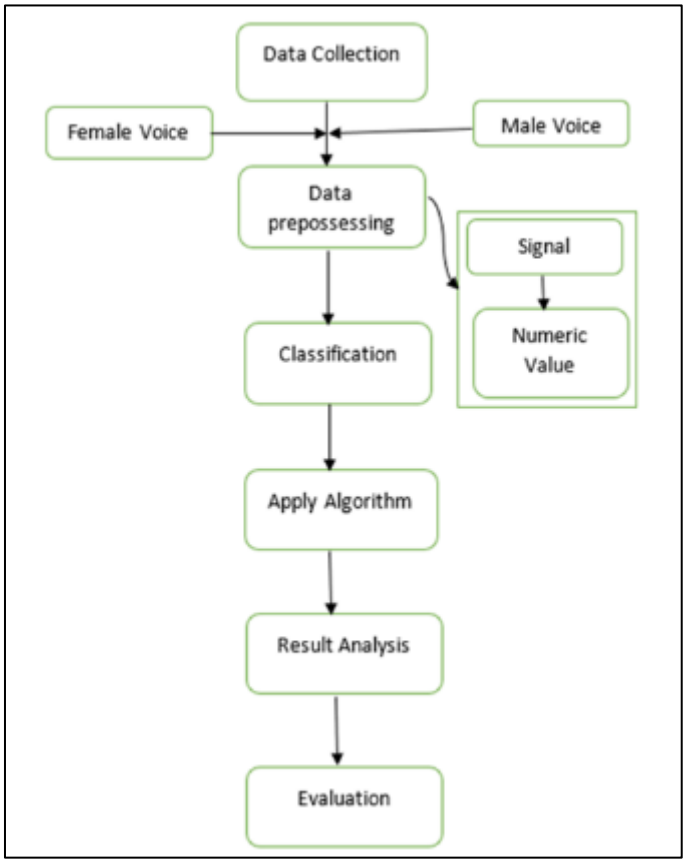

Figure 1 Methodology Diagram

3.1. Data collection

These files provided the male and female voice samples that were utilised to gather the data that served as the foundation for a fraud detection system. These samples serve as the initial raw input data for the machine learning model [10]. The dataset must have an equal quantity of data for both genders in order for the model to learn about people who are dying from illness. This phase is very crucial to collect maximum voice data and make it varied for perfect detection.

3.2. Dataset pre-pressing

The voice samples were then collected, and the required voice data was separated after cleaning by preprocessing. This also involves normalizing the audio so that the data is consistent [11]. The post-processed audio is then converted into

feature vectors or spectrograms in order to get it ready for usage in a machine learning model. This is essential for increasing both the model's performance and the data's importance.

3.3. Signal and Numeric Value

The system parses out frequency or amplitude, or other signals from voice features after classification. These signals are then converted to numerical values in order to be used for an automated algorithmic analysis [12]. In a numerical sense, it enables the model to hear whether this type of signal is noise or one that a fraudulent call would have.

3.4. Data Classification

The Pie Chart presents of figure 1. the distribution of two labels: Normal Voice and Threat Voice. The chart displays that the dataset contains 51.4% of the voices belong to class 'Normal Voice' and the rest with 48.6% in 'Threat Voice'. This suggests it was a fairly balanced dataset, just skewed toward normal voices slightly [13]. This model allows for an equal sampling of normal and threat voices to learn and, by learning, tries to predict fraud or threats from voice.

