IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Deepfake Video Identification and detection method using Artificial Intelligence

Alap Mahar[1], Dr. Pushpneel Verma[2], Dr. Ajit Singh[3]

[1]*Department of Computer Science and Engineering, Bhagwant University, India*
alapmahar@gmail.com

[2]*Department of Computer Science and Engineering, Bhagwant University, India*
pushpuneelverma@gmail.com

[3]*Department of Computer Science and Engineering, VMSB University, India*
erajit@rediffmail.com

Abstract

This survey explores the landscape of deepfake video identification and detection methods, leveraging the capabilities of artificial intelligence (AI). As the proliferation of deepfake technology poses significant challenges to the veracity of digital content, the need for robust and efficient detection mechanisms becomes paramount. The survey delves into various AI-driven approaches employed for discerning deepfake videos from authentic content, including Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), and recurrent architectures like Long Short-Term Memory (LSTM) networks. Based on the comparative analysis, the hybrid LSTM-CNN method is ineffective in detecting deepfakes, with a low accuracy rate of 82%. However, the hybrid method combining CNN with the Jaya Algorithm significantly improves accuracy to 99.3%, outperforming all other methods in detecting deepfake videos. This method also achieves the highest precision rate, though its recall rate, while robust, is slightly lower than some other techniques. The study addresses the limitations and challenges associated with these methods, including the constant evolution of deepfake techniques and the ethical considerations surrounding privacy and consent. By providing a comprehensive overview of the current state of deepfake video identification, this survey aims to contribute valuable insights for researchers, practitioners, and policymakers navigating the intricate landscape of AI-powered deepfake detection..

Keywords:

Deepfake video identification, challenges and limitations, AI

1. Introduction

In recent years, the proliferation of deepfake videos, which involve the use of artificial intelligence to fabricate realistic yet entirely synthetic content, has underscored the pressing need for robust identification and detection methods[1].Deepfake is a system that use deep learning, an artificial intelligence approach, to alter and produce specific videos[2]. Deepfake films often include the manipulation of videos to replace the original person's face with someone else's. Utilisation of deep learning algorithms to substitute individuals' facial features in films. A significant number of individuals using deepfake technology for malicious intentions. Two instances of deepfake

videos are now extensively spreading on various social media platforms. The first case pertains to the use of deepfake technology for creating pornographic content, while the subsequent case involves the use of deepfakes in a malicious campaign targeting political opposition[3]. In the first scenario, pornographic movies undergo processing using deepfake technology, whereby the visages of the performers in the videos are substituted with those of artists or prominent individuals, with the intention of tarnishing the reputation of the targeted individuals. In contrast, the second scenario involves the alteration of videos of individuals making contentious remarks, where the faces of the speakers are substituted with those of other political figures, typically those involved in a political competition, in order to diminish the electability of the original person[4].Multiple techniques have recently been suggested to identify modified material by examining spatial and frequency data in pictures, as well as temporal and frequency data from audio and video[5]. Several benchmarking datasets have been made publicly accessible to enhance the current capabilities of DeepFake detection. By using these databases and established methodologies, cutting-edge techniques have developed to capitalise on the notion of information fusion, enabling reliable identification of counterfeit media.Artificial Intelligence (AI) emerges as a pivotal technology in this realm, offering innovative solutions for deepfake video identification and detection[6]. Leveraging advanced neural network architectures, machine learning algorithms, and computer vision techniques, AI plays a crucial role in scrutinizing the subtle anomalies and inconsistencies within deepfake videos. This intersection of AI and video forensics not only aims to protect the integrity of multimedia content but also signifies a critical step in fortifying the digital landscape against the potential misuse of synthetic media.This introduction sets the stage for exploring the methodologies and technologies employed in the ongoing efforts to counter the challenges posed by deepfake videos through the lens of artificial intelligence.

1.1 Deepfake Technology

Deepfake refers to the manipulation of photos and videos in order to create deceptive material that seems convincingly authentic. Prior to the advent of deepfake technology, conventional methods for manipulating images and videos included the process of splicing. picture splicing involves the alteration of a picture by replacing certain objects[7]. Overwritten objects may originate from either the same picture or a distinct image. The outcome of the overwriting procedure is a picture that include newly introduced or replicated elements, as well as those that have been changed or eliminated[8]. Regarding video splicing, the modification process involves either adding frames or deleting frames, depending on the intended purpose of the spliced video. Copy-move forgery is the commonly used term for the technique of picture and video splicing. In order to identify instances of image and video splicing, a common approach involves using a feature extraction algorithm to perform a matching process and find duplicated portions within the picture and video[9][10]. One other approach to identify picture and video splicing involves doing statistical analysis of the information to identify regions with anomalies[11]. Deepfake content creation often involves using a neural network structure called Deep Autoencoder[12], as opposed to utilising picture and video splicing techniques. The deep autoencoder architecture is a kind of artificial neural network that consists of two components: an encoder and a decoder. The encoder takes a picture as input and converts it into a vector representation in latent space, while the decoder reconstructs the vector back into the original image. The deep autoencoder is taught to minimise the discrepancy between the reconstructed picture and the original image. Deepfake material is generated by using two autoencoder designs that use a common encoder. Assume that our objective is to substitute face A with face B. We construct two autoencoders, A and B, using an identical encoder and distinct decoders, namely decoder A and decoder B. Every autoencoder is taught to accurately recreate both pictures A and B. Deepfake material is generated. By feeding picture A into autoencoder B. The purpose is to generate face picture B as the output given face image A as the input. Figure 1 illustrates the process of generating deepfakes using an autoencoder.

