Review on Detection of Deepfake in Images and Videos

Yaramasa Gautham
Department of ECE
AMC Engineering College
Bangalore, India

Sindhu R.
Department of ECE
AMC Engineering College
Bangalore, India

Jenitta J.
Department of ECE
AMC Engineering College
Bangalore, India

ABSTRACT
Deepfake technology can manipulate and superimpose existing images or videos onto other images or videos, creating realistic-looking but fabricated content. This technology has raised concerns as it can be used to create deceptive or misleading media, potentially causing harm by spreading false information or manipulating public perception. A detailed review is done on the detection of Deepfake in images and videos and it is presented in this paper. Various methods with which the detection of deepfake can be performed are image-based, video-based, frequency-based, Machine learning algorithm-based and Generative Adversarial-based methods. Various databases used, advantages and drawbacks of each literature are discussed in detail. After thorough research, it was found that the Attentive-pooling methods are giving better results than all the other methods that were proposed in the literature.

Keywords
Attentive pooling, Deep-fake, Deep Learning, Generative Adversarial Network, Machine Learning

1. INTRODUCTION

Deepfake technology represents a cutting-edge facet of artificial intelligence and machine learning that has garnered significant attention recently. Deepfake technology gained prominence around the mid-2010s with the growing availability of powerful graphical processing units (GPUs) and large datasets. The term "deepfake" combines "deep learning" and "fake," reflecting its reliance on sophisticated neural networks. Deepfake technology has been used in various contexts. In the entertainment industry, it has been employed for creating special effects and realistic visualizations in movies and TV shows. Additionally, deepfakes have been used for impersonations, allowing actors to portray historical figures or celebrities convincingly. However, deepfakes can also be misused for malicious purposes. They have been used to create fake news, spread misinformation, defame individuals, or even engage in fraud.

The significance of deepfake technology lies in its ability to generate highly realistic and difficult-to-detect fake media. This technology poses risks to society, including the erosion of trust in digital content, the potential for political manipulation, and the threat to personal privacy and security. As deepfake technology advances, it becomes imperative to develop robust detection methods, regulations, and ethical guidelines to address these challenges. It involves swapping one person's face onto another's, making it seem like the latter is doing or saying things they never did. Some people misuse deepfakes to spread false information, manipulate public opinion, or conduct deceptive campaigns, leading to ethical and societal concerns. These issues include the need for consent, protection of privacy, and addressing the spread of misinformation, and they are central to ongoing discussions about the responsible use of this technology.

To tackle the problems caused by deepfake technology, deepfake detection methods have been developed. These aim to identify manipulated content by analyzing inconsistencies in facial movements, audio, or other patterns indicative of digital alteration. Deepfake detection is pivotal in preserving trust and combating misinformation. Techniques for detection include forensic analysis, deep learning, and digital watermarks to verify content authenticity. While detection methods continue to evolve, they represent a crucial defense against the deceptive use of deepfakes.

Deepfake detection is classified into many different methods as it is an evolving field that approaches continuous development to combat increasingly sophisticated deepfake techniques. The common classifications are (i) Image-based methods (ii) Video-based methods (iii) Machine learning algorithm based (iv) Generative adversarial network(GAN) based (v) Frequency-based. Fig. 1 shows the classification of deepfake detection methods. Fig. 2 shows the classification of deepfake detection methods with reference paper numbers.

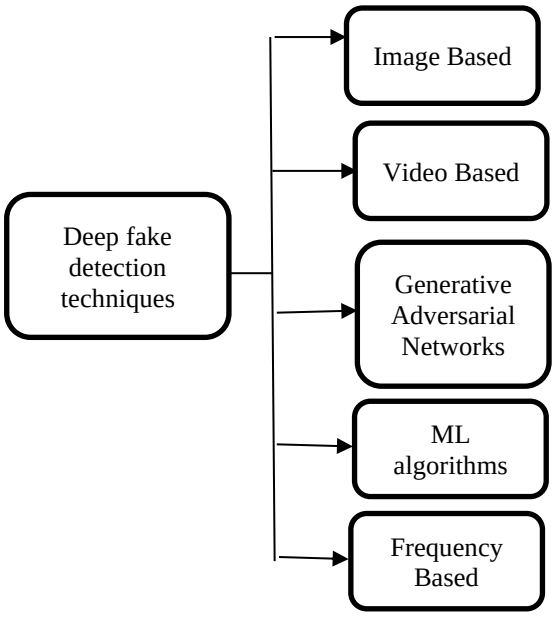

Figure 1: Classification of Deep fake detection techniques

In[1] The authors proposed Deepfake face detection using Deep Inception Net, using its sharp eye for details to catch subtle tricks in videos and images. This helps protect against sneaky attempts to deceive, making sure digital content stays trustworthy in the age of fake media. This process starts by gathering a dataset containing both real and deepfake images.

