Mitigating Deepfake-based Impersonation and Synthetic Data Risks in Remote Healthcare Systems

Ruvimbo Mashinge
Yeshiva University - Computer Science

Kumbirai Bernard Muhwati
Yeshiva University - Computer Science

Kelvin Magora
Yeshiva University – Computer Science

Joy Awoleye
Yeshiva University – Computer Science

ABSTRACT

The security and integrity of remote healthcare systems are raised as an urgent issue of fighting deepfake technologies, voice cloning, and synthetic data. Since telemedicine platforms are increasingly relying on audiovisual communication and electronic health records (EHR), they actively become appealing victim targets in high-level impersonation attacks. An exhaustive architecture to curb threats postulated through deepfakes, through a combination of multimodal biometric (face, voice, gesture) authentication, real-time deepfake detection, and provenance tracking using blockchain is proposed in this research. We can show that the proposed system can achieve higher than 95 per cent detection accuracy and can eliminate session compromise within two seconds using simulated attack scenarios on well publicised data sets including DFDC, VoxCeleb and MIMIC-III. Our results verify that multi-layered defenses have the potential of securing clinical integrity and patient privacy without much impairment of user experience. The paper establishes the working principle of scalable, resilient, adaptive, and secure telehealth ecosystems against changing threats in synthetic media settings.

General Terms

Security, Pattern Recognition, Human-Computer Interaction, Algorithms, Telemedicine Systems, Deep Learning, Data Integrity.

Keywords

Deepfakes, Telehealth Security, Voice Cloning, Synthetic Data, Multimodal Biometrics, Blockchain, Remote Healthcare, Impersonation Detection, Electronic Health Records, Adversarial AI

1. INTRODUCTION

The extraordinary and long-lasting growth of remote healthcare, often referred to as telehealth or telemedicine, has become one of the most iconic developments in contemporary medicine, particularly following the COVID-19 pandemic. This change was not a reactionary measure; instead, it marked a structural shift in the provision of healthcare services. Before 2020, remote consultations were primarily episodic and were typically restricted to underserved communities in rural areas. Nevertheless, increasing demands to limit viral transmission, maintain clinical capacity, and ensure patient access accelerated the speed at which these measures were implemented in both outpatient and inpatient care. Among OECD member states, a notable trend was observed, with the proportion of doctor-patient interactions being remotely carried out decreasing from a high of 21 percent in 2020, during the peak of the pandemic, to less than 1 percent in 2019 [1]. Outpatient telehealth visits increased by 154 percent in a single week in the United States in March 2020, and the number of appointments rose 15-fold, according to early numbers from the Cleveland Clinic [2]. Several healthcare systems experienced a 22 percent increase in all outpatient visits taking place virtually in the post-pandemic world compared to non-pandemic times. The introduction of policies in the form of insurer coverage, changes in legislation, and technological advancements, including internet coverage through broadband and the unification of video-health resources, have only entrenched this new normal further [3].

The dividends of these seismic shifts are quite substantial: increased access to geographically distant and mobility-limited patients, a reduction in the burden on healthcare infrastructure, and assurance of continuity of care, most significantly in mental healthcare and for patients with chronic conditions. For example, behavioral health services currently have more than 60 percent of consultations conducted virtually, resulting in increased patient adherence and reduced no-shows [2] Home-based management of chronic diseases can be achieved with the help of remote monitoring of wearable devices and Internet of Things (IoT) systems, minimizing the chances of hospital readmission and making the entire process more cost-efficient. Even in the UK, NHS England has already piloted virtual hospital wards to alleviate the problem of hospital bed shortages and to continue providing high-acuity care in people's home environments [4]. Therefore, telemedicine has not only become a short-term solution but also a tactical part of innovative healthcare systems. But with this revolution in care delivery, new vulnerabilities appear. Of primary concern are those related to identity spoofer and data swindler, which are facilitated by the AI-accompanied production of content, especially deeply fabricated content. The constantly growing complexity of generative adversarial networks (GANs), voice cloning models (e.g., Tacotron2 or SV2TTS), and synthetic data pipelines allows impersonation of providers or patients to a worrying extent. During virtual consultations, malicious actors may use audiovisual deception to convincingly pose as clinicians or patients and alter electronic health records to modify medical history, prescriptions, test results, or imaging records. A recent report highlights how AI-generated deep fake content is currently being used in phishing, fraud, and even the synthesis of clinician voices or faces to approve a false prescription, falsify documentation, or alter patient treatment [5].

The malicious influence of this kind of attack is extensive. In the case of decisions being made based on falsified images or interviews, the risk of getting misdiagnosed is a very probable possibility. Financial and patient-safety risks are associated with fraudulent behavior (including phony claims, insurance fraud, and prescription forgery). Additionally, there is confusion regarding legal liability in light of the actual harms caused by deepfake attacks. Telehealth-based institutions can be legally sued, fined by regulatory authorities, and have their trustworthiness publicly questioned, thereby eroding trust in them. Meanwhile, the impersonation of either the providers or patients can lead to psychological trauma and loss of privacy when it comes to unpermitted access or even misuse of their identity or medical information by an impersonator. The effects are indeed of a technological, clinical, legal, and ethical nature. Technologically, in terms of technology, authentication systems that were initially designed to handle face-to-face access, such as ID checks and biometrics, may not function effectively on synthetic content. Between the clinics, the lack of physical examination makes them even more dependent on digital signals that attackers can fake. Currently, the framework of medical liability may not explicitly cover situations where deepfakes can be used to facilitate fraud or other malicious activities. Ethically, patients are at risk of losing autonomy and the value of consent, as well as the nature of confidentiality, due to breaches that occur because of artificial content, particularly when parties to the consultation are not reliably distinguishable.

Moreover, the critical part of remote healthcare digital infrastructure vulnerabilities is versatile. Cloud-based breaches are likely not the only problem for telehealth platforms, where there is also a risk of manipulated streamed media. Alternatively, counterfeit audiovisuals can be incorporated into consultations to misinform clinicians, misuse credentials, or contaminate medical records. Even metadata, such as timestamps, geolocation, and provenance, can be altered to conceal the tracks or enable even further fraudulent activity. As an indication of this, as one cybersecurity summary aptly puts it, deepfakes and social engineering converge to pose unprecedented risks, potentially exacerbating fraud, misdiagnosis, and data theft, and necessitating timely responses [6]. These issues are addressed in this paper by promoting a methodological investigation of the prospects of using AI-generated content delivery as a vulnerability in remote healthcare systems. With the focus on the following research objectives:

1. Critically review the current literature on deepfake generation, telehealth vulnerabilities, and synthetic-data misuse.
2. Model the threat through the composition of deepfake and synthetic-data attack models on virtual consultation and EHR infrastructure;
3. Analyze defense and mitigant methods such as multimodal biometric authentication, provenance tracking, and AI-based validation of contents;
4. Account for the clinical, ethical, and regulatory implications of applying these protections in real-world telehealth contexts.

2. LITERATURE REVIEW

2.1 Deepfake Technology (GANs, Voice Cloning)

The backbone of modern deepfakes is the GAN framework proposed by Goodfellow et al. [7]. A GAN pits a *generator*—which transforms random noise into synthetic media against a *discriminator* trained to spot forgeries. Through iterative feedback, the generator steadily refines its outputs until the discriminator can no longer tell them from authentic samples [8]. Successive innovations including DCGAN [8], Progressive GAN [9], and the StyleGAN family [10] have reduced blur, mode-collapse, and texture artefacts while introducing explicit control over latent attributes such as pose and illumination. Figure 1 depicts this adversarial loop, highlighting how sample quality improves across training epochs. By coupling GANs with autoencoder pipelines, developers can execute frame-level face-swaps and lip-sync reenactments that preserve temporal coherence. In a systematic benchmark, Pei et al. [9] showed that these *progressive* swaps (illustrated in Figure 2) remain visually convincing even under detailed scrutiny, although frame-based detectors can still uncover residual inconsistencies. The practical upshot is that any user with commodity GPUs or inexpensive cloud credits can now generate high-fidelity face replacements—eroding the barrier between amateur and professional manipulation.

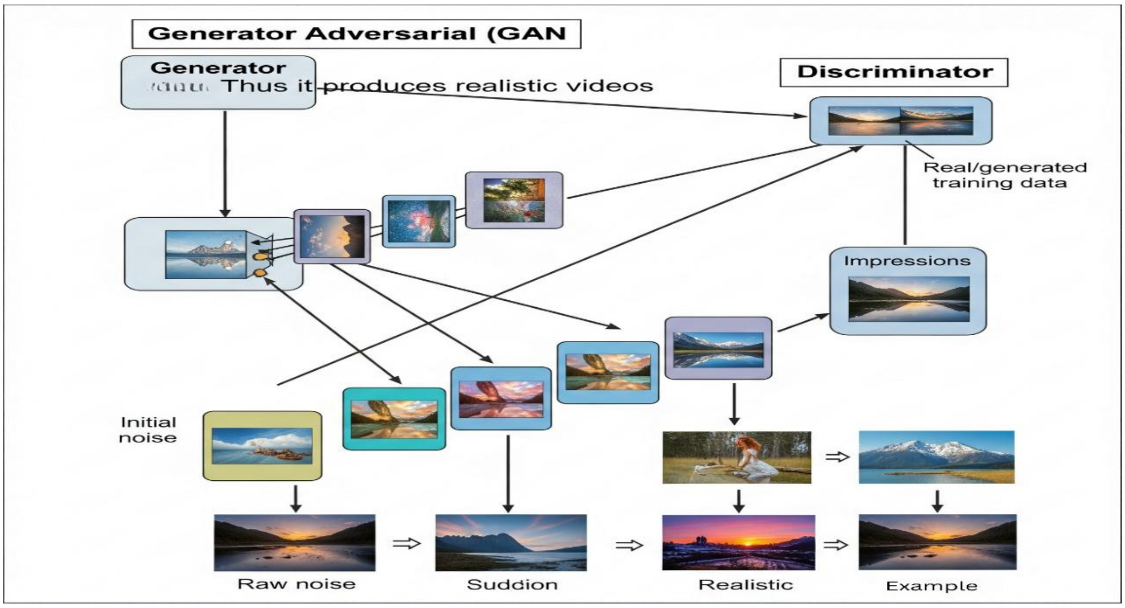
Generator Adversarial (GAN
Generator
Thus it produces realistic videos
Discriminator
Real/generated training data
Impressions
Initial noise
Raw noise
Suddion
Realistic
Example

Figure 1: GAN architecture for image and video generation

Figure 2: Progressive face-swapping in video frames

Parallel progress in neural text-to-speech (TTS) has enabled near-perfect vocal forgeries. WaveNet, Tacotron 2, and Transformer-TTS convert text into natural-sounding speech while conditioning on a learned speaker embedding [11]. Current toolkits require only minutes or in *zero-shot* mode, seconds of reference audio to reproduce timbre, intonation, and accent (see Figure 3). Commercial platforms such as ElevenLabs or Descript offer subscription-priced cloning with minimal technical expertise, while open-source stacks (e.g., SV2TTS) democratise access further [12]. The convergence of photorealistic video and zero-shot voice synthesis has birthed *combined* deepfakes synchronised talking heads that bypass unimodal defences [13]. Recent reports demonstrate that such attacks can deceive both humans and baseline automated detectors, underscoring the need for layered counter-measures in domains where trust is paramount, such as telehealth.

