JESICA: Journal of Enhanced Studies in Informatics and Computer Applications
Journal homepage: jesica.itsk-soepraoen.ac.id
ISSN: e-ISSN: 3046-6938 p-ISSN: 3046-6997
Vol. 3, No. 1, February 2026, Pp. 34-42

A Systematic Literature Review of Artificial Intelligence Algorithms for Deepfake Detection

Aulia Roessati Putri*[1], Bintang Aulia Novala[2], Deva Muhamad Syaiful Arifin[3], Zulhilmi Luthfiah[4], Risqy Siwi Pradini[5]
1,2,3,4,5 Institut Teknologi, Sains dan Kesehatan RS.DR. Soepraoen Kesdam V/BRW, Indonesia

***Corresponding author**
E-mail address:
23102007@student.itsk-soepraoen.ac.id

Keywords:
Algorithm, artificial intelligence, deepfake, detection, SLR

Abstract
The evolution of information technology has positioned multimedia content as a pillar of digital communication, but at the same time, it has opened a gap for serious threats in the form of deepfakes. This highly realistic media manipulation challenges information authenticity, privacy, and cybersecurity, which, for Information Technology professionals, presents both technical and ethical challenges. This Systematic Literature Review (SLR) aims to map the development of Artificial Intelligence based algorithms in deepfake detection. Using the PRISMA methodology on 20 selected primary articles (2021-2025), this study aims to identify trends in the use of AI algorithms for deepfake detection, determine the most effective approaches, and analyze the factors contributing to their effectiveness. The analysis results show a paradigm shift from single models (such as CNN) to hybrid architectures (CNN-LSTM-Transformer) and complex multimodal fusion systems. It was found that hybrid algorithms are the closest approach to best practice due to their ability to handle spatial and temporal dimensions simultaneously. Key contributing factors include hierarchical feature extraction, generative data augmentation, and the integration of Explainable AI (XAI).

1. Introduction
The development of information technology has brought society into the digital era, where multimedia content has become a major pillar of communication. However, this progress has been accompanied by the emergence of deepfakes as a serious digital threat that undermines information accuracy, privacy, and global data security [1]. For information technology professionals, the deepfake phenomenon is not only a technical challenge in cybersecurity but also an ethical challenge in maintaining the integrity and authenticity of content in the digital era [2].

The complexity of the resulting manipulation makes fake content increasingly difficult to distinguish from genuine assets, necessitating stronger defence mechanisms at the IT infrastructure level [3]. Efforts to stem this threat focus on the development of various detection algorithms using Artificial Intelligence (AI) technology to analyse fake content in the form of videos, images, and audio [2]. Study [4] suggests that the development of detection algorithms is focused on identifying artifacts and subtle patterns resulting from digital engineering. This effort aims to strengthen the verification system and maintain the integrity and security of user identities.

Based on this background, this study aims to conduct an investigation using the Systematic Literature Review (SLR) method to identify and map various algorithms used in deepfake detection. This research is very important to select the best algorithms among the various existing algorithms, so that the most accurate and efficient detection solution can be found currently in dealing with deepfake threats, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) [5]. Each architecture has characteristics in handling datasets, and determining which algorithm is most resilient to continuously evolving attacks is crucial for the loss of information security systems [6]. Without a systematic comparison, IT professionals will have difficulty determining effective protocol detection standards to handle the latest manipulation variants [7]. This study uses the SLR method to answer this need by analysing the results of previous studies.

The results of this systematic review are expected to provide in-depth findings regarding the effectiveness of each algorithm and formulate best practice guidelines for technology policymakers in handling deepfakes. The urgency of formulating these guidelines is in line with previous research findings that emphasize the need for a comprehensive approach and multi-layered governance in responding to the risks of deepfake technology [1],[8]. In an effort to fill this literature gap, this review will map the extent of the capabilities of current AI models in detecting various types of manipulation. This research is specifically designed to answer questions regarding the development of detection algorithms, the identification of the best algorithms, and factors supporting their effectiveness. The next section will

describe the SLR methodology, followed by a structured discussion of the results to provide a comprehensive overview of the future of deepfake detection technology.

2. Research Method

This study applies the SLR method to identify developments in digital engineering detection technology. The literature review process is structured based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The use of the PRISMA method is intended to ensure that the entire research process, from literature identification to synthesis of results, is transparent, structured, and replicable.

2.1. PICO Framework

This study is based on the PICO (Population, Intervention, Comparison, and Outcomes) framework, which serves as a guide in determining the study population, type of intervention, and comparison factors. This study was structured using the PICO framework, as shown in Table 1, to determine the accuracy of each study found. Referring to the established PICO framework, this study formulates 3 Research Questions (RQ) that form the basis of the analysis process.

- (RQ1): What are the algorithms applied in the detection and analysis of deepfake content using AI?
- (RQ2): Which is the best algorithm that can be used in detecting deepfakes according to the latest research?
- (RQ3): Why can this algorithm be stated as the best method? (effectiveness supporting factors)?

