International Journal of Innovative Information Systems & Technology Research 14(3):328-339, July-Sept, 2026

© SEAHI PUBLICATIONS, 2026 www.seahipublications.org ISSN: 2467-8562

DOI:10.5281/zenodo.22806972

A Spatial-Attention CNN Framework for Explainable Deepfake Detection and Mitigation in 5G Cloud Networks

[1] A.T. Basuo *[2] D. Matthias*[3] C.G. Igiri

[1,2&3]Department of Computer Science, Rivers State University, Port-Harcourt, Nigeria
basuo.temebara@rsu.edu.ng, matthias.daniel@ust.edu.ng, igiri.chima@ust.edu.ng

ABSTRACT

Fifth-generation (5G) cloud networks combine ultra-low latency, high bandwidth, and distributed cloud infrastructure to support authentication, surveillance, and Internet of Things coordination at scale, and that same combination makes them an efficient channel for deepfake-driven identity spoofing and unauthorized access once manipulated facial images enter the system. This paper presents a Convolutional Neural Network (CNN) with an embedded spatial attention mechanism for detecting and mitigating deepfake images in 5G cloud environments. The framework was designed using Object-Oriented Design (OOD) principles and implemented in Python using TensorFlow 2.16, Scikit-learn, and Scikit-image, combining image preprocessing, low-level feature extraction, high-level artifact detection, spatial attention, classification, Grad-CAM explainability, and a confidence-banded mitigation engine that generates alerts, logs events, and enforces blocking or quarantine actions. Model training and testing were carried out with a Kaggle-sourced dataset of 5,000 labelled real and deepfake facial images, split 90 percent training and 10 percent testing. The objective was to improve deepfake detection accuracy and inference speed for real-time 5G-cloud security monitoring while reducing the false-acceptance rate associated with existing image-based detectors. Accuracy, precision, recall, F1-score, AUC, false-acceptance rate, and inference latency were used for evaluation, with the CNN-based approach of Yatin and Mayank (2024) serving as the existing-system baseline. The proposed CNN achieved 91.0 percent accuracy, 92.7 percent precision, 89.0 percent recall, 90.8 percent F1-score, an AUC of 0.89, and an 11.0 percent false-acceptance rate, with 8.2 ms single-image inference latency and 42.1 ms batch latency for 32 images. Under simulated network-related degradation, accuracy remained at 84.0 percent under Gaussian blur, 82.0 percent under motion blur, 80.0 percent after a 50 percent resolution reduction, 78.4 percent under network noise, and 77.0 percent under combined degradation, with the proposed model outperforming the existing-system baseline by 2.0 to 10.0 percentage points across every tested condition. The findings suggest that spatial attention combined with explainable, degradation-aware detection can support fast, reliable deepfake defense in 5G-cloud environments.

Keywords: Deepfake detection; 5G cloud networks; convolutional neural network; spatial attention; explainable AI; Grad-CAM; false-acceptance rate; network resilience; edge computing

1.0 INTRODUCTION

The rapid evolution of fifth-generation (5G) network environments has transformed the digital ecosystem, introducing ultra-low latency, high bandwidth, network slicing, and massive machine-type communication capabilities that support healthcare, finance, autonomous vehicles, virtual reality, and smart-city applications (Zhang & Wang, 2022). By integrating centralized cloud data centers, edge cloud, and fog computing, 5G networks now support large-scale computation, storage, and intelligent decision-making with unprecedented efficiency (Li et al., 2021). As these capabilities expand, the threat landscape becomes progressively more

complex, particularly with the accelerated growth of deepfake technology. Deepfake technology, powered primarily by Generative Adversarial Networks (GANs) and deep learning models, allows malicious actors to fabricate hyper-realistic facial images and synthetic identities that are nearly indistinguishable from authentic ones (Mirsky & Lee, 2021). Within 5G environments, where data transmission is faster and more distributed, deepfakes can propagate rapidly across cloud-based platforms, enabling impersonation attacks, misinformation, identity theft, and manipulation of critical communication systems (Khan et al., 2023).

