Jurnal Informatika dan Rekayasa Perangkat Lunak ISSN 2656-2855
Vol. 8, No. 1, March 2026, Page. 102-111 e-ISSN 2685-5518

A Systematic Review of Machine Learning and Deep Learning Techniques for Deepfake Image Detection: Trends, Challenges, and Future Directions

Samuel Rhesa[1*], **Aditiya Hermawan**[2]
[1,2]Department of Informatics, Faculty of Science and Technology, Universitas Buddhi Dharma
[*]Email: samuel.rhesa@ubd.ac.id

Abstract

The rapid development of deep learning based face manipulation techniques has produced synthetic images that are increasingly realistic and visually indistinguishable from authentic ones. The deepfake phenomenon poses serious challenges to digital information authenticity and cybersecurity. This research presents a Systematic Literature Review (SLR) of publications from the 2020–2025 period to map trends, methodological approaches, and key challenges in machine learning and deep learning based image deepfake detection. Through an analysis of 24 empirical studies, this review identifies a shift in research direction from conventional convolutional architectures toward hybrid and attention based approaches that emphasize efficiency, adaptivity, and cross domain generalization. Findings show that although recent models such as Vision Transformer and hybrid CNN–LSTM are capable of achieving high accuracy under controlled conditions, their performance remains limited when tested on new domains. Key challenges identified include limited generalization against new manipulation types, vulnerability to image distortion and compression, and low transparency in model decision-making. This study fills research gaps by providing a comprehensive methodological map of architectural evolution, feature representation strategies, and evaluation metrics. Theoretically, this research expands the understanding of deepfake detection research dynamics, while practically, the results provide direction for developing adaptive, transparent, and efficient detection systems for real-time implementation.

***Kata kunci**: Deepfake, Deep Learning, Image Detection, Machine Learning*

INTRODUCTION

The rapid advancements in deep learning-based facial manipulation have produced synthetic images that are increasingly realistic and visually indistinguishable from authentic ones. This development presents a significant challenge to the field of computer vision, as deepfake imagery can effectively mask manipulation artifacts that were previously detectable by conventional detection methods (Guarnera et al., 2020). This phenomenon also raises serious ethical and social implications, such as the spread of misinformation, reputation tarnishing, and manipulation of public opinion, thus encouraging the scientific community to develop artificial intelligence-based detection approaches (Hsu et al., 2020). As attention to such threats increases, various studies have begun to focus on the development of methods for deepfake image detection by utilizing machine learning and deep learning.

Approaches based on CNN emphasize spatial features to recognize facial manipulation patterns (Sharma et al., 2024), whereas temporal approaches using LSTM are developed to detect inconsistencies between frames (Sagar & Arukonda, 2025). The subsequent development shifts to Transformer architectures that rely on attention mechanisms to obtain more comprehensive feature representations (Q. Liu et al., 2024), as well as exploration of the frequency domain that highlights hidden patterns through spectral analysis (X. Liu et al., 2022).

Although prior research has produced significant contributions, existing literature remains fragmented and has yet to present a conceptual relationship between architecture, feature domains, datasets, and the evaluation strategies used. In addition, differences in dataset characteristics such as FaceForensics++, Celeb-DF, and DFDC complicate performance comparisons across

studies. Based on these conditions, this Study conducts a Systematic Literature Review (SLR) to (i) analyze developmental trends in deepfake image detection approaches based on machine learning and deep learning, as well as (ii) map the methodological landscape covering architecture, feature domains, datasets, and evaluation metrics used. This study is expected to provide conceptual contributions in the form of a comprehensive mapping of the methodological landscape, as well as practical contributions as a basis for developing detection methods that are more adaptive and accurate against the ever-evolving complexity of visual manipulation.

LITERATUR REVIEW

The deepfake phenomenon is a result of deep learning technology advancements that enable realistic facial image manipulation through Generative Adversarial Networks (GAN) and Autoencoder algorithms. This system produces synthetic images that closely resemble original faces, thus posing serious challenges to digital media authentication and visual information security (Guarnera et al., 2020; Hsu et al., 2020). Deepfake detection efforts are then focused on the identification of synthetic artifacts such as texture inconsistencies, color distribution, and lighting (Guo et al., 2021).