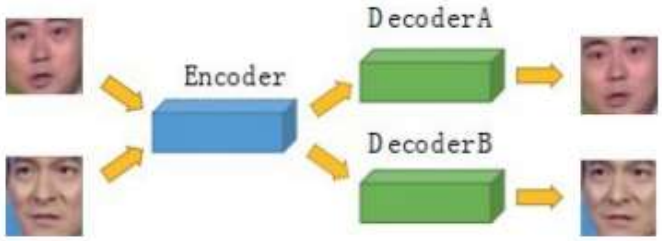

Figure 1. Deepfake Generation Illustration Using Autoencoder

Another method commonly used in producing deepfake content is Generative Adversarial Networks (GAN). GAN is a two-part neural network architecture called generator and discriminator. Generator is part of GAN which functions to produce fake content (fake) from a random vector. Discriminators are part of GAN which functions to detect whether content is original content or fake content produced by a generator[13]. GAN is trained to improve the performance of the generator so that the content produced is as natural as possible. Fig. 2 shows an illustration of the mechanism of the GAN.

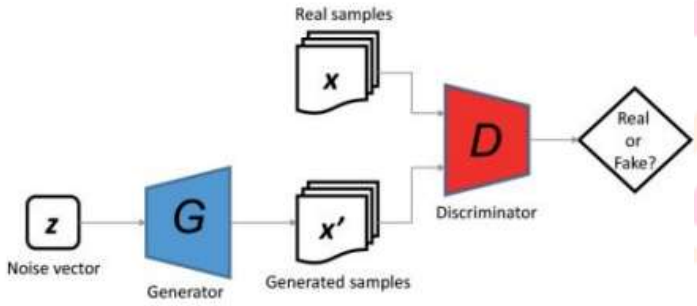

Figure 2. Illustrator of GAN [14].

1.2 Deepfake video generation

Following the first release of deepfake videos, further modification algorithms have been promptly introduced, with a majority of them relying on generative networks. These approaches use the utilization of deepfake algorithms to generate fabricated information that violates personal privacy, resulting in significant detrimental consequences for society. This section will provide an overview of the development of deepfake algorithms and thereafter delineate two distinct categories of deepfake algorithms.

1.2.1 Development of deepfake technologies

Facemanipulation is not a recent technological innovation. The first documented instance of facial alteration in literature may be seen in the renowned 1865 picture of the United States President, Abraham Lincoln. Due to advancements in computer graphics technology, the alteration of facial features in digital photographs has become readily attainable[15][16]. The development of facemanipulation technology has been significantly enhanced by recent developments in the area of deep learning. Based on the various objectives of face manipulation algorithms, current deepfake algorithms may be categorised into two distinct groups: face swapping and face reenactment.

1.2.2 Face swapping

In recent years, face swapping videos, which involve exchanging the identities of individuals in two recordings, have garnered significant interest. Research studies on this topic have been conducted since 2017. The research conducted by Korshunova et al. [17] included training convolutional neural networks (CNNs) to recognise the visual characteristics of a certain individual from a disorganised collection of photographs. This training allowed for the creation of face-swapping pictures of superior quality. However, the lack of consideration for temporal consistency renders this technique unsuitable for high-quality video creation. In the same year, Olszewski et al. [18]

introduced an innovative method for producing films using just one RGB picture and a source video sequence. A deep generative network was used to deduce perframe texture distortions of the desired identity by using source textures and the only target texture. In December 2017, a Reddit user shared the first movie created using the deepfake technique, which included changing faces. This video caused a remarkable sensation worldwide. Subsequently, a global trend of producing face-swapping movies emerged, irrespective of their intended impact, whether good or bad.

1.2.3 Face reenactment

In contrast to face-swapping technology, face reenactment algorithms aim to manipulate individuals' facial expressions in recordings. This allows attackers to create movies in which someone is coerced into doing actions that never occurred. The first face recreation algorithm may be traced back to 2006. Vlasic et al.[19] suggested conducting facial reenactment by using a face template that was adjusted according to various expression criteria. The majority of the ensuing research builds upon these techniques, in which a parametric model is used to modify face photographs. While these technologies have the ability to produce facial pictures that are very realistic, the outcomes often lack consistency across time. The study conducted by Suwajanakorn et al. [20] somewhat addressed this deficiency. Their objective was to acquire a sequence mapping that connects audio and film in order to control actors into speaking the exact same words as the speech material. The audio sequence was used to extract features, which were then inputted into a recurrent neural network (RNN). The RNN produced a sparse mouth shape for each frame of the video output. The oral sensations are also combined and integrated into the original videos.

2. Review of Literature

Seraj et al., (2024)[21]studied with the advent and popularity of generative models such as GANs, synthetic image generation and manipulation has become commonplace. This has promoted active research in the development of effective deepfake detection technology. While existing detection techniques have demonstrated promise, their performance suffers when tested on data generated using a different faking technology, on which the model has not been sufficiently trained. This challenge of detecting new types of deepfakes, without losing its prior knowledge about deepfakes (catastrophic forgetting), is of utmost importance in today's world. In this paper, we propose a novel deep domain adaptation framework to address this important problem in deepfake detection research. Our framework can leverage a large amount of labeled data (fake / genuine) generated using a particular faking technique (source domain) and a small amount of labeled data generated using a different faking technique (target domain) to induce a deep neural network with good generalization capability on both the source and the target domains. Further, deep neural networks are data-hungry and require a large amount of labeled training data, which may not always be available in the context of deepfake detection; our framework can also efficiently utilize unlabeled data in the target domain, which is more readily available than labeled data. We design a novel loss function and use the stochastic gradient descent (SGD) method to optimize the loss and train the deep network. Our extensive empirical studies on the benchmark FaceForensics++ dataset, using three types of deepfakes, corroborate the promise and potential of our framework against competing baselines.

