IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 October 2023, accepted 4 December 2023, date of publication 12 December 2023,
date of current version 21 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3342107

SURVEY

Deepfake Generation and Detection: Case Study and Challenges

YOGESH PATEL[1], SUDEEP TANWAR[1], (Senior Member, IEEE), RAJESH GUPTA[1], (Member, IEEE), PRONAYA BHATTACHARYA[2], (Member, IEEE), INNOCENT EWEAN DAVIDSON[3,4], (Senior Member, IEEE), ROYI NYAMEKO[3,4], SRINIVAS ALUVALA[5], AND VRINCE VIMAL[6,7], (Member, IEEE)

[1]Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat 382481, India
[2]Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University, Kolkata 700135, India
[3]African Space Innovation Center, Department of Electrical, Electronic and Computer Engineering, Cape Peninsula University of Technology, Bellville 7535, South Africa
[4]Department of Electrical, Electronic, and Computer Engineering, French South African Institute of Technology, Cape Peninsula University of Technology, Bellville 7535, South Africa
[5]Department of Computer Science and Artificial Intelligence, SR University, Warangal, Telangana 506371, India
[6]Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun 248002, India
[7]Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, Uttarakhand 248002, India

Corresponding authors: Sudeep Tanwar (sudeep.tanwar@nirmauni.ac.in), Innocent Ewean Davidson (Davidsoni@cput.ac.za), and Rajesh Gupta (rajesh.gupta@nirmauni.ac.in)

ABSTRACT In smart communities, social media allowed users easy access to multimedia content. With recent advancements in computer vision and natural language processing, machine learning (ML), and deep learning (DL) models have evolved. With advancements in generative adversarial networks (GAN), it has become possible to create fake images/audio/and video streams of a person or use some person's audio and visual details to fit other environments. Thus, deepfakes are specifically used to disseminate fake information and propaganda on social circles that tarnish the reputation of an individual or an organization. Recently, many surveys have focused on generating and detecting deepfake images, audio, and video streams. Existing surveys are mostly aligned toward detecting deepfake contents, but the generation process is not suitably discussed. To address the survey gap, the paper proposes a comprehensive review of deepfake generation and detection and the different ML/DL approaches to synthesize deepfake contents. We discuss a comparative analysis of deepfake models and public datasets present for deepfake detection purposes. We discuss the implementation challenges and future research directions regarding optimized approaches and models. A unique case study, *IBMM* is discussed, which presents a multi-modal overview of deepfake detection. The proposed survey would benefit researchers, industry, and academia to study deepfake generation and subsequent detection schemes.

INDEX TERMS Artificial intelligence, Deepfake generation, Deepfake detection, fake content, generative adversarial networks.

I. INTRODUCTION

Rapid technological advancements and easy access to the web have allowed users and communities to interact with each other on social platforms. Coupled with advancements in generative artificial intelligence (AI) models, it has enabled the creation of digital content (audio, video, and text) with a realistic flavor. This synthetic and fake content generation (termed as deepfakes) uses different machine learning (ML), and deep learning (DL) algorithms to look and sound real, and works on the superimposition of the face and voice of some person on another person [1]. This leads to the generation of fake news in social communities, spread hatred and misinformation manipulates public opinion, and can be

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa M. Fouda.

further extended to malicious uses like extortion, blackmail, spoofing, identity theft, character assassination, and deepfake pornography. This can cause detrimental effects, where it can significantly hurt the emotional and psychological state of a person, and cause him to face shame, humiliation, and social outrage in public. At the community front, deepfakes are used to generate fake propaganda and communal hatred (political and religion-based) and can lead to violence and outbreaks among communities [2]. Thus, deepfakes remain a significant threat to individuals, businesses, and society as a whole.

Several unfortunate incidents of deepfakes have happened recently over social media platforms, which have caused a potential concern about its open misuse by the general public. In 2019, a video circulated on social media showed the United State speaker of the House Nancy Pelosi, slurring her words and appearing drunk [3]. Similarly, in 2020, a fake video of United States ex-president Donald Trump emerged where he was seen endorsing his opponent Joe Biden [4]. In 2021, a deepfake video of actor Tom Cruise circulated on the Tiktok platform engaged in unethical behavior [5]. Similarly, a lot of deepfakes are used for the creation of fake pornographic content, which includes famous celebrities [6], [7], [8]. This creates a damaging impression in the minds of their fans, and sometimes these videos are used to harass these celebrities publicly. Another area where deepfakes are targeted is towards election campaigns, where fake videos of a political candidate are circulated to affect his public sentiments [9]. Thus, there is a stringent need to identify deepfakes from genuine sources.

Deepfake models are developed to create likeliness or fake versions of an individual in an image, speech, or video. Deepfake comes from the underlying deep learning (DL) technology that swaps faces in digital content to create a fake impression of a person in a realistic environment. Deepfakes involve deep neural networks, convolutional neural networks (CNNs), autoencoders, and generative adversarial networks (GANs) as popular generation techniques [10], [11]. As deepfakes look realistic, it poses a great threat and questions the authenticity of the published content. Deepfake manipulations can be done on audio content of any video that allows live videos of people making expressions and saying things they have never spoken before. The manipulations on the video/image contents are possible to swap faces, change expressions, lipsync, and many others. Once deepfake content is created, it is circulated on social platforms to propagate fake news. In recent times, it is noticed that deepfakes of popular celebrities, politicians, and sportspeople are frequently created by online users for fun, personal vendetta, or malicious propaganda. One emerging technique of deepfake generation is GAN It exploits a generative adversarial process having generative G and discriminative D models. In G, data distribution is captured and D estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. GAN framework closely resembles the minimax two-player game [12].

The GAN model, owing to its adversarial learning, has caught the attention of masses for the creation of deepfakes. However, GAN accommodates a small space of total AI deepfake models. On a positive note, GAN-enabled deepfakes are used to generate photos of imaginary models for clothing, branding, and related fashion accessories [13]. GANs are also used in the healthcare domain, where it can create artificial medical images of brain tumors in MRI scans for testing and validation models [14]. Deepfakes use cases have shifted towards Industry 5.0 AI models, where deepfakes are used to create realistic training and simulation environments for humans and cobots to collaborate and learn with dynamic safe response procedures. It can be used to create improvised chatbot services, where employees can learn customer service from realistic generated customer interactions. In marketing and business segments, deepfakes generate promotional videos with personalized messages for customers, creating a higher engagement and connectivity of the customer with the brand. Deepfakes can create highly realistic capture of face and body movements, generating realistic avatars for users in metaverse ecosystems [15]. The deployment of a particular GAN model for application purposes depends on the underlying neural network and the selected dataset. To cater with this need, deepfake generation tools are used to generate synthetic content from videos and images. Some notable video tools include Faceswap, Faceswap-GAN, DeepFaceLab, DFaker, and many others [16]. For an audio generation, deepfakes tools like WaveNet, MelNet, Char2Wave, and WaveGlow are mainly preferred [17]. The tools are easy for users to work with, and thus deepfake content is prominently surfacing in large numbers on social media communities, posts, and blogs. As per the report by Cybernews, the deepfake content over the Internet doubles every six months, and most of the contents use the GAN and the deep convolutional GAN (DCGAN) model [18]. FIGURE 1 shows the progressive increase of deepfake content (captured till April-2021). The number of deepfake generation contents has risen exponentially to 87,324 million content from 8,342 content in June-2018. The prominent difference between GAN and DCGAN is that the GAN generator uses a fully connected network, and DCGAN uses a transposed convolutional network, which upscales the images [19].

Deepfake audio tools have the ability to create entirely realistic lip sync videos in which the audio modalities are manipulated and mapped with the video frames [20]. One prominent example is Obamanet [21], which has the ability to create photo-realistic lip sync videos. The tools are mainly used in negative aspects for fake news, election propaganda, and character assassination, but can be used positively in movies where dialogue generation or dubbing can look realistic on the face of celebrities. FIGURE 2 shows the trend of users shifting towards deepfake tools, and the search trends have evidently increased in the last five years.

Thus, it becomes highly crucial to study the core techniques behind deepfake generation as well as deepfake detection. Thus, the researcher has shifted towards AI models

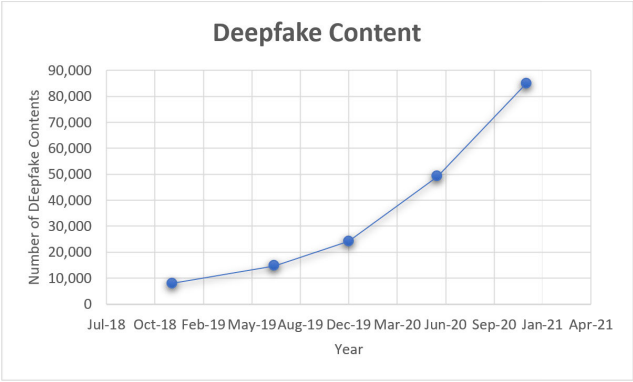

FIGURE 1. Doubling of deepfake content every six months.

FIGURE 2. Trends of google search of past five years for deepfake.

for deepfakes content generation, and classification models for detection [22]. Generally, the detection of deep fake image/video has been treated as a binary classification problem for which CNN models has been used. Whereas for the detection of audio manipulations, audio spoofing is mainly used. Audio spoofing detection has also been treated as binary classifications for which Gaussian mixture models (GMM), and deep neural networks (DNN) based solutions are explored. To detect video/image deepfakes, researchers have used spatial features and steganalysis features extracted using CNN. Spatial features include visible inconsistencies in images like contrast difference, artificial smoothness, and facial texture. Another direction is to perform steganalysis, which allows us to analyze the hidden features using low-level feature extraction techniques [21]. In CNN models, mainly the CNN-Xception network is found as a potential candidate for deepfake detection [21].

Deepfake generation and detection tools have become flexible and highly accurate. Some prominent deepfake generation tools are listed as follows.

1) *StyleGAN*: A prominent GAN model and it is considered an advanced version of basic GAN. It generates highly realistic images from sources with high degree of diversity. StyleGAN is mainly used in gaming designs, art platforms, and face generation.
2) *GANimation*: This model is used to generate highly realistic facial expressions and movements. The model is trained on a facial expression dataset and thus is perfectly used to create realistic deepfake videos.
3) *Face Swapping GAN*: The model, face swapping GAN (FSGAN) is used to create facial reenactment in videos. The facial expression of one person can be copied to another person, which helps in creating and enacting deepfakes.
4) *SimSwap*: This model is used for swapping purposes. In this, the face of one person can be swapped with another, and thus it is mainly used in deepfake pornography.

Next, we discuss the deepfake detection models as follows.

1) *XceptionNet*: XceptionNet uses deep CNN to detect deepfakes, and is trained on a large dataset of real and fake images and videos to learn the features that distinguish between real and fake images. It has a reported accuracy rate of over 90
2) *FaceForensics++*: The model uses a combination of handcrafted and learned features for deepfake detection. It works on facial manipulations, face pixel color correction, face composition and alignment, and other facial extractions to identify a real image from a generated image.
3) *FakeSpotter*: Fakespotter was developed at the University of California, Berkeley, and uses a combination of CNN and long short-term memory (LSTM) networks to analyze the spatial and temporal features from deepfake videos. It can distinguish the differences in real and fake features based on a probability score which presents the likelihood that the presented video is deepfake or not.
4) *GANomaly*: An anomaly detection model that uses GANs to detect abnormalities in data. It can be used for deepfake detection by identifying anomalies in deepfake videos or images.

We present some key works in deepfakes detection and generation. For example, Xuan et al. [23] proposed a two-phase model which pairs a real and fake image together. The pairs learn from discriminative common fake feature networks (CFFN). This discriminative CFFN is used to identify the authenticity of the image. To detect fabricated videos, models are trained that try to detect the inconsistencies in blinking that deepfake cannot mimic accurately. Also, several intraframe and temporal inconsistencies can be detected in the fabricated videos. Thus, various deep learning models using recurrent neural networks (RNN), or convolutional LSTM (C-LSTM) structures have been used to create this manipulation detection system [21]. In the case of audio spoofing detection, generally, there are two approaches that are widely common. Either standalone DNN-based models are trained over large datasets of spoofed audio or an ensemble model-based approach is used where capabilities of multiple individual models are combined to improve generalizability. Nasar et al. [24] first converted the audio file into a corresponding spectrogram, and then this spectrogram image is then used to classify the audio using CNN-based architecture. Other significant examples of single model architectures were presented by Zhang et al. [25], who proposed the use of ResNet-18 with a special loss function which improves the performance of the model. Similarly Ling et al. [26] introduced attention modules over each

residual bottleneck to improve performance. To improve the accuracy further, ensemble models are also used. For example, Dua et al. [27] created an ensemble model made up of three different models and aggregated their results. Authors in [28] created three different ensemble modes using 10 different individual models.

However, even with such prolific research in deepfake detection, real-world data training has shown less accuracy. DL-based models require a large amount of fabricated as well as authentic images/videos or audio files as datasets to train models. Along with it, deepfake technologies are also developing to evade detection systems. DL-based models try to find spatial inconsistencies such as blinking inconsistencies in the case of videos. In the case of images, the models find inconsistencies in face blurs, background contrast, and other discriminative features. In the case of audio spoofing detection, the spectrograms are the most common feature that is being used to detect spoofing. Each year new GAN models are designed that work towards the creation of realistic fabricated content with fewer inconsistencies, which makes it even more difficult to detect the deepfakes accurately [29].

A. MOTIVATION

Recently, the rise in generated synthetic content has made deepfake generation and detection techniques an interesting area of research. With the rise in data, DL models are used increasingly to support the adversarial learning of deepfakes, which makes researchers create deepfake tools which are seamless in audio and video content creation. The increased usage of these tools has made them a suitable choice for illicit activities at large. Although some positive use cases are also present, but it is mostly used for fake content creation and dissemination. As bad spreads fast than good, deepfake study has become crucial in the current context. There is a requirement to study these systems, their inherent capabilities, and the available datasets that are used in deepfake generation. The past survey articles have mostly presented the basics of AI models in the context of deepfakes and discussed how the models process the data, with the discussion of limited datasets. Deepfake classification is an area that is less researched, and this motivation drives us to present a systematic survey that discusses both deepfake generation and detection models. The survey presents an exhaustive discussion of the underlying principles of deepfakes (in terms of both audio and video generation), as well as current tools for deepfake detection. The survey is supplemented with a solution taxonomy of deepfake detection, and open issues and challenges with future research directions are presented.

B. SCOPE OF THE SURVEY

This subsection outlines the state-of-the-art (SOTA) surveys on deepfakes. TABLE 1 presents a comparative analysis of our survey with other surveys. As indicated, most of the surveys discussed the extent and capabilities of DL and ML-based deepfake models, the adversarial learning requirements of these models, and the open challenges in the quality of the datasets. For example, Yadav et al. [20] conducted a survey about the advantages and disadvantages of deepfakes. The positive and negative implementations of this technology were discussed as well as detection techniques. However, less emphasis is given to the detection portion in the survey, with no reference to audio detection. Ramadhani et al. [8] discussed deepfake generation as well as deepfake detection. The authors provided the taxonomy of deepfake detection techniques. They also have provided a brief list of popular research datasets for deepfake detection. However, audio spoofing detection was not considered part of the survey.

Katarya et al. [21] discussed all three types of deepfake detection which are image, video, and audio. Although the detection section was brief as well as list of datasets available was not included in the survey. In the survey by Tolosana et al. [30], the authors have discussed in detail facial manipulations and their types. They also discussed corresponding detection techniques for each type of facial manipulation. The survey focused mostly on facial manipulation and detection and thus modalities-based detection was not discussed. As a result, audio detection was not taken into account for the survey. Yu et al. [31] proposed an excellent survey on deepfake detection. They have discussed the general process of deepfake generation as well as detailed deepfake detection. They also have also discussed briefly deepfake datasets available for research. Again, audio spoofing detection was not taken into account in the survey.

Mirsky et al. [37] presented a survey on deepfake creation and detection. The authors have discussed deepfake creation explicitly in minute detail. However, the deepfake detection was brief in the survey. All these surveys lacked a discussion of audio spoofing techniques. Nguyen et al. [32] have discussed deepfake generation tools, and the taxonomy of deepfake detection and they have also discussed video deepfake detection in detail. Abdulreda et al. [33] discussed deepfake detection and various facial manipulations. The detection section was not outlined in detail, but the survey presented an excellent taxonomy of the deepfake detection models for all three modalities (images, videos, and audio).

C. REVIEW METHOD

Our survey articles bridge the gaps in previous surveys by discussing of deepfake detection for all the aforementioned three modalities and have considered a systematic literature review of academic databases like IEEE Xplore, ACM, Springer, and ScienceDirect to filter survey articles. In general, the recency of the survey articles is kept (over the period of 2019 to 2023). The reason for this is that recent surveys discuss up-to-date models. The keywords used to filter out the surveys mainly include words like deepfakes, deepfake generation, deepfake detection, types of deepfakes, deepfake models, GANs, AI-based deepfakes, autoencoders, deepfake GAN variants, deepfake applications, and others. Based on this keyword

TABLE 1. Comparison between various surveys.