Table 1. PICO framework

Population (P)	Intervention (I)	Comparison (C)	Outcome (O)
The deepfake phenomenon is a digital threat that impacts information integrity, privacy, and data security, as well as poses ethical challenges for IT professionals in the detection process.	Various deepfake detection algorithms utilize AI technology to analyze fake content.	Comparison of the performance of various deepfake detection algorithms.	Effectiveness of algorithms and best practices in handling deepfakes.

2.2. Research Procedures

This systematic review procedure was conducted through a series of systematic steps to select, evaluate, and analyse studies relevant to the effectiveness of various AI architectures in detecting deepfake threats. The evaluation focused on comparing algorithm performance to find the best detection solution. The identification stage begins with formulating research questions (RQ1, RQ2, and RQ3) and determining relevant key terms.

The literature search process is carried out using a combination of Boolean operators such as "Deepfake Detection AND Artificial Intelligence", "Convolutional Neural Networks AND Recurrent Neural Networks", "Digital Manipulation AND Image Forensics", "Deep Learning OR Machine Learning AND Detection Accuracy". The search was conducted on major databases (such as Google Scholar, Science Direct, Scopus, IEEE Xplore, and ResearchGate), resulting in a total of 250 initial records. Next, a screening step is performed to remove articles that do not meet the established inclusion and exclusion criteria. To ensure a high-quality review, strict selection criteria for both inclusion and exclusion can be established. The following table lists the inclusion and exclusion criteria, as presented in Table 2.

Table 2. Inclusion and exclusion criteria

Inclusion	Exclusion
Scientific articles published in journals, conference proceedings, or literature reviews that have gone through a peer review process.	Theses, dissertations, final project reports, blogs, white papers, or non-peer-reviewed articles.
A study that discusses deepfake detection using AI algorithms.	Articles that do not discuss deepfakes or only discuss deepfakes in general without detection algorithms.
The article presents an evaluation of the algorithm performance (accuracy, F1-score, precision, recall) or a comparison between algorithms.	Articles that do not include a performance evaluation or do not explain the effectiveness/weaknesses of the algorithm.
Articles in English or Indonesian.	Articles in languages other than Indonesian and English.
Available in full-text and accessible form.	Full text is not available, or only the abstract is accessible.
Articles published in the period 2021–2025.	Articles published before 2021.

During the identification stage, 59 articles were removed due to being identified as duplicates. The remaining data were further reviewed, considering the suitability of the marked titles and abstracts. Four data sets were declared non-compliant based on automated screening results for the 2021–2025 period. Additionally, 33 data sets were eliminated for other reasons, including tier classification (Q1 and Q2). No data were eliminated due to the absence of abstracts during the screening stage.

After the initial screening, 154 articles passed the screening stage. In this phase, 88 articles were removed, leaving 66 articles to proceed to the next stage. In this process, 66 articles were retrieved, while 46 articles were excluded. Next, the 20 successfully collected articles were evaluated for their suitability for inclusion in the review. Ultimately, 20 articles were included in this systematic review, with 20 articles from several studies. This number represents the result of a rigorous and systematic selection process. The PRISMA diagram showing the stages of identification, screening, eligibility, and inclusion of articles in this study is shown in Figure 1 below.