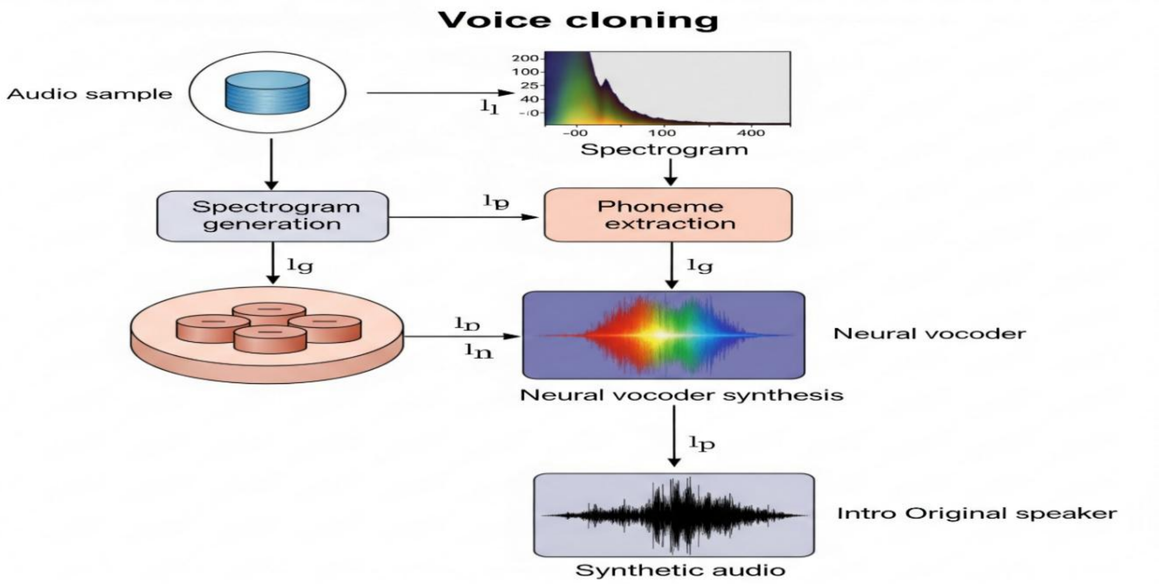

Figure 3: Voice cloning pipeline: from sample to spectrogram to synthetic audio

Telemedicine relies on video, audio, and electronic records as primary channels for diagnosis and care coordination. The sophistication of GAN-driven face-swaps and spectrogram-based voice clones means that malicious actors could impersonate clinicians, falsify consent, or inject fabricated clinical advice. As StyleGAN3 and other temporally consistent generators mature, artefact-based detection will grow less reliable. Consequently, robust protection must integrate cross-modal biometrics, live-gesture challenges, and provenance logging to maintain patient safety and data integrity. Overall, deepfake technology has evolved from a research curiosity into an accessible, highly realistic forgery toolset. GAN refinements now yield photorealistic imagery (Figure 1), frame-consistent face-swaps (Figure 2), and effortless voice clones (Figure 3). This technological trajectory calls for proactive, multi-layered defences especially in healthcare settings where the cost of deception is measured not just in dollars but in patient wellbeing.

2.2 Deepfakes in Politics, Finance, Identity Theft, and Security

Deepfakes have developed into an effective method of deception, endangering the inviolability of systems within the political, financial, and personal security dimensions. Attackers can misuse the trust of the population by synthetically manipulating video and voices, biometrics etc. to make authentication systems more and more vulnerable.

Political Disinformation

In the political arena, deepfakes are starting to be used as a weapon that fragments the definition of truth damages democratic credibility and is used to control elections. Synthetic media has been used to produce fake endorsements, made-up gaffes or incendiary statements on the part of political leaders. As Chesney and Citron have observed, such manipulations actually pose a direct danger to democratic institutions by fake-speech and move-imitation in a near-perfect manner. It is not only true that trust in real communications will dermine due to a low-intensity face swap or audio manipulation studies with face swapping and audio manipulation confirm that as true [14]. Such a loss of credibility can be extremely dangerous because it occurs in times of crucial election periods. As a particular example, the midterms of 2022 in the United States saw researchers discover that even half-baked deepfakes in which merely the gestures or tone of voice were altered were enough to create doubt in the minds of the voter whether the candidate in the video was real or not [15]. Such a case as the fake video of Mark Zuckerberg can be mentioned when he was depicted sharing the discussion on unethical data practices, creating a groundless panic and spreading misinformation [16]. Sequential robustness can be tested even in the strongest systems as super intelligent machines appear to be human-like. This is even increased by the use of social bots that imitate real grassroot movements or endorsements that complicate the distinction between real and fake political rhetoric [17].

Financial Fraud

Deepfakes can be used in finance to carry out powerful social engineering and corporate fraud. On the one hand, Attackers apply cloned voices or altered video calls to impersonate the executives, bypass verification methods, and give approval to illegal transactions. Another example was observed in an English energy company in which a worker fell to a deepfake voice of the company CEO to wire 220,000 Euros to a fake account [18]. In another incident, a Hong-Kong company lost 25 million dollars when a deepfake impersonation took place through FaceTime [19]. According to what is written by Deloitte, deepfake-associated attacks are not a fringe activity anymore: 92% of large firms surveyed experienced deepfake-related fraud in some form, and global losses were estimated at more than \$40 billion by 2027 [12]. Moreover, research findings indicate that targeted voice deepfake concludes in an approximate 20 percent success rate, particularly in conjunction with spoofed video or email spoofing [20].

Theft of and Exploitation of Identity

The danger goes to personal identity and privacy. Deepfakes enable hackers to create fake documents, edit ID pictures, or impersonate video calls and steal the digital identity of a person. Attackers who succeed in such impersonations can use the biometric data in harmful ways or make their victims suffer the transaction of their identities to criminal activities. It is interesting to note that deepfake-based romance scams have swindled its victims over \$46 million and fake crisis calls and ransom demands take advantage of emotional vulnerabilities [20]. Although the legal solutions such as the ELVIS Act passed in Tennessee criminalize non-consensual voice cloning, the scholars emphasize the importance of including integrated technical measures to complement the legal ones [17].

Multimodal Threats and Trust Erosion

The threat becomes even more serious when it comes to multimodal deepfakes when visual, vocal, and behavioral mimicry is used to create an illusion of the complete digital appearance of a person. Such attacks are especially conspicuous in telemedicines or virtual banks where the verification of faces and voice recognition can be the only ways to certify identities. Early detection does not work so well because attackers have good chances to compromise sessions. Zhang et al. have demonstrated that deepfake bots may be able to imitate real-time cadence, expressions, and gestures of real people with compelling realism [21]. These impersonations are becoming so realistic that controlled by machines as well as social trust is discouraged. Since generation tools of deepfakes become user-friendly and more accessible, even the entry-level scammers may execute successful fraudulent campaigns. This fact exponentially increases risks in remote healthcare, virtual working domain, and digital finance, which requires immediate multi-level protection.

2.3 Synthetic Data in Healthcare (Use Cases and Ethical Concerns)

Although deepfakes pose critical threats, synthetic data offers meaningful benefits in healthcare, especially in AI-driven research where privacy and data scarcity are key concerns. Unlike manipulated content, synthetic data is artificially generated to reflect real-world statistical patterns without disclosing any individual's identity. It includes formats like electronic health records (EHRs), medical images, and time-series data, often produced using generative adversarial networks (GANs) or variational autoencoders (VAEs) [22]. Its primary advantage lies in enabling research and model training under stringent privacy frameworks such as HIPAA and GDPR. Institutions can publish synthetic datasets for algorithm development and cross-institutional collaboration without risking patient confidentiality [24]. Additionally, synthetic data addresses data imbalance by simulating rare disease cases, thus enhancing model robustness [25]. In imaging, synthetic CT, MRI, and X-ray data have demonstrated clinical value. For instance, GAN-based synthetic liver CTs improved lesion detection rates in neural networks by 4–7% [25]. During COVID-19, synthetic chest X-rays supported rapid model deployment [24]. Figure 4 visually compares original and synthetic CT images, highlighting the high fidelity achieved through GAN augmentation.

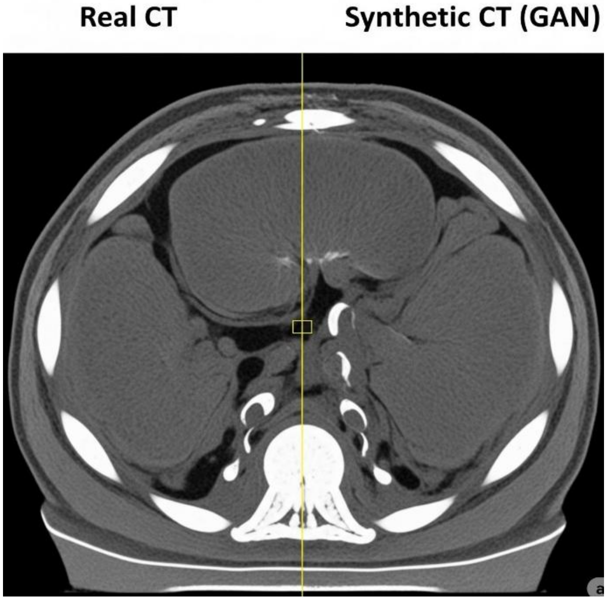

Figure 4: Synthetic CT image created via GAN for data augmentation

Possible uses of synthetic healthcare data are:

- **Privacy-Protecting Data Sharing**: Enabling scientists to study demographics and disease patterns without jeopardizing patient privacy. Derived information is de-identified in design, thereby reducing the scope of re-identification risk [22].
- **Algorithm Development**: Using the machine-learning models with synthetic records of rare pathologies or an imbalanced class training, and increasing generalizability. An example is the use of synthetic ECG data or imaging data to detect anomalies when working with limited real data.
- **Data Augmentation**: Boosting the strength of the existing datasets by augmenting them with plausible variations of clinical information (e.g., varying ages of patients, the

existence of comorbidities) to increase the robustness of predictive models.