Image-based methods for deepfake detection analyze individual images to spot any signs of manipulation or unusual After making sure the data is consistent, a model, such as the

Deep Inception net Learning Algorithm, is trained. This model learns to recognize distinct facial characteristics, making it good

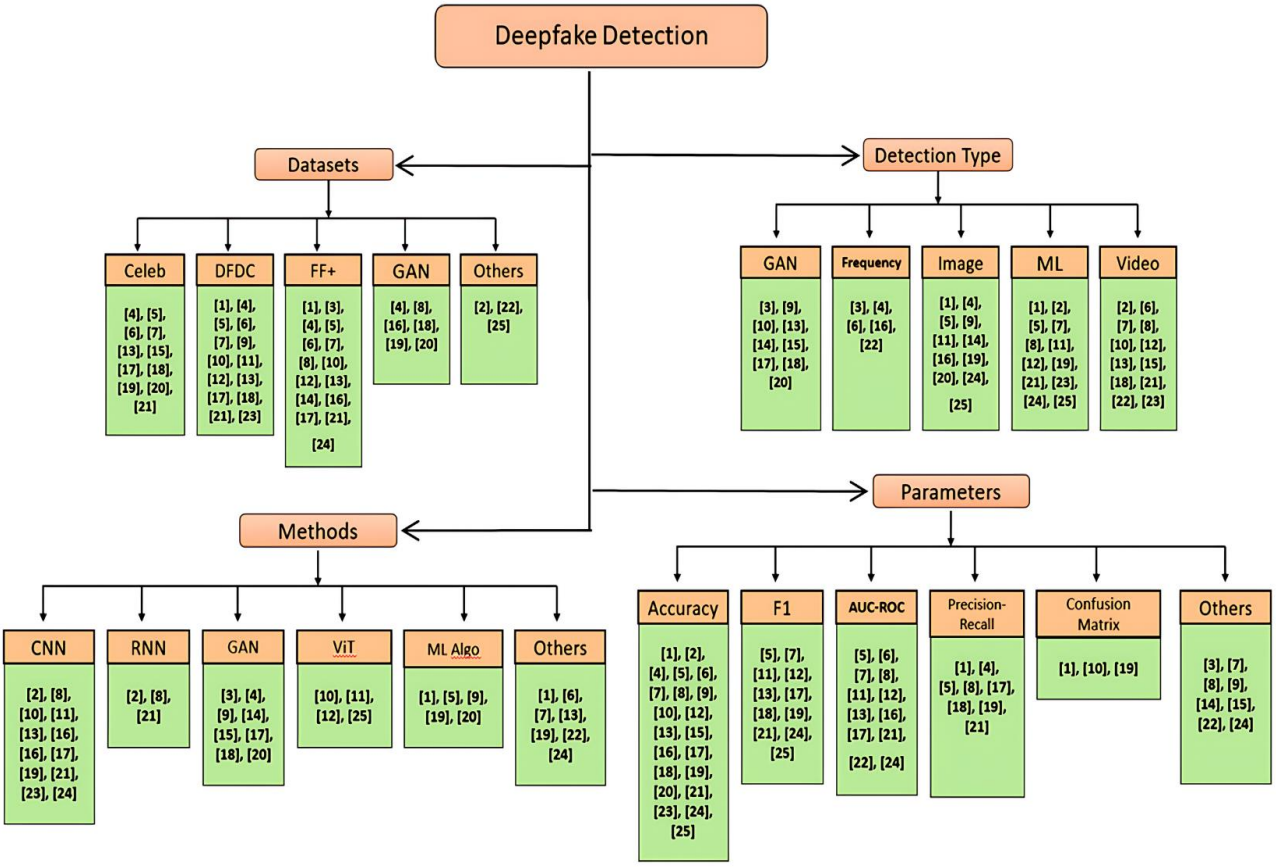

Figure 2- Classification of deepfake detection methods with reference paper.

at distinguishing real from manipulated faces. Once rigorously tested, it can be used in applications like content moderation on social media, making deepfake detection more accurate. However, there are downsides like the method being computationally complex and sensitive to compression rates, which can affect its performance.

Video-based methods analyze sequences of frames in videos to detect temporal inconsistencies or artifacts that indicate deep fake manipulation. They may also examine facial expressions, eye movements, or lip-syncing accuracy. using recurrent neural networks (RNNs) is a process that aims to identify manipulated videos created using deepfake technology. RNNs are a type of neural network that can process sequential data, such as videos frame by frame. In this video-based method, the RNN is trained on a dataset of both real and deepfake videos, learning to differentiate between the two through pattern recognition. This process begins by inputting a video into the RNN, which then breaks it down into individual frames. The RNN analyzes each frame and compares it to its learned patterns from the training dataset. By examining the sequential data of the frames, the RNN can identify inconsistencies or anomalies that are indicative of a deepfake video. This method does not detect which face was swapped or manipulated [2].

The other methods are, Generative adversarial networks (GAN) -based, Frequency-based and Machine-learning algorithm-based. The GAN-based method involves a two-step process. First, it uses a GAN to create synthetic deep fakes from social media data. Then, another GAN-based model is employed to discern real from generated content, enhancing detection accuracy. One drawback is the computational intensity of training and running dual GAN models [3]. The frequency-based method likely follows a process involving the use of a GAN-based approach. It incorporates perturbations at the frequency level in deep fake detection. A potential drawback is that it may require extensive computational resources for processing frequency-based perturbations in real-time applications[4]. The ML-based method focuses on the process of detecting deepfakes using machine learning methods. It likely involves data collection of authentic and deepfake content, feature extraction using machine learning algorithms, and model training. The machine learning models learn to distinguish between genuine and manipulated media, enhancing accuracy in identifying deepfakes[5].

