Volume: 13 Issue: III Month of publication: March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67796

The Impact of Deepfakes on Digital Media Authenticity

Aditi Indre[1], Jayesh Shinde[2], Dr.Srivaramangai Ramanujam[3]
Department of Information Technology, University of Mumbai, Kalina, India

***Abstract**: With digital manipulation blurring the lines between reality and fabrication, deepfakes have become one of the most shocking threats to media authenticity and public trust. Fed by advanced artificial intelligence techniques like Generative Adversarial Networks (GANs) and deep learning, deepfakes can create hyper-realistic videos, images, and audio that convincingly imitate real people. These forgeries have far-reaching implications, from spreading political misinformation and financial fraud to non-consensual explicit content and identity theft. As deepfake technology becomes more sophisticated and accessible, traditional detection methods struggle to keep pace, sparking an urgent technological arms race between creators and defenders.*

This survey discusses the rapidly changing landscape of deepfake detection, focusing on multimodal approaches that combine visual, audio, and physiological cues to improve detection accuracy. Combining facial landmark analysis, voice consistency checks, and even heart rate variability extraction, multimodal detection methods provide a strong defense against increasingly complex forgeries. We discuss state-of-the-art models, including Convolutional Neural Networks (CNNs), Vision Transformers, and XGBoost classifiers, along with their evaluation against challenging datasets like Celeb-DF and FaceForensics++. Lastly, the paper emphasizes the problems in current detection techniques, especially in real- time processing and the generalization problems of deepfakes across a wide variety of generation methods. This paper helps researchers and practitioners navigate the battlefield of digital deception by identifying the critical gaps of existing research and proposing future directions.

Keywords: Deepfake detection, generative adversarial networks (GANs), multimodal detection, Vision Transformers, audio-visual analysis, facial forgery, FaceForensics++, Celeb-DF dataset, media authenticity, real-time detection, image forensics, machine learning.

I. INTRODUCTION

Deepfakes, aportmanteau of deep learning and fake content, have changed the way digital media is created and consumed over the past few years. Using advanced artificial intelligence techniques, such as Generative Adversarial Networks (GANs), deepfakes can swap faces convincingly, change voices, and manipulate expressions in videos and images [5]. The technology innovative but threatens media integrity, privacy, and public trust. Deepfakes have been used to create politically sensitive fake news, non-consensual explicit content, and even fraudulent financial communications, highlighting their potential for widespread harm [3][12].

The process of deepfake creation involves two neural networks: a generative network, which synthesizes fake media, and a discriminative network, which evaluates the authenticity of that media. This interlaps between the networks and makes it possible to produce highly realistic forgeries [6]. The most popular tools, such as FaceSwap and Face2Face, make the process easier for even non-experts to produce convincing fakes [21][22]. Democratization of synthetic media generation further worsens concerns about misinformation and digital identity theft.Although the misuse of deepfakes is scary, the entertainment, education, and accessibility aspects are significant [12]. For instance, in the film industry, deepfakes are used to de-age actors or bring back historical figures. In such positive applications, the challenge lies in the fact that as deepfake generation techniques advance, so must the detection methods [10].

This paper surveys the latest advancements in deepfake detection technologies. It covers a wide range of methodologies, from traditional machine learning classifiers like XGBoost [1] to advanced deep learning models such as Convolutional Neural Networks(CNNs) and Vision Transformers [6][13].The paper also discusses the multimodal detection approaches, which involve both visual and auditory cues [8][9]. In addition, the paper reviews key datasets like Celeb-DF and FaceForensics++ [3][22],which are critical for training and benchmarking detection models. This paper attempts to present a comprehensive overview of the current state of the deepfake detection landscape by discussing the current challenges it is facing, like real-time detection [6], model generalization [8], and ethical implications [44], and thus proposes future directions concerning its risks.

II. LITERATURE SURVEY

The function of deepfakes detection is light and understandable. With the combination of heart rate data, it can use the XGBoost classifier, that is very effective in distinguishing between a deepfake and real videos. It outperforms existing approaches on the World Leaders Dataset(WLDR)and has a high AUC score indicating strong classification, as obtained in[1] The Forgery Quality Score(FQS) aids in the selection of better training samples for the detection of deepfakes. It also utilizes Frequency Data Augmentation (FreDA) to create the reality of low-quality samples. Training samples were gradually increased in difficulty to aid models in learning and improving better detection of deepfakes. Detection performance has improved drastically across datasets[2]. The Celeb-DF dataset consists of 5,639 high-quality videos of celebrities that have outperformed the former datasets with poorer quality videos and even noticeable flaws. This dataset provides better visual quality, making the deepfakes more realistic. Tests on different detection methods using Celeb-DF showed that many existing techniques have trouble staying accurate, highlighting the need for better detection algorithms. [3] Deepfake detection models have been evaluated using audio-video datasets, but many have been found to overestimate performance due to biases like the leading silence in video data. This bias allows classifiers to rely on irrelevant features, raising concerns about the real-world applicability of these models. Some methods, like unsupervised learning and self-supervised techniques are being explored to improve robustness and generalization[4]Deepfake technology can pose risks such as misinformation and identity theft, it can also be used in film, education, and healthcare industries. It uses techniques like GANs, autoencoders and CNNs, used for generating deepfakes and detecting them. It also highlights tools like FakeApp and DeepFaceLab and the need for advanced detection methods[5]. GANs and Autoencoders, detection techniques using CNNs and Vision Transformers, imposes the challenges of real-time detection, model interpretability, and keeping up with evolving deepfake methods.[6] Deepfake detection in audio, especially singing voices, is becoming increasingly important. Traditional methods struggle to differentiate between real and synthetic voices. Whisper encodings, which are pre-trained speech representations, to improve detection accuracy. Whisper-based embeddings outperform older audio feature extraction methods, showing promise for deepfake audio authentication.[7]