DHANARAJ et al., (2024)[22]Generative AI, also known as GenAI, has the ability to produce intricate and high-quality material that imitates human ingenuity. This technology proves advantageous for several sectors like gaming, entertainment, and product creation. Lately, deepfakes, which are artificial intelligence-generated counterfeit films, have grown more prevalent and persuasive. Face warping is an extra deepfake method that use digital processing to deliberately alter facial features in a noticeable manner. It is essential to track the distortion in photos and videos to avoid its malicious use. A method is suggested for identifying and pinpointing distorted regions of the face in video. The input video is retrieved in order to apply several image pre-processing algorithms that enhance the video and make it more suitable for effective classification of the classes. Transfer learning is used, and a pre-trained model is applied to train a Convolutional Neural Network (CNN) utilising the source videos in

order to detect face warping. According to the testing findings, it was concluded that the suggested model effectively identifies and pinpoints the distorted regions of the face with a precision of 89.25%.

Elpeltagy et al., (2023)[23]studied that the rapid progress in deep learning-based technology has led to the emergence of many synthetic video and audio creation techniques that can produce very realistic deepfakes. Deepfakes may be used to mimic the identity of an individual in films by substituting the face of the original person with that of the desired target. Deepfakes may also be used to replicate the speech of a certain individual by using audio samples. If used with malevolent intent, these sophisticated deepfakes might potentially endanger society. Therefore, the crucial matter at hand is to differentiate between deepfake visual video frames and cloned voices from real ones. This study introduces an innovative intelligent technique for detecting deepfake videos. The video frames and audio are taken from the provided videos. Two feature extraction approaches are suggested, one for each modality: visual video frames and audio. The first approach involves using an enhanced XceptionNet model to extract spatial characteristics from video frames. It generates a feature representation for visual video frames. The second model is a customised version of the InceptionResNetV2 model that uses the Constant-Q Transform (CQT) approach. It is used to extract time-frequency information from audio data. It generates a feature representation for the audio. The collected features from both modalities are combined in an intermediate layer to create a bimodal feature representation based on the information contained in the whole video. The Gated Recurrent Unit (GRU) based attention mechanism uses three separate representation levels to separately learn and retrieve significant temporal information at each level. Subsequently, the algorithm verifies whether the fake only affects video frames, audio, or both, and renders the ultimate determination about the genuineness of the video. The recently proposed technique has been assessed using the FakeAVCeleb multimodal videos dataset. The study of the experimental findings confirms the superiority accuracy which is 98.5% of the novel approach compared to the present state-of-the-art procedures.

Ikram et al., (2023)[24] examined in the current era, many fake videos and images are created with the help of various software and new AI (Artificial Intelligence) technologies, which leave a few hints of manipulation. There are many unethical ways videos can be used to threaten, fight, or create panic among people. It is important to ensure that such methods are not used to create fake videos. An AI-based technique for the synthesis of human images is called Deep Fake. They are created by combining and superimposing existing videos onto the source videos. In this paper, a system is developed that uses a hybrid Convolutional Neural Network (CNN) consisting of InceptionResnet v2 and Xception to extract frame-level features. Experimental analysis is performed using the DFDC deep fake detection challenge on Kaggle. These deep learning-based methods are optimized to increase accuracy and decrease training time by using this dataset for training and testing. We achieved a precision of 0.985, a recall of 0.96, an f1-score of 0.98, and accuracy of 0.968.

Malik et al., (2023)[25] analyzed that Deep Learning (DL) is a sophisticated and efficient technique that is extensively used in several sectors, such as medical imaging (MI). Data Mining (DM), Image Processing (IP), and Machine Vision (MV). Deepfake use deep learning technology to modify films in a manner that makes them impossible to differentiate from the actual human subjects. Researchers have lately focused on the efficacy of deep-fake technology, leading to the development of many deep learning-based methods to detect deep-fake movies. This research introduces a new approach for detecting deep-fake videos. The study used the Deep Fake Detection Challenge (DFDC) and Face Forensic databases. Furthermore, a frequency-based approach was used to extract frames from each movie during the preprocessing phase. Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) were used to detect counterfeit movies. The accuracy of the LSTM-CNN technique reached 82%. This study will be beneficial to researchers in detecting counterfeit videos using deep learning methods.

Hussain et al., (2023)[26] intended that the proliferation of deepfake videos is a growing worry due to their capacity to disseminate false information and inflict damage. This research presents a new method for precisely identifying deepfake films by integrating Convolutional Neural Networks (CNNs) with the Jaya algorithm

optimisation. The method is assessed using two publicly accessible datasets, namely the DeepFake Detection Challenge (DFDC) dataset and the Celeb-DF dataset, and achieves the highest level of performance on both datasets. The proposed method attainedaaccuracy of 99.3% on the DFDC dataset and 97.6% on the Celeb-DF dataset. Additionally, the high F1 scores obtained demonstrate a commendable level of precision and recall in the identification of deepfake films. In addition, our strategy exhibits greater resilience against adversarial assaults compared to current cutting-edge techniques. The integration of Convolutional Neural Networks (CNNs) with the Jaya algorithm optimisation allows for efficient extraction of temporal information in video sequences. Additionally, the use of reliable evaluation criteria guarantees objective assessment and comparison with other established techniques. The technique that indicate presents a very efficient way for identifying deepfake videos, which might serve as a helpful tool in media forensics, content moderation, and cyber security.

Patel et al., (2022)[27] stated in the present day, software technologies using deep learning have simplified the process of creating believable video face swaps with few indications of tampering, sometimes referred to as "DeepFake" movies. The use of visual effects has allowed for manipulation in digital media for many years. However, recent advancements in deep learning have greatly increased the ability to create realistic false material or content using simple methods. These are media created by Artificial Intelligence, often referred to as DF. Employing artificial intelligence methods to generate the DF is a straightforward endeavour. Nevertheless, identifying these DFs is a substantial obstacle. Teaching the algorithm to identify the DF is challenging. We have achieved advancements in identifying the DF by using Convolutional Neural Networks and Recurrent Neural Networks. The system uses a Convolutional Neural Network (CNN) at the frame level to extract information. These observations are recorded and may be used to train a Recurrent Neural Network (RNN), which has the capability to learn and categorise if a video has been manipulated and detect the temporal abnormalities in the frame caused by DF tools. By using a straightforward design, our system may achieve competitive results in this project.

