Detection of Speech Oriented Fraud Using Supervised Machine Learning Algorithms

Sridhar S K and Jayesh Ranjan
Department of Computer Science and Engineering, School of Engineering, Dayananda Sagar University, Bengaluru, India

Keywords: Artificial Intelligence (AI)-Generated Voice, Fraudulent, Artificial Neural Network, Convolutional Neural Network, Long Short-Term Memory, Cyber Security.

Abstract: Through the usage of deep learning methodologies, this paper focuses on the detection of AI-synthesized voices utilized for fraudulent activities. The methodology aims to apply an innovative approach based on deep neural networks to differentiate between the genuine and artificially generated voice, the project showcases the productiveness of this technique in unmasking AI generated voices. The outcomes of this project provide valuable insights to support voice authentication systems and boosting security protocols in the face of the rising popularity of AI-simulated speech technologies, making a significant advancement in fighting possible misuse of such technologies.

1 INTRODUCTION

The rise of artificial intelligence (AI) has progressed various industries, including the field of voice synthesis. AI- synthesized voices, created through deep learning algorithms and neural networks, have found widespread applications in voice assistants, entertainment, customer service, and even in audio deepfakes for deceptive purposes. However, the increase of AI-generated speech has raised concerns regarding the potential misuse of this technology for fraudulent activities such as voice impersonation, fraud, misinformation, and social engineering attacks. In response to these emerging challenges, this research project delves into the domain of voice authentication and security by exploring the application of deep learning techniques to detect and expose AI-synthesized voices. The primary objective of this study is to develop a novel methodology that leverages deep neural networks to differentiate between authentic human voices and AI-generated speech patterns with high accuracy and reliability. By focusing on the distinctive characteristics and nuances of AI-synthesized voices, this research aims to enhance the capabilities of voice verification systems and security protocols in identifying and mitigating the risks associated with synthetic speech. Through the implementation of cutting-edge deep learning models and data analysis techniques, the project seeks to shed light on the underlying patterns and features that distinguish AI- generated voices from genuine human speech. The findings and outcomes of this project are expected to contribute to the advancement of machine learning research, particularly in the realm of sequence data processing and predictive modelling. By elucidating the nuances of Artificial Neural Network (ANN) and Convolutional Neural Networks Long Short-Term Memory (CNN-LSTM) models and their performance disparities, this project seeks to enhance the understanding of deep learning methodologies and their applicability in real-world predictive tasks, paving the way for informed decision-making and model selection in future endeavours.

1.1 Scope

Voice scams involve the use of social engineering techniques to deceive individuals into performing actions that risk their security or revealing sensitive information. Common tactics include impersonating trusted entities, creating urgency, and exploiting emotional responses. One way of tackling this is through a system which can accurately differentiate between human and AI-generated voice which can help hinder any potential damages to security or revealing sensitive information. This project focuses on the use of deep learning models to accomplish the differentiation task.

S K, S. and Ranjan, J.
Detection of Speech Oriented Fraud Using Supervised Machine Learning Algorithms.
DOI: 10.5220/0013598300004664
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 2*, pages 609-615
ISBN: 978-989-758-763-4
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

1.2 Background

Traditional methods of scam detection often rely on manual review or rule-based systems, making them less adaptive to evolving scam tactics. In recent years, with the rapid advancements in artificial intelligence and machine learning technologies, deep learning models have emerged as powerful tools for predictive analytics and pattern recognition. The ANN and CNN-LSTM models stand out as popular architectures known for their effectiveness in handling complex data patterns and sequential information. The motivation behind this project stems from the growing interest in exploring the capabilities of different deep learning architectures for predictive tasks. While ANN models have been extensively used in various machine learning applications, the integration of CNNs and LSTMs in the form of CNN-LSTM models has shown promising results in processing sequential data, such as time series, audio signals, and natural language text.

1.3 Need

In real-world predictive tasks, paving the way for informed decision-making and model selection in future endeavours.