Author	Year	Pros	Cons
Yadav *et al.* [20]	2019	Discussed advantages and disadvantages of deepfakes and deepfake generation	Deepfake detection was very brief, taxonomy not discussed, audio detection not included, list of datasets not discussed
Ramadhani *et al.* [8]	2020	Presented taxonomy for deepfake detection, covered deepfake video detection, briefly discussed the list of datasets	Deepfake audio detection was not discussed at all
Katarya *et al.* [21]	2020	Briefly discussed image, video, and audio deepfake detection	Not too detailed, list of datasets was also not discussed
Tolosana *et al.* [30]	2020	Discussed in detail facial manipulations, their types, and detection methods	Considered facial manipulation detection methods only, audio detection was not considered, list of datasets was also not discussed
Yu *et al.* [31]	2020	Discussed deepfake detection in detail	A brief list of datasets was discussed, but audio detection was not discussed
Mirsky *et al.* [37]	2020	Discussed deepfake generation in very detail	Deepfake detection was brief, taxonomy was not discussed, audio detection was not included, and datasets were not discussed
Nguyen*et al.* [29]	2021	Discussed deepfake generation tools, Provided taxonomy for deepfake detection, discussed video deepfake detection	Audio deepfake detection was not discussed, list of datasets was not discussed
Abdulreda *et al.* [33]	2022	Discussed facial manipulation	The detection section was brief, neither audio detection was discussed nor dataset list
Masood *et al.* [34]	2022	Focused on the use of ML-based deepfake tools for audio and video contents	Discussion on basics of adversarial training is not presented
Tao Zhang [1]	2022	Discusses the datasets perspective and challenges in deepfake detection, mainly oriented towards benchmarking datasets	potential discussion on GAN-based adversarial training is not presented
Patil *et al.* [35]	2023	Biological classifiers importance in deepfake detection is outlined with distance based metrics	Taxonomy of the classifiers are not presented
Dhesi *et al.* [36]	2023	Usage of adversarial learning techniques in the classification of deepfakes is presented	Further discussion on generating perturbations and the minimization of likelihood errors is not presented
Our Survey	2023	In-depth discussion of taxonomy-oriented deepfake generation and detection for both audio and videos, with an exhaustive listing of datasets, features, and their importance in classification	Signal level feature detection and effective transfer learning approaches are not discussed

criteria, a total of 223 articles are selected (inclusion). Next, we filtered (excluded) based on article keywords that are in line with our survey, and a total of 178 articles are left. Next, we included case reports, scientific blogs, and other web sources (inclusion), which took the count to 211 articles. Next, we performed abstract-based exclusion, which eliminated 74 articles that were not relevant to the survey. Further exclusion of 26 articles is done based on the introduction, the article scope, and the contributions, which finally left us with 111 articles.

D. SURVEY CONTRIBUTIONS

Following are the contributions of the survey article

- The article outlines the deepfake generation models for multi-modal data, where the principal foundations of autoencoders, and GANs (the possible variants) are discussed. Audio deepfake generation is presented with key details of different GAN variants, and pre-trained models in use.
- Post the deepfake generation, the type of deepfake contents and associated body movements are discussed, which forms the founding basis of deepfake detection methods.
- Deepfake detection is presented via a solution taxonomy on three pillars- audio detection, image/video detection, and multi-modal detection with relevant examples. The available datasets for the detection purpose is outlined next.
- Based on the generation and detection models, the open issues and future research directions are outlined.
- A case study based on the incompatibility between multiple modes (IBMM) is presented. The model architecture, the loss functions, and the evaluation parameters of the case study are presented.

E. SURVEY LAYOUT

The layout of the survey is presented as follows. Section II presents the details of deepfake generation techniques. Section III presents the type of deepfake contents used in diverse applications. Section IV presents the multi-modal deepfake detection methods, which is followed by section V that outlines the available deepfake datasets. Section VI

presents the open issues and challenges in deepfake generation and detection, which is followed by section VII which presents the future research directions for deepfakes. Section VIII presents the case study on IBMM which presents the inconsistencies and incompatibilities between different types of input data and models. Section IX presents the concluding remarks and future scope of the survey.

II. DEEPFAKE GENERATION

The section discussed the underlying principles and models for deepfake generation. Deepfake is a method of manipulating images/videos or audio content and generating entirely synthetic content that looks completely authentic and realistic to everyone. DL is the technology fueling this rise of deepfakes [21]. We discuss the image/video deepfake generation, followed by audio-based deepfake generation. The details are presented as follows.

A. IMAGE/VIDEO DEEPFAKES GENERATION

Before deepfakes, images/videos were manipulated using images/video splicing, which was also known as a copy-move forgery. For images, certain parts would be cut and pasted over another region. Thus, images would be manipulated by overwriting another image. For videos, certain frames would be removed or inserted to generate manipulated videos [8]. Mostly, deepfakes use deep neural networks to manipulate images/video content. Two major DL models, i.e., autoencoders and GANs are the driving force for deepfake generation [21], [38]. We discuss the basic principles of these models as follows.

1) AUTOENCODERS

Autoencoders are termed as the earliest model which is used to generate deepfake content [21]. It was first developed in 2017 when it was available as a script, but afterward, it was developed as a user-friendly application known as FakeApp [39]. Traditionally, autoencoders have been used for dimensionality reduction, image compression, and learning generative models. Thus, autoencoders were able to generate the most compact representation of images with minimized loss function than other image compression methods [40]. They are excellent at learning compressed representations of images or latent vectors. Due to this capability, autoencoders were sought as the initial model for the face-swapping method in deepfake generation.

Mathematically, an autoencoder is used to recreate images that it was earlier trained on. The aim of autoencoder is to train a function $E : \mathbb{R}^m \rightarrow \mathbb{R}^l$ (encoder) and $D : \mathbb{R}^l \rightarrow \mathbb{R}^m$ (decoder) to satisfy condition which is presented in equation 1 [41].

$$\arg\min_{E,D} e[\Theta(x, (D \circ E)(x))] \tag{1}$$

where e denotes the expectation over the distribution of x, Θ is the reconstruction loss function, and $(D \circ E)(x)$ denotes a composition $(D(E(X)))$.

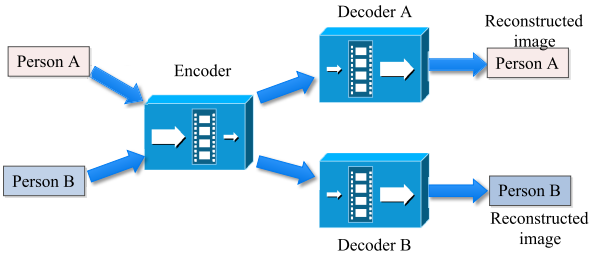

FIGURE 3. Working principle of autoencoders [8].

Primarily, an autoencoder works in three phases: encoder, latent space, and decoder. Encoder is responsible for compressing the image that was provided as input. It compresses the input image while encoding special features such as skin tone, skin texture, facial expression, structure, state of eyes, and other relevant features. These compressed data points are provided to latent space, which helps in learning patterns and structural similarities among various data points. Finally, the responsibility of the decoder is to recreate the original image on which it was trained based on the data points in the latent space. Its job is to create as realistic an image as the original image [40], [42]. These basic phases are also used in the deepfake creation [40].

For the generation of deepfakes, two different autoencoders are trained for two different faces. FIGURE 3 shows the working principle of an autoencoder. Both of the autoencoders share the same encoder but have different decoders. These two autoencoders are used to swap the face of Person A to Person B. To swap the face of Person A to Person B, the image of Person A is compressed using the encoder of Person A, and the decoder of Person B is provided to recreate the image. Thus, an image with the face of Person B could be recreated with the image of Person A [40]. Various deepfake technologies such as DFaker, DeepFaceLab, and TensorFlow-based deepfakes work on this type of concept [21]. For generating deepfake videos, the face of Person A needs to be replaced with the face of Person B frame by frame [43].

2) GENERATIVE ADVERSARIAL NETWORKS

GANs were first proposed by Goodfellow et al. in 2014 [21], [38]. The GAN consists of two networks: the generator and the discriminator. Both networks have their own purpose. The role of the generator is to produce synthetic/fake content. This fake content is then mixed with original content and fed to the discriminator. The role of the discriminator is to differentiate between synthetic and original contents [8]. The feedback of the discriminator is passed to the generator. After multiple cycles/epochs, the generator learns to create synthetic content that looks as realistic as the original content. Once the generator successfully fools the discriminator for $\approx 50\%$ of the inputs, the learning is said to be complete. The model is termed as trained, which can generate synthetic content as close to the real content [21].

These GAN architectures follow the game-playing strategy during the training phase. More specifically, a min-max method is adopted to train the GAN. The networks are competing in the game, where the generator G is presented with a noise vector z which generates synthetic content, defined as $G(z)$. Then this synthetic content and original data (denoted as x) are provided to the discriminator network. At the simplest level, a classification network classifies input into real or fake content. Thus, the output of the discriminator network for real content is denoted as $D(x)$, and for the synthetic content, it is denoted as $D(G(z))$. Thus, the min-max problem solves the value function $V(G, D)$ where G stands for the generator and D for the discriminator, which is presented in equation 2 as follows.

$$\min_G \max_D V(D, G) = Ex_{\sim\rho(x)}[\log D(x)] + Ez_{\sim\rho(z)}] \log(1 - D(G(z)))] \quad (2)$$

where E stands for expectation. The generator wishes to minimize this entire value function, whereas the discriminator tries to maximize this value function. Discriminator wants to maximize the $Ex_{\sim\rho(x)}[\log D(x)]$, which essentially means maximizing $D(x)$ as it suggests high confidence in the prediction of real data as real. The generator is not involved in any case for the first term. Whereas for the second term, which is $Ez_{\sim\rho(z)}] \log(1 - D(G(z)))]$ is for output of discriminator for synthetic content. If we get $D(G(x)$ 0, it means the discriminator is confident that the given input is synthetic and which also means that the generator is failing to trick the discriminator. Thus, the discriminator wants to minimize $D(G(x))$. To make the direction of both terms to be consistent in the equation, $(1 - D(G(z))$ is used. Now the discriminator wants to maximize the second term as well. In the process, generators become better and better at generating synthetic content and essentially fool the discriminator at the end. FIGURE 4 presents the basic generator-discriminator model of GAN.

Thus, GANs are useful for data generative type of applications. The advantage of GANs over autoencoder is that it has wider scope in terms of creating new data [21]. GANs have been used for many applications, such as image-to-image translation, image completion, and text-to-text image generation [44]. Earlier implementations of GANs produced low-resolution images which were often blurred. Karas et al. resolved this issue and came up with a new GAN model termed as progressively growing GAN (ProGAN) [45], which generated high-resolution images of $1024 * 1024$ pixels. It had the concept of progressively increasing generator and discriminator. Kingma and Dhariwal proposed another flow-based generative model [46]. TABLE 2 presents a list of various deepfake generation tools, their objectives, and the type of synthetic manipulations these models can derive from the real inputs.

The concept of the generator-discriminator game has made researchers to present novel solutions in GAN progress for the creation of more realistic deepfakes images/videos. Some

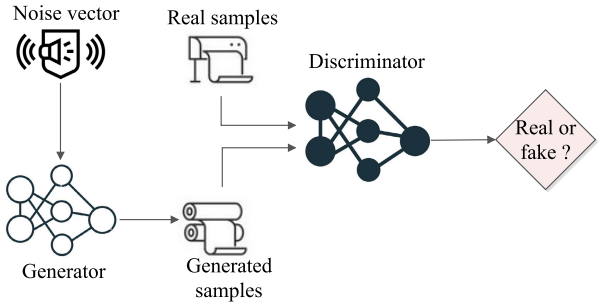

FIGURE 4. The generator-discriminator min-max game in GAN [8].

of the notable examples include the variational autoencoders (VAE) [58], VAE-GAN [59], CycleGAN [54], and others. Autoencoders are trained to recreate images by memorizing them, whereas the VAE generates latent vectors that follow a Gaussian unit distribution. By doing this, it allows us to generate new images by sampling a latent vector from the Gaussian distribution, which could then be passed to the decoder network. The aim of this approach to optimize the loss function. There is an alternate approach for the construction of VAE-GAN. The basic concept of this approach is similar to VAE, but instead of a decoder, we use a generator to recreate images from feature space. This approach uses the discriminator to distinguish the real image from synthetic content. The generator gets better and better at generating more realistic content [41].

CycleGAN is used for image-to-image translation. In this approach, two different generators are used, which are accountable for generating images of diverse domains, i.e., X and Y, respectively. So, it converts the domain X image to domain Y, vice versa. However, it is not feasible to just depend on adversarial GAN loss as it may lose to import features from the original class. Hence, the concept of cycle consistency is applied to the loss function of GAN. The main idea behind it is that the image is converted by the generators from one domain to another, and when reversed back, it should look exactly the same as its original class at the beginning. Thus, GANs become efficient for image translations, such as a change in art style, coloring greyscale images, changing the landscape season in an image, and other powerful utility tasks [41].

Face swapping GAN (FSGAN) [52] is another model which allows users to replace the face of one person with another while keeping the pose and expression intact. In FSGAN, the source image is I_s, and the target image is T_t. Corresponding to I_s, F_s represents the face in the source image, and F_t represents the face in the target image. FSGAN is made up of three different components. The first component consists of reenactment generator G_r and segmentation CNN G_s. G_r accepts heatmaps encoding of the facial landmarks of face F_t, and using these, it generates the reenacted image. It also generates the segmentation mask of the reenacted face. G_s generates the segmentation of face and hair of F_t. These generated reenactment images, however,

TABLE 2. Comparison of images/video-based deepfake generation tools.

Tools	Type of Manipulation	Objective
Face2Face [47]	Expression Swap	Animates the facial expression of target video based on the source input/actor
face swap [48]	Identity Swap	It recognizes faces in images/videos and swaps them
face swap-GAN [49]	Identity Swap	GAN for face swapping
Neural Textures [50]	Expression Swap	Understand neural texture of the target person
DeepFaceLab [51]	Identity Swap/Attribute Manipulation	It allows to swap faces, de-age faces, and also replace head
FSGAN [52]	Identity Swap	It simplifies the manipulation generation process. It eliminates laborious tasks of subject-specific data collection and model training
STGAN [53]	Attribute Manipulation	Useful for editing attributes of the image such as expression
CycleGAN [54]	Attribute Manipulation	Utilizes concept of image-to-image translation without input-output pair i.e. unpaired image-to-image translation
StyleGAN [55]	Entire Face Synthesis	Capable of generating entirely aritificial faces
AttGAN [56]	Attribute Manipulation	Allows manipulation of specific attributes such as applying a mustache, changing the hair color, etc.
Dfaker [57]	Identity Swap	Swaps face. Inputs 64x64 image and outputs a pair of images: One is a reconstructed RGB image, and the second is B/W for mask

might miss a few parts of the face. Thus, face inpainting network G_c is used to estimate the missing pixel values using the target segmentation. The final component blends the generated face into the target image to obtain face swapped image.

StarGAN version 2 (v2) [60] is a GAN-based image translation framework for different domains. The domain is an image set that is classified as a visually distinctive category. StarGAN v2 consists of a generator, discriminator, mapping network, and style encoder. The responsibility of the style encoder is to extract style code from the input image corresponding to the input domain. The mapping network extracts style code from the latent vector and the corresponding domain given as input. Both of these networks benefit from a multi-task learning setup. Finally, the generator network translates an input image into an output image based on the style code provided to the generator from either the mapping network or the style encoder. The task of the discriminator is to predict whether the input image is from its corresponding domain or is generated by the generator.

Another architecture known as STGAN [53] allows users to manipulate certain attributes of the images. For this purpose, STGAN has two components which are the generator and discriminator. The generator is essentially an encoder-decoder pair. Where the encoder is made up of five convolutional layers with kernel size 4 and stride 2, which generates abstract latent representations, and the decoder generates the target image. A selective transfer unit (STU) is applied after the first four encoder layer. On the other side, the discriminator has two different branches. D_{adv} network predicts whether the image is real or fake, whereas D_{att} is responsible for predicting the attribute vector. AttGAN [56] also follows a similar kind of architecture. It also contains a generator which is an encoder-decoder pair, along with an attribute classifier and a discriminator.

