Talluri Tarun Kumar
x23231262
MSc Cyber Security

A security application for social media and web platforms to identify the sound based deepfakes using signal processing infused deep learning framework

Abstract

Deepfake audio is a severe danger to digital communication since the AI used to create fakes can imitate voices and intonations at a high level of accuracy. This research fits the current gaps in the detection framework focusing on signal processing and deep learning as a preferred, more efficient methodology. Moreover, Mel-Frequency Cepstral Coefficients (MFCCs) and performance metrics (jitter and shimmer) having been incorporated into convolutional (CNN) and recurrent (LSTM) neural network would help the framework detect both the temporal and spectral changes in power audio. The study compares multiple models across various datasets: XGBoost outperforms others with clean audio database of 'In-the-Wild', 99.20% accuracy; CNN+LSTM on noisy audio database: Fake-or-Real, 69% precision. In improving the detection performance, preprocessing and integration of different modalities contribute well toward scalability and generalization on other novel deepfake generation algorithms. Realization in near real-time through an API recognizes the practical relevance of the application of the framework. This work lays the groundwork for adaptive, scalable audio deepfake detection systems that are necessary in emerging trust and secure societies.

Research Questions

RQ1: How can signal processing techniques be optimized to extract features that effectively distinguish genuine audio from deepfake audio?

RQ2: What deep learning architectures are most effective for capturing temporal and spectral characteristics of manipulated audio signals?

RQ3: How can the proposed framework ensure generalization across different deepfake generation techniques and unseen datasets?

Objective

The objective of this research is thus to design a strong and adaptive framework for detecting audio-based deepfakes using signal processing integrated deep learning framework. The main goals are to achieve the best feature extraction with MFCCs, jitter, shimmer to get close acoustic differences, to develop better deep learning architectures like CNNs, LSTMs, and combined models to increase detection rates and to produce a generalized solution when tested on different datasets and the different approaches to creating deepfakes; limiting the construction of a real-time API that will protect online communication services.

Table of Contents

Chapter 1: Introduction

Due to the progressive development of artificial intelligence there are both colorful innovations and, unfortunately, great threats, for example deepfakes. Being one of the most advanced forms of application AI deep fake can manipulate audio, video, and image to counterfeit personalities. The threat is particularly most apparent with audio-based deepfakes whose sound capabilities are precise that distinguishing them from original content is difficult especially for casual users.

Figure 1: An example of the deepfake. This can be found with the audio and the video (**Source: NewScientist**)

Link to a real time scam video: Mukesh Ambani deep fake video with voice over

Considering this situation, the identification of such falsifications has emerged as one of the increasing needs in the modern informational environment. Here's an overview of this study's approach, emphasizing the current challenges and solutions:

- **Exponential Growth in AI Manipulation**: Since basic tools of AI are available, deepfakes are now easier to produce – a problem for security on all forms of media (**Li et al., 2023; Kumar & Kundu, 2024**).

- **Unique Risks of Audio Deepfakes**: While compared to image deepfakes, speech fakes are much more dangerous due to their ability to copy not only voices, but also tones, intonations, per Centowski, and many other things, which are more difficult to notice (**Raza et al., 2023**).
- **Real-World Consequences**: From fake ID usage for phone con jobs to social engineering and ID theft the effects cover all the way from monetary scams to the undermining of credibility in virtual interactions.

With respect to metrics derived from time intervals between signal samples like jitter and shimmer which are characteristic of instability in synthetic audio, one is able to make very fine discriminations between vocal artifacts. Even these flowless variations of the pitch and tone of any material enrich the expressions of their forgery (**Almestekawy et al., 2024; Li et al., 2023**).

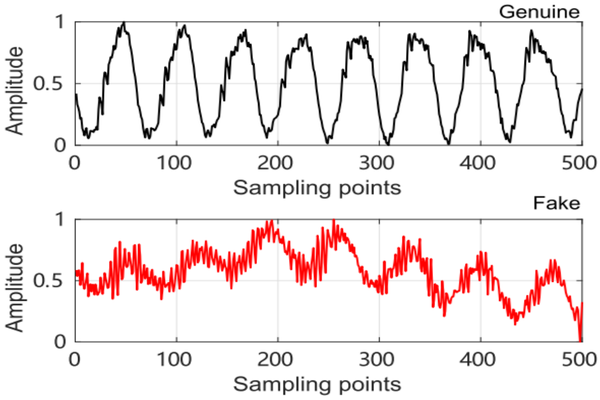


Figure 2: Example of different sampling in case of deep audio fakes detection (**Almestekawy et al., 2024)**

1.1 Research Problem

The recent appearance of audio deepfakes has put a new security threat in front of social networks and web-services. Deepfake audio can be used to mimic the voices of people, to disseminate fake news, or to perpetrate identity theft jeopardizing the credibility of the existing modes of digital communication. Current detection algorithms provide low generalization capabilities across the diverse deepfake generation approaches and do not consider the improvement of methods in fake audio construction. The main research issue is embedded in how and what kind of architecture can be made to have high TPR, low FPR, be simultaneously scalable and flexible enough to accommodate increasing volumes of audio deepfakes. Deep learning is combined with signal processing to contain temporal and frequency information of the sound; this enhances the signal's reliability and sensibility to detection (**Khalid et al., 2024**). The integration of spatial representations (deep learning features) and temporal variations (signal processing metrics) allow even the finest manipulations to be detected at even greater levels of efficiency (**Kumar & Kundu, 2024**).

1.2 Motivation

The problem of audio deepfakes is quickly growing, and its consequences are prospective threats to privacy, businesses, and society as a whole. It has also resulted to so many incidences of fraud and identify theft thus making such technologies vulnerable in need of detection methods that are reliable. This research is motivated by the need to defend digital communication platforms from threats such as jamming through the use of current developments in signal processing and; deep learning. Their integration shall lead to a construction of a enhanced detection procedure that can identify even increasingly minor manipulations in audio, to protect the safety of social media and web platforms.

1.3 Research Background

In the last few years, scientists tried to review a number of methods to identify the deepfakes audios where feature sets such as MFCC and spectro-temporal analysis comprehended. Research has also explored employing deep learning models as CNNs and RNNs for analysing temporal and spectral cues that exist in synthesised sound [3]. However, these methods present a low generalization to newly developed deepfake generation methods or unprecedented methods. Despite the development of various signal processing methods and state of the art machine learning algorithms rather than applying both approaches provide a promising way for improving detection accuracy and making it less sensitive to noise [2].

1.4 Research Solution

In this research, a security application that combines signal processing and deep learning frameworks to overcome the drawbacks of existing audio deepfake detection is proposed. The solution employs the state-of-the-art feature extraction methods, MFCC and spectral contrast, along with transformers and convolutional models' deep learning. As a temporal and spectral analysis, the framework should be able to detect manipulations which are deepfake audio inherently possesses. The solution also employs rigorous preprocessing and hyperparameters tuning techniques for better scalability in a different pre-processing environment and sets of real-life scenarios.

Chapter 2: Literature Review

2.1 A security application for social media and web platforms to identify the sound based deepfakes using signal processing infused deep learning framework

Due to the technology development in deep learning, synthetic media which is also referred to as deepfake has risen tremendously. Whereas original risks were focused on the video deepfakes, which modify the visual content, new kinds of deepfake attacks, audio deepfakes, are being considered as potentially malicious kinds of deepfakes at the moment. Voice clones, which refer to spoofing using the respective AI to create realistic audio voices associated with real people, apply various ramifications of phishing, fraud, and identity theft. These audios can be used by bad actors to mimic someone's voice, evade voice authentication, or manipulate people or firms in a range of personal or business settings (**Tiwari et al., 2023**).

2.2 The Growing Threat of Audio Deepfakes

Much of the work in deepfake detection has thus far been on videos, but audio deepfake is even more dangerous because it can convincingly fool both human and artificial intelligence. As an example, the real-time attackers can mimic the voice of the CEO or some other authoritative personnel to perform fictitious transactions, which has been observed in specific research works. This underlines the importance of the improved detection methods more focused on the audio deepfakes as the future research study conducted by (**Agnew et al., in 2024**). Unlike modifying of the messages in visual media, detecting the audio deepfakes one needs to analyze not only the content of the speech but also the acoustic aspect of the message, so the detection is harder.

One major problem is that audio deepfakes can easily defeat current voice authentication solutions that are commonly deployed for banking, calling center and security applications. Previous techniques of vocal recognition involve aspects like elevation, tone, and rhythm by basically mimicking the exact pattern through use of AI trained on large audio platforms (**Triantafyllopoulos et al., 2023**). Such elaborated models hint at the necessity of increasing the security level and developing the detection method researches and ways (**Yu et al., 2024**).

2.3 Deepfake Detection Techniques

A lot of research has been done on the deep fake detection techniques where authors have used CNN with random forest approaches. Among them, the one that combines CNNs with Recurrent Neural Networks (RNN) for temporal inconsistencies in videos and audios turned out to give better detection ratios. In this method, it is shown that temporal features—the variations in audio or video data across time—have significant importance in detecting deepfakes. Through analysing these disparities, the model will be able to differentiate the authentic media content from the forged content better than conventional methods proposed by (**Nailwal et al., 2023**).

Subsequent studies have extended the entropy-based costing in more detail especially when combined with CNNs and RNNs. The entropy based methods calculate the level of disorder or randomness of a system and thus, in case of deepfake detection they give an estimate of the extent of manipulation done. Several researchers have revealed that through using entropy based methods, one and the same results can be achieved when using benchmark data sets, thought provoking researchers to consider the other options as possible for increased general use in real situations (**Yu et al., 2024**).