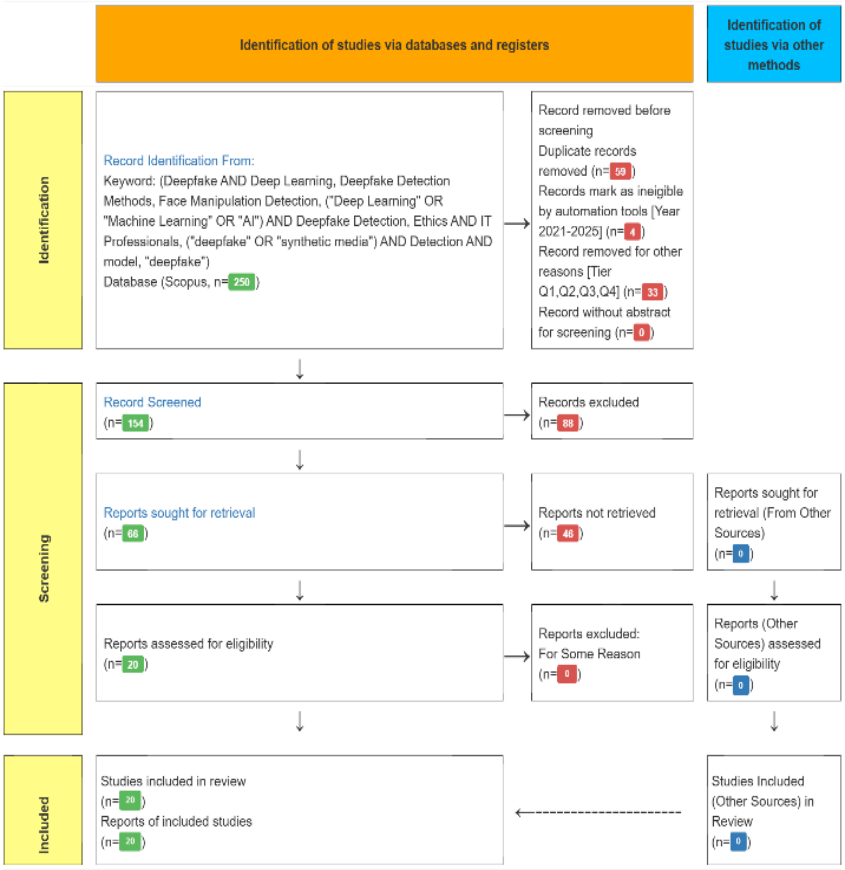

Figure 1. PRISMA diagram

3. Results and Discussion

Based on an analysis of 20 recent articles, covering the period from 2021 to 2025, this study identified various approaches to deepfake detection. Table 3 below summarizes the data synthesis results from each reviewed article.

Table 3. Data synthesis results

Studies	Research algorithms	Research subjects	Research result
Lei et al., 2025 [9]	FG-TEFusionNet, Diffusion Model, Adversarial Training (I-FGSM)	Dataset: FaceForensics++, Celeb-DF Type: Video and Image	The FG-TEFusionNet integration demonstrated superior performance with an accuracy rate of 99.49% and an AUC of 99.87%. The application of the Diffusion Model was proven to significantly reduce the Attack Success Rate (ASR) to 10.15%, strengthening the model's resilience against adversarial attacks.

Nelson et al., 2025 [10]	XceptionNet, CNN-LSTM, XceptionNet-LSTM	Dataset: CelebA, DEEP-VOICE, FF++, DFDC, Celeb-DF Type: Audio, Image, Video	This three-stage multi-modal framework successfully detects manipulation across various media types with an accuracy rate of 95.56% for images, 98.5% for audio, and 97.57% for video, respectively, through optimizing microscopic artifact extraction.
Wang et al., 2024 [11]	CSDSA (CNN & Swin Transformer)	Dataset: FaceForensics++ (FF++) Type: Video	The CSDSA network implementation successfully combined local spatial features with global context, resulting in a detection accuracy rate of 95.62%. The results showed that the application of the Swin Transformer module consistently improved model stability.
Ismail et al., 2021 [12]	YIX (YOLO, InceptionResNetV2, XGBoost)	Dataset: CelebDF-FaceForensics++ (c23) Type: Video	The use of XGBoost as a final classification tool in the YIX architecture has been proven to be able to increase detection precision to reach an accuracy of 90.73% and an AUC value of 90.62%.
Karathanasis et al., 2025 [13]	CNN (VGG-based) with pruning compression techniques, knowledge distillation, quantization, adapter for transfer learning	Dataset: Synthbuster, 140k Real dan Fake Faces, DeepFake dan Real Images, ForenSynths Type: Image and Video	This study demonstrates that the Knowledge Distillation method maintains detection accuracy above 97% despite significantly reducing the model size. This demonstrates the effectiveness of the compression model for implementation on resource-constrained devices.
Petmezas et al., 2025 [14]	3D ResNet18 and Transformer	Dataset: FF++, Celeb-DF, ReenactFaces Type: Video	Combining features from facial microexpressions using a two-branch fusion model yielded an accuracy rate of 99.81% and a perfect ROC-AUC score (100%). These results demonstrate that microexpression patterns are a strong indicator for distinguishing between real and manipulated faces.
Sohail et al., 2025 [15]	CNN, CNN-LSTM, CNN-GRU, TCN, GAN-Autoencoder	Dataset: FaceForensics++ (FF++) Type: Image and Video	Forensic analysis using the TCN and CNN-LSTM architectures achieved 96% accuracy and a 98% F1-score. This research demonstrates that using data augmentation with GAN-based methods significantly improves the model's ability to detect temporal inconsistencies.
Chen et al., 2025 [16]	MGA-Net (HWGAN and PDGAN)	Dataset: CIFAKE, GenImage Type: Image	The efficient MGA-Net architecture with a limited parameter set (0.48M) successfully achieved a detection accuracy of 97.89%. In this system, the dual graph attention module proved effective in extracting artifact features from AI-generated images.
Alhaji et al., 2024 [17]	Hibrida ACO-PSO and Deep Learning	Dataset: DFDC (Kaggle/Meta), Type: Video	The application of the ACO-PSO hybrid optimization algorithm improves the detection system's accuracy to 98.91%, with an F1-score of 99.12%. This method demonstrates better performance in selecting relevant spatiotemporal features for detecting video manipulation.
Awotunde et al., 2023 [18]	5-layer CNN with ReLU	Dataset: DeepFake, Face2Face, First-Order Motion Type: Video	The developed CNN model achieved an average accuracy of 86.32%. However, when it came to classifying specific face-swap (deepfake) content, the system demonstrated high effectiveness, with a detection success rate of 98%.
Petmezas et al., 2025 [19]	Hybrid CNN-LSTM-Transformer and 3DMM	Dataset: VoxCeleb2, DFD, Celeb-DF, FF++ Type: Video	Through the integration of 3D Morphable Models (3DMM), the developed hybrid model achieved an Area Under the Curve (AUC) value between 97% and 99%. Furthermore, the system demonstrated high robustness when tested on both low-quality and high-resolution videos.

