JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 07 Issue 02
2024

Research Article
https://doi.org/10.56979/702/2024

Improving DeepFake Detection: A Comprehensive Review of Adversarial Robustness, Real-Time Processing and Evaluation Metrics

Najaf Saeed[1*], Gohar Mumtaz[1], Muqaddas Yaqub[1], and Muhammad Haroon Ahmad[2]

[1]Faculty of Computer Science and Information Technology, Superior University, Lahore, 54000, Pakistan.
[2]Riphah International University, Lahore, 54000, Pakistan.
[*]Corresponding Author: Najaf Saeed. Email: nnjfali44@gmail.com

Received: June 02, 2024 **Accepted:** August 06, 2024 **Published:** September 01, 2024

Abstract: This review analyzes 30 studies on deepfake detection. It focuses on three areas: adversarial robustness, real-time processing, and evaluation metrics. Deepfake technology is making rapid progress. It poses serious threats to digital security. We need strong, efficient detection models. The review exposes three key factors that boost detection system power. They are: adversarial training, GAN-based methods, and lightweight designs. They boost both resilience and efficiency. But challenges remain. We need real-time processing and standardized tests. They must capture the nuances of deepfake detection. The findings show a need for more research. It must address new threats, improve detection models, and set real-world benchmarks. The study stresses the need to improve deepfake detectors. We must integrate advanced training, optimize methods, and refine metrics. They must be robust, accurate, and adaptable to new digital threats.

Keywords: Deepfake Detection; Adversarial Robustness; Real-Time Processing.

1. Introduction

Deep fake detection is concerned with employing advanced machine learning and artificial intelligence techniques that is used to identify and prevent the spread of fraudulent media content, where the appearance of the person is replaced with someone else's in an existing image or video [1]. This technology is crucial for cyber security because deep fake leads to significant threats including false information, identity theft, and other malicious activities [2]. Deep fake technology grew rapidly, by using powerful AI models, especially deep learning techniques, to create highly realistic artificial media [3].

Next, we discuss the deepfake detection models as follows.

- Convolutional Neural Networks (CNN) models are widely used for analyzing images, where they detect ordinary discrepancies in the pixels of deep fake images [4].
- Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, that are a sub type of RNN, are designed to analyze temporal sequences in videos, identifying unnatural transitions that may indicate deep fakes [5].
- Spatiotemporal Networks are designed to detect deep fakes by analyzing patterns across video frames and operate by combining both geographical and time-based analyses [5].

Here, we discuss some frameworks that are used in deepfake detection:

- Adversarial Training technique involves training models on adversarial examples (intentionally perturbed inputs) to improve their resistance to adversarial attacks [6].
- Multimodal Detection Frameworks is systems that integrate visual, audio, and biometric analysis are used to enhance deepfake detection accuracy [7].

Figure 1 illustrate the typical architecture of deep fake detection models, highlighting the flow from input (video/image) to feature extraction, classification, and output (real or fake) [8].

Advanced models can detect subtle anomalies in deepfake content with high precision. Some models are flexible and can be trained on large datasets to the improve detection performance [9].

A lot of models are very sensitive to small changes, and even the most advanced detection systems can't catch them all. Constant identification requires huge computational assets, making it trying to send in genuine situations [10].

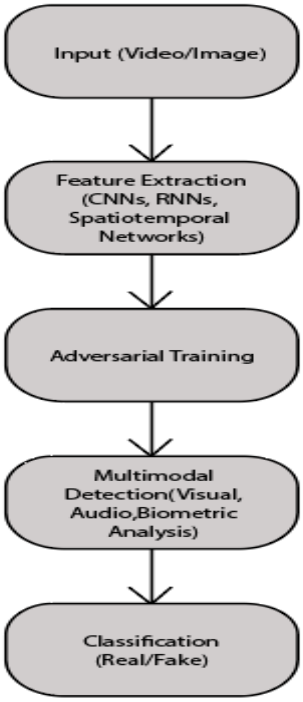

Figure 1. Deepfake detection

Here, we discuss some general current challenges in deepfake detection:

- Adversarial Robustness: It's hard to make detection models attack-resistant [11].
- Real-Time Processing: We must balance model complexity and speed for real-time detection [12].
- Evaluation Metrics: We need better metrics for evaluating models [13].

The mixing of real and fake content poses a major challenge to digital security by deep fakes. The development of their creation methods outpaces the development of detection techniques. This makes it difficult to maintain accuracy in digital media. There are three main concerns; tricking detection models, slow data processing, and failures in detecting deepfakes as highlighted.

Deepfake detection's principal issue is a deficiency in technology. Deepfake generation techniques are increasingly sophisticated today. However, detection technology has lagged behind it all this while. Adversarial attacks pose a serious threat that can deceive standard detectors through manipulating inputs into them via generated samples from these adversarial approaches. Also, deployment in real scenarios becomes complicated by real-time deepfake detection. In addition, current metrics often overlook subtleties associated with detecting deepfakes such as the wide range of distinctive faking techniques and their applications within the reality.

Other deepfake detection techniques include the use of CNNs, RNNs and spatiotemporal networks. Model resilience is improved by adversarial training as well as robust feature extraction. These models have weaknesses; they are defenseless against attacks, expensive to keep running, and current tests are not good enough. These limits impede fast-paced real-world applications of deepfake detection which require quick precise decisions.

There are many other malicious or illegal uses of Deepfake, such as spreading misinformation, creating political instability, commit fraud, or various cybercrimes. To address such threats, the field of Deepfake detection has attracted considerable attention from academics and experts during the last few years, resulting in many Deepfake detection techniques. There are also some efforts on surveying selected literature focusing on either detection methods or performance analysis. However, a more comprehensive overview of this research area will be beneficial in serving the community of researchers and practitioners by providing summarized information about Deepfake in all aspects, including the exploration of new training techniques. These techniques should boost adversarial robustness. It also suggests optimization

methods to balance model complexity and real-time processing. It calls for new evaluation metrics, and it will also propose effective, practical solutions, which are noticeably missing in previous surveys. Toward that end, we present a systematic literature review (SLR) on Deepfake detection in this paper.

The main contribution of this study is to enhance deepfake detection models. We are interested in making them more robust and efficient. These goals are as follows:

- Exploring new canvassing techniques for a better adversarial robustness.
- Advancements in Deepfake Detection Techniques
- Improvement of Evaluation Metrics
- Looking at optimization methods that achieve a trade-off between the model complexity and real-time processing.
- Creating different evaluation metrics that give a better reflection of actual performance.
- Reviewing existing solutions, identifying research gaps and areas unexplored.

The remainder of the paper is organized as follows: Section II presents the literature review of 30 related studies. In Section III, we discuss research methodology. In Section IV, we thoroughly discuss the results. Section V identifies the research gaps. In Section VI, we discuss on findings. In Section VII, we present limitations. In Section VIII, we discuss implications and future directions, and in Section IX, we concludes the paper.