Deepfake detection approaches can be segmented into machine learning, deep learning, and the hybrid approach. It describes advantages and disadvantages and challenges related to generalization along with adversarial attacks. Strengths of using multimodal cues (audio and visual) while detecting compared with single- modality models for getting better accuracy [8]. Importance of GANs in terms of producing nearly indistinguishable synthetic content from real ones. It points out the challenges in deepfake detection, especially the non generalizability of the current models and the ethical implications and societal impact of deepfake technology. The Incompatibility Between Multiple Modes (IBMM) and advocates for multi- modal detection approaches identified by Patel et al[9] Deep learning models like CNNs and Vision Transformers have shown effectiveness, although challenges persist in real time detection, generalization, and model interpretability. The paper by Croitoru et al emphasizes quality datasets for the training of such detection models.[10]It also reveals that the state-of-the-art face recognition systems like VGG and Facenet have high false acceptance rates, with 95%. The current methods of detection like audio-visual synchronization fail to differentiate between a deepfake video and a real one. Urgent need for better automated detection systems exists as deepfakes increasingly affect media and public perception according to Korshunov et al.[11] Deep neural networks (DNNs) and GANs generate very realistic fake content. This technology addresses the challenges deepfakes pose in the form of misinformation and manipulation, while considering the potential benefits it can offer in entertainment and marketing. Kietzmann et al introduces the R.E.A.L. framework to address the risks associated with deepfakes.[12] Local- & Temporal-aware Transformer- based Deepfake Detection (LTTD) framework, which enhances deepfake detection by focusing on local low-level cues and temporal information in videos. It improves detection accuracy across various datasets and is more robust to common post-processing methods. The framework uses innovative techniques like the Local Sequence Transformer (LST) and specialized loss functions to improve detection performance.[13] Self-Blended Images (SBIs) as synthetic training data to improve deepfake detection. SBIs aid models in identifying subtle inconsistencies present in deepfake content through the blurring of slightly altered versions of the same image. Models are trained with more accuracy and greater generalization, especially when faced with new or unseen types of manipulation.[14] Through the use of attention mechanisms in deepfake detection, models that incorporate attention layers are shown to exhibit strong improvement in terms of detection accuracy. These models outperform the traditional methods by focusing more on image-related features so that the manipulation of real material can be differentiated more effectively among various deepfaketechniques.[15]

DFREC is the latest developed method for recovering both source and target faces with maximum accuracy and fidelity from a deepfake image. It directly recovers both the faces from manipulated content as in existing algorithms. The scheme exploits identity segmentation that separates and recovers identities from the forged image.[16] The research approach proposed to identify deepfakes based on pose differences of head poses. Deepfakes generally have a higher difference in 3D head positions than the real images. Based on this, the approach uses an SVM classifier to correctly classify real and fake images.[17][18]Solopova at el Video Information Bottleneck Attribution (VIBA) method, which is an adaptation of the Information Bottleneck for Attribution (IBA) framework to video sequences. Unlike traditional explain ability methods that focus on static images, VIBA produces temporally and spatially consistent explanations for video classification, which is especially useful for tasks such as deepfake detection. The method uses Xception for spatial feature extraction and VGG11 for motion dynamics via optical flow.[19] The Face X-ray method detects face forgeries by focusing on blending boundaries in manipulated images, making it highly effective even against unseen manipulation techniques. It uses a greyscale representation that identifies discrepancies in blended areas, offering better generalization than traditional detection methods.[20] The Face2Face system introduces a real-time facial reenactment approach using monocular RGB video. It captures facial expressions from a live webcam feed and transfers them onto a target video, achieving high accuracy in facial identity recovery and expression transfer. The method outperforms offline techniques in both video quality and efficiency.[21]The FaceForensics++ framework uses deep learning methods, namely, convolutional neural networks (CNNs) and attention mechanisms, in order to detect manipulated facial images. Rössler et al set up a large- scale dataset of authentic and manipulated images for training, which was shown to be superior to the state of the art in terms of accuracy and generalization over various types of facial manipulations, like deepfakes.[22]

FaceForensics presents an extensive dataset targeted towards facial video forgery detection, especially in deepfakes. Cozzolino at el covers two types of video manipulations source-to-target reenactment and self-reenactment by benchmarking detection algorithms at different compression levels. The dataset employs advanced reconstruction of 3D models and image-based rendering techniques to become a highly effective resource for improving digital forensics while helping fight misinformation. [23]FaceShield is a defense mechanism that uses proactive protection towards facial images being manipulated by deepfakes. It works in the way it disrupts deepfake generation either by manipulating facial features, going after feature extraction models, and using adversarial perturbations. Techniques of Gaussian blur and low-pass filtering are used, making the defense imperceptible, achieving the state-of-art results in both robustness and imperceptibility against various models of deepfakes.[24]FaceShifter is a system with two steps involved in better face swapping. The Adaptive Embedding Integration Network known as AEI-Net will generate high- quality face images while the Heuristic Error Acknowledging Refinement Network referred to as HEAR-Net will refine these images with masks fixed on hidden areas. This method improves face swapping by maintaining realistic results and identity preservation, without needing manual input, offering more natural outcomes than older techniques.[25] FaceProtect, a proactive deepfake detection method that uses dynamic watermarks based on facial features leverages facial characteristics to create dynamic watermarks, enhancing detection accuracy. It makes use of GAN- based One-way Dynamic Watermark Generating Mechanism (GODWGM) and a Watermark-based Verification Strategy (WVS) for embedding and recovering watermarks without compromising the visual integrity.[26]FRIDAY, a novel approach to deepfake detection in that it corrects the unintentional learning of facial identity features by detection models. Most of the existing detectors for deepfakes focus on facial identities and have poor generalization across datasets. FRIDAY uses a face recognizer that trains detectors to focus on synthetic artifacts rather than facial identities to improve detection accuracy and generalization both in-domain and cross-domain datasets. [27]