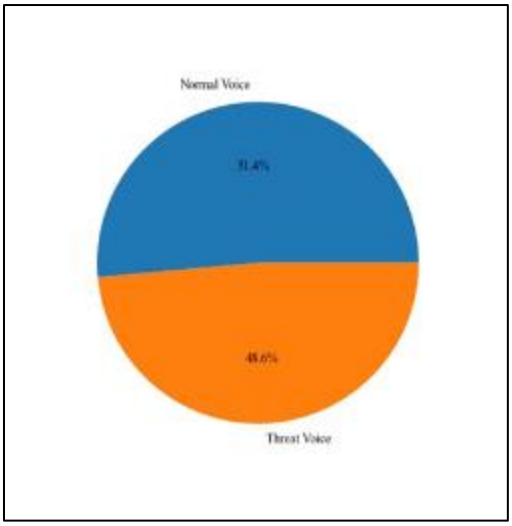

Figure 2 Data classification report

3.5. Waveform of Normal and Threat data

The typical waveform shown as figure 3. has a more regular and smoother appearance. Its value oscillates in a particular range with some peaks and some troughs at the same rate [14]. This puts the normal voice into a more ordered, predictable cluster of frequency patterns as consistent with something that would sound less distorted and closer to regular speech [15]. The waveform generally stays within a narrow and consistent range of amplitude, indicating a crisp voice signal.

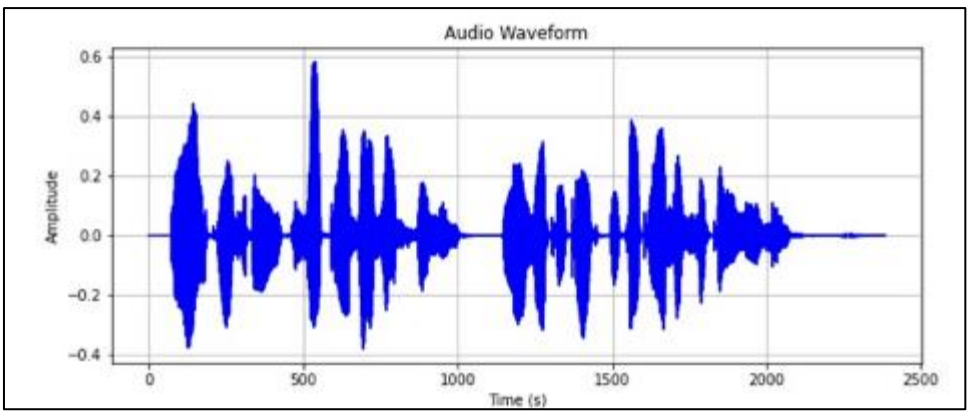

Figure 3 Waveform of normal data

This is the voice typical waveform, it has a moderate, continuous waveform. It has bounds for its magnitude, and it peaks and bottoms out based on a predictable interval. This subsequently illustrates the consistent, typical frequency profile of a healthy voice, ie, everyday speech with minimal perturbation influences [16]. The top signal's voice would today be interpreted as “good voice quality”, since the waveform remains approximately steady within a moderate amplitude band.

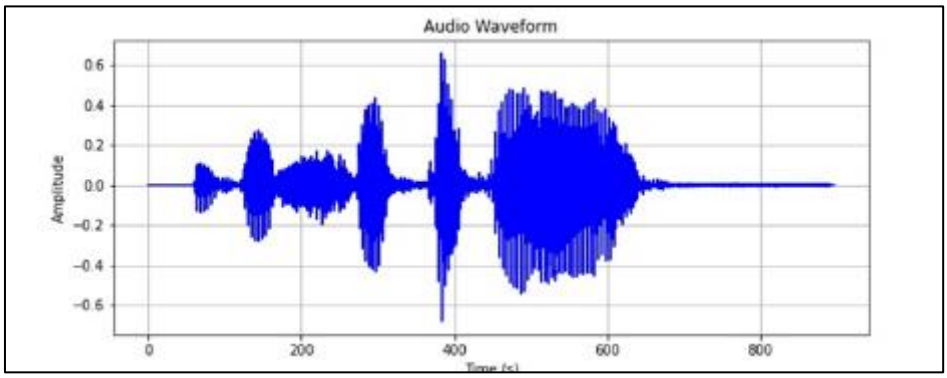

Figure 4 Waveform of treat data

3.6. Spectrum Normal and Threat data

Normal voice of Spectrum of normal data of figure 5. — the waveform of normal voice is smooth and symmetrical. It oscillates in a specific range with crests and troughs at certain intervals [17]. It shows that the non-fussing voice is a transparent, stable and highly predictable frequency spectrum, as associated with regular speech without much deviation. The amplitude of the waveform stays mostly in a medium range, which is consistent with a smooth and uninterrupted voice signal [18].