Other important models include the face swap-GAN [49], which uses trained autoencoder models to swap the faces of two persons, A and B, where both A and B share the common encoder. The decoders of A and B are different. A commonly shared layer is used for encoding A and B faces, but different decoders allow the separation of A and B faces. For swapping, the key point is to exchange B's decoder with A's decoder unit. Dfaker [57] extends the GAN concept by presenting a multi-domain GAN (MD-GAN) that can learn multiple distributions simultaneously. This means that it can generate synthetic data that is diverse and covers a range of different data types and distributions. For example, it can generate synthetic data for images, text, and structured data such as tables. The benefit of using Dfaker is that it can generate large amounts of synthetic data that are diverse and realistic, which can be used to train ML models. This can be particularly useful when real data is scarce or privacy concerns limit access to real data. DeepFaceLab [51] uses autoencoders for video data. The source video S_v is considered from where the face frame is to be swapped, and the target video T_v is where the face is to be inserted. The autoencoder learns the facial features and movements of the individual videos per-frame unit and then generates a new video where the face of S_v is swapped to T_v. The result is a high-quality face swap video that appears natural and realistic. Next, we present the audio-based deepfake generation models.

B. AUDIO DEEPFAKE GENERATION

Synthetic audio Generation uses similar concepts of encoder-decoder pair or the GANs from the deepfake image/video generation. Synthetic speech generation has

two major types of speech generation, namely- Text-To-Speech and voice conversion. Text-to-Speech needs to accept textual data as input and generate the corresponding synthetic audio. Voice conversion allows users to change the voice of one person to sound like another. The system used for the generation of synthetic speech is known as a vocoder. In Text-to-Speech, all the vocoders at the most basic level utilize the same concept, which is similar to the encoder-decoder pair. So, at the first component, the vocoder needs to encode the textual features into acoustic and prosodic features. The second component then needs to decode these features and synthesize speech waveform. Thus, GAN-based approaches nicely fit the audio generation landscape as well. The modified version of CycleGAN and StarGAN have also been proposed for the tasks of synthetic speech generation. Some of the prominent architectures are discussed as follows.

Char2Wav [61] is one of the synthetic audio generators which accepts text and generates corresponding synthetic audio. It comprises two modules, namely, the reader and the neural vocoder. The reader is essentially an attention-based encoder-decoder pair. The job of the encoder is to accept text or phonemes, which is a bidirectional recurrent neural network (RNN). The responsibility of the decoder is to produce acoustic features for the vocoder, which is essentially an RNN. Vocoder is nothing but a conditional extension of SampleRNN. The responsibility of the vocoder is to generate raw waveform samples out of the intermediate features. Thus, Char2Wav can generate synthetic audio out of textual input.

Another approach, WaveNet [62] combines the capabilities of two different networks, which are Tacotron and WaveNet. The system is composed of two different components where the first component is responsible for predicting the sequence of mel-spectrogram based on input characters. For this, an RNN sequence-to-sequence feature prediction network with attention is used. This first component which predicts spectrogram, is essentially a pair of encoder and decoder. The encoder inputs the sequence of characters and generates an intermediate feature, which the decoder utilizes to generate the corresponding spectrogram. The second component is a modified version of WaveNet. It is responsible for the generation of time-domain waveform samples from the input from the first component.

WaveNet predicts the distribution for an audio sample using a probabilistic approach. Essentially WaveNet calculates the joint probability of a waveform which can be factorized as a multiplication of conditional probabilities of samples of all previous timesteps. It is similar to PixelCNN [63], where a stack of convolutional layers is used to model conditional distribution. The output of the network is of the same time dimension as that of the input. Thus, there are no pooling layers used in the model. The main ingredient of WaveNet architecture is dilated causal convolutions. It ensures the ordering of the emitted predictions. Causal convolutions are faster than RNNs to train. Staked dilated convolutions allow networks to have the very large receptive field while at the same time maintaining the original resolutions of the given input. Causal dilated convolutions also allow models to capture long-term dependency. And another key component is the SoftMax function at the output layer, which makes it a classification problem rather than a regression. The authors have used gated activation as well as residual and skip connections. These enable faster convergence during the training phase as well as adds powerful non-linearity.

WaveGlow [64] is a generative model used to create audio clips by sampling from a distribution. The authors utilized a zero-mean spherical Gaussian distribution with same dimensions, which is desired for the output. These samples are then passed through a few layers that convert the sample distribution to the required flow. A flow-based network is used, and the model is trained with aim of minimizing the negative log-likelihood of data. This network consists of 12 coupling layers and 12 invertible 1×1 convolutions. Affine coupling layers are used to create invertible neural networks. The condition put in the affine coupling layers is that the channels in one half never modify one another. Thus, to mix information across channels, 1×1 invertible convolution layers are used before coupling layers.

Another Network is MelNet [65], which is a generative network with the capabilities to generate audio/music as well as text-to-speech generation. The inspiration was from earlier proposed autoregressive models for images.

Deep Voice 3 [66] is a fully convolutional sequence-to-sequence model for Text-to-Speech generation task. The aim of this model is to predict vocoder parameters which are then used by the audio waveform synthesis model. The architecture is made up of 3 components. The first encoder generates internal learned representations from the input, which are essentially textual features. The second component is the decoder which converts these learned representations from the encoder to low-dimensional audio representations such as mel-scale spectrogram. For this purpose, the decoder uses a multi-hop convolutional attention mechanism. The third component is a converter, whose responsibility is to predict the final vocoder parameters from the hidden states of the decoder. Hence, the converter is a post-processing network. For the optimization problem, the goal is to optimize the linear combination of losses of both the decoder and converter.

A prominent audio generation model, HiFi-GAN [67] is made up of one generator network and two discriminator networks, namely multi-period discriminator (MPD) and multi-scale discriminator (MSD). The generator is just a CNN that accepts mel-spectrogram as input. The job of the generator network is to upsample the received input until the length of the output is equal to the temporal resolution of the raw waveform. It is done using transposed convolutions. The output of these transposed convolutions is followed by a multi-receptive field fusion (MRF) module. The role of the MRF module is to summarize the output of multiple

residual blocks. As for the discriminator, there are two types of discriminators used. MPD is a combination of multiple sub-discriminators, each responsible for handling a portion of periodic signals of input audio. MSD is used to capture consecutive patterns and various long-term dependencies. Since MPD has several sub discriminators which take disjoint samples as input, MSD helps to consecutively evaluate the audio sequence. The loss function is the summation of Feature matching loss, mel-spectrogram loss, and GAN loss.

MelGAN [68] is a non-autoregressive convolutional model that generates audio waveforms using the concept of GANs. It is useful for generating high-quality text-to-speech. As the concept of GANs is used, the model has two networks which are the generator and the discriminator. A generator network is a convolutional feed-forward neural network that accepts mel-spectrogram as input and results a raw waveform for the corresponding input. The resolution of the input mel-spectrogram is lower, and thus stack of transposed convolutional layers is used to upsample the given input. After these transposed convolutional layers, a stack of residual blocks is followed by dilated convolutions. The benefit of dilated convolutions is that the receptive field increases exponentially with each increased layer. Unlike the regular GANs, noise vector is not used by the generator. Weight normalization is used in the generator. Whereas for Discriminator, MSD is used which had three sub-discriminators. The architecture of these three sub-discriminators was identical but each operates on different audio scales. Out of this, one discriminator works at a normal raw audio scale whereas the other two operate at downsampled audio scale by a factor of 2 and 4. These are done using average stride pooling layers. The benefit is that the discriminator trained over downsampled audio does not have access to high-frequency components, and hence that discriminator learns features of low-frequency components. Each discriminator is a Markovian window-based discriminator.

Voice impersonation is also a type of manipulation that differs slightly from voice Conversion in the fact that voice impersonation not only mimics the target speaker in terms of speech and signal qualities but also mimics the style of the target speaker. For this type of task, GAN networks have been used for style transfer. One such approach is using DiscoGAN [69]. DiscoGAN was originally developed for neural style transfer on images. In the case of voice impersonation, there are some constraints that need to be maintained. Firstly, linguistic information must be retained and modified, and secondly, the model must work on variable-sized audio signals. To retain linguistic information, they are mostly encoded in the details of the spectral envelope. The reconstruction loss is modified to ensure that this information is retained. For the second constraint, a variable length input generator and discriminator are needed. The generator of DiscoGAN is capable to handle variable generators, whereas the extracts style information and it ensures that the generated data has this style information in it. Essentially, it ensures that the generated signals have the style embedded into them. Multiple discriminators can also be used to ensure the different types of style aspects.

Non-parallel voice conversion using CycleGAN has also been proposed [54]. For voice conversion, mel-cepstrum, fundamental frequency, and aperiodicity bands are used. Each of these features is processed and converted separately. For the mel-cepstrum conversion, it is firstly divided into higher and lower order sub-components. The higher order is analogous to the spectral fine structure whereas the lower order is to the spectral envelope. Higher-order components s directly copied to converted speech as they do not carry any significant speaker information. In contrast, the lower-order coefficients carries linguistic information and speaker identity. CycleGAN efforts are made on this component. A fundamental frequency ($F0$), a common approach of linearly transforming the source speaker?s log $F0$ by equalizing the mean and the variance of the target speaker?s $logF0$ is used. Aperiodicity is copied without any conversion.

Another approach, StarGAN-VC [70] allows non-parallel voice conversion. The model is often compared to CycleGAN-VC [71], which has a limited one-to-one mapping capability, compared to many-to-many of StarGAN. The generator network of StarGAN learns n:m mappings using just a single pair of encoder-decoder. Contrary to CycleGAN, the generator network of StarGAN uses adversarial loss. StarGAN has a generator that accepts a sequence of acoustic features as well as attribute labels which essentially can be identities of different speakers. Other than this, there is also a domain classifier whose responsibility is to predict the class of the provided input. Thus, essentially, the discriminator's role is to predict the probability of whether the input is real or fake whereas the domain classifier predicts class probabilities. Discriminator and domain classifier both are devised using gated CNNs.

Another approach of parallel Voice Conversion has been discussed for the task of singing known as SINGAN [72]. Essentially this model allows users to convert the voice of singer A to singer B while keeping the same lyrics. For this purpose of singing voice conversion, the authors have proposed their framework based on the concept of GAN. Essentially during the training phase, the concept is that GAN would learn important differentiating features between the source singer and target singer using the discriminative approach. The training phase is composed of three steps. The first step is to perform WORLD analysis to extract spectral and prosody features. Secondly, the authors implement a dynamic time-warping algorithm for temporal alignment of the source and target singer spectral features. Finally, GAN is used for these two aligned singer features. During the run-time conversion phase, first, the features of source singers are extracted using WORLD analysis, then the spectral features are generated using GAN for the target singer (already trained GAN for a singer), and lastly, the singing

waveform is generated by WORLD synthesis. This model can also be used for inter-gender singing voice conversion. TABLE 3 presents a comparative analysis of the tools, their objective, and their potential limitations.

III. TYPES OF DEEPFAKE CONTENTS

The manipulated content can be of any form that has its own challenges while detecting them correctly and the risk of damage that it can do to an individual/organization. Various deepfakes generate manipulated content with diverse approaches. Deepfake content can be categorized as image, video, and audio deepfakes. Although image and video deepfakes can be of the same category as image deepfakes, where the frame-by-frame manipulate video content are essentially images [39]. There are deepfakes that manipulate audio as well as video content altogether to create lip-sync videos [21], [42]. The description and objectives of deepfake are mentioned in TABLE 4.

Although, for image and video manipulations, there exist different techniques. For creating image and video deepfakes, majorly face manipulations have been preferred. For face manipulation, there are multiple diverse types of manipulations possible, which are categorized as entire face synthesis, identity swap, attribute manipulations, and expression swap [30]. Although full-body puppetry can also be seen in deepfakes manipulated videos [42]. Full body puppetry means the movement of a person's body is transposed over another person's body. Let?s briefly discuss four categories of face manipulations one by one. *Entire face synthesis* means that an entirely new synthetic face is generated which is actually nonexistent. Generally, this is achieved using some powerful GAN, such as *StyleGAN* [73]. This category of deepfake can be used to create fake profiles on different online platforms. The second type of deepfake is *identity swap* which means that the face of another person is swapped over in a video to create fake videos/images. Its perfect example is *ZAO* mobile application that replaces the face of any popular person with anyone's face. It can be used for spreading fake news [74] as well as questionable content on social media. The third type of face manipulation technique is *attribute manipulation*. In this manipulation, certain attributes of the face such as skin color, hair color, or gender can be manipulated, which is known as face editing or face retouching. The fourth and final type of face manipulation is the *expression swap* where the facial expression of a person is manipulated. Its examples are *Face2Face* and *NeuralTextures* applications.

There is some manipulations that can be performed on audio files as well. The classic examples are voice-swapping and text-to-speech [42]. In *voice-swapping*, the voice of any person can be manipulated and changed to create forged content. Whereas, in *text-to-speech* technique, the contents in the audio can be changed by typing in the text. In Voice Conversion, these manipulations have a subdivision of their own, such as parallel and non-parallel voice conversions. In parallel conversion, the voice of the source target will be changed to sound like the target speaker, but the contents of the speech remain the same. Whereas for non-parallel conversion, the speech contents can be changed from the source audio along with the voice. It could be dangerous to create fake news as well as scam people by sending recordings of the people they know. And then there are types of manipulations done on both video and audio content to create extremely realistic lip-sync videos. Its viable example is *Obamanet* where you can create extremely realistic lip-sync videos [21].

All above-mentioned deepfake manipulations have their own set of challenges while detecting the deepfake. It becomes extremely difficult to come up with a system that can detect all types of manipulations accurately over different deepfake generation approaches. It is an active field of research nowadays and the researchers are thinking of coming up with a system that has generalizability in terms of detecting deepfake content. Different approaches have been used to detect deepfake content and they are discussed in detail in subsequent sections.

IV. DEEPFAKE DETECTION

Various deepfake detection approaches have been used by researchers to detect deepfake content (as shown in FIGURE 5). When these deepfakes are created, certain inconsistencies and traces of manipulations are left behind. The detection systems could use these traces to classify whether it is deepfake manipulated or authentic content [8]. At the same time, the detection needs to be done for images/videos as well as audio content. To detect such manipulations, the detection approaches are briefly classified as image/video-based detection, audio spoofing detection, and multi-modal-based approaches. Image/video-based detection is the most commonly used approach for investigating deepfake content. Generally, the inconsistencies left during the generation phase of these contents are leveraged to detect them. For image/video-based detection, three different approaches are proposed in the literature. They are physical/physiological-based approaches, signal-level features-based approaches, and data-driven models. In physical/physiological-based approaches, the visible inconsistencies in the image/video are leveraged, such as inconsistent head pose, blinking pattern, or incomplete visual artifacts. In contrast, in signal-level feature-based approaches, image feature descriptors/algorithms are used to extract deep signal-level features, which are then used to detect real or fake content. Whereas in data-driven models, specific models are trained over a large volume of such original as well as synthetic content and after some time, these models learn to differentiate real and fake content.

In audio spoofing detection, manipulations performed on audio contents are analyzed. There have been competitions, such as ASVspoof 2017 [75] and ASVspoof 2019 [76] challenges, where manipulated datasets are released along with some baseline solutions. And researchers are invited to improve over the provided baseline solutions. Manipulations

TABLE 3. Comparison of audio-based deepfake generation tools.

Tool	Type of Manipulation	Objective	Limitations
Char2Wav [61]	Text-to-speech	Generate speech audio from text input using a deep neural network architecture that learns to predict the spectral features of speech signals.	Limited to the English language requires careful tuning of hyperparameters to achieve high-quality results.
Wavenet [62]	Audio generation	Generate high-quality audio samples by modeling the waveform directly using a generative model that employs dilated causal convolutions.	Requires significant computational resources for training and inference, can be slow to generate samples.
Waveglow [64]	Audio Synthesis	Synthesize speech or singing voices using a flow-based generative model that maps Gaussian noise to mel-spectrogram representations.	Requires pre-training of a WaveNet vocoder, which can be sensitive to noise in the input signal.
Melnet [65]	Speech synthesis	Generate high-quality speech audio from text input using a deep neural network architecture that learns to predict the mel-spectrogram features of speech signals.	Limited to the English language requires significant computational resources for training and inference.
DeepVoice 3 [66]	Text-to-speech	Generate human-like speech from text input using a multi-speaker generative model that incorporates phonetic, prosodic, and speaker embedding features.	Limited to a certain pre-defined speaker and language combinations, can require extensive fine-tuning to achieve high-quality results.
HiFi-GAN [67]	Audio synthesis	Generate high-quality audio with high fidelity by training a generative adversarial network (GAN) to learn the mapping from a low-resolution waveform to a high-resolution waveform.	Can require a large amount of training data and significant computational resources, can be sensitive to noise in the input signal.
MelGAN [68]	Audio synthesis	Generate high-quality speech or music by training a GAN that learns to generate mel-spectrogram representations of audio signals and then converts them back to the waveform domain.	Can be slow to generate samples, can be sensitive to noise in the input signal.
DiscoGAN [69]	Image-to-image translation	Convert images from one domain to another (e.g., horses to zebras) using a GAN architecture that learns to map images from one distribution to another using cycle-consistent loss.	Can be sensitive to noise in the input signal, requires significant computational resources for training and inference.
CycleGAN-VC [71]	Voice conversion	Convert voice characteristics from one speaker to another using a GAN architecture that learns to map the source speaker's mel-spectrogram features to the target speaker's mel-spectrogram features while preserving linguistic content.	Limited to certain pre-defined speaker and language combinations, can require extensive fine-tuning to achieve high-quality results.
StarGAN-VC [70]	Voice conversion	Convert voice characteristics across multiple speakers using a multi-domain GAN architecture that can perform many-to-many voice conversion with a single model.	Limited to a certain pre-defined speaker and language combinations, can require extensive fine-tuning to achieve high-quality results.
SINGAN [72]	Image generation	Generate high-quality images with high resolution by training a GAN that uses a progressive growth strategy to increase the resolution of generated images over time.	Can require a large amount of training data and significant computational resources, can be sensitive to noise in the input signal, and may suffer from mode collapse during training.