2. LITERATURE SURVEY

In[6], Binh M .lee and Simon S. Woo proposed a single universal model for quality-agnostic deep fake detection that works on both high and low-quality images and videos instead of developing separate models for each quality type as it is impractical, this paper addresses, An intra-model-collaborative learning framework that trains the model on multiple quality versions of the same images is used as a framework. The authors introduce a collaborative framework that consists of a primary model and multiple auxiliary detectors. The primary model learns to distinguish real images from deep fake images without considering the quality of the fakes. The auxiliary detectors focus on distinguishing between high-quality and low-quality fake images. The proposed framework achieves state-of-the-art results on benchmark datasets of FaceForensic(FF++), Celeb Deep-Fake Videos 2 (CelebDFV2), Wild Deep fake, Deeper Forensics Faceshifter, and Deep-Fake Detection Challenge (DFDC), demonstrating its effectiveness in detecting deep fake images

regardless of their quality. The detection of very low-quality or heavily compressed data remains challenging here, and that achieved 97.82 percent Area Under the ROC Curve (AUC) on FF++,98.47 percent with efficient net backbone, 90.8 percent average AUC on unseen datasheets. This computational overhead of training uses multiple quality levels.

In [7], Aminollah Khormali, and Jiann-Shiun proposed a Deepfake detection method that shows promise in identifying forgeries within known datasets but struggles when faced with unseen samples that use self-supervised pre-training models. To address this, a reliable deepfake detection system should exhibit impartiality towards forgery types, appearances, and quality for reliable generalization. This study introduces a novel framework using self-supervised pre-training, a vision Transformer-based feature extractor, a graph convolution network with a Transformer discriminator, and a graph Transformer relevancy map. This framework excels in generalization and offers feature explainability. This proposed method achieved 99.4% AUC on CELEB DB and 81.3% on wild deepfake dataset where high computational requirements are achieved during training due to transformer architecture. In[8], Rahul katerya and Anushka Lab, proposed a method called Sentiment140(SST) Net, which stands for spatial, temporal and steganalysis network using a database of FF++ & GAN methods that address the growing risk of using deepfakes for malicious purposes. Its advantage lies in a proactive approach to understand and counteract potential dangers, offering insights into detection and prevention methods. However, limitations may include the evolving nature of deepfake technology, posing challenges in anticipating and addressing novel forms of weaponization.

"Latent Forensics" tackles the challenge of creating a faster and more resource-efficient deepfake detection method within the style GAN latent space, that performs face detection & alignment on input video frames. Matthieu Delmes, Amine Kacete, Stephene Paquelet, Simon Leglaive, and Renaud Seguier proposed an aligned face image into the latent space of style-GAN using an optimization-based inversion process. The compact latent space representation captures semantic features and removes background noises. They faced slow test time due to optimization-based StyleGAN inversion, the effect of compression, rotations, and occlusions are not analyzed which doesn't provide temporal inconsistencies in videos[9]. The paper[10], "Key video Frame Extraction Using GAN" aims to detect deep fakes by focusing on keyframes extracted from input videos using ResNext50 GNN, whose features are fed into an LSTM network with GAN technology. The datasets used are the DFDC dataset and Forensics ++. The problem it addresses is the need for effective video-specific detection methods. The advantage lies in improved accuracy through key frame analysis, but challenges include reliance on GAN effectiveness and the importance of frame selection for reliable detection.

Young-Jin Heo, Young-Ju Choi, Byung-Gyu Kim, and Young-Woon Lee proposed a deepfake detection scheme based on Vision Transformer (ViT) and distillation[11]. The authors leverage ViT, a state-of-the-art deep learning model for image classification, to extract high-level features from video frames. They also employ distillation techniques to transfer knowledge from a larger teacher network to a smaller student network. That resulted in lower loss and clearer discrimination of fake videos with an F1 score of 91.9% on DFDC. But requires a long training time of 2 days on a single GPU where this model is evaluated only with the DFDC dataset which remains as its drawback. The efficient Net and Vision Transformers system was proposed by David Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi about video Deep-Fake detection. That explains about detecting deepfake. Efficient Net is used to extract frame-level features from videos, while Vision Transformers are employed to capture global spatio-temporal information. The combination of these two models improves the performance of deep fake detection by effectively capturing both local and global visual patterns maintaining 80% accuracy, using DFDC and FaceForencis ++ as database. This system did not evaluate moving datasets and lacks in the testing of real-world deep fake videos[12].

Deressa Wodajo, Solomon Atnafu, and Zahid Akhtar propose a method called generative convolution Vision Transformers [13], Which has two main components (i) Generative port and (ii) Feature extraction port. The generative part uses an Autoencoder (AE) and a Variational Autoencoder (VAE) to learn the data distribution of training examples. Whereas the feature extraction part uses a ConvNet Model and a Swin Transformer model to extract relevant visual features from the input images and the reconstructed images from AE/ VAE. The datasets used on this system are DFDC with over 100000 videos of real-fake, FaceForencis with 1000 original YouTube videos manipulated using deep fakes, Face2Face, Faceswap, Neural Textures, Multiple Compression & Resolution, Celeb-DF(V2) with 890 real and 5639 fake videos, Trusted Media with 4380 fake and 2563 real videos with Diverse manipulation techniques and from Deepfake TIMIT with real and fake generated video using Generative adversarial networks (GAN)'S and VAE'S. Resulting in an accuracy of 98.5% on DFDC,97% on FaceForencis,98.28% on Deepfake TIMIT and 90.94% on CelebDF. The Average AUC of the model is 99.3% but the training generative model remains unstable and testing on videos remains as a challenging case. In[14]Chaofei Yong, Leah Ding, Yiren Chen, and Hai Li proposed a key idea to generate adversarial perturbation on the target faces such that training a deep fake model on these face degrades the quality of synthesized fake faces. Databases used are FaceForencis++ having 4 Male and 4 Female faces used as targets and sources for swapping. This results in increased Generative loss and adversarial or edge losses compared to the original model. Transformation-aware perturbation gives robustness to image transforms. The Model adds high-frequency noise in the FFT spectrum of fake faces.