Li et al., 2023 [20]	HuRawNet2 (HuBERT and RawNet2)	Dataset: ASVspoof 2021/2019, FMFCC-A, FAD Type: Audio	The use of the pre-trained HuBERT model resulted in an Equal Error Rate (EER) of 2.89% on the ASVspoof 2021 dataset. This result demonstrates excellent cross-language generalization capabilities in detecting voice forgery (deepfake audio).
Bhandarkawthekar et al., 2025 [21]	RLNet (ResNet50 + LSTM) and Grad-CAM	DFDC (Deepfake Detection Challenge) Type: Video	The RLNet architecture achieved an accuracy of 95.2%. The implementation of the Grad-CAM technique in this study contributes to the transparency model by providing a visualization of facial areas indicated by digital infection.
Maheshwari & Paulchamy, 2024 [22]	XAI-ART Framework: CNN + Adversarial Robustness Training (FGSM/PGD) + Explainable AI (SHAP, LIME) + automatic deletion proposal	DFDC (Kaggle) Type: Image and Video	This research resulted in a detection system with 97.5% accuracy integrated with Explainable AI (XAI). The use of SHAP and LIME enabled accountable interpretation of the model's decisions to support automated content removal.
Naskar et al., 2024 [23]	Stacking Ensemble and Meta-learning	Dataset: Celeb-DF (V2) and FaceForensics++ (FF++) Type: Video	The feature fusion strategy using Stacking Ensemble achieved 96.33% accuracy on Celeb-DF and 98.00% on FF++. These results confirm that combining multiple deep learning models provides more stable performance than a single model.
Qadir et al., 2024 [24]	ResNet-Swish-BiLSTM	Dataset: FF++, DFDC, Celeb-DF Type: Video	Achieving 98.99% accuracy on the FF++ dataset demonstrates the superiority of the Swish activation function over ReLU. This model has proven effective in capturing complex visual manipulation patterns on various benchmark datasets.
Reis & Ribeiro, 2024 [25]	ResNet-100, ArcFace, Bayesian Likelihood Ratio (LR)	Dataset: Celeb-DF (v2) Type: Image and Video	From a digital forensic perspective, the use of facial similarity scores yielded an AUC value of 0.994. The application of Bayesian evaluation provides a strong scientific basis for determining the probability of authenticity of digital evidence.
Gao et al., 2024 [26]	Dual-Stream TAD (Texture & Artifact Detector)	Dataset: WildDeepfake, FF++ Type: Image and Video	Through texture and artifact decomposition, the model achieved an intra-dataset accuracy of 93.32%. The most important finding was an increase in cross-dataset generalization ability to 81.44%, demonstrating the model's readiness for real-world scenarios.
Alsolai et al., 2025 [27]	Vision Transformer (ViT) + LSTM + Attention Mechanisms	Dataset: CDDB Benchmark Type: Image and Video	The Guardian-AI system achieved 95.8% accuracy with 96.2% precision. The use of an attention mechanism allows the model to focus on spatial anomalies that frequently appear in content manipulated by generative algorithms.
Javed et al., 2025 [28]	Multimodal Framework: CNN, ViT, Diffusion Models, and BiLSTM	Dataset: FakeAVCeleb, AV-Deepfake1M, TVIL, LAV-DF Type: Audio-Visual	This multimodal model achieves state-of-the-art accuracy of 99.87% on the FakeAVCeleb dataset. The use of Diffusion Models as a preprocessor has proven effective in removing noise from audio-visual feature representations.

The synthesis results in Table 3 show that the development of deepfake detection methodology, which initially used a single Convolutional Neural Network (CNN) architecture, has shifted towards the integration of advanced hybrid and multimodal models. The application of Diffusion Models and Vision Transformers (ViT) has been empirically proven to increase detection accuracy to exceed the critical threshold of 99% [9]. In addition, there is a shift in research focus from simply optimizing accuracy to strengthening the functionality and resilience aspects of the system, which include defense mechanisms against adversarial attacks [9], model efficiency for the Internet of Things ecosystem [13], increasing decision accountability through the Prepared AI framework [22], and strengthening generalization capabilities across datasets through texture decomposition and precise spatiotemporal analysis [26]. The following

section will discuss the answers to the three previously defined RQs. These answers are based on the results of the literature synthesis conducted and summarized in Table 3.