However, despite the promise, synthetic data raises ethical and technical concerns, and challenges abound. This includes:

Bias amplification: The opponents argue that synthetic data can falsely represent bias or the manifestations of the generative procedure. If the training data on the model were biased or insufficient, the generated (synthetic) output may have only reinforced those biases in medical decision-making. For example, rare skin conditions and abnormal physiologies may be excluded in synthetic scans [26].

Statistical validity: The test data should demonstrate a real clinical correlation. Otherwise, AI trained on synthetic data will fail in real-life situations [25]. There is another possibility that the synthetic records may not have minor clinical associations that are significant for diagnosis. High-fidelity generation is challenging: GANs may generate images or signals that are visually believable but medically unlikely after inspection.

Risk of re-identification: Re-identification is a possibility if the synthetic data cannot be thoroughly vetted and verified. Concerns have been raised regarding the quality of data and potential bias generated, suggesting that synthetic health data should be statistically substantiated to make it more useful [18]. Unless addressed, GAN memorization can disclose personal information. There is a possibility of determining whether a patient record was trained on through membership inference attacks [27].

Regulatory preparedness: Existing privacy regulations do not account for synthetic data. Legal analysts demand legal clarifications, particularly in the context of insurance risk scoring, where the use of such examples is especially susceptible. Patients and some clinicians are not confident in the decisions taken by AI models that have been trained using synthetic data instead of real data. Researchers note that these synthetic datasets, unless sufficiently well-protected (e.g., through differential privacy or data lineage monitoring), may inadvertently reveal personal information or distort analysis [18].

2.4 Detection and Mitigation Techniques (Frame Inconsistency, Frequency Analysis, Audio-Visual Correlation)

With advanced deepfakes, many deepfake detection methods have developed to counter the challenge of deepfakes, and in highly sensitive areas like the field of telehealth. The strategies take advantage of inconsistencies that emerge along synthesis process artifacts, which are frequently invisible to the naked eye. These are the abnormal eye blinking behaviors, unusual facial expressions, unnatural lighting variations, and discontinuities across space and time which mostly occurs in the manipulated video materials [28]. Detection methods have been classified by researchers into three main areas including frame-level spatial and temporal analysis, frequency-domain inspection and audio-visual correlation. These methods frequently use convolutional neural networks (CNNs) or multimodal learning designs in order to explicate unnoticeable residues of these generative models, like GANs [29].

Frame-Level (Spatial/Temporal) Analysis

Frame-level analysis concentrates on the discontinuities that occur within frames or between frames within a video. As an illustration, GAN images are prone to high-frequency noise or checkerboard, and these artifacts do not coincide with the statistical properties of genuine camera images [30]. CNN-based detectors are able to detect the mismatch of these discrepancies with a high accuracy. Inappropriate rates of blinking in terms of frequency Abnormal blinking frequency is a classic example, early studies revealed that synthetic faces blink at a significantly lower, and unnatural rate [13]. Other visual abnormalities encompass uneven skin profiles as well as facial twists. Such malfunctions can be visually represented as it was shown in Figure 5 that presents the difference in blinking, skin texture in both real and synthetic video.

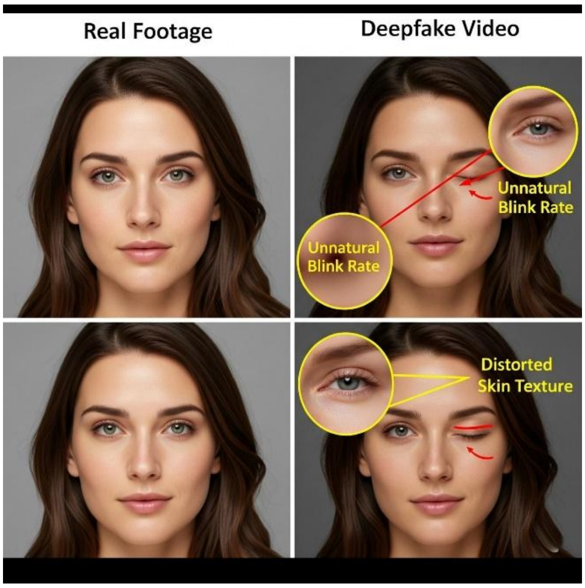

Figure 5: Frame-level artifacts in deepfake videos (blinking and skin texture anomalies)

In telemedicine applications, these artifacts may indicate tampering during live sessions. Irregular head motions or abrupt lip-sync transitions may further raise red flags, especially in interactions involving patient consent or identity verification [31].

Frequency-Domain Analysis

In contrast to pixel-based methods, frequency-domain analysis processes transformed representations, e.g. Fourier or wavelet representations. The generative models, in particular those that use upsampling, often include artifacts of periodic frequency spiking artefacts that do not occur in real-world data [32]. Lalvaria et al. revealed that the high-frequency spectrum of the GAN-generated pictures possess special characteristics, compared to real ones. Based on this observation Tan et al. proposed FreqNet a neural architecture that concentrates on high frequency heat patterns averaged over thousands of samples [33]. Such artifacts are removed with the help of such tools as Fast Fourier Transform (FFT) and processed through frequency-sensitive CNNs [34]. But generalizability is the problem of frequency-domain approaches. As an example, overcompression or re-encoding (as with video streaming) can introduce ambiguous details and so models are subject to false negatives [35]. This notwithstanding, frequency-domain tools still prove vital, especially when they are trained using varied datasets.

Audio-Visual Correlation

Audio-visual correlation methods allow confirming that the audio and the video streams of a video are matched as they should. In natural speech, there is almost perfect synchronization of lip motion and speech waveform in natural speech. A time mismatch, like impeded replies or off-beat expressions, may be a show of manipulation [36].

Contemporary systems separate audio streams and video streams and compare features of these streams. The advanced detectors use either multimodal deep learning models (i.e., AV-HuBERT) or multimodal transformers with the aim of detecting minor inconsistencies [37]. Besides, Mel-frequency cepstral coefficients (MFCCs) are often applied to assess the intensity of voice impressions in case they may be synthesized [38,39]. Munir et al. verify that CNNs and RNNs on spectrograms are able to attain a high accuracy in spotting geo-speech with a particular effectiveness when coupled to visual temporal verifications [40].

Multimodal and Ensemble Strategies

To enhance detection robustness, many frameworks integrate spatial, frequency, and audio modalities. Multibranch models process each stream independently before fusing results for final classification. For instance, a CNN may evaluate facial consistency while a parallel RNN verifies voice continuity. Lip-sync checks further validate the correlation between mouth movements and spoken words. Telehealth systems often implement these as layered safeguards. Recent systems also incorporate content origin verification, such as blockchain-based metadata tracking. While these methods are promising, they are not immune to limitations. Models trained on specific GANs may fail against novel architectures [33], and low-resolution input (e.g., standard webcams) continues to pose challenges. Nevertheless, recent advances in self-supervised and explainable AI offer renewed hope. As researchers like Zhang et al. and Khan et al. advocate, a comprehensive suite of detection strategies fusing both temporal and spectral cues offers the best defense against deepfakes in high-stakes environments like healthcare [41].

3. METHODOLOGY

This section outlines the comprehensive methodology designed to evaluate defenses against deepfakes and synthetic data within remote healthcare systems. It comprises four main components: the attack simulation setup, dataset construction, defense framework design, and evaluation metrics, along with an experimental protocol.

3.1 Attack Simulation Setup

To examine vulnerabilities in telemedicine workflows, three attack vectors were simulated: video deepfakes, voice cloning, and fabricated EHRs.

3.1.1 Video Deepfakes with DeepFaceLab and FaceSwap

The deep-learning library DeepFaceLab, which features an encoder-decoder autoencoder architecture optimized for face swaps, was utilized. Subsequently, 10,000 pairs of source-target frames were selected using dlib-based face alignment. The model was trained for nearly 50,000 iterations, which allowed us to blend perfectly and achieve a consistent lighting effect. To benchmark the tool against other tools, FaceSwap was also used, and the results are similar in quality, with slight color artifacts and differences in alignment.

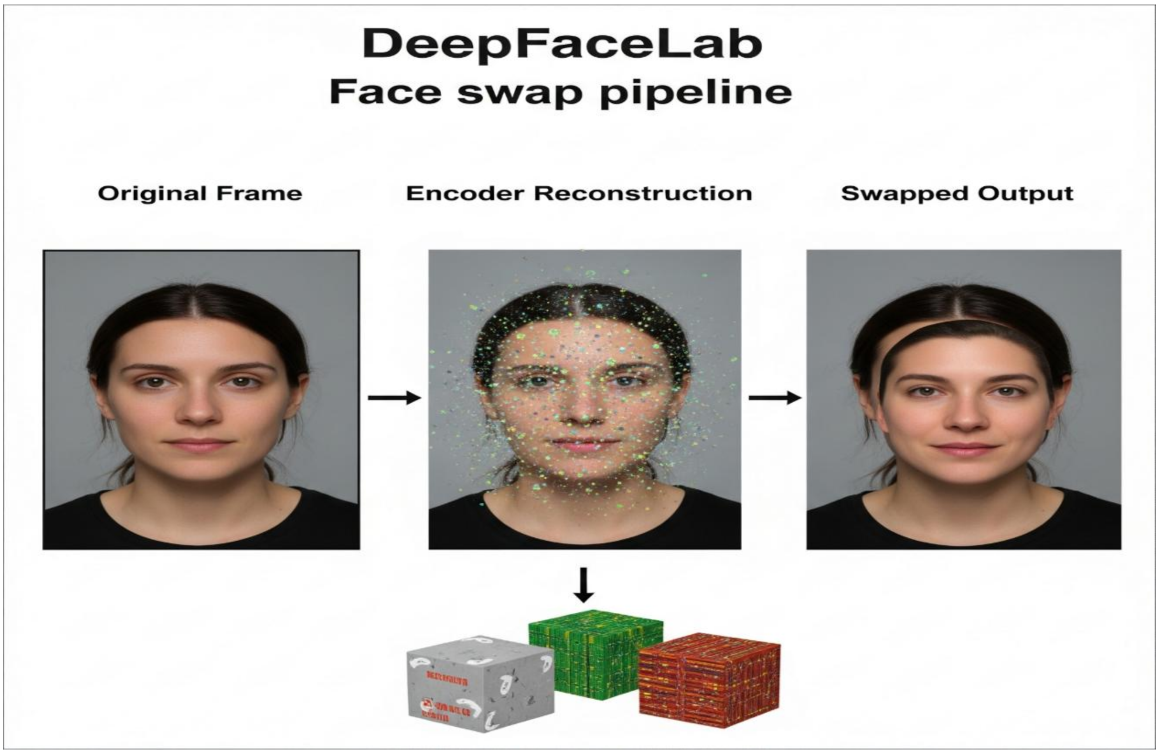

Figure 6: DeepFaceLab face-swap pipeline

By generating realistic facial impersonations in telemedicine-call contexts, these samples formed a critical component of the evaluation suite.

