ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

SYSTEMATIC REVIEW: DEEP FAKE DETECTION IN MEDICAL IMAGES

Ayesha Manzoor[1], Mariyam Amreen[2], Imra Shoukat[3], Dr. Umair Muneer[4], Dr. Imtiaz Hussain[5], Iqra Rehman[6], Iman Neha Butt[7]

[1,2,3,4,5,6,7]Department of Computer Science, University of Management and Technology, Sialkot 51310, Pakistan

ayeshamnzr297@gmail.com , mariyamamreen25@gmail.com , imrashoukat7@gmail.com , umair.muneer@skt.umt.edu.pk, imtiaz.hussain@skt.umt.edu.pk , Iqrarehman1109@gmail.com , imbutt1207@gmail.com

Corresponding Author: Ayesha Manzoor

ABSTRACT

The rapid advancement of generative models, particularly Generative Adversarial Networks (GANs), has led to the rise of highly realistic synthetic content, including medical deep fakes. Although such technologies offer significant potential for applications in medical training, simulation, and image improvement, they raise critical concerns regarding the authenticity, security, and reliability of medical data. Despite growing awareness of these threats, there is currently no comprehensive systematic review focusing exclusively on deep fake detection techniques within the domain of medical imaging. To address this gap, the present study presents the first known systematic literature review (SLR) on medical deep fake detection in images. The review was conducted through a structured and methodical examination of top-tier scientific databases, including IEEE Xplore, Elsevier ScienceDirect, SpringerLink, ACM Digital Library, and arXiv. Following a rigorous selection process based on relevance, novelty, and contribution, thirty peer-reviewed studies were analyzed. The main contributions of this work are as follows: (1) a comparative analysis of traditional machine learning and deep learning-based detection techniques such as ResNet, DenseNet, CNNs, YOLO architectures, and ensemble frameworks; (2) benchmarking of detection models against real-world datasets including CT-GAN, DFDC, and LIDC-IDRI; (3) assessment of core evaluation metrics such as accuracy, precision, recall, AUC, and model robustness; (4) identification of ethical concerns surrounding privacy, potential misuse, and the absence of clinical validation standards; and (5) recommendations for future improvements in model robustness, real-time operability, and interpretability. This review aims to serve as a foundational reference for researchers in the field of medical deep fake detection by highlighting current advancements, unresolved challenges, and prospective research directions.

INDEX TERMS *Deepfake detection, medical imaging, generative adversarial networks, machine learning, systematic literature review.*

Introduction

The integration of Artificial Intelligence (AI) into the healthcare sector has led to significant advancements, especially in medical imaging. Deep learning approaches have become essential for tasks such as disease classification, image segmentation, and anomaly detection, significantly enhancing diagnostic accuracy, minimizing human error, and streamlining clinical workflows. Among these technologies, Generative Adversarial Networks (GANs) and diffusion models have emerged as leading techniques due to their ability to generate highly realistic synthetic medical images. These artificially generated images have proven valuable for data augmentation, and preserving patient privacy in data sharing scenarios. Such capabilities are particularly useful situations involving small datasets or ethically sensitive medical data, ensuring effective model

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

training without compromising confidentiality [1], [3], [5].

However, despite their benefits, these generative technologies are increasingly being misused for malicious purposes. A major concern is the generation of medical deep fakes highly realistic but artificially manipulated

diagnostic images. These altered images can insert, remove, or modify clinical features in modalities such as CT, MRI, and X-ray, often remaining invisible to both clinicians and AI-based diagnostic systems. The implications are significant, including the risk of misleading diagnoses, falsifying medical records, disrupting clinical trials, and undermining the integrity of scientific research [2], [4], [8]. A significant concern is the absence of robust, real-time defense mechanisms within medical imaging infrastructures, which increasingly exposes clinical, legal, and insurance systems to potential security vulnerabilities [7], [10].

Moreover, medical deep fake detection tools remain underdeveloped. Most existing detection systems are adapted from general-purpose applications such as facial recognition and fail to perform adequately in medical contexts due to unique challenges, including grayscale intensity distributions, modality-specific features, and high inter-patient variability [9], [11]. Additionally, many of the current detection methods suffer from limited generalization ability, poor real-time performance, low interpretability, and limited applicability across different imaging modalities. Clinical integration remains minimal, as most solutions are still limited to experimental or research settings. Therefore, a comprehensive review of existing literature is crucial to assess progress, find critical gaps, and outline future directions for developing effective medical deep fake detection systems.

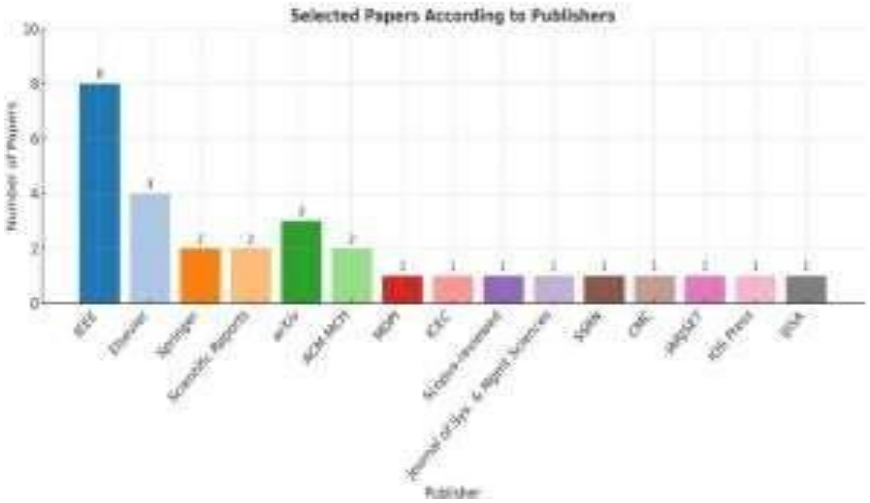


Figure 1. Artical type distribution in searched databases

To meet this need, this review systematically compares recent developments in the generation and detection of deep fake medical images across a diversity of imaging modalities. The literature encompasses a wide array of techniques, including GAN-based pathology simulation [1], convolutional neural network (CNN)- based detection methods [3], Zero-Shot Learning (ZSL) approaches [2], diffusion-based anomaly detection frameworks [5], and hybrid models that integrate generators with discriminators [6], [11]. For instance, a study involved the synthesis of over 320,000 knee X-ray images, which not

only misled experienced radiologists but also improved the performance of diagnostic models

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

trained on limited real data [1]. An additional investigation proposed a GAN-based system capable of altering retinal and chest X-ray images by adding or removing pathological evidence, all while preserving patient identity [4]. Several studies reported that even trained clinicians were often not capable of distinguish between real and fake images, whereas deep learning classifiers achieved accuracy rates exceeding 95% [3], [7], [10].

The detection techniques explored a variety of efficient CNN-based architectures to advanced anomaly detection models leveraging diffusion mechanisms. Some studies propose hybrid frameworks where architectures such as DenseNet serve as discriminators [6], while others employ U-Net and CycleGAN to produce challenging alterations for robustness testing [7]. Unsupervised learning approaches have also gained power due to their capacity to detect forgeries without depending on labeled data [5]. Nonetheless, challenges persist regarding model interpretability, generalization across diverse datasets, and real-world clinical deployment. Numerous studies advocate for the integration of security solutions such as digital watermarking, block chain technologies, and embedding detection systems within Picture Archiving and Communication Systems (PACS) to safeguard image integrity [12]. Furthermore, the absence of standardized datasets and evaluation protocols makes it difficult to compare detection techniques effectively [2], [9].

This article provides a comprehensive examination of machine learning and deep learning-based techniques for the detection of medical deep fakes, along with a summary of publicly available medical imaging datasets, key performance evaluation metrics, and the ethical and technical challenges reported in the existing literature. Furthermore, the main contributions of this systematic review, along with the justification for its significance, are outlined in the following sections.

A. Problem Statement

AI-generated fake medical images, commonly known as deep fakes, are becoming increasingly realistic and pose a significant threat to patient diagnosis and treatment if not correctly detected. Although numerous detection systems have been proposed, many rely on narrowly scoped datasets, limiting their generalizability to diverse or unseen medical imaging scenarios. Furthermore, these systems are infrequently validated in clinical settings, raising concerns about their real-world reliability and trustworthiness. Most existing methods are ineffective in detecting deep fakes within medical video data and frequently perform well only on training datasets but fail in real-world applications. Therefore, the current methods lack the robustness and reliability essential for safe and effective deployment in healthcare environments. There is an urgent need for more generalized, clinically validated, and trustworthy deep fake detection systems tailored specifically to the medical domain.

B. Key Contributions

This study offers the following key contributions:

- It presents a comprehensive literature review of current research on deep fake detection within the area of medical imaging.
- It systematically classifies detection techniques based on architectural frameworks, including convolutional neural networks (CNNs), GAN- based discriminators, hybrid models, anomaly detection approaches, and diffusion-based methods.
- It provides a comparative examination of performance metrics such as accuracy, recall, AUC, and generalization capability across several imaging modalities including CT, X-ray, MRI, and retinal images.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

- It identifies major limitations of current models, including limited generalizability, inadequate clinical deployment, insufficient dataset diversity, and a lack of model interpretability.

The paper is structured as follows: In Section II, a review of the literature is discussed. The detailed methodology is described in Section III. The paper selection criteria, comparison framework, and evaluation metrics are described in Section IV. The model-wise and modality-wise performance comparisons and limitations are presented in Section V. In Section VI the remaining gaps are emphasized, and future research directions are proposed. Section VII concludes the research by summarizing key findings and reflecting on the importance of developing clinically practical detection systems.

II. Related work

The growing attention in both the generation and detection of medical deep fakes is power by the increasing reliance on AI- assisted diagnostic systems and the pressing need for secure and trustworthy medical imaging. Numerous models and datasets have been proposed to improve detection accuracy and address related ethical and technical challenges. Abdel Rahman and Omar Al-Kadi [13] directed a comparative analysis of 13 pre- trained Deep Convolutional Neural Network (DCNN) architectures, including ResNet50V2 and DenseNet169, using the LIDC-IDRI and CT-GAN datasets. Their finding identified ResNet50V2 as the best effective model for detecting tampered CT images. Likewise, Siddharth Solaiyappan and Yuxin Wen [14, 25] evaluated eight machine learning models and know that traditional classifiers, such as Random Forest, performed well under low-data conditions, whereas CNN-based models outperformed others when localized image regions were used as input. Similarly, Saleh Albahli and Mariam Nawaz [15]
suggested the MedNet model, integrating EfficientNet-V2 with spatial-channel attention, which attained strong performance on the CT-GAN dataset. These studies collectively underscore the significance of both model architecture and dataset design in enhancing classification performance.

