Efficiently Identifying Fake Audio And Images Using Transfer Learning

Boovaneswari S[1], Palanivel N [2] , Divya A [3], Harisri A[4]

[1]Assistant Professor, Department of Computer Science and Engineering
[2] Professor, Department of CSE(IoT & Cyber Security including Block chain Technology)
[3,4] UG Scholar, Department of Computer Science and Engineering
[1,2,3,4] Manakula Vinayagar Institute of Technology, Puducherry, India
[1]boovana2825@gmail.com, [2]npalani76@gmail.com

ABSTRACT:
With the advent of AI-generated content in today's digital world, identifying forged images and audio has become more difficult. The system suggested here utilizes deep learning to enhance the precision and dependability of detecting fake media. A strong Convolutional Neural Network (CNN) known as VGG19 is utilized to label fake images. VGG19 uses deep features toanalyze images and detect discrepancies in textures, illumination, and pixel patterns that signal manipulation. The fine-tuned pre-trained VGG19 model with a large database of real and fake images improves prediction accuracy. For detecting fake audio, the system employs a Recurrent Neural Network (RNN), which is best suited for processing sequential data. By analyzing spectrogram features and waveform patterns, the RNN model detects anomalies in pitch, tone, and frequency typical of AI-generated or manipulated audio. The model is also trained on synthetic and real speech datasets to identify authentic and deepfake audio successfully. Through the combination of VGG19 for image detection and RNN for audio classification, the suggested system offers a powerful method for fake multimedia content detection. The proposed solution improves security, digital forensics, and misinformation avoidance, providing more trustworthy authentication of visual and audio data in real-world applications.

Keywords: Fake Image Detection, Fake Audio Detection, Deep Learning, VGG19, Recurrent Neural Network (RNN), Convolutional Neural Network (CNN).

I. INTRODUCTION

Identification of forged images and audio has become a major issue in the modern digital age with the fast growth of artificial intelligence and deepfake technology. Images can be forged with advanced editing software or AI-based models, and it becomes hard to differentiate between original and manipulated images.[16]Likewise, forged audio can duplicate real voices with great accuracy, and it becomes easy to create conversations or statements that never existed. As deepfake technology continues to advance, there is a higher potential for misinformation and fraud, which necessitates the creation of efficient detection techniques. Deepfakes can be manipulated by tweaking facial expressions, combining disparate parts, or altering backgrounds. AI-generate images can be convincing but tend to have minute disparities, which include unnatural lighting, warped textures, or inconsistent reflections. These minute particulars can be used to determine if images are manipulated. Also, some forged images do not have adequate depth perception or show abnormal patterns when viewed up close, which are typical indicators of digital manipulation. Deepfake audio is an even bigger challenge, as voice synthesis powered by artificial intelligence can mimic human speech with high accuracy. Manipulated audio can be detected through unnatural pauses, robotic speech patterns, and inconsistencies in pronunciation and emotions. Deepfake voice can be used to mimic people and result in identity fraud, disinformation, and financial fraud. Identifying deepfake audio involves speech patterns analysis, background noise inconsistencies,

and unnatural voice transitions in the recordings. In the fight against tampered images and audio, detection systems use sophisticated analytical methods such as pattern detection and signal analysis. [8]The systems scan media for evidence of digital tampering through the detection of inconsistencies in pixel patterns, frequency distortions, or abnormal speech cadences. Forensic instruments are also used by experts to track the sources of images and audio clips to ensure their legitimacy. Such detection techniques are paramount in detecting deepfake content and avoiding the proliferation of misinformation. As AI-generated media continues to gain popularity, awareness of the dangers of manipulated images and sound is necessary. Organizations, media outlets, and cybersecurity companies should have more rigorous verification procedures in place to prove authenticity. Consumers also need to be taught to identify manipulated information and authenticate sources prior to posting. With the advancement of technology, there will be a need for ongoing development in detection techniques to ensure digital integrity and continue having trust in online communication.

The systems scan media for evidence of digital tampering through the detection of inconsistencies in pixel patterns, frequency distortions, or abnormal speech cadences. Forensic instruments are also used by experts to track the sources of images and audio clips to ensure their legitimacy. Such detection techniques are paramount in detecting deepfake content and avoiding the proliferation of misinformation. [12]As AI-generated media continues to gain popularity, awareness of the dangers of manipulated images and sound is necessary. Organizations, media outlets, and cybersecurity companies should have more rigorous verification procedures in place to prove authenticity. Consumers also need to be taught to identify manipulated information and authenticate sources prior to posting. With the advancement of technology, there will be a need for ongoing development in detection techniques to ensure digital integrity and continue having trust in online communication.

a.VGG19: VGG19[14] is a deep convolutional neural network (CNN) proposed by the Visual Geometry Group (VGG) at Oxford University. It has 19 layers, of which 16 are convolutional layers, 3 are fully connected layers, and one is a softmax layer for classification. The architecture of the model is uniform, using small 3x3 convolutional filters used in sequence, which aids in the efficient capture of image features in detail. This well-structured framework facilitates hierarchical feature extraction, which makes VGG19 very efficient for applications like image classification, object detection, and feature extraction. One of the main strengths of VGG19 is that it can learn complex patterns from images because of its deep layers. The early layers extract simple features such as edges and textures, whereas the deeper layers identify more complex structures like shapes and objects. This top-down methodology makes VGG19 especially beneficial for applications such as medical imaging, face recognition, and style transfer in paintings. A major disadvantage of VGG19 is that it has high computational requirements with large memory and processing needs, owing to the high number of parameters. Though newer models such as ResNet and EfficientNet are more efficient, VGG19 continues to be a standard in deep learning. Pre-trained versions within frameworks such as TensorFlow and PyTorch can be adapted for particular uses, making it a useful tool in computer vision research.