Deepfake image detection approaches can be classified into three main categories, namely CNN based, Transformer based, and Hybrid or Ensemble based. The CNN approach became the initial foundation due to its ability to extract spatial patterns locally (Lee et al., 2021). However, the limited receptive field makes it difficult for CNN to capture global context, causing its performance to often decline on different datasets. As an alternative, Transformer architectures such as Vision Transformer (ViT) and Swin Transformer use self attention mechanisms to understand long range dependencies between image patches (Alsolai et al., 2025; Khormali & Yuan, 2022).

Although more adaptive to domain variations, these models demand high computational resources. Meanwhile, Hybrid and Ensemble approaches combine the strengths of CNN spatial features with Transformer global context to improve accuracy (Sagar & Arukonda, 2025).

Nevertheless, performance improvements are often accompanied by training complexity and overfitting risks, especially on small-scale datasets. In addition to architecture, current research also explores feature representation domains which include spatial, temporal, and frequency domains. Frequency analysis enables the detection of frequency inconsistencies that are not visually apparent (X. Liu et al., 2022; Man & Cho, 2025), whereas multi domain integration combining spatial, frequency, and semantics is becoming a new direction in improving model generalization (Chen et al., 2025; Wang et al., 2025).

Overall, although prior literature has discussed various approaches, most are still descriptive and taxonomic in nature, without a systematic comparative evaluation regarding the effectiveness of each category based on datasets, architectures, and performance metrics. Therefore, this study develops a systematic literature review that not only summarizes developmental trends, but also fills the methodological gap through mapping relationships between approaches and thematic analysis of their strengths and limitations.

METHOD

The formulation of research questions in this study refers to the PICOC framework (Population, Intervention, Comparison, Outcome, and Context). The PICOC framework helps researchers systematically break down the review objectives into traceable key elements, while guiding the formulation of research questions based on the components of Population, Intervention, Comparison, Outcome, and Context. This approach ensures that the focus of the study, scope limitations, and direction of analysis are established in a structured manner according to the literature review requirements (Carrera-Rivera et al., 2022). As shown in Table 1, this framework is adapted to describe the conceptual elements that form the basis for formulating research questions in this study.

Table 1. PICOC Framework

PICOC Component	Description in This Study
Population (P)	Studies examining deepfake image detection based on machine learning or deep learning in the facial domain (still image).
Intervention (I)	Detection approaches or architectures used.
Comparison (C)	Comparison between approaches based on detection performance, model complexity, computational efficiency, and cross dataset generalization ability.
Outcomes (O)	Main outcomes in the form of accuracy, F1-score, AUC, cross domain robustness, model interpretability, and detection efficiency in real world conditions.
Context (C)	Research focusing on the facial image domain (image level.

3.1 Research Question

After identifying research gaps in previous studies related to image based deepfake detection, this systematic literature review formulates three research questions (RQ) to establish the scope, direction of analysis, and contributions of the study. The formulation of RQs is an essential stage because the first and most fundamental step in a systematic review is developing clearly formulated research questions, where well structured questions become the foundation of the entire review process and guide the search strategy, study selection, and data synthesis (De Cassai et al., 2025). Table 2 summarizes the research questions used along with the primary motivation for each question.

Table 2. Research Question

Research Question (RQ)	Primary Motivation
RQ1. What are the publication trends and patterns related to image based deepfake detection during the 2020–2025 period?	To map temporal developments, as well as to identify the research focus and intensity in this field.
RQ2. What machine learning models are used in image based deepfake detection, and what are the characteristics of the datasets, evaluation metrics, and their performance?	To synthesize the technical approaches used, review dataset diversity, and assess the influence of methodology on model performance comparatively.
RQ3. What are the main challenges and research gaps that still exist in image based deepfake detection studies, and what research directions are suggested for the future?	To identify recurring methodological limitations, reveal unexplored areas, and provide future research directions to improve detection effectiveness.