Several researchers have also investigated hybrid and ensemble models to leverage complementary feature extraction abilities. Kranthi `Kumar G. et al. [18] established an ensemble of DenseNet121 and VGG16, achieving 90% accuracy in detecting deep fake lung CT scans. Similarly, Rajat Budhiraja and Manish Kumar [19] introduced the Convolutional Reservoir Network (CoRN), a hybrid model that integrates CNNs with Reservoir Computing, performing exceptionally well on ultra-lesser datasets. In a different study, Pradeepan and Gladston Raj [21] joined spatial feature extraction through EfficientNet-B0 with frequency-domain analysis via Discrete Wavelet Transform (DWT), resulting in an outstanding 99.6% accuracy. Reinforcing these insights, MeenaPrakash et al. [11] presented a DenseNet -enhanced GAN framework, using the discriminator's ability in DenseNet to detect subtle manipulations in CT, MRI, and X- ray images. Overall, these studies indicate that merging spatial, statistical, and frequency-based features yields improved performance in identifying tampered medical images.

Generative techniques for data augmentation represent another key domain of interest. Roa'a Al-Emaryeen and Sara Al-Nahhas [20] utilized CycleGAN to generate artificial brain tumor images, which were segmented using a U-Net model, resulting in better Dice scores. Likewise, Vajira Thambawita et al.
[22] established Pulse2Pulse and WaveGAN models to produce ECG signals that preserve diagnostic utility while protection patient privacy. Nawaz Waqas et al. [16] presented improvements to PGGAN by incorporating self-attention and spectral normalization, which

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

helped the generation of realistic knee MRI images and improved segmentation performance. Fabi Prezja et al. [1] trained GANs to generate over 320,000 synthetic knee X- rays, significantly improving model accuracy under data- constrained conditions. Hussain et al. [26] conducted a systematic review highlighting GANs' critical role in data augmentation and image augmentation, particularly where real data is scarce. These findings collectively affirm that GAN-based augmentation is instrumental in reducing data scarcity and strengthening diagnostic model resilience.

Unsupervised and low-data detection approaches have also gained prominence, especially in real- world scenarios where labeled datasets are narrow. Grabovski et al. [5] presented a novel "back-in-time" diffusion model for anomaly detection, capable of reversing the generation process to categorize synthetic tumor alterations, achieving AUC scores of 0.90 and 0.96. Likewise, MRA Parate et al.

[2] emphasized the applicability of Zero-Shot Learning (ZSL) for detecting image manipulations without requiring hugelabeled datasets, reporting YOLO as the best-performing model in traditional metrics. MA Arshed et al. [3] addressed the problem of insurance fraud in skin cancer diagnosis using a diffusion model combined with CNNs and histogram-based analysis, reaching nearby-perfect classification accuracy. Despite their promise, these models often face problems in generalizing across heterogeneous datasets, pointing to the necessity for scalable and adaptive detection methods.

In addition to technical aspects, the ethical and dual-use implications of medical deep fake technologies have been discovered. Neal Mangaokar et al. [4] presented "Jekyll," a GAN-based adversarial model able of injecting artificial diseases into diagnostic images such as retinal scans and X- rays. likewise, Waier et al. [8] highlighted the erosion of diagnostic trust resulting from such manipulations and proposed the acceptance of technologies like digital watermarking and block chain to ensure image authenticity. On the other hand, Zhu et al. [29] demonstrated a beneficial use of deep fake technology by anonymizing videos of Parkinson's patients through face-swapping techniques, preserving clinical cues critical for diagnosis. Kaur et al. [27] strained the significance of ethical governance and multidisciplinary collaboration, particularly as deep fakes become integrated into patient communication and medical training. S. Agarwal et al. [30] further supported for comprehensive policy frameworks to stabilize innovation with security. Together, these studies highlight the crucial need for regulatory and technological safeguards in parallel with scientific growth.

From a technical standpoint, several studies have focused on improving or comparing CNN backbones for superior detection accuracy. DT Alhalabi et al. [9] evaluated four CNN architectures VGG16, EfficientNetV2, InceptionV3, and a progressive model on CT scan datasets, eventually identifying VGG16 as the most effective, with a recall rate of 98%. A. Alsaheel et al. [10] linked six different models and concluded that an improved ResNet101 offered the highest accuracy (99%) in identifying untampered and manipulated CT scans. Similarly, Patel et al. [6] demonstrated that their dense CNN model achieved higher accuracy through five GAN-generated datasets, although its application was limited to static images. YS Kim et al. [7] suggested a custom CNN for detecting manipulated fundus images created with U-Net and CycleGAN, achieving an AUC of 0.913 outperforming human ophthalmologists. Nevertheless, the study noted the constraints posed by small datasets and highlighted the need for broader clinical validation. These results suggest that both classical and customized CNN models remain effective for deep fake detection when accurately adapted and validated.

The establishment of standardized datasets and collaborative benchmarking initiatives is also

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

important for allowing fair comparisons across detection frameworks. Dolhansky et al. [28] presented the Deep Fake Detection Challenge (DFDC) dataset, which, although centered on facial deep fakes, laid foundational work for transfer learning and cross-field benchmarking in medical applications. Bhuvan Gowda et al. [12] highlighted the need of adapting deep fake detection approaches across diverse modalities (CT, MRI, X-ray) and integrating them into hospital systems by leveraging CNNs, GANs, and anomaly detection methods. Deep this, Benjamin Phipps and pearse A. Keane et al.
[17] called for the development of standardized models and validation protocols to confirm the safe and effective deployment of generative AI in ophthalmology and beyond.
Lastly, many comprehensive overviews have synthesized the existing landscape of medical deep fake research while identifying future directions. Lakshmi and Hemanth [23] offered a broad review covering deep fake applications in medical imaging, including synthesis, enhancement, modality transfer, and augmentation. Karaköse et al. [24] supported for the integration of deep fake detection mechanisms directly within clinical infrastructures to simultaneously improve diagnostic accuracy and security. Collectively, these studies deliver a unified vision: the future of medical deep fake research must combine innovation, model transparency, dataset diversity, and strong ethical and regulatory frameworks to ensure safe and consistent adoption in clinical practice.

III. Research Methods

A wide range of literature has been reported for detecting deep fakes in medical images, covering various imaging modalities such as CT, MRI, X-rays, ECG, and fundus images. The primary objective of this systematic review is to examine which deep fake detection techniques, among traditional machine learning, deep learning, and anomaly detection approaches, have demonstrated superior performance in identifying tampered medical images. An essential aspect of this review is to investigate the influence of professional and synthetically generated datasets on the detection performance and generalizability of proposed methods.
Another critical objective of this SLR is to highlight the key challenges reported in the literature that hinder the clinical application and real-time deployment of deepfake detection systems in medical imaging. To the best of the authors' knowledge, this systematic review on deepfake detection in medical images is one of the first comprehensive attempts from 2020 to 2025 that consolidates evidence from multiple studies, identifies gaps, and suggests future research directions.
This SLR is performed using the guidelines provided by [31]. According to the guidelines, an SLR involves planning, conducting, and reporting the available research relevant to a specific topic or research area. The motivation behind performing this review is to identify the strengths and limitations of existing detection frameworks, examine dataset availability and ethical constraints, and propose a direction for future research in developing robust, interpretable, and clinically deployable deepfake detection systems [32]. As recommended by the literature, the systematic review process was executed in the following stages

1) Defining the research questions.
2) Identifying a few relevant studies and conducting a pilot study.
3) Searching for data across reputable scientific databases including IEEE, Elsevier ScienceDirect, SpringerLink, ACM Digital Library, and arXiv.
4) Documenting the entire search strategy.
5) Appraisal and selection of studies.
6) Analyzing and presenting the extracted results.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

7) Discussing the generalized conclusions, limitations, and challenges identified across the studies.
8) Making informed recommendations for future work.

The overall objective of the planned systematic review is to critically analyze and summarize the findings on deep fake detection in medical images, identify the limitations present in existing detection frameworks, and highlight potential opportunities for developing clinically reliable and ethically responsible solutions in the healthcare domain.

A. Research Question

Constructing clearly defined research questions is important in favor to evaluating the progress that has been made in the detection of deep fakes in medicine and the relevant research. The goal of this systematic literature review (SLR) is to identify critical gaps, trends, challenges, and main dynamics related to the medical deep fake justification and their identification. This review was conducted by the PIOC framework as proposed by [33], which ensures that all relevant essentials are taken while analyzing a phenomenon. In accordance with the scope outlined in Table 2, this SLR addresses the following questions.

TABLE 1
Criteria for Research Questions

Population	Medical imaging datasets (CT, MRI, X-ray images containing original and deepfake/manipulated images
Intervention	Techniques and methods for Deepfake detection in medical images
Outcome	Detection accuracy, reliability, robustness, and clinical applicability
Context	Healthcare settings, radiologydepartments, AI-based medical imaging research

RQ1: What empirical evidence of the benefits and limitations of deep learning and machine learning approaches currently exist to support the effectiveness of different deep fake detection techniques in medical imaging?

RQ2: What performance measures have been taken for measuring the accuracy of medical deep fake detection techniques?

RQ3: What deep/machine learning approaches currently exist to support the effectiveness of different techniques in Deep fake detection in medical images?

- RQ3a: What techniques have been reported for the detection of deep fake medical images?
- RQ3b: What deep fake detection approaches are reported to be superior for medical images?
- images?

RQ4: What are the potential challenges highlighted in existing studies to build a robust and clinically applicable deep fake detection models for medical images?

RQ5: What are thecritical characteristics of thedatasets used in this study? Do their features seem to affect the results?

- RQ5a: Which type of dataset has been been used for this research (Professional or self-acquired)?
- RQ5b: What are the main aspects of a dataset for the Deep fake detection in

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

medical imaging? Do they affect results?