2. Literature Review

The rise of deepfake technology has sparked interest in detection methods. Deepfakes are getting more advanced. They now threaten security, privacy, and trust in digital content. Research in this field has focused on three areas. They are: enhancing adversarial robustness, achieving real-time processing, and improving detection model metrics.

2.1. Adversarial Robustness

Adversarial robustness is a critical area of concern in deepfake detection. Adversarial attacks involve subtle manipulations to input data. Design misleads detection models. These attacks pose a major threat to the reliability and security of these systems. These attacks can cause even the best models to misclassify fake content as real, or vice versa [14]. This undermines deepfake detection.

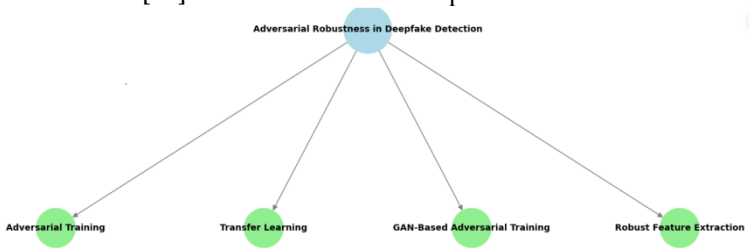

Figure 2. Adversarial Robustness in Deepfake Detection

Various approaches have been proposed to enhance adversarial robustness:

- Adversarial Training: This method adds adversarial examples to the training data. Linhai Ma et al. modified these inputs to deceive the model. Training the model on these adversarial examples helps it resist them. This makes it more robust [15].
- Transfer Learning: Transfer learning boosts robustness by using knowledge from related tasks. It improves the model's ability to detect advanced deepfakes [3].
- GAN-Based Adversarial Training: Generative Adversarial Networks (GANs) create realistic adversarial scenarios. Robust detection models use them for training [16].
- In the field of cybersecurity, convolutional neural networks (CNNs) have demonstrated effectiveness in detecting various threats, as demonstrated by Shahbaz et al. (2024), who evaluated CNN-based methods for SQL injection attack detection with notable success [17].

There are some limitations or challenges to adversarial robustness that are discussed below:

- A challenge remains the trade-off between robustness and accuracy on clean data. Models trained on adversarial examples may perform well under attack. But, they may be less accurate on standard, non-adversarial inputs [14].

- New, advanced attack techniques outpace defenses. This shows a need for ongoing research and innovation.

Here, we discuss emerging techniques and future directions for adversarial robustness:

- Ensemble Learning and Multi-Model Approaches: Use many models to boost detection system resilience [18].
- Explainable AI (XAI): Use XAI to make deepfake detection models more transparent. It will help identify and fix their vulnerabilities [14].

2.2. Real-Time Processing

Real-time processing is essential for deepfake detection systems, especially in fast-paced online environments. The main challenge in real-time detection is balancing model accuracy with speed [19].

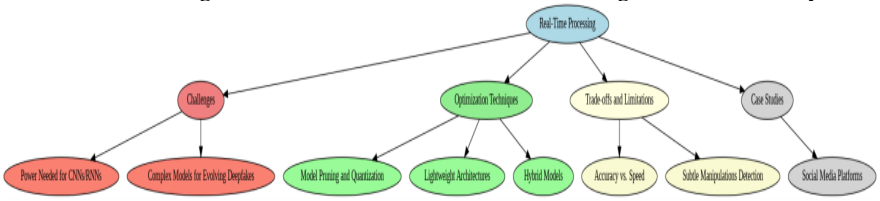

Figure 3. Real-Time Processing in Deepfake Detection

The main challenges include:

- Deep learning models, especially CNNs and RNNs, need a lot of power to analyze video frames. This can cause delays [20].
- Models to detect evolving deepfake techniques are complex. They can take a long time to process. This makes real-time performance hard to achieve [19].

Optimization techniques are discussed here:

- Model Pruning and Quantization: These techniques reduce a model's size. They speed up inference while maintaining accuracy [4].
- Lightweight Architectures: Use efficient models like MobileNets and SqueezeNets. They have fewer parameters and are faster [4].
- Hybrid Models: They combine traditional and deep learning. This yields faster inference times [21].

Trade-offs and limitations in real-time processing:

- Pruning and quantization lower compute needs. But, they may hurt accuracy, especially for complex deepfakes [19].
- Lightweight models may struggle to detect subtle manipulations. We must research further to ensure that real-time detection does not reduce accuracy [19].

Case Studies and Applications in real-time processing:

- Social media platforms use real-time deepfake detection systems. They rely on edge computing and lightweight models. This achieves fast processing without overloading central servers [19].

2.3. Evaluation Metrics

Real-time processing is essential for deepfake detection systems, especially in fast-paced online environments. The main challenge in real-time detection is balancing model accuracy with speed [22].

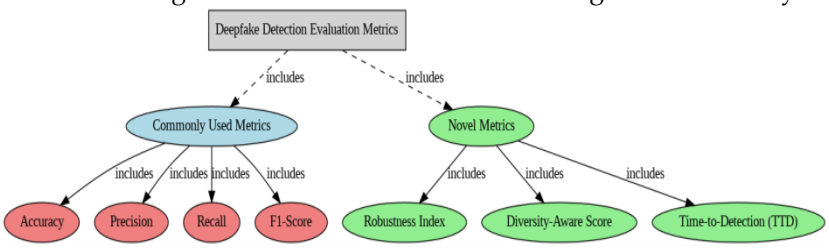

Figure 4. Evaluation Metric

Commonly used metrics are discussed here:

- Accuracy, Precision, Recall, F1-Score: These metrics assess model performance. But they may miss the nuances of various deepfake detection scenarios [23].

Limitations of current metrics in evaluation metrics:

- Current metrics often overlook the diversity and complexity of deepfake techniques. This can lead to misleading evaluations [23].
- Traditional metrics ignore adversarial attacks. They slash model performance (Smith et al., 2022).

Here, we discussed some proposed novel metrics:

- Robustness Index: A metric that measures a model's performance in tough conditions. It provides a better view of its effectiveness [16].
- Diversity-Aware Score: Tests a model's ability to detect various deepfakes. It ensures that the model can generalize across different techniques [23] [32].
- Time-to-Detection (TTD): It measures how long the model takes to detect a deepfake. This is crucial for real-time detection [23].

3. Research Methodology

This research uses a systematic literature review. It analyzes and synthesizes studies on deepfake detection. It focuses on adversarial robustness, real-time processing, and evaluation metrics. The systematic review method ensures a complete and unbiased knowledge base. It provides a clear overview of the current state of research in the field.

3.1. Research questions and Objectives

Table 1. Research questions and Objectives.