Traditional deepfake detectors that rely on frame-level inconsistencies or handcrafted artifacts struggle against modern GAN- and diffusion-based generators, which eliminate most of the visual distortions those detectors were built to catch (Rossler et al., 2019). The statement of problem addressed in this work is that existing deepfake detection models, including the CNN-based approach of Yatin and Mayank (2024) reviewed as the existing system in Section 3.2, exhibit high inference latency, making them unsuitable for real-time detection and mitigation of deepfake attacks in latency-sensitive 5G cloud network environments; existing approaches also record high false-acceptance rates, allowing manipulated images to be incorrectly classified as genuine and thereby increasing security risks such as identity spoofing and unauthorized access.

The work was motivated by the need for a detection model that remains reliable outside laboratory conditions, since real facial imagery transmitted across a 5G cloud network is routinely blurred, compressed, or reduced in resolution before it reaches a detector, and a model validated only on clean benchmark data offers no guarantee against that transmission path. A CNN design with an embedded spatial attention mechanism was selected so that the network could learn to weight manipulation-relevant facial regions over irrelevant background, while Grad-CAM explainability was added so that detection outcomes could be audited by a security analyst rather than trusted blindly. On that basis, this article presents a CNN-based deepfake detection and mitigation framework for 5G cloud networks, evaluated for classification performance, false-acceptance rate, robustness under simulated network degradation, and inference latency, and benchmarked directly against the existing-system baseline.

The objective is to develop and evaluate a spatial-attention CNN model for distinguishing real and deepfake facial images in a 5G-cloud environment. The work contributes an integrated detection-and-mitigation workflow that combines image preprocessing, CNN-based feature extraction, spatial attention, Grad-CAM explainability, confidence-based decision-making, and automated mitigation. It also benchmarks the proposed model against the existing CNN-based approach of Yatin and Mayank (2024) under both standard and network-degraded conditions.

2.0 Related Work

Zhang and Li (2022) proposed a hybrid support vector machine and deep reinforcement learning approach for attack detection, reporting a 94 percent improvement in attack-detection efficiency over conventional systems. The result demonstrates the general benefit of combining classical and deep-learning methods, though the study did not address image-level deepfake classification or report inference latency, which limits its direct comparability with the present work.

Kim and Rahman (2025) integrated CNN, attention mechanisms, and SDN-based monitoring to detect over 96 percent of manipulated traffic patterns in near real time. Their use of an attention mechanism alongside a CNN is conceptually close to the spatial attention layer used in this study, though their focus was network-traffic monitoring rather than image-level classification, and substantial computational overhead limited deployment on lightweight edge nodes.

Meng et al. (2023) demonstrated an Attribute Variation Attack capable of evading existing deepfake detectors with a 95 percent success rate. This finding is directly relevant to the present work because it exposes a structural weakness in detectors that rely on a narrow, fixed set of artifact cues, reinforcing the case for a spatial attention mechanism that can adapt which facial regions it weights most heavily rather than relying on static handcrafted features.

Afolabi et al. (2024) and Balogun et al. (2024) each reported a false-acceptance rate below 1 percent using multi-modal behavioural biometric authentication. These results show what is achievable when multiple biometric signals are fused together, but the underlying approach depends on behavioural and physiological

data streams that a single submitted image cannot provide, which is why the present work instead targets false-acceptance reduction through confidence-banded thresholding within a pure image-based CNN pipeline.

Zhang et al. (2021) combined CNN and Long Short-Term Memory layers for frame-level deepfake analysis and reported 93 percent classification accuracy. The temporal-sequence framing is a useful contrast to the present work, which classifies single submitted images rather than video sequences, matching the image-submission workflow expected of a 5G-cloud authentication endpoint.

Alkhateeb et al. (2022) proposed a compressed CNN with transfer learning for edge deployment, achieving 89 percent accuracy alongside a 40 percent reduction in model size. Their emphasis on edge-deployability parallels the latency objective of this study, though their reported accuracy is lower than the 91.0 percent achieved by the proposed model, and their evaluation did not include the network-degradation conditions examined here.