Ge et al., (2022)[28]examined that the identification of deepfake videos has become more complex due to the development of more sophisticated deepfake techniques. It has been noticed that deepfake videos often display anomalies in the appearance of certain face components between frames. This leads to distinct patterns in the spatial and temporal features among the semantic-level feature maps. Based on this discovery, we suggest a proactive method of learning called Latent Pattern Sensing to identify the specific features of semantic changes in order to detect deepfake videos. The methodology employs a Convolution Neural Network-based encoder, a ConvGRU-based aggregator, and a single-layer binary classifier in a cascading manner. The encoder and aggregator undergo pretraining in a self-supervised manner in order to generate spatiotemporal context characteristics that serve as representations. Next, the classifier undergoes training to accurately detect the contextual characteristics, effectively differentiating fabricated films from authentic ones. Ultimately, we suggest using a discerning self-distillation fine-tuning technique to enhance the durability and effectiveness of the detector. By using this approach, the recovered features may comprehensively capture the underlying patterns of films both geographically and temporally, resulting in a powerful and resilient deepfake video detector. Empirical tests and thorough analysis provide evidence of the efficacy of our method. For instance, our approach achieved an exceptional Area Under Curve (AUC) score of 99.94% on the FaceForensics++ benchmark, surpassing at least 12 state-of-the-art methods by 7.90% and 8.69% on the challenging DFDC and Celeb-DF(v2) benchmarks, respectively.

Ismail et al., (2021)[29] stated that there has been a proliferation of deepfake methods that enable the seamless substitution of faces, facilitating the effortless production of very authentic counterfeit films. The importance of verifying the genuineness of a video has escalated due to its possible adverse repercussions on a global scale. Introducing novel research called You Only Look Once Convolution Recurrent Neural Networks (YOLO-CRNNs) for the purpose of detecting deepfake films. The YOLO-Face detector identifies facial areas in each frame of the video, while a fine-tuned EfficientNet-B5 is used to extract the spatial characteristics of these faces. The input sequences are processed as a batch and fed into a Bidirectional Long Short-Term Memory (Bi-LSTM) to extract the

temporal properties. The novel approach is then assessed on a substantial dataset called CelebDF-FaceForencics++ (c23), which is a fusion of two widely-used datasets, FaceForencies++ (c23) and Celeb-DF. The pasting data strategy yields an AUROC score of 89.35%, an accuracy of 89.38%, a recall of 83.15%, a precision of 85.55%, and an F1-measure of 84.33%. The experimental study confirms the superiority of the suggested approach in comparison to the state-of-the-art methods.

Cozzolino et al., (2021)[30] examined a significant obstacle in the identification of DeepFake forgeries is that cutting-edge algorithms are primarily taught to identify a certain fraudulent technique. Consequently, these methods exhibit inadequate ability to apply to various forms of facial alterations, such as exchanging faces or reenacting facial expressions. In order to achieve this objective, we provide ID-Reveal, a novel methodology that use metric learning in conjunction with an adversarial training strategy to acquire temporal face characteristics that are unique to an individual's speech-related movements. The benefit lies in the fact that we may train just on authentic films without requiring any training data for counterfeit ones. In addition, we use advanced semantic characteristics, which enhance the resilience against prevalent and disruptive methods of post-processing. We do a comprehensive experimental investigation on many benchmarks that are publicly accessible. Our technique surpasses the current state of the art by enhancing generalisation and exhibiting more resilience to low-quality movies, which are often disseminated over social networks. Specifically, we achieve a mean enhancement of over 15% in accuracy while doing face reenactment on highly compressed footage.

Lewis et al., (2020)[31]evaluated that the verification of digital material has become an increasingly urgent need for contemporary culture. With the advent of Generative Adversarial Networks (GANs), the task of distinguishing synthetic media has gotten ever more challenging. Deepfakes refer to synthetic films that manipulate the appearance and/or sounds of individuals, posing a significant risk to trust and privacy in the realm of digital media. Deepfakes may be used as a means to gain political benefit, defame individuals, and destroy the credibility of public figures. Although deepfakes have flaws, individuals find it challenging to differentiate between genuine and altered pictures and videos. Hence, the presence of automated systems that can precisely and effectively categorise the authenticity of digital material is crucial. Several contemporary deepfake detection techniques use individual video frames and primarily analyse the spatial characteristics of the picture to deduce the genuineness of the video. Several effective methods leverage the temporal irregularities present in edited movies. Nevertheless, the main emphasis of study is in the analysis of spatial characteristics. Our proposition entails using a hybrid deep learning methodology that integrates spatial, spectral, and temporal elements in a coherent manner to discern between authentic and counterfeit films. Our study demonstrates that the use of the Discrete Cosine transform may enhance the detection of deepfake videos by effectively capturing the spectral characteristics of each frame. This study involves the development of a multimodal network that investigates novel characteristics for the purpose of identifying deepfake videos. The network achieves an accuracy of 61.95% on the Facebook Deepfake Detection Challenge (DFDC) dataset.

2.1 Comparison of reviewed technique

There is a wide range of authors who studied on a survey on deepfake video identification and detection method using artificial intelligence and give their findings as shown below.