Besides temporal analysis and entropy-based methods, other signal processing including Mel-frequency cepstral coefficients (MFCC), chroma short-time Fourier transform (STFT), and spectral contrast, have also been integrated with machine learning models for detecting audio deepfakes. MFCCs use short frames of audio data and are speaker specific; chroma STFT and spectral contrast give information about harmonic content and pitch shifts in the audio signal. When combined with deep learning approaches, such as CNNs and random forests, these signal processing features appear to hold much potential for identifying highly complex audio manipulations (**Zhang et al., 2024**).

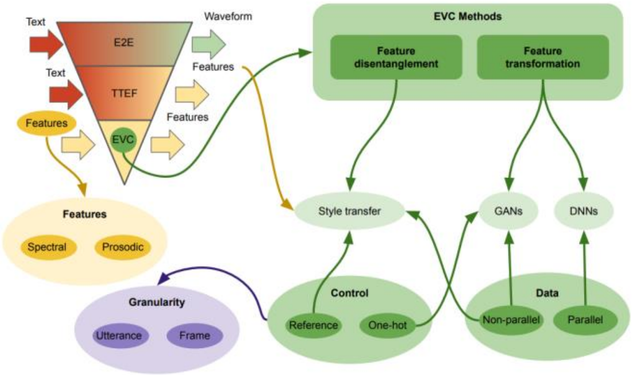


Figure 3: An explanation of how the features extraction process help in the deep fake detection (**Zhang et al., 2024)**

[7]Taxonomy of deep emotional speech synthesis approaches. Approaches can be primarily differentiated according to the following ways: (a) How many steps of the synthesis they incorporate, which is in term determined by their input and output, accordingly categorised as end-to-end (E2E), text-toemotional-features (TTEF), or emotional voice conversion (EVC) methods. (b) How control is achieved, as well as the level of granularity that this control can achieve. (c) For EVC methods, on whether they use parallel or non-parallel data. (d) For non-parallel data EVC methods, based on whether they rely on disentangling speech components or directly mapping features to capture the target emotion, as all parallel data methods use the latter form of conversion; TTEF methods instead primarily fall under the style-transfer category

2.4 Challenges in Multimodal Deepfake Detection

Unlike previous works that rely on the unimodal deepfake detection, whether based on images or text, the growth of the multimodal deepfakes will serve as a potent challenge. Multimedia deepfakes contain some components of two different forms of media; video and audio or video and text or audio and text making them harder to identify. One of the chief challenges germane to fusion of information from different modalities is how to cross-reference information in one modality puts in another to point out disparities (Zhang et al., 2024).

One of the approaches developed toward solving this problem is the Consistency-Checking Network (CCN). This is like human decision-making because CCN breaks-down incoming stimuli into different modalities, and then compares the stimuli for their consistency. For instance, when a deepfake video has sound, the CCN will check whether the speaker's lips align with the words spoken, tiny irregularities that signify that deepfake is at play (Tiwari et al., 2023).

Likewise, HAMMER which has hierarchical multimodal manipulation reasoning transformer has been put forward to assess the interaction between the modalities. The modeling capability of HAMMER enables it to understand relations between various media types and is a better depiction of the manipulation used. This approach assists to meet increased complexity of deep fake work especially when they gets to be multimodal deep fakes (Nailwal et al., 2023).

2.5 The Role of Social Media Platforms

There are six types of disinformation that have been identified to exist in TikTok, these ones include; low volume disinformation, satire/parody, manipulated content and lastly fake content. It was concluded that recommendation algorithm of the platform based on the "For You" page helps to increase the circulation of the content related to the conflict areas, such as Ukraine, which may involve the fake or misleading information. This amplification may be especially ill-suited when deepfakes are present, given that the platform is centered around trends and audio-based engagement (**Bösch and Divon, 2024**).

2.6 Future Directions for Deepfake Detection

The creation of deepfake increases the complexity of generation methods, and therefore, strenghtens the need to come up with improved detection methods. Thus, without question, future work can easily get ahead on identification and subsequent eradication of deepfake effectively, additional studying needs to consider deepfake technologies as complex entities that must be defeated using versatile, efficient and scalable solutions. This entails designing detectors that are robust to both the unimodal and multimodal deepfakes and the ability of the detection algorithms to respond to such new forms of manipulation from (**Ganga et al., 2024**).

The common goal for countering deepfakes will require a combined effort of working researchers, policymakers, and social media platforms apart from developments in technology. This paper has indicated some of the potential harms that are likely to arise from synthetic media including identity theft, political manipulation, scams, and cyberbullying (**Opdahl et al., 2023**).

2.7 Research on detecting sound based deepfakes using signal processing infused deep learning framework

The paper by (**Khalid et al., 2024**) presents ExplaNET, a framework to improve deepfake detection models interpretability through interpretable prototypes of manipulated facial features. This approach also enables the model to indicate some features in the manipulated images while doing so in a more comprehensible and recognizable manner to users. In this way ExplaNET robust improves user trust, a key component for the application of these systems in real-world scenarios**.**
In the audio domain **Li et al. (2023)** consider two pitch-contour within-phoneme variations as primary for detecting deepfakes: "jitter" and "shimmer." The authors of the work state that due to significant differences of these voice characteristics in real and synthetic audio they can be used as markers of audio manipulation. Combined with jitter and shimmer, other acoustic measures raise the level of true synthetic audio distinction in their method. This methodology underlines the actual need to mask the

differences between exact audio features in fighting deepfake audio, especially in the applications that need actual and precise detection **(Li et al., 2023)**.

Deepfake detection system proposed by (**Kumar and Kundu, 2024**) based on big data techniques primarily to focus on cybersecurity and is scalable for social media platforms. Their system uses these two categories of algorithms whereby the big data frameworks are used on large amounts of media contents to enable accurate live monitoring and threat detection which is crucial to social media applications. This approach solves for scalability, another issue that arises in high-throughput environments whereby, more data requires a faster rate of processing in order to maintain cybersecurity and platform sustainability.

In this study, Raza et al. (2023) introduce HolisticDFD, a deepfake detection model based on ST-Transformer where spatiotemporal features in videos are learned using transformer embeddings. By doing this, their model can identify manipulations related to frames that consist of a video, providing a holistic detection of deepfake videos. It becomes apparent that this method does a good job of incorporating the image-based features with temporal variations and constraints, a problem of static image models that seldom consider manipulations over time. I have shown that HolisticDFD's spatiotemporal nature leads to a better estimation of content at deeper levels and a better identification of deep fake videos **(Raza, Malik, & Haq, 2023)**.

Almestekawy et al. (2024) extend the discussion of video-based detection by integrating spatiotemporal texture analysis technique with deep learning. Their proposed method designed specifically for deepfake detection in videos' is founded on the concept of texture analysis over the frames of the original videos and the CNNs to extract deep features within those frames. The fusion of spatiotemporal texture and deep learning feature provides the model with a good and balanced feature which enables the model to detect if a video is real or a deepfake. They found that combining spatiotemporal dynamics with the spatial features' extraction enhances the model's capacity to detect slight manipulations in videos, and therefore they suggested that the current conventional spatial and temporal methods are inadequate **(Almestekawy, Zayed, & Taha, 2024)**.

2.8 Research Gap

Despite significant advancements in deepfake detection, the focus has primarily been on video-based deepfakes, leaving audio-based deepfakes relatively underexplored. This is concerning given the increasing sophistication and malicious potential of audio manipulations, such as voice cloning, which can bypass voice authentication systems and enable identity theft, fraud, and phishing attacks. Current approaches to detecting audio deepfakes often struggle with the complexities of analyzing acoustic features and temporal variations inherent in audio signals. Additionally, the rise of multimodal deepfakes, which combine multiple forms of media like audio, video, and text, introduces further challenges, as most existing models are unimodal and fail to address the interdependencies between different modalities. Scalability remains another critical issue, particularly in real-time applications on platforms like social media, where the high volume of data demands faster and more efficient processing. Furthermore, the interpretability of existing detection models is limited, reducing user trust and their practical applicability. Finally, the rapid evolution of deepfake generation techniques continues to outpace current detection capabilities, necessitating the development of robust and adaptable frameworks.

2.9 Research Contribution

This research addresses critical gaps in audio deepfake detection by proposing a novel signal processing-infused deep learning framework. Advanced techniques like Mel-Frequency Cepstral Coefficients (MFCCs) and acoustic markers such as jitter and shimmer are combined with deep learning architectures to capture subtle temporal and spectral manipulations in audio. For example, jitter, a measure of pitch variability, is calculated as:

$$Jitter = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{|T_{i+1} - T_i|}{T_i}$$

where T_i is the fundamental period in frame i. This helps identify synthetic alterations in audio signals. To address the complexity of multimodal deepfakes, a Consistency-Checking Network (CCN) is introduced, ensuring alignment between modalities such as audio and video. The system minimizes inconsistency through a loss function:

$$L_{CCN} = \left||E_a - E_v|\right|_2^2$$

where Ea and Ev represent audio and video embeddings, respectively. By incorporating scalable architectures and interpretable models, this framework enhances detection accuracy and real-time applicability, particularly in high-throughput environments like social media platforms.

Chapter 3: Methodology

This work utilizes the two already existing datasets called ‘In-the-Wild’ and Fake-or-Real (FoR) in creating a more reliable and encyclopedical framework of detecting fake audio deepfakes. Each dataset has its own difficulties and features and the features, which are specific for the dataset, allow testing machine learning and deep learning models under different circumstances, clean audio samples and re-recorded distorted audio. The following are the process of the methodology; preparation and preprocessing of the dataset, creating the model, evaluating the performance of the model and the deployment of the model.

3.1 Dataset Descriptions

3.1.1 ‘In-the-Wild’ Dataset

This dataset contains 20.8 hours of real audio and 17.2 hours of fake audio, with a total of 19,963 real files and 11,816 fake files. The audio samples, derived from social media and video-sharing platforms, represent diverse real-world conditions. On average, each speaker has 23 minutes of real and 18 minutes of fake audio. This dataset is ideal for evaluating models in scenarios mimicking real-world deployment environments.