In existing literature, SLRs seem to follow a three-step sequence: planning or structuring, executing, and reporting each containing several executable sub-tasks under them. This review uses an approach advocated by [33] who recommends starting with an experimental study. The aim of this early step is to assess how well formulated the research questions are in terms of significance as well as examine whether collecting appropriate data for analysis would be feasible. To achieve these objectives, a pilot study was conducted using a subset of selected documentsto determine whether sufficient information was available to address the formulated research questions and to verify the feasibility of the proposed analytical strategy. Based on the results, necessary reviews were made to improve thereview outline. Following this, a full-scale SLR was conducted, concentrating on techniques for detecting medical deep fakes, relevant datasets, and their applications within medical imaging.

B. Search Strategy

To conduct a systematic literature review (SLR), an efficient search technique is required to find all relevant studies. A detailed literature search was conducted to address the formulated research questions. This aligns with [34], where search term construction has multiple stages which include

1) From the research questions, primary search terms were gained by identifying components of Population, Intervention, Outcome, and Context (PIOC).
2) Existing literature on the topic was used to collect relevant keywords.
3) Dictionaries and other language resources were perused for synonyms or alternative spellings as opposed to them being listed.

1) Distinct keywords could be combined using Boolean operator AND which reduces the search scope.

To increase search coverage when necessary, all grouped terms with similar meaning using Boolean operator-OR

C. Search String

For this review, a search string was created to identify literature focused on the detection of deepfakes in medical images. The parts include

DEEPFAKE: ("deepfake" **OR** "deep fake" **OR** "synthetic image" **OR** "AI-generated image" **OR**"manipulated image" **OR** "fake medical image" **OR** "GAN-based image") **AND DETECTION** ("detection"**OR** "identification" **OR** "recognition" **OR**" diagnosis" **OR**" classification"
OR" analysis") **AND MEDICAL**: ("medical" **OR** "clinical" **OR** "healthcare" **OR** "diagnostic" **OR** "radiology") **AND IMAGES**: ("images" **OR** "scans" **OR** "X-rays" **OR** "CT" **OR** "MRI" **OR** "medical imaging")

Although some expressions may appear too specific or indefinite as per Table 3, all of them were included so that as many relevant studies as possible could be covered. Still, in the section step, it was ensured that only such articles which clearly discuss whether deepfakes in medical images are detected or identified are included.

TABLE 2
Keyword synonyms

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Keyword	Synonyms
Population	Medical imaging datasets (CT, MRI, X-ray images containing original and deepfake/manipulated images
Intervention	Techniques and methods for Deepfake detection in medical images
outcome	Detection accuracy, reliability, robustness, and clinical applicability
context	Healthcare settings, radiology departments, AI-based medical imaging research

The search strategy was typed to certain important considerations. We set a tradition publication gap from January 2020 to April 2025, which is the period during which deepfake generation and detection methodologies have been established within the medical imaging domain. Therefore, studies published after April 2025 are excepted since they would make the review timeframe unreliable; this is shown in Table 4.

D. String Refinement

After generating the initial search string, results from the selected databases had to be validated. The aim was to ensure that project and relevant works on medical deep fake detection appeared in theresults. If hardly any such works were returned, then further refinement wasneeded in the search string. These adjustments took place not only the refinement of synonyms identified but also in the parameters for searching in each database.

As part of enhance process, an impact evaluation was conducted on different factors: Inclusion/exclusion of synonyms, types of publications, date ranges, language filters, research domains, and journal sources. Adjustments made iteratively with individual tests until acceptable comprehensive results achieved. Fig 2 illustrates step-by-step search string development; Table 5 shows number of papers recovered from each database after applying final filters

TABLE 3
Search strategy decisions

Searched Databases	IEEE Xplore[8], SpringerLink[2], Elsevier[4], Scientific reports[2], arXiv[3], ACM-MCH[2], MDPI[1], ICEC[1], Scopus-reviewed[1], journal of Sys & Mgmt sciences[1], SSRN[1] CMC[1], IARJSET[1], IOS press[1],IJISA[1]
Searched Items	Journal papers, conference papers, reviews, book chapters
Search Applied On	Full-text peer-reviewed journal and conference papers were selected. Titles and abstracts were initially screened for relevance to deepfake detection in medical imaging, with preference given to technically focused ML/DL studies
Publication Period	Since Jan, 2020

D. Study Selection

In order to bring out this systematic literature review, an expressed and deep search was made

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

within the leading academic databases in the catalog of IEEE Xplore, Science Direct, Springer Link, ACM Digital Library, and arXiv. The extreme objective was to find studies that were explicitly carried out on deep fake detection in medical images. According to certain inclusion criteria, a total of 30 peer- reviewed works were selected for in-depth analysis. While other surveys may include hundreds of publications to fill up their more generalized charge, this review is limited to focused 30 studies that have significance and quality as well as empirical contribution regarding detection of medical image manipulation. This selected work extends a variety of healthcare picture types, featuring CT, MRI, X-ray, fundus, and ECG images. Each study was carefully looked at to pull out important details about data set characters, model structures, evaluation measures, noted pros and cons, and recommended next steps. The distribution of papers selected over the years is shown in Fig.2.

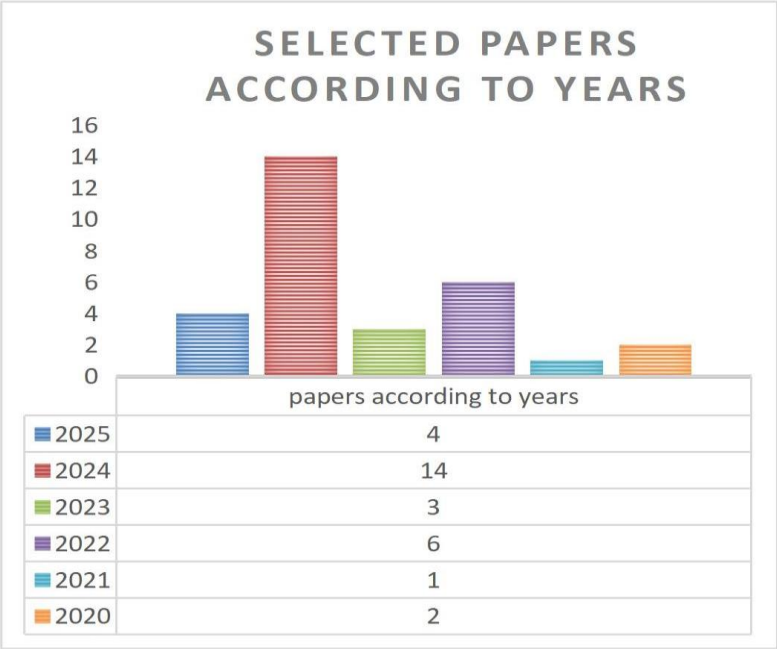


Figure 2. Year wise distribution of selected studies.

TABLE 4
Search limits on searched databases.

Category	
Year Range	Filter papers by publication (e.g 2020-25).
Language	Includes only papers written in English
Publication Type	Includes only peer-reviewed journals, conferences, or reviews (excludes newsletters, editorials, etc.).
Subject Area	Limit results to specific field like medical learning, AI or healthcare.
Access Type	Only includes full-text available papers

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

1) INCLUSION CRITERIA

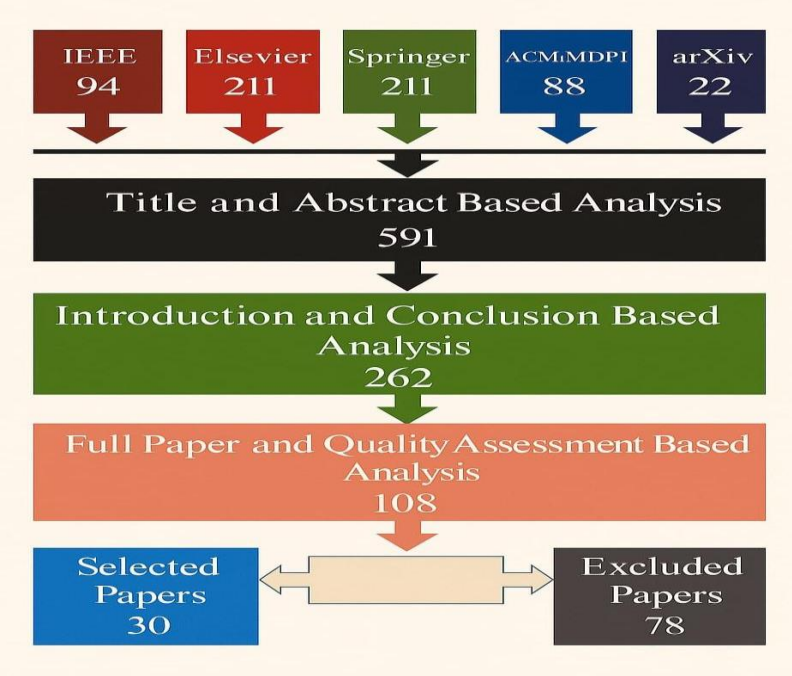
We used the following criteria for studies to be included in this review

1. The study must focus on the deepfakes detection in the medical imaging in particular.
2. Detectors proposed should use data-driven techniques such as machine learning, deep learning, or a combination of them.
3. The studies should have experimental validation on either real/publicly available or synthetic medical image datasets.
4. Studies were included if they were published in English and in peer review.
5. We prioritized publications that were published in the period 2020-2025, to consider the most recent technological developments.

2) **EXCLUSION CRITERIA**

The following criteria were applied to exclude studies

1. Deepfake generation without considering the impact of detection.
2. Review/survey/theory papers (without experiments)".
3. Publications available at address specified which are not found in the search results of the Funder and Research method specifics (see both complete or single searches).
4. Duplicate studies or papers that were just the abstracts were not analyzed in detail.
5. The time distribution of the included studies show a significant increase in related publications in 2024 and 2025, reflecting the increasing attention on protecting clinical diagnostics.
6. The year-wise distribution of the selected studies showed a rise in interest, particularly in 2024 and 2025, reflecting the group's growing concern with securing clinical diagnostics against AI-driven image manipulation



ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Figure 3. Summary of search process

IV. Research Methods

The outcome and results are presented in this section removed from the reviewed papers to answer the research questions. All the research questions are answered according to the relevant studies emphasized during the SLR.

