IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Classification Of AI Generated Speech For Identifying Deepfake Voice Conversions

Mohan Krishna Kotha[1], Sai Anjan Tati[2], Snehitha Pellimari[3], Vagolu Rani[4], Akash Kumar Raju Thangella[5]

[1]Associate Professor, Department of CSE, Vasireddy Venkatadri Institute of Technology, Guntur, AP

[2,3,4,5]UG Students, Department of CSE, Vasireddy Venkatadri Institute of Technology, Guntur, AP

Abstract:

In the world with advanced technology, where few confidential tasks are done through the medium of voice there is need for Voice Recognition. There are many chances that may outbreak with improper recognition of one's voice such as security, authentication, accessibility, convenience and communication. Deep Learning Technique is one of the efficient techniques that will make use of artificial neural networks using Machine Learning Algorithms to detect tasks with provided training data. Here we use Deep Learning Technique to recognize the difference between True and Fake voice considering the 50 neurons that are associated in the hidden layer architecture for the purpose of pattern recognition with 80 is to 20 of percent training and testing data. The extracted features also include the MFCC (Mel-Frequency Cepstral Coefficients) with thirteen coefficients for each frame. Further the performance is evaluated based on different parameters such as accuracy, precision, recall, F1-score, and confusion matrix analysis, approaching to be an effective model to distinguish the True and Fake voice signals in a detailed manner.

Keywords: Real and fake voice classification, Deep learning, Pattern recognition network, Mel-Frequency Cepstral Coefficients (MFCC), Accuracy, Precision, Recall, F1-score, Confusion matrix.

1.Introduction:

Voice recognition serves a range of purposes, enhancing security, accessibility, convenience, and communication. It promotes broader access to digital technology and the internet, proving especially advantageous for individuals with disabilities, including those with visual impairments or motor challenges. Regarding convenience, voice recognition streamlines interactions with devices, catering to individuals facing mobility issues or those who prefer not to type. This technology extends its usefulness to tasks such as controlling smart homes, managing smart speakers, operating phones, and tablets, setting reminders, and interacting with personal technologies hands-free. Additionally, given the widespread use of voice in everyday communication, the utilization of voice-based biometric authentication offers a simple means for users to verify their identities, seamlessly integrating security into their regular interactions. With the extent advantages of Voice, it is more important for an individual to differentiate the difference between True and Fake voice for advantageous advantage.

Deep learning, a facet of AI and machine learning, mimics human learning to handle diverse data like photos, text, and audio, performing tasks such as classification and pattern recognition honored to automate traditionally human-intensive tasks like image description and audio transcription. This advent technology is vital in data science, speeding up and streamlining the collection, analysis, and interpretation of large datasets. Unlike the human brain's interconnected neurons, deep learning utilizes neural networks with layers of software nodes, trained on labeled data and diverse architectures. Recent advancements in generative Artificial Intelligence (AI) highlight its growing importance for which an example of Cutting-edge systems can now instantly transform a speaker's voice using advanced deep learning models and a microphone.

This once-futuristic technology, now accessible through consumer-level computing, raises security concerns despite its entertainment value which even include the Voice, crucial for social recognition and biometric authentication, can be unethically manipulated for privacy breaches and identity theft. To tackle such adverse situations, the article contributes significantly in three areas that include original audio classification dataset.

This research aims to build a deep learning classification model distinguishing between genuine and fraudulent voice signals. It involves evaluating deep neural networks, exploring feature extraction methods like MFCCs, creating a tailored pattern recognition network, and assessing the model's performance metrics. The study also tests the model's resilience against spoofing attacks and real-world scenarios. Additionally, it explores deep learning-based countermeasures for synthetic speech impersonation.

We also include DEEP-VOICE1 dataset, promoting interdisciplinary research. AI-generated speech patterns, the analysis and identification process is crucial. We also consider few real-time models including implementation of warning systems in phone or conference calls where synthetic voices might be used with malicious intent.