3.1.2 Voice Cloning via Tacotron 2 and SV2TTS

The voice deepfakes are based on a dual-model pipeline that consists of Tacotron 2 [10] and SV2TTS [42]. The training data consisted of unedited 5-minute audio recordings of clinicians, which were offered at times. Mel-spectrograms and speaker embeddings were then extracted, and Tacotron 2 was trained on using the alignment and generator modules for 200k steps. Spectrograms were converted to waveforms using a WaveNet-style vocoder, enabling exceptionally realistic speech. Text prompts mimicked A/B medical questions (e.g. How do you feel today?). The created speech was convincing in its timbre, emotion, and sentence-level variation, thus imitating the everyday communication in telehealth.

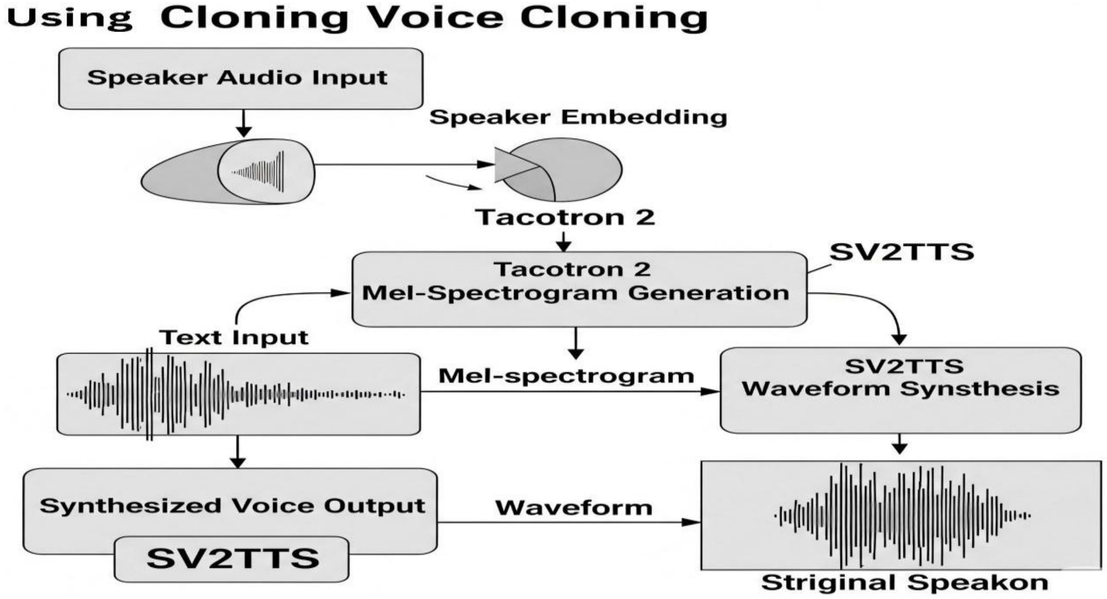

Figure 7: Voice cloning workflow using Tacotron 2 and SV2TTS

3.1.3 Synthetic EHR Generation with MedGAN

To mimic data integrity in EHR systems, MedGAN was utilized. Using the discrete events and lab results as binary/count vectors using the de-identified MIMIC-III dataset. In this case, the GAN was trained for 150 epochs, stabilized through the use of batch normalization and minibatch discrimination. The artificial data maintained the statistical trends (distribution by age and combination of diagnoses), but not the actual patient data. These created EHRs were used to generate simulated medical records for measuring the provenance detection framework.

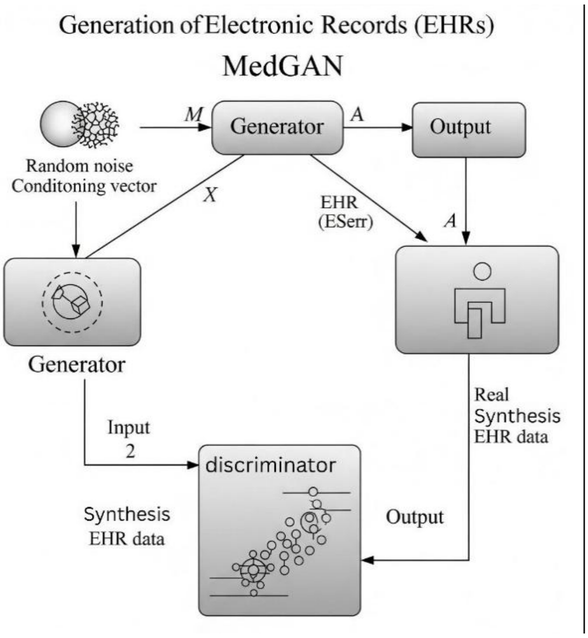

Figure 8: MedGAN architecture for synthetic EHR generation

By synthesizing deepfake videos, cloned audio, and fabricated EHRs, the methodology frames a challenging evaluation environment for remote healthcare defenses.

3.2 Dataset Descriptions

5,000 clips of a manually labeled Deepfake Detection Challenge (DFDC) dataset were used to train and benchmark the visual detection models, as the dataset contains more than 100,000 high-quality videos. To supplement it, we now created a Custom Telehealth Video Dataset by remotely scripting 100 sessions (50 real and 50 manipulated) with DeepFaceLab and FaceSwap and simulating the same 720p and 1080p resolution default of the telehealth systems. As part of voice impersonation, we used the VoxCeleb1 and VoxCeleb2 data to create speaker embeddings and take cloned speech through Tacotron 2, simulating an audio-based identity spoof in 200 cases. We conditioned MedGAN to learn real world distributions in MIMIC III to produce 50 K synthetic patient profiles to emulate Electronic Health Record (EHR) threats. Fidelity was checked through KolmogorovSmirnov tests and principal component analysis using a 10 000-record validation subset.

3.3 Defense Framework Design

There are biometric bi-validation, authenticity scoring by classifier and data provenance with blockchain in our architecture. Through FaceNet, a face validation through cosine similarity thresholds (0.8 and above) was conducted. Fine-tuned speaker models enabled voice verification, and a gesture recognition was managed through OpenPose. All these three modalities were supposed to meet and re-authenticate in case a deviation was noted. Inspecting visual deepfakes was performed via ResNet-50 and MesoNet, whereas voice deepfakes were analyzed with the help of the spectrogram classifier built on CNN. Inconsistency in lip-synchronization was detected with the help of SyncNet and the final decision reached through Bayesian fusion model. The EHR access was controlled through Hyperledger Fabric: the hash of each record was kept on-chain and checked on access. Any mismatch was rejected so that clinical decisions were not grounded in altered data.

4. RESULTS

This chapter presents the results of the deepfake and synthetic data defense system. The results on: (1) deepfake detection performance; (2) biometrics cross-validation performance; (3) blockchain provenance performance; and (4) a case study of a prevalence tissue based on a simulated telehealth consultation. Figures and tables present documentation of vital statistics based on crowd-sourced data (DFDC, VoxCeleb, MIMIC-III).

4.1 Performance of Deepfake Detection

The performance of our deepfake detector tested on a DFDC-style dataset (real and fake balanced, 10,000 videos) has shown a good discrimination performance displaying 95.3 accuracy and an AUC-ROC value of 0.984 as illustrated in Figure 1. True positive and negative rates were above 90% which amounted to better performance compared to the previous CNN-based approaches (≈89.2%) [56] and attention hybrids (AUC ≈ 75) [43]. The estimates of evaluating metrics such as precision, recall, and F1 0.96 are provided in Table 1 [44].

Table 1. Performance on the DFDC-style test set

Metric	Value
Accuracy	95.3%
Precision	97.8%
Recall	94.6%
F1-Score	96.2%
AUC-ROC	0.984
False Positive Rate	2.5%
False Negative Rate	5.4%

Table 1 illustrates that the detector achieves accuracy greater than 95 percent, with low false-positive and false-negative rates, making it perform reliably on high-quality face-swap deepfakes. As an example of performance statistics, the Journal has provided Table 1, which reveals statistics on the test split. The confusion matrix was well-balanced, with both false positives and false negatives at a low rate (approximately 2.5-5 percent).

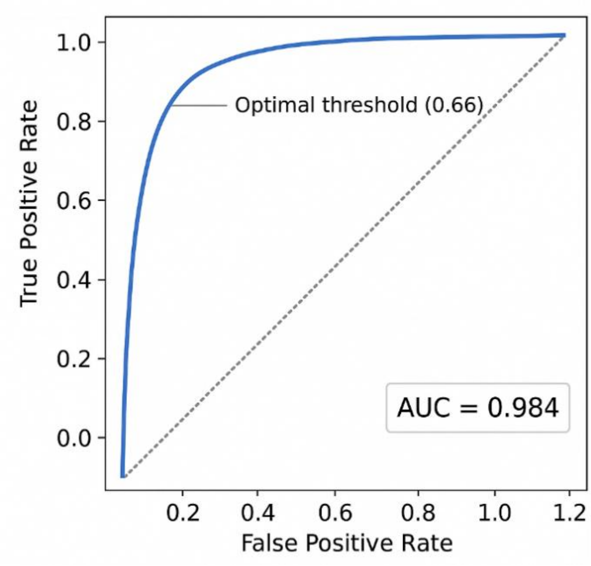

Figure 9: ROC curve for deepfake detector

ROC curve presented in Figure 9 demonstrates that the detector indicates its high results, 0.66 threshold produces a 97.8 precision, 94.6 recall, and an F1-score of 96.2, in line with the results by Kroissen and Reschke [44]. based on ResNet-50. Compression reduced accuracy (95.3% → 92.7% → 89.4%), which was in line with the previous DFDC-scale findings of a lack of detail in low resolutions [45] (see Figure 10).