Understanding the strengths and weaknesses of ANN and CNN-LSTM models is crucial for selecting the most appropriate architecture for a given predictive task. ANN models excel at capturing complex relationships in structured data but may struggle with sequential information due to the lack of temporal context. On the other hand, CNN-LSTM models leverage the spatial hierarchies learned by CNNs and the long-term dependencies captured by LSTMs, making them well-suited for tasks where both spatial and temporal features are essential. By conducting a comparative analysis of ANN and CNN-LSTM models in this project, we aim to gain valuable insights into their performance characteristics, predictive accuracy, and computational efficiency. This comparative study will provide a deeper understanding of how these architectures process and learn from data, ultimately guiding us in selecting the most efficient and effective model for our specific predictive task. The findings from this project have the potential to advance the field of deep learning and predictive modelling by offering empirical evidence and practical guidance on choosing the optimal architecture for similar tasks in the future. Through this exploration, we aim to contribute to the ongoing research efforts aimed at enhancing the capabilities of deep learning models and their applications in diverse domains.

1.4 Technology

In real-world predictive tasks, paving the way for informed decision-making and model selection in future endeavours.

Technology Stack for the Project: The project, encompassing the implementation of Artificial Neural Network (ANN) and Convolutional Neural Network Long Short-Term Memory (CNN-LSTM) models for predictive tasks, was executed within the Anaconda environment, which provides a comprehensive platform for managing Python packages and environments. The following technologies and libraries were utilized for the successful development and analysis of the models:

- **Python Programming Language:** Python played a central role in coding the machine learning models, data manipulation, and analysis tasks due to its simplicity and extensive libraries for deep learning.
- **NumPy and Pandas:** NumPy and Pandas were utilized for efficient numerical computations, array operations, and data manipulation tasks, including data preprocessing and cleaning.
- **Matplotlib and Seaborn:** Matplotlib and Seaborn were employed for data visualization, enabling the creation of informative plots and graphs to analyse model performance and results effectively.
- **Scikit-Learn:** Scikit-Learn was used for model evaluation and metrics calculation, including accuracy scores, classification reports, and confusion matrices, providing valuable insights into the model performance.
- **Jupyter Notebooks** Jupyter Notebooks served as the interactive development environment for coding, experimenting with different models, and documenting the project workflow. The notebook format facilitated seam- less integration of code, visualizations, and explanatory text.
- **Seaborn and Matplotlib:** Seaborn and Matplotlib libraries were used for data visualization, aiding in the analysis and interpretation of model results through various plots and charts.
- **Anaconda Environment:** The project was executed within the Anaconda environment, which offers a comprehensive suite of tools for data science, machine learning, and deep learning tasks, streamlining package management and environment setup.
- **TensorFlow:** TensorFlow, an open-source deep learning library, was utilized for implementing and training the neural network models. Its flexibility and scalability made it suitable for

handling complex architectures like CNN-LSTM.

1.5 Design

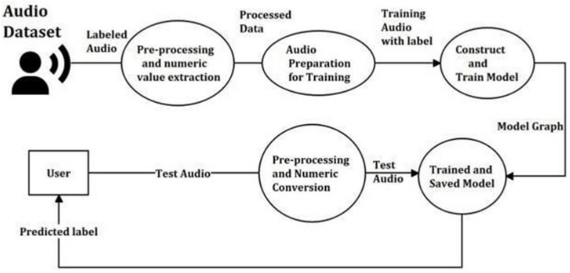

Audio Dataset
Labeled Audio
Pre-processing and numeric value extraction
Processed Data
Audio Preparation for Training
Training Audio with label
Construct and Train Model
Model Graph
User
Test Audio
Pre-processing and Numeric Conversion
Test Audio
Trained and Saved Model
Predicted label

Figure 1: Design Flow Diagram

Raw Audio: The process begins with raw audio data.