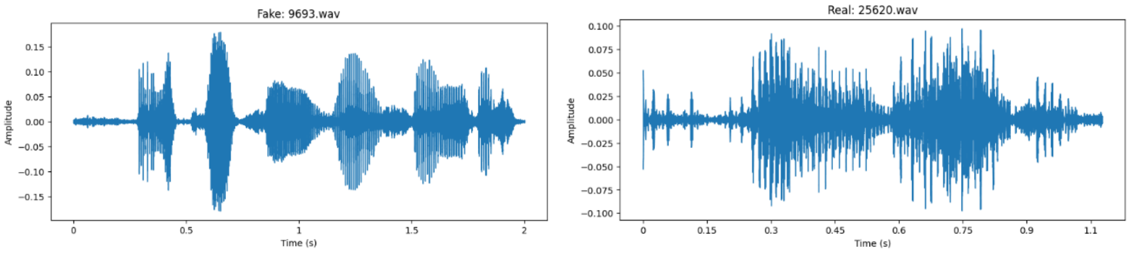


Figure 4: In the Wild Dataset example

3.1.2 Fake-or-Real (FoR) Dataset

The ‘for-rerec’ subset of the FoR dataset is used in this study. It comprises audio samples processed through voice channels (e.g., phone calls), introducing distortions and noise that simulate real-world challenges. This dataset offers 195K audio samples, split across categories of real and fake speech synthesized using advanced text-to-speech (TTS) models like Deep Voice 3 and Google Wavenet. It includes unique subsets such as re-recorded and normalized audio, allowing the study to train models for robust detection.

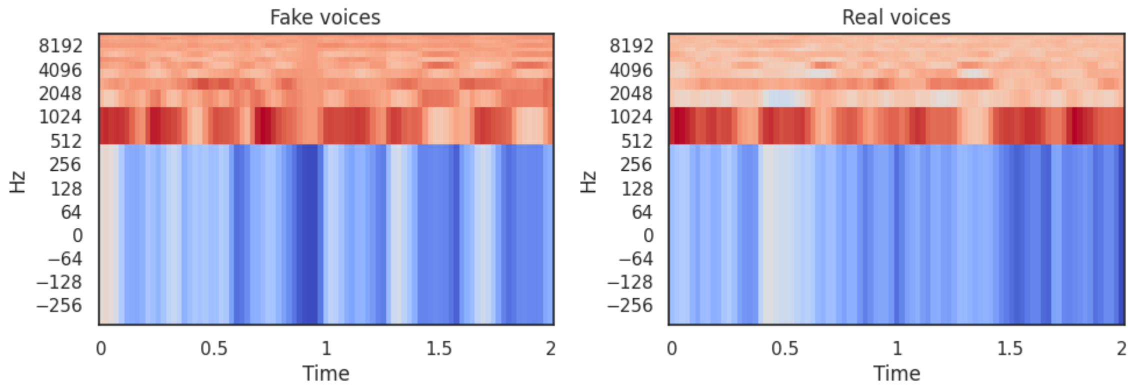


Figure 5: FoR Dataset example

3.2 Data Preprocessing

3.2.1 Feature Extraction

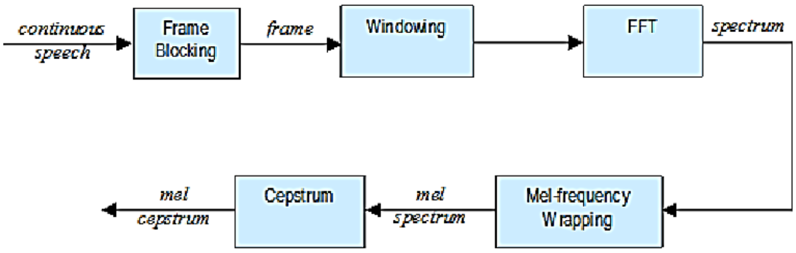


Figure 6: Features Extraction from the sound signals process (**Source: Towards Data Science**)

Chroma Short-Time Fourier Transform (Chroma STFT): This feature used to describe the harmonic or Chroma STFT, Spectral Centroid, Spectral Rolloff-related content of audio and helps analyze Chroma STFT, Spectral Centroid, Spectral Rolloff-related patterns.

Root Mean Square (RMS): Indicates the intensity of the audio signal. It calculates the average power of the waveform over time.

Spectral Centroid: This identifies the 'centre of mass' of the sound spectrum, indicating brightness of the audio.

Spectral Rolloff: It helps identify how the energy is distributed in the spectrum.

Zero-Crossing Rate: This feature is used to measure the number of times the audio waveform crosses the zero line in a defined time period.

Mel-Frequency Cepstral Coefficients (MFCCs): Here, we calculate 20 coefficients per audio file that summarizes essential tonal quality and texture aspects of the audio. These coefficients represent the short-term energy patterns of sound.

3.2.2 Standardization and Label Encoding

Features were standardized to ensure uniform scaling with a mean of 0 and a standard deviation of 1. Labels were encoded numerically as **1 for real audio** and **0 for fake audio**, ensuring compatibility with machine learning models.

3.2.3 Dataset Splitting

For the **'In-the-Wild' Dataset,** Split into training (80%) and test (20%) sets while in case of **FoR Dataset,** Split into training (70%), validation (15%), and test (15%) sets for robust evaluation.

3.3 Machine Learning and Deep Learning Models

Logistic Regression is a statistical model used for binary classification. It predicts probabilities using a logistic (sigmoid) function, making it suitable for problems with linearly separable data. While simple and interpretable, it struggles with non-linear relationships unless features are carefully engineered.

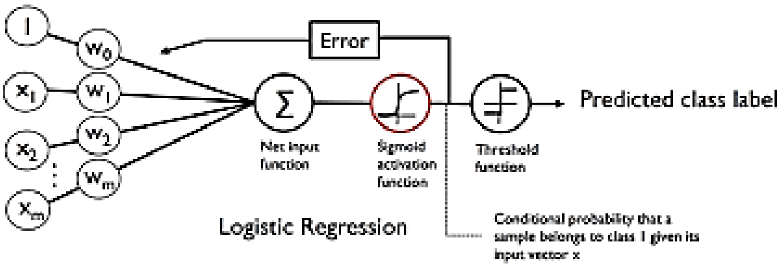


Figure 7: Illustration of Logistic Regression with the use of sigmoid function (**Source: Wikipedia**)

Random Forest is an ensemble learning technique that combines multiple decision trees to improve performance. By training trees on random subsets of data and features, it handles non-linear data well and reduces overfitting through averaging, though it can be computationally intensive for large datasets.

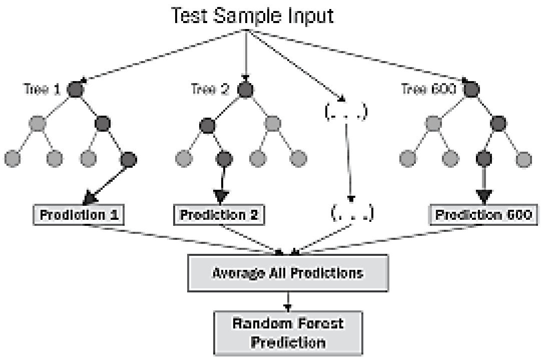


Figure 8: Illustration of Random Forest (**Source: Geeks For Geeks**)

XGBoost is a gradient boosting framework that builds decision trees sequentially to minimize errors. It's highly efficient and accurate, especially on structured data, but requires careful tuning to avoid overfitting, especially on small datasets.

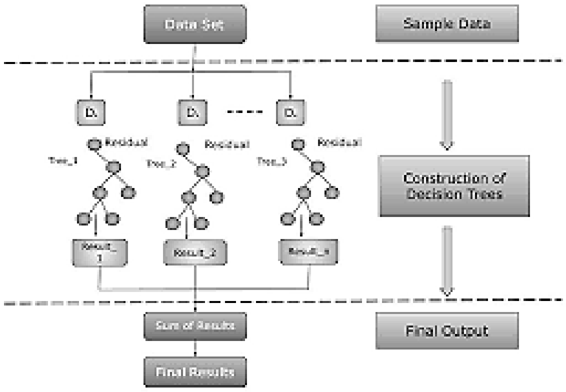


Figure 9: XGBoost (**Source: Nvidia**)

LightGBM is another gradient boosting framework, optimized for speed and memory efficiency. Unlike XGBoost, it splits trees leaf-wise and supports categorical features natively, making it faster on large datasets, though it may underperform on smaller ones.

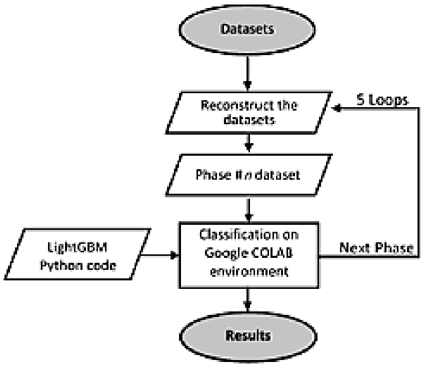


Figure 10: Flow diagram for Light Gradient Boosting (**Source: Javapoint**)

CNNs (Convolutional Neural Networks) excel at tasks involving spatial data like images. They extract hierarchical features using convolutional and pooling layers, followed by classification layers. While they deliver state-of-the-art results in image analysis, they demand large datasets and high computational power.

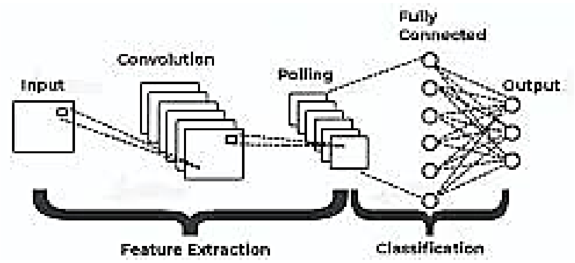


Figure 11: CNN architecture (**Source: Deeplearning.io**)

LSTMs (Long Short-Term Memory Networks) are specialized for sequential data, such as time series and text. They address the limitations of traditional RNNs by retaining long-term dependencies, though they are slower to train and may face challenges with very long sequences.

