Artificial Intelligence Approaches for Deepfake Detection: A Comprehensive Review

R. Dhanunjaya Rao[1*], K. Nagabhushan Raju[2]

*[1]Research Scholor, SKU, Ananthapur, Assistant Professor, Department of Digital Forensics, School of Sciences, Malla Reddy University, Hyderabad, India.

[2]Professor, Department of Instrumentation, Sri Krishnadevaraya University, Anantapur, Andhra Pradesh, India.

dhanunjayarao.reddy@mallareddyuniversity.ac.in*[1], knrbhushan@yahoo.com[2]

Abstract:

Deepfakes, synthetic media generated using artificial intelligence, pose serious challenges to information authenticity, legal systems, and public trust. This review paper explores the landscape of deepfake detection techniques, encompassing both image and audio modalities. It examines state-of-the-art methods such as CNNs, RNNs, transformers, and audio-visual fusion models, alongside traditional forensic approaches. The study also evaluates the performance of these models across various datasets, highlights challenges like dataset bias and generalizability, and discusses future directions including quantum-based detection and real-time robustness.

Keywords: Deepfake Detection, Video Forensics, Audio Tampering, CNN, GAN, Lip-Sync, AI Forensics, Quantum Detection, Dataset Bias

1. Introduction

Many people are concerned about deep fakes, especially because it is easy to be fooled by fake videos. Most of the time, it is hard for people to tell the difference between real and deep fake videos just by looking or listening. Deep fakes are created using several methods to make videos appear real, such as merging or replacing images and video clips. With advanced AI tools like generative adversarial networks (GANs), deep fakes can add realistic audio to videos, making both the visuals and sounds seem authentic. Facial manipulation systems and techniques have progressed to the point where even individuals without training in digital arts or picture retouching may utilise them. Indeed, there has been a recent uptick in the free distribution of nearly-automatic programs and libraries. One the one hand, new creative opportunities (in fields like filmmaking, visual effects, the visual arts, etc.) are being made possible by this technological progress. Concurrently, meanwhile, it makes it easier for bad actors to create video forgeries.

The problem of determining whether a video has undergone editing is not new. Specialists in multimedia forensics have been studying this area for a long time, and they've come up with a wide variety of answers to issues related to it. For example, many approaches to identify duplicate or missing frames, trace copy-moves using block-based or dense

D. R. Reddy et al. (eds.), *Proceedings of the First International Conference on Advances in Forensics and Cyber Technologies (ICFACT 2025)*, Advances in Computer Science Research 127,
https://doi.org/10.2991/978-94-6239-610-4_44

algorithms, and analyse video coding histories are all part of their range. The problem is that forensics footprints aren't often super obvious.

Generative adversarial networks (GANs), social media, and automated video and audio editing technologies have recently advanced to the point that high-quality manipulated video content may be quickly created and disseminated. Already, this kind of material has given rise to what is commonly known as "fake news," or purposeful disinformation that is having an effect on the political environments of many nations.

Deep fake is a method that leverages deep learning algorithms to generate deceptively realistic-looking false photographs. This is achieved, typically, by superimposing the face of another person onto a source image. Deep learning encoders and decoders, which have seen heavy use in the field of machine vision, form the basis of our deep fake creation technique. In order for encoders to produce the false image, they first need to extract all of the image's features. Nowadays, it's easy to find huge datasets of images on social media. Thanks to this abundance of data, more advanced deep fake techniques have been developed. Tensor flow is used to build many of these algorithms. However, training deep learning models used to be a difficult task. A free and open-source software library called Tensor Flow can be used to do numerical calculations with data-flow graphs. While Google created it for internal use in its own machine learning and deep neural network research and development, the system's generalisability made it suitable for use in many other domains, and it quickly gained popularity for machine learning applications after being made publicly available and free to use.

2. Literature Review

Generally, Deep Fake detection can be classified into two categories: temporal and spatial analysis for video DFD and frame forgery analysis for image DFD.

Afchar et al. [5] focused on developing detection systems that use a deep learning approach to examine the microscopic features inside photos. To tell the difference between actual and fraudulent films or images, they used two detection algorithms, Meso-4 and MesoInception 4, and two activation functions. One hidden layer in Meso-4's dense network made use of the Rectified Linear Unit (ReLU) activation function to enhance generalisation; this was followed by four convolution and pooling layers in succession. The authors of the MesoInception-4 architecture, on the other hand, used inception models in place of the first two convolutional layers before testing their performance on the Deep Fake and Face2Face images. Findings demonstrated a very high detection success rate, with 98% for the Deep Fake Dataset and 95% for the Face2Face dataset.