The Face Security Foundation Model (FSFM) leverages self-supervised learning to create robust facial representations, improving face security tasks like deepfake detection and anti-spoofing. FSFM performs by demonstrating superior generalization across various datasets, learning from unlabeled data without requiring large annotated datasets. This model is designed to address the limitations of existing methods in handling diverse facial manipulations.[28] This study presents a new method for detecting audio deepfakes, utilizing Latent Space Refinement (LSR) and Latent Space Augmentation (LSA). These techniques aim to enhance generalization and improve detection accuracy across multiple datasets, outperforming current state-of-the-art detection methods. By refining and augmenting the latent space of spoofed audio, the approach captures a broader range of deepfake patterns..[29] A hybrid of local and global frequency domain features with deepfake detection that helps improve accuracy. It exploits the properties of both types of features to help it better detect manipulations that are inaudible to the human ear, overcoming that limitation of its predecessors. The proposed method increases the accuracy by 2.9% over the best currently existing techniques.[30] GP-GAN: A framework that combines gradient-based methods with GANs to create high-resolution image blends. It offers the advantage of superior color consistency in the merged images with less number of artifacts as compared to other earlier techniques. The Gaussian- Poisson equation it uses holds on to real details and thus, improves the realism of composite images.[31] FacePoison is a defense framework that disrupts face detection and prevents DeepFake video generation. It does this by applying small, imperceptible changes to images that interfere with face detection systems. This weakens the training and operation of DeepFake models, effectively sabotaging face detectors essential for creating DeepFakes.[32] Zhang et al introduces a new method for detecting deepfake audio by improving the model's resilience to different attacks and audio distortions. Using the DeepFakeVox-HQ dataset, which is thelargest public voice dataset, and Frequency-Selective Adversarial Training (F-SAT), this approach focuses on high-frequency audio components to improve detection accuracy. The results are that this method outperforms previous models, especially in difficult scenarios.[33]

A novel method for detecting deepfake videos by analyzing eye blinking patterns, which are often unrealistic or absent in AI-generated content. The approach leverages natural physiological signals that are typically overlooked by deepfake creation technologies, aiming to provide a more reliable detection solution.[34] Chung et al presents a new approach using the "Watch, Listen, Attend and Spell" (WLAS) network, trained on the large "Lip Reading Sentences" (LRS) dataset, to transcribe video of mouth movements into characters.[35] The Markov Observation Model (MOM) for deepfake detection outperforms methods like GANs, SVMs, and particle filtering, demonstrating better accuracy and consistency in distinguishing real from fake sequences, including those created by GANs or simulators.[36] The multi- attentional deepfake detection method surpasses traditional binary classification approaches. This method focuses on identifying subtle and local differences between real and fake images, achieving impressive results on datasets like FaceForensics++, Celeb-DF, and DFDC.[37]Yan et al deals with the overfitting problem in AI-generated image detection where traditional methods usually fail to detect new, unseen fakes. Effort employs orthogonal subspace decomposition to enhance generalization and outperforms existing detection methods on several benchmarks.[38] multi-modal approaches in deepfake detection, showing the limitations of current passive methods that focus only on one type of media (such as images, videos, or audio). These approaches are not flexible enough to be applied to various generative models and fail to generalize well. Multi-modal approaches are considered more robust and adaptable to the changing landscape of deepfaketechnologies.[39] SPADE (Spatially-Adaptive Normalization), a new approach that improves image synthesis from semantic layouts. SPADE preserves semantic information during the normalization process, leading to more realistic and accurate images. [41]

Zang et al. introduce the SingFake dataset, the first one that includes both real and deepfake singing clips in different languages, placing importance on specialized methods for detecting deepfake singing voices.[42] Liu et al. introduce Spatial-Phase Shallow Learning (SPSL), a novel face forgery detection method that focuses on the phase spectrum in the frequency domain. Unlike traditional methods, which use amplitude spectrum data, SPSL is more sensitive to up-sampling artifacts and enhances detection accuracy.[43] Deepfake technology is on the rise, carrying with it both potential dangers of misinformation and lost trust, as well as potential benefits. This paper reviews deepfakes' societal impacts on politics, business, and privacy, suggesting solutions through detection, laws, and public awareness.[44] F3-Net: A Novel Approach for Face Forgery Detection Based on Frequency-Aware Clues for Detecting Subtle Forgery Artifacts in Low-Quality Media. It practically enhances detection and surpasses traditional methods that depend solely on the RGB domain by using frequency decomposition and local frequency statistics. [45] Dynamic Facial Forensic Curriculum (DFFC), a new approach to improving deepfake detection by focusing on harder samples during training. DFFC enables the model to learn better by gradually introducing challenging samples, thereby enhancing its ability to generalize across different datasets and detect deepfakes more effectively. [46] Yan et al introduce Latent Space Data Augmentation (LSDA) to further the improvement of the generalization of deepfake detection models. LSDA simulates variations in the latent space to help the model adapt to new and unseen forgery types that reduce overfitting to specific artifacts.[47]The "arms race" between deepfake generation and detection technologies. As generation techniques, such as GANs, improve, detection methods struggle to keep up, creating a constant need for innovation on both sides.[48] FakeSTormer, a method for detecting deepfake videos by focusing on both spatial and temporal vulnerabilities. It uses a multi- task learning framework that captures subtle features across video frames and provides interpretability with a focus on the areas where manipulation could take place.[49] WildDeepfake dataset. The dataset of real-world deepfake videos crawled from the internet. It puts forward how more sophisticated, realistic deepfakes can evade state-of-the-art detection models which were trained under controlled datasets. The dataset thus helps in making improvements in more robust detection methods. [50]