Data Analysis: The raw audio undergoes data analysis, resulting in a pre-processed image.

Audio to Numeric Conversion: The pre-processed image is converted into scaled numeric data.

Feature Extraction: Feature extraction extracts relevant features from the data, creating a feature vector.

Split Dataset: The dataset is divided into an 80% training set and a 20% testing set.

Model Training: Using Training set, Models are trained here.

Validate Model: The model's performance is validated using the test set features.

2 PROBLEM STATEMENT

In today's digital age, voice-based authentication and verification systems play a crucial role in various applications, from voice assistants to online banking and security access. However, the rise of voice impersonation and synthetic voices has made it imperative to develop robust voice authentication systems capable of distinguishing between real and fake voices. The goal of this machine learning project is to address the challenge of selecting the most suitable deep learning architecture, specifically comparing Artificial Neural Networks (ANN) and Convolutional Neural Network Long Short-Term Memory (CNN-LSTM) models, for predictive tasks. With the increasing complexity of data patterns and the need to process sequential information effectively, determining which architecture demonstrates superior predictive performance is crucial for optimizing model accuracy and efficiency. The primary problem revolves around identifying the strengths and limitations of ANN and CNN- LSTM models in handling predictive tasks, particularly in scenarios where spatial and temporal features play a significant role. By conducting a comparative analysis of these two architectures, the project seeks to provide empirical evidence and insights into their performance characteristics, enabling informed decision-making in selecting the most appropriate model for similar predictive tasks in various domains.

- **Caller Authentication:** Verifies the authenticity of the caller through voice biometrics and behavioural analysis.
- **Contextual Analysis:** Examines the context of the conversation, identifying inconsistencies or suspicious elements.
- **Real-time Monitoring:** Provides instantaneous analysis during the call to enable swift intervention if a potential scam is detected.

3 RELATED WORK

The paper “Open Challenges in Synthetic Speech Detection” provides the insights about the challenges faced in detecting synthetic speech and it emphasizes on the importance of improving detection methods to address real-world scenarios (Cuccovillo et al. 2022).

In the paper “Representation Selective Self-distillation and wav2vec 2.0 Feature Exploration for Spoof-aware Speaker Verification”, the major focus is on the countermeasures against synthetic voice attacks, which are becoming increasingly indistinguishable from genuine human speech due to advancements in text-to- speech and voice conversion technologies. The authors also explored, which feature space effectively represents synthetic artifacts using wav2vec 2.0 (Lee and Jin Woo 2022).

In the paper “ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild”, the central focus is on the spoofing and deepfake detection in speech. It aims to evaluate countermeasures against manipulated and fake speech data (X. Liu et al., 2021).

In the paper “One-Class Learning To- wards Synthetic Voice Spoofing Detection”, the authors propose an anti-spoofing system to detect unknown synthetic voice spoofing attacks i.e., text-to-speech or voice conversion) using one-class learning. The key idea is to compact the bona fide speech representation and inject an angular margin to separate the spoofing attacks in the embedding space.

Without resorting to any data augmentation methods, our proposed system achieves an equal error rate (EER) of 2.19% on the evaluation set of ASVspoof 2019 Challenge logical access scenario, outperforming all existing single systems (Y. Zhang, F. Jiang and Z. Duan, 2021).

