FOUNTAIN JOURNAL OF NATURAL & APPLIED SCIENCES

A Publication of the College of Natural & Applied Sciences

Fountain University, Osogbo, Nigeria

A comprehensive review of deepfakes in digital media forensics

Azeez, N. A*, Aaron, D. S., Akinbooro, S.A, Isiekwene, C. C.

Department of Cybersecurity and Software Engineering, Faculty of Computing and Informatics, University of Lagos, Nigeria.

Corresponding author: nazeez@unilag.edu.ng

ABSTRACT

Deepfake technology has grown fast in recent years. It started as a simple experiment but is now widely used to create very realistic fake images, sounds, and videos. This development emerged from advanced Artificial Intelligence (AI) methods, including generative adversarial networks (GANs), autoencoders, and diffusion models. This paper examines how deepfakes are made, the tools behind them, and different ways people use them, both for good and bad. Ninety-five papers in all were gathered from various sources. After carefully reviewing each publication, it was found that some shared similar objectives and methods. As a result, the most pertinent papers were chosen for adoption and review. This article examines the challenges of detecting deepfakes using AI tools, including CNNs, time-based models, attention mechanisms, and multimodal models. The paper points out real problems with these detection systems, such as biased data, difficulty handling new kinds of deepfakes, and attacks that try to fool the detectors. In the end, it emphasises that fixing the deepfake problem is not just a technological issue. It also needs laws, better rules, and more public understanding.

ARTICLE INFO

Article history:
Received January 2026
Revised March 2026
Accepted April 2026

Keywords:
Cybersecurity, Vulnerability, AI-based Detection, Multimodal Forensics. Synthetic Media, Artificial Intelligence

Introduction

The term "deepfakes," a combination of "deep learning" and "fake," refers to a class of artificial intelligence (AI) and machine learning (ML) algorithms used to produce remarkably realistic-looking fake audio, video, or image content. The rapid development of deepfake technology has sparked serious concerns across a number of domains, including politics, media, ethics, and law [1].

When a Reddit user produced a deepfake movie featuring celebrities' faces placed on the bodies of adult film actresses in 2017, the idea of deepfakes was born. Since then, technology has advanced rapidly, leading to the development of complex tools and algorithms that make it easier to produce highly realistic deepfakes. With potential repercussions including identity theft and impersonation, misinformation and disinformation, and national security, deepfakes represent a serious danger to media legitimacy and trust.

At first, people mostly used it for fun or research. But now, because computers and AI tools have gotten stronger, it has become much easier for anyone to make fake media that looks very real. This has caused a lot of concern. Deepfakes can trick people because they seem real and can even fool normal checks or security systems. There have been cases in which scammers used fake voices to impersonate company bosses and steal large sums of money [1]. Deepfakes can also be used in harmful ways. For example, they can spread false political news, create explicit videos without permission, or make people doubt what they see online [2].

As deepfake technology gets better, it becomes harder to spot and stop fake media. Older methods for detecting fake content don't work as well against new AI-generated videos and images. Although there have been efforts to develop AI-based detection systems, most are limited by dataset biases, hallucination, generalizability issues, or a lack of cross-modal robustness [3]. Furthermore, the legal and ethical

DOI: https://doi.org/10.53704/fujnas.v15i1.1135

frameworks for governing deepfakes remain underdeveloped in many jurisdictions, compounding the risks they pose.

This review work seeks to:

i. Clarify the concept and evolution of deepfakes across image, audio, and video domains.
ii. Examine the underlying technologies and models used in deepfake generation.
iii. Explore the legitimate and malicious applications of deepfakes in real-world scenarios.
iv. Analyse AI-based detection techniques and their performance.
v. Discuss challenges, limitations, and potential directions for future research and policy.

In line with these objectives, the following research questions guide this review:

i. How have deepfake technologies evolved across image, audio, and video domains between 2020 and 2025?
ii. What are the technical foundations and tools driving deepfake generation?
iii. In what ways are deepfakes being applied for both constructive and malicious purposes?
iv. What AI-based techniques are currently being used for detecting deepfakes, and how reliable are they?
v. What challenges remain in the detection and forensic analysis of deepfakes, and what future directions can be taken to address them?

Methodology

Undoubtedly, this is a literature survey-based article. The methodology adopted is a comprehensive content analysis of relevant and related papers.

Initially, a total of ninety-five articles were obtained from different sources. After a thorough review of all the papers, it was observed that there was some overlap in content, with the aims and methodologies sharing common features. Consequently, a decision was made to adopt the most relevant papers. The statistics for the papers, as obtained from various sources, are given in Table 1. The 76 publications selected during the search period are shown in Table 1.

Table 1: Distribution of the study's publications based on databases checked following screening

S/N	Database	URL	Count	% Count
1	IEEE Xplorer	https://ieeexplore.ieee.org	12	15.78%
2	Springer	https://link.springer.com	7	9.21%
3	Elsevier (ScienceDirect)	https://sciencedirect.com	10	13.15%
4	MDPI	https://mdpi.com	8	10.52%
5	SSRN-JETIR	https://www.ssrn.com	6	7.89%
6	Taylor&Francis	https://www.taylorandfrancis.com	4	5.26%
7	Others		29	38.16%

Strategy for searching and the keywords used

Seven scientific databases—MDPI, IEEE, ScienceDirect (Elsevier), Springer, Taylor & Francis, SSRN-JETIR, and Wiley Online Library — were searched. To ensure this review was more thorough, comprehensive, and detailed, an effort was made to scan the reference lists of the publications. Only English-language articles released between the year 2015 and 2025 were included in the search [4]. The prominent keywords used in searching are Deepfakes, Digital Media, Forensics, AI-based Detection, Multimodal Forensics, Synthetic Media and deep learning for Deepfakes.

Concepts and evolution of deepfake technology

The evolution of deepfake technology reflects the convergence of multiple innovations in machine learning, digital media processing, and data availability. In the past decade, the ability to generate synthetic content that closely mimics human behaviour and appearance has progressed from rudimentary image-blending techniques to sophisticated multimodal systems capable of producing hyper-realistic faces, voices, and gestures [5].

At the heart of deepfakes is a concept known as generative modelling. This kind of technology works by training AI systems on large amounts of real examples, so they learn what things usually look or sound like. As pointed out in Wood et al., the aim isn't just to copy something that already exists; it's to make new stuff that seems so real, most people wouldn't know it is fake [6].

