Deep Fake Detection: A Comprehensive Survey and Comparative Analysis

Dhavan S

Computer science Engineering Malnad College of Engineering Hassan, India dhavanj8@gmail.com

Deeksha S V *Computer science Engineering Malnad College of Engineering Hassan, India* deekshasv6@gmail.com

Dr. Kavyasri M N*Associate professor, Computer science Engineering Malnad College of Engineering*
Hassan, India mnk@mcehassan.ac.in

Dhanushree T S *Computer science Engineering Malnad College of Engineering Hassan, India* dhanushreets4@gmail.com
Deeksha K*Computer science Engineering Malnad College of Engineering* Hassan, India deekshakumar2106@gmail.com

***bstract*—DeepFake technology uses artificial intelligence to produce altered images, videos, or audio with one person's face or voice replaced by another. While it has some entertainment applications, it poses great risks in spreading misinformation, impersonation, and other malicious activities. The creation of convincing media utilizes deep learning, while detection focuses on the identification of subtle inconsistencies. The project builds a system that can not only create these deepfake videos using limited data but also detect whether a video is fake. The advancement of both creation and detection techniques continues to shape the landscape of digital media.**

I. INTRODUCTION

DeepFake technology applies artificial intelligence to create realistic but altered images, videos, or audio, wherein a person's face or voice is replaced by another. Its applications in entertainment make it useful; however, the risks it brings along make it dangerous, allowing for misinformation, imper- sonation, and other maliciously oriented activities. This tech- nology finds deep learning applications in creating convincing media, but detection methods focus on subtle inconsistencies. The project builds a system that can not only create these deepfake videos using limited data but also detect whether a video is fake. The advancement of both creation and detection techniques continues to shape the landscape of digital media.

II. PROBLEM STATEMENT

The rapid development of Deep Fake technology has allowed for the creation of highly realistic but altered media, includ- ing images, videos, and audio, where one person's face or voice can be seamlessly replaced with another's. While this technology offers creative opportunities in entertainment and media production, it also introduces significant risks, such as the spread of misinformation, identity theft, and malicious impersonation. Current systems for the creation of DeepFakes rely on huge data and resource, while the methods of detection cannot keep pace with the content generation. The work will therefore target the dual challenge of developing a system that can produce Deep Fake videos from limited data, and also a system that can accurately detect fake media.

III. LITERATURE SURVEY

We analyzed four recent literature papers and Compared to identify effective method for our project.

A. Deep Fake Face Detection Using Machine Learning and Various GAN Models

Deepfakes through GANs, in recent times, pose an immense threat with the high degree of realism in video and images. The scope of this study lies in exploring a robust methodology of detection with the use of GANs and focusing on the data preprocessing through MNIST and CelebA datasets at various resolutions (e.g., 4x4, 8x8, 16x16) to enable the effective training of models. Three GAN architectures are used. DC- GAN is used to generate realistic images, StyleGAN for high-resolution photorealistic outputs with controllable styles, and StarGAN for multi-domain image-to-image translation, making it possible to change attributes like hair color and gender. In the adversarial training framework, the discriminator should classify real images from fake images, and the generator should modify its output in order to cheat the discriminator. The stability of training is justified by discriminator loss and generator loss. Results indicate a detection accuracy of up to 95percent in distinguishing real and synthetic images, with StyleGAN and DCGAN showing superior realism and Star- GAN being highly versatile in attribute modification. Though the methodology holds promise for applications such as image synthesis and manipulated media detection, there are still challenges such as computational demands, dependence on dataset quality, and generalization to diverse datasets. This work emphasizes the GAN architectures in deepfake detection and how more diversity and resolution in datasets need to be developed..