Feature extraction and a two-class classifier to distinguish altered from unaltered films are two components of the AV integrated system. Here, **Korshunov et al**. [6] employed the same patterns and built their detection pipeline using Mel-frequency cepstral Coefficient (MFCC) as an audio feature and distance between mouth landmarks as a visual feature. The digital presentation attacks in Deep Fake films include PCA, LDA, IQMs, and SVMs, among others. In order to distinguish between altered and non-altered movies, PCA was used to reduce the dimensionality of the feature blocks used for merging the audiovisual. Then, the data was sent

into LSTM for classification. **Korshunov et al.** [6] In addition, we looked at baseline face-swap detection systems and discovered that the lip-sync-based method was unable to identify cases where the speaker's lip movements did not match their words. In addition, they confirmed that a Support Vector Machine (SVM) classifier can detect high-quality Deep Fake films with an equal error rate of 8.97% when measuring picture quality.

Mittal et al. [7] developed a technique that utilises a combination of a Convolution Neural Network and a Recurrent Neural Network to identify altered or synthetic faces by extracting critical temporal information from them. The most trustworthy frames for identifying fake faces were automatically selected using a Gated Recurrent Unit (GRU), a weighting mechanism, and Automatic Face Weighting (AFW).

Mittal et al. [7] proposed a model for a deep learning network that can detect false films using principles from the Siamese network and triplet loss. They used the area under the curve (AUC) measure on two massive DFD datasets—the DF TIMIT and the DFDC datasets—to validate the model. Afterwards, they compare it against various SOTA DFD methods like Two-Stream, MesoNet, HeadPose, FWA, VA, Xception, Multi-task, Capsule, and DSP-FWA. On the DFDC dataset, they achieve an AUC of 84.4%, while on the DF-TIMIT dataset, it is 96.6%. In order to detect emotions in a DFD, this method was the first to use video and audio fusion modalities at the same time.

Chugh et al. [8] suggested a method that uses the modality dissonance score (MDS) to categorise DeepFake video forgeries according to the similarities and differences between the audio and visual modalities. In order to examine the proximity characteristics of the audio and video, the contrastive loss was employed. Another step in identifying the specific modality, be it audio or video, is to apply entropy loss to the characteristics.

Symeon et al. [9] DL models were trained to detect fake mage or video content using the DFDC dataset. The focus of this work was on improving detection accuracy through the extraction of facial features, with a particular emphasis on false positive photos that contain a lot of noise. Three deep learning architectures were suggested by the authors: MesoInception-4, XceptionNet, and EfficientNet. Two pre-processing phases were included in the process: a data augmentation layer and an image filtering layer. Beginning with a horizontal and vertical flip, they proceeded to pre-process the dataset by applying various transformations such as random cropping, rotation, compression, Gaussian and motion blurring, and brightness, saturation, and contrast transformation. You can boost the image's quality by using this staging layer. During the second pre-processing layer, images with sizes smaller than or equal to N/2 are removed from consideration when they are in a linked form. Here, Ni is the number of extracted frames per movie following face extraction. Finally, DL models train on DFDC and combine sigmoid activation in the last layer with the Adam optimiser and minimisation of the log loss error.

Guarnea et al. [10] developed a method to examine Deep Fakes of people's faces in order to unearth a forensic fingerprint concealed within photographs by means of Expectation Maximisation (EM) algorithms. A set of local features is extracted from images using EM methods. To validate the results, a naive classifier is tested on five architectures: GDWCTS,

STARGAN, ATTGAN, STYLEGAN, and STYLEGAN2. The architectures are evaluated against the CELEBA datasets.

Masi et al. [11] introduced a two-stage network-based DFD approach that taught itself to enhance artefacts while suppressing high-level face material, thereby isolating digitally altered faces. In its current state, this procedure employs a pre-processing step that isolates spatial frequencies. With the Laplacian of Gaussian (LOG) acting as a bottleneck layer, one branch of this two-branch structure increased multi-band frequencies while the other reduced the face content. The LOG operator enhances artefacts by masking the picture material shown in low-level feature maps and functioning as a band-pass filter. New loss functions allowed for better and broader face manipulation and promoted compact representations of real faces. A new cost function for natural face variability was derived and a technique for unrealistic feature space facial samples was proposed by the authors. A densely linked layer that learns to merge colour and frequency domain information using a multiscale Laplacian of Gaussian (LOG) operator forms the basis of the two-branch representation extractor.

Li et al. [12] developed a new approach named Face X-ray to identify fabricated faces in photos and give the blending boundaries of a binary mask-based fabricated face. In the thorough design of the detection approach, the writers merged the changed face with an existing backdrop image, and there were obtrusive image disparities across the blending boundaries. The Face X-ray method takes the greyscale picture as input and finds out if it can be made of two photos from different sources. As a result of picture blending, the binary face boundaries will be created during testing with a fake image, and it will produce a blank image when the image is authentic.