Most deepfake technologies rely on advanced deep learning architectures such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and, more recently, Diffusion Models [7]. GANs are especially popular because they

DOI: https://doi.org/10.53704/fujnas.v15i1.1135

use a two-part system, a generator that creates fake content and a discriminator that judges how real it looks, to fine-tune the realism of the results. Many authors have argued that this back-and-forth setup helps the system improve over time, enabling it to produce realistic faces, voices, and even full-motion videos.

When it comes to swapping faces or generating facial images, many systems turn to autoencoder-based models. These are good at breaking down a face into simpler components, then reconstructing it while keeping the original structure and textures intact. Meanwhile, newer technologies such as Transformers and Diffusion Models are beginning to replace GANs, especially for high-resolution results.

Deepfakes aren't just an issue in entertainment or online content. They also pose serious concerns in areas like digital forensics. Because they can look so real, deepfakes make it harder to trust traditional signs of authenticity, things like camera metadata or subtle flaws in audio. As Mourad [8] points out, this growing realism puts pressure on experts to develop better detection methods and clearer ethical standards to help navigate this new digital landscape.

Differentiating deepfakes from traditional media manipulation

The manipulation of digital media predates the rise of deepfake technology, with traditional techniques having long been employed in domains such as entertainment, advertising, and political propaganda [9]. Traditional media manipulation typically involves manual or semi-automated editing techniques, including image splicing, audio dubbing, morphing, chroma key compositing, and frame-by-frame video alteration using software tools like Final Cut Pro or Adobe Premiere Pro. While such techniques can produce misleading content, they are labour-intensive and often detectable through forensic analysis due to visible inconsistencies, unnatural transitions, or metadata discrepancies [10].

In contrast, deepfakes represent a fundamental shift in how digital media is synthesised and manipulated. Rather than manually editing existing content, deepfake systems leverage machine learning, particularly deep generative models, to synthesise new content that mimics the statistical patterns of real data. As explained by researchers, these systems are trained on large datasets of faces, voices, or movements and learn to generate new, highly realistic outputs that match the target subject's appearance, speech patterns, or expressions. Once trained, such models can generate fake content without human intervention, and at scale.

What makes deepfake technology even more powerful and concerning is not just how realistic it looks or sounds, but how easy it is to use. Tasks like face-swapping, lip-syncing, and voice cloning that once required expert skills can now be done with just a few clicks, thanks to user-friendly tools like DeepFaceLab and Descript [11]. With these tools, it's now possible to make dozens, even hundreds, of fake videos or voice recordings in just a short time. In the past, doing that kind of editing would've taken hours or even days [12]. Because it's so fast and mostly automatic, deepfakes can be unsafe when people use them for the wrong reasons, like spreading false news, pretending to be someone else, or pulling off online scams.

Table 2 below shows some of the differences between older forms of media manipulation and modern AI-generated deepfakes: As this comparison makes clear, deepfake technology is not merely a

Table 2. Comparison between traditional media manipulation and deepfake synthesis

Feature	Traditional Media Manipulation	Deepfake Technology
Editing Method	Manual or semi-automated editing using software tools	AI-based generative modelling (e.g., GANs, VAEs, Diffusion)
Automation Level	Low	High
Realism	Moderate, detectable with trained observation	High, often indistinguishable from authentic media
Required Expertise	Video or audio editing skills	Minimal (via pre-trained models or apps)
Temporal Consistency	Often inconsistent across frames or audio segments	Preserves frame-to-frame or waveform consistency
Scalability	Time-intensive, limited to single edits	High-volume content generation possible
Forensic Detectability	Easier due to editing artefacts or metadata gaps	Difficult due to neural-level content realism
Example Tools	Photoshop, Premiere Pro, Audacity	DeepFaceLab, FaceSwap, Descript, ElevenLabs
Primary Use Cases (legacy)	Advertising, film editing, basic hoaxes	Fraud, misinformation, synthetic media, impersonation

DOI: https://doi.org/10.53704/fujnas.v15i1.1135

progression of previous editing techniques, it is a technological leap that automates and scales manipulation to levels previously unattainable. The shift from editing things by hand to using AI to make content has really changed the way people create and share media.

Historical evolution and technological milestones

Over the years, deepfake technology has evolved across different stages. The following subsections highlight those stages.

Early generative models and face manipulation (Pre-2015)

Older methods, such as morphing, warping, and basic image transformations, were mostly used before deep learning became popular. Those techniques often take a lot of time, and in most cases, the results usually don't look very real. However, we could see tangible improvements when encoder-decoder models were introduced. They enable basic face swaps to occur automatically, but only at low quality [13].

GAN era and open-source democratisation (2015-2020)

Fake images and videos have become more realistic with the advent of tools like GANs and VAEs. Around this time, applications such as DeepFake, DeepFaceLab, and FaceSwap emerged. Building on the culture of open-source projects and communities, many people, both experts and non-experts, began experimenting with the tools. Even an average person with not-so-much technical know-how could immediately start making faces and videos.

The introduction of NVIDIA's StyleGAN further advanced the field, giving users greater control over customisation.

High-fidelity and multimodal deepfakes (2021-2025)

Recent models such as GANSynth, AudioLM, and Make-A-Video allow text-guided and cross-modal generation, opening new avenues in synthetic journalism, digital avatars, and even crime [14].

Other authors have noted that the combination of high computational power, growing datasets, and publicly accessible APIs (e.g., ElevenLabs, D-ID) has driven the rapid adoption of these technologies across sectors. As noted by George et.[15], the emergence of real-time rendering and mobile integration in tools like Zao and Reface signifies the move of deepfakes into mainstream media consumption.

Modalities of deepfake generation

Figure 1 presents a hierarchical taxonomy of deepfake generation methods, categorised into three primary modalities: audio, image, and video. Each one uses a different kind of AI setup, depending on the media it's trying to mimic. Image deepfakes are mostly about swapping faces, tweaking features, or even building fake identities from scratch. Audio deepfakes focus on copying how someone talks, cloning their voice or generating speech that sounds like them. Then there are video deepfakes, which mix visuals and audio to alter entire scenes in a highly convincing way.