3.2 Search Strategy

Literature searches were conducted across four main databases, namely Scopus, ScienceDirect, IEEE Xplore, and MDPI Journals, with a publication range of 2020–2025 and filters for English language articles, full text, and sources from journals or scientific proceedings. The search process used a combination of primary keywords “machine learning” and “deepfake image”, which represent the research focus on applying machine learning for deepfake image detection. The Boolean search format used was: (“machine learning” OR “deep learning” OR “artificial intelligence”) AND (“deepfake image” OR “synthetic face” OR “fake image” OR “manipulated facial image”) AND (detect OR classif* OR identif*).

This strategy is designed so that the search scope remains broad yet relevant to the context of image based deepfake detection, and all results obtained were then reviewed manually to remove duplicates and ensure suitability with the research topic. The use of multiple databases and the application of keyword combinations through Boolean search is based on recommendations stating that search strategy effectiveness is highly determined by the breadth of source coverage and the accuracy of search term selection,

including the importance of ensuring that the terms used are able to represent the research domain comprehensively so that relevant studies are not missed (Wohlin et al., 2020).

3.3 Inclusion and Exclusion Criteria

Inclusion and exclusion criteria were applied to ensure that only relevant and empirical studies were analyzed in this review. Studies were included if: (i) written in English and published between 2020–2025, as this period reflects the rapid development phase of deepfake technology and new generative models; (ii) proposing, implementing, or evaluating deepfake detection methods on facial images (still image) using machine learning or deep learning approaches; (iii) presenting experimental results with measurable evaluation metrics such as accuracy, precision, recall, or F1 score; and (iv) focusing on the facial image domain which has characteristics distinct from video based detection.

Conversely, studies were excluded if: (i) they were duplicates or other versions of the same publication; (ii) focusing on video based detection without image experiments; (iii) having no empirical contribution, including conceptual articles, opinions, or non technical papers; (iv) discussing other topics such as deepfake generation, social impact, or ethics without a direct link to detection; and (v) representing secondary literature such as survey papers or reviews without primary experiments. The application of these criteria ensures that the study selection process is conducted objectively and free from bias, as establishing inclusion and exclusion criteria before conducting the review is a vital step in maintaining the integrity of primary evidence selection; furthermore, the selection of primary studies depends entirely on those criteria (Carrera-Rivera et al., 2022).

3.4 Study Quality Assessment

Assessing study quality is a crucial step in a systematic literature review because the quality of an SLR is highly influenced by the quality of the included primary studies, thus quality assessment is required to identify weaknesses that may affect relevance and the level of confidence in the review results (Usman et al., 2023). This study uses several indicators to assess the credibility of each study, including: (i) the clarity of research objectives and contributions; (ii) the appropriateness of the methodological design and the machine learning or deep learning architectures used; (iii) the relevance and verifiability of the datasets; (iv) the suitability of evaluation metrics to the research objectives; and (v) the consistency of result reporting and model generalization capability.

The assessment is conducted qualitatively by considering methodological strength, reporting transparency, and result consistency. Each study is classified into high, medium, or low quality categories. High quality articles are retained for in depth analysis, while medium quality studies are considered if they have relevant conceptual contributions. Low quality studies are excluded from the synthesis stage to maintain the validity of the review results.

3.5 Data Extraction

Data extraction is performed to accurately and consistently obtain and record core information from each study, following a data extraction form specifically designed to record all information necessary to answer the research questions; as explained in SLR guidelines that extraction forms must be established in advance to determine what data needs to be retrieved from each primary study (Carrera-Rivera et al., 2022). Extraction is performed manually using the previously established data extraction form, ensuring that each article is analyzed within a uniform framework and avoided from individual interpretation bias.

The extraction form contains several main information components, namely: (i) publication identity (year, author, and journal or conference source); (ii) research objectives and scope; (iii) image detection methods used, including machine learning or deep learning architectures; (iv) types and characteristics of datasets used, such as DFDC, FaceForensics++, Celeb-DF, or specific datasets; (v) evaluation metrics, such as accuracy, precision, recall, and F1 score; as well as (vi) main results and experimental findings. To ensure traceability, each article is annotated with additional notes explaining the

experimental context if there are significant variations in training techniques or testing schemes. The extraction process is carried out after the article is declared to meet the inclusion criteria and passes the quality assessment. Each study is read thoroughly, followed by a specific review of the methodology and experimental results sections as the primary information sources.