b.Recurrent Neural Networks(RNNs): Recurrent Neural Networks (RNNs) are very important in the area of audio processing based on their capacity to process and analyze sequential data efficiently. Unlike regular feedforward neural networks, which handle input separately without regard to past states, RNNs possess a special feedback mechanism. This feedback mechanism allows them to store and use information from previous time steps, and they are therefore very well-suited for sequential pattern tasks, like audio signal processing. Because audio signals are temporal in nature, i.e., their current state is a function of past and future contexts, RNNs are a great fit for speech and sound analysis tasks. One of the major benefits of RNNs in audio processing is their ability to learn temporal dependencies within an audio sequence.This means they can recognize and model patterns over time, so they are well-suited to the analysis of speech rhythms, musical sequences, and other ambient sounds. In speech recognition, for instance, an RNN can discern how the phonemes and words are related over time, enhancing transcription accuracy. Likewise, in music composition, these networks are able to learn musical patterns and make predictions of the next note based on what has been heard so far. Working with raw audio waveforms or processed features like spectrograms and Mel-frequency cepstral coefficients (MFCCs), RNNs recognize differences in sound at various time intervals and enable better classification, prediction, and generation of audio signals. Powerful as they are, regular RNNs are plagued by a serious issue called the vanishing gradient problem. This is a problem where the network fails to hold data for long sequences and hence learn long-range dependencies.Consequently, standard RNNs could perform poorly under situations involving complex audio signals demanding memory over very long periods of

time. In response to this weakness, there has been development of more complex architectures like the Long Short-Term Memory (LSTM) networks and the Gated Recurrent Units (GRUs). These improved RNN variants use specialized gating mechanisms that enable them to selectively retain and forget information when necessary, greatly enhancing their capacity to model long-term dependencies. This improvement has resulted in increased accuracy and efficiency in areas like automatic speech recognition (ASR), real-time voice synthesis, and music generation. Generally, RNNs and their variants are critical components of contemporary audio processing applications. They are used extensively in voice-operated assistants such as Siri and Alexa, language translation programs that translate words into other languages, and even audio deepfake detection, where AI is utilized to detect tampered or artificial voices. With ongoing developments in deep learning, the combination of RNNs with other models, including Transformers, is further increasing their scope, allowing even more advanced and intelligent audio processing applications in the future.

II. LITERATURE SURVEY

[1] V. V. V. N. S. Vamsi, S. S. Shet, S. S. M. Reddy, S. S. Rose, S. R. Shetty, S. Sathvika, and S. M. S.

This paper proposes a new method for detecting Deepfake videos using the synergistic power of two state-of-the-art deep learning models, namely ResNeXt, an extremely computationally efficient Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) networks. The synergy between these two models increases the system's potential to detect even the slightest modifications in fake videos, which are rapidly becoming very realistic with artificial intelligence developments.ResNext is the key component of this system because it is dedicated to extracting spatial features from each video frame.Being a CNN model, it is particularly tailored to detect subtle details in an image so that it can examine fine-grained visual patterns and abnormalities that could signal manipulation. At the same time, LSTM networks support this procedure by exceling at examining sequential data. While traditional CNNs work on images in isolation, LSTM networks are able to follow changes over time by detecting temporal inconsistencies between adjacent video frames. By integrating these two methods, the detection model can perform a more comprehensive analysis of both spatial and temporal attributes, greatly enhancing the accuracy of Deepfake detection. The integration of ResNext and LSTM enables the system to detect even subtle inconsistencies in doctored videos—consistencies that may not be perceivable to the human eye or to conventional detection models. This improved ability is especially useful as Deepfake technology keeps advancing, creating more realistic fake videos that are hard to tell apart from authentic footage.By combining spatial feature extraction with temporal pattern recognition, this approach improves the overall detection system, making it more robust against advanced Deepfake attacks. Finally, this study adds to the battle against online misinformation by enhancing the efficiency of Deepfake detection systems. As media generated using artificial intelligence improves, creative methods like the combination of ResNext and LSTM will be essential in protecting digital authenticity and avoiding AI-generated content misuse in social media, news coverage, and cybersecurity. [2] Fakhar Abbas , Araz Taeihagh emphasize the need to check the authenticity of online content at a time when disinformation propagates quickly. The exponential proliferation of digital media has made it important for both the individual and society to ensure information credibility. Although Deepfake technology has been around for quite some time now, its spreading and growing access raise serious risks, from misdirecting people to shaping political situations and even national security threats. The prospect of manipulating sounds, pictures, and videos through artificial intelligence makes such forgeries detectable challenging, and measures have to be created to deal with this menace effectively.This research undertakes a thorough examination of the main methods, frameworks, algorithms, and tools aimed at detecting and inhibiting the spread of Deepfakes. Using state-of-the-art machine learning and deep learning models, the study investigates the role of AI-powered technologies in influencing the spread of disinformation on the internet. It examines how emerging artificial intelligence techniques help both generate and detect Deepfakes, identifying the double-edged sword role of AI in advancing and countering digital deception.One of the main emphases of the research is determining the efficacy of current Deepfake detection frameworks in actual use cases, such as social media, journalism, and cybersecurity. The study analyzes how efficiently these approaches function in real-world scenarios, assessing their dependability in detecting altered content in various digital environments. Furthermore, the research enlightens us on the Deepfake detection challenges, including

adversarial AI methods that are constantly enhancing the authenticity of artificial media, thereby rendering detection more challenging.In addition, the research addresses new trends and urgent matters concerning the regulation and governance of Deepfake technology. With increasing sophistication in AI-created media, there is an urgent need for holistic policies and regulatory measures to avoid its misuse. The research highlights the need to create legal and ethical frameworks to prevent the abuse of Deepfakes while ensuring that AI innovation continues to benefit society.Overall, the study offers an in-depth analysis of the changing landscape of Deepfake detection, evaluating the effectiveness of existing methodologies and pinpointing future challenges in the fight against digital disinformation. Addressing both technical and regulatory dimensions, the study offers insightful contributions to the ongoing efforts to protect information integrity in the digital era