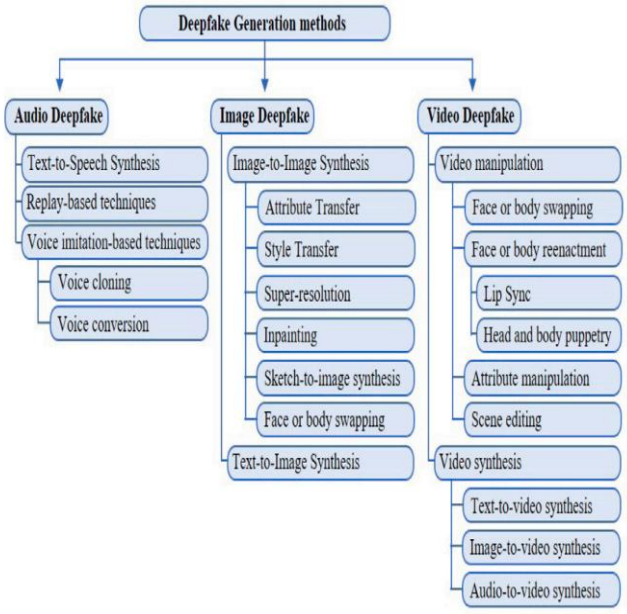

Figure 1. Modalities of deepfake generation with examples [1]

Each category contains specific subtypes based on the synthesis task. Audio deepfakes can include turning text into speech, replaying recorded voices, or cloning or converting someone's voice. With image deepfakes, tasks include transforming one image into another, swapping styles, tweaking features, filling in missing parts, or even creating images from sketches or written descriptions. When it comes to video, deepfakes usually fall into two categories: manipulation (like face-swapping, changing expressions, syncing lips to different audio, or editing scenes) and full video generation (like turning text or sound into videos). All these categories show just how advanced and varied deepfake techniques have become and why they're so tough to detect or regulate fairly.

DOI: https://doi.org/10.53704/fujnas.v15i1.1135

Image-based deepfakes

Image-based deepfakes focus on tasks such as swapping faces, altering facial features (e.g., expressions), or even creating new faces altogether [16]. Tools like StyleGAN2 and StyleGAN3 are often used to generate high-quality fake images that preserve a person's identity. As Dantcheva [17] notes, these techniques aren't just used for fun or art; they're also used in virtual makeup, creative photography, and even to trick biometric systems.

These tools can create entire identities that never existed, a phenomenon known as synthetic identity fraud. Undoubtedly, this poses challenges for face recognition systems, as AI-generated faces can bypass verification tools with high accuracy.

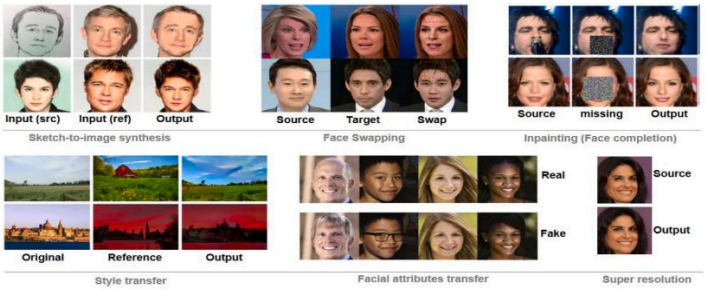

Figure 2. Common image-based deepfake generation techniques [1]

Figure 2 illustrates common image-based deepfake generation techniques, including sketch-to-image synthesis, face swapping, inpainting (face completion), style transfer, facial attribute transfer, and super-resolution. Each example highlights how AI can manipulate or enhance visual content with a high degree of realism.

Audio-based deepfakes

Audio deepfakes rely on text-to-speech (TTS) systems and voice conversion models [18]. Tacotron2, WaveNet, and FastSpeech2 are commonly used architectures, often paired with vocoders like MelGAN or HiFi-GAN. [19] Conducted a comparative study on deepfake audio detection, revealing that AI-generated voices can convincingly mimic pitch, intonation, and even speaker-specific prosody.

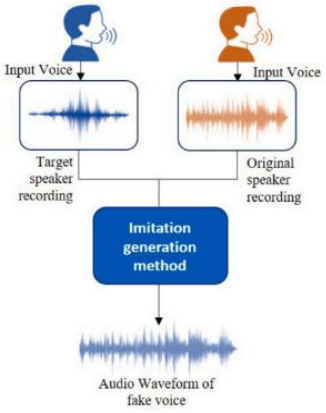

Figure 3. A typical audio-based deepfakes generation technique [2]

Figure 3 provides illustrative examples of audio-based deepfakes, showing how synthesised speech can mimic a target speaker's vocal characteristics. These systems use short voice samples to generate speech that matches the pitch, tone, and cadence of the original, often making it difficult to distinguish real from fake. The figure highlights the realism and potential deceptive power of voice cloning and text-to-speech models.

Schmitt and Flechais [20] discuss how voice cloning has been exploited in social engineering attacks, with adversaries synthesising executive voices to authorise fraudulent bank transactions. Zhang et al. [21] also point out that voice authentication systems are becoming easier to fool, especially now that deepfake tools can create fake voices using just a small sample of someone's real voice

Video-based deepfakes

Video deepfakes combine what you see and hear, often using techniques such as copying facial movements, syncing lips to new audio, or even animating a person's entire body. Tools like the First Order Motion Model (FOMM), for example, can take a still image of someone's face and make it move by following the motions from another video. As noted by Azeez et al. [22], the use of 3D face reconstruction and optical flow has significantly improved the temporal consistency of video deepfakes.

Von et al. [23] highlight the role of deep video portraits and neural head avatars in the

DOI: https://doi.org/10.53704/fujnas.v15i1.1135

Figure 4. A video-based deepfake generation instance [2]

creation of synthetic influencers, digital entities capable of interacting with viewers in real-time using AI-generated video content.

Figure 4 illustrates video-based deepfakes by showing AI-generated sequences where visual scenes evolve over time. In part (a), a dog in a superhero outfit is animated to fly across the sky, and in part (b), shifting sunlight animates a pile of books on a table by a window. These scenes showcase how generative models can simulate motion and lighting dynamics, producing highly realistic and coherent video content frame by frame [24]. This form of deepfake involves both temporal and spatial manipulation, enhancing the realism of synthetic videos.

Platforms for enabling deepfake generation

Numerous open-source and commercial platforms enable deepfake generation. Listed below are a few of them:

i. DeepFaceLab – Offers frame-by-frame face swapping and identity control.
ii. FaceSwap – Built on TensorFlow, widely used for research and education.
iii. Avatarify – Uses OpenPose and GANs to animate photos in real time.
iv. Zao – Chinese app that uses cloud-based face reenactment.
v. Descript Overdub – Enables voice editing for audio content.
vi. ElevenLabs – Offers multilingual voice cloning with emotion control.

Gamage et al. [25] note that many of these platforms include ethical disclaimers, but few enforce active mechanisms to prevent misuse.

Ethical and social implications

The social impact of deepfakes extends across disinformation, political manipulation, online harassment, and identity theft. The ease with which deepfakes can be generated and disseminated poses a threat to democratic processes, journalistic credibility, and the integrity of digital evidence.