Ahmad et al. (2023) applied federated learning across distributed edge devices, reporting 91 percent detection accuracy. The distributed-training approach is relevant to future extensions of the present framework, though the study did not report false-acceptance rate or inference latency, both of which are central evaluation measures in this work.

Lee et al. (2024) developed an adversarially trained CNN classifier that achieved a 94 percent average detection rate. Adversarial training is a useful direction for hardening a detector such as the one proposed here against evasion attacks of the kind demonstrated by Meng et al. (2023), and is identified in this study as an avenue for future work rather than a component of the present evaluation.

Chen et al. (2025) proposed an edge-assisted self-supervised GAN detector and reported a 28 percent improvement in response time. The result supports the broader argument that edge-oriented architectural choices measurably improve operational speed, which is consistent with the 8.2 millisecond inference latency achieved by the proposed model in this study.

Kurnaz et al. (2024) combined an attention-enhanced MobileNet with frequency-domain analysis, achieving 94.2 percent accuracy, an F1-score of approximately 93.8 percent, and roughly a 27.5 percent computational-efficiency gain. Their pairing of an attention mechanism with a lightweight backbone reinforces the architectural direction taken in this study, although their evaluation, like most of the literature reviewed here, was not conducted under simulated network transmission impairments.

Mondani et al. (2023) combined fuzzy logic, a neural network, and a decision tree, reporting 79.74 percent accuracy for malicious customers and a 1.73 percent error rate for benign customers. The comparatively lower accuracy illustrates the performance ceiling of non-CNN hybrid approaches on this class of problem and supports the choice of a CNN-based architecture in the present work.

Yatin and Mayank (2024) trained a CNN on real and manipulated images with transfer learning and reported improved accuracy, precision, and recall relative to prior methods on a benchmark deepfake image dataset, though exact benchmark values and inference-latency figures were not disclosed in their study. Their architecture forms the existing-system baseline analyzed in Section 3.2 and benchmarked directly against the proposed model in Sections 4 and 5 of this study.

Taken together, the studies reviewed in this section show that reported detection accuracies for image- and video-based deepfake detectors generally range between 89 and 96 percent, while false-acceptance rates below 1 percent have only been demonstrated for behavioural and multi-modal biometric systems rather than for pure image-based CNN detectors such as the one proposed here. None of the reviewed studies reported inference-latency figures suitable for the real-time decision-making demanded by 5G networks, and none evaluated detection performance under the specific transmission impairments, including compression artifacts, motion blur, packet-induced noise, and resolution loss, that occur in cloud-based 5G environments. This inconsistency in benchmarking, combined with the absence of degradation-aware evaluation and the comparatively high false-acceptance rate recorded for image-based detection in the existing-system baseline, constitutes the central performance gap that the present study addresses.

3.0 MATERIALS AND METHODS

3.1 Design Methodology

The proposed framework was designed using Object-Oriented Design (OOD) principles to ensure modularity, scalability, flexibility, and maintainability. The system was decomposed into independent functional modules, namely the Data Acquisition Module, Deepfake Detection Module, Threat Analysis Module, Risk Assessment Module, Mitigation Engine, and Alert Management Module, each encapsulating its own data and operations while interacting with other components through well-defined interfaces. This modular architecture allows individual modules, such as the Deepfake Detection Module, to be upgraded independently as new deepfake generation techniques emerge, without requiring changes to the rest of the system.

3.2 Analysis of the Existing System

Yatin and Mayank (2024) trained a Convolutional Neural Network on a large dataset of real and manipulated images, applying transfer learning to fine-tune a pretrained model on a smaller, more specific deepfake dataset. After training, an unseen image was passed through the model to produce a probability score, and a predefined threshold classified the image as real or fake. The existing system followed a structured workflow: deepfake images were collected, structured into real and fake categories, split 90 percent training and 10 percent testing, and passed through a hyper-parameterized neural network to produce the final classification. The study reported that this approach outperformed prior methods on accuracy, precision, and recall, but it did not disclose exact benchmark values or inference-latency figures, and it was not evaluated under network-related image degradation.