3.6 Data Synthesis

Data synthesis in this study is conducted using a narrative synthesis approach, by interpreting findings from all collected studies as part of the final stage of the systematic review. This approach is focused on identifying patterns, consistency, and differences between studies so that the review results can produce meaningful conclusions and reflect the accumulated evidence (De Cassai et al., 2025). Narrative synthesis is conducted by grouping studies based on themes, such as architecture types, feature extraction strategies, or image manipulation types. This approach produces thematic mapping that allows for comprehensive conclusions regarding trends, strengths, limitations, and research directions of image based deepfake detection methods. To clarify the relationships between studies and maintain consistency in cross theme analysis, the review results are further classified into three main dimensions: (i) model architectures, including CNN, Transformer, and hybrid/ensemble approaches; (ii) dataset characteristics, such as FaceForensics++, Celeb-DF, DFDC, CIFAKE, and custom made datasets; and (iii) evaluation metrics, including accuracy, F1 score, and AUC. This classification facilitates thematic analysis to identify performance patterns, limitations, and methodological trends across studies in a measurable way.

RESULTS AND DISCUSSION

4.1 Research Trends and Patterns of Image Based Deepfake Detection (RQ1)

In the initial phase (2020–2022), research focused on increasing the accuracy of Convolutional Neural Network (CNN) models in recognizing visual manipulation artifacts through convolutional traces and pairwise learning, although cross dataset generalization capabilities were still limited (Guarnera et al., 2020; Guo et al., 2021; Hsu et al., 2020). Since 2023, research focus has shifted toward hybrid and contextual architectures, such as CNN–LSTM which combines spatial and temporal features to detect micro artifacts, as well as GAN–CNN Ensemble and DeepGuardNet that emphasize computational efficiency and stability through multi-model integration (N & Simon, 2025; Sagar & Arukonda, 2025; Sharma et al., 2024). This shift marks a transition from single approaches toward ensemble learning systems that are more robust against manipulation variations. In the recent period, Vision Transformer (ViT) architectures and attention based models have emerged, which learn interpatch relations to recognize global forgery patterns (Alsolai et al., 2025; Çınar & Doğan, 2025; Khormali & Yuan, 2022). Frequency domain based approaches, such as FAD-Net and GAN-Transformer, have also evolved due to their ability to extract spectral patterns that are difficult to observe spatially (X. Liu et al., 2022; Man & Cho, 2025).

Overall, research trends are moving toward cross domain integration and improved model generalization. Recent studies such as SupCon-MPL and SMNDNet demonstrate a shift from architectural exploration toward the development of detection systems that are adaptive, efficient, and robust against new generative manipulations (Moon et al., 2024; Wang et al., 2025).

4.2 Models, Datasets, and Performance Evaluation of Image Based Deepfake Detection (RQ2)

Deepfake detection is generally classified as a binary problem, where the model determines whether an image is a result of manipulation or authentic. Early approaches heavily relied on Convolutional Neural Network (CNN) to recognize visual synthesis artifacts, with models such as SVM Expectation Maximization and AMTENnet reaching accuracies above 98% on standard datasets (Guarnera et al., 2020; Guo et al., 2021). Similar approaches are refined through Shallow FakeFaceNet (SFFN), which is more efficient on limited datasets (Lee et al., 2021).

As research progressed, studies shifted toward transfer learning and hybrid models to strengthen generalization capabilities. The combination of VGG16-CNN and NASNetLarge increased accuracy to 96.7%, while FAD-Net expanded analysis to the frequency domain to highlight deconvolution artifacts (Ilhan et al., 2022; X. Liu et al., 2022). Spatial–temporal integration is also applied through CNN-LSTM, which achieves accuracies of more than 96% on the FaceForensics++ and DFDC datasets (Sagar & Arukonda, 2025; Soundarya & Gururaj, 2025).