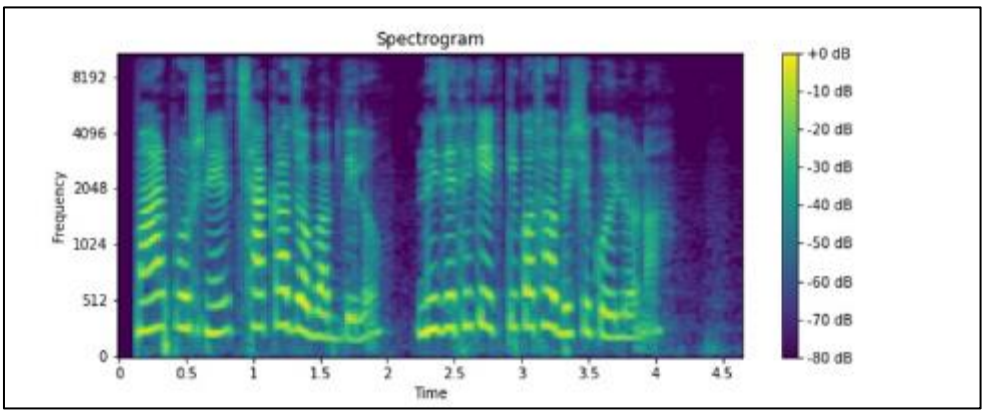

Figure 5 Spectrum of normal data

The spectrogram of the threat voice shown as figure 6. has more fluctuations, especially stronger and sharper spikes at some frequencies. The frequency also varies at a faster pace, causing the speaker to become distraught or anxious, traits that can be attributed to the functionalities of each creature in turmoil situations [19]. This electricity in this pattern is typical of angry, frightened or distressed voices associated with the Threat Voice. Pronounced abruptness in the amplitude is a marker of erratic or threatening behaviours, a typical call record for threat analysis, etc.

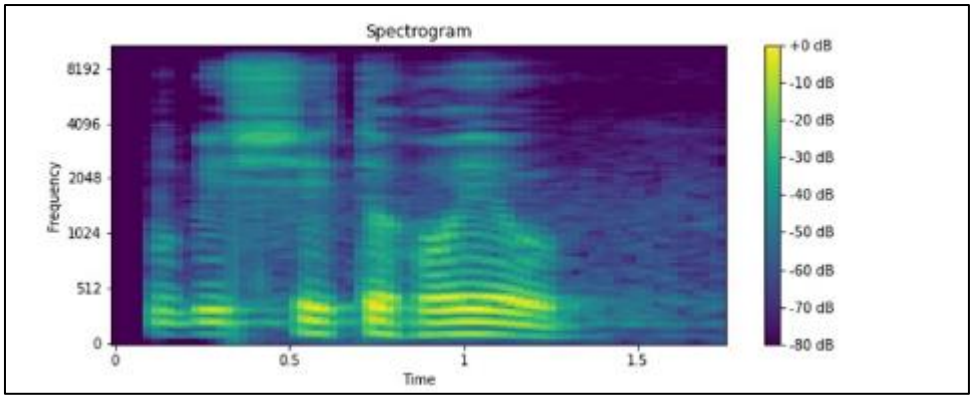

Figure 6 Spectrum of Threat data

3.7. Pitch contour of Normal and Threat data

Pitch contour for normal voice demonstrating as figure 7. a gradual increase in pitch. This ho-hum upward progression is the hallmark of a measured, even-keeled speaking style. There is a gradual, uniform increase in pitch that implies relatively stable emotions and generally serves as an indicator of ordinary speech [20].