Preethi, Manoj Kumar and Hitesh Kumar Sharma proposed a GAN-based model to detect Deep-Fake in social media, here they used deep convolutional Generative adversarial networks (GAN) Architecture with CelebA dataset for training the model. GAN accuracy reaches 100% by the 10^{th} iteration where its Inception score (IS) is 1.074 and Frechet Inception Distance (FID) is 49.3. IS measures the diversity and clarity of individual generated images while FID compares overall distributions between real and generated images. The drawback of this proposed model gives results that are evaluated only on Celebrity images and not on generalized datasets. There is a lack of study on model components and training strategies[15]. Young Oh Bang and Simon S. Woo proposed Domain Adversarial Face De-Identification and Forensics Through Neural Networks (DA-FDFTNET) [16]dual attention fake detection fine-tuning network to detect various AI-generated fake images, explains a key idea of pretraining Convolutional Neural Network (CNN) like Visual Geometry Group (VGG), ResNet, etc. on top of pretraining Convolutional Network (CN), Fine-tunning Transformer, MobileNet Block V3, Channel Attention Module and Classification Head are added. The model uses FaceForencis++, GAN-generated images and 2000 images(1000 real,1000 fake) datasets and achieved over 97% Accuracy and 99% Area Under the Receiver Operating Characteristic Curve (AUROC) on some datasets. The performance depends heavily on the pretraining, choice of baseline model and fine-tuning data and also there are no major insights on why this model works better.

Pratik Kumar Prajapati and Dr.Chris Pollett proposed a generalized approach to detect Deep-Fakes using perceptual image assessment. The used datasets are the DFDC dataset, the main dataset which contains real and deep fake videos subset of 400 videos.Celeb-DF-V2 contains Deep-Fake videos of celebrities and Face De-Identification Forensics (FDF)and Face

Detection from High-Quality images (FFHQ) datasets contain real images. Magnetic Resonance Imaging Generative Adversarial Network (MRI-GAN) highlights manipulated face areas at the pixel level and works for general video manipulation, not just on face swaps. This uses a perceptual similarity metric to train the GAN and achieves a reasonable accuracy of 74% on the datasets[17]. In [18], the aim is to detect Deep-Fakes using a method called GAN Discriminators. It tests multiple discriminator architectures and uses an ensemble of discriminators for improved robustness. The method achieves high accuracy on datasets that the discriminators were trained on. However, it does not work well for general Deep-Fake detection and has limited generalizability due to basic discriminator architecture and small datasets. The pre-processing and face-detection steps could be improved, and there is no evaluation on real-world Deep-Fakes. The results show a 92% accuracy when discriminators perform well on trained datasets, but only 69.7% accuracy on unseen DFDC Deep-Fakes.

Mahsa Soleimani, Ali Nazari, and Mohsen Ebrahimi Moghaddam propose a method that combines the Viola-Jones algorithm, MTCNN (Multi-task Cascaded Convolutional Networks), CNN (Convolutional Neural Network) called GRAM-NET, and ImageNet for Deep-Fake detection. The proposed method achieves high accuracy early on and performs well even with increasing occlusion ratios. It improves accuracy by assigning weights to different facial patches, giving more importance to the mouth and eyes. The proposed method achieves 84.9% accuracy, outperforming Lin et al. (80.3%) by 4.6%. It also converges faster than other methods. The fabricated image diversity is limited, as the evaluation does not include real-world manipulated images from StyleGAN, StarGAN, or PGGAN[19]. In Paper [20], Luca Guarnera, Oliver Giudice, and Sebastiano Battialo propose a method for detecting Deep-Fake images by analyzing convolutional traces. The method uses a combination of the EM algorithm, GAN, SVM, KNN, and LDA for classification. It achieves high accuracy in detecting state-of-the-art GAN fakes, such as style, and can work with varying image sizes. The method also discriminates well between different GAN architectures. However, it has limitations such as not considering the effects of compression and image processing and susceptibility to counter-forensic attacks. The results show high accuracy for styleGAN and styleGAN2, as well as good accuracy for other GAN architectures such as ATTGAN, GDWCT, and STARGAN. styleGAN and styleGAN2 also exhibit good separation owing to generator differences.

In[21], the Authors proposed an effective and fast Deep-Fake detection method based on Hybrid, CNN, and Long Short-Term Memory(LSTM). It utilizes both spatial and temporal inconsistencies by incorporating optical flow. The method can detect deep fakes even with fewer frames, allowing for early detection. The research uses FaceForensics++, celeb-DF, and DFDC datasets for evaluation. The advantages include the use of Google Colab Pro and libraries such as OpenCV, Keras, and Sklearn. However, the method has limitations like limited frames analyzed and the need for more diverse datasets. The results show that the method achieves an accuracy of 74.87% to 91.21% on FaceForensics++, and 63.24% to 79.49% on celeb-DF, with DFDC peaking at 66.26% accuracy using 30 frames.