Next approaches adopt ensemble and adversarial strategies to improve cross domain stability (Alrajeh & Al-Samawi, 2025; Sharma et al., 2024). A significant evolution occurred with the presence of Vision Transformer (ViT), which through models like DFDT and PV-ISM achieves accuracies up to 99% with superior global representation (Çınar & Doğan, 2025; Khormali & Yuan, 2022). Attention based and graph models such as MGA-Net and GAN–Transformer Fusion (FFC) also show high efficiency, while contrastive learning approaches like SupCon-MPL improve robustness against generative variations (Chen et al., 2025; Man & Cho, 2025; Moon et al., 2024).

This confirms the importance of developing detectors that are not only accurate, but also adaptive and robust against real world data variations, as summarized in Table 3, which maps the relationship between models, performance, and the datasets used.

Table 3. Summary of models and image based deepfake detection performance

Author/Year	Model	Performance	Dataset
(N & Simon, 2025)	DeepGuardNet (Separable CNN)	Celeb-DF: Acc. 91%, Prec. 92%, Rec. 88%, F1=90% MesoNet (72%), ResViT (80.48%)	Celeb-DF
(Gura et al., 2024)	Customized CNN	DFDC: Acc. 91.47%, Loss 0.342, AUC 0.92 XceptionNet (73.06%), VGG16 (81.03%), DST-Net (90.94%)	Deepfake Detection Challenge (DFDC)
(Sharma et al., 2024)	GAN-CNN (DCGAN + VGG16)	Train acc 98.67%, test acc 70.08% Precision 68–72%, Recall 66–74%, F1≈70%	Kaggle Real and Fake Face Detection
(Sagar & Arukonda, 2025)	CNN-LSTM (ResNeXt50 backbone)	FF++: Acc. 96.67%, AUC 96.68% Celeb-DF: Acc. 91.8%, AUC 92%	FaceForensics++, Celeb-DF V1
(Soundarya & Gururaj, 2025)	Dense-Swish-CNN + Bi-LSTM	DFDC: 97.76%, CelebDF: 96.98%, ForgeryNIR: 96.5%	DFDC, CelebDF, ForgeryNIR, UADFV
(Man & Cho, 2025)	FFC (GAN + Transformer + FFT)	FF++: Acc. 98.75%, AUC 99.43% Unggul atas Xception & fCNN	CelebDF, FF++, DFDC
(Q. Liu et al., 2024)	Dual-Branch CNN (Self-Blending)	DFD: AUC 99.3%, CelebDF: AUC 94.25%, DFDCP: AUC 89.2%, DFDC: AUC 78.36%	DFD, DFDC, CelebDF,
(Wang et al., 2025)	SMNDNet (CNN empat modul)	LDFI: Acc. 99.07%, Prec. 99.93%, Rec. 98.22%, AUC 99.97%	LDFI (EFSF, ISFF, ESFF, AMFF)
(Çınar & Doğan, 2025)	PV-ISM (Patch-based Vision Transformer)	CIFAKE: Acc. 96.6%, Prec. 96.7%, Rec. 96.23%, F1=96.61%	CIFAKE (CIFAR-10 + Stable Diffusion 1.4)