III. OBSERVATION

The lightweight design allows for quick identification of deepfakes without the heavy computational load typical of many deep learning models. Additionally, the XGBoost classifier not only performs well but also provides clear insights into its decision-making process, making it easier to interpret, which is effective at detecting deepfakes at both the frame and segment levels, showcasing its versatility. [1] The FQS is an important aspect that gives some good insights in sample difficulty, which is necessary in the construction of robust detection models. The FreDA technique puts forward a vital point concerning the significance of frequency processing, which standard methods seldom pay heed to. It impressively enhances the framework in terms of detection capability and might be useful for real-world multimedia forensics applications.[2].Current deepfake detection methods often rely on visual artifacts, which are less frequently observed in high-quality deepfakes, thus causing poor performance on the Celeb-DF dataset. Superior visual quality of the dataset is measured by metrics like Mask-SSIM, posing a major challenge for the detection technique, hence something more robust and adaptable is required. [3] Unsupervised learning method AVH-Align uses only real data to avoid exploiting bias. This method improves detection accuracy by aligning audio and visual features without relying on shortcuts.[4].The deepfake creation algorithms are becoming more sophisticated, making detection challenging. As these technologies grow, the risks of misuse, particularly in politics and social media, increase. Mahmud at el stresses the need for better detection techniques, regulation, and public awareness to minimize harm while leveraging the technology’s benefits.[5] Deepfakes are becoming increasingly realistic, making detection harder. While current methods like CNNs show promise, they struggle with real-time processing and generalizing across different types of manipulated content. High-quality datasets are essential for improving detection accuracy. [6]

A milestone in audio detection arrived with Whisper encodings, which have greatly enhanced the accuracy of detection of artificial singing voices. Though this technique has demonstrated great results on multiple datasets, it leaves scope for further optimization to better suit singing voice detection. [7] The present state of detection models is found to have excellent performance in laboratory settings but has difficulty detecting unknown or adversarial deepfakes. Although CNNs and RNNs are promising, their data-intensive nature and susceptibility to attacks are still a cause for concern. Multimodal methods that integrate audio and visual data are more reliable but are challenged in wider usage. [8] Deepfake generation has been improving significantly, and detection methods seem to be not catching up. It makes use of multimodal approaches like visual and physiological cues; yet, these are not very robust and generalize to other datasets. Interpretability also becomes essential when it comes to forensic use cases. [9][10] The fast growth of deepfake synthesis has exceeded detection strategies. Although visual-physiological indicator-based approaches are promising, they usually suffer from applicability in more general scenarios across various datasets. Interpretability becomes especially important in forensic use cases [9][10] The increasing level of sophistication in high- quality deepfake videos continues to present considerable challenges for face recognition systems. While detection techniques based on image quality and machine learning have advanced,they're still playing catch-up with increasingly sophisticated fakes. This creates grave concerns about media credibility and public trust, underscoring the need for more effective countermeasures [11][12].

Conventional detection techniques tend to fail when confronted with diverse generationmethods and post-processing manipulations. The LTTD model presents a new direction byfocusing on local and temporal patterns, thereby being less vulnerable to distortions. Likewise,the use of SBIs has improved model robustness by facilitating the learning of more generalfeatures [13][14]. The incorporation of attention mechanisms, and more specifically, self-attention and multi-head attention, has been useful to enable models to concentrate on relevant image areas for improved feature extraction and cross- dataset performance [15].DFREC has used an Identity Segmentation Module (ISM) to separate the face information and reconstruct source and target faces with specialized modules that improve the quality of the recovery process.[16] 3D head pose estimation based on facial landmarks shows profound differences between deep fake and realistic images. Deep fakes reflect larger mismatches in head poses, while for real images differences are smaller based on the use of cosine distances. The presented SVM classifier results in high detection accuracy for differences between the used types of image, especially its performance on UADFV test set. Cosine distances calculation for these differences seems to provide fairly good results, indicating that this approach could be useful for the detection of deepfakes.[17][18]VIBA produces relevance and optical flow maps that point to manipulated regions in videos and points of motion discontinuity, leading to interpretability without sacrificing any model accuracy. The addition of VIBA does not incur much loss on the model as well, since the explanations so produced are compatible with human labeling, showing in excess of over 50 percent overlap in explained regions.[19]