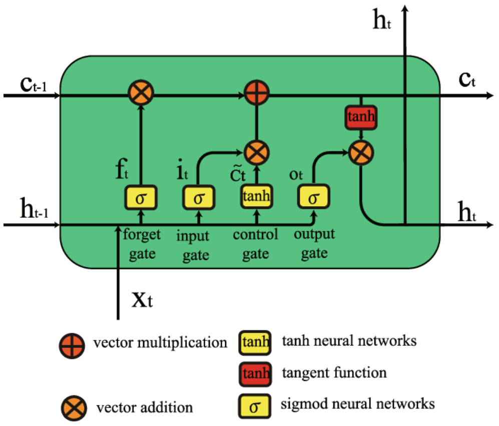


Figure 12: LSTM Cell (**Source: Researchgate**)

CNN + LSTM models combine the strengths of CNNs and LSTMs. CNNs extract spatial features, which are then passed to LSTMs for learning temporal patterns. This hybrid architecture is ideal for tasks like video analysis but comes with increased computational complexity.

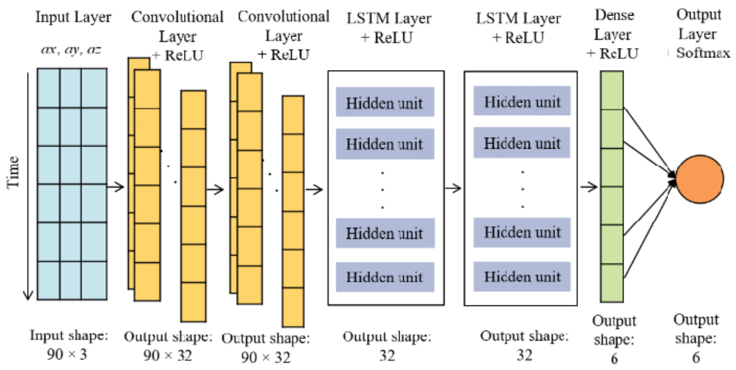
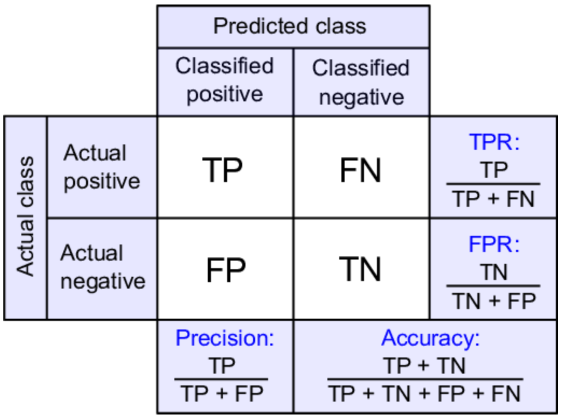


Figure 13: CNN based LSTM architecture (**Source: Research gate**)

3.4 Performance Evaluation

The models were evaluated using metrics such as **accuracy**, **precision**, **recall**, and **F1-score**. The **'In-the-Wild' dataset** allowed higher accuracy across models due to its clean audio samples, while the **FoR dataset** highlighted the need for robust models capable of handling distortions and noise.



		Predicted class		
		Classified positive	Classified negative	
Actual class	Actual positive	TP	FN	TPR: $\frac{TP}{TP + FN}$
	Actual negative	FP	TN	FPR: $\frac{TN}{TN + FP}$
		Precision: $\frac{TP}{TP + FP}$	Accuracy: $\frac{TP + TN}{TP + TN + FP + FN}$	

Figure 14: Confusion Matrix (**Source: Research gate**)

To evaluate the performance and effectiveness of our models in classifying Deepfake audio, we employed two primary metrices:

Accuracy: It measures the overall correctness of the model across all classes. It indicates the fraction of correct predictions made by the model. In this task of binary classification, achieving high accuracy is very important so that true samples as well as manipulated audio samples can be correctly labeled and misclassifications minimized.

$$Accuracy = \frac{\textit{True Positives} + \textit{True Negatives}}{\textit{Total Predictions}}$$

Precision: Precision is a metric that measures how often a machine learning model correctly predicts the positive class. For audio classification in deepfakes, high precision is highly necessary to ensure the majority of cases reported as 'fake' are actually fraudulent.

$$Precision = \frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Positives}}$$

Chapter 4: Implementation

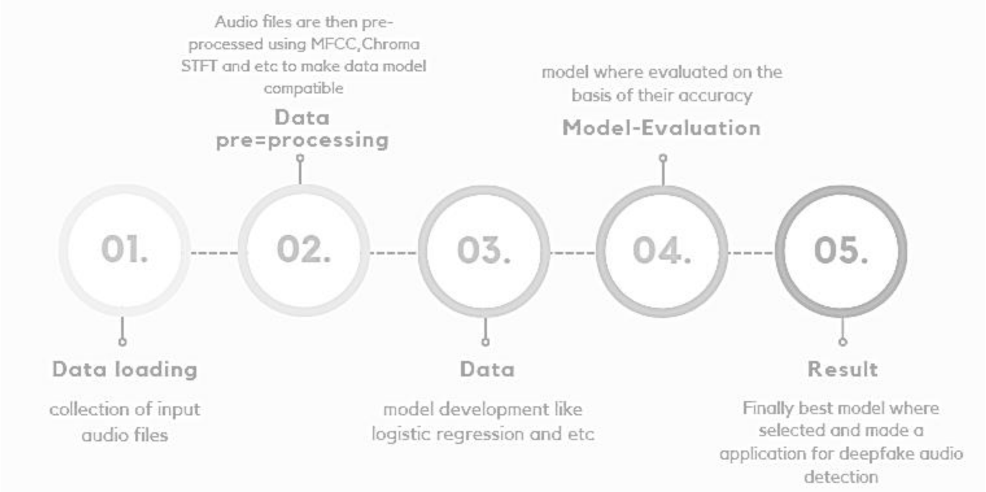

Audio files are then pre-processed using MFCC,Chroma STFT and etc to make data model compatible
Data pre=processing
model where evaluated on the basis of their accuracy
Model-Evaluation
01.
02.
03.
04.
05.
Data loading
collection of input audio files
Data
model development like logistic regression and etc
Result
Finally best model where selected and made a application for deepfake audio detection

Figure 15: Implementation Framework for Deepfakes detection using the audio samples

The implementation phase of this study involves the development, training, and evaluation of multiple machine learning and deep learning models for detecting audio deepfakes using two datasets: **'In-the-Wild'** and **Fake-or-Real (FoR)**. The implementation pipeline was meticulously designed to handle the unique challenges posed by both datasets, ensuring robust and scalable model development.

4.1 Dataset Preparation

Both datasets were preprocessed to ensure consistency and usability: **'In-the-Wild' Dataset:** Audio recordings were processed to extract key features like **Chroma STFT**, **Spectral Centroid**, **MFCCs**, and **RMS**, capturing spectral and temporal properties critical for distinguishing real and fake audio. The dataset was split into training (80%) and testing (20%) sets. **Fake-or-Real (FoR) Dataset:** The **'for-rerec'** subset was used due to its real-world relevance, with distortions and noise mimicking conditions such as phone calls. Similar features were extracted, and the dataset was divided into training (70%), validation (15%), and testing (15%) sets. Both datasets underwent standardization, ensuring feature values had a mean of 0 and a standard deviation of 1. Labels were encoded numerically (1 for real audio and 0 for fake audio).

4.2 Data Pre-Processing

The data preprocessing focused on extracting important audio features to describe each file. These features include:

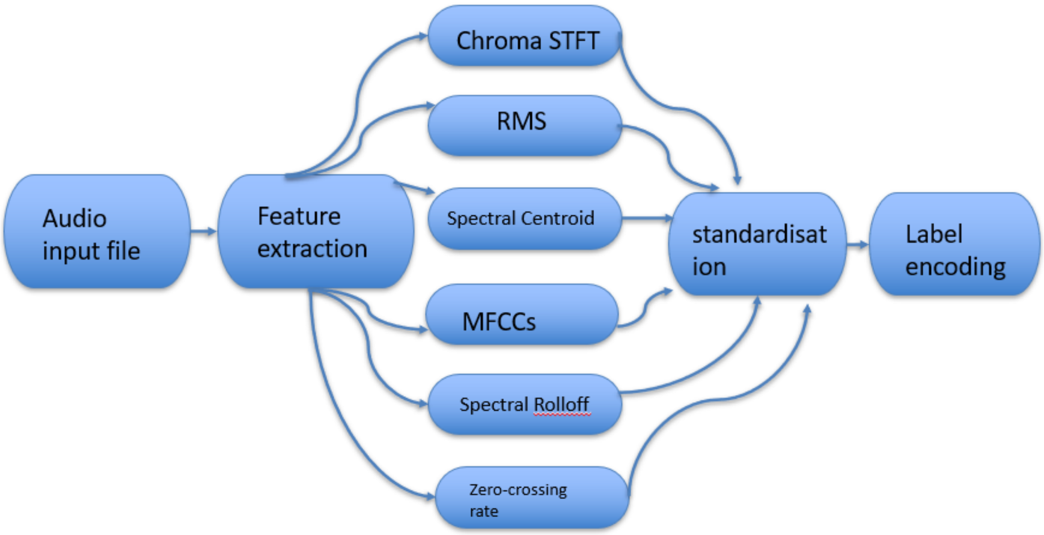


Figure 16: Features Extraction Process flow

4.2. Model Development

Using Machine Learning Models: Logistic Regression: Implemented as a baseline, Logistic Regression modeled the linear relationship between features and class labels. It achieved **94.83% accuracy** on the 'In-the-Wild' dataset but struggled with the distorted FoR dataset, achieving only **58% accuracy**. **Random Forest:** An ensemble model leveraging 100 decision trees achieved **98.28% accuracy** on the clean 'In-the-Wild' dataset but faced challenges with distorted audio, achieving **60% accuracy** on the FoR dataset. **XGBoost:** The gradient boosting algorithm achieved **99.20% accuracy** on the 'In-the-Wild' dataset, making it the best performer on clean audio. However, it achieved only **47% accuracy** on the FoR dataset, indicating difficulty with distortions.