2. Literature Survey:

Juefei-Xu et al. held a thorough survey on combating malicious deepfakes, encompassing detection methods, mitigation strategies that has a drawback of implementation details and real-world performance evaluations [1].

Truby and Brown discuss the concept of human digital thought clones and their implications for artificial intelligence and big data. While the paper presents intriguing theoretical concepts, it may lack empirical evidence or practical applications [2].

Beard examines the protection of digital personas and associated legal rights. While the paper emphasizes the importance of safeguarding digital identities, it may not offer specific solutions for addressing deepfake-related threats [3].

Wells-Edwards thoroughly examines the legal consequences of voice cloning technology, focusing on the realms of privacy and intellectual property rights. While the paper identifies potential legal challenges, while lacking comprehensive solutions for addressing issues related to voice cloning [4].

Dale explores the dynamics of voice synthesis in 2022, providing insights into industry trends, challenges, and advancements. However, the paper may lack specific technical details regarding the underlying algorithms and methodologies employed in voice synthesis. [5]

Wang et al. introduce Tacotron, an end-to-end speech synthesis model that converts text input into speech signals. While Tacotron signifies a noteworthy advancement in speech synthesis, it may encounter limitations in generating natural-sounding prosody and handling complex linguistic structures [6].

3. Methodology:

Here we make use of Real Human Recordings along with Fake-Generated Recordings from which the audio feature captured is extracted, later it is given to the Training process where the Deepfake Audio is detected and further classified either as Real or Fake.

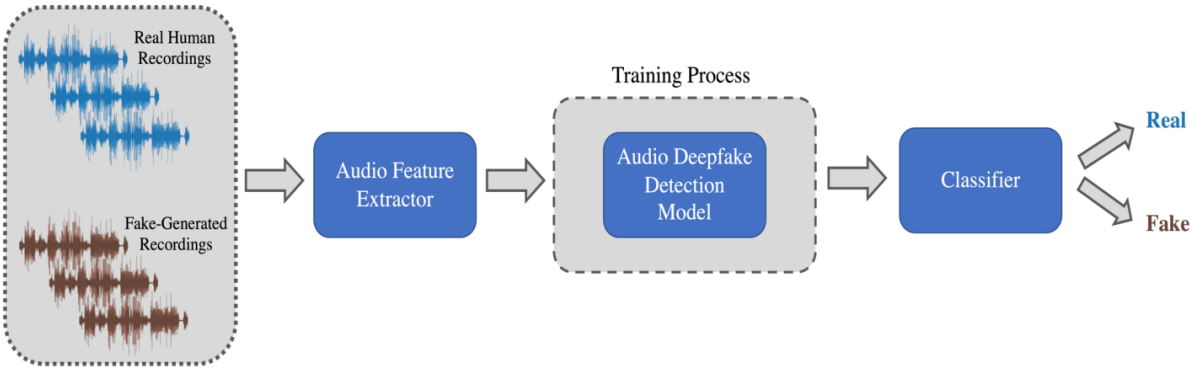


Fig1: Basic block diagram of proposed method

Algorithm:

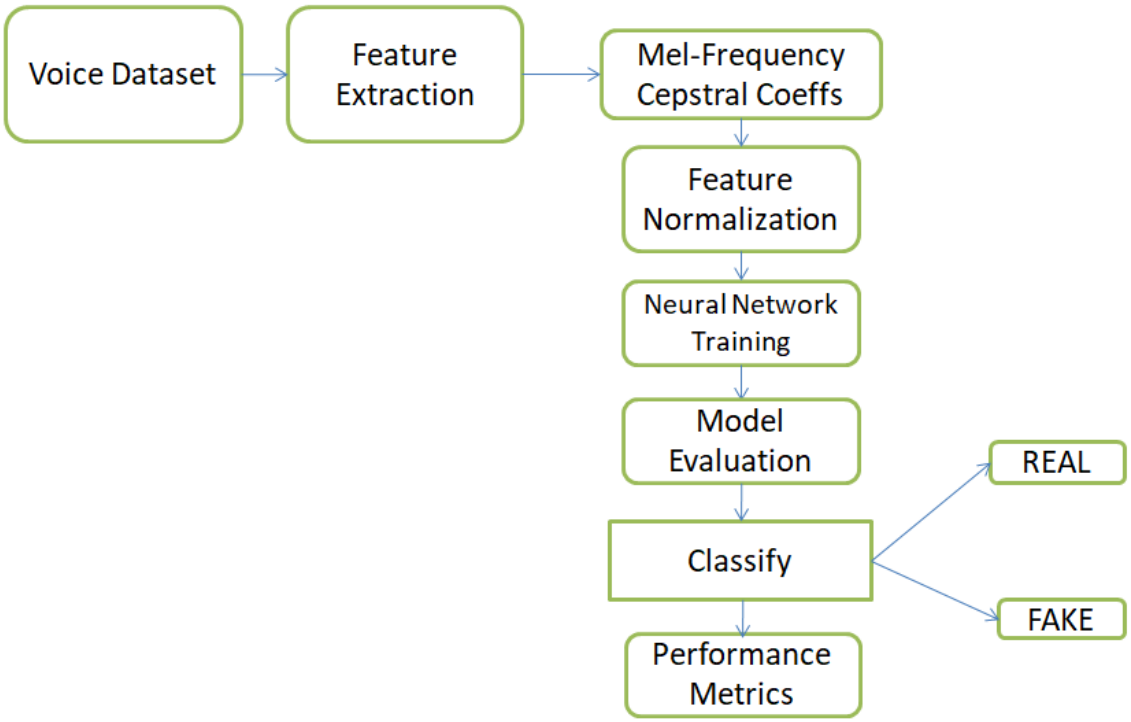


Fig2: Algorithm for Voice Classification

Step1: voice dataset

The initial stage of the algorithm includes the collection of Voice Dataset where the dataset containing the voice recordings balanced between the real and fake categories is used as an input with target labels either indicating True or false.

Step2:Feature Extraction:

In the Feature extraction stage, voice features are extracted based on the Mel-Frequency Cepstral Coefficients parameters, which is one the pre-dominant technique for estimating feature which undergoes calculation under six stages where it starts with Pre-Emphasis stage to amplify the higher frequencies, Framing Stage where the audio signal or the audio waves is divided into short overlapping frames, Windowing stage to reduce the spectral leakage, FFT stage to convert into frequency domain, Mel-Filtering Stage to apply filter bank on Mel-scale frequency representation, DCT stage to compute the Cosine transform of long-filter bank energies to retrieve the MFCCs.

Step3:Feature Normalization:

The Feature normalization technique of extracted features aims to improve consistency and enhance model performance and to avoid features with bigger magnitudes having a disproportionate impact on the model training process, retrieved MFCC features are often limited to a uniform range, usually between 0 and 1. Widely used techniques for normalization encompass min-max scaling or z-score normalization.

Step4:Neural Network

The Neural Network Training is adapted with feedforward neural network (FNN) with 50 neurons that are associated in the hidden layer is trained by using patternnet tool with normalized features and labels, where the architecture includes input and output layers, with activation functions like ReLU. The training process involves forward propagation, backpropagation for parameter updates, and optimization using categorical cross-entropy and algorithms like Adam or SGD.

Step5:Model Evaluation:

The Model Evaluation of the trained model is evaluated using a distinct dataset that was not used during the training phase is used to evaluate the trained model using the trained neural network.

Step6:Performance Metrics:

The quality of training data and detection of Voice is calculated using performance metric parameters such as accuracy, precision, recall, and F1-score to assess the effectiveness.

Step7:Result

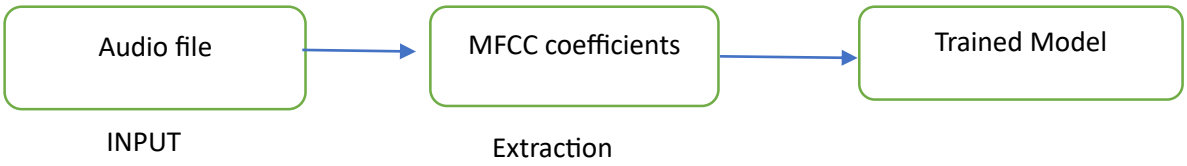
The Output Analysis include the presentation of results, featuring the confusion matrix, classification of audio file and performance metrics for interpretation.