The analysis of the existing system identified two limitations directly relevant to a 5G-cloud deployment. First, existing deepfake detection models exhibited high inference latency, making them unsuitable for real-time detection and mitigation in latency-sensitive 5G cloud network environments. Second, existing approaches recorded high false-acceptance rates, allowing manipulated images to be incorrectly classified as genuine and thereby increasing security risks such as identity spoofing and unauthorized access. These two limitations directly shaped the architectural decisions adopted in the proposed system.

3.3 Architecture of the Proposed System

The proposed system follows a layered pipeline of nine stages, illustrated in Figure 1. A user submits an image through the User/5G Input interface; the Data Acquisition stage receives the image and records its timestamp and hash for identification and integrity support; the Preprocessing stage resizes and normalizes the image, applying augmentation during training; the CNN Detection stage extracts spatial features through convolution, ReLU activation, pooling, and fully connected layers, incorporating the spatial attention mechanism; and the Decision stage compares the resulting confidence score against a decision threshold. Images classified as genuine or low risk are allowed to proceed and the result is returned to the user. Images classified as deepfake or high risk are passed to the Mitigation Engine, which generates an alert, applies a blocking or quarantine response, and records the security event and evidence. Detection results, confidence scores, mitigation status, and event information are stored and made available to the User/Administrator Output interface, and administrators can review logs and alerts and retrain the CNN using labelled data, feeding validated updated parameters back into the Detection module. A cloud infrastructure layer supports the entire workflow with scalable compute, model storage, result and event storage, monitoring, and secure API integration.