Zhao et al. [13] suggested an Interpretable Spatial-Temporal Video Transformer (ISTVT) that could capture spatial artefacts and temporal inconsistency for strong Deep fake detection. It included a new deconstructed spatial-temporal self-attention mechanism and a self-subtract mechanism. Extensive experiments employing large-scale datasets, such as Face Forensics++, Face Shifter, Deeper Forensics, Celeb-DF, and DFDC, shown increased Deep Fake detection performance.

Wang et al. [14] suggested using a deep convolutional transformer to absorb the key picture features on a global and local scale. The authors improved the efficacy and enriched the extracted features by applying convolutional pooling and re attention. They used picture key frames in model training to boost performance, and they showed how video compression affects the feature quantity gap between key and regular image frames, which helps in Deep Fake detection.

Jia et al. [15] devised a technique to detect face forgeries by combining global and local information. The goal of face forgery detection is to find cases when one person's face has been altered or copied. The term "global features" is most often used to describe the comprehensive traits that identify a person's facial anatomy and look. Facial landmarks, colour distributions, texture patterns, and statistical information are all examples of characteristic data. Using global characteristics, which give a bird's-eye view of the face, one can spot irregularities or discrepancies caused by face counterfeiting. When it comes to the face, local

features zero down on certain areas or patches. Local patterns, textures, and edges are just some of the finer characteristics captured by these features. This approach can identify forgeries or minor inconsistencies in facial features by examining local aspects. By merging global and local features, the GLFF method makes use of supplementary data. The goal of merging global and local characteristics is to improve detection accuracy by obtaining both the face's general structure and finer details. Experiments to measure the efficacy of the GLFF approach may round out this investigation. Doing so may necessitate running the method on reference datasets that include both natural and artificially enhanced face photos and videos. To prove that the GLFF method is better than other methods, the assessment might look at measures like recall, accuracy, precision, and F1 score.

Recurrent Convolutional Strategies (USC Institute)[15] To make use of the time-related data contained in video streams, Recurrent Convolutional Models are employed. It is possible to differentiate the most effective strategies for integrating variations and face pre-processing approaches by consulting Recurrent Convolutional Strategies for face manipulation detection. The two-step technique of cropping and aligning faces from video frames is the general approach for manipulation detection. (i)alignment that is explicitly driven by face landmarks, and (ii)alignment that is implicitly driven by a Spatial Transformer Network (STN). The state-of-the-art is surpassed by a blend of a face alignment method and a recurrent convolutional model. For the greatest results in detecting face manipulation in photos and videos, use a landmark-based face alignment using bidirectional-recurrent-densest.

Deep fake video detection (**Navdeep Singh Hada**) [26] Based on Deep Fake Detection, the project report evolved. When it comes to deep fake detection solutions, multi-modal detection techniques are utilised to determine if the target media has been edited or created artificially. For important artificial intelligence research in deep fake discovery, they showcased their deep convolutional neural network and built a model that was authorised using the Deep Fake Detection Challenge dataset. Results demonstrate that our method can, on average, detect DF on the web in real-world dispersion scenarios with a reliability of 94.63%. Many people are counting on real-time data to be absolutely correct. Recognising DF-made appearances relies heavily on the eyes and mouth, according to their notable empirical findings.

Fake locator (IEEE) Fake Locator is a method developed for robustly localizing GAN-based face manipulations. With the proliferation of deep learning techniques, particularly Generative Adversarial Networks (GANs). Fake Locator addresses the challenge of detecting such manipulations by focusing on localizing the specific regions of the face that have been altered. By accurately identifying these regions, it provides a means to scrutinize and potentially mitigate the spread of fake images in various contexts.

Effective and Fast Detection of Deep Fakes Based on Feature Point Defects(FFR_FD) The study primarily discusses the use of feature point defects as a reliable indicator of deep fake manipulation. By analysing the deviations in feature points between real and fake images/videos, the proposed method aims to accurately identify the presence of deep fakes. Overall, the research emphasizes the importance of feature point defects as a reliable cue for deep fake detection. The method described in the research study "FFR_FD: Effective and Fast

Detection of Deep Fakes Based on Feature Point Defects" uses feature point detector-descriptors to extract discriminative features at the pixel level and proposes the Fused Facial Region Feature Descriptor (FFR_FD) as an informative vector for deep fake detection (**Wang et al., 2021)** [14]

Walid El-Shafai (2023) [17] provides a comprehensive classification of video forgery detection techniques, ranging from active and passive methods to anti-forensics and deepfake identification. The study highlights the concerns over video authenticity, particularly in critical fields like law and security. It examines various tampering techniques, including copy-move, splicing, frame deletion, and deepfake creation, as well as machine learning approaches such as CNNs, RNNs, and GAN-based models. Key limitations discussed include poor generalization across video formats and the complexity of deepfake detection. The study underscores the need for adaptive deep learning models, standardized datasets, and real-time detection tools, making it highly relevant for advancing video forensic analysis.