Using Deep Learning Models: Convolutional Neural Networks (CNNs): CNNs were designed to capture local spectral patterns in audio. They achieved **98.23% accuracy** on the 'In-the-Wild' dataset and **58% accuracy** on the FoR dataset. **Long Short-Term Memory Networks (LSTMs):** LSTMs modeled temporal dependencies, achieving **95.12% accuracy** on the 'In-the-Wild' dataset and **63% accuracy** on the FoR dataset. Their sequential modeling capability improved robustness against distortions. **Hybrid Model (CNN+LSTM):** This model combined CNNs for feature extraction and LSTMs for temporal modeling. It achieved **69% accuracy** on both datasets, demonstrating consistent performance across clean and distorted audio.

Using Additional Models: LightGBM: Achieved moderate performance with **49% accuracy** on the FoR dataset, reflecting its efficiency on structured datasets. **Logistic Regression:** Served as a baseline model, providing a strong starting point for feature evaluation.

4.3 Training and Evaluation

Each model was trained on the preprocessed datasets using relevant hyperparameters: **Logistic Regression:** Trained using the fit() method from scikit-learn, with predictions evaluated using accuracy and precision metrics. **Random Forest and XGBoost:** Ensemble models were optimized to handle the diversity of features, with Random Forest achieving better results on clean audio and XGBoost excelling in gradient-based learning. **CNN and LSTM Models:** Deep learning models were implemented using TensorFlow/Keras, with Adam optimizer and binary cross-entropy loss. Training involved 10 epochs with a batch size of 32, monitoring loss and accuracy trends to prevent overfitting. **Hybrid CNN+LSTM:** Combined architectures to leverage both spatial and temporal feature extraction, resulting in robust performance across datasets.

Chapter 5: Result and analysis

Case 1: Using In the Wild dataset

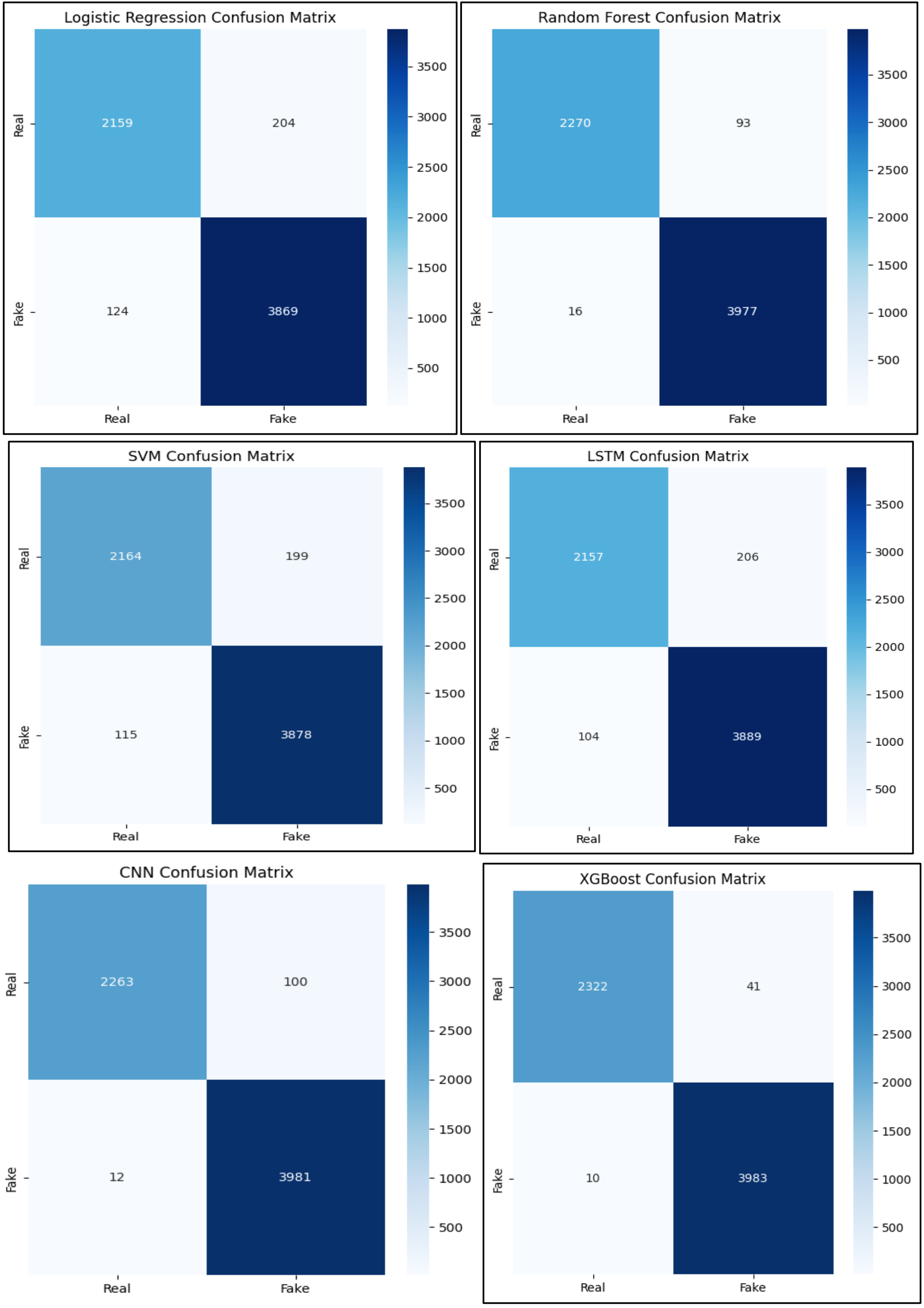


Figure 17: Confusion Matrix comparison of (a) Logistic Regression (b) Random Forest Classifier (c) Support Vector machines (d)LSTM (e) CNN (f) XGBoost

From the confusion matrix of the logistic regression model on ‘In-the-Wild’ dataset, it is seen that 6028 out of the total 6356 test samples have been correctly classified, achieving an accuracy of 94.84%. From the confusion matrix of the Random Forest Classifier model on ‘In-the-Wild’ dataset, it is seen that 6247 out of the total 6356 test samples have been correctly classified, achieving an accuracy of 98.29%. From the confusion matrix of the SVM model on ‘In-the-Wild’ dataset, it is seen that 6042 out of the total 6356 test samples have been correctly classified, achieving an accuracy of 95.06%.

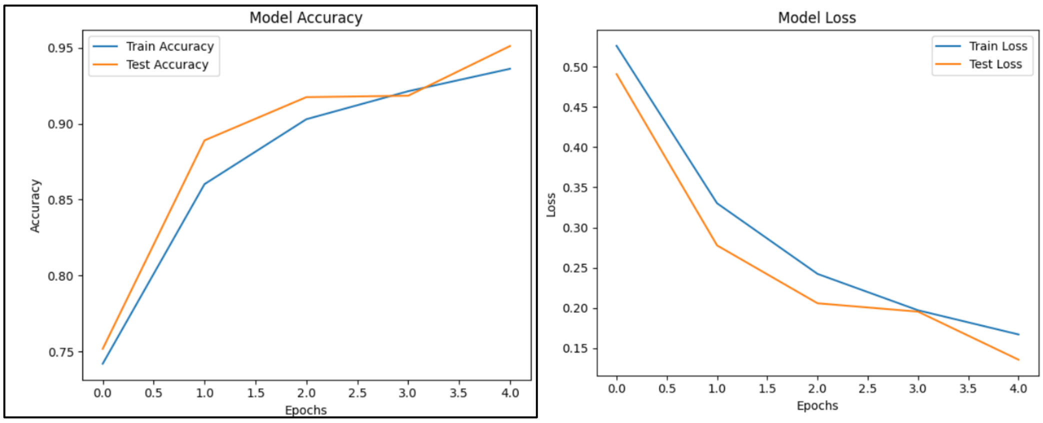


Figure 18: Loss and Accuracy Curved over Epochs for the LSTM model

This graph shows the progression of training and test accuracy over epochs. Initially, in epochs 0 to 1, the model has low training accuracy since it's trying to learn intrinsic patterns in the data, whereas it has a much faster improvement for test accuracy which means it learns the overall trends very quickly. In the middle phase (epochs 1 to 3), train accuracy is steady due to the learning of more accurate representations by the model, and test accuracy also rises a little beforehand than train accuracy, which indicates that this model generalizes well and is not overfitting. For the final stage, epoch 3 through epoch 4, learning continues to improve but at a slower pace, indicating that the model is fine-tuning its learning process. The second graph illustrates how the losses on the training and test cases progress to decrease over epochs as the model learns. The train loss is high at first, meaning many errors during training but test loss drops rapidly, indicating that the model is very well capturing general patterns. Both the losses drop steadily as training progresses onward, but test loss stays a little lower than the train loss, meaning good generalization without overfitting. In the final stages, the training loss decreases more slowly where the model has improved its knowledge. The test loss remains at a lower value, which indicates that the model is doing very well on new data. This steady decrease in loss shows that the model is well-trained and stable.

From the confusion matrix of the SVM model on ‘In-the-Wild’ dataset, it is seen that 6046 out of the total 6356 test samples have been correctly classified, achieving an accuracy of 95.12% which is marginally better than SVM model.