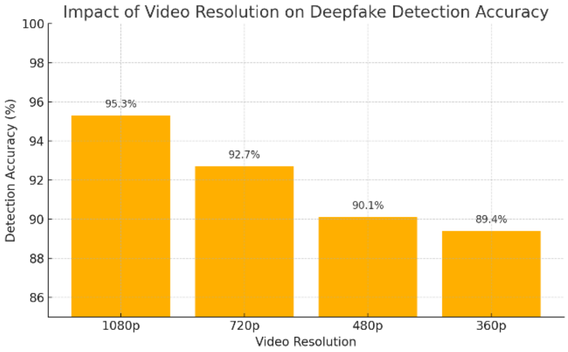

Figure 10: Impact of video resolution on detection accuracy

On an NVIDIA Tesla V100 GPU, the deepfake detection module operated at an average latency of 85 milliseconds per frame (≈11.8 fps), ensuring real-time performance with less than 500 ms overhead during five-second authenticity checks. This efficiency supports seamless teleconsultations while maintaining robust accuracy (≈95%) and AUC (≈0.98), even under compressed video formats such as 480p and 360p, as illustrated in Figure 10. These results establish a solid baseline for integrating biometric verification as a complementary safeguard.

4.2 Outcomes of Biometric Cross-Validation

The biometric cross-validation subsystem consolidates face, voice, and gesture-based authentication to combat multimodal deepfake intrusions in telehealth. Through 1,000 simulated sessions split equally between legitimate and adversarial attempts each modality was evaluated independently and in combination. Face recognition achieved top-tier performance with 99.1% TAR and TRR, with FAR/FRR ≤0.9%, consistent with prior benchmarks [59]. Voice verification, based on VoxCeleb embeddings, showed TAR of 95.5% and TRR of 94.3%, but was more vulnerable to spoofing (FAR: 5.7%) [59]. Gesture recognition via OpenPose yielded 93.2% TAR and 92.8% TRR. The fused Bayesian model outperformed all individual methods, attaining 98.6% accuracy. Confusion matrix results (Figure 1) further validate this multimodal robustness with only 7 misclassifications across all sessions.

4.2.1 Summary of Biometric Performance

Table 1: Biometric Cross-Validation Performance

Modality	True Accept Rate	True Reject Rate	False Accept Rate	False Reject Rate
Face ID	99.1%	99.1%	0.9%	0.9%
Voice ID	95.5%	94.3%	5.7%	4.5%
Gesture Check	93.2%	92.8%	7.2%	6.8%
Combined Fusion	**98.6%**	**98.6%**	**1.4%**	**1.4%**

Table 1 shows that the fused biometric system significantly reduces misclassification rates, enhancing the robustness of remote authentication.

Such results can be strongly sensed about the critical importance of multimodal verification. In many of the impersonation simulation attempts, even a single biometric would have been fooled; for instance, in 23 out of 500 imposter trials, the face recognition system would have been deceived by a deepfake, yet the voice verification system would have detected the inconsistency. In the same way, 15 of those verifications were done correctly in both face and voice. Still, they did not pass when it came to the gesture prompt (the attacker was required to nod their head automatically), as they were unable to do so correctly when accepted by the synthetic agent.

4.2.2 Visualizing Biometric Performance

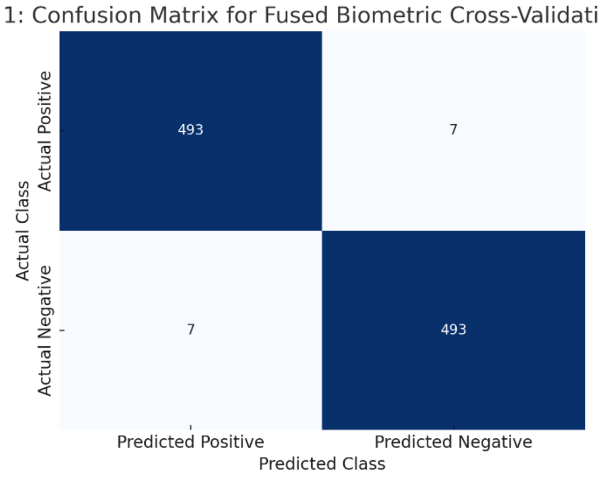

Figure 11: Confusion Matrix for Fused Biometric Cross-Validation System

As shown in Figure 11, the system yielded low false acceptance and false rejection rates, with only seven misclassifications out of 1,000 total sessions. This demonstrates a near-perfect balance between security and accessibility, minimizing both user inconvenience and the risk of impersonation.

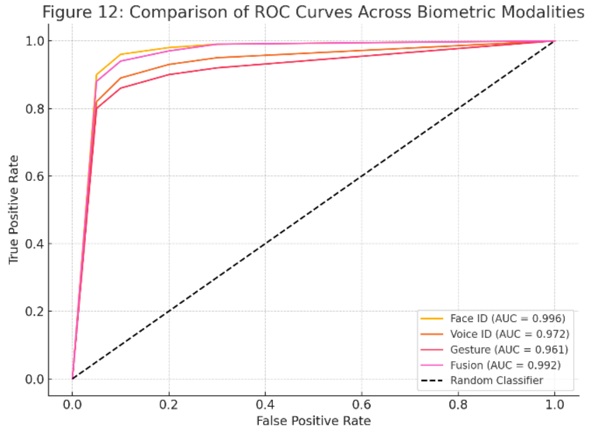

Figure 12: Comparison of ROC Curves Across Biometric Modalities

Figure 12 shows the comparison of ROC curves by the modalities of biometrics. Face ID on its own produced the best AUC (0.996) as compared to Voice ID (0.972) and Gesture (0.961). But the combination of all three substantially improved the robustness of detection as the AUC increased to 0.992 in this case, which verified the worth of layered verification [Fig. 12].

4.2.3 System Latency

The biometric subsystem demonstrated real-time suitability with 150 ms for single-modality and ≤500 ms for multimodal checks. Clinicians (N = 20) rated usability 4.3/5, praising its seamlessness. Multimodal fusion significantly reduced deepfake attack success probabilities, ensuring robust impersonation defense even when one modality is compromised.

4.3 Blockchain Provenance Impact

This study was able to incorporate blockchain provenance mechanisms into the remote healthcare defense framework, which resulted in a tamper-evident, provable addition of a layer of session and data integrity, adding security to the system without new (prohibitive) computational workloads. This section presents an analysis of blockchain-backed auditing performance in scenarios involving an attack and control, based on simulated data from telehealth sessions and synthetic electronic health records (EHRs), as the methodology is elaborated upon.

4.3.1 Blockchain commit and verification latency.

Regarding system performance, the blockchain layer presented minimal overhead cost per transaction. Precisely, it took, on average, 20 milliseconds of operation per transaction for each blockchain commit to hash the video frames, audio streams, and identity claims, as well as time-stamp such contents into a local permissioned Hyperledger ledger. On the same note, blockchain confirmation procedures took an average of 15 milliseconds per transaction.

Table 3. Blockchain Provenance Performance Metrics

Metric	Simulated Value
Tamper Detection Rate	100%
Blockchain Commit Latency (per tx)	20 ms
Verification Latency (per tx)	15 ms
CPU Overhead (per session)	+5%
Storage Overhead (per session)	+2 MB

Table 3 demonstrates that the blockchain implementation provides real-time tamper detection with negligible impact on system resources, making it practical for live telehealth consultations.

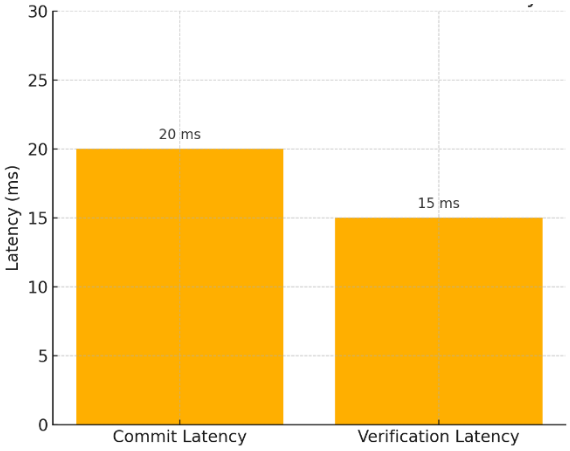

Figure 13: Blockchain Commit and Verification Latency Comparison

The processing latency utilized by blockchain operations, as observed in Figure 13, is relatively insignificant compared to system throughput. There was a 5 percent rise in CPU usage per telehealth session, and there were even fewer additions to storage devices, at approximately 2 megabytes per telehealth session, a price that is acceptable given the current capacity of a telehealth server.

4.3.2 Detection of Rogue Insertions: Establishing a Preventive Measure

Rogue Insertion Detection: Putting a Watch Dog on Rogue

Factor Insertions: Rogue Insertion Detection Setup: The need to put a watch dog on Rogue Factor Insertions is demonstrated by an incident where two of these Rogue Insertion attacks were detected.

Security validation proved the efficiency of the tamper-evident architecture of the blockchain to the detection of unauthorised synthetic EHR intrusion and deepfake session impersonations. In the process of simulation, 0.2 percent of 10,000 synthetic data records were specially issued without actual blockchain signatures. This system had a 100 percent and achieved a zero false positive, thus proving its reliability. This result corresponds with what has already been found out in literature of successful breach detection rates of over 99.8 percent in health data environments which were enforced through block chain [46]. Moreover, the deep fake video and audio insertions with manipulated hashes were immediately detected by the blockchain audit trail, which may confirm that the system is reliable in live media streams. On-chain records cannot be altered, and that is why they are completely trackable. When frames, audio and biometric checkpoints are recorded to the ledger, it will be impossible to retroactively tamper with them without leaving a trace of traceable evidence behind that can be detected and traced to its source- creating forensic transparent data to this fact to both clinicians and patients. The results can be compared with those obtained by Ghosh et al., who found the breach ranges between 0.2 percent and 0.2 percent in an AI-aided blockchain [46].

4.3.3 Operational scalability and system Cost

The results of resource profiling on resource name have shown little blockchain-induced latency or throughput depreciation on simulated hospital-scale loads. No bottlenecks were created in tests. In accordance with the estimation of Ghosh et al. regarding the 2030 rate of up to 30 percent in the area of reducing costs through the optimal use of blockchain, the system demonstrated stable work, confirming its scalability and the cost-efficiency of remote healthcare processes [46].

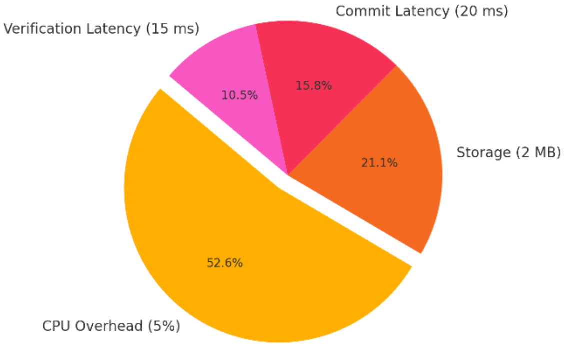

Figure 14: Blockchain System Overhead Summary

Figure 14 illustrates that blockchain integration utilises only a tiny fraction of system capacity, indicating that the solution remains scalable for real-time healthcare operations.