Qureshi et al. [26] argue that the challenge lies not only in detection but also in managing public perception. Even real content can now be dismissed as fake, leading to a "liar's dividend" where bad actors evade accountability by citing deepfakes.

Even with all the risks, deepfakes can also be used in helpful ways. For example, in healthcare, they're being used to help patients with motor neuron disease speak again by recreating their voice. In education, deepfakes can bring history to life by simulating historical figures or creating avatars that speak different languages, making learning more accessible worldwide.

Khalid et al. [27] suggest that if we have the right ethical guidelines and policies in place, this kind of synthetic media could really improve how people interact with technology and make digital tools more personal and inclusive.

Deepfake generation techniques and real-world use cases

Deepfakes come in different forms depending on what they're trying to do. Some swap faces, some copy people's expressions, and others create completely fake videos or images, sometimes just tweaking little details to change the meaning. These tools have opened up new ways to communicate, learn, and make things more accessible, but they also bring serious problems. They can be used maliciously, such as to spread false information, commit identity theft, and run scams.

Generation tasks and technical mechanisms

Rather than only just classifying deepfakes based on the subject media type, such as image, audio, or video, we can also employ a task-based classification approach. i.e., they get classified based on what they are trying to do. Many researchers have emphasised that this kind of task-based approach makes more

DOI: https://doi.org/10.53704/fujnas.v15i1.1135

sense when looking at how deepfakes are used in real situations or how they can be detected. The next subsections present some of the most common methods for creating deepfakes.

Face swapping

This means swapping one person's face with another's while keeping the original head position, lighting, and facial expressions unchanged. Usually, this is done using encoder-decoder setups, where two autoencoders are trained separately on two different faces, and then the decoder from the target face is combined with the encoder output from the source face [28]. Popular programs like DeepFaceLab and FaceSwap use this method, often speeding up processing with GPUs and adding finishing touches such as colour correction and blending to make the swap look more natural [29].

It is important to note that using specialised loss functions and image comparison methods helps ensure that textures appear consistent and that hidden parts are handled more effectively. Face swapping is popular in movies and mobile apps, but it's also used for harmful things like creating non-consensual explicit content or pretending to be someone else to commit fraud [30].

Facial reenactment

Facial reenactment means changing the facial expressions of someone in a video to match another person's expressions or a set of input movements [31]. Unlike face swapping, the person's identity stays the same, but features like their expressions, where they're looking, or how their mouth moves can be changed. This is usually done using techniques like tracking key facial points, 3D face models, or AI methods such as the First Order Motion Model [32].

People use facial reenactment for tasks such as dubbing movies in different languages while keeping lip movements in sync, creating avatars for live video chats, or even producing fake interviews and speeches [33].

Face synthesis

Face synthesis involves generating entirely new identities or images of faces that do not correspond to real individuals [34]. This is commonly done with models such as StyleGAN2 and StyleGAN3, which offer control over facial attributes such as age, ethnicity, hairstyle, and expression. These models are trained on large-scale face datasets and can interpolate between latent space vectors to produce novel outputs [35].

Synthetic faces are widely used in privacy-preserving advertising, virtual influencers, and simulation training systems. However, they are also used to create fake social media accounts and bypass facial authentication systems.

Attribute editing

Attribute editing allows the modification of specific facial or voice attributes (e.g., adding glasses, changing hair colour, modifying voice pitch or accent) while preserving other core features. This task relies on disentangled latent representations that enable control over single variables. Conditional GANs and encoder-decoder transformers are often used in this context.

Some authors have pointed out that attribute editing is integrated into user-facing applications such as Snapchat filters and TikTok's face effects, but also underpins tools for manipulating biometric characteristics in evasion attacks.

Constructive applications

Despite their controversial reputation, deepfakes are not inherently malicious. For instance, deepfakes can be utilised in entertainment to bring historical personalities to life or produce realistic special effects in films and TV series. Lastly, deepfakes can be utilised in marketing to produce customised ads or product demos. It can further be favourably utilised in the health and education sectors. Their potential in ethical and innovative applications is well-documented:

Film and entertainment

Studios increasingly use face swapping and reenactment to create digital doubles, de-age actors, or simulate scenes with unavailable cast members. Ali et al. [36] highlight the adoption of mobile apps like Reface and Zao, which brought deepfake capabilities to casual users and short-form content creators.

Education and cultural preservation

Ho et al. [37] document the use of deepfake avatars in virtual learning environments. These include historical re-enactments, AI instructors for multilingual lessons, and interactive teaching tools in STEM subjects. The ability to customise language, age, and teaching style enhances accessibility to learning.

DOI: https://doi.org/10.53704/fujnas.v15i1.1135

Assistive technology and healthcare

Voice cloning tools help patients with degenerative diseases (e.g., ALS) recover their voice or communicate using AI-synthesised speech. Almutairi and Elgibreen [38] describe how high-fidelity voice synthesis, trained on short samples, now matches emotional tone and pacing, enabling personalised speech aids.

Advertising and localisation

Brands use synthetic spokespersons to create region-specific marketing without hiring local actors. These avatars can be customised for age, language, and brand identity while maintaining message consistency across demographics [39]. This has proved highly advantageous for actors and stakeholders in the advertising industry.

Malicious use cases and threat scenarios

The growing accessibility of deepfake tools and lack of regulatory control have led to a surge in misuse across the globe:

Non-consensual content and harassment

Duquette [40] reports that one of the earliest uses of face-swapping technology was in creating fake explicit videos of celebrities and public figures. These media are distributed to harass, extort, or discredit individuals.

Fraud and financial scams

Other researchers highlight voice-cloning fraud in which attackers simulate a company executive's voice to authorise huge financial transfers. These real-time impersonation scams have affected corporate and banking sectors, bypassing traditional phone-based verification systems [41].

Political disinformation

Mangotra [42] describes fabricated speeches and doctored campaign videos used to spread false information during elections. The rapid virality of such content on social media platforms complicates early detection and public correction.

Bypassing biometric systems

Face synthesis and attribute editing are used to generate images that match multiple individuals in biometric databases. Several authors have identified this as a key concern for national ID systems, e-passports, and facial access controls.

Documented case studies

Several real-world cases underscore the societal implications of deepfake misuse:

a) In 2019, a German energy firm reported a case in which attackers used an AI-generated voice to impersonate the CEO and fraudulently direct a €220,000 transfer. This remains a pivotal example of audio deepfake fraud [43].

b) In 2022, during the Russia–Ukraine conflict, there was a circulation of a deepfake video showing the Ukrainian President Zelensky urging surrender. Though debunked within hours, the video demonstrated the use of facial reenactment for wartime propaganda.

c) In India, a tech startup launched a platform where synthetic avatars delivered online courses in rural dialects. In 2024, Gupta noted the program's success in boosting educational outreach, although it raised questions about consent and content traceability [44].