Face X-ray outshines other available forgery-detection methods: it maintains maximum accuracy even at the face of new, unheard-of manipulation tricks. It does not rely on specific artifacts, allowing for broader application and better generalization. [20] A non-rigid model- based approach is usedto align facial expressions from the source to the target. With GPU based optimization, it delivers real-time performance even on standard hardware. This approach works well for various applications, though it faces challenges in low-light conditions and occlusions due to its reliance on RGB input.[21] RGB reenactment method identifies subtle artifacts and inconsistencies in facial images, demonstrating strong performance in detecting manipulation. The use of attention mechanisms and a custom dataset enhances the model's ability to focus on crucial image areas, improving detection results. However, continuous updates to the dataset and model are necessary to handle evolving manipulation techniques.[22] FaceShield defends against deepfakes while preserving the quality of the original image. It relies on attention mechanisms and adversarial perturbations, but it requires updates to fight new deepfake techniques such as cropping.[24] FaceShifter boosts the realism of face- swapped images by overcoming occlusions and incorporating target attributes effectively. While it has high performance, its potential limitations might be in the handling of extreme facial occlusions or lighting conditions, and performance may vary between different demographics and face pairs.[25] FaceProtect shows higher accuracy, precision, and F1-score compared to existing state-of-the-art methods of deepfake detection. Its method makes the identification of real vs. manipulated images at least more reliable through dynamic watermarks that depend on facial features. However, the method presumes that the watermarks will still be maintained after image processing or manipulation, thus having room for improvement concerning advanced deepfake techniques. [26]

FRIDAY follows a two-phase training process that greatly reduces the reliance on facial identity features in deepfake detection. Experimental results demonstrate the superiority of FRIDAY over state-of-the-art deepfake detection methods and present better adaptability and performance in both in-domain and cross- domain settings. However, the study does not address some of the possible limitations of the face recognizer's performance in real- world applications or changing techniques in developing deepfakes.[27] FSFM uses masked image modeling and instance discrimination to learn effective facial features, achieving significant improvements in detection performance on ten public datasets. Its major strength is that it generalizes well across different face manipulation tasks, but may not perform as well with new deepfake techniques or scenarios that have significant facial variations or occlusions. [28] The integration of LSR and LSA provides learnable prototypes for spoofed audio along with augmentation applied within the latent space, which increases the model's potential to achieve better generalization by applying variations of deepfake audio. The results exhibit competitive performance on four datasets with superior robustness to existing methods.[29] Integration of Discrete Wavelet Transform for local features and Fast Fourier Transform for global features with the attention mechanisms ensures more accurate detection of deepfakes. In particular, this method has high performance over datasets generated by a variety of models, proving cross-domain robustness. Moreover, its plug-and-play design is easy to combine with the CNN classifiers used by existing works. [30] GP-GAN works in two stages. Firstly, it applies a low resolution to an image using a Blending GAN. It then enhances the image at higher resolutions with a special math equation referred to as the Gaussian- Poisson Equation. The approach has greatly improved the outcome of image blending, and experiments show users prefer the result of GP-GAN over the traditional methods.[31]FacePoison uses methods that can tamper with both face detection and DeepFake creation. VideoFacePoison upgrades this by modifying frames in a video; it is therefore able to enhance the outcome. It has been tested on five face detectors and eleven DeepFake models. This significantly degrades the quality of DeepFakes. The proposed method therefore undermines the training and use of DeepFake systems, providing good protection.[32] Voice augmentations and adversarial examples focusing on high-frequency audio components, enabling the model to effectively differentiate between real and fake audio. Experimental results reveal that the new approach significantly improves detection accuracy, particularly when dealing with out-of-distribution samples and corrupted audio.[33]

Convolutional Neural Networks (CNN) and Long-term Recurrent Convolutional Networks (LRCN) to analyze video frames and detect eye blinking. Focusing on temporal dynamics, the system was able to differentiate between real and fake videos, with a higher accuracy rate and fewer false positives than in traditional detection methods.[34] The WLAS network, which includes curriculum learning, is able to perform better than all the previous lip reading methods. It even surpasses all previous benchmarks and also outperforms a professional lip reader on videos from BBC television. The model shows that visual information can assist in speech recognition, even in the presence of audio.[35] MOM was tested against a coin-flip dataset and proved to be more reliable than GANs and SVMs, which failed to reliably identify fake content. The paper shows that GANs and SVMs are less reliable across different types of fake sequences.[36] Effort enhances detection by separating the feature space into two orthogonal subspaces: one for pre-trained knowledge and another for forgery patterns. This reduces overfitting and boosts generalization to detect both known and unseen fakes effectively.[38] The survey categorizes detection methods into various types like forensic based and data-driven techniques. It finds that while many methods perform well, they lack generalization and fail to handle multi- modal deepfakes (those manipulating both audio and visual components).[39] By using an adaptive phoneme pooling technique and a graph attention network (GAT), the proposed method successfully captures discrepancies between real and synthetic speech. This results in improved detection accuracy by learning temporal relationships in phoneme features.[40]