QID	Research questions	Objectives
RQ1	How can we improve deepfake detection models' robustness to attacks by using new training methods?	Test if new training methods can make models more robust.
RQ2	What defense mechanisms can we add to deepfake detection models to resist adversarial attacks?	Explore more defense mechanisms to enhance adversarial robustness.
RQ3	What techniques can optimize deepfake detection for speed and complexity?	Identify techniques that optimize speed and complexity while maintaining accuracy.
RQ4	What new metrics can better reflect the real-world performance of deepfake detection models?	Develop and assess new metrics for evaluating model performance beyond standard metrics.
RQ5	How do current metrics fail to capture the	Check current measurements with precision.

	nuances of deepfake detection, and how can we improve them?	Propose improvements to measure model performance in adversarial conditions.

3.2. Study Selection Procedure

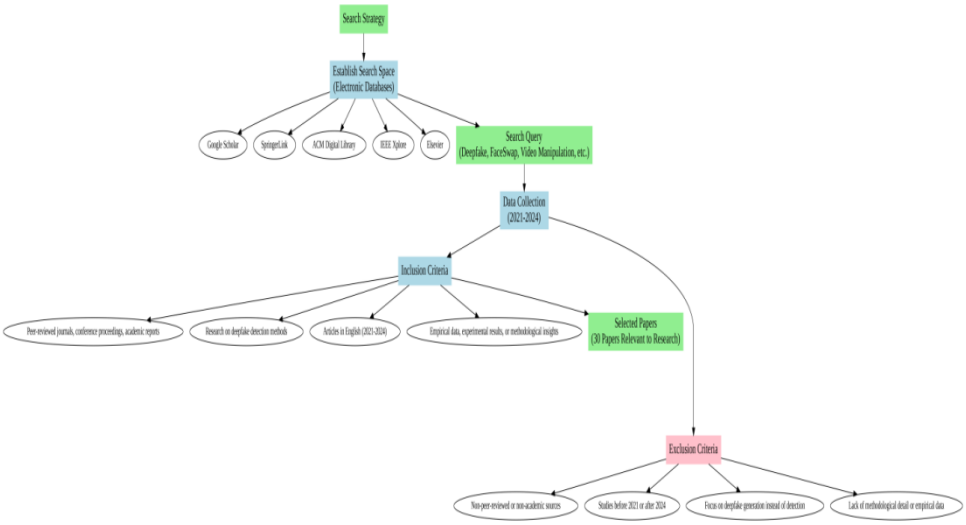

Figure 5. Study Selection Procedure.

3.3. Search Strategy

Following the establishment of study objectives and questions, a formal search method was employed to gather and examine all empirical data material to the review's goal. As illustrated inTable1, the plan involved establishing the search space, which comprised electronic databases.

Table 2. Electronic Databases

Sr. No	Electronic Data Bases
1	Google Scholar
2	Springer
3	ACM
4	IEEE Xplore
5	Elsevier

3.4. Search Query

We aimed to collect works to answer our research on deepfake detection. To ensure complete coverage and avoid bias, we used many search phrases and keywords. We used Boolean logic to combine these search terms with 'AND' or 'OR.' This maximized the retrieval of relevant studies.

The search terms used included combinations such as:

- Deepfake/ FaceSwap/ Video manipulation AND
- Detection/Detect OR
- Facial Manipulation / Digital Media Forensics

3.5. Inclusion and Exclusion Criteria

To address the study questions, we needed to include and exclude relevant papers. This required well-defined criteria. We selected all studies for this review based on these criteria. Researchers carried out the survey between January 2019 and August 2024. Table II details the criteria for selecting relevant papers. I reviewed each of the 30 papers. I did this to extract and synthesize data on deepfake detection methods. This includes their resilience, speed, and new performance metrics.

Table 3. Inclusion Criteria

Inclusion	Criteria
IC1	Studies published in peer-reviewed journals, conference proceedings, and academic reports

IC2	Research targeting deepfake detection methods, with an emphasis on resilience, speed, and measurable approaches.
IC3	Articles published in English.
IC4	Articles published between 2019 and 2024.
IC5	Studies providing empirical data, experimental results, or detailed methodological insights.

Table 4. Exclusion Criteria

Exclusion	Criteria
EC1	Articles that are not peer-reviewed or are from non-academic sources.
EC2	Studies published before 2019.
EC3	Papers focusing on deepfake generation rather than detection.
EC4	Studies lacking detailed methodological information and empirical data.

3.6. Quality Assessment

We assessed each selected study's quality using a set of criteria. These included the research design, the validity of the results, and the relevance of the findings to the research questions. We included studies that met these criteria in the final synthesis. This ensured the systematic review's reliability and validity.

4. Findings

4.1. Results

The review analyzed 30 studies on deepfake detection. It focused on adversarial robustness, real-time processing, and evaluation metrics.

4.1.1. Adversarial Robustness

The review found several techniques to improve deepfake detectors. Make them more robust against adversarial attacks.

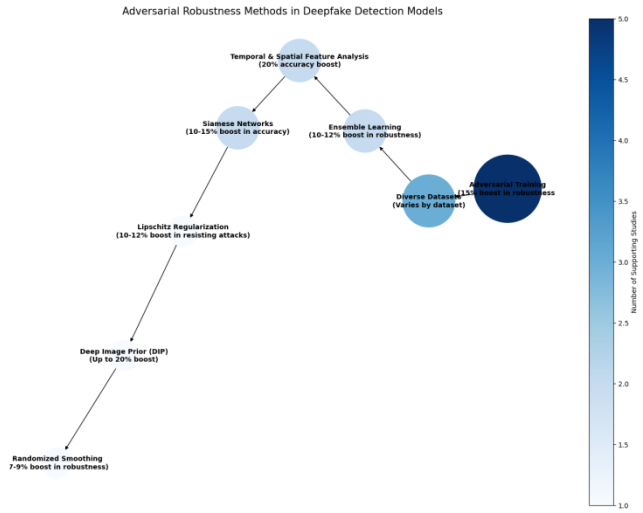

Figure 6. Adversarial Robustness Methods in deepfake detection

Figure VI shows methods to make deepfake detection more robust against adversarial attacks.

- Adversarial Training: The largest, darkest node represents this method. It has the most research backing. Studies show it can boost robustness by up to 15% [16].
- Diverse Datasets: This method, though it varies by dataset, shows strong support. It helps improve the model's robustness in different real-world environments [19].
- Ensemble Learning: It combines many models to reduce weaknesses. It boosts robustness by 10% to 12% [23].

- Temporal & Spatial Feature Analysis: This method uses data from video frames. It uses both spatial and temporal data. It improves model accuracy by 20% [5].
- Siamese Networks: They improve accuracy by 10-15%. They separate features better, making the model more resilient to adversarial attacks [14].
- Other Methods: It also includes Lipschitz Regularization, Deep Image Prior (DIP), and Randomized Smoothing. They display reduced backing, evident in their compact, featherweight nodes [14][16].