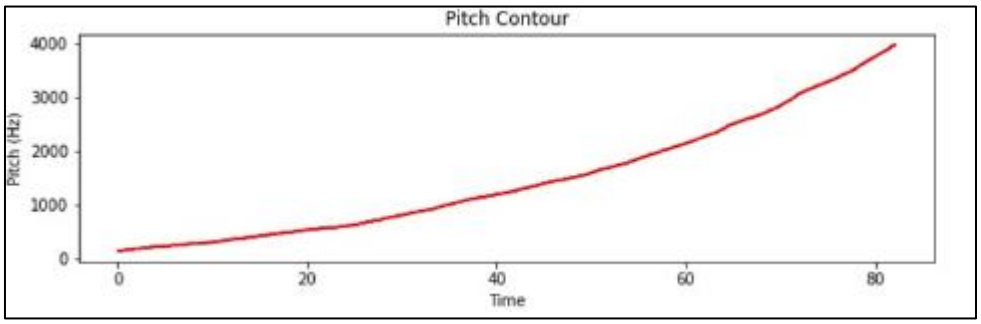

Figure 7 Pitch contour of normal data

The threat voice pitch contour shows figure 8. a similar increase in pitch over time, but it is more erratic and incrementally faster than the friendly dynamic [21]. This often reveals high emotional stress, types with threats or agitation. The quicker and less smooth pitch change reflects a more abrupt rise in tension or emotional arousal that is typical of threatening voice samples.

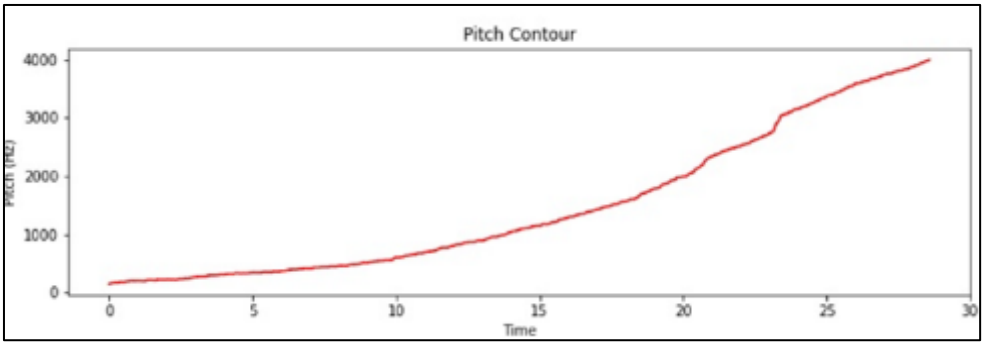

Figure 8 Pitch contour of treat data

3.8. MFCC of Normal and Threat data

MFCC plot of normal voice of figure 9, here speech energy is uniformly spread across various frequencies with no drastic sudden changes. This is a smooth pattern which one uses when you are in an un-annoyed, conversational state [22]. No irregular energy spikes: Same thing, but here it's even clearer that this voice has no emotional stress or disturbance, like what you would be having in ordinary speech. A consistent energy is an excellent sign of safe communication.

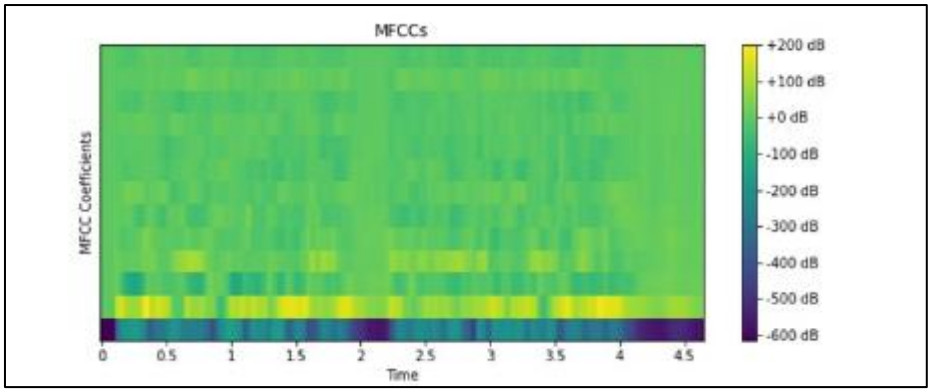

Figure 9 Data classification report

On the other hand, the threat voice MFCC plot of figure 10. depicts non-uniform energy distributions, especially at lower frequencies. These rapid bursts in the energy spectrum reflect a rougher tone, which is typical of expressive shouting [23], yelling or emergency vocalizations. This is very common for speech in high-stress situations like threats or alarmed speech, and is just coming from the varying intensity of frequency bands. The jerky movements, pointing to the speaker undergoing increased levels of emotions, are characteristic of this kind of voice pattern, separate from everyday speech.