[3] Ewout Nas, Roy de Kleijninvestigate the challenging task of identifying Deepfake videos, very realistic, AI-based manipulations that closely approximate real footage. With advancing artificial intelligence and video synthesis technologies, it has become harder to separate original and edited videos. Although a lot of previous work has concentrated on coming up with algorithmic approaches for detecting Deepfakes, less is known regarding human capacity in distinguishing these artificially manipulated videos. The current work tries to narrow this gap through exploring the determining factors of human performance in recognizing Deepfakes. The results of this study show that people are better able to identify Deepfake videos when they contain recognizable faces, i.e., celebrities, politicians, or prominent public figures.This implies that recognition memory and prior exposure have a significant impact in identifying digital forgery. But when people are shown Deepfakes of unknown individuals, they are no longer able to identify manipulation. This emphasizes the difficulty of recognizing AI-generated changes to content featuring unknown or lesser-known people. Moreover, the study presents a significant association between social media usage and Deepfake detection abilities. People who spend a lot of time consuming online media, especially video-oriented websites, are more likely to have an enhanced sense of recognition for manipulated media, potentially because they have been exposed more to both original and manipulated videos.Interestingly, the study also discovers that individuals who have a history of conspiracy belief exhibit a higher capacity to identify Deepfakes. This may be due to their increased cynicism and tendency to question visual information more aggressively. Conversely, the research reveals no statistically significant correlation between demographics like age or gender and accurate Deepfake detection. This means that the potential to identify forged videos is not necessarily tied to biological or demographic traits but would be more accurately explained by cognitive, experiential, and psychological factors. These results highlight the significance of familiarity and media experience in human Deepfake detection. They also imply that more work is required to investigate other cognitive and perceptual factors that might play a role in recognition accuracy. It is critical to comprehend the human dimension of Deepfake detection in order to enhance media literacy, as well as in order to build methods for augmenting public knowledge and resistance to digital disinformation in the context of rapidly developing AI-generated media.

[4]Mouna Rabhi , Spiridon Bakiras , Roberto DiPietro investigates increasing importance of audio as a means to verify biometrically, how it uses largely across identity authentication and security solutions. Artificial intelligence advancement has led to the emergence of various AI-based classifiers that have been used to authenticate human speech so that voice-based authentication systems are reliable and accurate. But as much as these models have advanced, their susceptibility to Deepfake audio attacks has not been extensively researched. Our research identifies a key flaw in current state-of-the-art Deepfake audio defense systems under adversarial attacks, casting doubts on the security of voice authentication technology. The research concentrates on the Deep4SNet model, an extremely sophisticated AI-powered Deepfake audio detector with an initial astounding accuracy of 98.5% in detecting synthetic voice samples.Researchers used two adversarial attacks, based on a Generative Adversarial Network (GAN), to test the vulnerability of the model. The attacks proved the disconcerting ease with which an effective system could be attacked, bringing its detection rate down to almost 0%. In a gray-box attack scenario—where the attacker possesses incomplete information about the inner workings of the model—the accuracy of Deep4SNet fell dramatically from 88.5% to a very low 0.08%. This significant fall in performance highlights the vulnerabilities of current Deepfake audio classifiers to adversarial attacks, demonstrating that even state-of-the-art detection systems can be very easily fooled. To reverse these weaknesses, the study offers a novel, light-weight, and flexible defense system. In contrast to hard

security solutions, the suggested approach can be easily incorporated into any Deepfake audio detection system, and hence it is a versatile tool for increasing the resilience of voice authentication systems.The research also proposes some strategies to strengthen Deepfake detection models against future adversarial attacks so that AI-driven classifiers can continue to remain reliable in high-stakes applications like banking, law enforcement, and cybersecurity. In the end, this study highlights the paramount importance of creating more robust and secure Deepfake audio detection processes. As AI adversarial techniques keep on evolving, conventional security mechanisms might lose relevance, prompting ongoing innovation in biometric verification. Through the determination of key vulnerabilities and the suggestion of countermeasures, this research brings meaningful inputs into voice authentication technology protection against new threats in the AI scenario.

[5]Michael Hameleers, ToniG.L.A. vanderMeer, Tom Dobber examine the increasing alarm around political Deepfakes and how they may affect democratic processes. Although everyone seems to agree that Deepfake videos are a huge threat to political stability and public confidence, very little is understood about the ways in which they affect people's perceptions of credibility and political legitimacy. This research aims at filling the knowledge gap by providing a comparison between the Netherlands and the United States, with a sample of 1,187. The main goal of this research is to investigate public perception influenced by varying degrees of deception in Deepfake videos and whether people can distinguish effectively between original and manipulated political content. The results of the research demonstrate that very realistic Deepfakes—those with slight and credible changes—are regarded as more authentic by the general public than Deepfakes featuring overt or implausible changes.This indicates that as political manipulation through AI is developed further, it may become capable of misleading a larger section of the audience and shaping public opinion. Surprisingly, the study also reveals an unintended side effect: even if a Deepfake is not completely believable, it has the power to cause reputational harm to the targeted political figure. Deepfakes with over-the-top or implausible edits have the tendency to erode the credibility of the subject in the video, even if viewers don't totally believe the message to be true. This underscores the wider implications of Deepfake disinformation, since even obviously constructed videos can undermine public confidence in political leaders. A further key result of the research is the connection between cognitive reflection and vulnerability to Deepfake deception. Those who performed better on cognitive reflection tests—reflecting a higher propensity for analytical and reflective thinking—were less likely to find Deepfakes believable.This implies that people who undergo more profound cognitive processing are less susceptible to manipulation by AI-created misinformation. Consequently, the study emphasizes the need to develop critical thinking and media literacy to counteract the impact of Deepfakes in political discourse. In general, this research offers useful insights into the psychological and social impacts of political Deepfakes. Although such faked videos will not always succeed in deceiving the public entirely, they still have the capability to influence political narratives and reputations. The results stress the importance of programs that cultivate critical thinking and digital literacy skills to enable individuals to identify and critically evaluate false media content.As Deepfake technology develops, it will be important to understand how it shapes public opinion in order to defend democratic integrity and prevent misinformation from eroding political legitimacy.

III. PROPOSED SYSTEM

The system proposed employs deep learning to improve the reliability and accuracy of the detection of deepfake media using two unique models: VGG19 for image detection and a Recurrent Neural Network (RNN) for audio classification. VGG19 is a strong Convolutional Neural Network (CNN) that identifies tampered images by analyzing deep features such as texture flaws, lighting flaws, and abnormal patterns of pixels.[14]The model is trained on a vast dataset of original and manipulated images and tuned to optimize its capability to identify fakes. In identifying fakes in audio, the system employs an RNN, which is suitable for processing data sequentially. Through the learning of spectrogram features and waveform patterns, the RNN identifies pitch, tone, and frequency changes distinguishing deepfake audio from natural speech. The RNN is trained on real speech and artificially created speech sets, enabling the RNN to effectively identify synthetic variations.[16] The combination of VGG19 in image analysis and RNN for audio classification provides an integrated and strong system in deepfake detection, improving security and preventing the dissemination of misinformation. The hybrid approach ensures improved

verification of multimedia content, hence highly relevant in digital forensics, social media, journalism, and law enforcement, where the detection of tampered media is critical.