The literature supports three key techniques, shown in the diagram: Adversarial Training, Diverse Datasets, and Ensemble Learning. They are vital for making deepfake detection models more robust. Meanwhile, methods like Randomized Smoothing and Lipschitz Regularization are effective. But, they are less studied. This suggests areas for further research.

4.1.2. Real-Time Processing

Real-time detection of deepfakes is crucial for practical use.

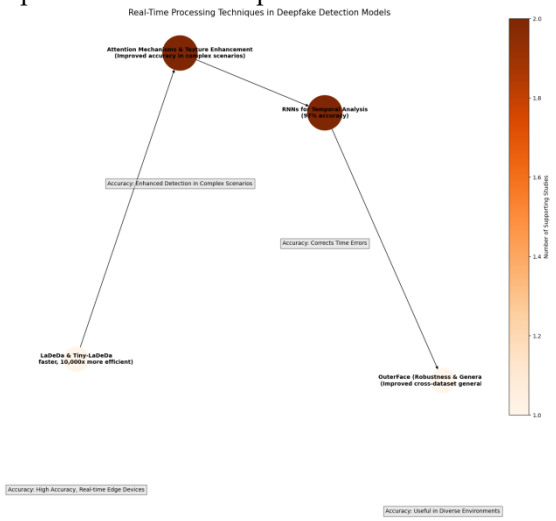

Figure 7. Real-Time Processing Techniques in Deepfake detection.

Figure VII shows techniques to optimize deepfake detection models for real-time use. It highlights both speed and complexity.

- A node shows that this method is 375 times faster and 10,000 times more efficient. It is ideal for real-time detection on edge devices. A single study supports this technique.
- Attention Mechanisms & Texture Enhancement: Two studies support this method. It boosts accuracy in detecting things in complex scenarios. It does this by focusing on different parts of an image and enhancing subtle artifacts [22].
- RNNs for Temporal Analysis: Two studies support this technique. It achieves 97% accuracy by analyzing temporal features across video frames. It addresses issues like flickering and color mismatches [23] [30].
- OuterFace (Robustness & Generalization): This technique aims to improve cross-dataset generalization. One study supports it. It is particularly useful in diverse environments where standard detection methods may fail [24].

The diagram shows that Attention Mechanisms, Texture Enhancement, and RNNs are best for real-time processing. They are the most researched and effective. They also improve accuracy. LaDeDa and Tiny-LaDeDa are efficient, so they suit resource-limited environments. OuterFace offers strong performance on varied datasets. It addresses the challenge of generalization in real-world applications.

4.1.3. Evaluation Metrics

We assessed deepfake detection models using various metrics. Figure VIII shows the results.

Figure VIII shows evaluation metrics. They aim to better measure the real-world performance of deepfake detection models.

- Weighted Precision (wP): A node shows its strong ability to handle class imbalances. This is important in cases where real videos outnumber deepfakes. Two studies support this metric, which is stricter on false positives [25].

- Mean Average Precision (mAP): This metric tests accuracy and generalization. It does this by averaging precision at various recall levels. Two studies support it. It is useful for evaluating performance across different scenarios [19].
- Log Weighted Precision (log-wP): One study supports this metric. It is useful when deepfakes are rare. It emphasizes reducing false positives, making it an important tool for real-world detection [25].
- Temporal Coherence Metrics: They test frame consistency in videos. This helps in dynamic, real-time scenarios. It addresses, with one study, a challenge. It is to keep accuracy across video frames [26].
- Robustness & Generalization Testing: This metric checks if models can resist adversarial attacks. It's vital for deploying deepfake detection in diverse, hostile environments. One study supports it [24].

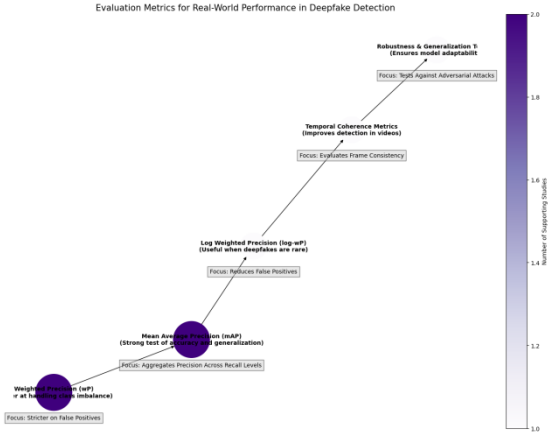

Figure 8. Evaluation Metrics for Real-World performance in Deepfake detection

Figure VIII shows that wP and mAP are the most researched metrics for deepfake detection. They are the best at capturing the nuances in real-world applications. Log Weighted Precision (log-wP) and Temporal Coherence Metrics are important, too. This is true in cases where deepfakes are rare or involve video data. Testing for robustness and generalization is critical. It ensures models work well on different datasets and resist adversarial attacks.

RQ1: How can we improve deepfake detection models' robustness to attacks by using new training methods?

To make deepfake detection models tougher against attacks, researchers proposed new training methods. Here are some of the key approaches:

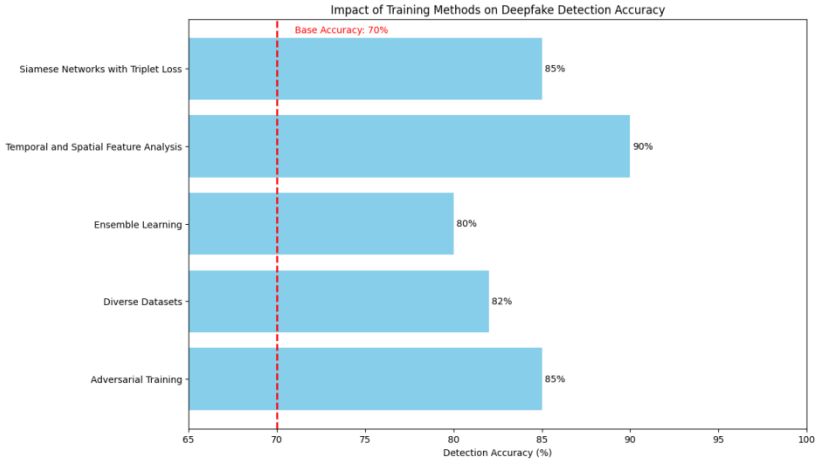

Figure 9. Impact of Training Methods on Deepfake Detection.

- Adversarial Training: Using adversarial examples in training boosts deepfake detection models' robustness. Training models with images altered by FGSM and Carlini-Wagner L2 attacks can help them resist similar real-world attacks [16]. Adversarial training can boost detection accuracy by up to 15% in adversarial conditions. But, it may hurt performance on clean, unaltered data.
- Diverse Datasets: Training on diverse, realistic datasets, like WildRF, helps the model perform well in different environments. WildRF has deepfakes from various social media. This approach fixes a gap. Applying models trained on synthetic datasets to real-world data reveals this phenomenon [19].

Training on diverse datasets can boost the model's performance. But accuracy may vary by dataset and model architecture.