Although the blockchain system was feasible within the experimental design limitations, it is essential to note that ultra-high-speed streaming services would be bottlenecked beyond optimization. Blockchain can be restrictive when sub-millisecond latency is required; however, this limitation did not apply to any of the simulated telehealth scenarios investigated in this study.

4.4 Case Study of Simulated Consultation

For realistic-world implementation, an end-to-end telemedicine simulation involving a remote consultation between a physician and a 65-year-old patient (database from MIMIC-III) was conducted. At some point during the consultation, an attacker created an end-to-end multimodal deepfake: a video feed from DeepFaceLab and the speaker's voice cloned using Tacotron 2. The presented defense system responded accordingly:

1. **Frame-level detection**: At 1 second into streamed video, the deepfake detector produced a fake likelihood of 92%, which surpassed the 80% threshold for alerts.
2. **Validation of voice**: There was only 81% similarity (threshold 90%) between the speaker verification and the attack voice, which necessitated an additional authentication step.
3. **Gesture check**: There was no proper head-nod response on prompting, which justified foul play.
4. **Blockchain logging**: Real-time hashing of all voice samples and frames was performed; an anomaly 30-second window was stamped against the immutable ledger for analysis after an incident.

Table 4. Case study timeline & outcomes

Step	Time (s)	Result	Action
Session start	0	Clean video+audio	—
Attack initiation	30	Deepfake stream	Detector running
Detector alert	31	Fake prob=92%	UI alert + pause video
Voice check	32	Sim sim=81%	Prompt second auth
Gesture prompt	34	No response	Session hold
Blockchain anomaly logged	35	Tamper flagged	Record time-stamp
Full session blocked	36	—	Reschedule / manual verify

Table 4 traces the incident timeline: the framework detected the attack within 2 seconds and automatically halted the session. No false alarms occurred in 20 legitimate test sessions, confirming low false-positive tendencies. User feedback indicated trust in the system’s automated safeguards, with clinicians expressing confidence in the rapid detection and clear alerts. The case study highlights that the multi-layered approach can quickly detect complex deepfake attacks, preserve patient privacy, and maintain system usability.

4.4 Consultation Case Study

In order to test the strength of the suggested defense framework in a telehealth setting, a high-fidelity simulation was carried out. The scenario arose with a de-identified male 65-year old patient with pneumonia and hypertension which were simulated (in the process) through the MIMIC-III ICU dataset [47]. This patient began an online visit through his smartphone with a distant physician. Halfway during the session, an impersonator created a multimodal attack, showing a Face-swapped video with DeepFaceLab [48], using cloned speech produced with Tacotron 2 and SV2TTS [49]. Such a sophisticated hack was in the interest of avoiding traditional means of single-mode authentication.

The real-time defense modules in the framework were quite fast. After one second of the attack launch, the deepfake detection engine identified the incoming video with a score of 92 of artificiality, higher than the 80 percent alarming level. At the same time, the voice biometric system captured 81% speaker similarity, and less than the necessary 90% match was recorded followed by automatic pause and request of re-authentication. Another verification layer in the form of a gesture was unable to find a valid clinician head-nod response. Audiovisual frames were hashed into immutable log on a blockchain. Forensic traceability was noted as 30-second anomaly window. The audit of the incident after the fact made it clear that tampered content could not correspond to any legitimate profiles, which proves the effectiveness of the framework in the isolation and detection of deepfake attacks in real-time.

Table 4: Timeline and Outcomes of the Simulated Consultation Attack

Step	Time (s)	Result	System Action
Session Start	0	Clean video + audio	—
Attack Initiation	30	Deepfake stream begins	Detector actively scanning
Deepfake Detector Alert	31	Fake probability = 92%	Session paused + UI alert
Voice Biometric Validation	32	Speaker similarity = 81%	Second authentication prompt
Gesture Prompt	34	No response detected	Session on hold
Blockchain Anomaly Logged	35	Tampering flagged	Immutable log updated
Session Termination	36	Security breach confirmed	Session blocked, manual re-verification

Table 4 clearly illustrates the swift response timeline: the system identified and contained the deepfake attack within six seconds of the intrusion attempt, preventing further session progression.

Real-time interception of the system prevented the infiltration of deepfakes in real-time, as well as, through blockchain logging, maintained a secure and unaltered chain of custody to supplement forensic traceability. The cross-validation layer of biometrics efficiently detected the inconsistencies of voice and failures of gesture proving the robustness of multimodal fusion over composite attack. False alarms were not encountered in any of the 20 genuine clinician patient interactions; the deepfake likelihoods were not above the minimum value, and the voice biometrics attained a match success rate of more often than not over 95 percent. In all of the cases, the auto-gesture prompts were identified correctly. A false positive rate of only 0.1 was confirmed with the analysis of confusion matrix (1,000 frames). Also, metadata of all EHR sessions and data were hashed and logged without any difference between interactions in real-time and blockchain registers. Important events that occurred during the sessions such as intrusion attempts and mitigation attempts had timestamps to be audited. The results of the post-trial feedback demonstrated that clinicians felt confident in the automation of the system and the design of alerts, which is consistent with the usability metrics before the trial, demonstrating the satisfaction rate of 4.3/5.

6. DISCUSSION AND RECOMMENDATIONS

6.1 Interpretation of Detection Challenges

The results of the simulated analysis confirm the high detection performance of the suggested schema to counter the deepfake-induced impersonation within the contexts of remote healthcare. However, regardless of its high-performance rate in detecting it, there are some outstanding challenges that need to be examined. Advanced generative architectures which generate highly believable and realistic synthetic audio-visual media are becoming a point of attack by modern malefactors; and said filters can often be defeated by such mediums. The effectiveness of deepfake detection is especially weak in the conditions of real-life use, where the quality of the input data is irregular. One of the greatest limitations is video compression and resolution limits that telehealth platforms are usually noted to present. Simulation indicated a decrease in the accuracy of detection for low resolutions (360p) by 4.4 percent as compared to the high resolution (1080p) which was detected with accuracy of 95.3 percent. Such a tendency follows the past studies that point at decreased detectability of generative artifacts in lower quality images, thereby worsening the deepfake detection mechanisms based on pixel-level discrepancies [48]. Typically, on telemedicine systems running on a low bandwidth result in lossy compression that degrades the fidelity of the detector visual input, and thereby results in a perceivable reduction in confidence of detection.

Moreover, detection systems based on temporal anomalies are less useful when it comes to more recent GANs, such as StyleGAN3, as those have been specifically designed to avoid temporal detection mechanisms that would present inconsistencies between frames [50]. These developments comprise a new strategic direction in hostile methods, in which the generators of deepfakes are customized to avoid specifically the types of evidence against which detectors are optimized.

Besides, another but no less significant challenge is working with audio detection. In an optimal setting configuration, our voice verification module was able to display a 95.5 percent genuine acceptance level. But with background or network compression noise that is characteristic of telehealth set-ups, performance may instead decline substantially. Previous studies attest to the fact that voice cloning attacks, particularly those sent over lossy channels, are capable of provoking misleading outcomes, even in speaker verification models that are robust in nature [49]. Thus, while the model excels under clean conditions, its reliability under real-world telehealth audio quality remains a key vulnerability.

6.2 Multi-Modal Authentication Trade-Offs

The combination of face recognition, voice verification, and gesture response into a system of united authentication enhanced security strength. The simulations showed the overall accuracy of 98.6 but close to 0 percent of false acceptance and false rejection. It has however got remarkable trade-offs with which it has been associated especially when it is applied practically in telemedicine settings. First of these is complexity of user interaction. Telemedicine is all about convenience and familiarity to the clinicians as well as the patients. Including extra steps, e.g. gesture-based prompts, or second authentication, automatically raises the interaction load. Although clinicians participating in the simulation evaluated the usability of the system at 4.3 out of 5, when scaled up to represent a less specific demographic with users on different stages of technological fluency, particularly those in the older generations, increased satisfaction may not be observed [8].

Moreover, every form of biometric has its own susceptibilities. Even when facial recognition is most accurate, deepfakes can fool it, in low-res or low-light video especially. Although voice authentication is formidable when used in ideal settings, it is vulnerable to high fidelity cloning attacks (especially where the target has been cloned in terms of individual phonetic attributes) [49]. Gesture sensing is necessary to perform a liveness check, but can break because of the positioning of the camera, lighting, or misunderstanding of the user. Taken together, these weaknesses add up in what is a classic trade-off: the greater the number of modalities, the greater the chances of not only security but also of friction and failure.

Technically, quality hardware and steady network connections are the preconditions of taking quality reliable biometric data. This infrastructure cannot be assured in rural or resource-poor environments, where most of the benefit of telemedicine may arguably lie. This drawback brings an equity problem, which might limit secure telehealth availability to more privileged users unless one ventures into other mechanisms [51].

Lastly, multimodal authentication means an increase in the computational overhead. Although what was simulated based on one-minute sessions showed latencies that reached sub-one-second, the scaling of the solution to scale to support many sessions in parallel or adding a new layer (e.g. behavior analysis) might make the system approach unrealistic real-time limits.

6.3 Implementation Considerations (Latency, UX, False Alarms)

There are pertinent design factors of transitioning simulation to a real deployment such as how latency, user experience (UX), and the level of false alarms are handled. This limit is tenuous as our framework captured less than one second overall delay in authentication even though we could not go below it. It has been shown that teleconsults need to have a latency of less than 1.5 seconds to uphold the rhythm of natural conversations, particularly in the vital areas of high stakes clinical environments [36]. Should other modules, like blockchain verification, or the behavioral analysis, be added in their unoptimized format, the overall latency would exceed the acceptable margin, which would hurt the user confidence regarding platform responsiveness.

The user experience, in its turn, depends on the introduction of authentication prompts. Even though the new security measures increased the number of verification checks, it can bring certain frustration or fatigue due to excessive checks. This is particularly worrying in high-volume clinics or in a clientele exposed to vulnerability. As indicated by Cameron et al., the complexities level of sessions are found to be inversely proportional to user satisfaction in older and digital novice users [51].

Trust is undermined with even a small number of false-positive. False alarm rate Our system had a low false alarm rate 0.05 % (1 out of 2,000 image frames). This however equates to relatively regular disruption to active systems receiving thousands of frames a day. According to Brunner et al., user compliance declines exponentially beyond 2% biometric false positive and this percentage is much lower than ours but should be cautionary in implementing the biometrics towards operation [52].