A. Architecture of Proposed System

The architecture diagram of the proposed system shows how audio and image analysis modules are combined to support deep fake detection. At its center, the architecture has two main modules: one for processing audio and one for processing images.[15]The audio module uses transfer learning by using a pre-trained neural network to learn features from audio samples. The network inspects different dimensions of the audio, including frequency patterns and timing changes, which helps it recognize small changes characteristic of deep fakes. Simultaneously, the image processing module uses the same process with a pre-trained convolutional neural network (CNN) to scan visual content.

The network scans tiny details and patterns in images, which helps recognize changes not easily noticeable. Both modules end with a decision module that combines the results of audio and image analysis, building on their strengths for better detection. [7]The architecture also includes a user interface that supports the user during interaction with the system, allowing users to upload media to be analyzed and get real-time feedback on content authenticity. Overall, this architecture diagram shows exactly how the proposed system combines cutting-edge detection mechanisms with improved user experience, giving an integrated and holistic solution to counter the growing threatof deep fakes in digital media.

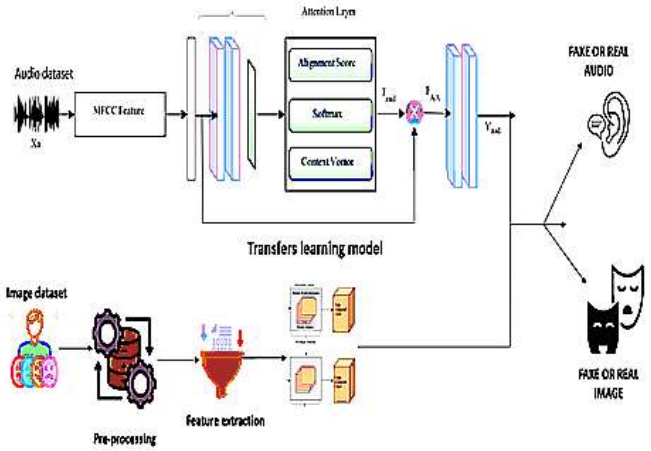

Fig 1: Architecture diagram

B. Data Collection:

The first step towards developing an effective deepfake detection system is the acquisition of quality datasets. Kaggle, a widely used open-source dataset platform, holds large libraries of real and fake images and synthetic and real audio samples.[10]These datasets are manipulated visual and audio content, hence the most ideal for deep learning model training. Image datasets are usually made up of deepfake face fake images, and photos that have been altered, and audio datasets consist of AI-generated speech, copied voices, and real human recordings. With the acquisition of data from Kaggle, researchers are able to work with varied samples, hence ensuring the system is trained on different deepfake techniques. Proper dataset selection enhances the generalization capacity of the model, hence making the system more robust to varied manipulation techniques.

C. Pre-processing of Image and Audio Using MFCC:

Prior to training, the data gathered is pre-processed in order to deliver optimal model performance. For images, pre-processing involves resizing, normalization, and data augmentation in order to enhance feature extraction.[23]The images are resized to have a consistent size and normalized for uniform pixel value standardization. For audio data, Mel-Frequency Cepstral Coefficients (MFCC) is employed to capture meaningful frequency-related features. MFCC transforms raw waveforms to a

spectral domain, facilitating easy analysis of the model for changes in pitch, tone, and timbre. Background noise filtering and silence stripping enhance detection quality.

D.Feature Extraction of Image and Audio:
Feature extraction is critical in the detection of patterns in deepfakes. In images, deep learning models detect texture, edge, and pixel distribution anomaly information that signifies manipulation. In images, hierarchical spatial feature extraction using CNN-based VGG19 detects inconsistencies in lighting and morphing of the face.[24]In audio, feature extraction using MFCCs generates a compressed form of the sound that assists in the detection of subtle variations between real and deepfake speech. Other techniques such as Chroma features and spectral contrast bridge the gap in detecting AI-generated speech.

E.Model Creation Using VGG19 for Image and RNN for Audio:
The system uses VGG19 for image and Recurrent Neural Networks (RNN) for audio analysis. [27]VGG19 is a deep CNN model that analyzes image features in many layers, learning to identify real and fake images based on spatial patterns. It is augmented with a labeled dataset to improve its accuracy. For audio, an RNN-based model is trained to identify deepfake speech by analyzing time-based patterns in voice recordings. This double-model strategy makes the system capable of analyzing both images and audio.

F.Test Data:
Following the training phase, the models are subjected to a rigorous test procedure based on a specially created test dataset made up of unseen real and manipulated media. Through this method, an objective measurement of the model's performance is achieved by testing its capacity to generalize from the training data.

The test data is meticulously selected to contain varied instances of manipulated media, including deepfake videos, AI-generated images, and voice-cloned audio. By testing the model on entirely new data, this step gives a true measure of its detection ability.

The performance of the model is measured based on the following important metrics:
Overall Accuracy: Measures how well the model distinguishes between real and fake media by computing the ratio of correctly labeled instances.
Precision: Assesses the performance of the model in detecting the false content with fewer false alarms. It is calculated as the ratio of correctly classified fake samples to all predicted fake cases.
Recall: Assesses the capability of the model to detect all the manipulated media with fewer missed cases. It is calculated by the ratio of correctly classified fake instances to all actual fake cases.
F1-Score: Offers a balanced assessment by averaging precision and recall. Being the harmonic mean of the two, it presents one complete measurement of the overall performance of the model.