Table 1. Comparison of reviewed technique

Authors [Ref.]	Technique	Outcome
Seraj et al., (2024) [21]	GAN	The proposed framework can also efficiently utilize unlabeled data in the target domain, which is more readily available than labeled data.
DHANARAJ et al., (2024)[22]	CNN	According to the testing findings, it was concluded that the suggested model effectively identifies and pinpoints the distorted regions of the face with a precision of 89.25%.
Elpeltagy et al., (2023)[23]	XceptionNet	The study of the experimental findings confirms the superiority accuracy which is 98.5% of the novel approach compared to the present state-of-the-art procedures.
Ikram et al., (2023)[24]	InceptionResnet v2+Xception	These deep learning-based methods are optimized to increase accuracy 96.8% and decrease training time by using this dataset for training and testing.
Malik et al., (2023)[25]	LSTM-CNN	The accuracy of the LSTM-CNN technique reached 82% and it will be beneficial to researchers in detection counterfeits videos using deep learning methods.
Hussain et al., (2023)[26]	CNN + Jaya algorithm	The proposed method attained a accuracy of 99.3% on the DFDC dataset and 97.6% on the Celeb-DF dataset.
Patel et al., (2022)[27]	CNN-RNN	By using a straightforward design, our system may achieve competitive results in this project.
Ge et al., (2022)[28]	CNN-based Encoder	Our approach achieved an exceptional Area Under Curve (AUC) score of 99.94% on the FaceForensics++ benchmark, surpassing at least 12 state-of-the-art methods by 7.90% and 8.69% on the challenging DFDC and Celeb-DF(v2) benchmarks, respectively.
Ismail et al.,	YOLO-CRNN	The experimental study confirms the

(2021)[29]		superiority of the suggested approach in comparison to the state-of-the-art methods.
Cozzolino et al., (2021)[30]	ID-Reveal	We achieve a mean enhancement of over 15% in accuracy while doing face reenactment on highly compressed footage.
Lewis et al., (2020)[31]	Multi-model network	The network achieves an accuracy of 61.95% on the Facebook Deepfake Detection Challenge (DFDC) dataset.

3. Challenges and solution for deepfake videoidentification using AI

Deepfake video identification and detection using Artificial Intelligence (AI) pose significant challenges and limitations, primarily due to the evolving sophistication of deepfake generation techniques which has been described given below.

- **Rapid Advancements in Deepfake Technology:** Deepfake techniques are continually evolving, making it challenging for AI-based detection methods to keep up with the latest advancements. New models with enhanced capabilities may outpace existing detection techniques.
- **Limited Labeled Training Data:** Developing robust deepfake detection models requires extensive labeled datasets, which are often scarce. The lack of diverse and comprehensive training data can hinder the ability of AI models to generalize effectively across various deepfake creation methods.
- **Transferability Across Different Content Types:** Deepfakes can take different forms, including face swaps, voice synthesis, and body replacements. Building a detection model that effectively generalizes across these diverse manipulation techniques poses a significant challenge.
- **Adversarial Attacks:** Deepfake creators can use adversarial techniques to deliberately manipulate content and evade detection. AI models may struggle to distinguish between subtly manipulated content and authentic videos.
- **Real-Time Processing Constraints:** The computational complexity of deepfake detection models may limit their applicability in real-time scenarios, such as social media platforms or live video streaming, where quick decision-making is crucial.
- **Ethical and Privacy Concerns:** The deployment of deepfake detection systems must navigate ethical considerations and privacy concerns. Balancing the need for detection with individual privacy and avoiding false positives is a delicate challenge.

Solutions:

- **Continuous Model Training and Updating:** Regularly updating detection models with new data and adapting to emerging deepfake techniques can help in staying ahead of evolving threats.
- **Data Augmentation and Synthesis:** Augmenting existing labeled datasets and synthesizing diverse data can address the challenge of limited training data, improving the model's ability to generalize across various deepfake types.
- **Ensemble Learning:** Combining multiple detection models through ensemble learning can enhance overall performance by leveraging the strengths of different algorithms, improving accuracy and robustness.

- **Explainable AI (XAI):** Integrating explainable AI techniques helps in understanding how models make decisions. This transparency can aid in identifying and addressing vulnerabilities, making the detection process more reliable.
- **User Education and Awareness:** Raising awareness among users about the existence and potential risks of deepfakes can contribute to a collective effort in combatting the spread of manipulated content.
- **Real-Time Detection Optimization:** Optimizing deepfake detection algorithms for real-time processing by leveraging hardware acceleration and efficient model architectures enhances their suitability for applications requiring immediate action.
- **Collaborative Efforts and Standardization:** Promoting collaboration between industry, researchers, and policymakers can lead to the development of standardized evaluation metrics and methodologies, fostering a more unified approach to deepfake detection.
- **Privacy-Preserving Technologies:** Implementing privacy-preserving technologies ensures that deepfake detection methods respect individual privacy, minimizing the risk of unintended consequences and ethical concerns.

Addressing the challenges and limitations in deepfake video identification and detection requires a multidisciplinary approach, combining advancements in AI research, data privacy considerations, and collaborative efforts across various stakeholders.

4. Comparative analysis

In this section, several authors provide their results following the accuracy performance metrics, which are described in table 2. According to Table 2, DHANARAJ and his fellow students were able to greatly boost the accuracy using the CNN method for detect the deepfake pinpoints the distorted regions of the face, which resulted in 89.25%. By using a XceptionNet method, Elpeltagy and his colleagues obtained 98.5% accuracy, while Ikram and his colleagues attained 96.8% accuracy using the (InceptionResnetV2+Xception), which is higher as compared to CNN. Malik and his fellow students attained the accuracy 825 by using the hybrid LSTM-CNN method. By using hybrid method (CNN +Jaya Algorithm), Hussain and his colleagues achieved a superior accuracy of 99.3% which is greater as compared to all other methods.