A. BENEFITS AND LIMITATIONS OF DEEP AND MACHINE LEARNING APPROACHES (RQ1)

Empirical observation shows strong evidence supporting the effectiveness of machine learning and deep learning approaches for detecting deep fakes in medical imaging. With sets such as CT-GAN and LIDC-IDRI, the performance of algorithms such as Deep Convolutional Neural Networks (DCNNs) for detecting deep fakes has reached an amazing accuracy of 97.86% on manipulated CT images that were transformed with models like ResNet50V2 and DenseNet169. MedNet, a study utilizing EfficientNet-V2 combined with attention mechanisms, was also delivered promising classification results. In addition to DCNN-based methods, other hybrid models based on more than one architecture, such as a combination of DenseNet121 and VGG16, also performed well and there were examples of systems achieving around 90% accurate deep fake detection on lung CT scans. The basic idea in these models is to combine the DCNN features with frequency-domain information, but this is not totally consistent across methods. For example, one method fuses the Discrete Wavelet Transform with EfficientNet-B0, and the model obtained 99.6% detection accuracies through combining spatial and frequency characteristics. There is also growing interest in methods based on Zero-Shot Learning (ZSL) or unsupervised methods, such as diffusion models were able to show promise in difficulties with limited datasets with increasing results and diffusion-based methods producing AUC of near to 0.96. So, the message here is that both supervised and unsupervised models have differences but they can be effective if properly modified to deep fake detection, especially in the medical domain.

While deep learning models have great capabilities in detecting deep fakes in medical imaging, there are limitations that constrain their practical use. The main concern surrounds generalizability; models that are effective on synthetic or curated datasets.

Furthermore, as many deep learning models are black box models, the challenge is that their rationale is difficult to interpret and could be highly unreliable. This lack of interpretability can reduce trust from medical professionals, especially if patients' outcomes rely on their decisions. Finally, many medical imaging deep fake detection models often do not have standard evaluation benchmarks or publicly available datasets that can help provide a level playing field for methods comparison, largely preventing consistent or fair comparisons. Overall, there is an urgent requirement for more interpretable, clinically validated, and generalizable deep fake detection approaches in the medical community.

B. PERFORMANCE MEASURES FOR MEDICAL DEEPFAKE DETECTION TECHNIQUES (RQ2)?

Different evaluation methods have been used in the medical research community to evaluate medical deep fake methods. Given many different computational models including Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), ensemble learning, auto encoders, diffusion, models, and modifiers the performance metrics are all different. It is valued to compile and summarize the evaluation criteria that are the best and most commonly used. From a systematic literature review, the most commonly reported metrics

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

include Accuracy, F1-score, Sensitivity, and Specificity. For picture manipulation and synthesis tasks, and specifically GAN based techniques, the most commonly used metrics include Area Under the Curve (AUC), Dice coefficient, Jaccard index, Inception Score, Fréchet Inception Distance (FID), and AM Score.

Table 6, Accuracy is well represented in more than 75% of the studies we reviewed, whereas F1-score and AUC have a strong presence in studies relying on classification tasks. Furthermore, measures like min support and confidence are used in studies analyzing pattern mining or rule-based decision systems. Structural Similarity Index (SSIM) and Peak Signal-to-noise Ratio (PSNR) are most often used to analyze image quality or performance of segmentation. Overall, accuracy is the most used metric for assessing the performance of models in medical deepfakes detection, but Sensitivity and Specificity are needed to test the medical diagnostic value, while image-based metrics such as FID, SSIM, and Dice, are necessary to test the realism and fidelity of generated images.

Table 5
Evaluation measures reported for medical deepfake detection

Study	Technique Compared	Preeminent Evaluation Metrics
Prezja et al. [1]	GAN (PGGAN)	**Accuracy**
Kim et al. [7]	CNN vs Human Radiologists	AUC, ROC
Patel et al. [6]	GAN (AttGAN, GDWCT, StyleGAN)	**Accuracy**
Alhalabi et al. [9]	CNN (VGG16, ResNet)	Accuracy, Precision, Recall, F1 Score
Parate & Jain [2]	YOLO, ZSL, GAN, CNN	Accuracy
Grabovski et al. [5]	Diffusion Model	AUC
Alsabbagh & Al-Kadi [13]	DenseNet variants	Accuracy, Precision, Recall, F1-score
Albahli & Nawaz [15]	CNN Hybrid	Accuracy, Sensitivity
Waqas et al. [16]	PGGAN + Self Attention	AM Score, Inception Score, FID
Phipps et al. [17]	GANs, CNNs	Accuracy, Sensitivity, Specificity (relative %)
Vardhan et al. [18]	CNN Ensemble	Accuracy, AUC, ROC, F1-score
Budhiraja et al. [19]	Convolution al Reservoir Networks (CoRN)	Accuracy, F1-score, Sensitivity, Specificity

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Al-Emaryeen et al. [20]	CycleGAN + U-Net	Dice, Jaccard, Sensitivity, Specificity
Pradeepan [21]	EfficientNet + DWT Hybrid	Accuracy, Precision, Recall, F1-score
Thambawita et al. [22]	Pulse2Pulse + WaveGAN	Accuracy, Sensitivity, Specificity
Dolhansky et al. [28]	Deepfake Video Detection Models (DFDC)	Log Loss, Precision@k
Zhu et al. [29]	GAN-based anonymizati on	AP@0.95, AUC, AR@0.95
Agarwal et al. [30]	CycleGAN, CNN	Accuracy (99.8%–98%)

C. .DEEPFAKE MEDICAL IMAGES DETECTION TECHNIQUES (RQ3A)

Several techniques have been suggested for detecting deep fakes in medical imaging, and these can generally be categorized into six main groups: Deep Learning Models, Traditional Machine Learning Techniques, Hybrid or Ensemble Approaches, Anomaly Detection Models, Detection through Generative Augmentation, and Zero- or Few-Shot Learning Methods. Figure 2 shows breakdown of these category types in the current literature. Deep learning models, and more specifically CNNs emerged as the most dominant category, representing just under 40% of all identified techniques for identifying different medical images.

A variety of deep learning architectures have been used to identify medical deep fakes, including ResNet50V2, DenseNet169, EfficientNet-V2 (also known as DeepMind), VGG16, Mobile Net–V 2, U-Net, InceptionV3, custom-designed CNNs and modified ResNET variants. CT-GAN, LIDC-IDRI, and DFDC datasets. [3],[9], [10], [13], [22].

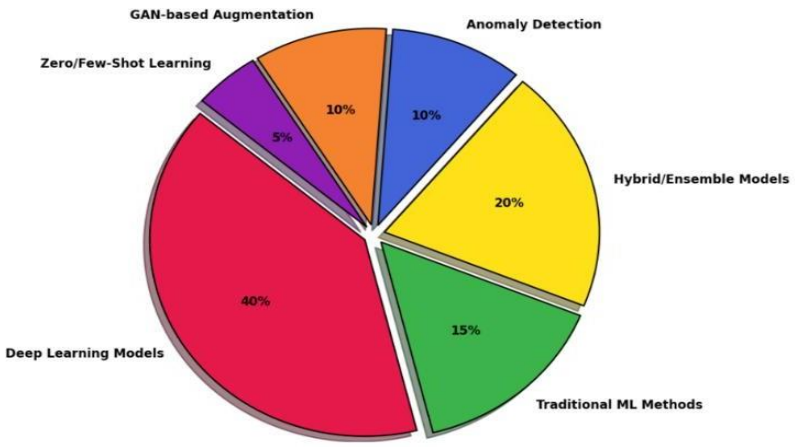


Figure 4. Distribution of deepfake detection techniques.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

The use of traditional machine learning algorithms like Random Forest (RF), Decision Trees (DT), and Support Vector Machines(SVM) has been proven to be effective, especially in situations where deep learning models may overfit [14], [25].
Hybrid approaches that merge various models, such as DenseNet121 with VGG16, EfficientNet-B0 and Discrete Wavelet Transform (DWT), CNNs with Reservoir Computing, and CNNS with GAN-based discriminators, have been developed to capture spatial and statistical features, often achieving detection accumulation peaks above 90%. [11], [18], [19], [21].
Unsupervised approaches, such as back-in-time diffusion models, have been developed to detect manipulated content without having to use annotated datasets. New or unidentified data types can be incorporated into these techniques using newer, more efficient methods. This makes them particularly promising.[2], [5].
Several GANs, including CT-GAN, StyleGNA and PGGANA, have been widely adopted for their ability to synthesize manipulated medical images; these are used in training contexts to increase model robustness and performance [1], [16], [20].
Zero-Shot and Few-Sound Learning-based techniques have been suggested as methods for identifying deep fakes with limited labeled examples. Certain models have shown high accuracy rates in detecting medical image manipulation using adaptations of YOLO [2], [24].

In addition to algorithmic approaches, several studies have highlighted the importance of ethical and conceptual frameworks, such as digital watermarking, blockchain- based validation, and integration with clinical systems like Picture Archiving and Communication Systems (PACS), to provide support mechanisms for maintaining data integrity and trust. Researchers have employed a range of datasets, including publicly available ones (such as LIDC-IDRI, CT- GAN, DFDC and BraTS), and proprietary datasets, such as custom-built collections of CT and MRI data for anomaly detection and simulation. Benchmarking often relies on publicly available, high-quality datasets, but training detection models depend on synthetic data produced with GANs when real data is rare. Why is this necessary? Even though certain models are near-perfect, with a 99.8% accuracy rate, there are still matters with their generalizability, real-time deployment, and interpretability in clinical settings. Therefore, future research should focus on developing explainable models that can function across imaging modalities, operate in real-time, and meet the criteria for clinical validation and data governance.

D. *SUPERIOR DEEPFAKE DETECTION APPROACHES FOR MEDICAL IMAGES (RQ3B)*

As the medical imaging industry faces evolving challenges due to the advanced deep fake technology, there is a critical lack of high precision and reliable detection systems. This research focuses on the best detection methods and considers thirty published works for purposes of evaluation. The approaches explored fall into six primary categories. deep learning techniques, hybrid or ensemble systems, conventional machine learning classifiers, detection methods enhanced by generative data augmentation, anomaly-based detection approaches, and zero-shot/few-short learning framework. Of all these categories, deep learning along with amalgam models are particularly powerful when tailored to particular imaging areas in medicine. Alongside those mentioned above, traditional SVMs and Random Forests showed promise in small dataset conditions. These simpler algorithms outperformed deep networks under limited data [14], [25]. Still, their ability to perform-diminished as

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

dataset size or augmentation increased, where deep models gained the upper hand.