TABLE 4. Types of deepfake contents [42].

Type	Manipulation	Objective
Photo	Face and Body-Swapping	Replacing the face of one person over another to create manipulated content
Audio	Voice Swapping	Changing the voice of one person to another
Audio	Text-to-Speech	Changing the content of the audio by typing in the text
Video	Identity Swap	Face of one person is swapped with another person?s face to create manipulated videos
Video	Face Morphing	Face of one person seamlessly transforms into another person?s face
Video	Full-body puppetry	Transposing movement from one person?s body to an other
Audio and Video	Lip-Syncing	Creating an extremely realistic lip-sync video

in audio data could also contribute to the spreading of fake news as well as defamation cases [77]. Another modality that also needs to analyze to detect manipulations. On the other hand, the third type of approach is a multi-modal approach which aims at utilizing more than a single modality to detect manipulation in the content. Multiple/different modalities analyze audio, video, as well as frame level, features all together in a single system to detect manipulations. This kind of approach uses more than one modality to analyze and detect manipulations. Each of these approaches is discussed in detail, along with corresponding research done in that direction.

A. IMAGE/VIDEO DETECTION

The fundamental idea of this approach is that the features from images and videos are extracted, which can be used to differentiate between authenticated and synthetic content. The main reason for these inconsistencies and gaps in deepfake creation is because of upsampling and affine transformation processes are performed. These inconsistencies could be something like resolution inconsistencies between the manipulated area and other areas of the image, it could be the incompatibility of the manipulated area with other areas of the image, or it could be a temporal discontinuity in videos, etc. These are the features that are being used to

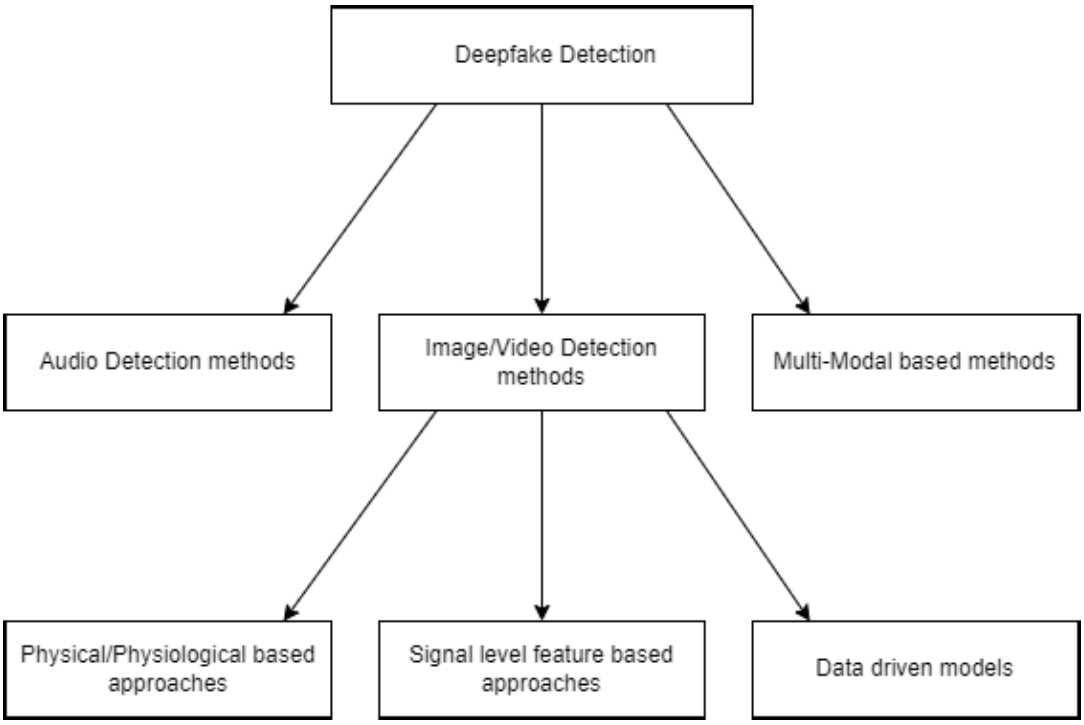

FIGURE 5. Taxonomy of deepfake detection approaches.

detect deepfake contents [8], [78]. In general, three different approaches are used to detect video deepfake content, which are physical /physiological discrepancies, signal level features, and data-driven models. Each of them is described in the following subsections.

1) PHYSICAL/PHYSIOLOGICAL ATTRIBUTE-BASED DETECTION

In this approach, physical inconsistencies are detected, which are left behind during the deepfake generation process. Thus, all inconsistencies are visible by the naked eye and are leveraged to detect whether the content is manipulated or authentic. These features could be inconsistency in the blinking of the eye in a manipulated video, color mismatch of face to rest of the body, lighting on manipulated are to other parts of the images, etc. [8]. One such example is [79], where inconsistency in the head pose was calculated to detect whether it is manipulated content or not. They used 68 facial landmarks to compare them with 17 landmarks that indicated pose directions. The 68 facial landmarks indicated the pose direction, whereas 17 landmarks indicated the pose directly from the center of the face. If both of those pose directions turn out to be different, it is classified as manipulated content.

Another approach was discussed in [38] by Li et al., where they tried to find inconsistencies in eye blinking. The assumption they made was that there is a difference in blinking patterns between authentic video and deepfake generated video. Another approach where multiple visual features called visual artifacts was proposed in [46] by Matern et al. In this paper, for visual artifacts, multiple features were used, such as the color of left and right eyes, reflection of light on the surface, inconsistencies and disproportionate shadows, or geometry of areas of image/video which are not properly detailed, such as teeth. These are some of the approaches utilizing physical/physiological or visual artifacts to detect deepfake content. But, the problem with this approach is that, with development in deepfakes, these visual artifacts are difficult to find. Thus, this type of deepfake detection is becoming obsolete.

2) SIGNAL-LEVEL FEATURE-BASED DETECTION

This type of detection uses feature extraction that is added to the content during the synthesis phase of the content. Thus, very local features are used at the pixel level to detect deepfakes [8]. Spatial features are related to visual inconsistencies, whereas steganalysis helps to extract features at a low level, extracting hidden information from the image/video. As discussed in [80], the image convolution, as well as the steganalysis feature, is used to detect manipulated areas in the image/video. It uses two streams and the first stream is GoogleNet for face classification. Whereas the second stream is a patch triplet stream used for local noise reduction. There is another study as well where Photo Response Non-Uniformity (PRNU) analysis with cross-correlation is used. But it only used 10 videos in the study [8].

In [78], the authors proposed a new approach for feature extraction known as Scale Invariant Feature Transform

TABLE 5. Literature review of image/video deepfake detection methods.

Types of Approach	Author	Year	Approach	Limitations
Physiological attributes-based detection	Li *et.al* [38]	2018	Detected deepfake videos using eye blink pattern	The assumption is that there is a time difference between the original and deepfake videos. Current advanced deepfake videos are arduous to detect with visual features
Physiological attributes-based detection	Yang *et.al* [79]	2019	Tries to use inconsistency between the head pose and other body parts using facial landmarks	Again visual features are not reliable with advanced deepfakes
Physiological attributes-based detection	Matern *et.al* [46]	2019	Used visual artifacts such as differences in eye colors, and disproportionate shadow in detecting deepfakes.	Again visual features are not reliable with advanced Deepfakes
Signal level features-based detection	Li *et.al* [78]	2020	Another feature extraction method is known as Scale-invariant Feature Transform (SIFT). It detects keypoint pixels and extracts their features.	It is a good method but with advanced DeepFakes, even local features are very difficult to detect
Signal level features-based detection	Zhou *et.al* [80]	2017	Two streams: (1) GoogleNet for Face Classification (2) Patch Triplet Stream for local Noise residual	Performs well but not with advanced DeepFakes
Signal level features-based detection	Kharbat *et.al* [81]	2019	Used multiple feature point descriptors such as HOG, BRISK, KAZE, etc for deepfake detection. HOG provides an accuracy of 94.5% with the SVM classifier	Hard to extract features With advanced deepfakes
Data-driven Models	Marra *et.al* [82]	2018	Performed a comparative study over InceptionNet, DenseNet, and XceptionNet. XceptionNet perform performs best	Lack of generalizability
Data-driven Models	Lee *et.al* [83]	2018	Proposed a five-layer CNN architecture called deep forgery discriminator	Lacked generalizability
Data-driven Models	Afchar *et.al* [39]	2018	It is a CNN model which utilizes the Inception module as the backbone of its architecture	It can work well with compressed videos but Xception outperforms it in every dataset
Data-driven Models	Guerra *et al.* [40]	2018	Tried to use RNN, to utilize Temporal features of video	Can only be used over videos. Other architecture outperforms it
Data-driven Models	Amerini *et.al* [84]	2019	Tried to exploit discrepancies in motion across frames at f(t) and f(t+1). And used CNN as a classification algorithm	Other algorithms outperform it

(SIFT). It detects all the key point pixels in the image and extracts its feature to classify whether it is manipulated or not [78]. Similar to SIFT, there are other feature-detector-descriptors used as well. Reference [81] used SIFT, HOG, ORB, BRISK, KAZE, SURF, and FAST to extract features. These features are then provided to the SVM classifier to classify the manipulated content. Where it turns out that it can be useful to detect manipulated contents with it. HOG turns out to provide 94.5 % accuracy with SVM classifier [81]. Another study was performed in [85], where firstly the frames are extracted from the videos and then the faces are detected using the Viola-Jones algorithm. Then, those extracted faces were normalized and then they were used to extract features via a few local image descriptors they considered for the study. They considered local binary pattern (LBP), local phase quantization (LPQ), pyramid of histogram of oriented gradients (PHOG), binary gabor pattern (BGP), speeded up robust feature (SURF), binarized statistical image features (BSIF), and image quality metric (IQM). The study focused on analyzing which method performs best for feature extraction to detect deepfakes. Among all, the IQM performed best [85] compared to PCM and LDA [8].

3) DATA DRIVEN MODELS

In this approach, instead of focusing on specific features or artifacts, DNNs are trained over the dataset of fake and authentic images/videos. Its prominent examples are MesoNet, XceptionNet, and MobileNets. With advanced deepfakes, instead of choosing features that can be used for deepfake detection, we use DNNs to extract more complex features. Based on the study [82] performed in 2018 in which InceptionNet, XceptionNet, and DenseNet were compared, where XceptionNet outperforms all of them. There exist, other models, such as deep forgery discriminator (DeepFD) which is a 5-layered CNN architecture for which the loss function used was in the form of contrastive loss. It provided impressive results [83]. The results of this DeepFD were further improved by using a pairwise learning approach. Due to this, it became possible to increase the generalizability of the model [8]. Another model was developed using this pairwise learning approach which calculated the difference between features of original as well as fake images using contrastive learning. It is called a common feature fake network (CFFN). Another CNN-based model was proposed known as MesoNet [39]. It uses InceptionNet as the backbone of its architecture and it is capable of detecting compressed images as well which can be useful for detecting deepfakes on social media. Capsule Network architecture has also proved to match the performance of MesoNet. Capsule Network consists of 3 primary capsules whereas it has 2 output capsules. The input from VGG-19 is given as input to capsule Network [86]. Capsule Network solves the weakness of unequivariance convolutional blocks.

To leverage the temporal component of video, many researchers have proposed RNN-based solutions in Ref. [40]. It has also been re-investigated by the researchers by using bidirectional RNN where features from DenseNet frames were provided. Other than RNN, the authors have also proposed an OpticalFlow model, which uses the CNN classification algorithm. Later, in [84] the authors proposed a model that uses a vector field across from $f(t)$ which essentially is a vector field at frame t to $f(t+1)$. So basically, the model tried to find discrepancies in motion across frames and CNN was used as a classifier [84].

TABLE 5 describe the comparative analysis of various existing image/video deepfake detection schemes.

B. AUDIO DETECTION METHODS

Audio contents are also manipulated to change one person's voice to another, or the contents of what a person is speaking might have been manipulated, or the audio might have been slowed down or it might have been speeded up. There is a common term known as Cheapfakes for these types of manipulated contents [21]. To detect whether the audio content has been manipulated or not, we need a high-level representation of the audio. So, the researchers classify whether the content is manipulated or not. There is an open-source tool known as Resemblyzer [21], which allows researchers/developers to find a high-level representation of the audio and compare it with two different samples. Generally, audio manipulation detectors used to use spectrograms to detect whether it is original or fake audio. They try to find the difference between spectrogram, which essentially is a visual representation of the audio content between original and fake audio. Thus, it could help to detect whether it is fake or original audio content [21], [87].

Generally, audio spoofing/manipulation can be classified into two different types: logical access (LA) and physical access (PA). LA has two different types of manipulations and they are Text-to-Speech (TTS) and voice conversion (VC). TTS generates entirely new human-like synthetic audio of the text that is entered into the tool. VC will convert the voice of one person to another while keeping the content of the speech same. Whereas PA consists of a replay attack, which means the prerecorded audio is used to impersonate the original speaker [26]. Generally, gaussian mixture model (GMM) classifiers were used to perform static feature analysis. But in a recent study, DNNs are used to extract features and are trained over datasets. They classify the audio as fake or original. These approaches outperform the traditional GMM approach [21].

Many researches have been conducted to detect audio spoofing. Also, different competitions have also been organized where the host releases a development dataset and baseline models. These baseline models are then used to compare the submissions in terms of how much they improve performance over these models. In [88], the authors proposed a novel human log - likelihood (HL) scoring method. They also have proved mathematically that their HLL scoring method is better than the baseline LLR method for audio spoofing detection. They then experimented with their new scoring method and implemented a DNN-HLL model trained over five different spectrum-based cepstral coefficients. They have shown improvement in the performance over the baseline model GMM-LLR and from various features, CCQC based DNN-HLL model performs the best on ASVspoof 2015 dataset.

Authors in [89] provided a solution for ASVSpoof 2019 challenge. The authors focused on creating a generalized model that performs well even when the type of attack is unknown. The authors focus on using acoustic features which can be derived using long-term constant-Q transform (CQT). Moreover, CQT offers higher frequency resolution for lower frequencies and higher temporal resolution for higher frequencies, unlike discrete fourier transform (DFT). The authors also investigated deep features using the DFT log power spectrum as input. The author aims at the fusion of multiple different systems that utilize these features. The outcome/conclusion turns out that deep features work efficiently for physical access attacks. Whereas long-range features work for logical access attacks and fusion of different systems improves performance [89].

Alzantot et al. [90] proposed a deep residual neural network for audio spoofing detection. They proposed three different variants of the model, which are identical except for the fact that they deal with different kinds of input features. These variants are MFCC-Resnet which takes MFCC as input, CQCC-Resnet which takes CQCC as input, and Spec-Resnet, which takes log-magnitude STFT as input. The model is inspired by the efficiency of residual convolutional networks in classification problems. These models are solutions for the ASU spoof 2019 competition. The model for Spec-ResNet takes input and passes the input to the 2D convolution layer with 32 filters. The output of this 2D convolution layer has 32 channels and it is then fed to a sequence of 6 residual blocks. The output received from the last residual block is passed through the dropout layer, which is received by a fully connected layer that has a leaky - ReLU activation function. The output of this fully connected layer is passed to a new fully connected layer with two units. These classification logits are transformed into probability distribution using the Softmax layer. The residual blocks are also discussed in detail. These three models have greatly improved t-DCF and EER than baseline algorithms.