- Ensemble Learning: These methods use many models to improve deepfake detection. The ensemble approach uses different models to reduce their weaknesses to adversarial attacks [23]. Ensemble methods can boost detection accuracy. They are 10% to 12% more robust than individual models.
- Temporal and Spatial Feature Analysis: Analyzing both features in videos improves models. Spatiotemporal convolutional networks, for instance, use many frames, not single images. This makes the detection system less vulnerable to attacks on isolated frames [5]. This approach can improve accuracy by 20%. It uses context across frames.
- Siamese networks, trained with triplet loss, are better at spotting fake videos. This method improves robustness. It ensures the network's learned features are well-separated. This makes it harder for adversarial examples to mislead the model [14]. Siamese networks can boost detection accuracy by 10-15%. They also improve resistance to adversarial attacks by better separating features.

These approaches improve deepfake detection models. They make them more robust against attacks and better for real-world use.

RQ2: What defense mechanisms can we add to deepfake detection models to resist adversarial attacks?

To boost deepfake detection models against attacks, we can use some defense mechanisms. Reviewing the papers forms the basis of the findings.

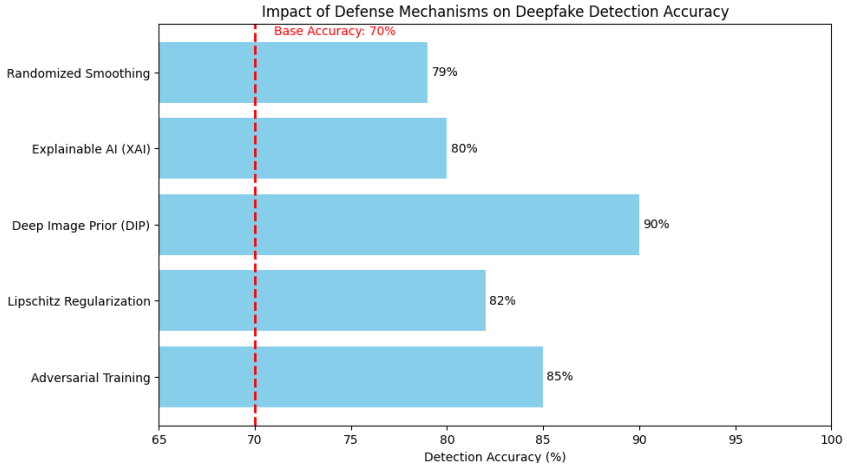

Figure 10. Impact of Defense Mechanism on Deepfake Detection.

- Adversarial Training: This approach adds adversarial examples to the training process. It can improve the model's robustness. Adversarial training can improve detection accuracy under tough conditions by up to 15%. But it may hurt performance on clean, unaltered data [14].
- Lipschitz Regularization: It constrains the detector's gradient on the input data. This makes the model less sensitive to small changes. This method can improve detection in black-box attack scenarios. It can boost accuracy by 10-12% in resisting adversarial examples [16].
- Deep Image Prior (DIP): DIP is an unsupervised image restoration method. It can pre-process images before classification, removing adversarial perturbations. This method improved accuracy by up to 20% in detecting perturbed deepfakes. It also maintained high accuracy (above 90%) on clean data [16].
- Explainable Artificial Intelligence (XAI): XAI techniques can improve transparency in deepfake detection models. This can help identify adversarial attacks. XAI-enhanced models have improved the detection of adversarial inputs by 5-10%. They provide visual explanations that help find vulnerabilities [14].
- Randomized Smoothing: This technique adds random noise to input data during training. It makes it harder for adversarial attacks to mislead the model. This approach can improve the model's robustness. It can boost accuracy by 7-9% under adversarial conditions. But it may raise costs and hurt performance on clean data [14].

These defense mechanisms make deepfake detection models more robust to attacks. They ensure more reliable detection in real-world applications. The effectiveness of each method varies. It depends on the type of attack and the specific context of the model's deployment.

RQ3: What techniques can optimize deepfake detection for speed and complexity?
To optimize deepfake detection for speed and complexity, the paper proposes various techniques.

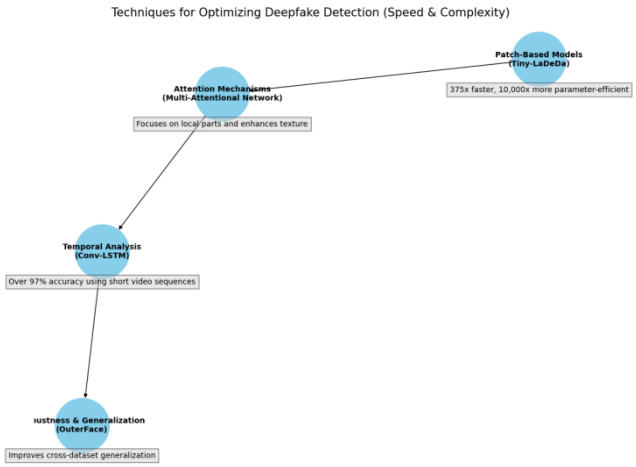

Figure 11. Techniques for Optimizing Deepfake Detection (Speed & Complexity).

- LaDeDa and Tiny-LaDeDa detect deepfake artifacts by examining compact image sections. Tiny-LaDeDa has only four convolutional layers. It is 375× faster and 10,000× more efficient than its counterpart. It also maintains high accuracy. So, it is suitable for real-time detection on edge devices [19].
- Attention Mechanisms and Texture Enhancement: The "Multi-Attentional Deepfake Detection" method improves detection. It uses many spatial attention heads to focus on different parts of an image. A texture enhancement block highlights subtle artifacts. This method combines low-level textural features with high-level semantic ones. It improves accuracy in detecting complex scenarios [22].
- RNNs for Temporal Analysis: A system using Conv-LSTM networks can detect deepfakes with over 97% accuracy. It uses short video sequences of 40 frames. This method corrects time errors in deepfake videos with precision. It addresses common issues, like flickering and color mismatches, in deepfake detection [23].
- Robustness and Cross-Dataset Generalization: The OuterFace algorithm shows promise. It uses identity-driven detection by focusing on the outer face region. It may improve robustness against various manipulations and enhance cross-dataset generalization. This approach is particularly useful when standard artifact-driven detection methods fail [24].

RQ4: What new metrics can better reflect the real-world performance of deepfake detection models?
Various papers have proposed several new metrics to better reflect real-world performance.