(RQ1): What are the algorithms applied in the detection and analysis of deepfake content using AI?

Based on a systematic review of the current literature, the algorithms used in deepfake content detection and analysis can be classified into several main approaches that utilize advanced artificial intelligence architectures. Table 4 below summarizes the classification of deepfake detection algorithms identified in this study.

Table 4. Classification of deepfake detection algorithms

Method Category	Description	Algorithms	Studies
Single Model	Uses a single underlying architecture (CNN or Transformer) to identify features from a single image or frame. Focuses on visual feature extraction and performance optimization.	CNN, ResNet, XceptionNet, EfficientNet, ViT, MGA-Net	[13], [16], [18]
Hybrid Model	Combining two architectures in one workflow for video analysis (spatial and temporal) or a combination of capabilities.	CNN-LSTM/BiLSTM, ResNet-Swish-BiLSTM, Hybrid CNN-Transformer, 3D CNN + LSTM/Transformer	[11],[19],[21], [24]
Multimodal Functions and Ensembles	Combining results (predictions or features) from multiple models or independent processing branches for a more robust final decision.	Audio-Visual Fusion, Cross-modal Attention, Multi-Branch Decision Fusion, Stacking Ensemble	[9], [10], [28]
Generative and Specialist Approaches	Using generative models as a specialized tool for specific application contexts, such as forensics.	GAN-based Detection, Diffusion Models, Similarity Scoring dan Likelihood Ratio	[17], [25], [28]

From Table 4 above, the current implementation of deepfake detection algorithms shows various strategies that emphasize efficiency in identifying invisible manipulation artifacts. First, single CNN-based models such as pure XceptionNet or Vision Transformer (ViT) generally focus on recognizing pixel anomalies and texture irregularities in a single image [13]. To overcome the limitations in capturing motion dynamics, hybrid models such as CNN-LSTM or CNN-Transformer combinations were developed specifically to extract spatial features while analyzing temporal inconsistencies between video frames [15]. Further evolution resulted in the multimodal fusion category, where the system cross-verifies visual and audio signals to detect phonetic differences with lip movements [10]. This strategy is often reinforced by a generative approach using Diffusion Models or GAN-Autoencoders that serve as pre-processing units to remove detrimental noise (denoising) and emphasize the differences between the original texture and artifacts from the manipulation results of generative algorithms [9]. Overall, this grouping confirms that the effectiveness of detection systems increases significantly as the model's ability to integrate various data aspects and advanced processing techniques increases.

(RQ2): Which is the best algorithm that can be used in detecting deepfakes according to the latest research?

Based on a systematic review of the current literature, research results indicate that no single method is universally the "best" answer. Detection effectiveness is highly dependent on the type of data (image, audio, video), the quality of manipulation, and the variety of datasets used [15]. However, comparative analysis shows that hybrid architectures and multi-modal fusion systems consistently demonstrate superior performance, robustness, and generalization capabilities compared to single-model approaches [19][26]. The following is a summary of the comparison of the best methods presented in Table 5.

Based on performance analysis from recent research, for deepfake videos, the hybrid CNN-LSTM-Transformer shows the best performance (99.81% accuracy) with deep spatial-temporal analysis capabilities [14]. For deepfake images, Vision Transformer with attention mechanism achieves 97.89% accuracy due to its ability to understand global context [16]. The multi-modal CNN-ViT-Diffusion approach equipped with BiLSTM audio achieves the highest overall accuracy (99.87%), despite its considerable implementation complexity [27]. The data shows that the “best method” is contextual and depends on the type of data being analyzed.

For video content, models that combine CNN with LSTM and are equipped with an attention mechanism or Transformer have proven to be the most efficient. For example, the CNN-LSTM-Transformer hybrid model proposed by Petmezas et al. [11] obtained an AUC between 97–99% on various datasets and showed good robustness to video

compression. Its advantage lies in its comprehensive capabilities: CNN captures spatial artifacts in each frame, LSTM models temporal inconsistencies between frames, and Transformer provides global context [19],[21]. This combination provides a powerful answer to detect subtle temporal manipulations.

Table 5. Comparison of the best methods based on data type

Data Types	Best Method	Highest Accuracy	Dataset	Superiority
Video	Hybrid CNN-LSTM-Transformer	99.81%	FaceForensics++	Comprehensively combines spatial, temporal, and contextual analysis
Image	Vision Transformer (ViT) with Attention	97.89%	CIFAKE	Global context understanding, subtle artifact detection
Audio	HuBERT + RawNet2 (modified)	EER 2.89%	ASVspoof 2021	Cross-language generalization, very high accuracy
Multi-modal	CNN-ViT-Diffusion with BiLSTM audio	99.87%	FakeAVCeleb	Advanced audio-visual integration, denoising with diffusion models

In terms of generalizability across datasets, the toughest challenge in deepfake detection is where methods that separate and analyze intrinsic features of content demonstrate superiority. The Texture and Artifact Detector (TAD) achieved the highest average accuracy (81.44%) when tested on seven different datasets [25]. Its success lies in its dual-stream design that separates the analysis between natural textures and artificial artifacts, allowing the model to learn the underlying characteristics of the manipulation [25]. The same principle is applied in multi-modal fusion, where the CNN-ViT-Diffusion framework managed to maintain accuracy above 98% on four different datasets, thanks to the fusion of local-global features and pre-processing using a diffusion model [28].