Deep Learning powered architectures ResNet50V2, DenseNet169 and EfficientNet-V2 were cited as having great outputs continuously across numerous research works. DenseNet169's performance claimed to achieve 97.86% accuracy on CT-GAN datasets, [13] while variants of ResNet consistently reported over 99% accuracy across multiple GAN-generated datasets [6], [22]. VGG16 was selected for high recall detection in cancer CT scans [9]. Models such as DenseNet121 combined with VGG16 and EfficientNet-B0 fused with Discrete Wavelet Transform (DWT) achieved remarkable precision, hitting accuracy levels of around 99.6 percent. The models in consideration effectively integrate spatial features and frequency-domain characteristics, which improves robustness in detection. Also powerful were the approaches that utilized Dense Net-based GAN discriminators for CT, X-ray, and MRI image manipulation; those techniques were quite effective at detecting fine-grained alterations [11]. Furthermore, patch-level CNN designs demonstrated training accuracy close to perfect for modified images of skin cancers [3]. In real-time scenarios, recall metrics unprecedentedly peaked at 99.7% with YOLOv5su [24].

In the realm of anomaly-based detection, Back-in-Time diffusion frameworks demonstrated notable success localizing forged tumor areas with AUC scores of 0.90 and 0.96 [5]. This highlights the effectiveness of unsupervised strategies within environments where there is a scarcity of annotated data.

Table 6
Superior Deepfake detection techniques

Model	Accuracy	Dataset Used	References
ResNet50V2	99.1%	CT-GAN	[25]
DenseNet169	97.86%	LIDC-IDRI	[13]
EfficientNet-V2+ Attention MedNet	85.49%	CT-GAN	[15]
DenseNet121+ VGG16	90%	CT Lung Scan	[18]
EfficientNet-B0+ DWT	99.6%	CT-GAN	[21]
DenseNet- Enhanced GAN	Better than traditional GANs	CT, X-ray, MRI	[11]
Patch-based CNN (Skin Cancer)	~100% (training), 68% (user study)	Skin Cancer Dataset (with diffusion fakes)	[3]
YOLOv5su (ZSL)	Recall 99,7%	X-rays, CT-GAN	[24]
Back-in-Time Diffusion	AUC: 0.90–0.96	Custom CT & MRI fake datasets	[5]

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Random Forest (ML)	98.7%	LIDC-IDRI	[14]
SVM+ResNet	99.1%	CT-GAN, LIDC-IDRI	[22]

Approaches focusing on Zero-Shot Learning (ZSL) and few-shot learning also recorded significant outcomes.ZSL systems using YOLO performed exceptionally well achieving near 99.7% accuracy even in the absence of heavily curated datasets underpinned by labeled data [2]. Generalizability and interpretability limitations problem still persists indicating there is more research to be done for future refinement.

As outlined previously, deep fake detection methods that work best combine state-of-the-art Convolutional Neural Networks with attention-based feature extraction, combination models, and other neural nets boosts which are specially tailored to the task at hand. Combined approaches that involves Dense Net with GANs, Efficient Net with DWT (Discrete Wavelet Transform), ResNet50V2 and Back-in-Time diffusion models have shown to surpass traditional techniques more reliably in accuracy and flexibility. Yet even with such precise frameworks as described above, there still exists a lack of defined benchmarking standards which poses limitations within clinical validation cycles, this indicates an important gap for further development towards practical use.

E.POTENTIAL CHALLENGES HIGHLIGHTED IN EXISTING STUDIES TO BUILD A ROBUST AND CLINICALLY APPLICABLE DEEPFAKE DETECTION MODEL FOR MEDICAL IMAGES (RQ4)

Developing deep fake detection systems that are clinically useful and trustworthy captures challenges discussed in the literature. One of the most significant gaps is the lack of clinical validation. Many sophisticated algorithms undergo evaluation only in controlled academic settings; none have been incorporated into hospital workflows or integrated into systems like PACS for real-time clinical imaging analysis [9], [10], [12]. This gap poses concerns regarding their operational integration in ever-changing healthcare environments.

Another prominent issue is the limited adaptability of numerous models. Algorithms built using specific datasets often struggle to process images from other modalities (e.g., MRI, CT, or fundus photography) or those captured at other institutions and from different patients [2], [3], 5]. This undermines robustness and directly impacts their widespread use.

Meeting real-time requirements poses an additional challenge. High-performing models based on diffusion methods or hybrid convolutional neural networks often need considerable computing power, which prevents their use in time-sensitive clinical workflows [5], [21]. This problem is amplified by the issue of model explain ability. Trust within the medical community can be hard to earn with applications that rely on deep learning due to its vague architecture and lack of decision transparency [2], [24]. Detection capabilities face new hurdles due to the advancement of generative techniques. With frameworks like Jekyll, there are alarming security and trust issues as it is possible for GAN's to inject deceptive and convincing diagnostic evidence into medical scans [4]. Alongside these technological concerns lie moral and legal problems such as the ethical concerns regarding the dual-use potential of generative tools. Recent studies emphasize there is an urgent need to address image integrity and misuse through regulatory

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

frameworks, digital watermarking, and block chain-based validation systems [8], [27], [30].
Another limitation is the lack of variety and scale within the training datasets. The majority of models are built on small, homogeneous datasets, which greatly limits their ability to cope with real-world variability and undermines their adaptability [13], [14], [23]. Lastly, insufficient established benchmarking frameworks slows other researchers from quantifying cross-study model performance in a unified way because there is no standard way to measure and benchmark outside the study's framework [17], [25], [28].
Addressing these matters is essential for shifting deep fake detection research towards practical medical usage. These technologies can become safe and reliable diagnostic tools if clinical validation, generalizability, explainability, ethical safeguards, and standardized evaluation frameworks are integrated.

F.Which type of dataset has been used for this research (Professional or Self-Acquired) (RQ5a)

An important issue highlighted in the systematic review of medical imaging deep fake detection refers to the sparse and inconsistent nature of precisely assembled public medical image datasets. During the review, it was evident that many researchers relied on certain datasets like LIDC-IDRI, CT-GAN, BraTS, DFDC, and OAI as they are well known for their quality within the medical imaging research community. However, in many instances, researchers did not mention dataset origins or resorted to using datasets created by PGGAN, CycleGAN, and StyleGAN aimed at anomaly detection and augmentation [1], [5], [16], [20].
Even when dataset provenance was provided, critical elements such as collection methods, ethical clearances obtained, or geographic origin were often missing. A number of authors explained these gaps citing concerns for patient privacy shielding data deemed too sensitive alongside institutional restrictions for revealing clinical photographs and images [4], [8], [23]. This review exhibited an overreliance on a small number of primary datasets such as LIDC-IDRI and CT-GAN which featured prominently in CT-based anomaly detection studies while other modalities like MRI ultrasound medical video datasets were relatively underrepresented in the available body of literature [9], [10], [18].
The goal of this systematic literature review (SLR) is to compile and present a detailed categorized analysis of datasets pertaining to medical deep fake detection. This includes professionally curated as well as synthetically generated datasets through GAN-based techniques. A structured summary capturing types, contexts of usage, and sources of the datasets is presented in Table 1. The analysis also indicates LIDC-IDRI and CT-GAN as the most popular datasets used for benchmarking performance. Others like OAI and DFDC serve more focused purposes, such as in the analysis of knee joint X-rays or the anonymization of surgical video documentation [1], [13], [28]. Figure 3 captures the breadth of imaging modalities represented by various studies including CTs, MRIs, X-rays, ECGs, and fundus imaging along with their respective datasets. These were either gathered from publicly available repositories or generated using methods such as GANs. With few prominent labeled medical deep fake datasets accessible, many researchers ended up creating them for augmentation and anomaly detection tasks.
As evidenced within other critical fields, there exists an urgent need to develop a diverse pool of ethically sourced standardized medical deep fake datasets. Such resources should include multi-modal imaging and effect data from multiple parts of the body, thus enhance the precision, dependability, and medical incorporation of future medical deep fake detection systems [2], [6],

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

[12], [17].

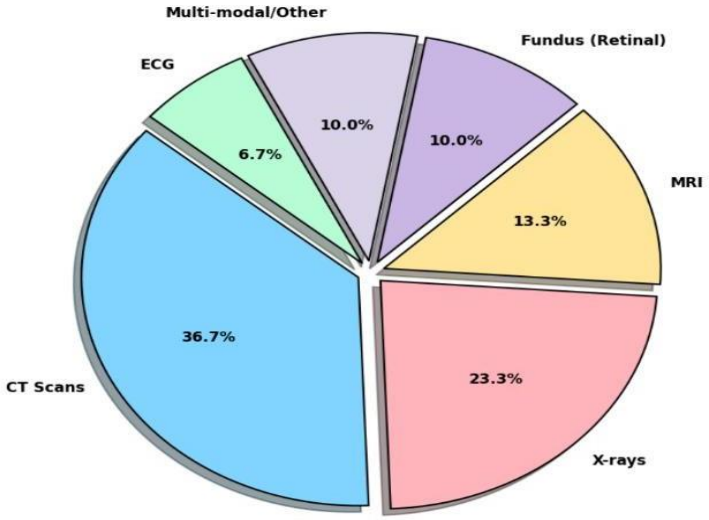


Figure 5. Studies on specific datasets types

G.WHAT ARE THE MAIN ASPECTS OF A DATASET FOR THE DEEPFAKE DETECTION IN MEDICAL IMAGES? DO THEY AFFECT THE RESULTS (RQ5B)

The clinical datasets considered in this systematic review were mostly obtained from well-known repositories or constructed through sophisticated generative techniques like PGGAN, CT-GAN, and CycleGAN. As pointed out by the authors of this review several datasets described in the review lacked essential metadata such as modality-specific annotations, imaging equipment details, and patient age- and-gender information [1], [5], [16]. With the advancement of AI algorithms for image generation and recognition, as we work to improve deep fake detection systems, using comprehensive metadata like modality tags (CT, MRI, X-ray), imaging parameters and detailed change descriptions becomes ever more important. Applying manipulation type with multi-modal imaging has been a recent focus leading to better interpretable and clinically useful models aimed at detecting altered medical content [2], [3], [10].