Table 2. Comparative analysis

Author	Year	Technique	Accuracy
DHANARAJ et al., [22]	2024	CNN	89.25%
Elpeltagy et al., [23]	2023	XceptionNet	98.5%
Ikram et al., [24]	2023	InceptionResnetV2+Xception	96.8%
Malik et al., [25]	2023	LSTM-CNN	82%
Lewis et al., [31]	2020	Multi-model network	61.95%
Hussain et al., [26]	2023	CNN+ Jaya Algorithm	99.3%

The highly achieved accuracy is revealed in Fig. 3., as can be seen in the following graph. The CNN and Jaya Algorithm (CNN+ Jaya Algorithm) has attained maximum accuracy which is 99.3% for detect the deepfake videos as compared to other methods as shown in the graph.

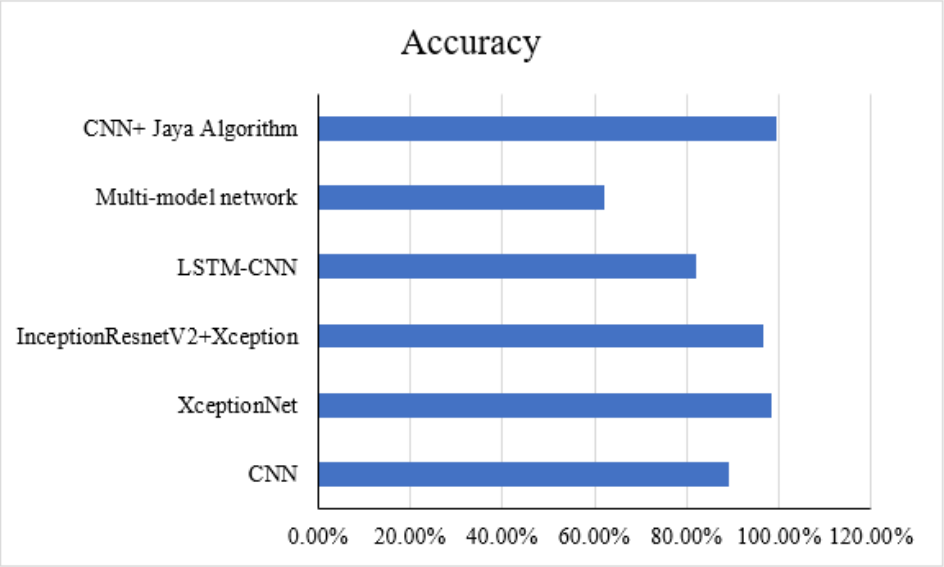

Figure 3. Comparative analysis

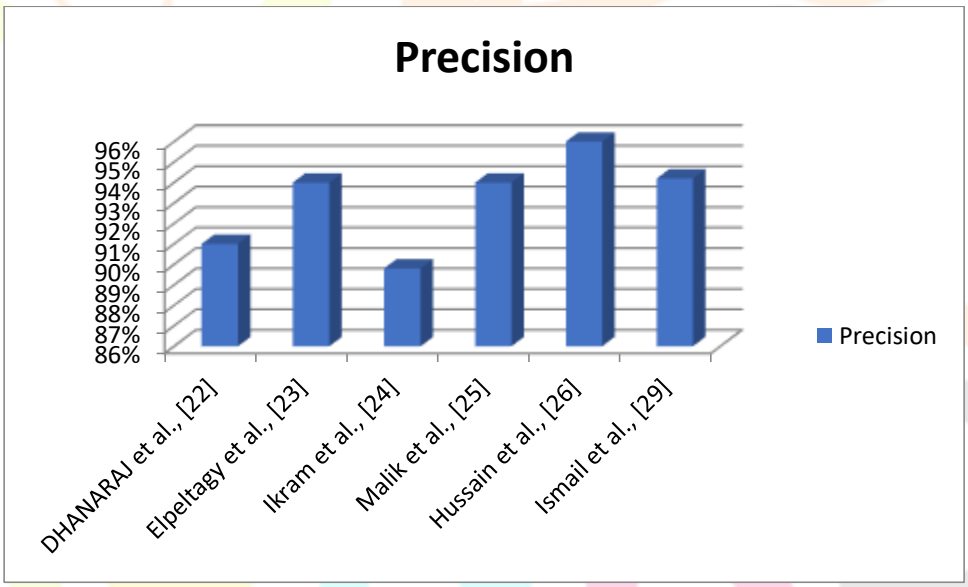

Figure 4. Precision Analysis

Among the techniques evaluated mentioned in figure 4, the method proposed by Hussain et al. (referenced as [26]) utilizing the CNN + Jaya algorithm achieves the highest precision rate. Specifically, the precision rate for Hussain et al. is significantly higher than several other methods, surpassing those of Dhanaraj et al. [22], Elpetagy et al. [23], Ikram et al. [24], and Malik et al. [25]. While Ismail et al. [29] also demonstrate a high precision rate comparable to Hussain et al., it is evident that the CNN + Jaya algorithm by Hussain et al. stands out for its superior performance. This high precision rate underscores the effectiveness of their approach in the context of the discussed application.