SPADE works by adjusting normalization based on semantic input, using a generator architecture with ResNet blocks and a multi-scale discriminator. Experiments on datasets like COCO-Stuff and Cityscapes show that SPADE outperforms existing methods in both quantitative and qualitative measures.[41] By retraining existing speech detection systems on the SingFake dataset, the authors find significant improvements in detection performance. However, these systems still struggle with unseen singers, languages, and complex musical contexts, especially when background music is involved.[42] SPSL combines phase spectrum analysis with convolutional neural networks (CNNs) to capture local texture information, improving detection performance across various datasets. The method shows strong results in cross- dataset evaluations and multi-class classification tasks, outperforming existing methods.[43] The research, based on 84 online articles, finds that deepfakes are often used maliciously for political manipulation and personal harm. However, they can also serve legitimate purposes like entertainment. The study stresses the need for improved detection methods and better public understanding of deepfakes.[44] F3- Net achieves superior performance in detecting forgery on datasets like FaceForensics++, especially under different compression qualities. The use of frequency-aware decomposition and local statistics helps uncover subtle manipulation patterns that typical methods might miss, showing significant improvements in accuracy.[45] The experiments show that DFFC significantly boosts the performance of deepfake detection models, both within the same dataset and across new ones. The strategy of prioritizing difficult samples during training results in improved accuracy and robustness, as it helps the model learn more discriminative features of deepfakes.[46] Experiments show that LSDA significantly enhances the detection performance of deepfake models across multiple datasets, making them more robust and accurate. The method helps in learning generalizable features by using both real and augmented forgery data.[47] As the generation techniques get more advanced, detection methods also play catch-up, with the use of deep learning and feature extraction. The robustness and generalizability of the models are lacking, and there is a lack of diverse datasets in detection models that hampers their real-world effectiveness.[48]FakeSTormer is better than other detection techniques. It shows improved generalization and interpretability, and it makes use of novel techniques such as Self- Blended Videos and TimeSformer to improve its detection capability. However, there are challenges in detecting new manipulation techniques and improving interpretability in complex cases.[49] Baseline detection models showed a significant drop in performance when tested on the WildDeepfake dataset, confirming that real-world deepfakes are much harder to detect. ADDNets, proposed in this paper, showed better performance than the baseline detection models, thereby demonstrating the advantages of using attention mechanisms to focus on key facial features.[50]

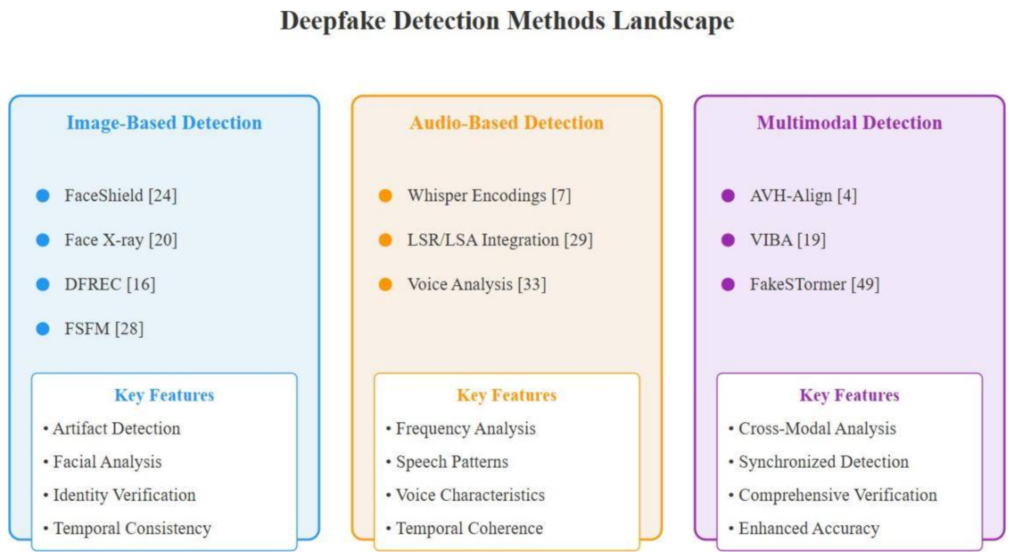

Fig. 1

IV. CONCLUSION

In summary, although deepfake detection research has achieved great milestones, more gaps still need to be filled in order to really respond to the ever-evolving complexity of deepfake technologies. The state-of-the-art methods, which integrate heart rate analysis, facial feature recognition, attention mechanisms, etc., have promising roles in improving detection accuracy and robustness; these are some good generalization frameworks with a capability for real-world adaptability. Other quality-oriented approaches, such as Forgery Quality Scores and synthetic data generation, have shown promise in boosting model performance to ensure that the detection methods are more robust on diverse datasets.

However, several key advancements need to be achieved to bridge the remaining gaps. First, more sophisticated and novel deepfake techniques are challenging to detect. Most of the models fail to generalize across new manipulation techniques, especially when it is unseen data or more sophisticated forgeries. Hence, adaptive methods such as adversarial training and dynamic curriculum learning are required in order to keep updating the model and improving the resilience against emerging threats. Better integration of multi-modal detection approaches that combine facial, audio, and behavioral cues will likely be essential for tackling a wide range of deepfake content. Although tremendous progress has been achieved in the accuracy of detection, the next steps would involve ensuring that such models are interpretable and transparent. Most current methods are very complex and not clear enough to be used in real-world applications, especially those in legal or forensic applications. Detection systems should be interpretable, explainable, and adaptive to new manipulation techniques for them to be useful in the long run.

Finally, there is an urgent need to increase the number of real-world datasets, particularly those that capture diverse and natural scenarios, as most of the existing datasets tend to focus on specific types of forgeries or controlled environments. Improving the diversity of the dataset and the use of real-time data will make models more adaptable and accurate in practical situations.