The authenticity and clinical significance of medical imaging datasets still poses a challenge problem, considering the efficiency of deep fake detection is directly proportional to the quality and authenticity of the data in question. The majority of scientists focus on original medical images in combination with edited versions, teaching algorithms to identify modality-specific intricate distinctions. Dataset volume, diversity in pathologies, resolution metrics, or anonymization degrees also significantly impact the accuracy and robustness of these systems when integrated into actual clinical workflows [6], [7], [17].

In order for datasets to be considered truly useful for the deep fake detection community within medicine, they need to be ethically anonymized: clinically validated across multiple imaging modalities alongside clear and coherent indicators on manipulation. While some datasets are derivatives from hospital PACS systems, others are synthetically created solely for research purposes. Proper curation aligned with patient confidentiality protocols is crucial for ensuring

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

trustworthiness described as responsible within bounds of patient anonymity guidelines strives towards making the data reliable.

Several studies [4], [8], [12] have report out the importance of having standardized evaluation frameworks and protocols for creating datasets. This is crucial for ensuring fair and consistent comparisons across different studies. To make this happen, we need collaborative efforts from medical societies, research institutions, and ethical review boards. These groups can help promote data-sharing initiatives, provide training on ethical data usage, and kick off multi-institutional projects to tackle the shortage of high-quality datasets. By working together in this way, we can significantly enhance the effectiveness, reliability, and clinical integration of deep fake detection technologies in healthcare.

V. Analysis

SLR goal to comprehensively assess and synthesize the current methods developed to detect deep fakes using medical imaging. The review is organized around purposefully constructed research questions, which were used to evaluate the studies we consider using systematic inclusion and exclusion criteria to assess relevance and quality.

Research questions 1 and 2 pertain to the various detection methods used in medical deep fake studies and measures of performance for these approaches. RQ3 explores the main limitations and technical issues shared in existing approaches, which can serve as direction for researchers going forward in this area. Using RQ4, we also explore the importance of medical datasets in deep fake detection as well as the principal factors that are important for training effective, reliable models.

The first step in the development of deep fake detection systems, specifically for medical purposes, is to correctly identify manipulated or synthetically generated content. Certainly, the advancement of generative technologies (including Generative Adversarial Networks (GANs), and diffusion models) has created a dual approach to the generation and detection of synthetic medical images. These models can leverage the different dimensions of spatial and structural information specific to the imaging modality to reliably identify the falsified areas and their relevance for clinical decision making. RQ1 deep dives into the categorization and evaluation of detection and evaluation methods focusing on leading methods and commonly evaluated performance metrics as presented in the existing literature.

Based on the literature reviewed, a broad range of approaches from machine learning and deep learning perspectives have been considered. Within these, classification models especially convolutional neural networks (CNN) and some-offshoot design concepts including GANs, feature a strong likelihood of effectiveness. Most prominently dense and YOLO based networks have been categorized as competitive models with proven accuracy and performance characteristics under in- laboratory settings. While positive progress has been achieved, and continued advancements are expected; there continues to be a gap from a performance perspective for our clinical datasets and therefore warrant to discuss further innovations going forward and improvements to these protocols.

The success of deep fake detection in medicine depends on a model's ability to determine subtle errors and inconsistencies generated through image manipulation. This ability is critical to preserve accuracy in diagnostics and to maintain clinician trust in clinical workflows. Although classic machine-learning-based models and hybrid methods have established a strong precedent, concerns about model generalization across different clinical scenarios remain a potential risk. Recently, promising new architectures such as diffusion models and temporal analysis have emerged that could help mitigate these failings by virtue of their greater flexibility and enhanced

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

detection performance.

To assess and compare different detection approaches fairly, research has relied on a variety of performance measures. The range of performance metrics consist of accuracy, area under the curve (AUC), precision, recall, and F1-score, but checking these measures allows researchers to determine which are as reliable and frequently used as possible so that their work can lay the foundation for future standards. Uniform use of validated performance measures will ensure deep fake detection methods can be benchmarked properly, taking into account transparent and objective comparison within and across studies and encouraging reproducibility.

Accuracy, F1-score, Sensitivity, Specificity, and Area Under Curve (AUC) remain some of the most frequently used metrics used to quantify performance for medical deep fake detection, especially considering classification tasks, and how well a model is able to distinguish manipulated images from real images. On the other hand, concerning model reporting in terms of performance measuring if the quality of images or imaging is synthetically generated in models like GANs or Diffusion, there are several metrics that have seen consistent use including Fréchet Inception Distance (FID), Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), Dice Coefficient and Jaccard Index. These aim to quantify how visually credible and structurally accurate the generated/altered images are on a structural basis. In general, using consistent metrics is important to uphold transparency, reproducibility and credibility in research findings.

(RQ3) is focused on the technology limitations current detection approaches are facing as well as new potential approaches of innovation. This aims to determine where the most important research gaps exist so that people are new participate into the area have more clarity on existing weaknesses thereby allowing them to make available successful contributions. Several important future directions are identified in this review including transferring learning to adapt models to various types of imaging, and improving detection model robustness associated with types of deep fake detection models that are based on CNNs, GANs and Transformer architectures.

TABLE 7.
Dataset and Their Characteristics Used for Deep Fake Detection

Dataset	Characteristics	Modality	GAN-Based
CT-GAN	Contains real/fake CT images; used for image tampering and forgery detection	CT	YES
LIDC-IDRI	Real thoracic CT scans; often paired with CT-GAN	CT	NO
OAI Dataset	Kee X-rays used for generating fake disease	X-rays knees	Yes
BraTS 2020	Brain tumor MRI data; used in segmentation	MRI	Yes
IXI Dataset	Unpaired brain MRI scans; used for modality transfer	MRI	Yes

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

GESUS + Inter99 (ECG)	ECG datasets; used in synthetic ECG generation	ECG	Yes
EyePACS	Fundus images used in Debatic retinopathy	Fundus Retinal	Yes
ProstateX Challange	Prostate MRI data used inImage enhancement GANs	MRI	Yes
DRIVE Dataset	Retinal images used for synthesis and segmentation in VAE-GAN	Fundus	Yes
STARE Dataset	Fundus image used with DRIVE/HRF	Fundus	Yes
HRF Dataset	High-resolution fundus images; used in opthalmology	Fundus	Yes
SPACE Dataset	MRI dataset used for multi-modal deep learning	MRI	Yes
NeuBI	Used in hybrid GAN/VAE	MRI	Yes
ADNI Dataset	Alzheimer's disease-related, MRI used in image	MRI	Yes
SRC Dataset	Reference for fake image generation	Mixed	Yes
Skin Cancer	Used in patched based CNN and stable diffusion model	Skin Dermatology	Yes
Parkinson video dataset	Used in GAN based facial de-identification	Medical video	Yes
DFDC Benchmark	Non medical video dataset used in adapted face swapping	General face video	No

Additionally, we advocate the use of Explainable Artificial Intelligence (XAI) as a strategy to be more interpretable and to enhance clinical trust in the AI-generated tools in terms of their use and purpose. We also recommend developing lightweight and real-time detection models, particularly for integration into clinical systems and workflows through systems like PACS.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

LSTM and 3D CNNs will be particularly valuable for sequential data forms like ECG signals and medical videos to model temporal patterns.

Meanwhile, some important areas to consider are ethical issues on fires involved with synthetic data generation and use; securing images authenticity through embedded digital watermarks or block chain; and the creation of large-scale, heterogeneous, annotated benchmarking datasets in order to train and test the deep fake detection algorithms used in real clinical contexts

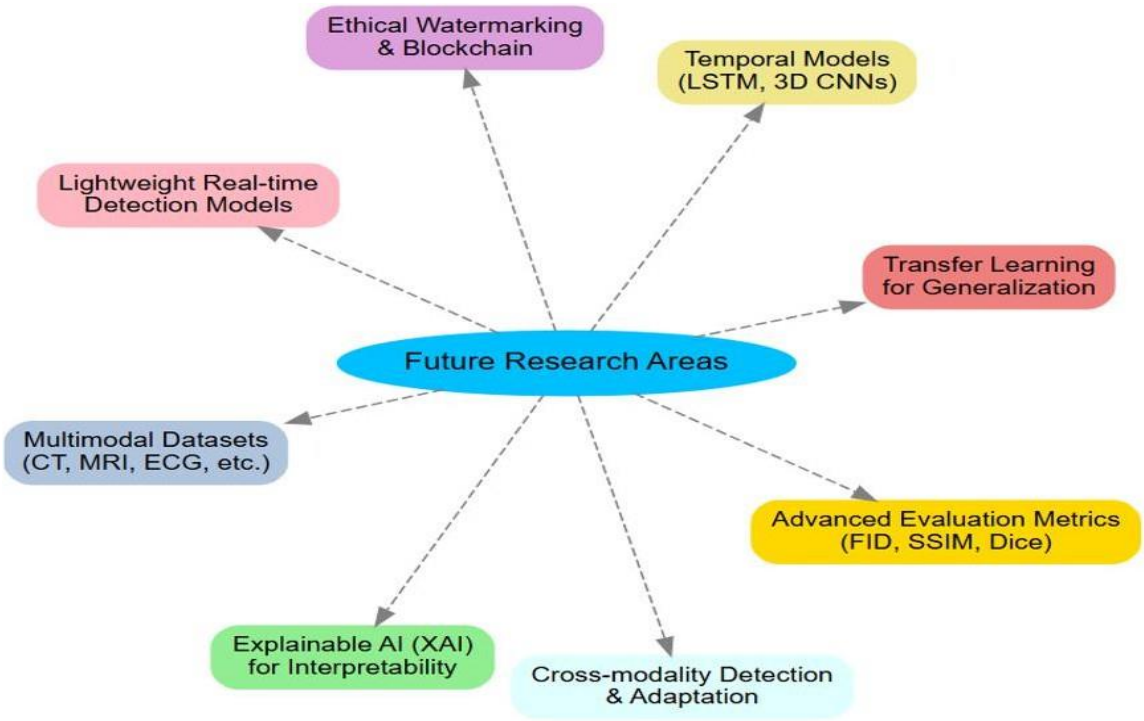


Figure 6. Potential areas for future research

The prerequisite of successful deep fake detection in medical images is the existence of well-labeled and thoroughly-annotated datasets. This review of the literature reveals that common are cases when explicit references to the datasets are missing or when institution-specific datasets are used that cannot be used by a wider audience of researchers. Such lack of willingness to share has been attributed by many researchers to ethical and compliance constraints, patient privacy laws, and restrictions on the part of institutional review boards or healthcare providers. Popular publicly available data in this area include CT- GAN, LIDC-IDRI, BraTS, OAI, IXI, EyePACS, and DFDC. These data sets encompass a diverse set of imaging modalities including, MRI, X-ray, fundus imaging, ECG, and video. Yet they may do so with varying quality of annotations, variation across fakes in type of deep fake manipulation, and extent of patient coverage. In many situations, datasets used are synthetic or modality-specific, leaving the general detection hard to achieve.