In terms of computational efficiency and implementation readiness, methods that combine high accuracy with model optimization can be considered "best" for real-time or edge computing situations. Knowledge distillation techniques effectively maintain accuracy >97% on CNN models that have been compressed to 10% of their original size [13]. This shows that the "best" model is not always the most complex, but the one that achieves an optimal balance between accuracy and efficiency [13], [23]. For audio-based detection, architectures that utilize self-supervised pre-training on a large amount of audio data are superior in terms of generalization. HuBERT modified RawNet2 (HuRawNet2) achieves the lowest Equal Error Rate (EER) (2.89%) on the ASVspoof 2021 dataset and demonstrates strong performance on Mandarin data, proving its ability to address linguistic variations [9].

Taking all these aspects into account, it can be concluded that the "best" methods according to current research are design systems that are:

- Hybrid and Hierarchical: Combine the strengths of more than one architecture (CNN, RNN, Transformer) to attack the problem from multiple levels of abstraction [11], [19], [20].
- Fused and Disjointed: Integrate multi-modal (audio-visual) or multi-branch (texture-artifact) information, but maintain separate processing pipelines for these fundamental features [10], [19], [21].
- Robust and Interpretable: Strengthened by adversarial training and equipped with Explainable AI (XAI) to increase robustness and trustworthiness [9], [28].
- Efficient and Adaptable: Optimized through techniques such as distillation and transfer learning to remain relevant in the face of new generation methods and real-world computational constraints [18],[20].

Therefore, the current research direction is no longer competing to create a single winning algorithm, but rather designing flexible and modular frameworks, where the best-tested components can be assembled and adapted to specific deepfake threats [11], [13], [17].

(RQ3): Why can this algorithm be stated as the best algorithms? (effectiveness supporting factors)?

The analysis in RQ2 identified hybrid architectures and multi-modal fusion systems as leading approaches to synthetic media threat mitigation. The superiority of these methods is supported by a comprehensive set of system design principles and implementation paradigms, which address the technical validity of why these approaches are classified as best-in-class methods in the current literature (2021–2025).

Based on a thorough analysis of the factors presented in Table 6, it can be concluded that the effectiveness of current deepfake detection methods no longer depends on a single parameter, but rather on the collaboration between architectural components that function. The main factor that makes hybrid architectures and multimodal fusion systems the best methods is their ability to perform multidimensional cross-validation that combines spatial, temporal, and audio modalities [19], [28]. The effectiveness of these systems stems from hierarchical feature extraction, where

the use of CNNs for local features and Vision Transformer (ViT) for global context allows the model to detect manipulative artifacts at various scales, ranging from pixel anomalies to overall facial structural mismatches [11],[27].

Table 6. Factors supporting effectiveness

Supporting factors for effectiveness	Principles/characteristics	Main benefits	Studies
Hierarchical and Exhaustive Feature Extraction	A combination of architectures that capture local (fine details) and global (overall context) features	Capable of detecting anomalies at the pixel/texture level while understanding structural inconsistencies within the overall content	[13], [16], [18], [19]
Smart Fusion Strategy	Multi-modal Fusion: Combining local (fine details) and global (overall context) features	Detects inter-modality mismatches (e.g., lip-sync errors) and separates the original signal from manipulated noise, improving accuracy and robustness	[10], [27], [28]
Utilization of Generative Techniques	Using GAN or Diffusion Model for Data Augmentation and Denoising Pre-processing	Creating more diverse and challenging training samples, and cleaning low-quality inputs, thereby improving the model's generalization ability	[9], [15], [27]
Training Paradigm for Robustness	Implementation of Adversarial Training and Strict Regularization Techniques	Strengthens the model against attacks designed to deceive detection (adversarial attacks) and prevents overfitting, resulting in more stable performance on new data	[9], [14], [21], [22]
Transparency and Optimization for Implementation	Integration of Explainable AI (XAI) and Model Optimization techniques (such as Knowledge Distillation)	Improve decision accountability and debugging capabilities through model decision interpretation, and enable real-time deployment on limited devices (edge computing)	[13], [21], [22]

The application of intelligent fusion strategies, particularly on audio-visual data, provides a robust defence layer, as generative technologies are notoriously difficult to precisely align lip movements with speech signals without introducing visible inconsistencies [10], [28]. The incorporation of novel generative techniques such as Diffusion Models serves both as a challenging data augmentation tool and a denoising preprocessing unit that can filter out features from low-quality or compressed inputs [9], [28]. Therefore, the system's maturity is also supported by the pillars of transparency through XAI and model optimization, which ensure that classification decisions are not only statistically sound but also forensically justifiable and efficient for deployment on resource-constrained devices [13], [22]. Overall, the combination of these five supporting factors forms a detection ecosystem that is modular, adaptive, and highly generalizable in the face of future deepfake threats.