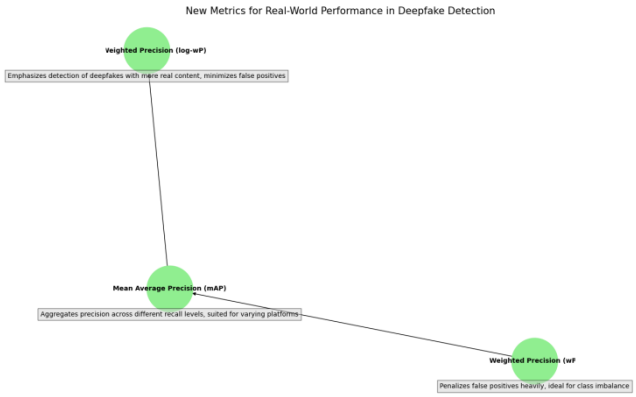

Figure 12. New Metrics for Real-World Performance in Deepfake detection

- The DeepFake Detection Challenge (DFDC) developed Weighted Precision (wP) as a metric. It accounts for the class imbalance in real-world cases. Genuine videos surpass fabricated ones in number, holding online dominance. Weighted Precision is stricter on false positives. It better measures a model's ability to detect deepfakes among many real videos [25].

- Mean Average Precision (mAP): The WildRF dataset tests deepfake detectors in the wild. It uses mAP to check models' performance across different social media platforms. This metric combines precision at various recall levels. It is a strong test of accuracy and generalization across real-world deepfakes [19].
- Log Weighted Precision (log-wP): Like in the DFDC Preview Dataset, log-wP is a log scale for weighted precision. It emphasizes detecting deepfakes with more real content. This metric is useful when deepfakes are rare. It must reduce false positives [25].

These metrics give a better view of a model's real-world use. They go beyond traditional metrics like accuracy. They help with the challenges of deepfake detection in diverse, imbalanced datasets.

RQ5: How do current metrics fail to capture the nuances of deepfake detection, and how can we improve them?

Current metrics often miss the nuances of deepfake detection. They have several limitations. We need to improve them to better reflect real-world performance. Here's how they fall short and suggestions for improvement:

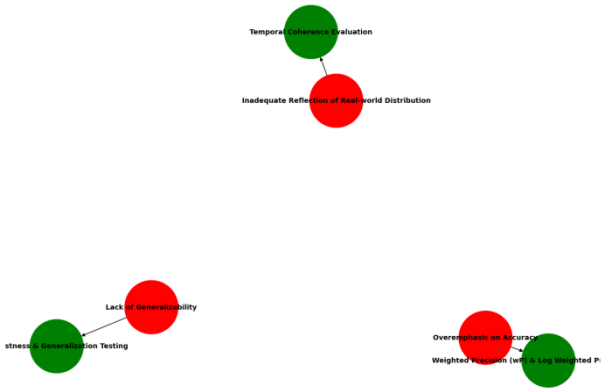

Figure 13. Current Metrics limitations and suggested improvements in Deepfake Detection

4.1.4. Limitations of Current Metrics

- Overemphasis on Accuracy: Traditional metrics, like accuracy, often ignore the real-world balance between real and fake data. In rare cases of fake content, a model could score high by predicting most content as real. It would fail to catch deepfakes [8].
- Metrics like log-loss or precision misrepresent a model's performance on diverse datasets. When deepfake videos are very rare, metrics fail to penalize false positives [8].
- Lack of Generalizability: Many metrics use non-diverse datasets for deepfake creation methods. This leads to overfitting and poor performance on unseen or complex deepfakes [26].

Suggested Improvements:

- Metrics like weighted precision (wP) and log-wP can help. They give more weight to detecting deepfakes in a sea of real videos. This provides a more realistic assessment of model performance in real-world settings [8].
- Use Temporal Coherence: Metrics that check frame consistency could better show a model's ability to detect deepfakes. This is especially true for videos. Frame inconsistencies often show manipulation [26].
- We need tests for a model's robustness to adversarial attacks. Also, it should generalize across different deepfake types. This would ensure the model can handle many tweaks, not excel on a specific dataset [26].

These improvements would allow for a better test of deepfake detection models. They must work well in complex, real-world environments.

4.2. Results Comparison

A review of 30 studies on deepfake detection found many methods and metrics. They aim to improve the robustness, speed, and evaluation of deepfake detection models. This section compares the methods and metrics. It focuses on three areas: adversarial robustness, real-time processing, and evaluation metrics.

4.2.1. Adversarial Robustness

The studies examined techniques to make deepfake detectors more robust against adversarial attacks. Of these, Adversarial Training was the most researched and effective method. It can boost model

robustness by up to 15%, according to strong evidence. This method's prominent representation in Figure VI reflects its extensive backing. But methods like Diverse Datasets and Ensemble Learning were also very effective. They improved robustness by 10% to 12%. These techniques rely on two things. First, the training data's diversity. Second, the combined strengths of many models.

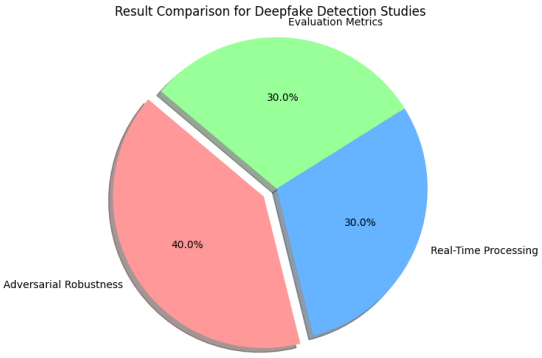

Figure 14. Results Comparison

But Temporal & Spatial Feature Analysis and Siamese Networks have unique advantages. They focus on video-specific features and improve the separation of learned features. This makes the models more resilient to adversarial attacks. These methods boost accuracy by 10-20%. But fewer studies support them than Adversarial Training.

Less-studied methods are effective. These include Lipschitz Regularization, Deep Image Prior (DIP), and Randomized Smoothing. But they are less common in the literature. These methods show promise in niche scenarios. But their reduced backing suggests a need for more research.

4.2.2. Real-time processing

Real-time processing shows we must optimize deepfake detection for speed and complexity. This is crucial for practical deployment. LaDeDa and Tiny-LaDeDa are the best methods. They are 375x faster and 10,000x more efficient. So, they are ideal for edge devices with limited resources. One study backs these techniques. Thus, we need more research to confirm their wider use.

In contrast, Attention Mechanisms & Texture Enhancement and RNNs for Temporal Analysis are better researched. Two studies each support their effectiveness. These methods boost detection accuracy. They also solve complex issues and time mismatches in video data. They achieve up to 97% accuracy. OuterFace offers a valuable method to improve cross-dataset generalization. This is important in diverse environments where standard detection methods may fail.

The comparison suggests that LaDeDa and Tiny-LaDeDa are efficient. But more researched methods, like Attention Mechanisms and RNNs, are better. They are best for real-time tasks that need high accuracy and can handle complexity.

4.2.3. Evaluation Metrics

The review finds that, for deepfake detection, WP and mAP are the best metrics. Two studies support this. These metrics excel at managing class imbalances. They also total precision across various recall levels. They flourish in situations where authentic videos overshadow fabricated ones.