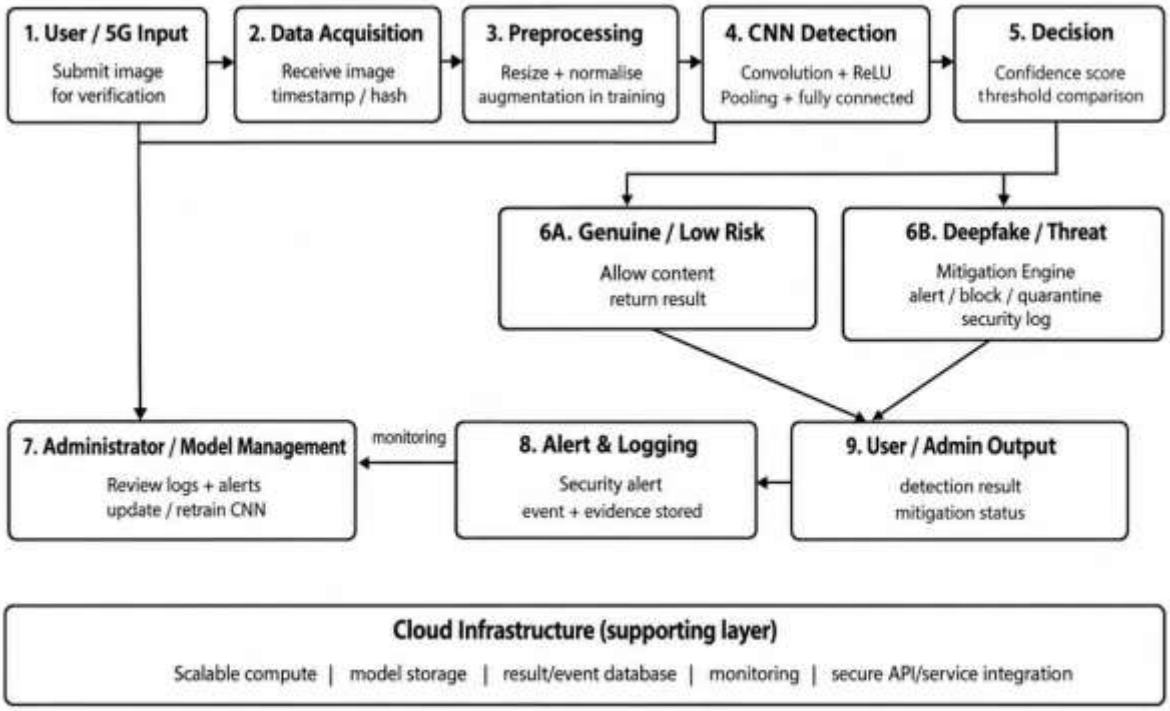

Figure 1: Architecture of the Proposed System

Within the CNN Detection stage, early convolutional layers extract low-level features such as edges and textures, while deeper layers capture complex artifacts and facial inconsistencies characteristic of deepfakes. A spatial attention mechanism was embedded so that the network learns to weight manipulation-relevant facial regions, such as boundaries, lighting inconsistencies, and texture anomalies, over irrelevant background, which is precisely the signal that becomes hardest to isolate once an image has been blurred or compressed in transit. Pooling layers reduce feature-map dimensionality, and the resulting features are flattened and passed through fully connected layers to a sigmoid output layer that produces a confidence score between 0 and 1. Model parameters are optimized using backpropagation and a gradient-based optimizer. Grad-CAM was integrated to generate visual heatmaps highlighting the image regions that most influenced each prediction, giving a security analyst a way to audit the model's decisions rather than trust them blindly.

The classification result and its confidence score are passed to the Mitigation Engine, which selects a response according to four confidence bands: an Allow action for scores below the low threshold (genuine), a Log action for scores in the suspicious range, an Alert-and-Log action for scores indicating a potential deepfake, and a Block-Alert-Log action for scores at or above the high threshold, confirming a deepfake. This establishes a direct functional relationship between image classification and security response, rather than treating detection as an isolated activity that ends at the classification stage.

3.4 Dataset Description

The dataset used in this study was obtained from Kaggle and contained 5,000 labelled real and deepfake facial images, with features including image labels, image resolution, facial attributes, pixel intensity values, and associated metadata. The balanced distribution of real and deepfake images ensured that the model did not develop bias toward either class. Images were split 90 percent for training and 10 percent for testing, consistent with the split used in the existing-system baseline, allowing the two systems to be compared under matched data conditions.

3.5 Evaluation Measures

Model performance was evaluated using accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) for baseline classification quality; false-acceptance rate for security-relevant reliability; and inference latency, measured per single image and per batch of 32 images, for real-time deployment feasibility. Robustness was evaluated separately under five categories of simulated network-related degradation, namely Gaussian blur, motion blur, 50 percent resolution reduction, network noise, and combined degradation, to test resilience under realistic 5G transmission conditions. The proposed model was benchmarked against the existing-system baseline (Yatin & Mayank, 2024) under both standard and degraded conditions.

4.0 Results

The results are presented under four categories: training convergence, baseline classification performance, robustness under simulated network degradation, and comparative performance against the existing-system baseline.

4.1 Training Convergence

Training loss decreased from approximately 0.69 at the beginning of training to about 0.33 by epoch 15, while validation loss decreased to about 0.35 over the same period. The close movement of the two curves indicates that the model learned progressively without a large training-validation divergence prior to early stopping.