4 METHODOLOGIES

In the context of this project, it is of utmost importance to transform the raw audio data into numerical samples. This pivotal step is accomplished through feature extraction, leveraging Python libraries that are specialized in handling audio data. Feature extraction is a critical process as it allows us to convert the audio's continuous waveform into a format that can be comprehended by machine learning algorithms. As we delve into the realm of data preprocessing, a series of essential techniques are meticulously applied to ensure the dataset's cleanliness and enhance its reliability. These techniques involve tasks such as removing noise, handling missing values, and normalizing the data. This diligent preparation phase sets the foundation for robust model training and evaluation. Following data preprocessing, the next step is data splitting. This step involves in dividing the dataset into two distinct subsets where the training set is utilized to train the deep learning models, allowing them to learn the underlying patterns in the data, while the validation set is used to assess the model's performance and make necessary adjustments. In order to gain a comprehensive understanding of the project's objectives and the effectiveness of the employed deep learning algorithms, we have devised a comparative analysis. This analysis involves the utilization of two deep learning algorithms, each with its own architecture and parameters. By conducting this comparative study, we aim to shed light on the strengths and weaknesses of these algorithms and determine which one is better suited for our specific task. This approach enhances the overall significance of the project by providing valuable insights into the performance of different machine learning techniques. Given that the project follows a classification-based approach, it is imperative to employ appropriate classification metrics during the evaluation phase. One such metric is the confusion matrix, which gives us the models predictions, which includes true positives, true negatives, false positives, and false negatives. By analysing the confusion matrix and other classification metrics, we can gain a deeper understanding of the model's performance, its ability to correctly classify instances, and its potential areas for improvement.

4.1 System architecture

The Figure 2 illustrates the process of training and validating a machine learning model using an audio dataset. Starting with raw audio data, the steps include data analysis, preprocessing, numeric conversion, and feature extraction.

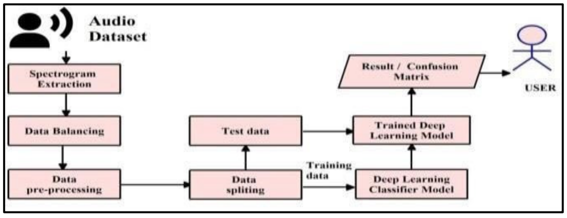


Figure 2: System Architecture

The dataset is split into training (80learning classifier model and the test data is then analysed using this model, resulting in a confusion matrix or other evaluation metrics. The final outcome is presented to the user, allowing them to query the model with new audio data for predictions.

4.2 CNN-LSTM

CLSTM is an innovative type of recurrent neural network (RNN) that merges the spatial processing capabilities of Convolutional Neural Networks (CNNs) with the temporal modelling abilities of Long Short-Term Memory (CLSTM) networks (Xingjian Shi et al.,2015). In a CNN- CLSTM network, the input data undergoes a sequence of Convolutional layers that extract spatial features from the input. These extracted feature maps are subsequently fed into CLSTM cells, which excel at capturing temporal dependencies within the data. Notably, the CLSTM cells are equipped with Convolutional connections, enabling them to selectively focus on specific regions of the input feature maps. The typical architecture of a CLSTM network involves multiple layers of Convolutional and CLSTM cells, with the output of each layer flowing into the subsequent layer. Ultimately, the final output is generated by a fully connected layer. CLSTM networks have demonstrated effectiveness across various applications, including video analysis, image captioning, and speech recognition.

4.3 ANN

The Artificial Neural Networks (ANNs) are computational models which consist of interconnected nodes or “neurons” organized in layers which structures the functioning of a human brain. The input layer receives the data, which is then pro- cessed through the hidden layers which results in the form of an output layer. Each connection between neurons has an associated weight, which adjusts during training to optimize predictions. ANNs excel at learning complex patterns in data, making them widely used in tasks like image recognition, natural language processing, and regression analysis. They possess the ability to generalize from training data to make predictions on unseen data. However, ANNs can be computationally intensive and require substantial data for effective training. Despite their complexity, they are a cornerstone of modern machine learning and have led to significant advancements in various fields.