Kaur, 2024[20] The study explores the creation of deepfake videos through techniques such as face swapping, lip-syncing, and attribute manipulation, highlighting the growing threat posed by AI-generated fake media. It reviews the evolution of deepfake fraud, detection strategies, and future research directions. The authors assess deep learning-based detection methods like CNNs, transformers, and biological signal-based approaches, identifying challenges such as dataset bias, high computational costs, and poor generalization across new deepfake techniques. Despite deep learning's dominance in detection, the study notes its reliance on large labeled datasets and performance issues with novel deepfake methods. The authors conclude that future research should focus on optimizing detection models for real-world applications, reducing computational complexity, and enhancing adaptability to evolving deepfake technologies.

fayyad-Kazan, 2021[19] This review emphasizes the critical role of audio forensics in authenticating digital audio evidence, which is essential in modern investigations where tampered recordings could distort legal proceedings. With advancements in AI-based speech synthesis and audio-editing tools, tampering has become increasingly sophisticated. The study covers various methodologies for enhancing, authenticating, and detecting alterations in audio recordings, including electronic measurement, waveform analysis, metadata assessment, spectral analysis, watermarking, and reverberation analysis. It critiques the strengths and weaknesses of these techniques, highlighting the ongoing challenges in detecting deepfake audio and similar forgeries. The study underscores the urgent need to continually improve forensic tools to stay ahead of evolving threats, making it highly relevant to our work on audio tampering and ensuring the reliability of digital audio in forensic investigations.

Weifeng Liu, 2024[25] This study classifies existing deepfake detection techniques into visual-only and audio-visual approaches, with a focus on detecting lip-syncing forgeries. As lip-syncing deepfakes become more prevalent, there is an urgent need for adaptive detection frameworks that can identify these manipulations in real-time. The proposed method, LipFD, addresses this by detecting temporal inconsistencies between lip movements and speech

patterns, using biological correlations of lip and head motions to enhance accuracy. The review highlights the limitations of current techniques, the development of LipFD, and its evaluation across multiple datasets. It suggests that future research should focus on dataset diversity, multilingual detection, and improved computational efficiency for real-time applications. This study is highly relevant to our work, as it provides valuable insights into detecting tampered videos through inconsistencies between audio and lip movements, a key aspect of audio-video forensics.

Milani, 2012[24] The study of video forensics is presented in this paper, with an emphasis on its significance in light of the growing accessibility of video manipulation and worries about the authenticity of digital evidence. Due to the fact that edited recordings could mislead investigations, it is vital in the media, security, and legal sectors to ensure the authenticity of video evidence. Among the forensic procedures included in the research are compression analysis, acquisition source identification, and methods for detecting forgeries, such as motion inconsistencies and double compression detection. Also discussed is the necessity for stronger defences against anti-forensic methods that hide manipulation evidence. The research notes that there are obstacles to discovering alterations in videos with low quality or that have been heavily compressed, even while forensic technologies are successful. We can now better comprehend forensic approaches for video modification detection thanks to these insights, which are extremely pertinent to our research. Improving detecting algorithms and incorporating AI-based methods should be the major goals of future research in order to achieve higher accuracy.

Delp, 2018[18] To identifying deep fake videos, a technology that is making it harder to differentiate between real and edited information due to its rapid evolution. Due to its potential for misuse in spreading false information, slander, and fraud, the proliferation of deep fakes is a major concern for public confidence. Given the increasing importance of deep fake detection in digital forensics, this project was selected to investigate a deep learning-based strategy for this problem. Though there are a number of detection methods available, many of them aren't very good at generalising their results to other datasets. Utilising a ResNet50 + LSTM model trained on the Deep fake Detection Challenge dataset, the study utilises Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to uncover spatial and temporal abnormalities in films. The study also highlights how important it is to do preprocessing processes like face detection and feature extraction in order to improve object detection performance. Due to the potential for overfitting when working with small datasets, there is room for improvement in object detection; this is a major drawback. This research proves that deep learning is effective in fighting deep false threats, which is a tremendous step forward. Because it delves deeply into the topic of using AI models for deep fake detection, its results are more pertinent to our research.

Manikandan (2019) [22] examines the challenges of video tampering in digital forensics, highlighting its potential to spread misinformation and disrupt legal proceedings. The study categorizes tampering methods into spatial, temporal, and spatio-temporal, with techniques like object addition, frame insertion, and reordering. It differentiates between active detection

methods, which rely on pre-embedded data such as watermarks, and passive methods, which identify inconsistencies in video content. The study also explores machine learning-based detection, such as Support Vector Machines (SVM), to automate the process. While active methods ensure authenticity, passive methods provide more flexibility in real-world applications. The study concludes that combining both techniques in a hybrid approach is crucial for effective video forensic analysis.