Author/Year	Model	Performance	Dataset
(Chen et al., 2025)	MGA-Net (Multi-Graph Attention Network)	CIFAKE: Acc. 97.89%, GenImage: 75.1% Parameter hanya 0.48M	CIFAKE, GenImage
(Raza et al., 2022)	DFP (VGG16 + CNN)	Acc. 94%, Prec. 95% VGG16 (90%), NAS-Net (83%)	Kaggle RFDF (CIPLAB @ Yonsei)
(Moon et al., 2024)	SupCon-MPL (ResNet50)	Known: 81.40% acc, CG unknown: 57.85%, PG unknown: 51,90%	FF++, DFDC, CelebDF, StyleGAN, NeuralTextures
(Alsolai et al., 2025)	Guardian-AI (CNN + ViT + LSTM Hybrid)	CDDB: Acc. 95.8%, Prec. 96.2%, Rec. 95.6%, F1=95.9%	CDDB (Custom Deepfake Detection Benchmark)
(Khalil et al., 2021)	iCaps-Dfake (HRNet + Capsule Network)	CelebDF: Acc. 91.7%, DFDC-P: AUC +20.25%	DFDC-P, CelebDF
(Khormali & Yuan, 2022)	DFDT (Vision Transformer)	FF++: 99.41%, CelebDFv2: 99.31%, WildDeepfake: 81.35%	FF++, CelebDFv2, WildDeepfake
(Hsu et al., 2020)	Pairwise Siamese DenseNet	Precision 0.988, Recall 0.948 (wajah), Precision 0.934, Recall 0.900 (umum)	CelebA, ILSVRC12, GAN: DCGAN–PGGAN
(Kolagati et al., 2022)	MLP–CNN Hybrid	DFDC: AUC 0.877, Val. acc. 83%	DFDC, Dessa Dataset (YouTube)
(Alrajeh & Al-Samawi, 2025)	ViT-CNNs Ensemble (Binary Tree)	FF++: Acc. 97.25%, F1=97.28%	FaceForensics++
(Lee et al., 2021)	SFFN (Shallow-FakeFaceNet)	HFM: AUROC 72.52%, GAN fake: 93.99%	HFM, RFF, CelebA
(Guarnera et al., 2020)	EM + SVM (Convolutional Traces)	CELEBA–STYLEGAN2: Acc. 99.81%	CELEBA, STARGAN, STYLEGAN, ATTGAN, GDWCT
(Guo et al., 2021)	AMTENnet (Adaptive Convolution)	HFF: Acc. 98.52%, FF++: 95.17%	Hybrid Fake Face (HFF), FaceForensics++
(X. Liu et al., 2022)	FAD-Net (Fourier Attention Detection)	ForenSynths: AP 94%, +42.65% accuracy	ForenSynths (ProGAN, CycleGAN, StyleGAN)
(Ilhan et al., 2022)	NASNetLarge CNN	CelebDFv2: Acc. 96.7%, Loss 0.09	CelebDFv2 (YouTube frames)
(Kawabe et al., 2022)	CNN Ensemble (ResNet-18 per bagian wajah)	FFHQ–StyleGAN2: Acc. 1.000, F1=1.000	FFHQ, StyleGAN2

4.3 Challenges, Gaps, and Research Directions in Image Based Detection (RQ3)

Image based deepfake detection continues to advance through the utilization of CNN, Vision Transformer (ViT), and ensemble approaches, yet real world application still faces several fundamental challenges. Models trained on a single dataset or generator often show a performance drop when tested on different domains, such as images generated by StyleGAN3 or recent diffusion models (Kawabe et al., 2022; N & Simon, 2025). Furthermore, most public datasets only cover full-face manipulations, leaving studies on partial or localized edits still limited. Vulnerabilities to compression, distortion, as well as low interpretability and computational efficiency, also restrict their application in real-time scenarios and mobile devices.

From these findings, several critical gaps emerge, including limited cross generator generalization, a lack of multimodal cues integration, the absence of standardized metrics and benchmarks, and the remaining weakness in

detecting partial and hybrid fakes. However, before directing future research, it should be noted that this study has certain limitations: the scope of analysis only covers image based detection, the heterogeneity of experimental designs across studies limits direct comparison, and no meta-analysis was conducted due to variations in architectures, metrics, and reporting configurations.

Based on these challenges and limitations, future research directions need to emphasize improving cross domain robustness through approaches such as domain generalization, meta-learning, and few-shot and zero-shot models that are adaptive to new manipulations. The development of adaptive local representations based on attention and integration with Explainable AI (XAI) can strengthen model transparency, while the exploration of lightweight architectures such as MobileNet-V3, TinyViT, or EfficientFormer will enable efficient implementation on edge and mobile devices. Overall, enhancing generalization, efficiency, and interpretability remains the top priority in building image based deepfake detection systems that are more adaptive and ready to face the complexities of upcoming generative technologies.

CONCLUSION

This research conducts a Systematic Literature Review (SLR) of 24 empirical studies on machine learning and deep learning based image deepfake detection for the 2020–2025 period. Review results show a methodological evolution from classical convolutional architectures toward hybrid, Transformer based, and ensemble learning approaches that utilize multi-domain representations to improve generalization. Models such as Vision Transformer (ViT), CNN–LSTM, and frequency domain based approaches demonstrate high accuracy under controlled conditions, yet still decline when tested on new datasets or generators.