Matrix of use cases across sectors

Table 3 below summarises how these generation tasks map to applications across various domains:

Table 3: Functional Mapping of Deepfake Tasks to Real-World Sectors

Generation Task	Constructive Use Case	Malicious Use Case
Face Swapping	Film dubbing, actor substitution	Non-consensual explicit content, identity fraud
Facial Reenactment	Multilingual news anchors, virtual avatars	Fabricated public statements, political manipulation
Face Synthesis	Synthetic identities for privacy, virtual training	Fake social media profiles, facial spoofing attacks
Attribute Editing	Augmented reality filters, accessibility design	Circumventing biometric verification, data poisoning

DOI: https://doi.org/10.53704/fujnas.v15i1.1135

AI-based detection of deepfakes: techniques, performance, and challenges

As deepfake generation techniques grow increasingly sophisticated and accessible, detecting synthetic content has become a critical research priority in digital media forensics. In the past, one could easily look at an image or video and tell whether it was fake. However, in recent times, that has become almost impossible.

Artificial intelligence is now mostly used to help handle this. Models are trained to notice every little details not obvious to an average human.

Key categories of AI-based detection techniques

AI-Based deepfake detection can be categorised into four main categories based on how the models are built and what they focus on.

Convolutional neural networks (CNNs)

Convolutional Neural Networks, or CNNs, are often used to detect deepfakes in images and individual video frames. They operate by analysing images in layers and then using specialised filters to flag features such as edges, textures, and unusual patterns. They are especially good at catching visual glitches such as uneven blending, skin-texture problems, or compression artefacts that often occur when faces are swapped.

Li et al. [45] show that CNNs, when trained on popular datasets such as FaceForensics++ and Celeb-DF, can correctly detect deepfakes with a 90% success rate in controlled settings. They also noted that various CNN architectures, such as XceptionNet, EfficientNet, and ResNet, are commonly used for this type of detection.

Often, these models analyse each video frame individually and then combine the results to determine whether the entire clip is real or fake. But a drawback of CNNs is that they can miss changes that occur over time, such as unusual movements or mismatched sounds, so they aren't as good at spotting problems that span multiple frames.

Recurrent and temporal models

To overcome the limits of models that can only process single images or frames, researchers introduced recurrent neural networks (RNNs) and their variants, such as long short-term memory (LSTM) and gated recurrent unit (GRU) networks. These models are good at understanding sequences, so they work better for spotting deepfakes in videos and speech [46].

Similar temporal models are also used to detect fake audio by tracking changes in pitch, rhythm, and tone. These models often analyse spectrograms or raw audio waveforms that are prepared using specialised signal-processing methods.

Attention and transformer-based models

Models such as Vision Transformers (ViT) and Swin Transformers can sometimes outperform traditional CNNs, especially when they have access to large amounts of data. They're better at detecting various kinds of changes or edits, since they examine entire images or sequences at once [47].

Many authors have established that attention layers are useful for detecting subtle changes in facial features, such as mismatches around the eyes or mouth, or a mismatch between someone's facial expressions and what they are saying.

Multi-modal detection systems

Some deepfakes combine multiple types of data—for example, a video with fake audio. Detecting these kinds of manipulations requires systems that can handle different types of information simultaneously. Multimodal detection systems do exactly that by analysing images, sounds, and sometimes even text to find inconsistencies across these modalities.

In his work [48], Francis describes models that combine image-based and audio-based classifiers. For example, models that use both Mel-Frequency Cepstral Coefficients (MFCC) and log-mel spectrograms tend to do better at spotting voice cloning, even across different languages [49].

Some advanced systems go a step further by combining CNNs, RNNs, and attention layers into one model. This layered approach helps the system perform better even when the input media is of poor quality, like blurry videos or noisy audio [50].

Datasets for training and evaluation

Deepfake detection models require high-quality, diverse, and labelled datasets for training and benchmarking [51]. Several datasets have been instrumental in advancing AI-based forensic research:

i. FaceForensics++: Over 1,000 real and manipulated video pairs with varying compression levels. Used for face manipulation detection.

DOI: https://doi.org/10.53704/fujnas.v15i1.1135

ii. DFDC (Deepfake Detection Challenge): More than 100,000 videos representing multiple actors and manipulation methods. Sponsored by Meta.
iii. Celeb-DF: High-quality deepfake video dataset with improved realism. Addresses issues of overfitting on FaceForensics++ and better reflects real-world scenarios.
iv. ASVspoof: is a dataset made up of both real and fake speech samples. The fake ones are created using text-to-speech and voice conversion techniques. This dataset is often used to train and test systems for voice cloning detection.
v. FakeAVCeleb: is a dataset that includes deepfakes with both audio and video, where the sound and images are synced up. It's used to help train models that analyse multiple data types simultaneously.

How well AI detection models work usually depends a lot on the variety and quality of the data they're trained on. If a model learns only from a small range of fake examples or very carefully chosen data, it might struggle when faced with new or different kinds of deepfakes in the real world [52].

Performance metrics and comparative results

It is really important to properly test how well AI deepfake detection systems perform to determine whether they can be trusted in real-world situations. Usually, their performance is measured using various metrics and scores that show how well they can distinguish real content from fake content. These measures also help us understand how the system handles tough cases, such as when videos are compressed, come from different sources, or are of poor quality.

Common evaluation metrics

i. Accuracy: shows the percentage of samples the system correctly labels out of all the test cases. While it's a common way to measure performance, accuracy can be misleading if the dataset isn't balanced or if certain types of manipulation appear more often than others.
ii. Precision and Recall: Precision tells us how many of the samples the system marked as manipulated were actually fake. Recall, sometimes called sensitivity or the true positive rate, indicates how many of the actual manipulated samples the system detected. Both of these measures are really important, especially when the cost of missing a fake or of wrongly flagging something real can vary.
iii. F1 Score: The harmonic mean of precision and recall, offering a balanced view of a model's ability to minimise both types of classification errors.
iv. AUC (Area Under the ROC Curve): AUC quantifies the ability of a binary classifier to distinguish between classes across all threshold settings. Higher AUC values (closer to 1.0) indicate better separability.
v. EER (Equal Error Rate): Frequently used in biometric and audio deepfake detection, EER is the error rate at which the false acceptance rate equals the false rejection rate. A lower EER suggests more accurate classification [53].