Matern, 2019[23] is about detecting deep fake videos and face manipulations using handcrafted visual artifact analysis rather than deep learning-based methods. As deep fake technology continues to evolve, it becomes increasingly difficult to distinguish between real and manipulated content, raising concerns about its misuse for political misinformation, financial fraud, and identity theft. This study was chosen because it presents an alternative approach to deep fake detection that does not rely on large labeled datasets or complex AI models, making it more interpretable and adaptable. The study identifies visual artifacts left by deep fake generation techniques, such as color inconsistencies, illumination mismatches, geometric distortions, and missing fine details. Unlike deep learning models, which function as black boxes, this method provides clear reasoning for detection and requires minimal training data. However, as generative models improve, these inconsistencies may become less noticeable, limiting the long-term effectiveness of this approach. The study concludes that deep fake detection does not always require deep learning and that simple visual inconsistencies can serve as reliable indicators of manipulation. This research is highly relevant to our project, as it provides an alternative forensic approach that can be integrated with audio analysis to enhance the overall detection of tampered videos.

Maksimovic, 2021[21] discusses partial audio match detection and localization, which are inherently important in media forensics, copyright management, and broadcast analysis. As digital audio is becoming one of the most important terrains of usage, the integrity of digital audio should be considered for any forensic investigation. Fingerprint-based methods cannot detect unknown partial matches and small-scale audio manipulations. This analysis has proposed a novel algorithm for partial audio matching without predefined queries, targeting both large-scale output analysis and fine-grained forensic examination. To analyze the dataset, methods of audio manipulation include segment insertion, removal, and copy-move forgeries, all while tracking the changes that occur between the datasets. The findings suggest that query-based matching of audio is, in fact, inefficient for forensic applications expecting some unknown alteration detection. The actual algorithm of this study exhibits higher accuracy and faces basic challenges, given the large nature of the datasets. This study advocates that a good number of detection algorithms competent in tracing unauthorized edits while verifying the authenticity of the documents at the back of it need be advanced. This work forms a basis for our project, as it highlights the forensic audio matching techniques in the analysis of tampered media content.

Aggarwal, 2017[16] examines the authentication of video content, a focal area of research in digital forensics, considering that with advances in editing tools, manipulation of videos has become relatively easy. Since digital videos represent legal evidence and key sources in journalism and intelligence, their authenticity is paramount. However, currently available

detection techniques suffer from poor generalizability, dependence on non-standardized datasets, and lack of a holistic understanding with respect to forgery detection. The study classifies forgeries into inter-frame and intra-frame manipulations, dealing with copy-paste forgeries, temporal splicing, and upscale-crop attacks. The structure of the research includes proposer forensic features and detection approaches such as forgery detection, re-capture analysis, and anti-forensic countermeasures. The results showed that most of the proposed models are heavily dependent on frame-by-frame analysis and perform poorly against sophisticated forgeries. The authors assess the different detection strategies in realistic forensic settings and identify the key challenges facing them, such as the lack of standardized benchmarks and the continual intricacy involved in manipulation tools. The main conclusion is that even though some of the existing methods are promising, a combined, multi-modal, context-relevant adoption of different forensic cues needs to be came up with it. The current study report of the findings is very closely related to our project, as it helps in building a cogent understanding of the video tampering-forensics domain.

Methodology of Deepfake Generation

Deepfakes are primarily created using encoder-decoder architectures and GANs. The encoder compresses features of the source media, while the decoder reconstructs it into a target framework. Techniques such as face swapping, lip-syncing, and expression alteration are common. These models require large datasets, often sourced from social media, making the process more accessible.

3. Detection Techniques

In response to growing worries about online privacy and security, several techniques have surfaced for identifying deep fake images; this research delves into both the creation and detection of deep fakes using deep learning, and it suggests a method for enhancing deep fakes using deep learning to make them even better.

Deep fake Creation: In order to compress images, deep fake uses an autoencoder-decoder pipeline. These encoders are based on deep neural networks; when a bottleneck is introduced to the network, the original input is compressed. As encoders become more advanced, high-quality image compression becomes possible, allowing deep fake to function with less computational power. To create deep fakes, two auto encoders are trained. The process begins with two encoders learning the features of the source and target images, which are shared between them. Then, to create the deep fake image, the decoder from the source image is used to reconstruct the target image, which will result in a picture of the target that has features from the source. Using these techniques, one can create deep false images and/or films; however, photos, due to their small size and ease of processing, are faster to manufacture than videos.