Main challenges include limited cross domain generalization, low interpretability, and the lack of standardized metrics and benchmarks across studies. Therefore, future research needs to emphasize domain generalization through adversarial training, meta-learning, and semi-supervised learning, accompanied by multimodal integration combining visual, physiological, and metadata features. The development of Explainable AI (XAI) and lightweight architectures such as MobileNet-V3, TinyViT, and EfficientFormer is also essential to support real-time detection on edge devices. Overall, future deepfake detection systems must be adaptive, efficient, and transparent, with cross-disciplinary collaboration between computer vision, machine learning, and digital forensics to maintain visual authenticity and information trust in the era of generative technology.

REFERENCES

Alrajeh, M., & Al-Samawi, A. (2025). Deepfake Image Classification Using Decision (Binary) Tree Deep Learning. *Journal of Sensor and Actuator Networks, 14*(2), 40. https://doi.org/10.3390/jsan14020040

Alsolai, H., Mahmood, K., Alshuhail, A., Ben Miled, A., Alqahtani, M., Alshareef, A., Alallah, F. S., & Alghamdi, B. M. (2025). Guardian-AI: A novel deep learning based deepfake detection model in images. *Alexandria Engineering Journal*, *126*, 507–514. https://doi.org/10.1016/j.aej.2025.04.095

Carrera-Rivera, A., Ochoa, W., Larrinaga, F., & Lasa, G. (2022). How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, *9*,101895.https://doi.org/10.1016/j.mex.2022.101895

Chen, G., Du, C., Yu, Y., Hu, H., Duan, H., & Zhu, H. (2025). A Deepfake Image Detection Method Based on a Multi-Graph Attention Network. *Electronics*, *14*(3), 482.https://doi.org/10.3390/electronics14030482

Çınar, O., & Doğan, Y. (2025). Novel Deepfake Image Detection with PV-ISM: Patch-Based Vision Transformer for Identifying Synthetic Media. *Applied Sciences*, *15*(12), 6429. https://doi.org/10.3390/app15126429

De Cassai, A., Dost, B., Tulgar, S., & Boscolo, A. (2025). Methodological Standards for Conducting High-Quality Systematic Reviews. *Biology*, *14*(8), 973. https://doi.org/10.3390/biology14080973

Guarnera, L., Giudice, O., & Battiato, S. (2020). DeepFake Detection by Analyzing Convolutional Traces. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2841–2850. https://doi.org/10.1109/CVPRW50498.2020.00341

Guo, Z., Yang, G., Chen, J., & Sun, X. (2021). Fake face detection via adaptive manipulation traces extraction network. *Computer Vision and Image Understanding*, *204*, 103170. https://doi.org/10.1016/j.cviu.2021.103170

Gura, D., Dong, B., Mehiar, D., & Said, N. Al. (2024). Customized Convolutional Neural Network for Accurate Detection of Deep Fake Images in Video Collections. *Computers, Materials & Continua*, *79*(2), 1995–2014. https://doi.org/10.32604/cmc.2024.048238

Hsu, C.-C., Zhuang, Y.-X., & Lee, C.-Y. (2020). Deep Fake Image Detection Based on Pairwise Learning. *Applied Sciences*, *10*(1), 370. https://doi.org/10.3390/app10010370

Ilhan, I., Bali, E., & Karakose, M. (2022). An Improved DeepFake Detection Approach with NASNetLarge CNN. *2022 International Conference on Data Analytics for Business and Industry (ICDABI)*, 598–602. https://doi.org/10.1109/ICDABI56818.2022.10041558

Kawabe, A., Haga, R., Tomioka, Y., Okuyama, Y., & Shin, J. (2022). Fake Image Detection Using An Ensemble of CNN Models Specialized For Individual Face Parts. *2022 IEEE 15th International Symposium on Embedded Multicore/Many-Core Systems-on-Chip (MCSoC)*, 72–77. https://doi.org/10.1109/MCSoC57363.2022.00021

Khalil, S. S., Youssef, S. M., & Saleh, S. N. (2021). iCaps-Dfake: An Integrated Capsule-Based Model for Deepfake Image and Video Detection. *Future Internet*, *13*(4), 93. https://doi.org/10.3390/fi13040093