This block diagram provides a high-level illustration of the classification process, delineating the data flow and operations utilized in distinguishing real and fake voice signals through deep learning techniques. The results, inclusive of the confusion matrix and performance metrics, are scrutinized and interpreted to evaluate the model's efficacy and identify possible improvements. Each step in this systematic process is crucial for the accurate classification of real and fake voice signals using deep learning techniques. Adhering to such an approach facilitates the creation of robust model's adept at precisely discerning between authentic and synthetic voice recordings.

3.1 Dataset:

For this study, we obtained the audio files from the Kaggle dataset for voice classification, which consists of two main folders: classified audio files and extracted coefficients. This dataset was obtained from a publicly available audio file on Kaggle having Fake Or Real audio files of 4GB [https://www.kaggle.com/datasets/birdy654/deep-voice-deepfake-voice-recognition]. It consists of fake audios[5890] and real audios[5889].It was easily labelled and prepared so that it could be used right away throughout the training stage. The dataset included utterances in.wav file format that were classified as authentic(real) and phony (fake).

From the extracted mfcc coefficients we train the model. After the model has been trained, it asks for the input of the audio files from the downloaded dataset.



3.2 FNN:

Feedforward Neural Networks (FNNs) is an Artificial neural network that transfer data in a single direction, from input to output, without the need for feedback loops. FNNs function as a network of linked brain cells that collaborate to solve an issue. Information from the outside environment is received by each cell, or neuron, which then processes it before sending it on to the following neuron.

According to science, FNNs are a class of artificial intelligence model modelled after the neural networks found in the human brain. They are made up of three layers: output, concealed, and input. Each linked neuron in these layers has a unique collection of weights and biases that affect how it reacts to incoming data. The input data is processed by the network using a process known as forward propagation, where each neuron applies a mathematical operation to the information it obtains. This makes FNNs effective tools for tasks like pattern recognition of the model, regression, and classification since they enable them to discover intricate patterns and correlations in the data.

3.3 PatternNet :

PatternNet is a popular neural network architecture in deep learning for pattern recognition applications. It is a member of the feedforward neural network family, which is characterised by unidirectional information flow from input layers to output layers. The capacity of PatternNet to recognize and understand intricate patterns in data sets it apart. It is made up of several linked layers of neurons, each of which produces output for the layer below it after receiving input from the layer above. During training, PatternNet modifies the weights of connections between neurons using a procedure known as backpropagation in order to reduce the discrepancy between the expected and actual outputs. Because of its ability to generalise effectively to new and unseen data, PatternNet is used for a wide range of tasks, including voice recognition, picture classification, and medical diagnosis.

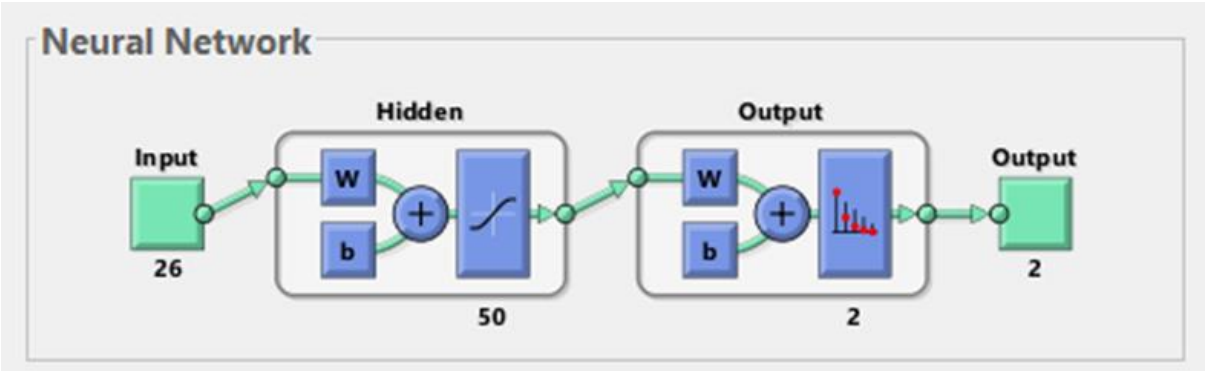


Fig 3 Neural Network.