4. Conclusion

Based on 20 systematic reviews of the current literature, research findings indicate that no single method is universally the "best" answer. Detection effectiveness depends heavily on the type of data (image, audio, video), the quality of manipulation, and the variety of datasets used. However, a comparative analysis shows that hybrid architectures and multi-modal fusion systems consistently demonstrate superior performance, robustness, and generalization capabilities compared to single-model approaches. Combinations of various architectures, such as combining the spatial feature extraction capabilities of CNNs with the temporal analysis capabilities of RNNs, have been shown to address vulnerabilities often present in individual models. Therefore, this study concludes that hybrid approaches and fusion systems are the closest to current best practices. This strategy is considered most effective because it can adapt to a wide variety of manipulation attacks simultaneously, providing IT professionals with a higher level of reliability in maintaining content authenticity at the infrastructure level.

Moving forward, further research needs to focus on developing models that are adaptive to evolving manipulation techniques, including leveraging transformer-based architectures and self-supervised learning approaches to improve cross-dataset generalization. Furthermore, exploration of XAI mechanisms is needed to enhance the transparency and accountability of detection systems, particularly in the context of implementation in the cybersecurity and digital identity

verification sectors. Standardizing evaluation datasets and testing in real-world deployment scenarios are also crucial to ensure the sustainability and readiness of this technology to address the increasingly complex deepfake threat.

References

[1] N. Hynek, B. Gavurova, and M. Kubak, "Risks and benefits of artificial intelligence deepfakes: Systematic review and comparison of public attitudes in seven European countries," *Journal of Innovation & Knowledge*, vol. 10, no. 5, p. 100782, 2025, doi: 10.1016/j.jik.2025.100782.

[2] F. Folorunsho and B. F. Boamah, "Deepfake technology and its impact: Ethical considerations, societal disruptions, and security threats in AI-generated media," *International Journal of Information Technology and Management Information Systems (IJITMIS)*, vol. 16, no. 1, pp. 1060–1080, 2025, doi: 10.34218/IJITMIS_16_01_076.

[3] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake Detection: A Systematic Literature Review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022, doi: 10.1109/ACCESS.2022.3154404.

[4] M. Alrashoud, "Deepfake video detection methods, approaches, and challenges," *Alexandria Engineering Journal*, vol. 125, pp. 265–277, 2025, doi: 10.1016/j.aej.2025.04.007.

[5] V. L. L. Thing, "Deepfake Detection with Deep Learning: Convolutional Neural Networks versus Transformers," *Proc. 2023 IEEE Int. Conf. Cyber Secur. Resilience*, CSR 2023, pp. 246–253, 2023, doi: 10.1109/CSR57506.2023.10225004.

[6] J. Alves, P. Sousa, T. Cruz, and J. Mendes, "A review of architecture features for distributed and resilient industrial cyber–physical systems," *Journal of Manufacturing Systems*, vol. 82, pp. 1069–1090, 2025, doi: 10.1016/j.jmsy.2025.07.012.

[7] S. U. Qureshi, J. He, S. Tunio, N. Zhu, A. Nazir, A. Wajahat, F. Ullah, and A. Wadud, "Systematic review of deep learning solutions for malware detection and forensic analysis in IoT," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 8, p. 102164, 2024, doi: 10.1016/j.jksuci.2024.102164.

[8] J. Pratama, "Ethical deepfake governance: EU AI Act benchmarks and Indonesia multi-stakeholder responses," *Jurist-Diction*, vol. 9, no. 1, pp. 69–90, 2026, doi: 10.20473/jd.v9i1.80885.

[9] S. Lei, J. Song, F. Feng, Z. Yan, and A. Wang, "Deepfake Face Detection and Adversarial Attack Defense Method Based on Multi-Feature Decision Fusion," *Appl. Sci.*, vol. 15, no. 12, 2025, doi: 10.3390/app15126588.

[10] L. Nelson, H. Batra, and P. Radha, "Deepfake Detection in Manipulated Images/ Audio/ Videos: A Three-Stage Multi-Modal Deep Learning Framework," *Intel. Artif.*, vol. 28, no. 76, pp. 20–39, 2025, doi: 10.4114/intartif.vol28iss76pp20-39.

[11] S. Wang, C. Du, and Y. Chen, "A New Deepfake Detection Method Based on Compound Scaling Dual-Stream Attention Network," *EAI Endorsed Trans. Pervasive Heal. Technol.*, vol. 10, pp. 1–10, 2024, doi: 10.4108/eetpht.10.5912.