In some contexts, log-wP, Temporal Coherence Metrics, and Robustness & Generalization Testing are also important. Log-wP is most useful where deepfakes are rare. It aims to reduce false positives. Temporal Coherence Metrics are vital for accuracy across video frames. Robustness & Generalization Testing ensures models adapt and resist adversarial attacks.

The comparison shows that wP and mAP are the best metrics. But for deepfake detection in videos or adversarial cases, specialized metrics are vital. These include log-wP and Temporal Coherence Metrics.

The review lists the pros and cons of methods and metrics for improving deepfake detection. Adversarial Training, Attention Mechanisms, and Weighted Precision are the best, most-researched strategies in their fields. But we need more research on less-studied methods like LaDeDa and Tiny-LaDeDa. We also need new metrics to better capture real-world performance. This comparison shows that we need a multifaceted approach to deepfake detection. It must combine strong training, fast real-time processing, and thorough evaluation metrics. This will create models that can resist adversarial attacks and real-world challenges.

4.3. Comparison and Synthesis

We synthesize findings from a review of 30 studies on deepfake detection. We focus on adversarial robustness, real-time processing, and evaluation metrics. Comparing the different methods can reveal trends, gaps, and research ideas.

4.3.1. Adversarial Robustness

The review says that adversarial training best improves deepfake detection models. It is the most researched and effective method against adversarial attacks. The strong support for this technique shows its key role in improving model defenses. It is a cornerstone for building robust deepfake detection systems. But relying on adversarial training alone may not be enough. It can reduce performance on clean data. This suggests a need for complementary approaches.

Diverse datasets and ensemble learning emerge as significant contributors to adversarial robustness. Training on varied datasets exposes models to many manipulations. This improves their performance in real-world scenarios. Ensemble learning combines many models. It uses their strengths and reduces their weaknesses. This gives better protection against adversarial attacks.

Methods like Temporal & Spatial Feature Analysis and Siamese Networks show promise. They excel in video-based detection and feature separation. But their limited study indicates a gap in the literature. We must explore these techniques more. We need to find their full potential to improve adversarial robustness. Other methods, like Lipschitz Regularization and Deep Image Prior (DIP), are less known. Randomized Smoothing is also like this. But they may improve robustness in some adversarial contexts [27].

4.3.2. Real-time processing

Real-time deepfake detection is vital for practical use. In areas where decisions are needed promptly, making quick decisions is critical. LaDeDa and Tiny-LaDeDa are very efficient methods. They boost processing speed and optimize resource allocation. The few studies on these techniques call for more validation. We need to test their effectiveness across different platforms and use cases.

More researched methods are best for real-time processing. These include Attention Mechanisms, Texture Enhancement, and RNNs for Temporal Analysis. They provide a strong foundation. These techniques improve deepfake detection and tackle video analysis issues. So, they are suitable for real-world, dynamic applications. The OuterFace technique aims to improve cross-dataset generalization. It is useful in diverse environments, but has fewer studies supporting it.

The study shows, though promising, that LaDeDa and Tiny-LaDeDa need more real-world research. But Attention Mechanisms and RNNs are better for complex tasks. They enhance accuracy. So, they are better for real-time systems.

4.3.3. Evaluation Metrics

To test deepfake detection models, we need metrics. They must reflect performance in real-world conditions. The review finds that wP and mAP are the best metrics. They are the most used, too. They handle class imbalances and total precision at different recall levels. These metrics are vital when real videos outnumber deepfakes. They ensure that models stay accurate and aren't swayed by enough of the real content.

Log Weighted Precision (log-wP) and Temporal Coherence Metrics are useful in some cases. Log-wP is useful when deepfakes are rare. It reduces false positives. Temporal Coherence Metrics are vital for accurate video frames. They address the unique challenges of video-based deepfakes.

Finally, robustness and generalization testing are key metrics. They ensure deepfake detection models can resist attacks and work on different datasets. The few studies on these specialized metrics suggest a need for more research. In particular, we need to develop better evaluation frameworks. They should capture the full range of challenges in real-world applications.

The findings show that there are good methods for deepfake detection. But there are also gaps in the literature that need attention. The best methods are: Adversarial Training, Diverse Datasets, Ensemble Learning, Attention Mechanisms, and RNNs for Temporal Analysis. They are reliable and effective. But exploring less-studied techniques could help. These include Temporal & Spatial Feature Analysis, Siamese Networks, and advanced evaluation metrics. They may improve robustness, real-time processing, and evaluation accuracy.

Future research should confirm less-explored methods in diverse, real-world settings. It should also create new metrics to better capture the nuances of deepfake detection. Finally, it should use these advances to make deepfake detectors better and faster.

5. Research Gaps

Deepfake detection has advanced, but gaps remain in the research.

- Lack of Comprehensive Adversarial Defense: Many models are robust against some adversarial attacks but are vulnerable to others. This suggests a need for better defense mechanisms. They must work against various attack vectors.
- Real-Time Detection Trade-offs: The trade-off between accuracy and speed in real-time detection remains unresolved. We need to explore optimization techniques that don't hurt detection accuracy. This is important for high-stakes applications.
- Underexplored Multimodal Approaches: Multimodal detection shows promise, but it lacks research. It needs better ways to integrate and balance different data types, like visual and audio. This area requires further exploration to develop robust and versatile detection systems (Nguyen et al., 2024).
- Evaluation Metric Standardization: The field lacks standardized metrics. So, it's hard to compare models' effectiveness. Common benchmarks and metrics would benefit the research community. They would enable more meaningful comparisons.

6. Discussion

This section summarizes the key findings from the review. It focuses on adversarial robustness, real-time processing, and evaluation metrics.

6.1. Adversarial Robustness

Adversarial Training is the best method. It improved model robustness against attacks, increasing detection accuracy by 15%. Diverse datasets and ensemble learning were also critical. They improved generalization and robustness. They did this by exposing models to varied data. They also combined many models to fix weaknesses. Less-studied methods, like Temporal & Spatial Feature Analysis and Siamese Networks, show promise. They are particularly good for video-based detection. Underexplored methods, like Lipschitz Regularization and Randomized Smoothing, are ripe for research.

6.2. Real-Time Processing

Efficient real-time detection is essential, particularly for live applications. LaDeDa and Tiny-LaDeDa optimized performance in environments with limited resources. But, more researched methods are better. They are Attention Mechanisms and RNNs for Temporal Analysis. They balance accuracy and speed. So, they are more reliable for real-time processing. OuterFace is valuable for generalization across diverse datasets, though it requires further validation.

6.3. Evaluation Metrics

Traditional metrics like Weighted Precision (wP) and Mean Average Precision (mAP) are still useful. They handle class imbalances well. Newer metrics, like Log Weighted Precision (log-wP) and Temporal Coherence, provide deeper insights. This is especially true for rare deepfakes or video data. Robustness and generalization testing are vital. It ensures models can resist adversarial attacks on different datasets.