Based on the Haar Wavelet Transform, Mohammed Akram Younus, and Taha Mohammed Hasan [22] proposed a method for effectively and quickly detecting Deep-Fakes using the Haar wavelet transform. It utilizes a Dlibs face detector library, regions of interest (ROI), and a Haar wavelet transform-based analysis. The method has advantages such as fast detection irrespective of uniform face backgrounds, generalizability to different Deep-Fake generation methods, and accurate discrimination of fake faces. However, it has limitations like not considering artifacts other than blur inconsistencies, lower detection accuracy for practical use, and lack of analysis on computational complexity. The method is evaluated on the Unconstrained Audio-visual Deep-Fake Videos (UADFV) dataset, achieving a 90.5% accuracy in detecting Deep-Fakes. Atmik Ajoy, Chethan U Mahindrakar, Dhanya Gowrish and Vinay A [23] proposed a Deep-Fake detection model based on a frame-by-frame approach using CNN. The model utilizes three Multi-task Cascaded Convolutional Networks (MTCNN) models to detect flaws in fake faces, such as distortion. It is trained on datasets like DFDC, YouTube dataset, ImageNet, and FaceForensics, recognizing unique Deep-Fake patterns. However, it has limitations like not working in real-time and requiring large training datasets. The proposed model achieves an accuracy of 85.8% with a loss of 0.3403. Despite its promising results, further validation on a larger variety of datasets is necessary.

It's, a deep learning approach, utilizing CNN, LSTM, and Recycle-GAN, for Deep-Fake detection. It trains and tests the model using databases like Faceforensics++(which contains 1000 manipulated YouTube videos and around 1 million images), Deep-Fakes, face2face, faceswap, and neural textures. The method offers advantages such as high accuracy, robustness, efficiency, generalizability, and automation. However, it also has limitations, including limited datasets, vulnerability to adversarial attacks, and limited explainability. The system's promising approach to address the growing concern in Deep-Fake is acceptable. In[24], the authors proposed a Multi-Spectral Class Center Network (MSCCNET) for face manipulation detection and localization. It uses multiple datasets including FF++, FACE2FACE, Faceswap, face-shifter, and neural textures. The proposed methodology introduces more accurate pixel-level annotations for FF++ and achieves good results for both localization and detection on this dataset. It also demonstrates better generalization to unseen manipulation and datasets compared to existing benchmarks. The paper suggests that joint optimization of the classifier and localizer in an end-to-end manner could be explored, and improving computational efficiency for high-resolution images is another potential area of improvement. The MSCC NET achieves a significant improvement of 4-25% in MIOU (mean intersection over union) compared to existing methods.

In[25], Fanglei Xue, Qiangchang Wang, Zichang Tan, Zhongsong Ma, and Guodong Guo proposed Vision Transformers (ViTs) with Attentive Pooling for Robust Facial Expression Recognition. Researchers introduce APViT, leveraging two novel Attentive Pooling modules with Vision Transformer to enhance facial expression recognition on limited datasets. The proposed approach focuses on discerning crucial features while discarding irrelevant ones, mitigating issues related to occlusion and noise. Across six in-the-wild FER datasets, APViT consistently outperforms state-of-the-art methods, as validated through visualization of its intuitive and robust attentive pooling mechanisms.

Table I: Comparison Of Existing Methods

S.NO	Proposed Method	Type of Detection	Database Used	Parameters Used	Advantages	Limitations
[1]	Deep Inception Net	ML Algorithm and Image-Based	Custom dataset of 401 deep fake videos, Tested on FF++ and DFDC Dataset	Accuracy, Confusion matrix, loss function, precision, and recall	Used for optimized multi-scale processing	Computational complexity, Sensitivity to different compression rates.
[2]	Convolutional LSTM, RNN	ML Algorithms, and Video-Based	300 videos collected from multiple video hosting websites, 300 pristine videos randomly selected from the HOHA DATASET	Accuracy	1. Uses a sample convolutional LSTM, RNN architecture 2. Exploits spatial and temporal inconsistencies	Does not detect which face was swapped or manipulation region
[3]	Deep fake model Discriminator	GAN and Frequency Based	Faceforencis++	Generator and Discriminator Losses, Adversarial, Edge, and Reconstruction Losses, FFT Spectrums and Visualization, MontraNet Forgery Detection Maps, Average Intensity of High Frequency	1. Ensemble generalizes to black boxes. 2. Robust to transformations. 3. Adds high-frequency noise	Lack of generalization, Absence of stealthiness analysis for adversarial perturbations,
[4]	FrePGAN	Image and Frequency Based	ProGAN, FFHQ, LSUN, ImageNet, CELEBA, COCO Datasets, Faceforencis++ deep fake Dataset	Accuracy, Average Precision	Frequency perturbation, Improved performance	Requires retraining if new, unknown artifacts appear.
[5]	DFFvector, SVM, Random Forest, Decision Tree, MLP (ML Algorithms)	ML Algorithms and Image-Based	FF++,DFDC, CELEB DF, VDFD	Precision, accuracy, Recall, F1 score, AUCROC	Low computational cost and faster training, utilizes SVM, RF, and ERT instead of deep learning.	No temporal analysis of frames in videos, limited robustness testing on videos
[6]	An intra-model Collaborative learning framework of ResNet50 and Efficient net	Frequency and Video-Based	FF++, CelebDFV2, Wilddeep fake, Deeper Forensic Faceshifter, DFDC	Classification accuracy, AUCROC	Outperforms quality-aware models like ADD & BZnet, A Single model works for various quality deep fakes.	Struggles with very low-quality or compressed data, Requires multiple quality versions for training.
[7]	Self-Supervised graph	ML Algorithm and Video-Based	Forensics++, Celeb DF, wild deep fake, Deeperforensics,	ACC, AUCROC, F1-Score, MIOU	represents facial images as graphs using self-	High computational requirements