Empirical performance on benchmark datasets

The empirical results provide a comparative understanding of how various detection techniques perform across benchmarks. Table 4 summarises selected findings from notable studies in video and audio deepfake detection. These findings reveal several important trends:

i. Models perform significantly better on the datasets they are trained on, but show degraded performance on unseen manipulations, confirming the existence of a generalisation gap.
ii. Temporal models (e.g., CNN-LSTM) are generally more robust for video detection due to their ability to capture motion-based anomalies such as unnatural blinking, inconsistent head movements, or desynchronized lip-sync.
iii. In audio detection, ensemble approaches that combine spectral and temporal features (e.g., MFCCs, log-Mel spectrograms) outperform single-stream models, particularly in noisy or adversarial conditions.
iv. Lightweight models like EfficientNet show promising performance with reduced computational cost, making them more suitable for real-time or mobile deployments.

Figure 5 provides a visual summary of the top performance metrics reported in the selected studies. It enables a comparative understanding of how models perform across different modalities and benchmark datasets.

Table 4. Performance summary of AI-based detection techniques

DOI: https://doi.org/10.53704/fujnas.v15i1.1135

Study	Media Type	Model Type	Dataset	Accuracy/F1/AUC	Notes
[23]	Video	CNN-LSTM hybrid	FaceForensics++	Accuracy: 94% (HQ); ↓71–80%	Accuracy drops on compressed/unseen data
[28]	Audio	Spectrogram CNN Ensemble	ASVspoof 2021	F1 Score: 0.92+	Multi-feature input improves robustness
[1]	Video	XceptionNet CNN	DFDC, Celeb-DF	Accuracy: 88–91%	Lower AUC on lower quality samples
[1])	Image	EfficientNet-B5	Celeb-DF, FaceForensics++	AUC: up to 0.98	High precision on high-res; generalisation untested

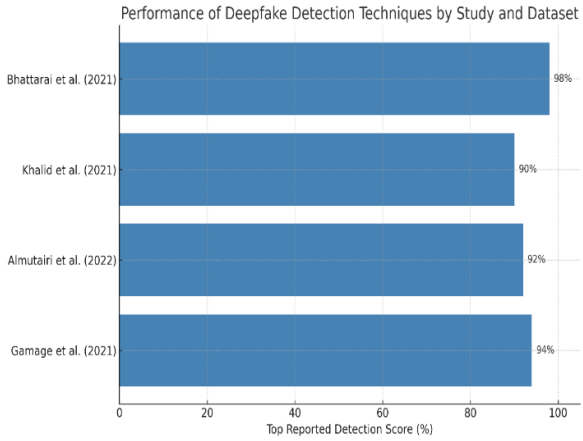

Figure 5: Comparative performance of selected AI-based deepfake detection models on benchmark datasets. Values represent the top reported metric (Accuracy, F1 Score, or AUC) under optimal testing conditions

Cross-dataset generalisation and performance drop-off

A key takeaway from deepfake detection research is that models often struggle with new data types or manipulation methods they weren't trained on. For example, Seng et al. [54] show that models trained on DFDC videos performed very well during validation, with AUC scores above 0.90, but their performance dropped to 0.78 or lower when tested on videos subjected to different manipulation techniques.

This issue is especially noticeable in audio deepfake detection. Some authors have pointed out that the effectiveness of these models can vary widely depending on the language, accent, or speaking style of the people involved. For example, a model trained on English speakers might perform very well, achieving high F1 scores, but its accuracy can drop significantly when tested on Arabic or Mandarin datasets, where intonation and phrasing differ markedly.

These performance differences show that although deep learning models can perform very well in controlled lab environments, using them effectively in real-world situations requires training them on diverse, representative datasets that reflect real-life conditions.

Concluding observations

After reviewing, a few key thoughts are stated as follows:

i. AI-based detection systems can work really well, but their effectiveness heavily depends on factors like the quality and variety of datasets, the type of manipulation being tested, and even the hardware resources used to run them.

ii. Deepfakes are becoming easier and faster to create, which means our detection systems need to evolve just as quickly. They have to be adaptable and ready to handle new, unseen forms of manipulated content.

iii. There's a strong need for multi-modal detection approaches, systems that can simultaneously analyse images, audio, and video for signs of manipulation.

iv. Adversarial deepfakes (those designed to trick detection systems) are a growing concern. Detection models need to be trained in ways that make them more resistant to these types of attacks.

v. Real-time detection is becoming more important. It's not just about accuracy; it's also about speed. We need tools that can catch deepfakes as they spread, not after the damage is done.

All of these points point to one thing: staying ahead of deepfakes will take continuous research, smarter training strategies, and stronger tools that can actually be deployed in real-world situations.

Current challenges

Generalisation and robustness:

Many of the leading models perform well on familiar data but lose accuracy when faced with new

DOI: https://doi.org/10.53704/fujnas.v15i1.1135

types of deepfakes or videos altered in different ways. For instance, ensemble models trained on spectrogram features for audio detection perform impressively on known spoofing patterns but degrade significantly under noisy conditions or new attack types [55].

Dataset limitations:

Most existing benchmarks are constrained by a limited number of subjects, fixed generation pipelines, or uniform manipulation quality. In their paper, some authors have demonstrated that datasets such as Celeb-DF still fall short of capturing the full spectrum of demographic diversity, post-processing, and cross-modal manipulations, leading to biased detection by disproportionately misclassifying specific ethnicities or age groups.

Adversarial adaptation:

As detection techniques evolve, so too do generation methods. Newer architectures, such as diffusion models and transformers, like StyleGAN3 and GANformer, can learn to evade specific forensic cues, prompting a cat-and-mouse dynamic between creators and detectors. This adversarial adaptation reduces the shelf life of any standalone detector.

Ethical and legal ambiguities

The impact of deepfakes on society is made worse by the lack of clear rules and regulations in many places. Although some countries, such as China, the U.S., and the EU, have begun developing policies to manage the spread of deepfakes, enforcement remains inconsistent and often occurs only after problems arise. Moreover, the line between protected expression and harmful manipulation remains blurry, complicating efforts to legislate effectively.

Future research direction

Cross-modal detection

A promising area of future work involves developing detection frameworks that simultaneously analyse visual, auditory, and linguistic signals. Multimodal fusion networks, for instance, can detect mismatches among facial expressions, vocal tone, and the meaning of spoken words. This helps these systems make more reliable decisions when identifying deepfakes.

Self-supervised and continual learning

Since labelled deepfake data is expensive and slow to generate, self-supervised learning offers a scalable alternative. Models trained on contrastive representations can detect anomalies without relying on curated labels. Continual learning methods would allow detectors to adapt to new attack types without catastrophic forgetting.