Deep fake Detection: A neural network developed with the express purpose of detecting deep fakes is Mesonet. There are a lot of deep fake videos floating around social media. Since most videos shared on platforms like Instagram are compressed and of low quality, it's not possible to do microscopic analysis based on image noise. Meso Net accounts for this. Since even humans have trouble detecting deep fakes at a higher semantic level, it uses a small-layer deep neural network as an intermediary method. This network starts with a four-layer convolution and pooling pattern, then moves on to a dense network with a single hidden layer. It is usual

practice to employ a convolution layer first, then a pooling layer, when extracting image features; this is because the convolution layer finds the features and the pooling layer generates a down-sampled version of the feature map. The convolutional layers employ non-linear ReLU activation functions and Batch Normalisation to regularise the output, enhancing generalisation; the fully-connected layers, meanwhile, employ Dropout to regularise and strengthen their resilience.

3.1 Convolutional Neural Networks (CNNs)

CNN-based approaches like EfficientNetB4, MesoNet, and XceptionNet extract spatial features for classification. Attention layers and Siamese training have been proposed to improve detection accuracy. Techniques focus on identifying visual artifacts, inconsistent lighting, or unnatural facial features.

3.2 Temporal Models and Recurrent Networks

LSTM and GRU networks are used to capture temporal inconsistencies across video frames. These methods detect manipulation by analyzing motion artifacts and sequential anomalies.

3.3 Audio-Visual Fusion Methods

Models like those proposed by Korshunov and Chugh integrate audio features (MFCC, lip-sync mismatch) with visual ones (mouth landmarks) to improve detection accuracy. These are effective in cases of audio tampering.

3.4 Transformer-based Models

Recent architectures such as Interpretable Spatial-Temporal Video Transformers (ISTVT) and convolutional pooling transformers aim to capture both local and global inconsistencies through self-attention and frequency domain analysis.

3.5 Handcrafted and Hybrid Approaches

Handcrafted methods exploit color mismatches, geometric distortions, or illumination differences. Some systems use image quality metrics (IQMs) combined with SVMs for classification.

4. Comparative Analysis of Detection Techniques

Table1: Comparative analysis of Detection Techniques

Approach	Strengths	Limitations	Accuracy (Sample Studies)
CNN + Siamese[1] [7]	High spatial accuracy	Needs large datasets	96.6% (DF-TIMIT)
Audio-Visual Fusion[23] [25]	Robust to lip-syncing	Fails with poor audio	~89% (Various datasets)

Transformers[14]	Handles global artifacts	High computational cost	94–98%
Handcrafted[23]	Interpretable	Poor performance on HQ deepfakes	<85%
Hybrid (e.g., GLFF, Face X-ray) [22]	Balanced detection	Complex architectures	Varies

5. Datasets Used for Evaluation

- **DFDC (Facebook)**
- **DF-TIMIT**
- **FaceForensics++**
- **Celeb-DF**
- **VidTIMIT**

These datasets include manipulated and real videos/images used to train and benchmark detection systems.

6. Challenges and Limitations

- Poor generalization across different compression formats and devices
- Dataset bias limiting real-world deployment
- High computational requirements
- Evasion via adversarial attacks
- Difficulty detecting unseen or soft perturbations

7. Emerging Trends and Future Directions

7.1 Quantum AI for Deepfake Detection

Quantum neural networks (QNNs) promise powerful classification capabilities for multimodal detection. Though still under research, they offer a new frontier.

7.2 Data Augmentation and Adversarial Training

GAN-based augmentation and noise injection methods are helping improve robustness.

7.3 Real-Time and Lightweight Detection

Future models must be optimized for speed and efficiency for deployment on mobile and edge devices.

To improve accuracy and resilience against modified material, it is essential to assess the effect of different data augmentation methods on the training of Deep Fake detection systems. Image manipulation, cropping, rotation, translation, noise injection, kernel filters, adversarial training, GAN-based data augmentation, neural style transfer, and meta-learning data augmentation are some of the methods that can be used. Because it relies on neural networks, generative AI, which includes Deep Fake technology, is easy prey for adversarial assaults. By taking advantage of these flaws, malicious actors can craft misleading content that could go undetected. A model's robustness against purposeful manipulations can be greatly improved using enhanced adversarial training, which integrates various adversarial examples during model training and defensive measures like adversarial loss.

Deep Fakes make use of a wide variety of transformation techniques to create fake media that fools human eyes into thinking it's real. Using machine learning, deep learning, and rule-based learning, three main approaches can be used to create and detect Deep Fakes. These approaches thoroughly examine the image or video using the extracted spatial and temporal data as embeddings to detect any assaults. Face swapping, lip-syncing, emotion changing, hair changing, false motions, fake voice synthesis, background substitution, and eye-gaze manipulation are common changes used to generate Deep Fakes. Soft or unseen attacks can also cause falsification, and they are difficult to spot using lightweight detection models or regular investigations, in contrast to hard or seen attacks, which can be detected by people or lightweight detection models. The combination of all these changes makes it difficult to create general algorithms that can identify any kind of disturbance in movies or photos. There are algorithms that work best with images only, and others that can handle both visual and aural input. As a result, certain attack detection algorithms focus on particular transformations during development and training, while others aim at different transformations. Detection models' performance could drop when testing and training sets are drastically different, hence there has to be further investigation into better model generalisation to make models that are resilient and adaptive across datasets.