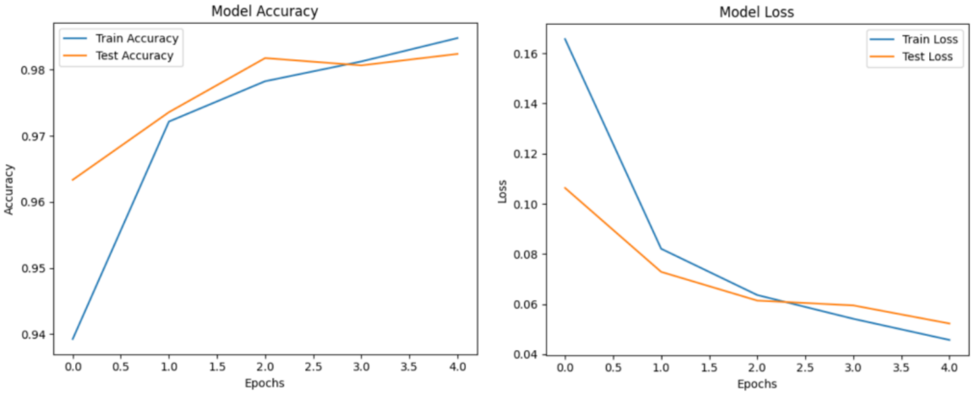


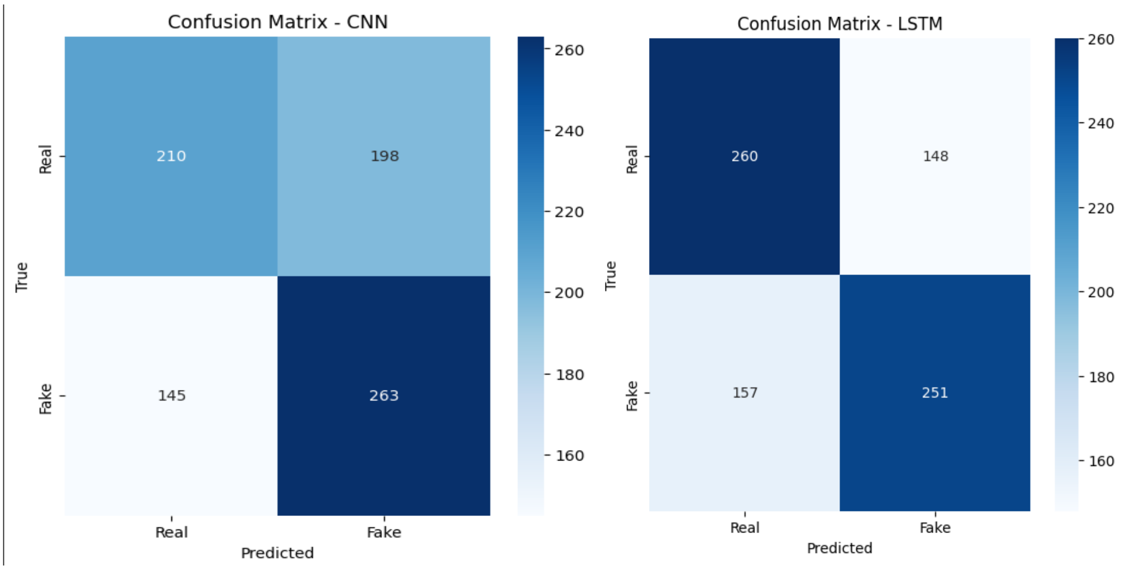
Figure 19: Accuracy and Loss curves for the CNN model over the epochs

The accuracy graph demonstrates that both the training and testing accuracies for the CNN model steadily improved over the course of 5 epochs. Initially, both accuracies started around 0.94 and quickly increased, with testing accuracy rising slightly faster than training accuracy. By the third epoch, both accuracies were close to 0.98, and they stabilized by the fourth epoch, indicating that the model was learning effectively from the training data and generalizing well to the test data. The overall trend shows that the model achieved high accuracy without overfitting, maintaining a consistent performance on both the training and testing sets. This loss graph depicts the amount of error, or loss, of the CNN model for the training data (blue line) and the testing data (orange line) over the course of 5 epochs of training steps. The training loss is high at first but drops drastically as the model starts matching the data. In any case, the test loss also goes down fairly well, showing the model does better on new data it hasn't seen yet. The final lines come together and stabilize, meaning that it's learned well and not overfitting. Overall, the steady drop in both losses, indicating that the model is doing a good job. From the confusion matrix of the CNN model on ‘In-the-Wild’ dataset, it is seen that 6244 out of the total 6356 test samples have been correctly classified, achieving an accuracy of 98.24%. From the confusion matrix of the XGBoost model on ‘In-the-Wild’ dataset, it is seen that 6305 out of the total 6356 test samples have been correctly classified, achieving an accuracy of 99.20% which is best among all the models.

Table 2: Precision and Accuracy comparison of all models for Wild dataset

MODEL	**ACCURACY**	**PRECISION**
Logistic Regression	94.83%	94.99%
Random Forest Classifier	98.28%	97.71%
Support Vector Machine	95.05%	95.11%
LSTM	95.12%	94.96%
CNN	98.23%	97.54%
XGBoost	99.19%	98.98%

5.2 Case2: Using FOR dataset



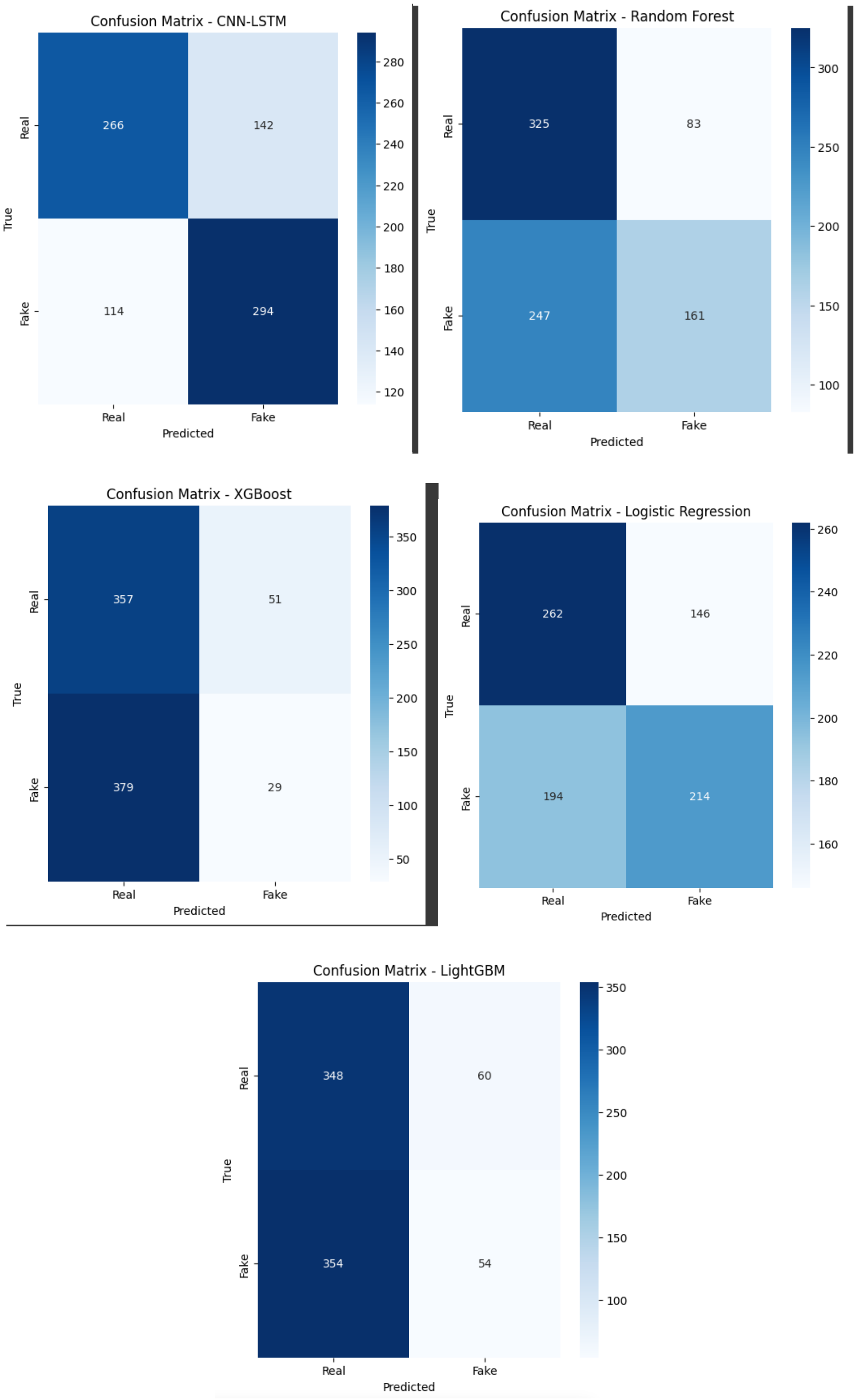


Figure 20: Confusion matrix comparison of all the models using the FOR dataset

The comparison of several models for classifying fake or real audios reveals a high performance of the proposed hybrid model CNN+LSTM. This model gave the greatest top precision of 69% and F1-measures 0.68 (Class 0) and 0.70(Class 1) than the other methods. As shown by Fig. 4, during training, both the accuracy and precision gradually rose and validation accuracy began to oscillate slightly after the sixth epoch or two, suggesting a mild case of over-training. Possible loss trends ensured that learning was effective by having a validation loss, rise slightly after epoch 7. The overall accuracy and the model learning capability of CNN are relatively high as compared to other models but the validation accuracy has the danger to oscillate after the 7th epoch and the model starts overfitting. The LSTM model does show good training/validation split balance with little overfitting with the accuracy of 63% and the F1-scores of around 0.62. Models such as Random Forest and logistic regression had decent practices of 60% and 58% respectively with random forest having a better recall rate in class 0. XGBoost and LightGBM gave relatively poor accuracies-47% and 49% respectively and low levels of both precision and recall.