In the other method proposed by Ling et al. [26] was built upon standard ResNet-based architecture as proposed in [91]. The standard ResNet architecture has first a spectrogram extracted, which passes through certain residual blocks, and then passes through the pooling layer and then loss functions are applied. In [26], the authors have built their system considering the original CNN architecture. They have added an attention module over each residual bottleneck of ResNet architecture. This attention module comprises two separate blocks, the Frequency attention block and the

channel attention block. These two blocks help to capture the frequency relationship and channel relationship. At each attention block, for frequency attention block specifically, the attention module for given input f_i generates inter frequency relationship matrix and creates weighted features by sequential matrix multiplication. And then, by element-wise summation, it outputs frequency refined feature f_f. And with a similar process also yields refined channel features f_c. This residual learning stabilizes the training process. Moreover, they use attentive temporal pooling. It helps to prioritize certain sections of input [26].

In [93], the authors leveraged the capabilities of CNN and RNN. The architecture used was made with 4 convolutional layers, followed by a mass pooling layer and a dropout layer. After this, a flattened layer is followed. This layer aimed to extract different features of the audio clip. Following upto this were 4 dense layers. And then, finally, an activation layer. It was responsible for the binary classification of fake and real audio. In [24], Nasar et al. proposed a CNN-based model comprising four different modules. These were the data preparation module, image Enhancement, CNN model generation, and detection Phase. This system could detect an image, video, and audio content. The pre-processing of content is performed based on that. For audio files, the spectrogram images will be generated. And then, this image will be used for the classification problems. These images were given to CNN Networks, consisting of convolutional layers, 3 max-pooling layers, and 5 activation layers. The accuracy achieved was around 90%. In [25], the authors proposed a special loss function that improves performance for audio spoofing detection. The authors utilized deep residual ResNet-18, replacing the global average pooling layer with attentive temporal pooling. The model takes extracted LFCC features as input. Experiments show that the loss function proposed outperforms the traditional Softmax and AM-Softmax. The authors claim at that time that their model outperforms all existing single models, not including fusion/ensemble models.

Another approach to detect audio spoofing is by ensemble models where the capabilities of multiple independent models are used to improve performance. In the paper [28], Chettri et al. proposed three ensemble models, E1, E2, and E3, using deep and shallow classifiers. In deep classifiers, 2 CNN, 1 CRNN, 1-D CNN, and 1 Wave-U-Net is used, whereas, in shallow models, 3 GMM models trained over MFCC, IMFCC, and SCMC and 2 SVMs trained using 1-vectors and long-term-average spectrum (LTAS). So, the authors proposed 3 ensemble models with different combinations of these 10 different classifiers. These models perform better than baseline models, namely LFCC and CQCC feature-based GMM models.

In [92], the authors proposed a hybrid approach in detecting spoofed audio. The system has two arms to predict whether the audio is spoofed, and then a combined prediction is made. For each arm, a separate GMM-UBM model is built. The extracted audio features are post-processed for the first arm and sent for prediction. Whereas for the second arm, the audio features are given as input to the autoencoder and then processed. And at the prediction of both, the arms are sent to a fusion module. This module calculates a hybrid estimation of authenticity. Another ensemble approach to detect audio spoofing by Dua et al. [27] The authors utilized three features, namely MFCC, IMFCC, and CQCC. At the same time, three models are used, which are TD-LSTM-DNN, TC-LSTM-DNN, and tSC-LSTM-DNN. The architecture of tD-LSTM-DNN consists of 3 dense layers with 24, 16, and 10 units. Then it is followed by a dropout layer. After that, 3 LSTM layers with 10, 20, and 30 units are placed sequentially. And at last, another drop-out layer before the output layer. The architecture of TC-LSTM-DNN consists of two 1D convolutional layers with 48 and 32 units, respectively. Then a dropout layer follows, followed by three LSTM layers sequentially with 24, 32, and 48 units. The model utilized a temporal convolution. And for the architecture of tSC-DNN, which has one layer of 2D convolution, after which batch normalization and another 2D convolutional layer follow up. After that, it follows a max pooling layer whose output is fed to the flattened layer and then to the base layer. A dropout layer follows it and then one more dense layer and flattened layers follow up. To calculate the final score, the ensemble model calculates a weighted confidence score using the scores of these three models.

TABLE 6 describe the comparative analyis of various existing audio spoofing detection techniques.

C. MULTI-MODAL APPROACHES

Another type of approach is used where authors have tried to bring both modalities together instead of just focusing on either video or audio. A single system detects both types of manipulation and sometimes even uses disharmony between the two modalities as an advantage to improve performance. Chintha et al. [94] proposed a system that detects deepfake videos and spoofed audio. But, both modules work separately in detecting the corresponding manipulation. For Video detection, authors have proposed 4 different variants of the model. So, the architecture is based on XceptionNet and recurrent processing used in ConvLSTM, and FaceNetLSTM. These variants are Xceptemporal(CE), XcepTemporal(KL), XcepTemporal(EN), and XcepTemporal(EN1+n). XcepTemporal(CE) and XcepTemporal(KL) differ by layer based on the loss function. Whereas, XcepTemporal(EN) classifies into two classes, i.e., real and fake. Whereas XcepTemporal(EN1+n) also predicts the type of fake content. The basic concept of the architecture was extracting faces from images, which are then given as Input to XceptionNet and then passed to a sequence of Bidirectional LSTMs. The layer changes based on the type of variant of XcepTemporal work exceptionally well on Face forensics ++ and Celeb DF. Thus, for the audio spoofing detection, two different modules were investigated. They were CRNN Spoof and WIRENet Spoof. For CRNN

TABLE 6. Comparison of existing audio spoofing detection techniques.

Author	Year	Approach	Advantage	Limitation
Yu *et al.* [88]	2017	Authors have proposed a new human log-likelihood scoring method.	Performs well on the ASVspoof2015 dataset and outperforms baseline GMM-LLR Models	Newer datasets have more advanced spoofed audios which are more difficult to detect than the ASVSpoof2015 dataset
Chettri *et al.* [28]	2019	Three ensemble models were proposed which were different combinations of 5 deep classifiers and 5 shallow classifiers	Performs well over LA attacks, Improvement over baseline models of competition	Can perform well over LA attacks but needs improvement over PA attacks which essentially suggests a lack of Generalizability
Das *et al.* [89]	2019	The authors investigate acoustic features extracted using long-term CQT as well as deep features. Also investigates the fusion of multiple models	Fusion of models increases performance, CQT offers few advantages over traditional DFT	Improvement in terms of accuracy and generalizability still exists
Balamurali *et al.* [92]	2019	A hybrid approach using two audio spoofing detection arms and then combining results of both arms to calculate the estimation of authenticity	Investigated almost 11 different types of features and pinpointed which features are most valuable for audio spoof detection in the wild	Recent datasets will have higher generalizability challenges then ASVSpoof2015. Moreover, other models outperform the proposed one
Alzantot *et al.* [90]	2019	Implemented three variants using ResNet based on input features	Improves Performance than baseline algorithms of ASVSpoof 2019 competition	Models perform better on PA subset of ASVSpoof 2019 dataset but performance drops over LA subset because of lack of generalizability ability
Wijethunga *et al.* [93]	2020	CNN and RNN-based architecture for audio spoof detection	The system has noise filtering module and speaker diarization module other than just synthesis detection	Outcomes of the experiments were not satisfying in terms of performance
Nasar *et al.* [24]	2020	Converts audio into spectrogram image and CNN-based architecture is used for classification	Simple Architecture	Lack of generalizability to detect various deepfakes
Dua *et al.* [27]	2021	The ensemble model is made up of three different models namely tD-LSTM-DNN, TC-LSTM-DNN and tSC-DNN. Finally calculates a weighted confidence score based on output of these three models	Ensemble models improve performance over single models	Models could be trained over different feature vectors and deeper CNN or LSTM could be investigated to improve accuracy
Zhang *et al.* [25]	2021	Proposes a special loss function which improves performance. ResNet-18 is used where the average global pooling layer is replaced by attentive temporal pooling	The proposed loss function outperforms Softmax and AM-Softmax	Ensemble models outperform single models
Ling *et al.* [26]	2021	Built upon ResNet but inserted attention module over each Residule bottleneck	Reduces redundancy of information, Attentive Temporal pooling	The experiment was performed over the LA subset of the ASVSpoof 2019 dataset only, Also other ensemble methods outperform proposed fusion methods

Spoof, it comprised 5 convolutional layers responsible for feature extraction and downsample input audio. Followed by a Bidirectional LSTM layer. The the output of these layers was fed to two fully connected layers sequentially and then followed by a log softmax function. Whereas WIRENet stands for wide inception residual network (WIRENet) Spoof. This architecture uses stridden convolution and max pooling operations to lower the size of the feature map. To obtain log-mel spectrogram, the audio clip is repeated to the fixed length of 4 seconds. Then a sequence of Wideblocks is used, which captures different levels of temporal information. To avoid overfitting the model, batch normalization and Dropout layers are used.

Agarwal et al. in [95] tried to detect lip-sync videos using the inconsistencies in Phoneme and Viseme. Visemes can be defined as the dynamics of mouth shape that forms during speaking of a particular phoneme. So, the authors have tried to leverage the inconsistency in Viseme with the spoken Phoneme. This essentially means that in general while speaking mother, brother or parent, it requires your lips to have to be completely closed, But, this type of Phoneme to Viseme mapping is not properly duplicated by the lip-sync generating tools. And thus, using this inconsistency, lip-sync videos can be detected. To detect, authors have selected M,B,P phoneme group (MBP). Now, for extracting location of phonemes, Google's speech-to-text API is used to generate the transcribe of audio track associated with video automatically. Then, the transcribe is manually checked for error and realigned to audio using P2FA. During this process, the phoneme along with their occurred is recorded. Then, the corresponding viseme is extracted. For extracting viseme three different approaches have been used. They are manually hand crafted profile feature or using CNN which predicts either mouth is closed or not. For classifier, XceptionNet is trained with Adam optimizer and cross-entropy loss function. CNN works much better than hand crafted profile features.

In [96], the authors lomnitz et al. have proposed a multi modal based deepfake detection which aims at investigating manipulation detection from various perspectives. The authors have proposed a system which is made up of different modules where these models investigate manipulation at single frame level, multi frame level, audio level and power spectrum. For a single frame level, multiple models such as XceptionNet, VGG-19 and deeper ResNet-152 were experimented with. Whereas for multi level frame level, XceptionNet based model was used since it was best performing for single frame detection. So, the model based on Xception Net which was followed up by two Bidirectional LSTMs sequentially. For audio detection, SincNet-based model was used. For power spectrum, authors extracted features from images spatial power spectrum and used a MLP to classify it into real or fake. Now, the results of these individual models were combined using a learnt weighted average as a part of a MLP.

Chugh et al. [97] proposed a system where the detection of deepfake content is based upon dis-harmony between audio and video contents. Thus, they proposed a modality dissonance score (MDS), which is used to classify content between real and fake. There are two different streams for audio and video content. The video and audio contents are divided into segments of length of 1 second. For visual stream, model similar to 3 D-ResNet is used. The feature representation extracted by the visual stream is used to calculate contrastive loss. And also, a 2-neuron layer is added at end to classify between real and fake. For Audio, MFCC features are used to detect manipulation. The first 13 MFCC are used and represented as heatmap. The, further architecture is based on CNN for image classification. The output from these networks is used to calculate Cross-Entropy Loss. Whereas contrastive Loss is employed to enforce higher dissimilarity between audio and video. And finally MDS is calculated by combining both streams dissonance and it is used to label content as either real/fake [97]. Whereas in another approach, Fan et al. have proposed a system with two neural networks responsible for processing audio and video content. Again, over here as well, for audio processing, first 13 MFCC features are extracted and represented in form of heatmap. Now both the models are processing image content. For video, inputs are down sampled whereas for audio, inputs are up sampled. This architecture also processes video and audio contents in segments of 1 second. Thus, finally a Video Confidence Score (VCS) is used to classify content between real/fake [98].

TABLE 7 describe the comparative analyis of various existing multi-modal approaches for audio/video spoofing detection. Also, TABLE 8 shows the comparative study of existing deepfake detection models.

V. DATASETS

There are certain datasets which are generally used for research purposes of deepfake detection of images and videos. They are mentioned in this section. They are:

- *UFDAV*: This dataset consists of 49 original videos which are form the YouTube and other 49 deepfake generate videos [79].
- *FaceForensics*: It is made up of 2008 manipulated content and 1004 real content. It has two types of dataset. First is Source to Video is created using traditional Face2Face method where two randomly chosen videos are reenacted. Whereas in second type is Self Reenactment where the manipulation is done on the same source video to generated pair of real as well as manipulated content [99].
- *FaceForensics* ++: It is also referred to as FF-DF. It contains 1000 original videos from YouTube whereas 1000 generated videos using faceswap [100].
- *Celeb-DF*: It consists of 1203 contents where 795 are manipulated content and 408 are authentic content. deepfake videos generated out of 590 original videos using faceswap [78].
- *Celeb-DF v2*: It consists of 5639 deepfake videos generated out of 590 original videos using faceswap [78].
- *DF-TIMIT*: It is divided into two different sets of videos. They are DF-TIMIT-LQ and DF-TIMIT-HQ. DF-TIMIT-LQ contains videos of lower quality whereas DF-TIMIT-HQ contains videos with higher quality. There are around 640 videos in this dataset generated using faceswap-GAN [101].
- *Google/Jigsaw Deepfake Detection (DFD)*: This dataset contains 3068 deepfake videos generated out of 363 original videos. These videos are generated with 28 individuals of different gender, age and race [102].
- *Facebook Deepfake Detection Challenge Preview (DFDC Preview)*: This dataset has 4113 deepfake videos generated out of 1131 original videos which include 66 individuals with different gender, age and race [103].
- Facebook Deepfake Detection Challenge (DFDC): This dataset consists of 104,500 deepfake content and 48,190 original content [104].
- *DeeperForensics* 1.0: It consists of 10,000 manipulated content as well as 50,000 Original content. 35 perturbations has been performed and it was created using various actors of different age, gender and race [105].
- *WildeDeepfake*: It is a relatively small dataset made of 707 deepfake contnet. It cou;d be used in addition to other datasets [106].
- *KoDF*: It is focused on Korean subjects. It is made up of 175,776 deepfake videos and 62,166 pristine videos. It was created using 6 different deepfake manipulation approach and 403 subjects [107].
- *Deepfake MNIST+*: It is mostly focused on facial animation. It consists of 10,000 deepfake and 10,000 original content [108].
- Wavefake: It is focused on audio manipulation. It was created using 6 different approaches and comprises 2 languages [109].
- ForgeryNet: It was created using 15 different approaches and consists of 121,617 deepfake content as well as

TABLE 7. Comparison of existing multi-modal approaches for audio spoofing detection.

Author	Year	Approach	Advantage	Limitation
Chintha *et al.* [94]	2020	Two different streams Where CRNNSpoof and WIRENetSpoof networks are used for audio spoofing detection whereas 4 different variants of XceptionNet for Video detection	Tried to combine video as well as audio manipulation in a single system	In video detection, XceptionNet outperforms their model over the FaceSwap dataset. Audio Performance decreases on the evaluation set of ASVSpoof2019. Hence indicating a lack of generalizability
Agarwal *et al.* [95]	2020	Tried to detect inconsistencies between phoneme and viseme in manipulated content	Higher interpretability	There are certain cases where lips are asymmetrically open, etc. where automatic techniques cannot classify whether the mouth is open or not
Lomnitz *et al.* [96]	2020	Four different modalities were used to classify whether the content was real or fake	Single frame level, multiple frame level, audio level as well as power spectrum level features were used to detect manipulation	Other systems have performed better in deepfake detection competition
Fan *et al.* [98]	2021	Tried to process audio and video streams differently using siamese NN and combining results of networks Video confidence score is calculated	Good interpretability of model, and good	Performance drops on low-quality videos of DF-TIMIT dataset
Chugh *et al.* [97]	2021	Tried to calculate Modality Dissonance Score (MDS) based on the disharmony between audio and video contents	Good interpretability of model, Good performance over DFDC dataset	There are other models well over the DF-TIMIT dataset, hence suggesting generalizability is still an issue

TABLE 8. Comparative study of various deepfake detection models [8].

Method	UADFV	DF-TIMIT	DF-TIMIT	FF-DF	DFD	DFDC	Celeb-DF
Two Stream	85.1	83.5	73.5	70.1	52.8	61.4	53.8
Meso4	84.3	87.8	68.4	84.7	76	75.3	54.8
Face warping artifact	97.4	99.9	93.2	80.1	74.3	72.7	56.9
Head pose	89	55.1	53.2	47.3	56.1	55.9	54.6
Visual artifact	70.2	61.4	62.1	66.4	69.1	61.9	55
XceptionNet	91.2	95.9	94.4	99.7	85.9	72.2	65.3
Multi-Task	65.8	62.2	55.3	76.3	54.1	53.6	54.3
Capsule Network	61.3	78.4	74.4	96.6	64	53.3	57.5

94,630 pristine content. 5400+ subjects were used to create this dataset and 36 perturbations were performed [110].