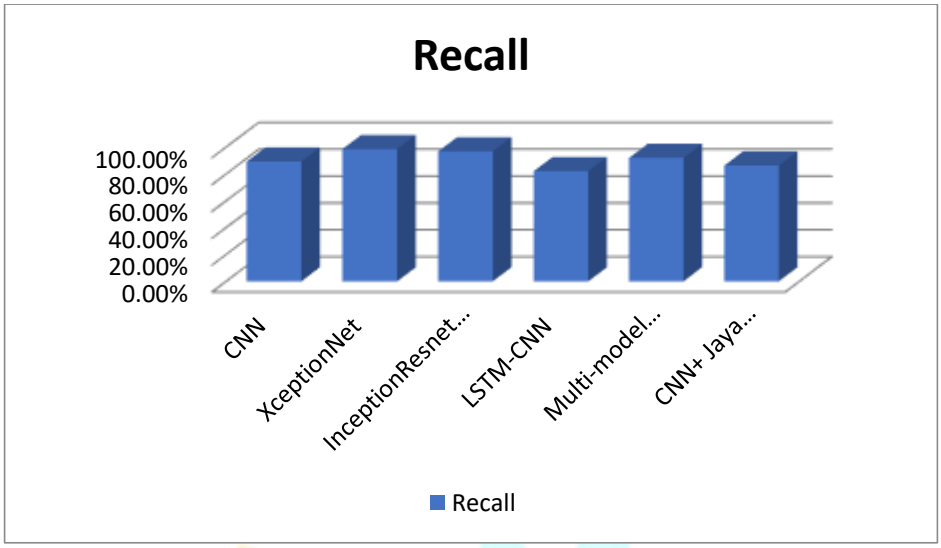

Figure 5. Recall Analysis

The recall rates of various techniques shown in figure 5 are compared in Figure 5, which provides insight into the effectiveness of these methods in correctly identifying relevant instances. Among the methods evaluated, InceptionResnet exhibits the highest recall rate, closely followed by XceptionNet and LSTM-CNN. These three techniques outperform the others in terms of recall, demonstrating their superior ability to capture relevant instances.

CNN, Multi-model approaches, and the combination of CNN + Jaya algorithm also show commendable recall rates but fall short compared to the top three methods. Specifically, the recall rate for the CNN + Jaya algorithm, despite being robust, is slightly lower than those of the standalone CNN, XceptionNet, and InceptionResnet techniques. This indicates that while the CNN + Jaya algorithm excels in precision as noted previously, its recall performance, while strong, is not the highest among the methods compared in this analysis.

5. Discussion

The survey on deepfake video identification and detection methods using artificial intelligence (AI) represents a crucial exploration into the rapidly evolving landscape of multimedia forensics. As the prevalence of deepfake content continues to rise, the development of effective techniques to identify and detect manipulated videos becomes imperative. The survey likely delves into various AI-driven approaches employed in this domain, such as machine learning algorithms, neural networks, and computer vision methods. One key focus is likely on the advancements in deep learning, particularly the utilization of Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), and recurrent architectures like Long Short-Term Memory (LSTM) networks.The survey may discuss the challenges associated with deepfake video detection, including the constant evolution of deepfake techniques and the need for diverse and annotated datasets for robust model training. It would likely cover the trade-off between computational complexity and real-time detection, addressing the resource constraints often faced in deploying these systems. Ethical considerations and privacy concerns related to the use of deepfake detection methods may also be highlighted, as striking a balance between technological advancement and responsible deployment is crucial.Furthermore, the survey is expected to provide insights into the state-of-the-art techniques for evaluating the authenticity of videos, potentially discussing metrics and benchmarks used to measure the performance of different detection models. The discussion may also touch upon the collaborative efforts in the research community to develop standardized evaluation protocols and countermeasures against emerging deepfake methods.

6. Conclusion

The survey on deepfake video identification and detection methods using Artificial Intelligence (AI) underscores the growing importance of addressing the challenges posed by the proliferation of manipulated multimedia content. As deepfake generation continues to advance, the development of robust and sophisticated detection techniques becomes imperative. This survey highlights a diverse array of AI-based approaches, including Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), and recurrent models like Long Short-Term Memory (LSTM) networks. While these techniques showcase promising results, the landscape of deepfake creation constantly evolves, necessitating ongoing research and adaptation of detection strategies. The comparative analysis reveals that the hybrid LSTM-CNN method is less effective for deepfake identification and video detection, with a lower accuracy rate of 82%. In contrast, the hybrid method combining CNN with the Jaya Algorithm significantly boosts accuracy to 99.3%, demonstrating superior performance in detecting deepfake videos. Furthermore, precision and recall results underscore the efficacy of these methods, with Hussain et al.'s approach using the CNN + Jaya algorithm achieving the highest precision rate, highlighting its accuracy in identifying true positives, though its recall rate, while robust, is slightly lower compared to some other techniques. The ethical implications of deepfakes, coupled with the potential consequences of their misuse, emphasize the urgency of developing comprehensive, accurate, and real-time detection mechanisms. Collaborative efforts between researchers, industry professionals, and policymakers are crucial to staying ahead in the ongoing arms race between deepfake generation and detection. The insights gleaned from this survey not only contribute to the current state of knowledge but also provide a roadmap for future endeavors aimed at fortifying our defenses against the ever-evolving challenges of deepfake threats in the digital age.

References

[1]. Kingra, Staffy, Naveen Aggarwal, and Nirmal Kaur. "Emergence of deepfakes and video tampering detection approaches: A survey." *Multimedia Tools and Applications* 82, no. 7 (2023): 10165-10209.

[2]. Westerlund, Mika. "The emergence of deepfake technology: A review." *Technology innovation management review* 9, no. 11 (2019).

[3]. Malik, Asad, Minoru Kuribayashi, Sani M. Abdullahi, and Ahmad Neyaz Khan. "DeepFake detection for human face images and videos: A survey." *Ieee Access* 10 (2022): 18757-18775.

[4]. Meskys, Edvinas, Julija Kalpokiene, Paul Jurcys, and Aidas Liaudanskas. "Regulating deep fakes: legal and ethical considerations." *Journal of Intellectual Property Law & Practice* 15, no. 1 (2020): 24-31.

[5]. Weerawardana, M. C., and T. G. I. Fernando. "Deepfakes detection methods: A literature survey." In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pp. 76-81. IEEE, 2021.

[6]. Maras, Marie-Helen, and Alex Alexandrou. "Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos." *The International Journal of Evidence & Proof* 23, no. 3 (2019): 255-262.