It is evident from the results that the best model for this task will be the CNN + LSTM since it exhibits the best overall performance of the recognized features due to its capability to extract the spatial and temporal qualities. Other models have issues of over fitted models, or issues of class imbalance, or issues with learning performances.

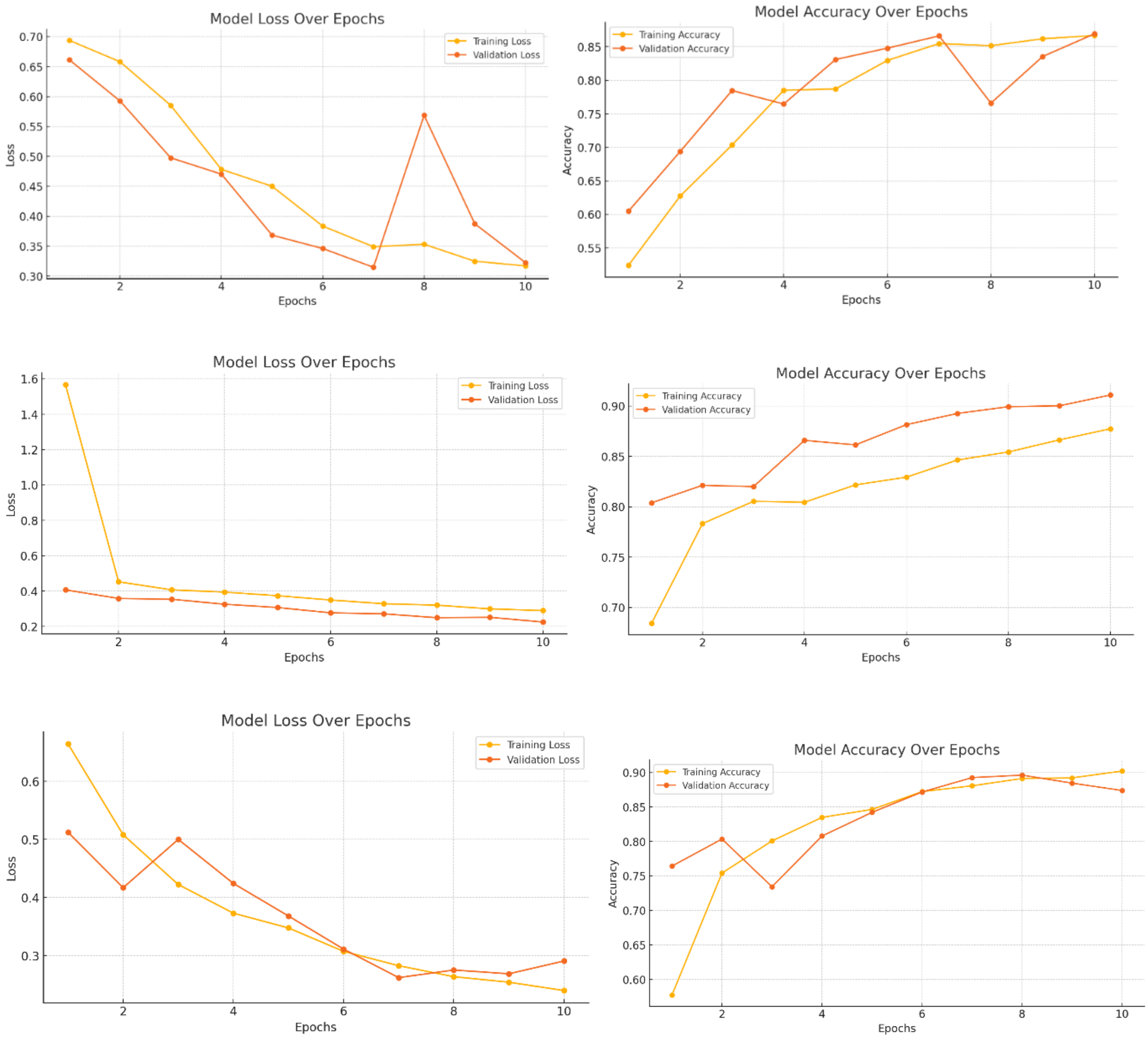


Figure 21: Loss and Accuracy curves for (a) CNN model (first 2 graphs) (b) LSTM (middle row) (c) CNN + LSTM (bottom row)

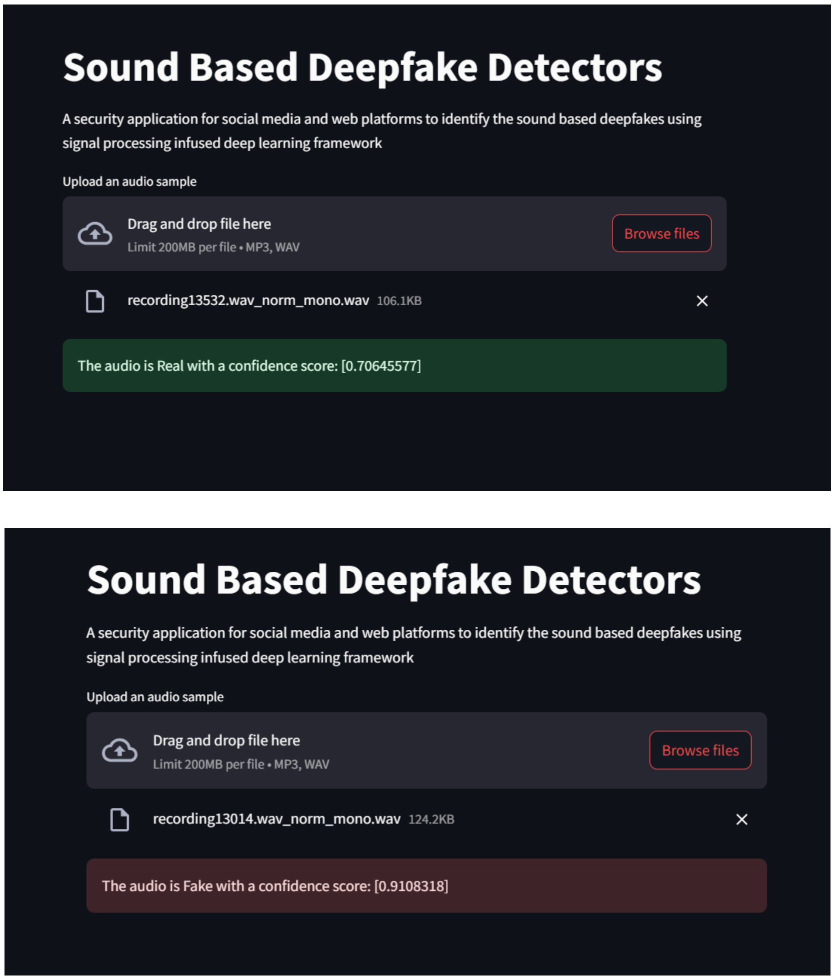
After examine all the model Hybrid model(CNN+LSTM) is identified best suited for the identification of fake or real audios.

5.3 Discussion

This study presents a comprehensive framework for detecting audio deepfakes, addressing a critical challenge in safeguarding social media and web platforms against malicious misuse. By leveraging two diverse datasets, **‘In-the-Wild’** and **Fake-or-Real (FoR)**, the research demonstrates a robust approach to handling both clean and distorted audio samples, ensuring adaptability to real-world scenarios. The **XGBoost** model emerged as the best performer on the **‘In-the-Wild’ dataset**, achieving remarkable accuracy and precision in detecting deepfakes in cleaner audio environments. Meanwhile, the **hybrid CNN+LSTM** model proved its robustness on the **FoR dataset**, effectively managing the complexities introduced by distortions and noise. This dual-model deployment strategy highlights the importance of tailoring detection systems to the specific characteristics of the data, ensuring both precision and scalability.

The study also underscores the significance of advanced preprocessing techniques and feature extraction methods in enhancing model performance. By capturing temporal and spectral features of audio signals, the models were able to distinguish subtle patterns between real and fake samples. Furthermore, the integration of these models into an API framework enables real-time detection, ensuring practical applicability across diverse domains.

5.4 Real time application



Sound Based Deepfake Detectors

A security application for social media and web platforms to identify the sound based deepfakes using signal processing infused deep learning framework

Upload an audio sample

Drag and drop file here

Limit 200MB per file • MP3, WAV

recording13532.wav_norm_mono.wav 106.1KB

The audio is Real with a confidence score: [0.70645577]

Sound Based Deepfake Detectors

A security application for social media and web platforms to identify the sound based deepfakes using signal processing infused deep learning framework

Upload an audio sample

Drag and drop file here

Limit 200MB per file • MP3, WAV

recording13014.wav_norm_mono.wav 124.2KB

The audio is Fake with a confidence score: [0.9108318]

Figure 22: Real Time application for the detection of Real or Deep Fake Audio

Link to the deployed code:

In conclusion, this research lays a strong foundation for the development of adaptive and scalable audio deepfake detection systems. While the results demonstrate significant progress, the evolving nature of deepfake technologies calls for continued innovation to stay ahead of potential threats. By bridging the gap between clean and distorted audio detection, this study contributes to the creation of secure digital ecosystems, ensuring trust and reliability in online communication.

Chapter 6: Conclusion and Future Work

This work developed a proper framework that can be used for detecting audio deepfakes, it uses signal processing and deep learning model, for detecting a new threat that is arising in cybersecurity. This is due to the inclusion of features like MFCCs in conjunction with metrics such as jitter and shimmer in conjunction with CNNs, LSTMs and the proposed hybrid model, the system detects fake and real audio with high accuracy. The models were not sensitive to the change between Clean (In-the-Wild) and noisy (Fake-or-Real) data sets; while XGBoost gave higher Precision for Clean data, CNN+LSTM proved its immunity against distorted samples. The real-time implementation through an API further illustrates how this research can be applied in real-world cybersecurity paradigms such as identity theft protection, scams and fake news prevention for online communication platforms.