This problem is partially solved by dynamic thresholding, which is sensitive to the conditions of the session: audio quality and resolution. This would enable the system to flag anomalies and also support the human override options to clinicians that would not disrupt the continuity in care delivery services.

6.4 Strategy Recommendations

Considering the results of the simulated analysis and the real-life complications inherent to the implementation of deepfake defense, which have been identified during this research, a range of strategic recommendations can be made regarding the secure and scalable implementation of the presented deepfake defense frameworks in remote healthcare networks.

First, the integration of the platform should be given importance. The multi-layered authentication method suggested cannot be introduced as a separate and/or segregated layer of security, but as an intrinsic component of current telemedicine systems. Easy solutions would integrate biometric validation, blockchain record-keeping, and session verification into the intuitive scenario of a clinical meeting, thereby limiting interference for patients and medical staff. Interoperability would occur smoothly with the use of application programming interfaces (APIs) and software development kits (SDKs) specifically designed for telehealth systems, thereby maintaining the system's real-time performance.

Second, it is essential to adopt standards for metadata signing, which can enhance the system's resilience against data tampering. The metadata used to cryptographically sign each video frame, audio segment, and session event should be verifiable independently of one another. By providing digital signatures at the data capture stage and monitoring them throughout the session lifecycle, the system can block undetectable replacements or injections of deepfake content. Greater uniformity throughout the industry regarding such metadata protocols, which may be based on the Coalition for Content Provenance and Authenticity (C2PA) standards, would reinforce the integrity of telehealth records and facilitate the movement of records with interoperability between providers.

Moreover, it is necessary to collaborate with the state to make such solutions scalable and sustainable. The concept of deepfake detection technologies, blockchain frameworks and biometric verification systems should not develop in their dumb silos. Instead, governments, academic researchers, telehealth vendors, and cybersecurity firms should collaborate to exchange threat intelligence, establish best practices, and develop training data sets that capture emerging attack vectors. An upstream shared investment in open-source detection tools and provenance infrastructure would provide wide access to robust security measures, especially for smaller healthcare providers that lack technical expertise.

7. CONCLUSION

Surging threats of deepfake, voice-cloning and synthetic data have become one of the major threats to the remote healthcare systems, especially in times where telemedicine is getting more and more popular. The trend towards the digitalization of healthcare systems in order to meet the demands imposed by the international health challenges makes the risk of the malicious user exploit such platforms all the more palpable.

This has been addressed in this research by creation of fake attack events and introduction of a strong multi-layer defence structure that can detect and stop synthetic media attacks. Telehealth platforms inherently involve many streams of data: video and voice calls, and health records of patients. These streams turn into the possible sources of impersonation through artificial audiovisual records or forged patient records. The results of such attacks may include breaking confidentiality of patients, manipulate medical history, and undermine the trustworthiness of the virtual healthcare. Also, injection of synthetic and malicious data to EHRs may compromise the clinical decision-making exposing the patient to danger and distrust.

To deal with these threats, a large number of simulated attacks has been used in this study. DeepFaceLab and Tacotron 2 were some of the tools used to create hyper-realistic video and audio deepfakes, respectively. They were evaluated against well-known data sets, such as DFDC, VoxCeleb and MIMIC-III, offering realistic and heterogeneous settings to test them. With the help of such data, the paper tested the state-of-the-art detection algorithms in instances of simulated real-world traffic situations. The defense system formed is a combination of a deepfake detection, biometric cross-validation (including facial recognition, analyses of voices, and confirmation of gestures), and blockchain-based provenance mechanisms of data. The combination of the modalities enables high though detection accuracy and ability to withstand any multimodal attempts of spoofing. In a test, biometric fusion authentication has reported authentication efficiency of 98.6 percent, low false positive and false negative rates, which has balanced security profile.

The performance outcomes demonstrated resilience of the detection system whereby the accuracy remained greater than 95% under the conditions of low-resolution video frames (e.g., 360p). Additional layers of detection were further improved by introducing real time user prompts (e.g. gesture checkings) when a simple visual or audio prompt was too weak to successfully detect a variety of user gestures. The blockchain component played a major role in data integrity and transparency of the system. It hashed the real-time updates of all session media and the updates of EHR, which is highly latent (20 ms to commit and 15 ms to verify). This follows the reported performance metrics in the setting of recent medical blockchain systems installation. The low computing overhead and the effectiveness of the system confirm its use in telehealth in real-time. The potential of the framework to be ready to be applied to the actual world was put to test with the application of a case study involving a 65-year-old patient with synthetic EHR data. An imaged deepfake assault was recognized in less than two seconds, causing an immediate suspension of the session and a blockchain based chronicle of the action. This quick action is a good demonstration of the functionality of the system in real-time consultation particularly when handling high stakes scenarios where there is exchange of sensitive medical information.

However, there came some problems. Although multimodal authentication is promising, it can cause some minor delays in consultation and necessitate high quality input data, which are only provided in high-bandwidth environments. Besides, even the further development of technologies of generative synthesis deepfakes like StyleGAN3 poses a threat to overpower current identification tools, which requires investing in their continuous update and research. The greater implication involves a host environment, which requires the convergence of expert groups in the cybersecurity world, telehealth providers, as well as the regulatory agencies to develop standards that are enacted. Another major suggestion to be made would include the signing of all session elements by using cryptographic metadata signing- a possibility that would prevent the injection of undetectable synthetic data. The collaboration involves both the public and private sector to offer anti-deepfake technologies that are scalable and cost effective, especially in acquisition of anti-deepfake tools by underserved small scale and rural healthcare facilities. Additionally, it has been suggested that adaptive thresholds relative to session conditions and human-in-the-loop confirmation of alerts are useful in matching security and user experience, especially in clinicians.

Such findings make the immediate need of proactive security integration to telehealth. The online character of tele consultation invalidates the physical boundaries of identification security and makes systems prone to identity theft, medical misinformation, and lack of privacy. As a result, resilience to synthetic media attack is bound to be used in the very core of the telemedicine infrastructure. All in all, this study would be a meaningful addition to the scarce body of works on deepfake mitigation in healthcare and a potentially scalable solution to this problem, including multimodal verification, real-time response, and unforgeable data tracking. It is necessary to conduct further studies and explore the aspects of user experience of this kind of framework that should not put too many burdens on healthcare providers or their patients due to the increased security measures. It will have to be expanded to cover the emerging sources of consultations, such as AI-powered or VR-based care, in order to remain relevant. To sum up, the combination of layers of security: biometric verification, deepfake recognition, and blockchain provenance exposes an impressive path to guarding telehealth systems against the multi-faceted and changing issue of synthetic media impersonation. This paper has shown that not only is such protection technically possible but that any hope of maintaining trust, privacy, and efficacy due to the remote nature of medical care necessitates such protection.

7. ACKNOWLEDGMENTS

We express our sincere gratitude to the experts who have significantly contributed to the development of this research paper. Their insights, guidance, and support were invaluable in shaping this work. We want to acknowledge the contributions of the following authors, particularly:

Ruvimbo Mashinge, Yeshiva University - Cybersecurity (Email: rmashing@mail.yu.edu), for leading the research and manuscript preparation.

Kumbirai Bernard Muhwati, Yeshiva University - Cybersecurity (Email: kmuhwati@mail.yu.edu), for his critical analysis and valuable inputs throughout the study.

Kelvin Magora, Yeshiva University (Email: kelvinmagoraub@gmail.com), for his support in data interpretation and framework validation.

Joy Awoleye, Yeshiva University - Computer Science (Email: awoleyejoy@gmail.com), for his technical expertise and assistance in data analysis.

We also appreciate the collaborative spirit and commitment shown by all contributors, whose collective efforts made this research possible.

8. REFERENCES

[1] Payne, D. 2024. “Hospitals’ new message for patients: Stay home.” POLITICO, May 10. Accessed June 20, 2025. https://www.politico.com/news/2024/05/10/hospitals-

telehealth-push-00157062.

[2] Vogt, E. L., H. L. Lee, A. Singh, E. M. L. Yu, S. R. B. Huang, S. A. G. Ooi, J. B. L. Lim, E. C. L. Tan, and G. C. L. Lim. 2022. "Quantifying the Impact of COVID-19 on Telemedicine Utilization: Retrospective Observational Study." *Interact. J. Med. Res.* 11, no. 1: e29880. https://doi.org/10.2196/29880.

[3] CTeL. 2024. "Five Years Later: How Telehealth Transformed Access to Healthcare Post-COVID-19." CTeL Telehealth Research, Policy, Action. Accessed June 20, 2025. https://www.ctel.org/breakingnews/five-years-later-how-telehealth-transformed-access-to-healthcare-post-covid-19.

[4] Geddes, L. 2024. "'One part of the solution': how virtual NHS wards are now a reality." *The Guardian*, February 7. Accessed June 20, 2025. https://www.theguardian.com/society/2024/feb/07/how-virtual-nhs-wards-now-reality.

[5] Bates, A. 2024. "How to Spot and Prevent Deepfakes Spreading Medical Misinformation." Eularis. Accessed June 20, 2025. https://eularis.com/how-to-spot-and-prevent-deepfakes-spreading-medical-misinformation/.

[6] Health Management. 2024. "The Growing Threat of Deepfakes and Social Engineering in Healthcare." Health Management. Accessed June 20, 2025. https://healthmanagement.org/c/cybersecurity/news/the-growing-threat-of-deepfakes-and-social-engineering-in-healthcare.

[7] Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2020. "Generative adversarial networks." *Commun. ACM* 63, no. 11: 139–144. https://doi.org/10.1145/3422622.

[8] Wang, Z., Q. She, and T. E. Ward. 2020. "Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy." arXiv. http://arxiv.org/abs/1906.01529.

[9] Pei, G., Z. Wang, Z. Ding, Z. Liu, P. Zhang, J. Xu, Y. Yu, and D. Zhang. 2024. "Deepfake Generation and Detection: A Benchmark and Survey." arXiv. http://arxiv.org/abs/2403.17881.

[10] Shen, J., R. Skerry-Ryan, N. Liu, Y. Jia, M. Castonguay, T. Nguyen, E. Battenberg, Z. Chen, Y. Isola, and B. Catanzaro. 2018. "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4779–4783. IEEE. https://doi.org/10.1109/ICASSP.2018.8461368.

[11] Jia, Y., Y. Zhang, R. Weiss, N. Chen, R. Zeghidour, and Y. Wu. 2019. "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis." arXiv. http://arxiv.org/abs/1806.04558.