B. Constrained Convolutional Neural Networks: A New Ap- proach Towards General Purpose Image Manipulation Detection

The detection of image manipulations is a very critical challenge in multimedia forensics, with the sophistication of forgery tools growing. Most traditional methods are only lim- ited to certain types of manipulation and hence require much time and are not very versatile. This paper proposes a general- purpose Constrained Convolutional Neural Network (CNN) for the detection of multiple types of manipulation while sup- pressing content-dependent features. It would create a dataset of manipulated grayscale image patches that would have the following operations with varying parameters applied to ensure the diversity: filtering, blurring, and compression. The novelty is within the constrained convolutional layer, suppressing the image content, but looking for manipulation traces. Later, layers will get spatial and hierarchical features that support robust detection. This model achieves an accuracy of up to 99.97percent, better than traditional approaches such as Spatial- Domain Rich Model (SRM). The model is also robust against variations in parameters and complex editing scenarios, making it scalable and adaptable for real-world applications. Some of the benefits include general-purpose applicability, content suppression to improve accuracy, and scalability with large datasets. Challenges remain with computational intensity, hyperparameter sensitivity, and dependency on dataset quality. Such technique, though prone to its vulnerabilities, constitutes significant progress toward practical multimedia forensics by promoting a scalable accuracy solution toward many forms of diverse image manipulations.

C. Deepfakes Detection Techniques Using Deep Learning: A Survey

The increasing prevalence of deepfake content, generated using advanced deep learning techniques like GANs, poses significant risks to public trust, security, and media integrity. Detecting deepfakes remains a critical challenge since tra-ditional detection methods cannot keep pace with the rapid advancements in generative models. This survey explores state-of-the-art deep learning-based techniques for deepfake detection, focusing on both image and video-based methods. The image detection models use spatial and statistical features, preprocessed to minimize artifacts for enhanced detection. Video-based methods make use of both temporal and spatial analysis, wherein CNNs are used to extract frame-level fea- tures and LSTMs to model temporal dependencies. Physiolog- ical signals, such as eye blinking and heart rate, are also used to detect anomalies in fake videos. Multi-modal detection ap- proaches combine audio and visual cues for further improved accuracy. Despite having a high detection accuracy, there are challenges here, such as the resource-intensive deep learning models, dependence on datasets, lack of real-time applicability, and vulnerability to adversarial attacks. Current models' great results on benchmark

datasets have been impressive, reaching up to 96percent accuracy in some cases, but still require work, especially on real-time detection and adversarial robustness.

Criteria	Detection using ML	Constrained CNN	Detection using Deep learning	Systematic Literature review
Research Objective	Develop algorithms for detecting Deepfake videos by analyzing traces from GAN engines.	Propose a constrained convolutional neural network (CNN) for detecting multiple image manipulations.	Conduct a systematic review of Deepfake detection techniques and evaluate their performance.	Provide a survey on Deepfake detection methods and dataset, emphasizing deep learning approaches.

Deepfake Detection: A Systematic Literature Review

Deepfakes-are a dangerous and newly emerging kind of AI- generated fake media that can be used to create realistic yet manipulated images, videos, and audio. The purpose of such media is mostly malicious: for spreading misinformation, political manipulation, or committing cybercrimes. Deepfakes have been a challenge because they evolve quickly with new generation techniques. This paper provides a systematic review of current deepfake detection methods by categorizing them into deep learning-based, classical machine learning-based, statistical, and blockchain-based techniques. Data collection makes use of extensive datasets such as FaceForensics++ and Celeb-DF which contain both authentic and manipulated media. Feature extraction seeks the presence of biological signals, spatial-temporal inconsistencies, and GAN induced artifacts. Deep learning models specifically the CNNs and RNNs automatically learn features, whereas hybrid models combine both spatial and temporal analysis for video-based detection. Statistical methods utilize techniques such as expec- tation maximization to measure differences, while blockchain- based solutions ensure the authenticity of content through de- centralized verification and tamper-proof records. Model eval- uation is performed using metrics such as accuracy, precision, recall, and AUC, benchmarking performance on datasets such as the Deepfake Detection Challenge (DFDC). The review shows that deep learning models, especially CNNs, achieve high detection accuracy, with over 90percent on benchmark datasets. Ensemble models, such as DeepfakeStack, further enhance performance up to 99.65percent. However, the ef- fectiveness of these models has challenges, including dataset biases, adversarial attacks, high computational costs, and real- time limitations. The paper concludes by underlining the need for further research into more robust, scalable solutions and exploring emerging trends, such as multi-modal approaches and advancements in adversarial robustness, to address these challenges.