Quantum neural networks (QNNs) and other quantum algorithms have a lot of promise for the future of quantum computing when it comes to distinguishing between real and fake media. Nevertheless, quantum computing is a very new discipline, and there may be a number of obstacles and restrictions that actual quantum computers must overcome. How a multi-modal corpus's pure-set and false-set are classified determines the quantum-based algorithms or models developed for Deep Fakes. Consequently, if quantum algorithms take advantage of the special features of quantum systems, classical computers might have a hard time detecting Deep Fakes. We must do continuous research to properly comprehend the implications of quantum computing for artificial intelligence, especially for vision intelligence, as its development progresses. This opens up a promising new line of inquiry into the design of Deep Fake algorithms that are compatible with and adaptable to quantum and classical infrastructures.

Most important is the capability to objectively identify video sequences where a face has been altered. The recent proliferation of sophisticated methods for altering faces in video clips highlights the continued importance of the problem of detecting such manipulation. In this

research, we look at how different trained CNN models assemble. In order to achieve better classification performance, the suggested solution uses two important concepts—attention layers, which improve the model's interpretability and learning capacity, and siamese training, which extracts deep features from data—to obtain various models from a base network, like EfficientNetB4. The main emphasis is on several modification techniques, such as deep fakes, Face2Face, Face Swap, and Neural Textures, specifically in relation to the Facebook DFDC dataset that was launched on Kaggle in December 2019. Two separate training methods, one based on the siamese paradigm, and the possibility of employing an ensemble of various CNN models are explored in the work. One of these models is an attention-enhanced variant of EfficientNetB4.

Some of the suggested methods rely on convolutional neural network (CNN)-based frame-by-frame analysis; one such method is MesoNet, a relatively shallow CNN trained to identify faux faces. Some methods use Long Short-Term Memory (LSTM) analysis to take advantage of how video frames change over time. Other approaches that aim to overcome pixel analysis' shortcomings centre on frame-level semantic analysis. As an example, one method may detect shallow fakes by comparing them to natural head poses, while another can detect deep fakes by analysing inconsistent lighting effects. The efficacy of semantic approaches declines, regrettably, as manipulation techniques provide more realistic outcomes. Furthermore, there are methods that give additional localisation data; for example, a multi-task learning approach can recognise altered faces in video and still images and produce a segmentation mask for each.

In the study "DeepFakes: a New Threat to Face Recognition? Assessment and Detection" by Pavel Korshunov and Sébastien Marcel, the ease of automatically replacing one person's face with another in a video using pre-trained generative adversarial networks (GANs) is highlighted. This research presents the first publicly available set of Deep Fake videos generated from the VidTIMIT database, utilizing open-source software based on GANs. The study emphasizes that training and blending parameters can significantly influence the quality of the resulting videos. The best-performing method, based on visual quality metrics commonly used in presentation attack detection, achieved an 8.97% equal error rate on high-quality Deep Fakes.

For detecting Deep Fakes, an audio-visual approach was initially employed to identify inconsistencies between visual lip movements and speech in audio, allowing for an assessment of how well generated Deep Fakes mimic mouth movements and whether lip synchronization with speech is accurate. Several baseline methods from the presentation attack detection domain were also applied, treating Deep Fake videos as digital presentation attacks. These methods included simple principal component analysis (PCA) and linear discriminant analysis (LDA), as well as approaches based on image quality metrics (IQM) and support vector machines (SVM). The image-based systems implemented included Pixels+PCA+LDA, which utilized raw faces as features with PCA LDA classifiers, retaining 99% variance and resulting in a 446-dimensional transform matrix, and IQM+PCA+LDA, which used IQM features with PCA-LDA classifiers, retaining 95% variance and resulting in a 2-dimensional transform matrix. The IQM+SVM system, which averaged scores from 20 frames for each video, demonstrated reasonably high accuracy in detecting Deep Fake videos, although high-quality

models posed a more significant challenge. Ultimately, techniques based on image quality measures with SVM classifiers were capable of detecting high-quality Deep Fake videos with an equal error rate of 8.97%.

8. Conclusion

The battle against deepfakes is ongoing and evolving. This review underscores the progress made in detection technologies but also highlights critical gaps. Continued advancements in AI, along with the development of standardized datasets and cross-modal models, are essential for addressing the threats posed by deepfake media. Collaborative research involving both computer vision and forensic science communities will be key to ensuring media authenticity in the digital age.