[12] A. Ismail, M. Elpeltagy, M. S. Zaki, and K. Eldahshan, "A New Deep Learning-Based Methodology for Video Deepfake," *MDPI, Basel, Switz.*, pp. 1–15, 2021, doi: https://doi.org/10.3390/s21165413

[13] A. Karathanasis, J. Violos, and I. Kompatsiaris, "A Comparative Analysis of Compression and Transfer Learning Techniques in DeepFake Detection Models," *Mathematics*, vol. 13, no. 5, 2025, doi: 10.3390/math13050887.

[14] G. Petmezas, V. Vanian, M. P. Rufete, E. E. I. Almaloglou, and D. Zarpalas, "A Dual-Branch Fusion Model for Deepfake Detection Using Video Frames and Microexpression Features," *J. Imaging*, vol. 11, no. 7, pp. 1–12, 2025, doi: 10.3390/jimaging11070231.

[15] S. Sohail, S. M. Sajjad, A. Zafar, Z. Iqbal, Z. Muhammad, and M. Kazim, "Deepfake Image Forensics for Privacy Protection and Authenticity Using Deep Learning," *Inf.*, vol. 16, no. 4, pp. 1–30, 2025, doi: 10.3390/info16040270.

[16] G. Chen, C. Du, Y. Yu, H. Hu, H. Duan, and H. Zhu, "A Deepfake Image Detection Method Based on a Multi-Graph Attention Network," *Electron.*, vol. 14, no. 3, 2025, doi: 10.3390/electronics14030482.

[17] H. S. Alhaji, Y. Celik, and S. Goel, "An Approach to Deepfake Video Detection Based on ACO-PSO Features and Deep Learning," *Electron.*, vol. 13, no. 12, 2024, doi: 10.3390/electronics13122398.

[18] J. B. Awotunde, R. G. Jimoh, A. L. Imoize, A. T. Abdulrazaq, C. T. Li, and C. C. Lee, "An Enhanced Deep Learning-Based DeepFake Video Detection and Classification System," *Electron.*, vol. 12, no. 1, 2023, doi: 10.3390/electronics12010087.

[19] G. Petmezas, V. Vanian, K. Konstantoudakis, E. E. I. Almaloglou, and D. Zarpalas, "Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification," *Multimed. Tools Appl.*, vol. 84, no. 33, pp. 40617–40636, 2025, doi: 10.1007/s11042-024-20548-6.

[20] L. Li, T. Lu, X. Ma, M. Yuan, and D. Wan, "Voice deepfake detection using the self-supervised pre-training model HuBERT," *Applied Sciences*, vol. 13, no. 14, p. 8488, 2023, doi: 10.3390/app13148488.

[21] V. Bhandarkawthekar, T. M. Navamani, R. Sharma, and K. Shyamala, "Design and development of an efficient RLNet prediction model for deepfake video detection," *Front. Big Data*, vol. 8, 2025, doi: 10.3389/fdata.2025.1569147.

[22] R. U. Maheshwari and B. Paulchamy, "Securing online integrity: a hybrid approach to deepfake detection and removal using Explainable AI and Adversarial Robustness Training," *Automatika*, vol. 65, no. 4, pp. 1517–1532, 2024, doi: 10.1080/00051144.2024.2400640.

[23] G. Naskar, S. Mohiuddin, S. Malakar, E. Cuevas, and R. Sarkar, "Deepfake detection using deep feature stacking and meta-learning," *Heliyon*, vol. 10, no. 4, p. e25933, 2024, doi: 10.1016/j.heliyon.2024.e25933.

[24] A. Qadir, R. Mahum, M. A. El-Meligy, A. E. Ragab, A. AlSalman, and M. Awais, "An efficient deepfake video detection using robust deep learning," *Heliyon*, vol. 10, no. 5, p. e25757, 2024, doi: 10.1016/j.heliyon.2024.e25757.

[25] P. M. G. I. Reis and R. O. Ribeiro, "A forensic evaluation method for DeepFake detection using DCNN-based facial similarity scores," *Forensic Sci. Int.*, vol. 358, p. 111747, 2024, doi: 10.1016/j.forsciint.2023.111747.

[26] J. Gao et al., "Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection," *Eng. Appl. Artif. Intell.*, vol. 133, no. PC, p. 108450, 2024, doi: 10.1016/j.engappai.2024.108450.

[27] H. Alsolai et al., "Guardian-AI: A novel deep learning based deepfake detection model in images," *Alexandria Eng. J.*, vol. 126, no. May, pp. 507–514, 2025, doi: 10.1016/j.aej.2025.04.095.

[28] M. Javed, Z. Zhang, F. H. Dahri, and T. Kumar, "Enhancing multimodal deepfake detection with local–global feature integration and diffusion models," *Signal, Image Video Process.*, vol. 19, no. 5, pp. 1–9, 2025, doi: 10.1007/s11760-025-03970-7.