Explainable and transparent forensics

There is a growing need for explainable AI in media forensics, particularly for legal and journalistic applications. Instead of binary real/fake classifications, detectors should offer interpretable rationales (e.g., "inconsistencies in lip synchronisation" or "frequency domain anomalies") to aid human verification and court admissibility.

Real-time detection in the wild

As deepfakes move into live-streaming platforms and real-time communications, there is a technical push toward lightweight, on-device detection. Research into neural network compression, knowledge distillation, and edge inference will be instrumental in building deployable solutions.

Future research direction

Watermarking and provenance protocols

Researchers and policymakers increasingly advocate for built-in watermarking during content creation. Standards like the Coalition for Content Provenance and Authenticity (C2PA) aim to ensure that media origins and modification histories are traceable. Incorporating metadata signatures into AI-generated content could complement detection tools.

Public education and digital literacy

One effective way to fight deepfakes is to help people learn more about digital media. Running public campaigns, especially for groups who might be more at risk, like older adults or political activists, can make it harder for fake content to be believed and spread.

Platform accountability

Technology companies and content-hosting platforms must assume greater responsibility. This approach involves actively filtering out deepfake content before it spreads, collaborating with research labs to share information on new threats, and being clear about how and when fake content is removed.

DOI: https://doi.org/10.53704/fujnas.v15i1.1135

Conclusion and final remarks

Deepfakes are artificial intelligence (AI)-generated audio, video, or image content that can be extremely lifelike, making it difficult to tell what is real and what isn't. Because deepfakes can be used to spread false information, commit identity fraud, and threaten national security and privacy, identifying them has become a major concern in digital media forensics. Frame-by-frame analysis, blink analysis, edge analysis, error level analysis, and speed analysis are a few of the popular detection methods. Convolutional Neural Networks (CNNs), which are effective at capturing subtle artefacts generated by deepfake production processes, are one of the Deep Learning-Based Detection Methods.

Generative Adversarial Networks (GANs), which are widely used to identify deepfakes by examining the differences between generated and real content, and Recurrent Neural Networks (RNNs), which are highly effective for analysing temporal aspects in videos. It is crucial to acknowledge the difficulties with this method. The most notable of these is the complexity of deepfakes. This problem makes it harder to identify deepfakes as deepfake technology develops. Additionally, the efficacy of deepfake detection models is hampered by the dearth of extensive datasets for training and testing.

Lastly, it can be difficult to tell whether a video or image is real or fake because deepfake detection techniques often fail to understand its context. Finally, one important future goal is to develop methods that provide insights into the decision-making processes of deepfake detection models. To improve the accuracy of deepfake detection, researchers are also focusing on integrating visual, audio, and other modalities.

Acknowledgement

The authors wish to acknowledge the grant —CRC/TETFUND/NO 2024/02 —received from the Office of the Deputy Vice Chancellor (A&R) through the Director of the Research Management Office (RMO) of the University of Lagos, Lagos, Nigeria, to execute this research project.

References

1. Bendiab G, Haiouni H, Moulas I, Shiaeles S. Deepfakes in digital media forensics: Generation, AI-based detection and challenges. *J Inf Secur Appl*. 2025;88:103935. doi:10.1016/j.jisa.2024.103935.
2. Nguyen TT, et al. Deep learning for deepfakes creation and detection: A survey. *Comput Vis Image Underst*. 2022. doi:10.1016/j.cviu.2022.103525.
3. Ghiurău D, Popescu DE. Distinguishing reality from AI: Approaches for detecting synthetic content. *Computers*. 2025;14(1):1. doi:10.3390/computers14010001.
4. Carpenter P. *Faik: A practical guide to living in a world of deepfakes, disinformation, and AI-generated deceptions*. Hoboken (NJ): John Wiley & Sons; 2024.
5. Farid H. Creating, using, misusing, and detecting deep fakes. *J Online Trust Saf*. 2022;1(4). doi:10.54501/jots.v1i4.56.
6. Wood E, Baltrušaitis T, Hewitt C, Dziadzio S, Cashman TJ, Microsoft JS. Fake it till you make it: Face analysis in the wild using synthetic data alone [Internet]. 2021 [cited 2026 Apr 18]. Available from: https://microsoft.github.io/FaceSynthetics
7. Bengesi S, El-Sayed H, Sarker MK, Houkpati Y, Irungu J, Oladunni T. Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. *IEEE Access*. 2024;12:69812–69837. doi:10.1109/ACCESS.2024.3397775.
8. Mourad B. Deepfakes and media integrity: Navigating the new reality of synthetic content. 2024.
9. Olanipekun SO. Computational propaganda and misinformation: AI technologies as tools of media manipulation. *World J Adv Res Rev*. 2025;25(1):911–923. doi:10.30574/wjarr.2025.25.1.0131.
10. Ogundiran A. A goal-oriented visualization approach to digital forensics evidence presentation. 2024.
11. Singh AK. Deep learning for deepfakes creation and detection. 2024.
12. Masood M, Nawaz M, Malik KM, Javed A, Irtaza A. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. 2021.
13. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J. Deepfakes and beyond: A survey of face manipulation and fake detection. *arXiv* [Internet]. 2020 [cited 2026 Apr 18]. Available from: http://arxiv.org/abs/2001.00179
14. Bandi A, Adapa PVSR, Kuchi YEVPK. The power of generative AI: A review of requirements, models, input–output formats, evaluation metrics, and challenges. *Future Internet*. 2023;15(8):260. doi:10.3390/fi15080260.
15. George AS, Hovan George AS. Deepfakes: The evolution of hyper realistic media manipulation. 2023. doi:10.5281/zenodo.10148558.
16. Waseem S, Abu Bakar SAR, Ahmed BA, Omar Z, Eisa TAE, Dalam MEE. Deepfake on face and expression swap: A review. *IEEE Access*. 2023;11:117865–117906. doi:10.1109/ACCESS.2023.3324403.
17. Dantcheva A. Computer vision for deciphering and generating faces [Internet]. 2021 [cited 2026 Apr 18]. Available from: https://hal.archives-ouvertes.fr/tel-03500318
18. Ferrari E, Spolaor R, Janeja VP. Audio deepfakes: A survey [Internet]. 2023 [cited 2026 Apr 18].
19. Warren K, Olszewski D, Layton S, Butler K, Gates C, Traynor P. Pitch imperfect: Detecting audio deepfakes through acoustic prosodic analysis. *arXiv* [Internet]. 2025 [cited 2026 Apr 18]. Available from: http://arxiv.org/abs/2502.14726
20. Schmitt M, Flechais I. Digital deception: Generative artificial intelligence in social engineering and phishing.