IV. COMPARISON TABLE

By comparing each literature survey paper, we analysed and created a comparison table by identifying key words.

Methodology	Use DCGAN and StyleGAN models for dataset analysis, combined with CNNs.	Design a constrained CNN architecture with a new convolutional layer that suppresses content and focuses on manipulation detection.	Review 112 research papers grouped into four categories: deep learning, classical machine learning, statistical, and blockchain techniques.	Analyze and compare deep learning techniques, including CNN, RNN, and LSTM, for detecting Deepfake media.
Key Findings	GAN-based deep fakes leave detectable traces that can be identified using DCGAN and StyleGAN models.	Achieved up to 99.97% accuracy in detecting manipulations using constrained CNNs, outperforming state-of-the-art detectors.	Deep learning-based methods outperform others in detecting Deepfakes. Challenges remain in dataset variability and real-world applicability.	Deep learning techniques provide state-of-the-art results for Deepfake detection, with detailed limitations and future directions.
Accuracy	Highlights accuracies ranging from 95% to 98% using Meso-4 and Misconception-4 architectures.	Achieved up to 99.97% accuracy for image manipulation detection.	Comparative evaluation provided, concluding that DL techniques outperform other methods.	focuses on analyzing state-of-the-art performance in Deepfake detection.
	explanation.		datasets and tools.	further research.
Disadvantages	Limited scope of datasets (MNIST and CelebA) and lack of emphasis on real-world application challenges.	Performance may vary with unseen manipulation types; computationally intensive for training large datasets.	Focuses on past works (2018–2020), missing more recent advancements	Limited discussion on implementation in practical real- world scenarios; primarily theoretical analysis.

The table showcases a progression from algorithm development to reviews and surveys . Papers 1 and 2 excel in technical innovation and accuracy, while Papers 3 and 4 provide broader overviews and insights. Each paper has distinct strengths and limitations, making them collectively valuable for understanding the landscape of Deepfake detection.

V .CONCLUSION

The Deepfake detection is essential in dealing with the issues of misinformation, security, and privacy. The advanced deep learning techniques and CNNs have proven to outperform in identifying the manipulation artifacts in a subtle manner. The results were also able to reach 99percent accuracy in controlled settings. However, it relies on diverse and high- quality datasets; otherwise, the lack of data limits generalization to real-world scenarios. Deepfake generation is dynamic, so the detection models have to be updated regularly to remain effective against the changing threats. The project has a potential to deliver a robust solution that can be integrated into platforms like social media or forensic tools for broader impact by emphasizing scalability, dataset diversity, computational efficiency, and real-world testing.

REFERENCES

[1] Riya Kaku and Upasana Mishra Tiwari , “Deep Fake Face Detection Using Machine Learning and Various GAN Models”,Research gate
,March 2022, DOI:10,46647//IJETMS.2022.V06I02.001.

[2] Abdulqader M.Almars, “Deepfakes DetectionTechniques UsingDeep Learning:A Survey”,Scientific Research Publishing,september 2021.

[3] Belhassen Bayar,Student member,IEEE and Matthew C.Stamn, Mem- ber,IEEE, “Constrained Convolutional Neural Networks: A New Ap- proach Towards General Purpose Image manipulation Detection”, IEEE Xplore.

[4] Md Shohel Rana and Andrew H. Sung, “Deepfake Detection: A Sys- tematic Literature Review”, IEEE , Jan 2022.