4.4 CNN-LSTM vs. ANN

In the experimentation phase of the project, two deep learning architectures, Artificial Neural Networks (ANN) and a hybrid Convolutional Neural Network Long Short-Term Memory (CNN-LSTM) model, were evaluated for their effec- tiveness in classifying real and fake voices. The ANN model achieved an accuracy of 90.66hybrid model outperformed it significantly with an accuracy of 97.54Through rigorous test- ing and validation procedures, the accuracy of each model was assessed, with the CNN-LSTM hybrid model demonstrating superior performance in distinguishing between real and fake voices. The results signify the effectiveness of leveraging a hybrid deep learning architecture that can effectively capture both spatial and temporal characteristics present in voice data, leading to enhanced classification accuracy. Furthermore, the experimentation process involved fine-tuning the models, optimizing hyperparameters, and conducting cross-validation to ensure robustness and generalizability of the findings. The significant difference in accuracy between the ANN and CNN-LSTM hybrid models underscores the importance of selecting the appropriate architecture for voice classification tasks, especially in scenarios where discerning between authentic and synthetic voices is critical for security and authentication purposes. Overall, the experimentation results highlight the potential of employing advanced deep learning architectures like the CNN-LSTM hybrid model to enhance the accuracy and reliability of voice authentication systems, offering valuable insights for future research and development in the field of voice-based security technologies. CNN-LSTM (Convolutional Neural Network - Long Short-Term Memory) architectures are often preferred over simple Artificial Neural Networks (ANN) for voice classification tasks due to several reasons:

Spatial and Temporal Features Learning: CNNs are excellent at capturing spatial features from data, while LSTMs are proficient in capturing temporal dependencies. In voice classification tasks, the input data (such as spectrograms or MFCCs) often contain both spatial and temporal patterns. CNNs are effective in extracting spatial features from these representations, while LSTMs can effectively model temporal dependencies in the sequence of features.

Hierarchical Feature Learning: Lower layers learn simple features like edges and corners, while deeper layers learn more complex features. This hierarchical feature learning is advantageous for voice classification tasks where the input data may have complex structures.

Robustness to Variability: Voice data can vary significantly due to factors like accent, pronunciation, back- ground noise, etc. CNN-LSTM architectures are more robust to such variability as they can learn invariant representations from data by capturing both local patterns (via CNNs) and long-term dependencies (via LSTMs).

Reduced Overfitting: LSTMs, with their ability to re- member information over long sequences, can help pre- vent overfitting by capturing relevant temporal patterns while disregarding irrelevant noise. This is especially useful for voice classification tasks where the input data may have a high degree of variability.

Interpretability: CNNs and LSTMs have interpretable components. CNNs learn filters that represent specific patterns in the data, while LSTMs learn sequential patterns. This can be beneficial for understanding which features the model is using for classification, aiding in model interpretability and debugging.