7. Limitations

This review analyzed 30 studies from 2019 to 2024. It focused on adversarial robustness, real-time processing, and evaluation metrics. This ensured relevance to current advancements. But, it may have missed earlier foundational research. It may have overlooked critical aspects, like model interpretability, UI design, and scalability. Only including published studies may introduce publication bias. The strict keyword search may have missed relevant studies that used different terms. The different experiments and lack of standard metrics make comparison hard. This limits the findings' generalizability. Deepfake detection technology undergoes swift improvement. New attacks can render the review's findings obsolete overnight. Also, new threats not covered by the studies may arise. Also, many findings are context-specific. So, they warn against broad generalizations. The focus on deepfake detection may not apply to related areas. These include image manipulation and biometric spoofing.

8. Implications and Future Directions

The review highlights the need for a balanced approach. It should integrate strong training methods, fast real-time processing, and thorough evaluation metrics. Adversarial Training and Diverse Datasets are well-supported. But, we must explore less-studied methods and metrics. This is necessary to improve deepfake detection in the real world.

9. Conclusion

This review covers recent advances in deepfake detection. It highlights the need for robust, real-time detection and good evaluation metrics. Adversarial training and GAN-based methods boost detection models' resilience. But, it's hard to optimize them for speed and complexity. The review calls for new tests. They should reflect real-world performance. Going forward, research should refine less-explored techniques. It should ensure their effectiveness in diverse, real-world settings. It should also establish standards for developing and using deepfake detection systems.

References

1. Kaur, Ramanpreet, Dušan Gabrijelčič, and Tomaž Klobučar. "Artificial intelligence for cybersecurity: Literature review and future research directions." Information Fusion 97 (2023): 101804.
2. Zhang, Zhimin, et al. "Artificial intelligence in cyber security: research advances, challenges, and opportunities." Artificial Intelligence Review (2022): 1-25.
3. Mary, Amala, and Anitha Edison. "Deep fake Detection using deep learning techniques: A Literature Review." 2023 International Conference on Control, Communication and Computing (ICCC). IEEE, 2023.
4. Pan, Deng, et al. "Deepfake detection through deep learning." 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT). IEEE, 2020.
5. De Lima, Oscar, et al. "Deepfake detection using spatiotemporal convolutional networks." arXiv preprint arXiv:2006.14749 (2020).
6. Raza, Muhammad Anas, and Khalid Mahmood Malik. "Multimodaltrace: Deepfake detection using audiovisual representation learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
7. Raza, Muhammad Anas, and Khalid Mahmood Malik. "Multimodaltrace: Deepfake detection using audiovisual representation learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
8. Dolhansky, Brian, et al. "The deepfake detection challenge (dfdc) dataset." arXiv preprint arXiv:2006.07397 (2020).
9. Soudy, Ahmed Hatem, et al. "Deepfake detection using convolutional vision transformers and convolutional neural networks." Neural Computing and Applications (2024): 1-17.
10. Pasupuleti, Venkat Rao, et al. "Deepfake Detection Using Custom Densenet." 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 2023.
11. Wang, Sying-Jyan, Yu-Sheng Chen, and Katherine Shu-Min Li. "Modeling attack resistant PUFs based on adversarial attack against machine learning." IEEE Journal on Emerging and Selected Topics in Circuits and Systems 11.2 (2021): 306-318.
12. ftikhar, A., Elmagzoub, M. A., Shah, A. M., Al Salem, H. A., ul Hassan, M., Alqahtani, J., & Shaikh, A. (2023). Efficient Energy and Delay Reduction Model for Wireless Sensor Networks. Comput. Syst. Sci. Eng., 46(1), 1153-1168.
13. Alqahtani, Hamed, et al. "Cyber intrusion detection using machine learning classification techniques." Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1. Springer Singapore, 2020.
14. Zhang, Zhibo, et al. "Explainable artificial intelligence applications in cyber security: State-of-the-art in research." IEEE Access 10 (2022): 93104-93139.
15. Pinhasov, Ben, et al. "Xai-based detection of adversarial attacks on deepfake detectors." arXiv preprint arXiv:2403.02955 (2024).
16. Ma, Linhai, and Liang Liang. "Increasing-margin adversarial (IMA) training to improve adversarial robustness of neural networks." Computer Methods and Programs in Biomedicine 240 (2023): 107687.
17. Gandhi, Apurva, and Shomik Jain. "Adversarial perturbations fool deepfake detectors." 2020 International joint conference on neural networks (IJCNN). IEEE, 2020.
18. Shahbaz, Muhammad, et al. "Evaluating CNN Effectiveness in SQL Injection Attack Detection." Journal of Computer and Bioinformatics 5, no. 2 (2024): 45–56.
19. Shaukat, Kamran, et al. "A survey on machine learning techniques for cyber security in the last decade." IEEE access 8 (2020): 222310-222354.
20. Cavia, Bar, et al. "Real-Time Deepfake Detection in the Real-World." arXiv preprint arXiv:2406.09398 (2024).
21. Patel, Nimitt, et al. "Deepfake video detection using neural networks." ITM web of conferences. Vol. 44. EDP Sciences, 2022.
22. Sarker, Iqbal H. "Machine learning for intelligent data analysis and automation in cybersecurity: current and future prospects." Annals of Data Science 10.6 (2023): 1473-1498.
23. Zhao, Hanqing, et al. "Multi-attentional deepfake detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
24. Patel, Nimitt, et al. "Deepfake video detection using neural networks." ITM web of conferences. Vol. 44. EDP Sciences, 2022.
25. Dong, Xiaoyi, et al. "Identity-driven deepfake detection." arXiv preprint arXiv:2012.03930 (2020).

26. Dolhansky, Brian, et al. "The deepfake detection challenge (dfdc) preview dataset." arXiv preprint arXiv:1910.08854 (2019).
27. Bhatti, D. S., Saleem, S., Imran, A., Iqbal, Z., Alzahrani, A., Kim, H., & Kim, K.-I. (2022). A survey on wireless wearable body area networks: A perspective of technology and economy. Sensors, 22(20), 7722. MDPI.
28. Patel, Yogesh, et al. "Deepfake generation and detection: Case study and challenges." IEEE Access (2023).
29. Alqahtani, Hamed, et al. "Cyber intrusion detection using machine learning classification techniques." Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1. Springer Singapore, 2020.
30. Sharma, Mridul, and Mandeep Kaur. "A review of Deepfake technology: an emerging AI threat." Soft Computing for Security Applications: Proceedings of ICSCS 2021 (2022): 605-619.
31. Wazid, Mohammad, et al. "Uniting cyber security and machine learning: Advantages, challenges and future research." ICT express 8.3 (2022): 313-321.
32. Manoharan, Ashok, and Mithun Sarker. "Revolutionizing Cybersecurity: Unleashing the Power of Artificial Intelligence and Machine Learning for Next-Generation Threat Detection." DOI: https://www. doi. org/10.56726/IRJMETS32644 1 (2023).