Here we consider a Neural Network Architecture with input weight of 26 with hidden and output layer

3.4 Calculation Metrics :

Precision: A classification model's positive predictions are evaluated using the precision metric. It evaluates the extent of accurately anticipated positive occurrences (true positives) among all examples anticipated as certain (true positives + false positives). In simpler terms, precision estimates the model's capacity to try not to label negative instances as positive. A high accuracy demonstrates a low pace of misleading false positives, implying that the model is usually correct when it makes a given prediction.

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}.$$

Recall: Recall is a measure to a model's accuracy at locating relevant instances within a dataset. Out of all real relevant cases, it counts the percentage of true positives, or accurately detected relevant situations. Recall quantifies how comprehensive the model's predictions are from a scientific standpoint. Recall, put more simply, provides an answer to the following query: "Of all the relevant items, how many were correctly identified by the model?" Better performance in collecting all relevant events is indicated by higher recall values.

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}.$$

F1 Score: The F1-score is used to calculate how well the proposed model is going to classify the binary classification. A single metric that combines recall and accuracy gives a model a more objective evaluation performance. Recall and accuracy are allowed equal weight for calculating the F1-score, which is the harmonic mean of both values. As a result, it could be useful in scenarios where there is an imbalance in the distribution of false positives and false negatives or among the classes. A high F1-score indicates both high accuracy and good recall in a well-performing model.

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Accuracy: The amount of certainty to which a model's predictions are accurate is measured. It expresses as a ratio the percentage of correctly identified cases among all analyzed cases. When it comes to math, accuracy is determined by dividing the total number of predictions that are true to the model makes to the number of precise estimates. Especially in binary classification problems, it is a straightforward and widely used indicator for assessing a model's overall performance.

$$\text{Accuracy} = (TP + TN) \ / \ (TP + TN + FP + FN)$$

4. Results:

The presented results are based on the categorization of real and fake voices using a DL (deep learning) model, likely a neural network, applied to a dataset. Here's an explanation of each metric:

Confusion Matrix:
- The confusion matrix shows the number of (TP-1177), (FP-19), (FN-13), and (TN-1146).

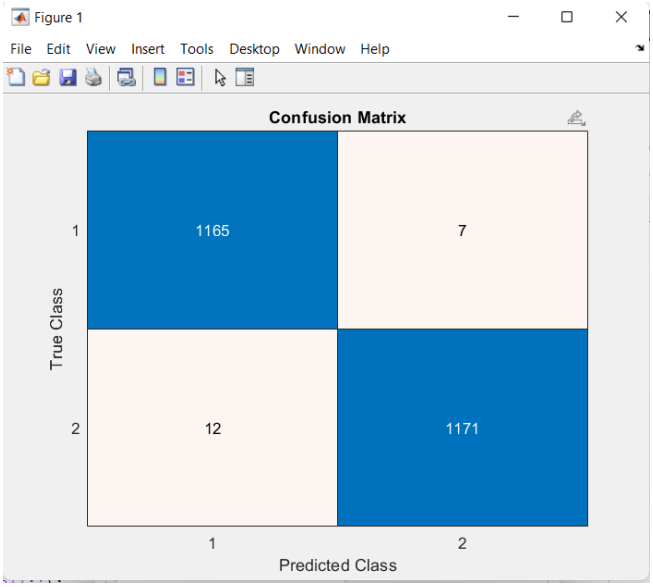


Fig 4 Confusion Matrix.