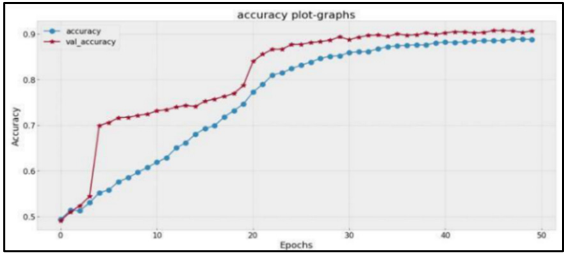


Figure 3: Accuracy plot-graph for ANN

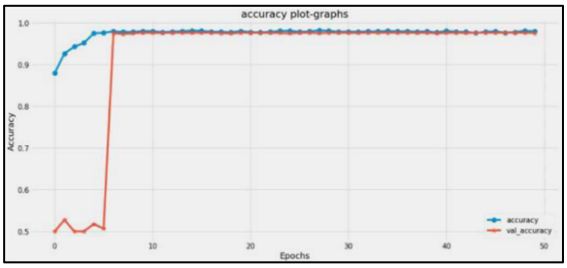


Figure 4: Accuracy plot-graph for CNN-LSTM

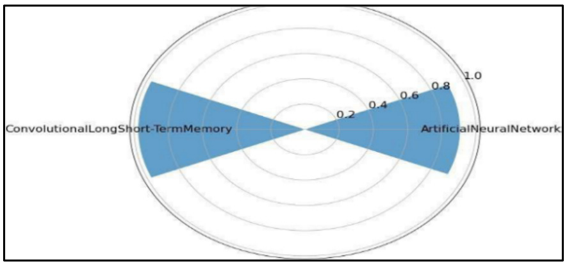
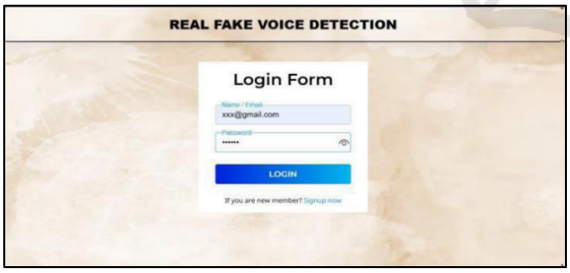


Figure 5: Comparison of model accuracy (Radar Chart)

5 RESULTS

5.1 Webpage

Login Page: This page is used for user authentication where the user enters valid mail and password. Additionally, we have provided the option to signup if the user already does not exist. The user credentials are stored in a excel sheet locally.


REAL FAKE VOICE DETECTION

Login Form

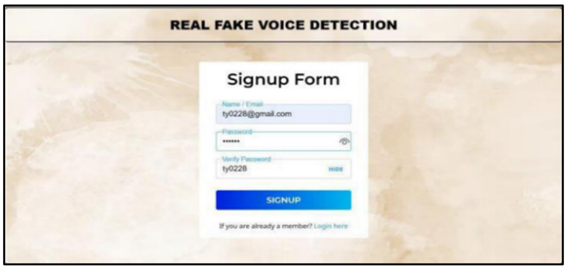
Name / Email: xxx@gmail.com

Password: ••••••

If you are new member? Signup now

Figure 6: Login page

5.1.1 Signup Page


REAL FAKE VOICE DETECTION

Signup Form

Name / Email: ty0228@gmail.com

Password: ••••••

Verify Password: ty0228

If you are already a member? Login here

Figure 7: Signup page

5.1.2 Prediction Page

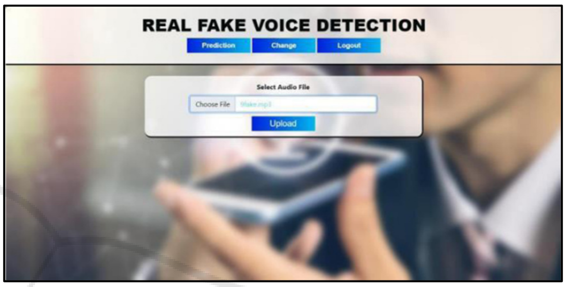
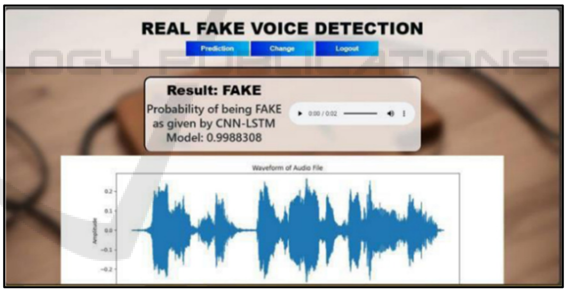

REAL FAKE VOICE DETECTION

Figure 8: Voice credibility prediction page

5.1.3 Result Page


REAL FAKE VOICE DETECTION

Result: FAKE

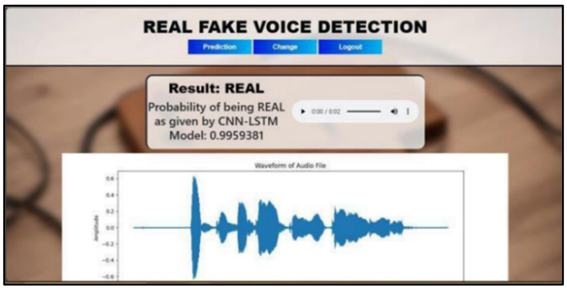
Probability of being FAKE as given by CNN-LSTM Model: 0.9988308

Figure 9: Fake voice detection

Fake voice: The image displays a computer interface for “REAL FAKE VOICE DETECTION” software. At the top of the interface, there are navigation options: “Prediction”, “Change” and “Logout”.