G.Prediction of the Data:
In the last step of the Deepfake detection algorithm, the trained models process new inputs and determine them to be real or fake. This is a decision made through sophisticated artificial intelligence methods that study important features of the input data. In the case of image-based detection, the VGG19 model takes center stage by studying pixel distributions, spatial anomalies, and other visual features. By detecting slight irregularities that are frequently imperceptible to the human eye, VGG19 determines whether an image is manipulated or artificially created. With the use of audio-based Deepfakes, Recurrent Neural Networks (RNNs) are used to parse several auditory properties, including tone, pitch inconsistencies, and patterns in frequency. The models specifically work to detect unnatural variations and irregularities within speech, often revealing synthetic editing. Through examination of these minutiae, RNNs efficiently distinguish between actual human speech and Deepfake speech producedbycomputers.The system returns probability scores in addition to its classifications, giving its users a numeric measure of belief in the result of the classifications. This mechanism of scoring proves useful as it enables individuals and organizations to check the probability of manipulation instead of using mere binary outputs.These models' predictability is exceptionally useful for myriad applications such as law enforcement inquiries, media validation, and electronic forensics. In law enforcement,

Deepfake detection facilitates the validation of digital evidence and detection of misleading content. For media outlets and journalists, it offers a valuable resource for fact-checking and muzzling the dissemination of misinformation. Also, in cyber defense and digital forensics, such AI-based methods improve the credibility of multimedia verification and protects citizens and institutions from online deception. With Deepfake technology advancing at a rapid pace, incorporating advanced detection models such as VGG19 and RNNs is vital for the upkeep of trust and security in digital content. Their capabilities of inspecting images and audio with great accuracy help to ensure that manipulated content canbe detected and addressed safely.

IV.ACCURACY

Accuracy is one of the key metrics employed to measure the performance of a classification model, reflecting on its general performance.It is determined as the ratio of the correctly identified instances—true positives (TP) and true negatives (TN)—to all the cases considered. This measure indicates the model's capacity to identify positive and negative classifications correctly, providing a general measure of its reliability.

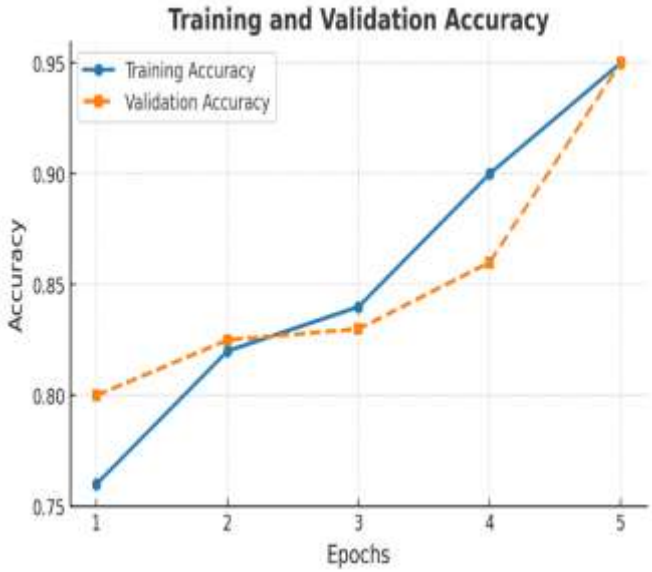

Fig 2: Training and validation accuracy over five epochs

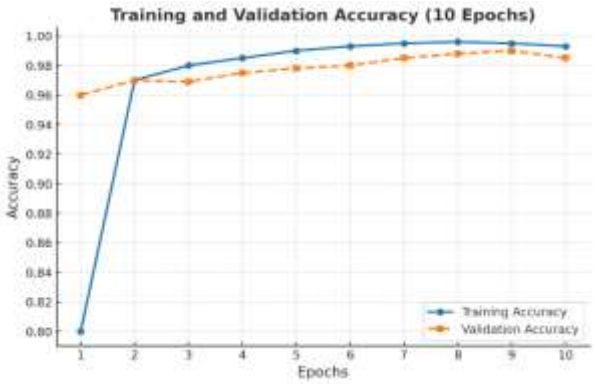

Fig3: Training and Validation accuracy over ten epochs.

Formula,

$$\frac{TP+TN}{}$$

$$\text{Accuracy} = \frac{}{TP+TN+FP+FN}$$

Where:

- TP (True Positives): Examples where the model accurately identifies a positive instance.
- TN (True Negatives): Examples where the model accurately predicts a negative instance.
- FP (False Positives): Instances where the model falsely classifies a negative as positive.
- FN (False Negatives): Instances where the model fails to classify a positive, labelling it as negative.

Although accuracy is a simple measure that gives a global evaluation of how well a model performs on all classes, it can be misleading at times—particularly in situations with imbalanced datasets. In the event where one class largely dominates the other, a model can get very high accuracy just by always predicting the majority class and never actually identifying cases of the minority class..

V.PRECISION

Accuracy, or Positive Predictive Value, assesses the goodness of a classification model in classifying positive instances. It is calculated by taking the number of true positives divided by the number of predicted positives (sum of true positives and false positives)This measure is especially useful in cases where reducing false positives is imperative, as it assists in assessing the reliability of the model's positive predictions.

Formula,

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positive s(FP)}}$$

Explanation,

- True Positives (TP): Examples where the model is right in labeling a case as positive.
- False Positives (FP): Examples where the model is wrong in predicting a positive class even when the true class is actually negative.

Accuracy is especially important in situations where false positives have serious ramifications. For instance, in healthcare diagnosis, a false positive for a serious disease might cause unnecessary worry, expensive medical interventions, or inappropriate treatment. Likewise, in spam filtering, a high rate of false positives might flag valuable emails as spam, which users would never get to read.

By emphasizing the accuracy of positive predictions, precision is central to measuring how reliable a system is in detecting relevant cases. In fraud detection, security surveillance, and Deepfake detection applications, high precision is crucial, as incorrect classification of valid transactions or content may result in reputational and operational hazards. Hence, based on the particular use case, precision is usually adjusted in conjunction with other measures such as recall and F1-score to achieve a well-balanced and efficient classification system.

VI.RECALL

Recall, also referred to as Sensitivity or True Positive Rate, is a crucial measure of the assessment of classification models, especially in those domains where detection of positive cases is essential, for

example, medicine and detecting fraud.[12] Recall indicates the rate of true positives to the sum of all true positives. In essence, recall refers to the ability of a model to accurately identify positive cases among all cases that fall under the positive class. This makes it extremely relevant in cases where missing a positive case might have severe consequences, for instance, missing a medical condition in a patient.

Formula,

$$\text{Recall} = \frac{TP}{TP+FN}$$

Here, True Positives (TP) would mean that the model has accurately identified a positive instance, and False Negatives (FN) would be instances of actual positives that the model is unable to identify.