In summary, deepfake detection has made tremendous progress. However, the technology itself continues to evolve, and there is still much to be discovered on developing adaptable, interpretable, and real-time solutions. Coordinative work across research, technology, and policy spheres will be crucial towards bridging the identified gaps so that detection systems stay up with the evolving nature of deepfake technology.

REFERENCES

[1] Muhammad Umar Farooq, Ali Javed, Khalid Mahmood Malik, Muhammad Anas Raza “A Lightweight and Interpretable Deepfakes Detection Framework”, arXiv:2501.11927, 2025

[2] Wentang Song, Zhiyuan Yan, Yuzhen Lin, Taiping Yao, Changsheng Chen, Shen Chen, Yandan Zhao, Shouhong Ding, Bin Li, “A Quality-Centric Framework for Generic Deepfake Detection”, arXiv:2411.05335, 202425

[3] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, SiweiLyu “Celeb-DF- A Large-scale Challenging Dataset for DeepFake Forensics”, , arXiv:1909.12962, 2019.

[4] Dragos-AlexandruBoldisor, Stefan Smeu, Dan Oneata, ElisabetaOneata “Circumventing shortcuts in audio-visual deepfake detection datasets with unsupervised learning”, arXiv:2412.00175, 2024.

[5] Bahar Uddin Mahmud, AfsanaSharmin, “Deep Insights of DeepfakeTechnology : A Review”, arXiv:2105.00192, 2021

[6] Amal Naitali, Mohammed Ridouani , Fatima Salahdineand Naima Kaabouch , “Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions”, In MDPI, Computers 2023.

[7] Falguni Sharma, Priyanka Gupta “Deepfake Detection of Singing Voices With Whisper Encodings”, arXiv:2501.18919v1, 2025

[8] Md ShoRana, Mohammad Nur Nobi, Beddhu Murali, Andrew H. Sung , “Deepfake Detection: A Systematic Literature Review” , In IEEE, Jan-2022

[9] Yogesh Patel, Sudeep Tanwar, Rajesh Gupta, Pronaya Bhattacharya, Innocent Ewean Davidson, RoyiNyameko “Deepfake Generation and Detection Case Study and Challenges ”, In IEEE December 2023 , Digital Object Identifier 10.1109/ACCESS.2023.3342107

[10] Florinel-Alin Croitoru, Andrei-Iulian Hiji, Vlad Hondru, Nicolae CatalinRistea, Paul Irofti, Marius Popescu, Cristian Rusu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Mubarak Shah “Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook”, arXiv:2411.19537, Nov 2024.

[11] Pavel Korshunov, Sebastien Marcel , “DeepFakes: a New Threat to Face Recognition? Assessment and Detection”, arXiv:1812.08685, Dec 2018

[12] Jan Kietzmann, Linda W. Lee , Ian P. McCarthy, Tim C. Kietzmann , “Deepfakes: Trick or Treat?”, Business Horizons, Vol. 63, 2020

[13] Jiazhi Guan, Hang Zhou, Zhibin Hong, Errui Ding, Jingdong Wang, Chengbin Quan, Youjian Zhao , “Delving into Sequential Patches for Deepfake Detection”, arXiv:2207.02803 , 2022

[14] Kaede Shiohara Toshihiko Yamasaki, “Detecting Deepfakes with Self-Blended Images”, In IEEE, June 2022 , DOI: 10.1109/CVPR52688.2022.01816

[15] MD Sadik Hossain Shanto, Mahir Labib Dihan, Souvik Ghosh, Riad Ahmed Anonto, Hafijul Hoque Chowdhury, AbirMuhtasim, Rakib Ahsan, MD Tanvir Hassan, MD RoqunuzzamanSojib, Sheikh Azizul Hakim, M. Saifur Rahman , “DFCon: Attention-Driven Supervised Contrastive Learning for Robust Deepfake Detection”, arXiv:2501.16704 , Jan 2025

[16] Peipeng Yu, Hui Gao, Zhitao Huang, Zhihua Xia, Chip-Hong Chang “DFREC: DeepFake Identity Recovery Based on Identity-aware Masked Autoencoder”, arXiv:2412.07260, Dec 2024

[17] Xinghe Fu, Zhiyuan Yan, Taiping Yao, Shen Chen, Xi Li, “Exploring Unbiased Deepfake Detection via Token-Level Shuffling and Mixing”, arXiv:2501.04376v1, Jan 2025

[18] Xin Yang, Yuezun Li, SiweiLyu, “Exposing Deep Fakes Using Inconsistent Head Poses”, arXiv:1811.00661, 2018

[19] Veronika Solopova, Lucas Schmidt, Dorothea Kolossa, “Extending Information Bottleneck Attribution to Video Sequences”, arXiv:2501.16889, Jan 2025

[20] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, Baining Guo, “Face X-ray for More General Face Forgery Detection”, arXiv:1912.13458, Apr 2020.