- DF-W: It was created using only the deepfakes present in the wild i.e. internet. It was created with the fact in mind that real-world content is harder to detect than research datasets. It consists of 1,689 deepfake content [111].

Each dataset has its own type of challenges. Especially the last three datasets namely DFD, DFDC and Celeb-DF, are difficult to achieve accuracy with. Celeb-DF is considered to be one of the most difficult datasets to work with. At the same time, the latest dataset, DF-W was created from deepfake videos in the wild. The deepfake videos/images in the wild i.e. from the internet, poses a greater challenge than any research dataset. For deepfakes in the wild, the tools and quality of content are more sophisticated. In addition to that, the manipulations done on content are also very specific. Thus, achieving accuracy over content from the wild is a very difficult task to achieve. TABLE 9 describe the comparison of existing datasets used for deepfake detection.

VI. OPEN ISSUES AND CHALLENGES

The detection of deepfakes is a complex task. It poses certain challenges in creating an effective deepfake detection system. These challenges are becoming more and more significant with the advancement in deepfake generation techniques. The issues are common for all kinds of modalities, be it audio, video or image content. These challenges have been discussed in detail in this section.

A. GENERALIZABILITY

The first and foremost issue with deepfake detection systems is generalizability. As different deepfake generation tools generate content in different ways, it becomes difficult to develop a model that efficiently detects fake content of all different kinds of deepfakes from different tools. And thus, research is required in this direction. There has been much research done to improve generalizability where capacities of two different approaches are combined to improve generalizability. Or where a feature extracted from one model is provided to another model for classification and provides really good accuracy and improves generalizability [8]. But still, the lack of generalizability is a great challenge that needs to be tackled. Even in audio spoofing detection, there exists much research where certain types of features and models are good for LA attacks but lack accuracy in PA attacks and vice versa [28], [90]. Thus, developing a model which is good for both kinds of attacks is still a challenge. This lack of

TABLE 9. Various datasets available for deepfake detection.

Dataset	Modality	Year	Size of Dataset	Fake/Real	Approaches used	Remarks
UADFV [79]	Video	2019	98	49/49	FakeApp	Has content divided into two classes: Fake and Real
FaceForensics [99]	Video	2018	3012	2008/1004	Face2Face	2 Types of Reenactment Dataset: Source to Video (original Face2Face approach) and Self Reenactment (Apply forgery on input video to create pair of real and synthetic content)
FaceForensics++ [100]	Video	2019	5000	4000/1000	Face2Face, Deepfakes, Faceswap, NeuralTextures	Provides 1000 manipulated content for 4 different approaches each which are applied over 1000 pristine videos
Celeb-DF [78]	Video	2019	1203	795/408	Unknown	Recent Version is available
Celeb-DF v2 [78]	Video	2020	6229	5639/590	Improved Deepfake synthesis algorithm	59 different subjects are used of different age, gender and race
DF-TIMIT [101]	Video	2018	620	620/ -	faceswap-GAN	Created from Vid-TIMIT dataset. Thus, Vid-TIMIT could be used as source of prisitne content
DFD [102]	Video	2019	3431	3068/363	Unknown	Created using 28 actors of different gender, age and race
DFDC Preview [103]	Video	2019	5244	4113/1131	Unknown	It was a Preview Dataset. Could be used as test dataset
DFDC [104]	Video	2020	1,52,690	1,04,500/48,190	faceswap	Large dataset published by Facebook
DeeperForensics 1.0 [105]	Video	2020	60,000	10,000/50,000	Unknown	35 perturbations has been performed. Various actors of different age, gender and race have been considered
WildDeepfake [106]	Image or Video	2020	707	707/ -	Unknown	Small dataset. Could be used in addition with other datasets
KoDF [107]	Video	2021	2,37,942	1,75,776/62,166	FaceSwap, FSGAN, DeepFaceLab, FOMM, ATFHP, Wav2Lip	Created with Korean subjects in mind
Deepfake MNIST + [108]	Video	2021	20,000	10,000/10,000	Unknown	Focused on facial animation
Wavefake [109]	Audio	2021	1,17,985	1,17,985/ -	6 different approaches	Focused on audio manipulation and comprises of 2 languages
ForgeryNet [110]	Image or Video	2021	2,21,247	1,21,617/94,630	15 different Approaches	36 perturbations were performed and created using 5400+ different subjects
DF-W [111]	Video	2021	1,869	1,869/ -	Content directly from the wild i.e. Internet	Provides greater challenge in terms of generalizability over research datasets

generalizability hinders the practical use of these deepfake detection systems.

B. ROBUSTNESS

These deepfake detection models need to be robust in nature. In practical use cases, the manipulations done on the contents will be more challenging than research datasets. In research datasets, manipulations are performed on every frame whereas in the practical world, these manipulations will not be necessarily done on every frame level. Or in real-world use cases, multiple tampering operations or other manipulations might have been performed on the content to remove or minimize the traces of manipulations. The media content on social media platforms will also be in lower quality. And it has been seen that the performance of these models drops significantly over lower-quality content. Thus, the robustness of these models becomes very important for these models to be implemented or used in real-world use cases [31].

C. LACK OF INTERPRETABILITY

For image/video content, the physical/physiological approaches have very high interpretability. It is easier for humans to understand what is actually happening in detecting these deepfake contents. For example, deepfake detection using eye blinking patterns or inconsistency in head poses. But with the advancements in the deepfake generation tools, these approaches have become unreliable and hence, data-driven models have been used. In this approach, we train a model over a large volume of both deepfake as well as original content and the model learns over time to detect these manipulated contents. But in these approaches, it becomes difficult for humans to interpret what is actually behind these models. It is an inherent problem with neural networks as they can't produce human understandable justifications for the predictions they make. For use cases in the forensic domain, human interpretable justifications are necessary for such models. The same problem exists for audio spoofing detection as well. The majority of the approaches

involve neural networks which lack human understandable interpretability. Hence, research in this direction is also necessary [31].

VII. FUTURE RESEARCH DIRECTIONS

For image/video content, as we have seen in the above sections, the detection methods can be categorized into 3 different approaches. These approaches have their own advantages as well as disadvantages. Out of the three, the first approach of utilizing the physical/physiological inconsistencies was the genesis point of detection of deepfake contents. The first approach of utilizing inconsistencies in blinking patterns was proposed around 2018 [38]. And then based on this type of approach, many other forms of deepfake detection methods were proposed. As discussed in paper [79], which focused on using head poses and inconsistencies in face generation with other bodies. In a similar way in [46], where all the inconsistencies, such as eye color, or geometry, which are not detailed, etc., were utilized to classify whether it is fake content or not.

But the disadvantage of the first approach, is that with new and highly efficient deepfakes generation tools, these visible inconsistencies are becomes hard to detect. It almost becomes impossible for the human eye to identify such inconsistencies [8]. And thus, the detection systems with this approach are not as reliable. And as a result, Signal Level features detection methods come into the scene. This approach relies on various feature extraction algorithms such as HOG, ORB, SURF, etc. And then classification algorithms are used to classify whether sample content is fake or original [81].

This approach has been proven to improve accuracy. But over time, deepfakes are getting better and better at generating more realistic deepfakes. Also, deepfakes are improved over time by trying to evade these detection methods. And as a result, deepfakes evolve over time to leave as little traces as possible behind during their synthesis process. Multiple tampering operations are performed and post-processing is done to destroy the traces left behind after manipulation [22]. Thus, we move towards data-driven models, that are trained over large volumes of data. They themselves figure out which features to extract, instead of us focusing on a specific feature. Whereas for audio spoofing detection, traditionally GMM-based models were used. But now more recent research uses NN-based models that are trained over large volumes of audio data sets. Still, there is a lot of room for improvement in these deepfake detection methods. We have discussed a few research directions that we have been able to identify.

A. TRANSFER LEARNING

There have been attempts made to overcome the lack of generalizability of these systems by using the concept of transfer learning. As discussed in [112], transfer learning approaches are used where the Xception model is used with pretrained weights of ImageNet. ImageNet has a very good capability of detecting people, animals, birds, etc. It is trained with over 20 million images and 20,000 sub-classes. So, the basic concept of the approach was that since ImageNet is so good at identifying all these categories, it might be able to extract more relevant features. And hence, improving generalizability [112]. Thus, transfer learning-based approaches do hold a lot of potential in developing a system that has higher generalizability.

B. ENSEMBLE MODELS

Ensemble models combine the results of multiple different models and after aggregating the results of these individual models, it predicts the final output. This type of approach tends to perform better than separate individual models. These ensemble models have also shown their potential in audio spoofing detection as well. A complex system made up of the concept of an ensemble model can greatly improve the generalizability as well as the robustness of the deepfake detection system. This could prove to be a great research direction for both the modalities of video as well as audio. Even a multi-modal approach could be investigated using the concept of ensemble models where each separate model analyzes different modalities of the content.

C. SPECIFICALLY DESIGNED MODELS

In addition to the above-mentioned methods, models designed specifically for the task of deepfake detection could also be investigated. It's example could be capsule networks or MesoInception-4 networks. So the models could be designed with the specific task of manipulation detection as well as by focusing on features related to deepfake manipulations. The primary goal for this approach would be to design a model that could leverage deepfake manipulation-related features.

D. TEMPORAL FEATURES

Temporal features could be leveraged for video or audio content. Instead of analyzing visual artifacts based intra frame features, a temporal feature-based approach could be investigated that could leverage features across different frames i.e. inter-frame features.

VIII. CASE-STUDY: INCOMPATIBILITY BETWEEN MULTIPLE MODES FOR DEEPFAKE DETECTION

For the sake of the case study, we have discussed a multi-modal-based approach for deepfake detection. It is known as incompatibility between multiple modes (IBMM) by Zhang et al. [113]. Since, being a multi-modal approach, the system accepts the input video, processes, it and using the inconsistences and incompatibility among different modes, predicts whether the video is real or fake. The system consists of EfficientNet and 3D-CNN to process three different modes of input data. Experimental results have shown an improvement in the performance of deepfake detection on the DFDC dataset compared to existing deepfake detection systems by 5.21%.

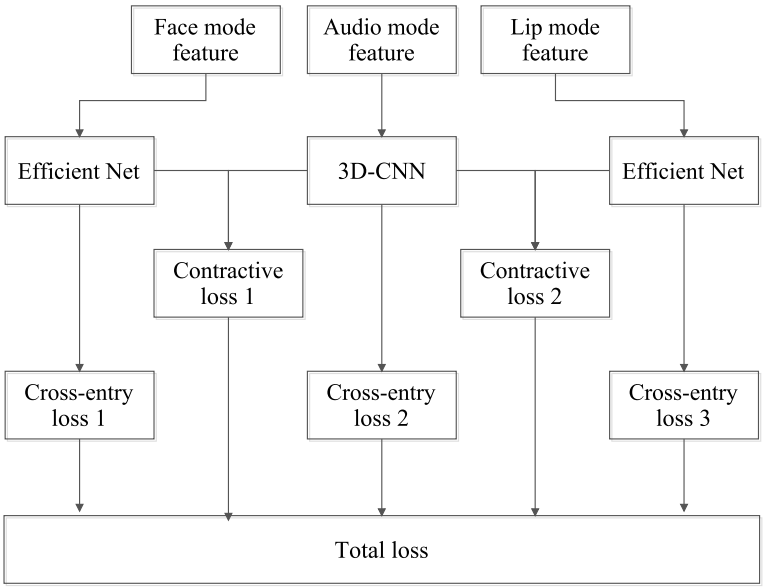

FIGURE 6. Incompatibility between multiple modes (IBMM).

A. MODEL ARCHITECTURE

IBMM-based deepfake detection system processes the input video to detect manipulations at face mode, audio mode and lip mode. For the sake of detecting forgery at face mode and lip mode, the EfficientNet-b3 network is used and corresponding features are fed as input to the network. Whereas to detect manipulations in audio mode, a 3D-CNN network is used. Each of these models are independent and detects forgery at its corresponding mode and level. But the outputs from the middle layers of these networks are also used and compared together to find inconsistencies across different modes such as asynchronized lip and audio, lack of coordination between face and lip movements, etc [112]. The network diagram is shown in FIGURE 6.

1) AUDIO MODE

For the audio mode, the audio content is segmented and for each segment of audio content, MFCC is extracted using the ffmpeg library. Extracted MFCC is then converted into heat maps of size (1 × 13x99) where 1 is the number of channels and 13 × 99 are the height and width of the heatmap. This heatmap is then fed as input to 3D-CNN architecture. For this image data, the output of FC8 layer is received Which is denoted as m_a and it is then passed through a pooling layer and a fully connected layer. Finally, the prediction of the audio mode is obtained and it is denoted as f_a.

2) FACIAL MOTION MODE

For the purpose of facial motion mode, the authors have used a lightweight EfficientNet-b3 because of the memory constraints. The advantages of EfficientNet over ImageNet are that it?s simpler, uses compound model scaling to enlarge CNNs, the number of parameters and computations are greatly reduced. And hence, it also performs better than ImageNet. For the purpose of extracting facial features, S3FD is used. The authors have claimed that due to memory constraints, EfficientNet-b3 has been used but the performance of the model can be greatly increased if a higher version of EfficientNet is used. To obtain the output from the network, coefficients of the network are modified and then output m_f is obtained. This output m_f is then fed to a pooling layer and connecting layer. And the final prediction f_f for facial motion mode is obtained.

3) LIP MOTION MODE

The architecture for lipmotion mode is almost the same as that of facial motion mode. It also consists of EfficientNet-b3. The only thing different is that object detection for this model is the lips instead of the entire face. So, for input greyscale image of lips is given as input and intermediate, as well as final output is obtained i.e., m_m and f_m. Thus, for each mode, intermediate results as well as final predictions are obtained. These intermediate outputs will be helpful to calculate the inconsistencies across different modes. The loss functions for each of the networks are discussed below.

B. LOSS FUNCTIONS

As it can be seen from the figure, there are two types of loss. One is Cross Entropy Loss and Contrastive Loss. And the combination of these two is a Total Loss function. Cross-Entropy is used to detect irrationality and inconsistency within the mode whereas contrastive Loss is used to finding inconsistency across two different modes. Each of these loss functions is discussed in detail.

1) CONTRASTIVE LOSS

Here for notation, l_i denotes the label of ith video and m is the threshold. So first contrastive loss is between lip motion mode and audio mode. Thus the output of these modes is used. The loss function is denoted as Con_1 and the Euclidean distance between lip movement audio is denoted as D_{w_1} [90]. Thus, the formulas will be,

$$D_{w1} = \|f_m - f_a\|_2 \tag{3}$$

$$\text{Con}_1 = \frac{1}{N}\sum_{i=1}^{N} l_i\,(D_{w1})^2 + (1-l_i)\max\,(m - D_{w1}, 0)^2 \tag{4}$$

Similarly, the second contrastive loss will be calculated which represents the inconsistency between face and audio mode. Its loss will be denoted as Con_2 and Euclidean distance will be denoted as D_{w_2} [90].

$$D_{w2} = \|f_f - f_a\|_2 \tag{5}$$

$$\text{Con}_2 = \frac{1}{N}\sum_{i=1}^{N} l_i\,(D_{w2})^2 + (1-l_i)\max\,(m - D_{w2}, 0)^2 \tag{6}$$

2) CROSS-ENTROPY LOSS

These cross-entropy losses are calculated for each mode. For facial motion, lip and audio motion, they are denoted as Cro_1, Cro_2 and Cro_3 respectively [90]. The formulas for these losses are,

$$Cro_1 = -\frac{1}{N}\sum_{i=1}^{N} l_i \log \hat{f}_f^i + (1-l_i)\log\left(1-\hat{f}_f^i\right) \tag{7}$$

$$Cro_2 = -\frac{1}{N}\sum_{i=1}^{N} l_i \log \hat{f}_m^i + (1-l_i)\log\left(1-\hat{f}_m^i\right) \tag{8}$$

$$Cro_3 = -\frac{1}{N}\sum_{i=1}^{N} l_i \log \hat{f}_a^i + (1-l_i)\log\left(1-\hat{f}_a^i\right) \tag{9}$$

3) TOTAL LOSS

It is the addition of all the losses mentioned earlier. Hence, the formula will be,

$$\text{Loss} = \text{Con}_1 + \text{Con}_2 + \text{Cro}_1 + \text{Cro}_2 + \text{Cro}_3 \tag{10}$$

C. DATA USED FOR TRAINING

For the purpose of training and evaluating the IBMM model, the authors have used DF-TIMIT and DFDC datasets. DF-TIMIT dataset has 320 real videos from the VidTIMIT dataset and 320 GAN-generated 320 videos. Moreover, the DF-TIMIT dataset has low-resolution images (64 × 64) LQ as well as high-resolution images (128 × 128) HQ. DF-TIMIT only has manipulation done over the facial region and not on audio [113].