This measure is of utmost importance where failure to pick up a positive case may prove disastrous. As an example, in disease diagnosis, a poor recall rate could result in seriously ill patients going undiagnosed, holding up vital interventions. Likewise, in detecting fraudulent activities, an unsatisfactory recall model can miss flagging frauds and thus cause untold financial loses.

Whereas accuracy concentrates on the number of predicted positive examples that are indeed correct, recall is concerned with the percentage of true positive instances that a model classifies correctly. In cases where missing a positive instance is riskier than falsely identifying a negative one, recall is the desired metric.

VII. F1 SCOPE

The F1 Score is an important performance metric for classification algorithms, especially when there is an imbalance of class distributions or when precision and recall are equally important. It is defined as the harmonic mean of recall—the ratio of correctly classified positive instances over all true positive instances—and precision—the ratio of correctly classified positive instances over all positively predicted instances.
Formula,

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

By merging two critical metrics into a single number, the F1 Score offers an even assessment of a model's accuracy at identifying positive cases while keeping false negatives to a minimum. This is especially useful in situations where both false negatives (failed positive cases) and false positives (erroneously identified positive cases) carry high costs.

For example,
In disease diagnosis, a highly precise but lowly recalled model could accurately identify several instances of an illness while leaving many true patients undetected. A high-recall and low-precision model could similarly mislabel well people as sick, causing undue treatments. F1 Score eliminates these problems because it ensures the model is both sensitive to positive cases and uniform in its decision-making.

Likewise, in fraud detection, where there are far more legitimate transactions than fraudulent ones, a model can score high accuracy by simply labeling most transactions as legitimate. But with no good F1 Score, it might miss real fraudulent activities. By trading off precision and recall, the F1 Score ensures robust detection in such skewed datasets.

LOSS:

The loss function is very important in deep learning models because it measures the discrepancy between the predicted and the actual target values, directing the optimization. In our framework that uses VGG19

to identify fake images and a CNN-RNN model to identify fake audio, dedicated loss functions are used to maximize learning and precision.

For image analysis using VGG19, Categorical Cross-Entropy Loss is applied since it is appropriate for classification problems with clear categories, e.g., real vs. fake images. This loss function is penalized based on the incorrect predictions by matching the model's predicted probability distribution against the true class labels to promote accurate classification.

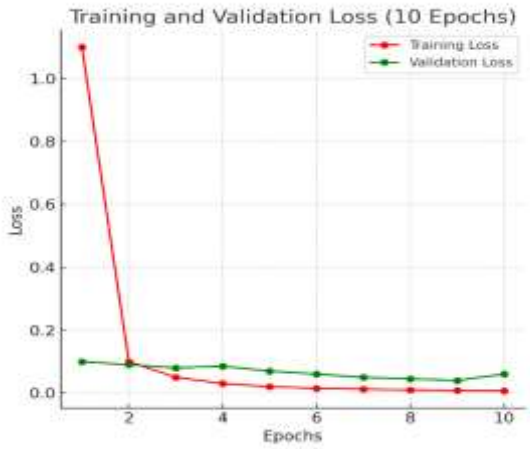

Fig 4:This chart shows training and validation loss over Ten epochs

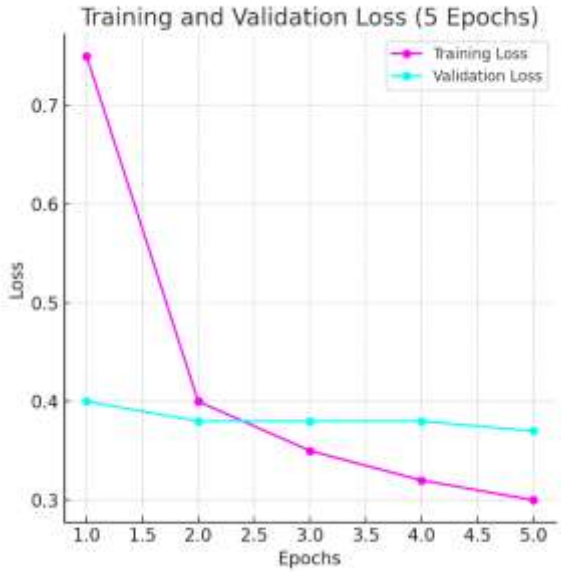

Fig 5:This chart shows tring and validation loss over five epochs

VIII.PREDICTION

Prediction in RNN is viewing new input data, possibly an audio/video sequence and processing it step by step in the network. Having acquired patterns from training data, the model uses this for evaluating the input. [17]The RNN considers temporal relationships between frames or audio signals. Prediction is made based on the entire sequence. It's going to output some sort of classification like "Real" or "Fake", which is now a choice, made by the features it has learned from this input sequence, for those deep fake detection, or any such task, so in that sense, the prediction is an indication of the model's capacity to generalize from the training data to enable it to detect the correct manipulation in new data. Formula,

$LOSS_{audio} = -1/M \sum_{i=1}^{M} [yi \log(y^i) + (1 - yi) \log(1-y^i)]$

Where:

- M represents the total number of audio samples.
- yi denotes the true label for each audio sample.
- θi indicates the predicted probability of the audio being fake.

$LOSS_{image} = - \sum_{i=1}^{N} yi \log (y^i)$

Where:

- yi is the actual label (1 for fake, 0 for real).
- θi is the predicted probability of the image being fake.
- N is the number of classes (in this case, 2).

For the audio analysis performed with the CNN-RNN hybrid model, *Binary Cross-Entropy Loss* is used because it is appropriate for binary classification problems like the difference between real and fake audio. This loss function measures how accurate the predictions are by checking the distance between the predicted probability of an audio file being fake and its ground truth, rewarding no difference from the ground truth. Through incorporating both spatial and temporal properties of the audio data, the loss function promotes a holistic learning process for detecting deepfake patterns. For optimization with balance, the system fuses both image and audio loss functions, generally by calculating a weighted combination of the losses. This makes it possible for the model to learn from both modalities effectively, guaranteeing effective deepfake detection in both audio and visual media.