[21] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, Matthias Nießner“Face2Face: Real-time Face Capture and Reenactment of RGB Videos”, arXiv:2007.14808, Jul 2020

[22] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner “FaceForensics++: Learning to Detect Manipulated Facial Images”, arXiv:1901.08971, Jan 2019

[23] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner , “FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces”, arXiv:1803.09179 , Mar 2018

[24] JaehwanJeong, Sumin In, Sieun Kim, Hannie Shin, JongheonJeong, Sang Ho Yoon, Jaewook Chung, Sangpil Kim, “FaceShield: Defending Facial Image against Deepfake Threats” rXiv:2412.09921, Dec 2024

[25] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, “FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping”, arXiv:1912.13457, Dec 2019

[26] Shulin Lan, Kanlin Liu, Yazhou Zhao, Chen Yang, Yingchao Wang, Xingshan Yao, Liehuang Zhu, “Facial Features Matter: a Dynamic Watermark based Proactive Deepfake Detection Approach”, arXiv:2411.14798, Nov 2024

[27] Younhun Kim, Myung-Joon Kwon, Wonjun Lee, Changick Kim, “FRIDAY: Mitigating Unintentional Facial Identity in Deepfake Detectors Guided by Facial Recognizers”, In IEEE Xplore January 2025, DOI: 10.1109/VCIP63160.2024.10849915

[28] Gaojian Wang, Feng Lin, Tong Wu, Zhenguang Liu, Zhongjie Ba, Kui Ren, “FSFM: A Generalizable Face Security Foundation Model via Self-Supervised Facial Representation Learning” , arXiv:2412.12032 , Dec 2024

[29] Wen Huang, Yanmei Gu, Zhiming Wang, Huijia Zhu, Yanmin Qian, “Generalizable Audio Deepfake Detection via Latent Space Refinement and Augmentation”, arXiv:2501.14240 , Jan 2025

[30] Jiazhen Yan, Ziqiang Li, Ziwen He, Zhangjie Fu, “Generalizable Deepfake Detection via Effective Local-Global Feature Extraction”, arXiv:2501.15253, Jan 2025

[31] Huikai Wu, Shuai Zheng, Junge Zhang, Kaiqi Huang, “GP-GAN: Towards Realistic High-Resolution Image Blending” ,arXiv:1703.07195 , Mar 2017

[32] Yuezun Li, Xin Yang, Baoyuan Wu, SiweiLyu, “Hiding Faces in Plain Sight: Disrupting AI Face Synthesis with Adversarial Perturbations”, arXiv:2412.01101, Dec 2024

[33] Zirui Zhang, Wei Hao, AroonSankoh, William Lin, Emanuel Mendiola-Ortiz, Junfeng Yang, Chengzhi Mao, “I Can Hear You: Selective Robust Training for Deepfake Audio Detection”, arXiv:2411.00121v1, Oct 2024

[34] Yuezun Li, Ming-Ching Chang, SiweiLyu, “In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking”, Published in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), January 2019

[35] Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman, “Lip Reading Sentences in the Wild”, Published in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017

[36] Jyoti Bhadana, Michael A. Kouritzin, Seoyeon Park, Ian Zhang, “Markov Processes for Enhanced Deepfake Generation and Detection”, arXiv:2411.07993, Nov 2024

[37] Hanqing Zhao, Tianyi Wei, Wenbo Zhou, Weiming Zhang, Dongdong Chen, Nenghai Yu, “Multi-attentional Deepfake Detection”, Published in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), November 2021, DOI: 10.1109/CVPR46437.2021.00222

[38] Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen,Taiping Yao, Shouhong Ding, Baoyuan Wu, Li Yuan, “Orthogonal Subspace Decomposition for Generalizable AI-Generated Image Detection”, arXiv:2411.15633v2, Jan 2025

[39] Hong-Hanh Nguyen-Le, Van-Tuan Tran, Dinh-Thuc Nguyen, Nhien-An Le-Khac, “Passive Deepfake Detection Across Multi-modalities: A Comprehensive Survey”, arXiv:2411.17911, Nov 2024

[40] Kuiyuan Zhang, Zhongyun Hua, Rushi Lan, Yushu Zhang, Yifang Guo, “Phoneme-Level Feature Discrepancies: A Key to Detecting Sophisticated Speech Deepfakes”, arXiv:2412.12619, 17 Dec 2024

[41] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, Jun-Yan Zhu, “Semantic Image Synthesis with Spatially-Adaptive Normalization”,arXiv:1903.07291, Mar 2019

[42] Yongyi Zang, You Zhang, MojtabaHeydari, ZhiyaoDuan,” SingFake: Singing Voice Deepfake Detection”, arXiv:2309.07525, Sep 2023

[43] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, Nenghai Yu , “Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain”, arXiv:2103.01856 , Mar 2021

[44] Mika Westerlund, “The Emergence of DeepfakeTechnology:A Review”, Technology Innovation Management Review, November 2019

[45] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, Jing Shao, “Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues”, arXiv:2007.09355 , Jul 2020

[46] Wentang Song, Yuzhen Lin, Bin Li, “ Towards General Deepfake Detection with Dynamic Curriculum ”, arXiv:2410.11162 , Oct 2024

[47] Zhiyuan Yan, Yuhao Luo, SiweiLyu, Qingshan Liu, Baoyuan Wu , “Transcending Forgery Specificity with Latent Space Augmentation for Generalizable Deepfake Detection”, arXiv:2311.11278, Nov 2023 \

[48] Hannah Lee, Changyeon Lee, Kevin Farhat, Lin Qiu, Steve Geluso, Aerin Kim, Oren Etzioni, “The Tug-of-War Between Deepfake Generation and Detection”, arXiv:2407.06174 , Jul 2024

[49] Dat Nguyen, Marcella Astrid, Anis Kacem, EnjieGhorbel, DjamilaAouada, “Vulnerability-Aware Spatio-Temporal Learning for Generalizable and Interpretable Deepfake Video Detection”, arXiv:2501.01184, Jan 2025

[50] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, Yu-Gang Jiang, “WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection”, arXiv:2101.01456 , 5 Jan 2021