	transforme r		Faceshifter, DFDC		supervised pre-training.	during training
[8]	SSTNET, Spatial CNN, RNN, LSTM	ML Algorithms and Video-Based	FF++, GAN based models	ACC, AUCROC, PRECISION, RECALL, MSE and PSNR	analyzed different modalities, (image/videos/ audio), and compared different GAN architectures.	Detect deep fakes created using supervision GAN'S
[9]	StyleGAN, Random Forest, Logistic Regression , MLP Algorithm s	GAN and Image-Based	DFDC datasets	Classification ACC, LPIPS	Mid-to-lake StyleGAN channels are more useful for detection, capturing semantic features, and removing background noise.	slow test time, Not analyzed effect of compression, rotations, occlusions
[10]	ViT, CNN, LSTM, GAN base model	Video and GAN-Based	DFDC and faceforensics++	Classification ACC and confusion matrix	Uses ResNet Xt 50 CNN per extraction, LSTM with GAN technology	Not Tested on moving videos, not analyzed the effect of keyframe solution strategy
[11]	ViT, CNN features, and Distillatio n	Image and ML Algorithms Based	DFDC	AUCROC, F1 SCORE	Hybrid CNN-ViT for Localization, Confusion Matrix Improvement.	Ignored temporal inconsistencies and lengthy training time on a single GPU.
[12]	Efficient net - ViT and Convoluti onal Cross ViT	Video and ML Algorithms-Based	DFDC and faceforensics++	AUCROC, F1 score, ACC	EfficientNet for effective feature extraction, Achieved high performance without ensemble or distillation.	slightly lower than top methods on DFDC, acks testing on diverse real-world deep fake video. face computational overhead for using two network branches.
[13]	AE, VAE, CNN, and a Swin transforme r model	Video and GAN-Based	DFDC, Face forencis++, face2face, faceswap, deep fake TIMIT, Celeb-DF(V2), Trusted media	Classification accuracy, F1-SCORE, AUCROC	Utilizes both generative and discriminative models, Dual network architecture provides ensemble benefits, Evaluated on diverse datasets, demonstrating generalization.	Complex dual network architecture increases computational demands, Training generative models can be unstable, Testing done on constrained videos, and challenging cases remain unaddressed.

[14]	Deep fake model, Discriminator DA	GAN and Image-Based	Faceforencis++	Generator and discriminator losses, Adversarial, edge, reconstruction losses, FFT spectrums, and visualization, The average intensity of high freq in FFT, MonTraNet forgery detection maps	The ensemble method extends to a black box setting, Enhanced deep fake quality degradation.	Limited to FaceForensics++ dataset, Lacks generalization to other methods, Stealthiness of adversarial perturbations not analyzed.
[15]	Convolutional GAN	Video and GAN-Based	CelebA	Inception score, FID, accuracy, and loss	Stable and effective GAN architecture, High accuracy in detecting GAN-generated fakes.	Limited study on model components and training, Risk of mode collapse with different settings, and Limited evaluation beyond celebrity images.
[16]	CNN, ResNET, FTT, Mobilenet Block V3	Frequency and Image-Based	Faceforencis ++, MTCNN, styleGAN &styleGAN2	Accuracy and AUCROC	Effective across diverse deep fake and GAN datasets, Utilizes transformer, MB Blocks, and channel attention modules for improved performance.	Performance heavily depends on pretraining, baseline model choice, and fine-tuning data.
[17]	U-net style encoder-decoder CNN and a Discriminator	Image and GAN Based	DFDC, CELEB-DF-V2, FDF & FFHQ	Accuracy, true positive rate, false positive rate, precision-recall, F1 score specificity, AUCROC	General video manipulation capability, Uses perceptual similarity metric for GAN training, Qualitative MRI images show differences for fake faces	Low accuracy compared to the state-of-the-art. MRI-GAN performs worse than the plain frames method. Hyperparameter tuning needs improvement
[18]	Uses GAN discriminators combined with ensemble techniques	Video and GAN-Based	DFDC, Celeb-A, styleGAN	Accuracy, precision, recall, F1-score	Evaluate diversity and effectiveness and enhance model resilience.	Basic discriminator architecture, limited generalizability, and basic discriminator architecture, limited generalizability. Challenges in data preprocessing and face