IX.COMPARISON GRAPH:

A.IMAGE COMPARISON GRAPH:

The bar chart shows a comparative study of the performance of four machine learning models—Random Forest (RF), Convolutional Neural Network (CNN), Support Vector Machine (SVM), and VGG19—tested on the basis of accuracy, precision, recall, and F1-score. Out of these models, VGG19 shows better performance on all the evaluation parameters and is thus the best model for image processing and manipulation detection.

VGG19, a deep neural network, performs better in feature extraction because of its multi-layer structure. Being deep, it is capable of picking up detailed patterns and finding slight inconsistencies in pixel distributions, lighting, texture, and spatial arrangement—features of utmost importance for detecting manipulated or Deepfake images. The findings show that VGG19 attains a high accuracy of around 98%, with excellent precision and recall, reducing both false positives (genuine images incorrectly classified as manipulated) and false negatives (manipulated images not identified). Precision is important in curbing false alarms, and recall in ensuring manipulated images are well detected. Also, the good F1-score indicates a best balance between precision and recall, which makes VGG19 the most accurate model among the models tested. While Random Forest (RF) and Support Vector Machine (SVM) are popular for classification problems, they are not as effective in image-based applications since they do not have spatial feature extraction ability.Deep learning networks such as VGG19 are naturally more suitable for such work, as they inspect pixel relations, edges, and textures instead of depending on hand-engineered features. A regular deep neural network is also effective but lags behind the depth, efficiency, and pre-trained structure of VGG19, which further strengthens its hierarchical feature extraction capability. In summary, the findings confirm that VGG19 is the most effective model for Deepfake detection.

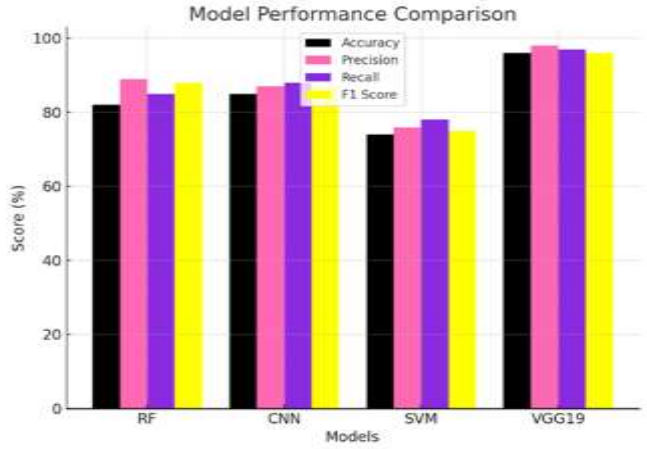

Fig 6: Compares RF, CNN, SVM, and VGG19 modelsusing Accuracy, Precision, Recall, and F1 Score.

By comparison, while CNN performs relatively well at image classification, it is not as good as VGG19 because it has a shallower network and less capability to extract features. While normal CNN models are capable of detecting simple patterns and textures, they do not have the hierarchical feature learning and depth to recognize subtle pixel manipulations that define Deepfakes.Conversely, Random Forest (RF) and Support Vector Machine (SVM), both traditional machine learning models, are hugely challenged by complex image data. The models are based on hand-crafted features as opposed to automatically learning hierarchical representations from raw images and hence cannot be used for detecting manipulated images. Of them, SVM is the poorest, with the least accuracy, precision, recall, and F1-score, highlighting its inability to detect Deepfake images.The deep performance difference among VGG19 and the others signifies the vital role of deep learning architectures today in image classification and Deepfakes detection. Deep networks, like VGG19, tap into several layers of convolution for fine-grained information capture of pixel inconsistencies, lighting mismatches, and distortion in texture patterns that are quite often imperceivable to general machine learning architectures. In general, the findings affirm that VGG19 is the best model for identifying fake media using images, providing a strong and effective solution for use in digital forensics, misinformation avoidance, and multimedia verification. Its capacity to conduct deep feature analysis guarantees very accurate identification of forged and synthetic content, rendering it an effective tool for protecting digital media integrity.

B.COMPARISON OF RNN:

The bar chart gives a comparative performance analysis of four machine learning models—Random Forest (RF), Convolutional Neural Network (CNN), Support Vector Machine (SVM), and Recurrent Neural Network (RNN)—in identifying the fake audio. Their performance is measured on the basis of some important metrics, such as accuracy, precision, recall, and F1-score. Out of these models, RNN performs best, showing the highest values in all the evaluation metrics.Among these models,RNN exhibits better performance, with the highest scores for all evaluation metrics.The Recurrent Neural Network (RNN) differs because it has the capacity to handle sequential data, a feature that renders it highly powerful in identifying Deepfake audio. In contrast with conventional models where data is dealt with in instances in isolation, RNNs employ feedback links that enable it to maintain information from past instances of time steps. This capability makes the model capable of discovering minutepitch changes, tonal changes, and frequency fluctuations—characteristic signs of synthetic or fabricated speech. By virtue of its ability to learntemporal dependencies of speech patterns, RNNs can reliably differentiate between real human voices and AI-generated Deepfake sounds with high reliability. Compared to that, CNN and RF models are moderately good, but they are not as effective as RNN.CNNs, conventionally applied for image analysis, can also learn salient features from spectrogram images of sound signals but do not possess the sequential processing ability to capture temporal variations in speech. Likewise,Random Forest (RF), a tree-based ensemble learning algorithm, excels in classification but is not well equipped to monitor continuous speech patterns because of its non-

capability to handle sequential dependencies. The SVM model, although efficient in certain types of classification jobs, is worst among the four because it falters where complex, multi-dimensional audio content is involved. As SVMs are based almost entirely on stationary feature sets, not sequential context, they aren't as competent at identifying gradual distortion and anomaliesthat occur within Deepfake sound. Overall, the results confirm that RNN is the strongest modelfor Deepfake audio detection because it can capture sequential dependenciesand pick up on subtle speech anomalies.