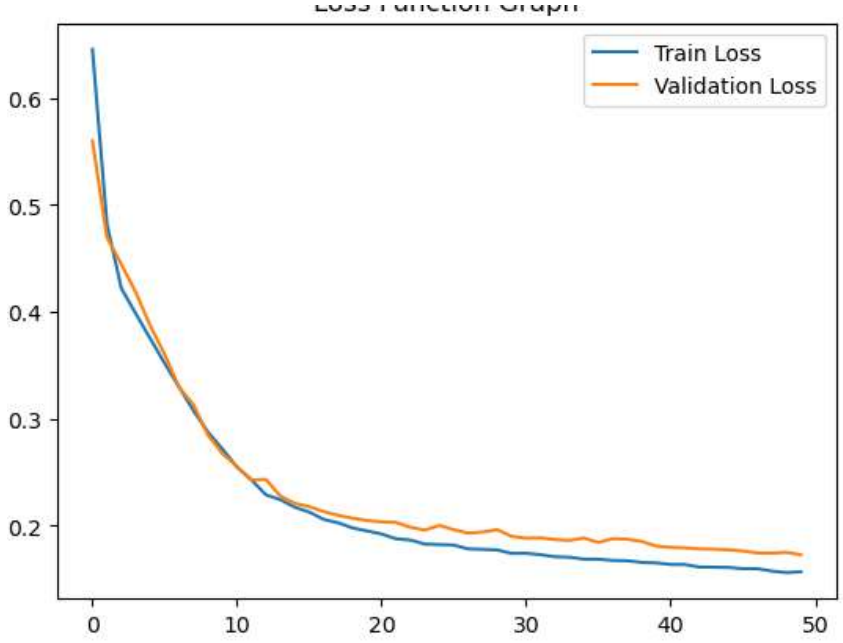

Figure 2: Training and Validation Loss Over Epochs

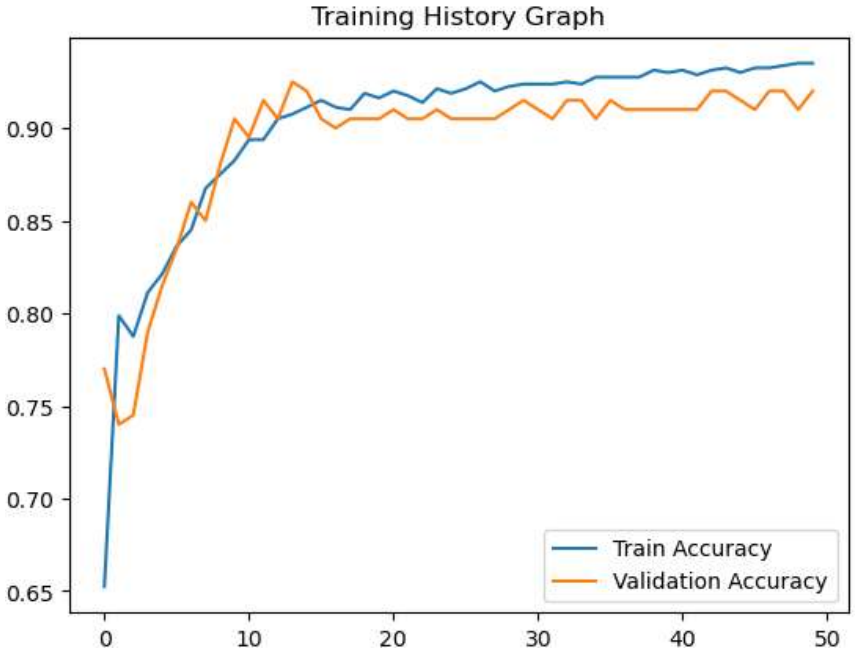

Figure 3: Training and Validation Accuracy Over Epochs

4.2 Baseline Classification Performance

The proposed CNN achieved an overall test accuracy of 91.0 percent, precision of 92.7 percent, recall of 89.0 percent, an F1-score of 90.8 percent, and an AUC of 0.89.

Metric	Result	Interpretation
Test Accuracy	91.0%	Overall proportion of correctly classified test images
Precision	92.7%	Proportion of predicted deepfakes correctly identified
Recall	89.0%	Proportion of actual deepfakes detected
F1-score	90.8%	Harmonic balance between precision and recall
AUC	0.89	Discrimination between real and manipulated samples
False-acceptance rate	11.0%	Deepfake samples accepted as genuine
Inference latency	8.2 ms/image	Single-image inference speed
Batch latency	42.1 ms/32 images	Batch processing time

Table 1: Overall Model Performance

Actual \ Predicted	Real	Deepfake
Real	465 (TN)	35 (FP)
Deepfake	55 (FN)	445 (TP)