						detection.
[19]	Viola-jones, MTCNN, CNN(GRAM-NET), and ImageNET	Image and ML Algorithms Based	CELEB-A, FFHQ, styleGAN, styleGAN2 ON FFHQ, STARGAN & PGGAN on CELEB-A	Accuracy, precision, recall, F1-score, confusion matrix	Achieves high early accuracy and handles occlusion.	Limited diversity in fabricated images, no evaluation of real-world manipulated images. Relies on pre-trained models, a small dataset limits generalizability.
[20]	EM Algorithm, GAN, SVM, KNN, LDA	Image and GAN Based	Celeb-A, STARGAN, ATTGAN, GDWCT, STYLEGAN, STYLEGAN2	Classification accuracy, Classification accuracy between different pairs of GAN architecture	Achieves high accuracy in detecting state-of-the-art GAN fakes like StyleGAN, Computationally efficient feature extraction using EM algorithm, and Discriminates well between different GAN architectures.	Doesn't consider the effects of compression and image processing. Susceptibility to counter-forensic attacks not analyzed. Limited testing on other datasets and in-the-wild images.
[21]	CNN & LSTM	ML Algorithms and Video-Based	Faceforensics++,celeb-DF, DFDC	Accuracy, recall, precision, F1-score, AUC-ROC	Spatial-temporal inconsistency capture, CNN & LSTM enhance early fake detection.	Limited frames due to computation constraints, Only 3 datasets evaluated; lacks diversity, Detection accuracy insufficient for practical use.
[22]	Dlibs Face Detector, ROI, HAT	Frequency and Video-Based	UADFV	AUC, NN, MESO-4, MESOINCEPTION-4, HEADPOX	Fast detection, Generalizable to GAN-based methods, and Accurate discrimination.	Limited accuracy; lacks computational analysis, Ignore post-manipulation artifacts, Limited dataset evaluation; needs validation.
[23]	CNN (MTCNN-3 Models)	ML Algorithms and Video-Based	DFDC, YOUTUBE dataset, ImageNet, Faceforensics	Accuracy loss	Deep learning and CNN usage, Frame-based flaw detection, Identifying unique deep fake patterns.	Limited to pre-recorded videos, Requires large datasets, and Needs more diverse validation.
[24]	MSCCNET	Image and ML Algorithms Based	FF++, FACE2FACE, Faceswap, face-shifter, and neural	ACC, AUCROC, F1-SCORE, and MIOU	Precise Localization, Joint Optimization, Robust	Computational Efficiency, Variable Impact of GCN

			textures		Generalization	
[25]	ViT with Attentive Pooling	ML Algorithm and Image-Based	AffectNet, RAF-DB, SFEW, ExpW	Accuracy, F1 Score	Attention pooling enhances pertinent features. Outperforms CNNs and vanilla ViT. Generalizes better across datasets	Still lags human performance, Testing is limited to constrained facial images, and the Pooling strategy is dataset dependent. Computationally intensive self-attention.

3. CONCLUSION

An extensive exploration of diverse Deep-Fake detection methods was discussed in this paper, emphasizing the challenge posed by heightened computational complexity during feature extraction from image frames in large datasets. This paper concludes that an attentive-pooling method outperforms all the other proposed methods in the literature. Attentive pooling is a combination of attentive patch pooling and attentive token pooing. Unlike conventional approaches that directly employ all extracted features for deep fake detection, the attentive pooling method utilizes attentive patch pooling and assigns scores to each feature. Through attentive token pooling, lower-scored features are dropped, maintaining a keep rate of k=0.8, retaining 80% of features while discarding 20%. This strategic approach significantly reduces computational complexity, enhancing the accuracy. The study indicates that this attentive-pooling method surpasses other techniques, leading to superior performance in Deep-Fake detection compared to the other existing methods discussed in the paper. The achieved high performance is attributed to the effective implementation of attentive pooling, providing a promising solution to the computational complexity in feature extraction for large datasets.

4. REFERENCES

[1] Prasannavenkatesan Theerthagiri, Ghouse basha Nagaladinne, “Deepfake Face Detection Using Deep InceptionNet Learning Algorithm”, *2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science*, Feb 2023.

[2] David Guera, Edward J. Delp, “Deepfake Video Detection Using Recurrent Neural Networks”, *15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Feb 2019.

[3] Vedant Jolly, Mayur Telrandhe, Aditya Kasat, Atharva Shitole, Kiran Gawande, “CNN based Deep Learning model for Deepfake Detection”, *2nd Asian Conference on Innovation in Technology (ASIANCON)*, pp.1-5, Aug 2022.

[4] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, Jongwon Choi, “FrePGAN: Robust Deepfake Detection Using Frequency-level Perturbations”, *Association for the Advancement of Artificial Intelligence*, pp. 1060-1068, Feb 2022.

[5] Md. Shohel Rana; Beddhu Murali; Andrew H. Sung, “Deepfake Detection Using Machine Learning Algorithms”, *10th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 458-473, July 2021.

[6] Binh M. Le, Simon S. Woo, “Quality-Agnostic Deepfake Detection with Intra-model Collaborative Learning”, Sep 2023.

[7] Aminollah Khormali, Jiann-Shiun Yuan, “Self-Supervised Graph Transformer for Deepfake Detection”, pp.1-13, Jul 2023.

[8] Rahul Katarya, Anushka Lal, “A Study on Combating Emerging Threat of Deepfake Weaponization”, *Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 485-490, Oct 2020.

[9] Delmas, M., Kacete, A., Paquelet, S., Leglaive, S. and Seguier, R., “LatentForensics: Towards lighter deepfake detection in the StyleGAN latent space,” arXiv preprint arXiv:2303.17222, 2023.

[10] Lalitha S, Kavitha Sooda, “DeepFake Detection Through Key Video Frame Extraction using GAN”, International Conference on Automation, Computing and Renewable Systems (ICACRS 2022), pp. 859-863, Dec 2022.

[11] Young-Jin Heo, Young-Ju Choi, Young-Woon Lee, Byung-Gyu Kim, “Deepfake Detection Scheme Based on Vision Transformer and Distillation”, Applied Intelligence, Apr 2021.