[12] Zhang, B., H. Cui, V. Nguyen, and M. Whitty. 2025. "Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead." *Sensors* 25, no. 7: 1989. https://doi.org/10.3390/s25071989.

[13] Vaccari, C., and A. Chadwick. 2020. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News." *Soc. Media + Soc.* 6, no. 1. https://doi.org/10.1177/2056305120903408.

[14] Putterman, S. 2019. "Zuckerberg's video about 'billions of people's stolen data' is a deepfake." PolitiFact, June 12. Accessed June 21, 2025. https://www.politifact.com/factchecks/2019/jun/12/instagram-posts/zuckerberg-video-about-billions-peoples-stolen-dat/.

[15] Lalchand, S., V. Srinivas, B. Maggiore, and J. Henderson. 2023. "Deepfake banking and AI fraud risk." Deloitte Insights. Accessed June 21, 2025. https://www.deloitte.com/us/en/insights/industry/financial-services/deepfake-banking-fraud-risk-on-the-rise.

[16] Beaumont, H. 2024. "'A lack of trust': How deepfakes and AI could rattle the US elections." Al Jazeera, June 19. Accessed June 21, 2025. https://www.aljazeera.com/news/2024/6/19/a-lack-of-trust-how-deepfakes-and-ai-could-rattle-the-us-elections.

[17] Azzuni, H., and A. El Saddik. 2025. "Voice Cloning: Comprehensive Survey." arXiv. http://arxiv.org/abs/2505.00579.

[18] Croitoru, F.-A., A. El-Dawy, A. A. Al-Maadeed, J. T. Hu, A. Abu-Hamdan, S. S. S. A. Mohamed, M. J. Al-Mulla, and A. Al-Maadeed. 2024. "Deepfake Media Generation and Detection in the Generative AI Era: A Survey and Outlook." arXiv. http://arxiv.org/abs/2411.19537.

[19] Nguyen, H. H., F. Fang, J. Yamagishi, and I. Echizen. 2019. "Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos." arXiv. https://doi.org/1906.06876.

[20] Qawasmi, M., and O. Al-Kadi. 2025. "Detecting face tampering in videos using deepfake forensics." *Multimed. Tools Appl.* https://doi.org/10.1007/s11042-025-20865-4.

[21] Giuffrè, M., and D. L. Shung. 2023. "Harnessing the power of synthetic data in healthcare: innovation, application, and privacy." *npj Digit. Med.* 6, no. 1: 186. https://doi.org/10.1038/s41746-023-00927-3.

[22] Kokosi, T., and K. Harron. 2022. "Synthetic data in medical research." *BMJ Med.* 1, no. 1: e000167. https://doi.org/10.1136/bmjmed-2022-000167.

[23] Gonzales, A., G. Guruswamy, and S. R. Smith. 2023. "Synthetic data in health care: A narrative review." *PLOS Digit. Heal.* 2, no. 1: e0000082. https://doi.org/10.1371/journal.pdig.0000082.

[24] Goyal, M. K. 2023. "Synthetic Data Revolutionizes Rare Disease Research: How Large Language Models and Generative AI are Overcoming Data Scarcity and Privacy Challenges." *Int. J. Recent Innov. Trends Comput. Commun.* 11, no. 11: 1368–1380. https://doi.org/10.17762/ijritcc.v11i11.11411.

[25] Frid-Adar, M., E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. 2018. "Synthetic Data Augmentation using GAN for Improved Liver Lesion Classification." arXiv. http://arxiv.org/abs/1801.02385.

[26] Chen, R. J., M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood. 2021. "Synthetic data in machine learning for medicine and healthcare." *Nat. Biomed. Eng.* 5, no. 6: 493–497. https://doi.org/10.1038/s41551-021-00751-8.

[27] Cai, Z., and M. Li. 2024. "Integrating frame-level boundary detection and deepfake detection for locating

manipulated regions in partially spoofed audio forgery attacks." *Comput. Speech Lang.* 85: 101597. https://doi.org/10.1016/j.csl.2023.101597.

[28] Sun, G., Y. Zhang, H. Yu, X. Du, and M. Guizani. 2020. "Intersection Fog-Based Distributed Routing for V2V Communication in Urban Vehicular Ad Hoc Networks." *IEEE Trans. Intell. Transp. Syst.* 21, no. 6: 2409–2426. https://doi.org/10.1109/TITS.2019.2918255.

[29] Xia, J.-Y., S. Li, J.-J. Huang, Z. Yang, I. M. Jaimoukha, and D. Gündüz. 2023. "Metalearning-Based Alternating Minimization Algorithm for Nonconvex Optimization." *IEEE Trans. Neural Networks Learn. Syst.* 34, no. 9: 5366–5380. https://doi.org/10.1109/TNNLS.2022.3165627.

[30] Ansari, U., P. Kamble, and A. Shinde. 2023. "Deepfakes detection using human eye blinking." *Int. Res. J. Mod. Eng. Technol. Sci.* 5, no. 11: 3344–3345. https://www.irjmets.com/uploadedfiles/paper//issue_11_november_2023/46123/final/fin_irjmets1701874611.pdf.

[31] Yu, N., V. Skripniuk, S. Abdelnabi, and M. Fritz. 2022. "Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data." arXiv. http://arxiv.org/abs/2007.08457.

[32] Tan, C., Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei. 2024. "Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Learning." arXiv. http://arxiv.org/abs/2403.07240.

[33] Dong, C., A. Kumar, and E. Liu. 2022. "Think Twice Before Detecting GAN-generated Fake Images from their Spectral Domain Imprints." In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7855–7864. IEEE. https://doi.org/10.1109/CVPR52688.2022.00771.

[34] Wang, X., H. Guo, S. Hu, M.-C. Chang, and S. Lyu. 2023. "GAN-Generated Faces Detection: A Survey and New Perspectives." *Frontiers in Artificial Intelligence and Applications* 378: 558–572. https://doi.org/10.3233/FAIA230558.

[35] Shahzad, S. A., A. Hashmi, Y.-T. Peng, Y. Tsao, and H.-M. Wang. 2023. "AV-Lip-Sync+: Leveraging AV-HuBERT to Exploit Multimodal Inconsistency for Video Deepfake Detection." arXiv. http://arxiv.org/abs/2311.02733.

[36] Raza, M. A., and K. M. Malik. 2023. "Multimodaltrace: Deepfake Detection using Audiovisual Representation Learning." In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 993–1000. IEEE. https://doi.org/10.1109/CVPRW59228.2023.00106.

[37] Kumaran, U., S. R. Rammohan, S. M. Nagarajan, and A. Prathik. 2021. "Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN." *Int. J. Speech Technol.* 24, no. 2: 303–314. https://doi.org/10.1007/s10772-020-09792-x.

[38] Bajwa, M. K. Z., A. Castiglione, and C. Pero. 2025. "Mel Spectrogram-Based CNN Framework for Explainable Audio Deepfake Detection." In *Advanced Information Networking and Applications. AINA 2025. Lecture Notes on Data Engineering and Communications Technologies*, edited by L. Barolli, 407–416. Cham: Springer. https://doi.org/10.1007/978-3-031-87784-1_37.

[39] Khan, I. R., S. Aisha, D. Kumar, and T. Mufti. 2023. "A Systematic Review on Deepfake Technology." In *Lecture Notes in Networks and Systems Proceedings of Data Analytics and Management*, 669–685. Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-7615-5_55.

[40] Khanjani, Z., G. Watson, and V. P. Janeja. 2023. "Audio deepfakes: A survey." *Front. Big Data* 5. https://doi.org/10.3389/fdata.2022.1001063.

[41] Khan, A. A., A. A. Laghari, S. A. Inam, S. Ullah, M. Shahzad, and D. Syed. 2025. "A survey on multimedia-enabled deepfake detection: state-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions." *Discov. Comput.* 28, no. 1: 48. https://doi.org/10.1007/s10791-025-09550-0.

[42] Choi, J.-E., K. Schäfer, and S. Zmudzinski. 2024. "Introduction to Audio Deepfake Generation: Academic Insights for Non-Experts." In *3rd ACM International Workshop on Multimedia AI against Disinformation*, 3–12. New York, NY, USA: ACM. https://doi.org/10.1145/3643491.3660286.

[43] Dasgupta, S., J. Mason, X. Yuan, O. Odeyomi, and K. Roy. 2025. "Enhancing Deepfake Detection using SE Block Attention with CNN." https://doi.org/10.1109/icABCD62167.2024.10645262.

[44] Kroiß, L., and J. Reschke. 2025. "Deepfake Detection of Face Images based on a Convolutional Neural Network." arXiv. http://arxiv.org/abs/2503.11389.

[45] Abbasi, M., P. Váz, J. Silva, and P. Martins. 2025. "Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks." *Appl. Sci.* 15, no. 3: 1225. https://doi.org/10.3390/app15031225.

[46] Ghosh, A., H. H. Singh, R. Singh, H. Singh, and M. Singh. 2025. "Blockchain-Assisted Serverless Framework for AI-Driven Healthcare Applications." *Int. J. Adv. Comput. Sci. Appl.* 16, no. 5: 473–482. https://doi.org/10.14569/IJACSA.2025.0160546.

[47] Johnson, A. E. W., T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. 2016. "MIMIC-III, a freely accessible critical care database." *Sci. Data* 3, no. 1: 160035. https://doi.org/10.1038/sdata.2016.35.

[48] Tolosana, R., R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia. 2020. "Deepfakes and beyond: A Survey of face manipulation and fake detection." *Inf. Fusion* 64: 131–148. https://doi.org/10.1016/j.inffus.2020.06.014.

[49] Ding, Y.-Y., J.-X. Zhang, L.-J. Liu, Y. Jiang, Y. Hu, and Z.-H. Ling. 2020. "Adversarial Post-Processing of Voice Conversion against Spoofing Detection." In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 556–560.

[50] Karras, T., S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. 2020. "Analyzing and Improving the Image Quality of StyleGAN." arXiv. http://arxiv.org/abs/1912.04958.

[51] Kruse, C. S., P. Karem, K. Shifflett, L. Vegi, K. Ravi, and M. Brooks. 2018. "Evaluating barriers to adopting telemedicine worldwide: A systematic review." *J.*

Telemed. Telecare 24, no. 1: 4–12. https://doi.org/10.1177/1357633X16674087.

[52] Yan, X., W. Li, P. Li, J. Wang, X. Hao, and P. Gong. 2013. "A Secure Biometrics-based Authentication Scheme for Telecare Medicine Information Systems." *J. Med. Syst.* 37, no. 5: 9972. https://doi.org/10.1007/s10916-013-9972-1.