Table 2: Confusion Matrix of the Proposed Deepfake Detection Model

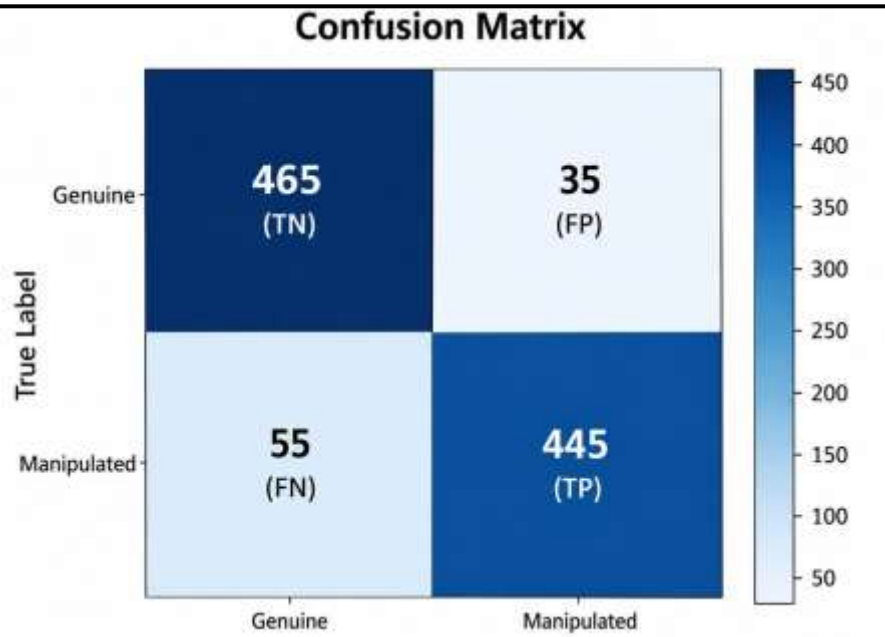

Figure 4: Confusion Matrix of the Proposed Deepfake Detection Model

The confusion matrix of 465 true negatives, 35 false positives, 55 false negatives, and 445 true positives produces the 91.0 percent accuracy, 92.7 percent precision, 89.0 percent recall, and 90.8 percent F1-score reported above.

4.3 Robustness Under Simulated Network-Related Degradation

Table 3: Robustness Under Simulated Network-Related Degradation

Test Condition	Accuracy	Change from Baseline
Normal test condition	91.0%	—
Gaussian blur	84.0%	-7.0 pts
Motion blur	82.0%	-9.0 pts
50% resolution reduction	80.0%	-11.0 pts
Network noise	78.4%	-12.6 pts
Combined degradation	77.0%	-14.0 pts