[12] Davide Coccomini, Nicola Messina, Claudio Gennaro, Fabrizio Falchi, “Combining EfficientNet and Vision Transformers for Video Deepfake Detection”, Image Analysis and Processing, Springer – Lecture Notes Series, volume 2, pp.1-11, Jan 2022.

[13] Deressa Wodajo, Solomon Atnafu, Zahid Akhtar, “Deepfake Video Detection Using Generative Convolutional Vision Transformer”, arXiv:2307.07036, Volume 1, pp.(1-11), Jul 2023.

[14] Chaofei Yang, Leah Ding, Yiran Chen, Hai Li, “Defending against GAN-based DeepFake Attacks via Transformation-aware Adversarial Faces”, *2021 International Joint Conference on Neural Networks (IJCNN)*, July 2021.

[15] Preeti, Manoj Kumar, Hitesh Kumar Sharma, “A GAN-Based Model of Deepfake Detection in Social Media”, *International Conference on Machine Learning and Data Engineering*, Volume 218, pp. 2153-2162,2023.

[16] Young Oh Banga, Simon S. Woob, “DA-FDFtNet: Dual Attention Fake Detection Fine-tuning Network to Detect Various AI-Generated Fake Images”, arXiv:2112.12001v1 [cs.CV], pp.1-9, Dec 2021.

[17] Pratikkumar Prajapati, Chris Pollett, "MRI-GAN: A Generalized Approach to Detect DeepFakes using Perceptual Image Assessment", *ArXiv* abs/2203.00108, Feb 2022.

[18] Sai Ashrith Aduwala, Manish Arigala, Shivan Desai, Heng Jerry Quan, Magdalini Eirinaki, "Deepfake Detection using GAN Discriminators", *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*, pp.69-77, Aug 2021.

[19] Mahsa Soleimani, Ali Nazari, Mohsen Ebrahimi Moghaddam, "Deepfake Detection of Occluded Images Using a Patch-based Approach", arXiv:2304.04537Apr 2023.

[20] Luca Guarnera, Oliver Giudice, Sebastiano Battiato, "DeepFake Detection by Analyzing Convolutional Traces", *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* pp.2841-2850, 2020.

[21] Pallabi Saikia, Dhwani Dholaria, Priyanka Yadav, Vaidehi Patel, Mohendra Roy, "A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features", *2022 IEEE World Congress on Computational Intelligence,* Jul 2022.

[22] Mohammed Akram Younus, Taha Mohammed Hasan, "Effective and Fast DeepFake Detection Method Based on Haar Wavelet Transform", *2020 International Conference on Computer Science and Software Engineering (CSASE),* pp.186-190, 16-18 April 2020.

[23] Atmik Ajoy, Chethan U Mahindrakar, Dhanya Gowrish, Vinay A, "DeepFake Detection using a frame based approach involving CNN", *3rd International Conference on Inventive Research in Computing Applications (ICIRCA-2021),* pp. 1329-1333, Sep 2021.

[24] ChangtaoMiao, QiChu,ZhentaoTan, ZhenchaoJin, WanyiZhuang, YueWu,BinLiu, HonggangHu, NenghaiYu, "Multi-spectral Class Center Network for FaceManipulation detection and localization", arXiv:2305.10794 **,** Version 2, pp.1-16, Sep 2023.

[25] F. Xue, Q. Wang, Z. Tan, Z. Ma and G. Guo, "Vision Transformer With Attentive Pooling for Robust Facial Expression Recognition," in IEEE Transactions on Affective Computing, vol. 14, no. 4, pp. 3244-3256, 1 Oct.-Dec. 2023, doi: 10.1109/TAFFC.2022.3226473.

5. AUTHOR'S PROFILE

Gautham Yaramasa is an undergraduate student pursued a Bachelor's degree in Electronics and Communication Engineering from AMC Engineering College, Bangalore. Gautham has demonstrated expertise in Machine Learning, Deep Learning, Natural Language Processing, and Computer Vision through internships. His project work includes developing a Wikipedia bot using Langchain and Transformers, Stock Price Forecast, Facial Emotion Recognition Model using OpenCV. His research interests include Machine Learning, Deep Learning and Quantum Computing.

Sindhu R is an undergraduate student pursued a Bachelor's degree in Electronics and Communication at AMC Engineering College. She possesses skills in Artificial Intelligence, Machine learning, Deep learning, Natural language processing and MySQL. Her expertise lies in machine learning and deep learning. Notably, Her project work includes developing a diabetes prediction and Sentimental analysis model. With a strong foundation in cutting-edge technologies, she aims to contribute to research and development in these domains at a reputed organization.

Dr. Jenitta J works as Professor in the department of ECE, AMC Engineering College, Bangalore. She completed her B.E from M.S. University in the year 2004, M.E. from Karunya University in the year 2006 and Doctoral degree from Anna University in the year 2016. Her area of research is Biomedical Signal Processing. Her area of interest includes Machine Learning, Deep Neural Networks and Artificial Intelligence. She is a reviewer in IET Signal Processing, IET Science, Measurement and Technology, IEEE Journal of Biomedical and Health Informatics, IEEE Access, IEEE Journal of Biomedical Signal Processing and Control and she has published more than 35 papers in reputed Journals and Conferences. Currently she is a mentor for Deeplearning.ai courses in Coursera. She is a senior member of IEEE. She is also a member of professional bodies like ISTE and IAENG.