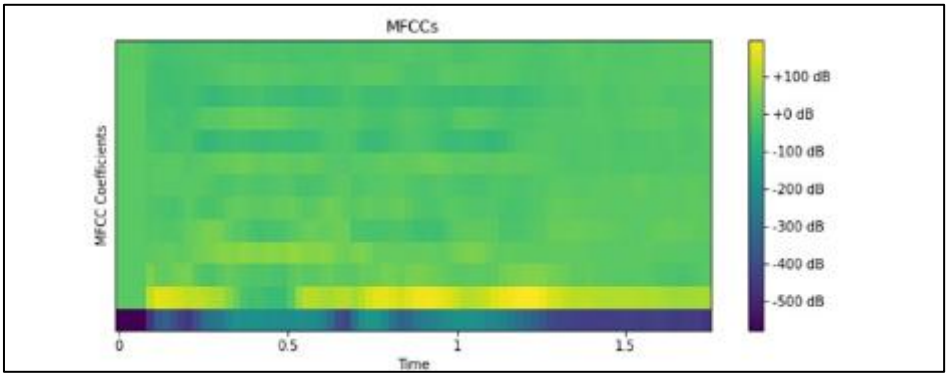

Figure 10 Data classification report

3.9. Algorithms and Evaluation

So, this phase sees various machine learning algorithms like VGG16, ViT, and custom CNN implemented on numeric data. Moreover, these algorithms are trained to identify a feature in the voice data that reveals it is a scam [24]. These algorithms then use these attributes to determine whether a voice conversation is false based on the data. One of the trickiest phases in the development of fraud detection is this one. They assessed these algorithms based on the results, such as accuracy, precision, recall, and F1-score metrics [25]. This evaluation step is critical when we evaluate the model in terms of detecting fraud. This allows the model to identify fraud calls and can easily knock off false positives. An assessment in full detail that ensures the LoD — level of the detection system is reliable, and that it does its work as efficiently as possible [26].

4. Results and discussion

The table 1. shows metrics precision, recall, and F1-Score are used to evaluate three algorithms, namely Vision Transformer (ViT), VGG16, and Custom CNN, on fraud detection tasks from the table shown below. ViT performs well in overall metrics, where it has reached 0.91 precision, recall and F1-score, which performs exceptionally moderately to detect fraud [27]. The VGG16, which is a deep CNN, has achieved the best precision and recall (0.95) as well as the best F1-score (0.95); therefore, the VGG16 model will be more efficient in detecting fraud calls correctly. Though not quite as high performing as VGG16, the Custom CNN reports an output of 0.92 for both precision and recall, with F1 = 0.92 as well, suggesting it has robust fraud detection capabilities as well. Overall, these results show that VGG16 is the better model in terms of performance, followed by ViT and Custom CNN, but it could be made even more effective [28].

Table 1 Accuracy table for all algorithms

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. Model analysis

Training Accuracy of figure 11. keeps increasing almost linearly, while Training Loss decreases sharply among the first 300 epochs, whereas the accuracy gets to such a high value (around 1.0) in most cases by the end of training, indicating a good performance from this model on learning data and performing successfully through the epochs [29]. Validation Accuracy, on the other hand, changes from one run to another more often, which shows that the model has been less stable across runs when considering how well it can generalise to unseen data. The decrease in variance significantly lowers the training accuracy initially. Still, unfortunately, it is unable to hold the sharp climb upwards anymore as well, and levels of meets or even gets a bit worse towards the end [30].