References

[1] A. Singh, "Video Face Manipulation Detection through Ensemble of CNNs," *arXiv preprint*, arXiv:2004.07676v1 [cs.CV], Apr. 2020.

[2] P. Korshunov and S. Marcel, "Deep Fakes: A New Threat to Face Recognition? Assessment and Detection," *arXiv preprint*, arXiv:1812.08685v1 [cs.CV], Dec. 2018.

[3] A. Author, "Deepfakes Creation and Detection Using Deep Learning," in *Proc. 2021 Int. Mobile, Intelligent, and Ubiquitous Computing Conf. (MIUCC)*, IEEE, 2021, doi: 10.1109/MIUCC52538.2021.9447642.

[4] G. Gupta, K. Raja, M. Gupta, T. Jan, S. T. Whiteside, and M. Prasad, "A Comprehensive Review of Deep Fake Detection Using Advanced Machine Learning and Fusion Methods," *Electronics*, vol. 13, p. 95, 2024. [Online]. Available: https://doi.org/10.3390/electronics13010095

[5] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: A compact facial video forgery detection network," in *Proc. 2018 IEEE Int. Workshop on Information Forensics and Security (WIFS)*, Hong Kong, China, 2018, pp. 1–7.

[6] P. Korshunov and S. Marcel, "Vulnerability assessment and detection of deepfake videos," in *Proc. 2019 Int. Conf. on Biometrics (ICB)*, Crete, Greece, 2019, pp. 1–6.

[7] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proc. 28th ACM Int. Conf. on Multimedia*, Seattle, WA, USA, 2020, pp. 2823–2832.

[8] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other: Audio-visual dissonance-based deepfake detection and localization," in *Proc. 28th ACM Int. Conf. on Multimedia*, Seattle, WA, USA, 2020, pp. 439–447.

[9] P. C. G. K. Z. Symeon and P. I. Kompatsiaris, "A face preprocessing approach for improved deepfake detection," *arXiv preprint*, arXiv:2006.07084, 2020.

[10] L. Guarnera, O. Giudice, and S. Battiato, "Deepfake detection by analyzing convolutional traces," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, Seattle, WA, USA, 2020, pp. 666–667.

[11] I. Masi et al., "Two-branch recurrent network for isolating deepfakes in videos," in *Proc. ECCV 2020: 16th European Conf.*, Glasgow, UK, Springer, 2020, pp. 667–684.

[12] L. Li et al., "Face X-ray for more general face forgery detection," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 5001–5010.

[13] C. Zhao et al., "ISTVT: Interpretable spatial-temporal video transformer for deepfake detection," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 1335–1348, 2023.

[14] T. Wang, H. Cheng, K. P. Chow, and L. Nie, "Deep convolutional pooling transformer for deepfake detection," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, p. 179, 2023.

[15] Y. Ju, S. Jia, J. Cai, H. Guan, and S. Lyu, "GLFF: Global and Local Feature Fusion for Face Forgery Detection," *arXiv preprint*, arXiv:2211.08615, 2022.

[16] R. D. Aggarwal, *Video Content Authentication Techniques: A Comprehensive Survey*, Springer, 2017.

[17] W. El-Shafai et al., "A comprehensive taxonomy on multimedia video forgery detection techniques: Challenges and novel trends," Springer, 2023.

[18] D. G. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. IEEE Int. Conf. on Advanced Video and Signal-Based Surveillance (AVSS)*, 2018.

[19] H. Fayyad-Kazan, "Verifying the audio evidence to assist forensic investigation," *Canadian Center of Science and Education*, 2021.

[20] A. Kaur, "Deepfake video detection: Challenges and opportunities," Springer, 2024.

[21] M. Maksimovic, "Detection and localization of partial audio matches in various application scenarios," Springer, 2021.

[22] R. H. Manikandan, "A Review: Video tampering attacks and detection techniques," *IJSRCSEIT*, vol. 11, 2019.

[23] F. Matern, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. IEEE WACVW*, 2019, p. 10.

[24] S. Milani, "An overview on video forensics," Cambridge University Press, 2012.

[25] W. Liu and T. Sun, "Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes," *Key Lab of Aerospace Information Security*, 2024.

[26] N. S. Hada, "Deepfake video detection," M.Tech Thesis, Jaypee Univ. of Information Technology, 2022. [Online]. Available: http://www.ir.juit.ac.in:8080/jspui/bitstream/123456789/3616/1/Deepfake%20Video%20Detection.pdf

[27] E. Sabir et al., “Recurrent convolutional strategies for face manipulation detection in videos,” in *CVPRW Media Forensics*, 2019. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2019/studys/Media%20Forensics/Sabir_Recurrent_Convolutional_Strategies_for_Face_Manipulation_Detection_in_Videos_CVPRW_2019_study.pdf