Confusion Matrix:
1177 19
13 1146

Precision:
With a precision of 0.98908 for the "FAKE" class, 98.91% of the voices labelled as false are, in fact, fake. With a precision of 0.98369 for the "REAL" class, 98.37% of the voices that are categorised as real are, in fact, real.

Precision:
FAKE REAL

0.98908 0.98369

Recall:
The recall for the "FAKE" class is 0.98411, which indicates that 98.41% of the fake voices were identified as fake.
with a recall of 0.98878 for the "REAL" class, 98.87% of the fake voices are identified as real Recall

0.98411
0.98878

F1-score:
For the "FAKE" class, the F1-score is 0.98659.
For the "REAL" class, the F1-score is 0.98623.
F1-score:
FAKE REAL

0.98659 0.98623

Overall Accuracy:
The system's overall accuracy is 0.98641 which implies that roughly 98.64% of the model's predictions in both classes are accurate.
Overall Accuracy: 0.98641

With this high accuracy and reliability, the model takes the input audio file and classifies it as real or fake and displays the dialogue box based upon the predicted accuracy range.

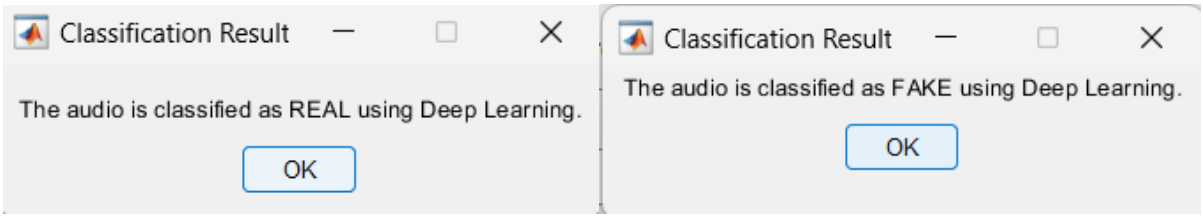


Fig 5 Classification result.

All of these outcomes illustrate how well the classification model works to distinguish between actual and false sounds, attaining high recall, F1-score, precision, and overall accuracy. Interpreting these findings in light of the particular dataset, model design, and assessment criteria employed is crucial, though. To fully comprehend any potential biases or limitations of the model, additional investigation could be required.

5. Result Analysis:

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	0.9352	0.935	0.931	0.933
PatternNet	0.9846	0.98	0.988	0.992

6. Conclusion and Future Scope:

The recognition of Voice signals from a particular individual have become a tedious and important task to ensure the security such that it does not fall in the prey of intruders. Here we have successfully built an algorithm based on Deep Neural Networks, a branch of Machine Learning to detect the voice whether it is true or fake voice. To add an additional security, feature the Deep Neural Network has also been utilized utilizing Mel-Frequency Cepstral Coefficients (MFCCs) has a security feature by which the system could achieve higher accuracy, precision, recall and F1 score during the classification task. Further, the efficiency of the system could be further improved by expanding the dataset that involves the larger range of voices and languages generalizing the capabilities of the system.

References

[1] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, “Countering malicious deepfakes: Survey, battleground, and horizon,” International journal of computer vision, vol. 130, no. 7, pp. 1678–1734, 2022.
[2] B. Borel, “Clicks, lies and videotape,” Scientific American, vol. 319, no. 4, pp. 38–43, 2018.
[3] J. Truby and R. Brown, “Human digital thought clones: the holy grail of artificial intelligence for big data,” Information & Communications Technology Law, vol. 30, no. 2, pp. 140–168, 2021

[4] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, “Protecting world leaders against deep fakes.,” in CVPR workshops, vol. 1, p. 38, 2019.
[5] R. Dale, “The voice synthesis business: 2022 update,” Natural language engineering, vol. 28, no. 3, pp. 401–408, 2022.
[6] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al., “Tacotron: Towards end-to-end speech synthesis,” arXiv preprint arXiv:1703.10135, 2017.
[7] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” in international conference on machine learning, pp. 4693–4702, PMLR, 2018.