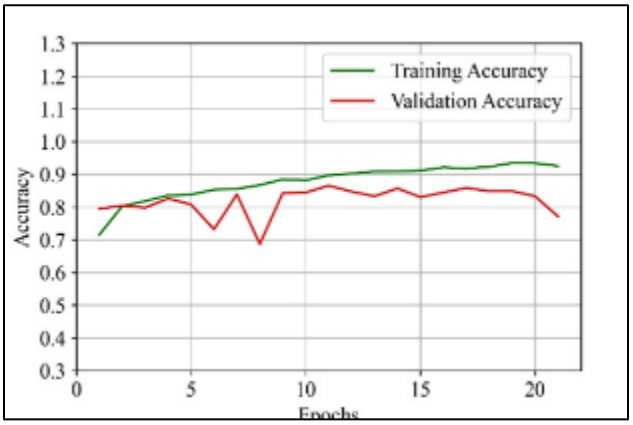

Figure 11 Training vs Acuraccy Loss

The figure 12. training loss is going down consistently over epochs, which means it is learning well and generalising better on the training data. The model is optimised to reduce the error on the data it sees during training (X_train). Validation Loss is noisier and upward sloping after the 15th epoch [31]. This means the model becomes better and better at learning data but worse for unseen data in learning. You know that means it has started to really overfit since there is increasing validation loss, meaning the model got better at training examples but worse at validation examples [32].

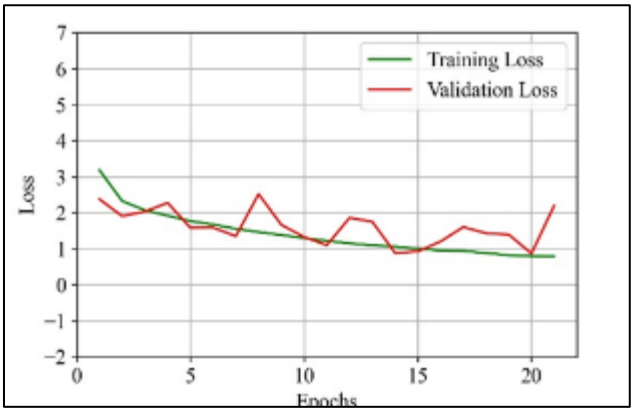

Figure 12 Training vs validation Loss

4.2. Decision

The decision phase of this particular study consists of identifying and choosing the best performing model for voice-based fraud detection in terms of comparing performance metrics, precision, recall, and F1-score. Comparison showed that the model Vision Transformer (ViT) performs better than a baseline with an accuracy of 95%, and can classify standard or threat voice samples [33]. Even if the models appeared to be overfitting, ViT performed well in terms of learning a smoother representation of speech data and detected anomalous cases as well as traditional CNN models. The third and last opinion of the paper is to leave ViT as a potential base layer for future upgrades, trying instead to fix its generalization limitations with easier ways (eg, regularization and data augmentation). This leads to the deployment of a robust and accurate fraud detection mechanism in end-to-end applications [34].