DOI: https://doi.org/10.53704/fujnas.v15i1.1135

Artif Intell Rev. 2024;57(12):324. doi:10.1007/s10462-024-10973-2.
21. Zhang Z, et al. Mitigating unauthorized speech synthesis for voice protection. 2024. doi:10.1145/3689217.3690615.
22. Azeez NA, Adefemi F, Olayinka, Fasina EP, Venter IM. Evaluation of a flexible column-based access control security model for medical-based information. *J Comput Sci Appl Niger Comput Soc*. 2015;22(1):24–31.
23. Von H, Verlag H, Scorzin PC. AI body images and the meta-human: On the rise of AI-generated avatars for mixed realities and the metaverse. *Interdiscip J Image Sci*. 2023;37(1).
24. Okonkwo I, Mujinga J, Namkoisse E, Francisco A. Localization and global marketing: Adapting digital strategies for diverse audiences. *J Digit Mark Commun*.
25. Gamage D, Ghasiya P, Bonagiri V, Whiting M, Sasahara K. Are deepfakes concerning? Analyzing conversations on Reddit and exploring societal implications. *arXiv*. 2022. doi:10.48550/arXiv.2203.15044.
26. Qureshi SM, Saeed A, Almotiri SH, Ahmad F, Ghamdi MAA. Deepfake forensics: A survey of digital forensic methods for multimodal deepfake identification on social media. *PeerJ Comput Sci*. 2024;10:e2037. doi:10.7717/peerj-cs.2037.
27. Khalid H, Tariq S, Kim M, Woo SS. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv*. 2021. Available from: http://arxiv.org/abs/2108.05080
28. Ogundiran A. A goal-oriented visualization approach to digital forensics evidence presentation. 2024.
29. George AS, Hovan George AS. Deepfakes: The evolution of hyper realistic media manipulation. 2023. doi:10.5281/zenodo.10148558.
30. Lin J, Latoschik ME. Digital body, identity and privacy in social virtual reality: A systematic review. *Front Virtual Real*. 2022. doi:10.3389/frvir.2022.974652.
31. Altuncu E, Franqueira VN, Li S. Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review. *Front Big Data*. 2024;7:1400024.
32. Lin C, Deng J, Hu P, Shen C, Wang Q, Li Q. Towards benchmarking and evaluating deepfake detection. *arXiv*. 2022. Available from: http://arxiv.org/abs/2203.02115
33. Deng J, Lin C, Hu P, Shen C, Wang Q, Li Q. Towards benchmarking and evaluating deepfake detection. *IEEE Trans Dependable Secure Comput*. 2024.
34. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf Fusion*. 2020;64:131–148.
35. Naffi N, Charest M, Danis S, Pique L, Davidson AL, Brault N, et al. Empowering youth to combat malicious deepfakes and disinformation. *J Construct Psychol*. 2025;38(1):119–140.
36. Ali G, Rashid J, Hussnain MRU, Tariq MU, Ghani A, Kwak D. Beyond the illusion: Ensemble learning for effective voice deepfake detection. *IEEE Access*. 2024.
37. Ho W, Woo HS, Lee DH, Kim Y. Development and distribution of deep fake e-learning content videos using open-source tools. *J Distrib Sci*. 2022;20(11):121–129.
38. Almutairi Z, Elgibreen H. A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms*. 2022;15(5):155. doi:10.3390/a15050155.
39. Asan E, Berardi FB, Dostliev A. Dispatches from Ukraine: Tactical media reflections and responses. 2022.
40. Duquette H. Digital fame: Amending the right of publicity to combat advances in face-swapping technology [Internet]. 2020 [cited 2026 Apr 18]. Available from: https://perma.cc/Z4LE-7HDF
41. Dolhansky B, et al. The DeepFake Detection Challenge (DFDC) dataset. *arXiv*. 2020. Available from: http://arxiv.org/abs/2006.07397
42. Mangotra BD. Evaluating deepfake detection models: A comprehensive framework for comparison across diverse datasets. 2025.
43. Reis PMGI, Ribeiro RO. A forensic evaluation method for deepfake detection using DCNN-based facial similarity scores. *Forensic Sci Int*. 2024;358:111747.
44. Gupta SK. Pedagogy and education management review. *Pedag Educ Manag Rev*. 2024;4(18):4–24.
45. Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-DF: A large-scale challenging dataset for deepfake forensics [Internet]. 2020 [cited 2026 Apr 18]. Available from: https://deepfakedetectionchallenge.ai
46. Azeez NA, Venter IM. Towards ensuring scalability, interoperability and efficient access control in a multi-domain grid-based environment. *SAIEE Afr Res J*. 2013;104(2).
47. Azeez NA, Adefemi F, Olayinka, Fasina EP, Venter IM. Evaluation of a flexible column-based access control security model for medical-based information. *J Comput Sci Appl Niger Comput Soc*. 2015;22(1):24–31.
48. Francis N. Deepfake detection and defense: An analysis of techniques and robustness. 2025.
49. Azeez NA, Vyver CV. Digital education: Assessment of e-learning and m-learning adoption in tertiary institutions in South Africa. In: *Proc IEEE IC3e*. Langkawi (Malaysia); 2018. p. 21–22.
50. Azeez NA, Iliyas HD. Implementation of a 4-tier cloud-based architecture for collaborative health care delivery. *Niger J Technol Dev*. 2016;13(1):17–25.
51. Venkateswarulu S, Srinagesh A. DeepExplain: Enhancing deepfake detection through transparent and explainable AI model. *Informatica*. 2024;48(8).
52. Azeez NA, Ajetola AR, Oyewole AS. Assessment of existing gap between industrial IT skill requirements and computer science curriculum in tertiary institutions. *Pac J Sci Technol*. 2009;10(2):326–336.
53. Azeez NA, Salaudeen BB, Misra S, Damaševičius R, Maskeliūnas R "Identifying phishing attacks in communication networks using URL consistency features" International Journal of Electronic Security and Digital Forensics. Vol. 12(2), 2020. pp. 200-21
54. Seng LK, Mamat N, Abas H, Ali WNHWA. AI integrity solutions for deepfake identification and prevention. *Open Int J Inform*. 2024;12(1):35–46.
55. Ahmad J, Salman W, Amin M, Ali Z, Shokat S. A survey on enhanced approaches for cyber security challenges based on deep fake technology in computing networks. *Spectr Eng Sci*. 2024;2(4):133–149.

DOI: https://doi.org/10.53704/fujnas.v15i1.1135