DFDC dataset has 119,146 videos, each having length of 10 seconds. In this dataset, manipulations are done over audio as well as image content. And thus, helps authors to show the true ability of their proposed system. Authors have only used 7,937 videos for training and testing [90].

D. ADJUSTING THE PARAMETERS

For this task, authors have used EfficientNet-b3 and 3D-CNN model for forgery detection. They have these models with a batch size of 2 for 49 epochs. For EfficientNet-b3, the authors have set DROP_PATH_RATE to 0.2 and CROP_CPT to 0.904.

E. MODEL EVALUATION

The authors have used the AUC index to evaluate the performance of the model. AUC index basically represents that the probability of predicting a positive sample is greater than the probability of predicting a negative sample on the test set. Its formula can be seen as:

$$I\left(P_{\text{psample}}, P_{\text{nsample}}\right) = \begin{cases} 1, & P_{\text{psample}} > P_{\text{nsample}} \\ 0.5, & P_{\text{psample}} = P_{\text{nsample}} \\ 0, & P_{\text{psample}} < P_{\text{nsample}} \end{cases} \tag{11}$$

Here $P_{psample}$ represents the probability of a positive sample in a pair of positive negative samples predicting a positive sample result and vice versa for $P_{nsample}$. Whereas $I(P_{psample}, P_{nsample})$ represents the number of positive samples whose probability of prediction is greater than that of negative samples in pair of positive-negative samples. The formula of the AUC index will be,

$$AUC = \frac{\sum_{j=1}^{M\times N} I\left(P_{\text{psample}}^j, P_{\text{nsample}}^j\right)}{M \times N} \tag{12}$$

F. RESULTS

The results of experiments show that the proposed system performs best in on DFDC dataset with an AUC index of 95.87% which is 5.21% higher than existing models. The performance of the system on the DF-TIMIT dataset is 98.1% on LQ images/videos and 97.2% on HQ images/videos.

G. LESSONS LEARNED FROM THE SURVEY

The survey has highlighted the potential of deepfake technology in generative AI, which has emerged as a game-changer in recent AI development. Deepfake has the potential to revolutionize different fields such as entertainment, education, and healthcare. However, the article has also discussed the lingering dangers of the technology, when it comes to deploying in real-life situations. As potential lessons learned from the survey, we discussed both deepfake generation and detection models.

A major drawback of generation models is their over-reliance on a large amount of available data, which is sometimes not possible in applications like healthcare, where data shared publicly is limited. However, the models require high-quality data for training. This means that they may not be suitable for generating deepfakes of individuals who have limited public images or audio recordings available, or for generating deepfakes in languages other than those which are mostly spoken worldwide. Further, these models

require significant computational resources for training, which makes them an expensive affair. Thus, for small organizations, the day-to-day accessibility and usage of these models are limited in scope. For large organizations, scaling the deepfake models is also expensive, and thus optimization in terms of fine-tuning model parameters is a future direction in deepfake generation research.

In the case of deepfake detection models, they highly depend on the specific characteristics of the deepfake being detected. This means that they may not be effective at detecting new types of deepfakes, or deepfakes that have been specifically designed to evade detection. Further, the detection models require further access to unmodified data for high accuracy, which sometimes is restricted owing to privacy and legislative concerns. Thus, the lack of original data in real-world scenarios leads to less accuracy of detection models, and increased cases of misconduct and alterations become a sad reality.

Finally, there are significant ethical and legal considerations surrounding the use of deepfake technology. The potential for deepfakes to be used for malicious purposes such as disinformation, revenge pornography, or financial fraud is a major concern, and there is a need for clear regulations and guidelines on the appropriate use and dissemination of deepfakes. Overall, while deepfake technology holds promise for many applications, it is important to carefully consider the limitations and potential dangers associated with this technology in real-life contexts. Further research is needed to address these issues and ensure that deepfake technology is developed and deployed in a responsible and ethical manner.

IX. CONCLUDING REMARKS

The survey highlighted the potential of deepfake technology, specifically in terms of generation and detection models. Different approaches of deepfake generation for multi-modal scenarios are presented (face swapping, voice conversion, audio synthesis), and also the surrounding challenges and limitations of these models are analyzed. The latest developments are discussed, and it is established that deepfake models are highly required with a high degree of generalizability. Going forward, deepfake would remain an active domain of research in the far future as well. The deepfake generation and detection models complement each other, with advanced generation strategies, the detection models would evolve, which would further augment the features of generation models. Thus, it forms an evolutionary cycle, where one benefits from the other. There are both positive as well as negative impacts of deepfakes on society. Although developing a deepfake detection system with good generalizability as well as high robustness is in itself a herculean task. Various approaches have already been investigated and developed but the results are still far from perfect. Other than traditional approaches, multi-modal systems are also pretty impressive. We also have provided an exhaustive list of available research datasets that could be used for future research.

Going forward, there are several areas for future research in this field. An important direction is to design more robust, highly scalable, and lightweight deepfake models. Another direction is to include validity and explainability to the deepfake models, which could address the ethical and social issues of deepfake usage. There is a need for more research on how to raise awareness and educate the public about the risks and dangers of deepfakes, and how to prevent their spread in the digital world. In totality, deepfake technology brings both exciting opportunities and significant challenges for society, and thus it is important as how researchers, policymakers, and organizations can drive and harmonize together to address common issues surrounding deepfakes.

REFERENCES

[1] T. Zhang, "Deepfake generation and detection, a survey," *Multimedia Tools Appl.*, vol. 81, no. 5, pp. 6259–6276, Feb. 2022.

[2] A. O. J. Kwok and S. G. M. Koh, "Deepfake: A social construction of technology perspective," *Current Issues Tourism*, vol. 24, no. 13, pp. 1798–1802, Jul. 2021.

[3] S. Awah Buo, "The emerging threats of deepfake attacks and countermeasures," 2020, *arXiv:2012.07989*.

[4] J. Ice, "Defamatory political deepfakes and the first amendment," *Case W. Res. L. Rev.*, vol. 70, p. 417, 2019.

[5] *Uncanny Deepfake Tom Cruise Fools People on Tiktok*. Accessed: Mar. 13, 2023. [Online]. Available: https://futurism.com/the-byte/deepfake-tom-cruise-tiktok

[6] V. Karasavva and A. Noorbhai, "The real threat of deepfake pornography: A review of Canadian policy," *Cyberpsychol., Behav., Social Netw.*, vol. 24, no. 3, pp. 203–209, Mar. 2021.

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

[8] K. N. Ramadhani and R. Munir, "A comparative study of deepfake video detection method," in *Proc. 3rd Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Nov. 2020, pp. 394–399.

[9] M. Safiullah and N. Parveen, *Big Data, Artificial Intelligence and Machine Learning: A Paradigm Shift in Election Campaigns*. Hoboken, NJ, USA: Wiley, 2022, ch. 11, pp. 247–261.

[10] Y. Patel, S. Tanwar, P. Bhattacharya, R. Gupta, T. Alsuwian, I. E. Davidson, and T. F. Mazibuko, "An improved dense CNN architecture for deepfake image detection," *IEEE Access*, vol. 11, pp. 22081–22095, 2023.

[11] R. Singh, S. Shrivastava, A. Jatain, and S. B. Bajaj, "Deepfake images, videos generation, and detection techniques using deep learning," in *Machine Intelligence and Smart Systems*, S. Agrawal, K. K. Gupta, J. H. Chan, J. Agrawal, and M. Gupta, Eds. Singapore: Springer, 2022, pp. 501–514.

[12] K. Patel, C. Mistry, D. Mehta, U. Thakker, S. Tanwar, R. Gupta, and N. Kumar, "A survey on artificial intelligence techniques for chronic diseases: Open issues and challenges," *Artif. Intell. Rev.*, vol. 55, pp. 3747–3800, Jun. 2022.

[13] C. Campbell, K. Plangger, S. Sands, J. Kietzmann, and K. Bates, "How deepfakes and artificial intelligence could reshape the advertising industry," *J. Advertising Res.*, vol. 62, no. 3, pp. 241–251, Sep. 2022.

[14] T. Dudeja, S. K. Dubey, and A. K. Bhatt, "Ensembled EfficientNetB3 architecture for multi-class classification of tumours in MRI images," *Intell. Decis. Technol.*, vol. 17, no. 2, pp. 395–414, May 2023.

[15] P. Bhattacharya, D. Saraswat, D. Savaliya, S. Sanghavi, A. Verma, V. Sakariya, S. Tanwar, R. Sharma, M. S. Raboaca, and D. L. Manea, "Towards future internet: The metaverse perspective for diverse industrial applications," *Mathematics*, vol. 11, no. 4, p. 941, Feb. 2023.

[16] C. C. Ki Chan, V. Kumar, S. Delaney, and M. Gochoo, "Combating deepfakes: Multi-LSTM and blockchain as proof of authenticity for digital media," in *Proc. IEEE/ITU Int. Conf. Artif. Intell. for Good (AI4G)*, Sep. 2020, pp. 55–62.

[17] S. Ji, J. Luo, and X. Yang, "A comprehensive survey on deep music generation: Multi-level representations, algorithms, evaluations, and future directions," 2020, *arXiv:2011.06801*.

[18] (1999). *Number of Expert-Crafted Video Deepfakes Double Every Six Months*. Accessed: Mar. 20, 2023. [Online]. Available: https://cybernews.com/privacy/report-number-of-expert-crafted-video-deepfakes-double-every-six-months/
[19] S. Karthika and M. Durgadevi, "Generative adversarial network (GAN): A general review on different variants of GAN and applications," in *Proc. 6th Int. Conf. Commun. Electron. Syst. (ICCES)*, Coimbatre, India, Jul. 2021, pp. 1–8.
[20] D. Yadav and S. Salmani, "Deepfake: A survey on facial forgery technique using generative adversarial network," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICCS)*, May 2019, pp. 852–857.
[21] R. Katarya and A. Lal, "A study on combating emerging threat of deepfake weaponization," in *Proc. 4th Int. Conf. I-SMAC, IoT Social, Mobile, Anal. Cloud*, Oct. 2020, pp. 485–490.
[22] R. Thakur and R. Rohilla, "Recent advances in digital image manipulation detection techniques: A brief review," *Forensic Sci. Int.*, vol. 312, Jul. 2020, Art. no. 110311.
[23] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, "Deep fake image detection based on pairwise learning," *Appl. Sci.*, vol. 10, no. 1, p. 370, Jan. 2020.
[24] B. F. Nasar and E. R. Lason, "Deepfake detection in media files–audios, images and videos," in *Proc. IEEE Recent Adv. Intell. Comput. Syst. (RAICS)*, Dec. 2020, pp. 74–79.
[25] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 937–941, 2021.
[26] H. Ling, L. Huang, J. Huang, B. Zhang, and P. Li, "Attention-based convolutional neural network for ASV spoofing detection," in *Proc. Interspeech*, Aug. 2021, pp. 4289–4293.
[27] M. Dua, C. Jain, and S. Kumar, "LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems," *J. Ambient Intell. Humanized Comput.*, vol. 13, no. 4, pp. 1985–2000, Apr. 2022.
[28] B. Chettri, D. Stoller, V. Morfi, M. A. M. Ramírez, E. Benetos, and B. L. Sturm, "Ensemble models for spoofing detection in automatic speaker verification," in *Proc. Interspeech*, Sep. 2019, pp. 1018–1022.
[29] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *Proc. IEEE 10th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2019, pp. 1–8.
[30] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020.
[31] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection," *IET Biometrics*, vol. 10, no. 6, pp. 607–624, Nov. 2021.
[32] P. Swathi, "DeepFake creation and detection: A survey," in *Proc. 3rd Int. Conf. Inventive Res. Comput. Appl. (ICIRCA)*, Sep. 2021, pp. 584–588.
[33] A. Abdulreda and A. Obaid, "A landscape view of deepfake techniques and detection methods," *Int. J. Nonlinear Anal. Appl.*, vol. 13, no. 1, pp. 745–755, 2022.
[34] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Int. J. Speech Technol.*, vol. 53, no. 4, pp. 3974–4026, Feb. 2023.
[35] K. Patil, S. Kale, J. Dhokey, and A. Gulhane, "Deepfake detection using biological features: A survey," 2023, *arXiv:2301.05819*.
[36] S. Dhesi, L. Fontes, P. Machado, I. K. Ihianle, F. F. Tash, and D. A. Adama, "Mitigating adversarial attacks in deepfake detection: An exploration of perturbation and AI techniques," 2023, *arXiv:2302.11704*.
[37] Y. Mirsky and W. Lee, "The creation and detection of deepfakes," *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1–41, Jan. 2022.
[38] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.
[39] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.
[40] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.
[41] M. Zendran and A. Rusiecki, "Swapping face images with generative neural networks for deepfake technology—Experimental study," *Proc. Comput. Sci.*, vol. 192, pp. 834–843, Jan. 2021.
[42] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, "Deepfakes: Trick or treat?" *Bus. Horizons*, vol. 63, no. 2, pp. 135–146, Mar. 2020.
[43] M. Khichi and R. Kumar Yadav, "A threat of deepfakes as a weapon on digital platform and their detection methods," in *Proc. 12th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Jul. 2021, pp. 01–08.
[44] L. Guarnera, O. Giudice, and S. Battiato, "Fighting deepfake by exposing the convolutional traces on images," *IEEE Access*, vol. 8, pp. 165085–165098, 2020.
[45] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.
[46] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2019, pp. 83–92.
[47] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2387–2395.
[48] *Faceswap*. Accessed: Jan. 4, 2022. [Online]. Available: https://github.com/deepfakes/faceswap
[49] *Faceswap-Gan*. Jan. 14, 2022. [Online]. Available: https://github.com/shaoanlu/faceswap-GAN
[50] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–12, Jul. 2019.
[51] *Deepfacelab*. Accessed: Jan. 14, 2022. [Online]. Available: https://github.com/iperov/DeepFaceLab
[52] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7183–7192.
[53] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, "STGAN: A unified selective transfer network for arbitrary image attribute editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3668–3677.
[54] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1–13.
[55] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2018, *arXiv:1812.04948*.
[56] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5464–5478, Nov. 2019.
[57] *Dfaker*. Accessed: Jan. 14, 2022. [Online]. Available: https://github.com/dfaker/df
[58] A. Vahdat and J. Kautz, "NVAE: A deep hierarchical variational autoencoder," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19667–19679.
[59] B. I. Ibrahim, D. C. Nicolae, A. Khan, S. I. Ali, and A. Khattak, "VAE-GAN based zero-shot outlier detection," in *Proc. 4th Int. Symp. Comput. Sci. Intell. Control*, NY, NY, USA, Nov. 2020.
[60] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.
[61] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," in *Proc. Workshoptrack-ICLR*. Montreal, QC, Canada: Montreal Inst. Learn. Algorithms (MILA), 2017.
[62] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*.
[63] A. Kolesnikov and C. H. Lampert, "PixelCNN models with auxiliary variables for natural image modeling," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, D. Precup and Y. W. Teh, Eds., Aug. 2017, pp. 1905–1914.
[64] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 3617–3621.
[65] S. Vasquez and M. Lewis, "MelNet: A generative model for audio in the frequency domain," 2019, *arXiv:1906.01083*.
[66] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," 2017, *arXiv:1710.07654*.

[67] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020, pp. 17022–17033.

[68] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 1–11.

[69] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, D. Precup and Y. W. Teh, Eds., Aug. 2017, pp. 1857–1865.

[70] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Athens, Greece, Dec. 2018, pp. 266–273.

[71] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 2100–2104.

[72] T. R. Shaham, T. Dekel, and T. Michaeli, "SinGAN: Learning a generative model from a single natural image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4569–4579.

[73] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4217–4228, Dec. 2021.

[74] D. Rohera, H. Shethna, K. Patel, U. Thakker, S. Tanwar, R. Gupta, W.-C. Hong, and R. Sharma, "A taxonomy of fake news classification techniques: Survey and implementation aspects," *IEEE Access*, vol. 10, pp. 30367–30394, 2022.

[75] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech*, Aug. 2017, pp. 2–6.

[76] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. Aik Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," 2019, *arXiv:1904.05441*.

[77] A. Gupta, N. Kumar, P. Prabhat, R. Gupta, S. Tanwar, G. Sharma, P. N. Bokoro, and R. Sharma, "Combating fake news: Stakeholder interventions and potential solutions," *IEEE Access*, vol. 10, pp. 78268–78289, 2022.

[78] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3204–3213.

[79] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 8261–8265.

[80] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1831–1839.

[81] F. F. Kharbat, T. Elamsy, A. Mahmoud, and R. Abdullah, "Image feature detectors for deepfake video detection," in *Proc. IEEE/ACS 16th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2019, pp. 1–4.