To counteract this shortcoming, in this review we examine Research Question 4 (RQ4) that highlights the urgent development of diverse, well-organized, and accessible datasets in order to consider various setting types and manipulation techniques. Not only it would improve reproducibility across applications, but it would also allow fair comparison of algorithm

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

performances. This SLR ends by discussing several challenging aspects concerning the reliability, clinical validity, diversity, and ethical transparency of the current medical data sets. These elements are critical to the performance and robustness of deep fake detection models. To further facilitate research in this sphere, the establishment and open release of rich, ethically curated medical deep fake databases that accurately represent realistic clinical contexts are undoubtedly an imperative.

VI.CONCLUSION AND FUTURE WORK

This study provides a systematic exploration of the existing state of medical deepfake detection, adhering to the systematic literature review (SLR) rules proposed by Keele [31]. To the best of the authors' knowledge, this is the first comprehensive SLR conducted between 2020 and 2025 that focuses completely on the detection of deepfakes in medical imaging. By analyzing 30 peer-reviewed studies crossing various imaging modalities—including CT, MRI, X-ray, ECG, and fundus images—this review consolidates knowledge and offers a structured overview of existing techniques and challenges.

The major objective of this review was twofold. First, it aimed to classify the deepfake detection methods employed in medical imaging and to identify the most accurate and practically applicable techniques. The assumptions reveal that deep learning models—particularly CNN-based architectures such as ResNet50V2, DenseNet169, and EfficientNet-V2—consistently outperform traditional methods, often achieving classification accuracies higher than 97%. Moreover, hybrid frameworks such as DenseNet121 combined with VGG16 and EfficientNet-B0 with Discrete Wavelet Transform (DWT) have demonstrated improved performance by integrating spatial and frequency-domain features. Emerging techniques such as diffusion-based anomaly detection and Zero-Shot Learning (ZSL) also demonstrate promising results, particularly in low-data scenarios and real-time detection tasks.

However, despite high reported accuracies, most of these methods show significant limitations. As detailed in the examination sections and Table 1, many models struggle with generalizability through diverse datasets, lack real-time processing capabilities, and have seen the least for clinical validation. The lack of standardized datasets and evaluation benchmarks further inhibits reproducibility and fair comparison. Additionally, the narrow interpretability of current models—often operating as black boxes— compromise their acceptance in clinical settings.

The second goal of this review was to evaluate the datasets used for training and testing medical deepfake detection models. Though publicly available datasets such as CT-GAN, LIDC-IDRI, and BraTS are commonly used, many studies rely on institution-specific or synthetically generated datasets with narrow transparency or accessibility. Furthermore, MRI and ultrasound modalities stay underrepresented compared to CT and X-ray images. The absence of metadata, inconsistent annotation protocols, and limited modality diversity pose extra challenges to clinical integration.

In conclusion, this SLR confirms that deepfake detection in medical imaging is an emerging and important area of research that demands advance innovation. Whereas current methodologies show assurance in controlled environments, their applicability in real-world clinical settings remains limited due to issues of generalization, interpretability, and standardization. Addressing these limitations requires a shift in focus toward the development of clinically robust and trustworthy systems. Future research should highlight the development of Explainable AI, Lightweight, real-time models, Ethically sourced, diverse, annotated datasets, Hybrid temporal models (e.g., LSTM, 3D CNNs) for detecting deepfakes in video and time-series medical data such as ECGs, Security-integrated methods and Standardized evaluation protocols and

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

benchmarking frameworks for fair and reproducible comparisons. Ultimately, collaboration among AI researchers, medical professionals, ethicists, and policymakers is imperative to ensure the growth of deepfake detection systems that are accurate, transparent, secure, and clinically viable.

Table
Modal and key characteristics

NO	Year + Citation	Proposed Model	Key Characteristics	Accuracy	Limitation	Future Direction	Dataset
1	2025 + [17]	GANs, Autoencoders, Diffusion Models	Review of generative model in Ophthalmology	13% increase	Bias, high computation, misuse	Improve explainability, data authentication	Ophthalmology Images
2	2025 + [26]	BPGAN, SD-GAN, MSG-CapsGAN, IGAN	GANs for medical image enhancement, segmentation, synthesis	SSIM, PSNR, Dice (high)	Mode collapse; lack of clinical deployment	Clinical validation; explainable AI	BraTS,LIDC, ProstateX, EyePACS
3	2025 + [09]	CNN Architectures (VGG16, EfficientNetV2, InceptionV3, Sequential)	Compared multiple CNNs to detect deep fakes in CT scans; VGG16 selected for best recall	93%	Limited to CNNs; no real-world testing; lacks use of transformer/hybrid models	Use more diverse datasets; integrate security in PACS; explore real-time application & advanced architectures	CT scan images (resized 256×256)
4	2025 + [11]	DenseNet-enhanced GAN Model	Uses DenseNet-based discriminator for improved fake image detection in medical images (CT, X-ray, MRI)	N/A	No focus on specific diseases, lacks real-time detection evaluation	Future models using transformer-based architectures, hybrid models, and real- time detection optimization	Medical images (CT, MRI, X-ray); dataset not specified

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

5	2024 + [13]	13 DCNN Models (e.g., DenseNet169 , ResNet50V2)	Evaluated 13 pre-trained CNNs for medical deep fake detection	97.86%	Overfitting , dataset limitations, lack of explainabili ty	Expand DCNN s, explore hyperparameter tuning	CT-GAN (LIDC -IDRI)
6	2024 + [18]	VGG16+ DenseNet121 Ensemble	Transfer learning + data augmentation	90%	Lower performan ce of individual models	Improve reliability in divers e conditions	CT Lung Scans (1,931 images)
7	2024 + [20]	CycleGAN + U- Net	Deep fake generation for tumor segmentation	0.87	Smalllesio n performan ce, noise sensitivity	Larger dataset s, real-world trials	BraTS 2020, IXI, Unpaired MR- CT
8	2024 + [21]	EfficientNet-B0 + DWT Hybrid	Combines spatial + frequency features	99.6%	CT-specific testing only	Test on diverse image types	CT-GAN
9	2024 + [23]	GANs, VAEs, Hybrid (VAE-GAN)	Medical image synthesis, modality transfer, augmentation	**N/A**	Detection research underrepres ent ed (only 2%)	Develop robust detection frameworks	DRIVE, ADNI, STARE, LIDC-IDRI, HRF, NeuB1, SPACE, SRC
10	2024 + [24]	YOLOv3–YOLOv8x	Real-time deep fake detection in medical images (X-ray, CT)	**0.997**	Limited fake data in CT, no DICOM handling	Expand datasets; clinical integration	CT-GAN, Osteoarthrit is X- rays
11	2024 + [27]	Ethical application models (avatars, GANs)	Conceptual use of deep fakes for therapy, education, telemedicine	**N/A**	No real-time systems; under-regulated space	Ethical frameworks, blockchain verification	Conceptual (no dataset)
12	2024 + [30]	GANs, VAEs,	Active/passive detection	**CycleG AN: 99.8%;**	Limited real-world	Passive detection models; legal	SPACE, ADNI, HRF,

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

		CNNs, PCA+SVM	approaches; synthesis risk evaluation	**CNNs: 80–98%**	application; ethical misuse	frameworks	NeuB1, STARE, SRC, LuNoTim
13	2024 + [02]	Zero-Shot Learning (ZSL)	Works with very little labeled data; adaptable to new/unseen fake images	**99.7%**	Hard to interpret (black-box) - Poor generalization across scan types Not real-time ready	Improve ZSL embeddings - Add Explainable AI for clinical use and trust	Not specified in the paper (review-based)
14	2024 + [03]	CNNs and Patch-Based Neural Networks	Uses deep learning to detect fake skin cancer images generated via stable diffusion (not just GAN)	**~100% (Training & Validation), 68% (User visual study)**	Thresholding is hard due to distribution shifts; challenging for visual detection; possible overfitting in training	Improve generalization and robustness; explore further detection beyond histogram and patch-level features	Publicly available skin cancer dataset, with fake images generated using stable diffusion
15	2024 + [05]	Back-in-Time Diffusion	Uses reverse diffusion process for unsupervised	**0.90 (injection), 0.96 (removal)**	Real-world deployment and generalization on diverse	Further validation, open-source code, encourage more research	Custom CT & MRI deep fake datasets
			medical deep fake detection		datasets not discussed		
16	2024	Improved	Uses a dense CNN model to detect	**98.33% (AttGAN), 99.33% (GDWC**	May not capture temporal	Extend to video deep fakes	Seven datasets with 5000 fake

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

	+ [06]	Dense CNN (D-CNN)	deep fake images from various sources	**T), 95.33% (StyleGAN), 94.67% (StyleGAN2), 99.17% (StarGAN)**	inconsistencies in videos, only works on images	and explore other optimization techniques	and 10000 real images (from GANs)
17	2024 + [08]	Not model-based (Systematic Literature Review)	Reviews how GANs can be used both positively (training AI) and negatively (malicious manipulation) in medical imaging	**N/A**	No practical detection framework proposed; purely conceptual and policy-oriented	Calls for development of policy, legal safeguards, and use of techniques like watermarking/block chain to detect deep fakes in medical data	No specific dataset used (theoretical /literature-based)
18	2024 + [12]	GAN + CNN + Anomaly Detection based Deep fake Detection System	Uses GANs For mage synthesis, CNNs for classification, and nomaly detection techniques to identify manipulated medical images (e.g., CT scans).	**N/A**	Not yet tested in real clinical settings; lacks evaluation on diverse imaging modalities	Deploy system in hospital settings; expand to other medical imaging types like MRI, X- ray	Not specified
19	2023 + [15]	EfficientNetV2 + Dense Layers	Attention mechanism highlights tampered regions	**85.49%**	Dataset bias, attention complexity	Applyto MRI, ultrasound	CT-GAN
20	2023		Generates realistic Deep fake knee X-ray images with different levels of	**96.21%**	Limited to OAI dataset, lacks	Expand dataset variety, improve detection methods, and explore clinical	OAI (Osteoarthritis Initiative)