[7]. Bappy, Jawadul H., Cody Simons, Lakshmanan Nataraj, B. S. Manjunath, and Amit K. Roy-Chowdhury. "Hybrid lstm and encoder–decoder architecture for detection of image forgeries." *IEEE Transactions on Image Processing* 28, no. 7 (2019): 3286-3300.

[8]. Thakur, Rahul, and Rajesh Rohilla. "Copy-move forgery detection using residuals and convolutional neural network framework: a novel approach." In *2019 2nd International conference on power energy, environment and intelligent control (PEEIC)*, pp. 561-564. IEEE, 2019.

[9]. Su, Lichao, Cuihua Li, Yuecong Lai, and Jianmei Yang. "A fast forgery detection algorithm based on exponential-Fourier moments for video region duplication." *IEEE Transactions on Multimedia* 20, no. 4 (2017): 825-840.

[10]. Jia, Shan, Zhengquan Xu, Hao Wang, Chunhui Feng, and Tao Wang. "Coarse-to-fine copy-move forgery detection for video forensics." *IEEE Access* 6 (2018): 25323-25335.

[11]. Aloraini, Mohammed, Mehdi Sharifzadeh, Chirag Agarwal, and Dan Schonfeld. "Statistical sequential analysis for object-based video forgery detection." In *IS and T International Symposium on Electronic Imaging Science and Technology*, vol. 2019, no. 5, p. 543. 2019.

[12]. Guo, YanHui, Xue Ke, and Jie Ma. "A Face Replacement Neural Network for Image and Video." In *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, pp. 163-167. 2019.

[13]. Goodfellow, J. "Pouget-Abadie, M." *Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio,"Generative adversarial nets," in Advances in neural information processing systems* 27 (2014): 2672-2680.

[14]. Dai, Bo, Sanja Fidler, Raquel Urtasun, and Dahua Lin. "Towards diverse and natural image descriptions via a conditional gan." In *Proceedings of the IEEE international conference on computer vision*, pp. 2970-2979. 2017.

[15]. Chesney, Bobby, and Danielle Citron. "Deep fakes: A looming challenge for privacy, democracy, and national security." *Calif. L. Rev.* 107 (2019): 1753.

[16]. Delfino, Rebecca. "Pornographic deepfakes—revenge porn's next tragic act–the case for federal criminalization." *Available at SSRN* (2019).

[17]. Korshunova, Iryna, Wenzhe Shi, Joni Dambre, and Lucas Theis. "Fast face-swap using convolutional neural networks." In *Proceedings of the IEEE international conference on computer vision*, pp. 3677-3685. 2017.

[18]. Olszewski, Kyle, Zimo Li, Chao Yang, Yi Zhou, Ronald Yu, Zeng Huang, Sitao Xiang, Shunsuke Saito, Pushmeet Kohli, and Hao Li. "Realistic dynamic facial textures from a single image using gans." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5429-5438. 2017.

[19]. Vlasic, Daniel, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. "Face transfer with multilinear models." In *ACM SIGGRAPH 2006 Courses*, pp. 24-es. 2006.

[20]. Suwajanakorn, Supasorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. "Synthesizing obama: learning lip sync from audio." *ACM Transactions on Graphics (ToG)* 36, no. 4 (2017): 1-13.

[21]. Seraj, Md Shamim, Ankita Singh, and Shayok Chakraborty. "Semi-Supervised Deep Domain Adaptation for Deepfake Detection." In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1061-1071. 2024.

[22]. DHANARAJ, Rachel, and M. Sridevi. "Face Warping Deepfake Detection and Localization in a Digital Video using Transfer Learning Approach." *Journal of Metaverse* 4, no. 1 (2024): 11-20.

[23]. Elpeltagy, Marwa, Aya Ismail, Mervat S. Zaki, and Kamal Eldahshan. "A Novel Smart Deepfake Video Detection System." *International Journal of Advanced Computer Science and Applications* 14, no. 1 (2023).

[24]. Ikram, Sumaiya Thaseen, Shourya Chambial, and Dhruv Sood. "A performance enhancement of deepfake video detection through the use of a hybrid CNN Deep learning model." *International journal of electrical and computer engineering systems* 14, no. 2 (2023): 169-178.

[25]. Malik, Mubasher H., Hamid Ghous, Salman Qadri, Syed Ali Nawaz, and Anam Anwar. "Frequency-based Deep-Fake Video Detection using Deep Learning Methods." *Journal of Computing & Biomedical Informatics* 4, no. 02 (2023): 41-48.

[26]. Hussain, Zahraa Faiz, and Hind Raad Ibraheem. "Novel Convolutional Neural Networks based Jaya algorithm Approach for Accurate Deepfake Video Detection." *Mesopotamian Journal of CyberSecurity* 2023 (2023): 35-39.

[27]. Patel, Nimitt, Niket Jethwa, Chirag Mali, and Jyoti Deone. "Deepfake Video Detection using Neural Networks." In *ITM Web of Conferences*, vol. 44, p. 03024. EDP Sciences, 2022.

[28]. Ge, Shiming, Fanzhao Lin, Chenyu Li, Daichi Zhang, Weiping Wang, and Dan Zeng. "Deepfake video detection via predictive representation learning." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, no. 2s (2022): 1-21.

[29]. Ismail, Aya, Marwa Elpeltagy, Mervat Zaki, and Kamal A. ElDahshan. "Deepfake video detection: YOLO-Face convolution recurrent approach." *PeerJ Computer Science* 7 (2021): e730.

[30]. Cozzolino, Davide, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. "Id-reveal: Identity-aware deepfake video detection." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15108-15117. 2021.

[31]. Lewis, John K., Imad Eddine Toubal, Helen Chen, Vishal Sandesera, Michael Lomnitz, Zigfried Hampel-Arias, Calyam Prasad, and Kannappan Palaniappan. "Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning." In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1-9. IEEE, 2020.