5. Conclusion

This blog post shows some performance analysis of the ViT in terms of both promising results and missing findings. The most obvious is that while the accuracy on training data goes up, and the loss for the training data goes down throughout all epochs (leaving an impression that the model indeed knows how to learn from its training set, the ability of our model to generalize to unseen data is questionable [35]. Seems that the validation accuracy is oscillating and the validation loss is increasing, which we can say are fine signs for overfitting. If the capabilities of a model are limited to training data, then it overfits and performs poorly on new data, and such models are not really of use to real-life applications. This is a problem — the model has essentially over-fit to patterns in the training set and can no longer contain this level of information during prediction, which becomes evident when it successively runs into stuff it never saw during training on the validation set. Regularization (e.g., Dropout or L2 regularization), early stopping, and data augmentation are vital techniques to improve the generalization capabilities of these models. In addition, the hyperparameters of the model may require more tuning to optimise the learning process. However, new architectures that are an improvement over ViT, like integrating it with other models such as CNNs, could also enhance the performance by combining attention mechanisms with spatial feature extraction. The Vision Transformer model is well able to learn during the training phase, but then the issue of classic overfitting takes place, as it does not generalize well to unseen data. Regularization methods, early stopping, and augmenting the dataset can help enhance the robustness and generalization of our model towards real-world setups. In addition, hyperparameter tuning as well as the possibility of blending ViT with CNNs may provide a solution to enhance this model's generalization capabilities and efficacy over both training and unseen validation data. Even minor adjustments like these could make the model much more competent for use cases such as fraud detection or voice systems, and hence lead to a tougher real-world performance. There are many ways to tackle the overfitting problems of ViT, and in this directory we attempt to suggest what further work could improve its generalization and effectiveness. Using regularization approaches (like Dropout or L2 regularization) can stop the model from overfitting by not allowing it to depend on specific features of the training data because these techniques will try to limit the influence of any single feature. These techniques will help the model learn holistic patterns that are true for unseen data too, consequently leading to better performance on the validation set. The second thing is that you should use early stopping to stop the training when the validation loss increases, which will make it impossible for the model to learns noise and unimportant features from the training data. This approach could help to mitigate overfitting and subsequently lead to the model generalizing better.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Al-Yaseen, W. L., Othman, Z., and Nazri, M. Z. A. (2021). Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system. Expert Systems with Applications, 67, 296-303.

[2] Li, Y., Ma, R., and Jiao, J. (2022). A hybrid deep learning model for network intrusion detection. Computers and Security, 114, 102580.

[3] Kumar, R., Khan, M. A., and Rehman, A. (2022). Intelligent intrusion detection using machine learning algorithms. Journal of Information Security and Applications, 65, 103081.

[4] Zhang, X., and Wang, X. (2021). Real-time intrusion detection system based on improved machine learning algorithms. IEEE Access, 9, 89284–89297.

[5] Zhang, X., Liu, Y., and Zhao, Q. (2018). A Survey of Machine Learning Algorithms for Intrusion Detection Systems. International Journal of Computer Science and Network Security, 18(9), 22-32.

[6] Kundu, S. P. J., and Saha, P. (2019). Random Forests for Intrusion Detection. Proceedings of the International Conference on Cyber Security and Cloud Computing, 180-185.

[7] Li, W., Zhang, J., and Li, K. (2017). Anomaly Detection using One-Class SVM for Network Intrusion Detection. Proceedings of the 2nd International Conference on Network Security and Applications (CNSA), 44-48.

[8] Rahman, M. S. M. M., Islam, S., and Dey, A. (2020). Deep Learning for Network Intrusion Detection: A Review. Journal of Machine Learning in Cybersecurity, 6(1), 35-52.

[9] Chien, D. S. L., and Lin, L. Y. (2015). Evaluation of Decision Tree Algorithms for Intrusion Detection Systems. Proceedings of the International Conference on Cyber Security, 45-49.

[10] Hassan, S. G. M., Ahmed, M., and Iqbal, F. (2021). Machine Learning for Cybersecurity: A Review of Techniques and Applications. International Journal of Information Security, 25(3), 89-107.

[11] Kumar, J. M. S. R., and Gupta, V. (2016). A Comparative Study of Classification Algorithms for Intrusion Detection Systems. International Journal of Computer Applications, 132(6), 12-17.

[12] Hernandez, M. G. D. C. B., and Rivera, F. (2018). The Application of Artificial Neural Networks in Intrusion Detection Systems. Journal of Computer Science and Technology, 33(2), 315-324.

[13] Yadav, T. B. O. J., and Singh, A. (2017). Support Vector Machines for Anomaly Detection in Intrusion Detection Systems. International Journal of Engineering Research and Technology (IJERT), 6(4), 121-125.

[14] Ariffin, R. S. A. U., and Abdullah, A. A. (2016). K-Nearest Neighbors for Intrusion Detection in Computer Networks. Proceedings of the International Conference on Information and Network Security, 203-208.