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

		CT-GAN			generalization, no fake image detection	applications with deeper testing	
	+ [01]	(Conditional GAN)	osteoarthritis by injecting/removing disease features		Method and has ethical/ privacy concerns		
21	2023 + [10]	Enhanced ResNet101, ResNet50, DenseNet121 /20 1, MobileNet/V 2	Evaluated 6 CNNs on 3- class data (Untampered, False Benign, False Malicious); ResNet101 achieved best results	**99%**	Focus on synthetic data only; real-world deployment not tested ; limited to classification task	Expand to real-time PACS integration; test on real-world datasets; adapt to evolving deep fake techniques	2000 images, 3 classes (Untampered, False Benign, False Malicious)
22	2022 + [14]	5 Deep Learning + 3 ML Models (DenseNet, ResNet, RF, SVM)	Comparative performance on tampered CT scans	**97.9%**	Limited data, overfitting, clinical generalization	Integration with PACS, use diverse clinical data	LIDC-IDRI
23	2022 + [16]	Enhanced-GAN (PGGAN + self-attention)	Improved knee image synthesis	**AM S core: 3.01**	Image quality artifacts, data dependence	Applyto other medical tasks	OAI Knee MRI, CelebA
24	2022 + [19]	Convolutional Reservoir Networks (CoRN)	Lightweight CNN + Reservoir Computing	**>90%**	Small dataset	Broader applications, dataset expansion	LIDC-IDRI + CT-GAN
25	2022 + [25]	SVM, RF, DT, ResNet,	ML/DL model comparison for CT-GAN deep	**ResNet5 0: 99.1%;**	Small datasets; weak raw image	Explore 3D models; real-time solutions	LIDC-IDRI, CT-GAN

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

		DenseNet, VGG	fakes	**RF: 98.7%**	performanc e		
26	2022 + [04]	Jekyll (GAN-based Style Transfer Framework)	Translates real biomedical images (e.g., X-ray, retinal) to fake ones with attacker- chosen disease indications	**N/A**	Lacks defense generalizati on; targeted at specific modalities (X- ray, retina); no full-scale dataset benchmark	Explore robust AI defenses, improve detection of GAN- based manipulations, expand to more modalities	Biomedical images (X-ray, retinal fundus)
27	2022	Deep learnin g- based manipulation detection model using U-Net	Detects manipulated medical images (fundus) - Uses U-Net and CycleGAN -	**Model AUC: 0.913 Ophthal molo gist: 0 .72 General doctor: 0.67**	- Only fundus images used - Not tested on real-time clinical settings - Small sample	Extend detection to other medical data types Real-world hospital testing	356 right fundus images (Normal, Diabetic Retinopath y, Glaucoma, Macular Degenerati on)

	+ [07]		Verified by doctors and ophthalmolo gi sts - Evaluation through AUC and ROC		Size		
28	2021 + [22]	Pulse2Pulse + WaveGAN	Generates synthetic ECGs fo r privacy	Sensitivity: 99.9%, Specificity: 100%	Focused on normal ECGs only	Generate abnormal ECGs, apply to EEG/EMG	GESUS + Inter99 (7,23 3 ECGs)

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

29	2020 + [28]	DFDC dataset; detection benchmark models	128K face-swap videos; public benchmark for detection	Log loss: 0.4279; Precision@0. 1: 0.9843	Not focused on healthcare; no baseline model	Adapt DFDC to healthcare; release medical variants	DFDC
30	2020 + [29]	Faceswap-based deep fake de-identification	Anonymizes medical videos (e.g., Parkinson's) while preserving facial/body keypoints	AP@0.95: 0.99; AR@0.95: 0.998; ROC AUC: >0.9999	Fails on side profiles or multi-person videos	Improve multi-face tracking; preserve hair/clothing	Parkinson's videos + DFDC (Google/Jigsaw)

REFERENCES

[1] Prezja F, Paloneva J, Pölönen I, Niinimäki E, Äyrämö S. Deep fake knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification. Sci Rep. 2022;12(1):18573.

[2] Parate MRA, Jain K. Advancements in fake medical image detection: A comparative analysis of YOLO, GAN, CNN, and zero-shot learning approaches. Cuest Fisioter. 2025;54(4):6618-25.

[3] Arshed MA, Mumtaz S, Gherghina ȘC, Urooj N, Ahmed S, Dewi C. A deep learning model for detecting fake medical images to mitigate financial insurance fraud. Computation. 2024;12(9):173.

[4] Mangaokar N, Pu J, Bhattacharya P, Reddy CK, Viswanath B. Jekyll: Attacking medical image diagnostics using deep generative models. In: 2020 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE; 2020. p. 139-57.

[5] Grabovski F, Yasur L, Amit G, Mirsky Y. Back-in-time diffusion: Unsupervised detection of medical deep fakes. arXiv. 2024; arXiv:2407.15169.

[6] Patel Y, Tanwar S, Bhattacharya P, Gupta R, Alsuwian T, Davidson IE, et al. An improved dense CNN architecture for deep fake image detection. IEEE Access. 2023;11:22081-95.

[7] Kim Y, Song H, Han J. A deep fake-based deep learning algorithm for medical data manipulation detection. J Syst Manag Sci. 2022;12(1):13-24.

[8] Waier J, Shillair R. Deepfaking medical images: Eroding trust in medical diagnosis. SSRN. 2024;4909781.

[9] Alhalabi DT, Alawida M, Chikhaoui B, Hamadeh HS. Deep fake detection in cancer medical imaging using CNN architectures. In: International Conference on Web Information Systems Engineering. Singapore: Springer; 2024. p. 299-312.

[10] Alsaheel A, Alhassoun R, Alrashed R, Almatrafi N, Almallouhi N, Albahli S. Deep fakes in healthcare: How deep learning can help to detect forgeries. Comput Mater Continua. 2023;76(2).
[11] MeenaPrakash R, Kamali B, Vimala M, Madhuvandhana K, Krishnaleela P. A DenseNet-enhanced GAN model for classification of medical images into original and fake. In: 2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI). IEEE; 2025. p. 1540-5.
[12] Gowda BN, Nandeshwar D, Raju DC, Hadadi MS. Deep-fake detection for medical images: A survey. Int Adv Res J Sci Eng Technol. 2024;11(5). doi:10.17148/IARJSET.2024.11595.
[13] Alsabbagh AR, Al-Kadi O. Comparative analysis of deep convolutional neural networks for detecting medical image deep fakes. arXiv. 2024; arXiv:2406.08758.
[14] Solaiyappan S, Wen Y. Machine learning based medical image deep fake detection: A comparative study. Mach Learn Appl. 2022;8:100298.
[15] Albahli S, Nawaz M. MedNet: Medical deep fakes detection using an improved deep learning approach. Multimed Tools Appl. 2024;83(16):48357-75.
[16] Waqas N, Safie SI, Kadir KA, Khan S, Khel MHK. Deep fake image synthesis for data augmentation. IEEE Access. 2022;10:80847-57.
[17] Phipps B, Hadoux X, Sheng B, Campbell JP, Liu TA, Keane PA, et al. AI image generation technology in ophthalmology: Use, misuse and future applications. Prog Retin Eye Res. 2025;101353.
[18] Vardhan GN, Sastry VS, Cherukumalli L. Identifying deep fakes in CT scans of lung cancer using an ensemble architecture. In: 2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC). IEEE; 2024. p. 1-6.
[19] Budhiraja R, Kumar M, Das MK, Bafila AS, Singh S. MeDiFakeD: Medical deep fake detection using convolutional reservoir networks. In: 2022 IEEE Global Conference on Computing, Power and Communication Technologies (GlobConPT). IEEE; 2022. p. 1-6.
[20] Al-Emaryeen RA, Al-Nahhas S, Himour F, Mahafza W, Al-Kadi O. Deep fake image generation for improved brain tumor segmentation. In: 2023 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT). IEEE; 2023. p. 6-11.
[21] Pradeepan P. Detection of deep fake medical images based on spatial and frequency domain analysis. In: 2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN). IEEE; 2024. p. 611-7.
[22] Thambawita V, Isaksen JL, Hicks SA, Ghouse J, Ahlberg G, Linneberg A, et al. Deep fake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. Sci Rep. 2021;11(1):21896.
[23] Lakshmi D, Hemanth DJ. An overview of deep fake methods in medical image processing for health care applications. Des Stud Intell Eng. 2024; p. 304-11.
[24] Karaköse M, Yetış H, Çeçen M. A new approach for effective medical deep fake detection in medical images. IEEE Access. 2024.
[25] Yoon J, Panizo-Lledot A, Camacho D, Choi C. Triple-modality interaction for deep fake detection on zero-shot identity. Inf Fusion. 2024;109:102424.
[26] Hussain J, Båth M, Ivarsson J. Generative adversarial networks in medical image reconstruction: A systematic literature review. Comput Biol Med. 2025;191:110094.

[27] Kaur A, Noori Hoshyar A, Wang X, Xia F. Beyond deception: Exploiting deep fake technology for ethical innovation in healthcare. In: Proceedings of the 1st International Workshop on Multimedia Computing for Health and Medicine. 2024. p. 70-8.
[28] Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, et al. The deep fake detection challenge (DFDC) dataset. arXiv. 2020; arXiv:2006.07397.
[29] Zhu B, Fang H, Sui Y, Li L. Deep fakes for medical video de-identification: Privacy protection and diagnostic information preservation. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020. p. 414-20.
[30] Agarwal S, Peta S, Panyam S. Deep fakes in healthcare: Reviewing the transformation potential and its challenges. Int J Intell Syst Appl Eng. 2024;12(4):3965-70.
[31] Keele S. Guidelines for performing systematic literature reviews in software engineering. EBSE Tech Rep. 2007;2.3:1-57.
[32] Weidt F, Silva R. Systematic literature review in computer science: A practical guide. Relatórios Técnicos Do DCC/UFJF. 2016;1(8):1-7.
[33] Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering. EBSE Tech Rep. 2007.
[34] Kitchenham B.Procedures for performing systematic reviews.Keele Univ Tech Rep. 2004;0400011T.1:1-12