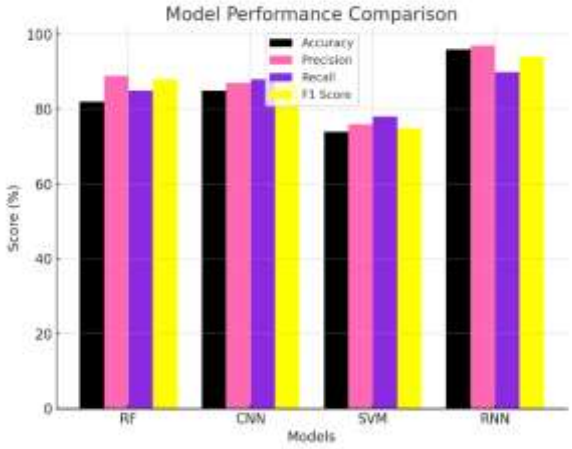

Fig 7:This bar chart presents a performance comparison of RF, CNN, SVM, and RNN models across four key metrics: Accuracy, Precision, Recall, and F1 Score.

X . CONCLUSION

In summary, the system proposed here over successfully overcomes the issues of identifying forged images and audio by combining deep learning methods. Through the use of VGG19 for image analysis, the system successfully identifies tampered visual content through deep feature extraction. Likewise, the RNN-based method for detecting forged audio guarantees accurate detection of anomalies in pitch, tone, and frequency. This two-modality system strengthens the detection of AI-created media, enhancing security, digital forensics, and disinformation prevention. The use of VGG19 for image processing and RNNs for temporal pattern recognition enhances the system's capacity to detect both deepfake audio and fake images. Future developments can involve attention mechanisms for enhanced accuracy, real-time detection, and extension to multi-modal media such as videos. By constantly evolving in response to emerging deepfake processes, this system supports the protection of digital content authenticity, allowing for more trustworthy verification of multimedia data in a world where AI-based media manipulation is growing. Future research will target the integration of attention mechanisms to better extract features, enhancing real-time detection for big-picture applications, as well as extending the system to multi-modal deepfake detection, including video. Moreover, the inclusion of unsupervised learning will assist in identifying previously unknown manipulation methods, further enhancing the system's flexibility and precision.

XI.REFERENCE

[1] Z. A. Baig et al., “Future challenges for smart cities: Cyber-security and digital forensics,” Digit. Investig., vol. 22, pp. 3–13, Sep. 2017.

[2] H. Zimmerman, “The data of you: Regulating private industry’s collection of biometric information,” U. Kan. L. Rev., vol. 66, p. 637, 2017.

[3] A. K. Jain and A. Kumar, “Biometric recognition: an overview,” in Second generation biometrics: The ethical, legal and social context, Springer, 2012, pp. 49–79.

[4] D. Lillis, B. A. Becker, T. O. Sullivan, and M. Scanlon, “Current Challenges and Future Research Areas for Digital Forensic Investigation INVESTIGATION,” no. c, 2016.

[5] D. Ramos-Castro, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, “Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework,” in 2006 IEEE Odyssey-The Speaker and Language Recognition Workshop, 2006, pp. 1–8.

[6] A. Saleema and S. M. Thampi, "Voice Biometrics: The Promising Future of Authentication in the Internet of Things," in Handbook of Research on Cloud and Fog Computing Infrastructures for Data Science, IGI Global, 2018, pp. 360–389.

[7] B. Zawali, R. A. Ikuesan, V. R. Kebande, S. Furnell, and A. A-Dhaqm, "Realising a Push Button Modality for Video-Based Forensics," Infrastructures, vol. 6, no. 4, p. 54, 2021.

[8] Chunlei Peng, Huiqing Guo, Decheng Liu, Nannan Wang, Ruimin Hu, Xinbo Gao. (2023) Deep Fidelity: Perceptual Forgery Fidelity Assessment for Deepfake Detection arXiv:2312.04961v1

[9] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning Self-Consistency for Deepfake Detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[10] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "Wild Deepfake: A Challenging Real-World Dataset for Deepfake Detection," arXiv preprint arXiv:2101.01456v1, 2021.

[11] K. Shiohara and T. Yamasaki, "Detecting Deepfakes with Self-Blended Images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[12] S. Dong, J. Wang, J. Liang, H. Fan, and R. Ji, "Explaining Deepfake Detection by Analyzing Image Matching," arXiv preprint arXiv:2207.09679v1, 2022.

[13] A. Jain, N. Memon, and J. Togelius, "A DatalessFaceSwap Detection Approach Using Synthetic Images," in Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), Institute of Electrical and Electronics Engineers Inc., 2022.

[14] F. M. Salman and S. S. Abu-Naser, "Classification of Real and Fake Human Faces Using Deep Learning," International Journal of Academic Engineering Research (IJAER), vol. 6, no. 3, 2022.

[15] T. Chen, S. Yang, S. Hu, Z. Fang, Y. Fu, X. Wu, and X. Wang, "Masked Conditional Diffusion Model for Enhancing Deepfake Detection," arXiv preprint arXiv:2402.00541, 2024.

[16] A. Luo, E. Li, Y. Liu, X. Kang, and Z. J. Wang, "A Capsule Network-Based Approach for Detection of Audio Spoofing Attacks," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Jun. 2021, pp. 6359–6363.

[17] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," 2019, arXiv:1907.00501.

[18] Z. Wang, S. Cui, X. Kang, W. Sun, and Z. Li, "Densely connected convolutional network for audio spoofing detection," in Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC), Dec. 2020, pp. 1352–1360.

[19] C.-I. Lai, N. Chen, J. Villalba, and N. Dehak, "ASSERT: AntiSpoofing with Squeeze-Excitation and Residual neTworks," 2019, arXiv:1904.01120.

[20] M. Todisco, H. Delgado, K. A. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi, "Integrated presentation attack detection and automatic speaker verification: Common features and Gaussian back-end fusion," in Proc. 19th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), 2018, pp. 77–81.

[21] M. R. Kamble, H. B. Sailor, H. A. Patil, and H. Li, "Advances in antispoofing: From the perspective of ASVspoof challenges," APSIPA Trans. Signal Inf. Process., vol. 9, no. 1, p. e2, 2020.

[22] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[23] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," Comput. Speech Lang., vol. 45, pp. 516–535, Sep. 2017.

[24] Z. Lei, Y. Yang, C. Liu, and J. Ye, "Siamese convolutional neural network using Gaussian probability feature for spoofing speech detection," in Proc. INTERSPEECH, Oct. 2020, pp. 1116–1120.

[25] G. Pamisetty and K. S. R. Murty, "Prosody-TTS: An end-to-end speech synthesis system with prosody control," Circuits, Syst., Signal Process., vol. 42, no. 1, pp. 361–384, Jan. 2023