6.1 Future Work

The future work will focus on improving the generalization for unfair examples across a new set of deepfake generation methods through transfer learning and, indeed, unsupervised techniques. Extension of the multimodal approach in terms of identifying the audio-visual consistency checks, for example, will enhance further defense to the advanced attacks effectively. Increasing sample size with respect to different languages, accents, and noisy environments will enhance the generality of the applicability. Effortless model building will help to expand lightweight to low-resource environments, for example, IoT devices. Thus, this study helps to protect against cybercriminal threats rooted in deepfake technology – phishing, social engineering, and fake news distribution, ensuring the security of digital communication and strengthening the integrity of interactions, as well as protecting individuals, organizations, and governments from AI threats.

References:

Agnew, W., Barnett, J., Chu, A., Hong, R., Feffer, M., Netzorg, R., Jiang, H.H., Awumey, E. and Das, S., 2024. Sound Check: Auditing Audio Datasets. arXiv preprint arXiv:2410.13114.

Bösch, M. and Divon, T., 2024. The sound of disinformation: TikTok, computational propaganda, and the invasion of Ukraine. New Media & Society, 26(9), pp.5081-5106.

Ganga, B., Lata, B.T. and Venugopal, K.R., 2024. Object detection and crowd analysis using deep learning techniques: Comprehensive review and future directions. Neurocomputing, p.127932.

Guo, S., Wang, Y., Zhang, N., Su, Z., Luan, T.H., Tian, Z. and Shen, X., 2024. A Survey on Semantic Communication Networks: Architecture, Security, and Privacy. arXiv preprint arXiv:2405.01221.

Khalid, F., Javed, A., Malik, K.M. and Irtaza, A., 2024. ExplaNET: A Descriptive Framework for Detecting Deepfakes With Interpretable Prototypes. IEEE Transactions on Biometrics, Behavior, and Identity Science.

Kumar, N. and Kundu, A., 2024. Cyber Security Focused Deepfake Detection System Using Big Data. SN Computer Science, 5(6), p.752.

Kumar, N. and Kundu, A., 2024. Cyber Security Focused Deepfake Detection System Using Big Data. SN Computer Science, 5(6), p.752.

Li, K., Lu, X., Akagi, M. and Unoki, M., 2023. Contributions of Jitter and Shimmer in the Voice for Fake Audio Detection. IEEE Access.

Nailwal, S., Singhal, S., Singh, N.T. and Raza, A., 2023, November. Deepfake Detection: A Multi-Algorithmic and Multi-Modal Approach for Robust Detection and Analysis. In 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE) (pp. 1-8). IEEE.

Opdahl, A.L., Tessem, B., Dang-Nguyen, D.T., Motta, E., Setty, V., Throndsen, E., Tverberg, A. and Trattner, C., 2023. Trustworthy journalism through AI. Data & Knowledge Engineering, 146, p.102182.

Raza, M.A., Malik, K.M. and Haq, I.U., 2023. Holisticdfd: Infusing spatiotemporal transformer embeddings for deepfake detection. Information Sciences, 645, p.119352.

Tiwari, A., Dave, R. and Vanamala, M., 2023, April. Leveraging deep learning approaches for deepfake detection: A review. In Proceedings of the 2023 7th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (pp. 12-19).

Triantafyllopoulos, A., Schuller, B.W., İymen, G., Sezgin, M., He, X., Yang, Z., Tzirakis, P., Liu, S., Mertes, S., André, E. and Fu, R., 2023. An overview of affective speech synthesis and conversion in the deep learning era. Proceedings of the IEEE, 111(10), pp.1355-1381.

Yu, X., Wang, Y., Chen, Y., Tao, Z., Xi, D., Song, S. and Niu, S., 2024. Fake Artificial Intelligence Generated Contents (FAIGC): A Survey of Theories, Detection Methods, and Opportunities. arXiv preprint arXiv:2405.00711.

Zhang, G., Gao, M., Li, Q., Zhai, W. and Jeon, G., 2024. Multi-Modal Generative DeepFake Detection via Visual-Language Pretraining with Gate Fusion for Cognitive Computation. Cognitive Computation, pp.1-14.

APPENDIX

1. Appendix – Research Summary

Table 1: Summary of different research papers that is being reviewed

Paper Title	Authors	Dataset Used	Model Used	Result Summary
Leveraging Deep Learning Approaches for Deepfake Detection: A Review	Tiwari, A., Dave, R., Vanamala, M.	Utilizes multiple datasets, such as FaceForensics++	Various deep learning models (e.g., CNN, RNN, GAN)	Provided a comprehensive review of deep learning models and their applications in deepfake detection. Highlighted advantages and challenges in detecting deepfakes.
Fake Artificial Intelligence Generated Contents (FAIGC): A Survey of Theories, Detection Methods, and Opportunities	Yu, X., Wang, Y., Chen, Y., Tao, Z., Xi, D., Song, S., Niu, S.	(text, images, audio, and video) used for training and benchmarking FAIGC detection, including datasets designed for deepfake detection, audio synthesis, and multimodal validation	Various detection methods including CNN, GANs	Provided a detailed survey of fake AI-generated content detection methods, challenges in detection accuracy, and future opportunities.
Multi-Modal Generative	Zhang, G., Gao, M., Li,	Celeb-DF, DeepfakeDetection	Multi-modal models with	Achieved high accuracy in detecting deepfakes by combining visual and textual

DeepFake Detection via Visual-Language Pretraining with Gate Fusion for Cognitive Computation	Q., Zhai, W., Jeon, G.		visual-language pretraining	features using gate fusion, outperforming traditional models.
The Sound of Disinformation: TikTok, Computational Propaganda, and the Invasion of Ukraine	Bösch, M., Divon, T.	TikTok videos, social media data	Propaganda analysis techniques	Semi-automated, scalable disinformation thrives, leveraging TikTok's audio elements
Sound Check: Auditing Audio Datasets	Agnew, W., Barnett, J., Chu, A., Hong, R., Feffer, M., Netzorg, R., Jiang, H.H., Awumey, E., Das, S.	Various audio datasets (e.g., LibriSpeech, AudioSet)	Auditing algorithms for audio datasets	Presented an auditing framework to ensure the quality of audio datasets, emphasizing dataset reliability and ethical concerns in machine learning tasks.
Deepfake Detection: A Multi-Algorithmic and Multi-Modal Approach for Robust Detection and Analysis	Nailwal, S., Singhal, S., Singh, N.T., Raza, A.	FaceForensics++, Celeb-DF	Multi-algorithmic model combining CNNs, RNNs	Developed a robust deepfake detection model integrating multiple algorithms and modalities, achieving significant improvement in detection accuracy.
An Overview of Affective Speech Synthesis and Conversion in the Deep Learning Era	Triantafyllopoulos, A., Schuller, B.W., İymen, G., Sezgin, M., He, X., Yang, Z., Tzirakis, P., Liu, S., Mertes, S., André, E., Fu, R.	Various speech synthesis datasets (e.g., VCTK Corpus)	Deep learning-based speech synthesis models (e.g., Tacotron, WaveNet)	Provided an extensive review of speech synthesis models and discussed how deep learning is revolutionizing affective speech generation.
Trustworthy Journalism through AI	Opdahl, A.L., Tessem, B., Dang-Nguyen, D.T., Motta, E., Setty, V., Throndsen, E., Tverberg, A., Trattner, C.	N/A	AI-driven journalism models	Explored the use of AI in journalism, focusing on enhancing trust and mitigating misinformation through automated fact-checking and content verification models.
A Survey on Semantic Communication Networks: Architecture, Security, and Privacy	Guo, S., Wang, Y., Zhang, N., Su, Z., Luan, T.H., Tian, Z., Shen, X.	N/A	Semantic communication models	Surveyed semantic communication networks, highlighting their architectural advancements and the security challenges posed by AI-driven communication systems.
Object Detection and Crowd Analysis using Deep Learning Techniques: Comprehensive Review and Future Directions	Ganga, B., Lata, B.T., Venugopal, K.R.	COCO, PASCAL VOC	Object detection models (e.g., Faster R-CNN, YOLO)	Reviewed the advancements in object detection and crowd analysis, emphasizing deep learning models and future research directions in improving real-time performance and accuracy.

ExplaNET: A Descriptive Framework for Detecting Deepfakes With Interpretable Prototypes	Khalid et al. (2024)	Custom Deepfake Dataset	Interpretable Prototype Network	Achieved high interpretability, allowing human validation alongside accuracy improvements in detecting altered facial features
Contributions of Jitter and Shimmer in the Voice for Fake Audio Detection	Li et al. (2023)	Audio deepfake datasets	Jitter and Shimmer-based Feature Extraction	Improved detection of synthetic audio by leveraging nuanced voice features like jitter and shimmer in synthetic voices
Cyber Security Focused Deepfake Detection System Using Big Data	Kumar & Kundu (2024)	Social media dataset	Big Data ML Framework	Enabled scalable, real-time detection across large social platforms, optimized for rapid processing for cyber defense
HolisticDFD: Infusing Spatiotemporal Transformer Embeddings for Deepfake Detection	Raza et al. (2023)	Video datasets (e.g., FaceForensics	Spatiotemporal Transformer	Enhanced accuracy by capturing both spatial and temporal patterns across video frames for more robust deepfake detection
Deepfake Detection: Enhancing Performance with Spatiotemporal Texture and Deep Learning Feature Fusion	Almestekawy, Zayed, & Taha	Video deepfake datasets	Spatiotemporal Texture Extraction and CNN Feature Fusion	Achieved superior performance by combining temporal texture changes with CNN-derived spatial features, providing a comprehensive approach for detecting subtle manipulations in videos