Khormali, A., & Yuan, J.-S. (2022). DFDT: An End-to-End DeepFake Detection Framework Using Vision Transformer. *Applied Sciences*, *12*(6), 2953. https://doi.org/10.3390/app12062953

Kolagati, S., Priyadharshini, T., & Mary Anita Rajam, V. (2022). Exposing deepfakes using a deep multilayer perceptron – convolutional neural network model. *International Journal of Information Management Data Insights*, *2*(1), 100054. https://doi.org/10.1016/j.jjimei.2021.100054

Lee, S., Tariq, S., Shin, Y., & Woo, S. S. (2021). Detecting handcrafted facial image manipulations and GAN-generated facial images using Shallow-FakeFaceNet. *Applied Soft Computing*, *105*, 107256. https://doi.org/10.1016/j.asoc.2021.107256

Liu, Q., Xue, Z., Liu, H., & Liu, J. (2024). Enhancing Deepfake Detection With Diversified Self-Blending Images and Residuals. *IEEE Access*, *12*, 46109–46117. https://doi.org/10.1109/ACCESS.2024.3382196

Liu, X., Liu, J., Guo, P., Tuo, D., Tian, S., & Jiang, Y. (2022). FAD-Net: Fake Images Detection and Generalization Based on Frequency Domain Transformation. *2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1–7. https://doi.org/10.1109/CISP-BMEI56279.2022.9980271

Man, Q., & Cho, Y.-I. (2025). Exposing Face Manipulation Based on Generative Adversarial Network–Transformer and Fake Frequency Noise Traces. *Sensors*, *25*(5), 1435. https://doi.org/10.3390/s25051435

Moon, K.-H., Ok, S.-Y., & Lee, S.-H. (2024). SupCon-MPL-DP: Supervised Contrastive Learning with Meta Pseudo Labels for Deepfake Image Detection. *Applied Sciences*, *14*(8), 3249. https://doi.org/10.3390/app14083249

N, A. D., & Simon, P. (2025). DeepGuardNet: A Novel CNN Architecture for DeepFake Image Detection. *Procedia Computer Science*, *258*, 811–818.

https://doi.org/10.1016/j.procs.2025.04.313

Raza, A., Munir, K., & Almutairi, M. (2022). A Novel Deep Learning Approach for Deepfake Image Detection. *Applied Sciences*, *12*(19), 9820. https://doi.org/10.3390/app12199820

Sagar, N. K., & Arukonda, S. (2025). A Novel CNN-LSTM Approach for Robust Deepfake Detection. *Procedia Computer Science*, *258*, 1844–1855. https://doi.org/10.1016/j.procs.2025.04.436

Sharma, P., Kumar, M., & Sharma, H. K. (2024). GAN-CNN Ensemble: A Robust Deepfake Detection Model of Social Media Images Using Minimized Catastrophic Forgetting and Generative Replay Technique. *Procedia Computer Science*, *235*, 948–960. https://doi.org/10.1016/j.procs.2024.04.090

Soundarya, B. C., & Gururaj, H. L. (2025). A Novel Dense-Swish-CNN With Bi-LSTM Framework for Image Deepfake Detection. *IEEE Access*, *13*, 89641–89653. https://doi.org/10.1109/ACCESS.2025.3570761

Usman, M., Bin Ali, N., & Wohlin, C. (2023). A Quality Assessment Instrument for Systematic Literature Reviews in Software Engineering. *E-Informatica Software Engineering Journal*, *17*(1), 230105. https://doi.org/10.37190/e-Inf230105

Wang, Q., Wang, X., Li, J., Han, R., Liu, Z., & Guo, M. (2025). SMNDNet for Multiple Types of Deepfake Image Detection. *Computers, Materials & Continua*, *83*(3), 4607–4621. https://doi.org/10.32604/cmc.2025.063141

Wohlin, C., Mendes, E., Felizardo, K. R., & Kalinowski, M. (2020). Guidelines for the search strategy to update systematic literature reviews in software engineering. *Information and Software Technology*, *127*, 106366. https://doi.org/10.1016/j.infsof.2020.106366