Below this, there is a result box with the label “Result: FAKE” indicating that the system has detected a fake voice. Accompanying text states: “Probability of being FAKE as given by CNNLSTM Model: 0.998808.”

An audio player interface is visible, featuring play and volume controls. However, the audio is not currently playing (indicated by “00:00 / 00:02”). Beneath the audio player, there is a visual representation of an audio file waveform in blue colour.



REAL FAKE VOICE DETECTION

Result: REAL

Probability of being REAL as given by CNN-LSTM Model: 0.9959381

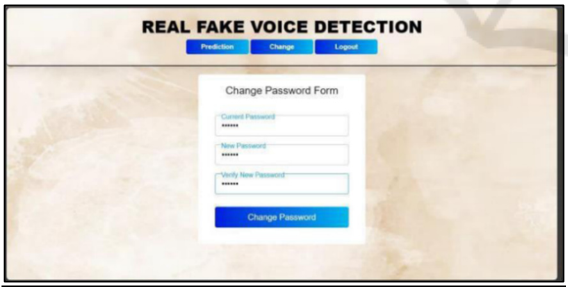
Figure 10: Real voice detection

<u>Real voice:</u> The image displays a computer interface for "REAL FAKE VOICE DETECTION" software. At the top of the interface, there are navigation options: "Prediction", "Change" and "Logout".

Below this, there is a result box with the label "Result: REAL" indicating that the system has detected a real voice. Accompanying text states: "Probability of being REAL as given by CNNLSTM Model: 0.9959381."

An audio player interface is visible, featuring play and volume controls. However, the audio is not currently playing (indicated by "00:00 / 00:02"). Beneath the audio player, there is a visual representation of an audio file waveform in blue colour.

<u>Change Password Page:</u>



REAL FAKE VOICE DETECTION

Change Password Form

Current Password: ******

New Password: ******

Verify New Password: ******

Figure 11: Change password page

6 CONCLUSIONS

Our study implies that CNN-LSTM networks is powerful for sequence prediction tasks where CNN's extract spatial features from input data while LSTM's capture temporal dependencies in sequences. In our project, AI-generated voice fraud detection, CNN can analyse audio spectrograms or other representations. LSTM processes sequential audio data to identify patterns indicative of fraud.

The hybrid architecture benefits from both spatial and temporal modelling, improving accuracy. The choice of the model is based on the need of the user, which is in-turn, based on speed and accuracy. The future work could be improvising speed while keeping accuracy as high as possible and building complete functioning software for the cyber security companies.

REFERENCES

L. Cuccovillo et al., 2022, "Open Challenges in Synthetic Speech Detection", IEEE International Workshop on Information Forensics and Security (WIFS), Shanghai, China, 2022, pp. 1-6, doi: 10.1109/WIFS55849.2022.9975433.

Lee, Jin Woo, et al., 2022, "Representation Selective Self-distillation and wav2vec 2.0 Feature Exploration for Spoof-aware Speaker Verification." arXiv preprint arXiv:2204.02639.

X. Liu et al., 2023, "ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2507-2522, doi: 10.1109/TASLP.2023.3285283.

Y. Zhang, F. Jiang and Z. Duan, 2021, "One-Class Learning Towards Synthetic Voice Spoofing Detection," in IEEE Signal Processing Letters, vol. 28, pp. 937-941, doi: 10.1109/LSP.2021.3076358.

Yi, Jiangyan, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao, 2023, "Audio deepfake detection: A survey." arXiv preprint arXiv:2308.14970.

Lv, Zhiqiang, et al., 2022, "Fake audio detection based on unsupervised pretraining models", ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.