[82] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of GAN-generated fake images over social networks," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Apr. 2018, pp. 384–389.

[83] C.-C. Hsu, C.-Y. Lee, and Y.-X. Zhuang, "Learning to detect fake face images in the wild," in *Proc. Int. Symp. Comput., Consum. Control*, Dec. 2018, pp. 388–391.

[84] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1205–1207.

[85] Z. Akhtar and D. Dasgupta, "A comparative evaluation of local feature descriptors for DeepFakes detection," in *Proc. IEEE Int. Symp. Technol. Homeland Secur. (HST)*, Nov. 2019, pp. 1–5.

[86] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2307–2311.

[87] J. K. Lewis, I. E. Toubal, H. Chen, V. Sandesera, M. Lomnitz, Z. Hampel-Arias, C. Prasad, and K. Palaniappan, "Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning," in *Proc. IEEE Appl. Imag. Pattern Recognit. Workshop (AIPR)*, Oct. 2020, pp. 1–9.

[88] H. Yu, Z.-H. Tan, Z. Ma, R. Martin, and J. Guo, "Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4633–4644, Oct. 2018.

[89] R. K. Das, J. Yang, and H. Li, "Long range acoustic and deep features perspective on ASVspoof 2019," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 1018–1025.

[90] M. Alzantot, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," in *Proc. Interspeech*, Sep. 2019, pp. 1078–1082.

[91] J. Monteiro, J. Alam, and T. H. Falk, "Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers," *Comput. Speech Lang.*, vol. 63, Sep. 2020, Art. no. 101096.

[92] B. T. Balamurali, K. E. Lin, S. Lui, J.-M. Chen, and D. Herremans, "Toward robust audio spoofing detection: A detailed comparison of traditional and learned features," *IEEE Access*, vol. 7, pp. 84229–84241, 2019.

[93] R. L. M. A. P. C. Wijethunga, D. M. K. Matheesha, A. A. Noman, K. H. V. T. A. De Silva, M. Tissera, and L. Rupasinghe, "Deepfake audio detection: A deep learning based solution for group conversations," in *Proc. 2nd Int. Conf. Advancement Comput. (ICAC)*, vol. 1, Dec. 2020, pp. 192–197.

[94] A. Chintha, B. Thai, S. J. Sohrawardi, K. Bhatt, A. Hickerson, M. Wright, and R. Ptucha, "Recurrent convolutional structures for audio spoof and video deepfake detection," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 1024–1037, Aug. 2020.

[95] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting deep-fake videos from phoneme-viseme mismatches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 2814–2822.

[96] M. Lomnitz, Z. Hampel-Arias, V. Sandesara, and S. Hu, "Multimodal approach for DeepFake detection," in *Proc. IEEE Appl. Imag. Pattern Recognit. Workshop (AIPR)*, Oct. 2020, pp. 1–9.

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

[98] Z. Fan, J. Zhan, and W. Jiang, "Detecting deepfake videos by visual-audio synchronism: Work-in-progress," in *Proc. Int. Conf. Embedded Softw.*, Sep. 2021, pp. 31–32.

[99] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: A large-scale video dataset for forgery detection in human faces," 2018, *arXiv:1803.09179*.

[100] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

[101] P. Korshunov and S. Marcel, "DeepFakes: A new threat to face recognition? Assessment and detection," 2018, *arXiv:1812.08685*.

[102] (1999). *Deep Fake Detection Dataset*. Accessed: Jan. 14, 2022. [Online]. Available: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

[103] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset," 2019, *arXiv:1910.08854*.

[104] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The DeepFake detection challenge (DFDC) dataset," 2020, *arXiv:2006.07397*.

[105] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2886–2895.

[106] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "WildDeepfake: A challenging real-world dataset for deepfake detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2382–2390.

[107] P. Kwon, J. You, G. Nam, S. Park, and G. Chae, "KoDF: A large-scale Korean DeepFake detection dataset," 2021, *arXiv:2103.10094*.

[108] J. Huang, X. Wang, B. Du, P. Du, and C. Xu, "DeepFake MNIST+: A DeepFake facial animation dataset," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1973–1982.
[109] J. Frank and L. Schönherr, "WaveFake: A data set to facilitate audio deepfake detection," 2021, *arXiv:2111.02813*.
[110] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, and Z. Liu, "ForgeryNet: A versatile benchmark for comprehensive forgery analysis," 2021, *arXiv:2103.05630*.
[111] J. Pu, N. Mangaokar, L. Kelly, P. Bhattacharya, K. Sundaram, M. Javed, B. Wang, and B. Viswanath, "Deepfake videos in the wild: Analysis and detection," 2021, *arXiv:2103.04263*.
[112] P. Ranjan, S. Patil, and F. Kazi, "Improved generalizability of deep-fakes detection using transfer learning based CNN framework," in *Proc. 3rd Int. Conf. Inf. Comput. Technol. (ICICT)*, Mar. 2020, pp. 86–90.
[113] Y. Zhang, J. Zhan, W. Jiang, and Z. Fan, "Deepfake detection based on incompatibility between multiple modes," in *Proc. Int. Conf. Intell. Technol. Embedded Syst. (ICITES)*, Chengdu, China, Oct. 2021, pp. 1–7.

YOGESH PATEL received the M.Tech. degree in computer science and engineering from the Institute of Technology, Nirma University, Ahmedabad, India. His research interests include blockchain, federated learning, machine learning, and deep learning.

SUDEEP TANWAR (Senior Member, IEEE) received the B.Tech. degree from Kurukshetra University, India, in 2002, the M.Tech. degree (Hons.) from Guru Gobind Singh Indraprastha University, Delhi, India, in 2009, and the Ph.D. degree with specialization in wireless sensor network, in 2016. He is currently a Professor with the Computer Science and Engineering Department, Institute of Technology, Nirma University, India. He is also a Visiting Professor with Jan Wyzykowski University, Polkowice, Poland, and the University of Pitesti, Pitesti, Romania. He has authored two books and edited 13 books, more than 250 technical articles, including top journals and top conferences, such as IEEE Transactions on Network Science and Engineering, IEEE Transactions on Vehicular Technology, IEEE Transactions on Industrial Informatics, IEEE Wireless Communications, *IEEE Networks*, ICC, GLOBECOM, and INFOCOM. He initiated the research field of blockchain technology adoption in various verticals, in 2017. His H-index is 71. He actively serves his research communities in various roles. His research interests include blockchain technology, wireless sensor networks, fog computing, smart grid, and the IoT. He is a Final Voting Member of the IEEE ComSoc Tactile Internet Committee, in 2020. He is a member of CSI, IAENG, ISTE, and CSTA, and a member of the Technical Committee on Tactile Internet of IEEE Communication Society. He has been awarded the Best Research Paper Awards from IEEE IWCMC-2021, IEEE GLOBECOM 2018, IEEE ICC 2019, and Springer ICRIC-2019. He has served many international conferences as a member of the Organizing Committee, such as the Publication Chair for FTNCT-2020, ICCIC 2020, and WiMob2019, a member of the Advisory Board for ICACCT-2021 and ICACI 2020, the Workshop Co-Chair for CIS 2021, and the General Chair for IC4S 2019, 2020, and ICCSDF 2020. He is also serving the editorial boards of *Computer Communications*, *International Journal of Communication System*, and *Security and Privacy*. He is also leading the ST Research Laboratory, where group members are working on the latest cutting-edge technologies.

RAJESH GUPTA (Member, IEEE) received the Bachelor of Engineering degree from the University of Jammu, India, in 2008, the master's degree in technology from Shri Mata Vaishno Devi University, Jammu, India, in 2013, and the Ph.D. degree in computer science and engineering from Nirma University, Ahmedabad, Gujarat, India, in 2023, under the supervision of Dr. Sudeep Tanwar. He is currently an Assistant Professor with Nirma University. He has authored/coauthored some publications (including papers in SCI indexed journals and IEEE ComSoc sponsored international conferences). Some of his research findings are published in top-cited journals and conferences, such as IEEE Transactions on Industrial Informatics, IEEE Transactions on Network and Service Management, IEEE Transactions on Network Science and Engineering, IEEE Transactions on Green Communications and Networking, IEEE Transactions on Computational Social Systems, IEEE Internet of Things Journal, *IEEE Network* magazine, *IEEE Internet of Things Magazine*, *Computer Communications*, *Computer and Electrical Engineering*, *International Journal of Communication Systems* (Wiley), *Transactions on Emerging Telecommunications Technologies* (Wiley), *Physical Communication* (Elsevier), IEEE ICC, IEEE INFOCOM, IEEE GLOBECOM, IEEE CITS, and many more. His research interests include device-to-device communication, network security, blockchain technology, 5G communication networks, and machine learning. His H-index is 31 and i10-index is 67. He is also a recipient of Doctoral Scholarship from the Ministry of Electronics and Information Technology, Government of India, under the Visvesvaraya Ph.D. Scheme. He is a recipient of Student Travel Grant from WICE-IEEE to attend IEEE ICC 2021, Canada. He has been awarded the Best Research Paper Awards from IEEE ECAI 2021, IEEE ICCCA 2021, IEEE IWCMC 2021, and IEEE SCIoT 2022. His name has been included in the list of Top 2% scientists worldwide published by the Stanford University, USA, consecutively in 2021, 2022, and 2023. He has participated the fully-funded the most prestigious ACM's Heidelberg Laureate Forum 2023 held in Heidelberg University, Germany. He was felicitated by Nirma University for their research achievements bagged, in 2019, 2020, 2021, and 2022. He is also an Active Member of the ST Research Laboratory (www.sudeeptanwar.in).

PRONAYA BHATTACHARYA (Member, IEEE) received the Ph.D. degree from Dr. A. P. J Abdul Kalam Technical University, Lucknow, Uttar Pradesh, India. He is currently an Associate Professor with the Computer Science and Engineering Department, Amity School of Engineering and Technology, Amity University, Kolkata, India. He has over ten years of teaching experience. He has authored or coauthored more than 100 research papers in leading SCI journals and top core IEEE COMSOC A* conferences. Some of his top-notch findings are published in reputed SCI journals, such as IEEE Journal of Biomedical and Health Informatics, IEEE Transactions on Vehicular Technology, IEEE Internet of Things Journal, IEEE Transactions on Network Science and Engineering, IEEE Transactions on Computational Social Systems, IEEE Transactions of Network and Service Management, IEEE Access, IEEE Sensors Journal, *IEEE Internet of Things Magazine*, *IEEE Communication Standards Magazine*, *ETT* (Wiley), *Expert Systems* (Wiley), *CCPE* (Wiley), *FGCS* (Elsevier), *OQEL* (Springer), *WPC* (Springer), ACM-MOBICOM, IEEE-INFOCOM, IEEE-ICC, IEEE-CITS, IEEE-ICIEM, IEEE-CCCI, and IEEE-ECAI. He has an H-index of 19 and an i10-index of 32. His research interests include healthcare analytics, optical switching and networking, federated learning, blockchain, and the IoT. He has been appointed at the capacity of a keynote speaker, a technical committee member, and the session chair across the globe. He was awarded Eight Best Paper Awards in Springer ICRIC-2019, IEEE-ICIEM-2021, IEEE-ECAI-2021, Springer COMS2-2021, and IEEE-ICIEM-2022. He is a Reviewer of 21 reputed SCI

journals, such as IEEE Internet of Things Journal, IEEE Transactions on Industrial Informatics, IEEE Transactions of Vehicular Technology, IEEE Journal of Biomedical and Health Informatics, IEEE Access, *IEEE Network* magazine, *ETT* (Wiley), *IJCS* (Wiley), *MTAP* (Springer), *OSN* (Elsevier), *WPC* (Springer), and others. He is also an Active Member of the ST Research Laboratory (www.sudeeptanwar.in).

INNOCENT EWEAN DAVIDSON (Senior Member, IEEE) received the B.Sc. (Eng.) (Hons.) and M.Sc. (Eng.) degrees in electrical engineering from the University of Ilorin, in 1984 and 1987, respectively, the Ph.D. degree in electrical engineering from the University of Cape Town, in 1998; and the master's Diploma degree in business management from the University of KwaZulu-Natal, in 2004. He also received Associate Certificate in Sustainable Energy Management (SEMAC) from the British Columbia Institute of Technology, Burnaby, BC, Canada, in 2011, and the Course Certificate in Artificial Intelligence from the University of California at Berkeley, USA, in 2020. He was a Full Professor and the Chair of the Department of Electrical Power Engineering; a Research Leader of the Smart Grid Research Centre, and the Program Manager of the DUT-DSI Space Science and CNS Research Program, Durban University of Technology (DUT), South Africa, from 2016 to 2022. Currently, he is a Full Professor and the Director of the African Space Innovation Center (ASIC), and French South African Institute of Technology (F'SATI), Cape Peninsula University of Technology (CPUT), Bellville, South Africa. He has supervised five postdoctoral research fellows, and graduated 56 Ph.D./master's students and over 1200 engineers, technologists, and technicians. He is the author/coauthor of over 370 technical papers in accredited journals, and peer-reviewed conference proceedings and book chapters. He has managed over U.S.$3 million in research funds. His current research interests include space science and technology innovation, applied artificial intelligence, and the grid integration of renewable energy. He is a recipient of numerous international best paper awards, and from DUT's Annual Research and Innovation. He is a C2-Rated Researcher from the National Research Foundation (NRF), South Africa. He is a fellow of the Institute of Engineering and Technology, U.K., and the South African Institute of Electrical Engineers; a Chartered Engineer in the U.K.; and a registered Professional Engineer (P Eng.) of the Engineering Council of South Africa. He is a member of the Western Canada Group of Chartered Engineers (WCGCE), the Institute of Engineering and Technology (IET Canada) British Columbia Chapter, and IEEE Collabratec Communities on Smart Cities and IEEE (South Africa Chapter). He was the General Chair of the 30th IEEE Southern Africa Universities Power Engineering Conference, in 2022. He is the Host and a Convener of the DSI-DUT-SANSA-ATNS Space Science and CNS Symposium, in 2021 and 2022; and a Guest Speaker in several forums, including the Science Forum of South Africa and the International Conference on Sustainable Development.

ROYI NYAMEKO received the National Diploma, and the B.Tech. and M.Tech. degrees in electrical and electronic engineering from the Cape Peninsula University of Technology (CPUT), in 2004, 2009, and 2013, respectively. He worked extensively with radio frequency and microwave circuits for more than 20 years. In the last ten years, he has been involved developing communication subsystems for CubeSats, missiles, and UAVs. From 2015 to 2017, he was with the African VLBI Network (AVN) Team, SKA SA responsible for the conversion of an old Ghana Satellite Station into a VLBI observation station, and radio astronomy telescope, and designed the DC power distribution network, including control and monitoring circuits for the VLBI receiver. He has designed and developed VHF, UHF and S-band transceiver's RF front-ends subsystems LNAs, down-convertor and up-convertors, and local oscillators. Since 2017, he has been with the African Space and Innovation Centre, Department of Electrical, Electronic and Computer Engineering, CPUT, as a Chief Engineer. He has been involved with the development, launch and commissioning of CubeSat missions, namely, ZaCube-2, in December 2018, and MDASat-1 constellation (1a, 1b and 1c), in January 2022. He is responsible for the concept design, development, and maintenance of the current ground station operating at UHF/VHF and S-band frequencies. His current research interest includes the hardware implementation of distributed multi-function RF sensors for nanosatellite applications.

SRINIVAS ALUVALA is currently the Director of the Center for Software Development and Digital Learning, and an Assistant Professor with the School of CS & AI, SR University, India. He is passionate to take up challenging tasks with targeted goals. Enjoy working with people and like to explore different fields of computer science and engineering and to develop practical applications on real-life projects. His professional experience includes technical training, delivery of academics, administration, and research guidance. His work experience include C, Java, JavaScript, HTML, PHP, Python, and SCALA, and also include IBM Worklight, Manual and Automation Testing (Selenium), and Hadoop. He is having expertise in guiding and preparation of NBA and NAAC accreditation documentation and NIRF Application process.

VRINCE VIMAL (Member, IEEE) received the B.E. degree in electronics and communication engineering from GCOE, North Maharashtra University, Maharashtra, India, in 2001, the M.Tech. degree in electronics and communication engineering from JRN Rajasthan, India, in 2006, and the Ph.D. degree in communication systems from the Department of Electronics and Communication Engineering, Indian Institute of Technology Roorkee, Roorkee, India, in 2019. He is currently an Associate Professor with the Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun, India, U.K. He has around 16 years of teaching, training, and research experience in India and Abroad. He has authored or coauthored several research papers in international peer-reviewed journals and conferences. He has six patents to his name. His research interests include cellular networks, 5G, the IoT, mobile ad-hoc networks, routing protocols, and soft computing. He is a Life Member of IAENG and NSBE.