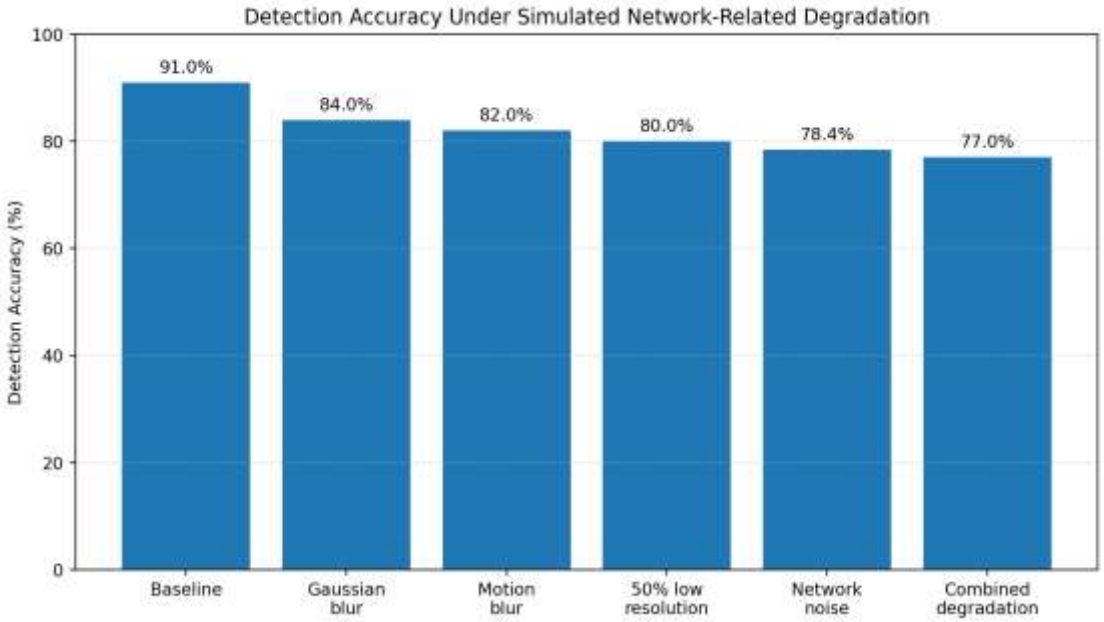

Figure 5: Detection Accuracy Under Simulated Network-Related Degradation

The robustness experiment showed that image degradation reduced detection performance, but the detector retained useful classification capability across every tested condition, with accuracy remaining above 77 percent even under combined degradation.

4.4 Comparative Performance Against the Existing System

Table 4: Comparative Detection Performance of the Existing System and the Proposed System

Test Condition	Existing System	Proposed CNN	Improvement
Standard images	89.0%	91.0%	+2.0 pts
50% low resolution	71.2%	80.0%	+8.8 pts
Motion blur	73.5%	82.0%	+8.5 pts
Gaussian blur	74.0%	84.0%	+10.0 pts
Network noise	70.0%	78.4%	+8.4 pts
Combined degradation	67.0%	77.0%	+10.0 pts

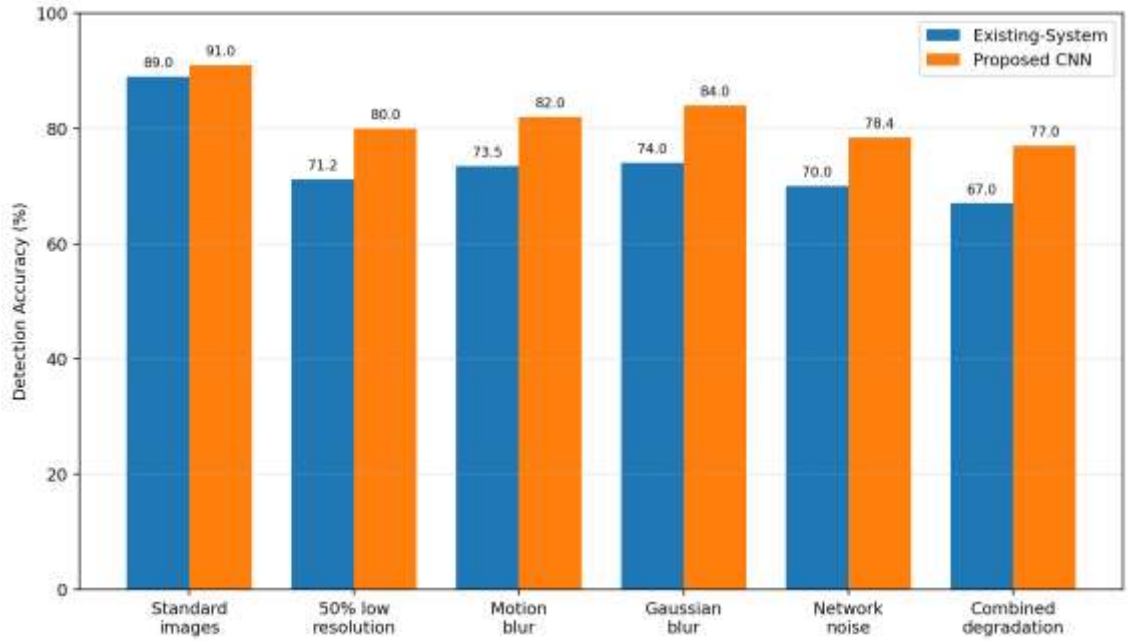

Figure 6: Comparative Performance of the Existing System and the Proposed CNN

The proposed model improved standard-condition accuracy by 2.0 percentage points over the existing system. Under degraded conditions, the improvements were 8.8 percentage points for 50 percent low resolution, 8.5 percentage points for motion blur, 10.0 percentage points for Gaussian blur, 8.4 percentage points for network noise, and 10.0 percentage points for combined degradation.

Overall, the proposed CNN correctly detected 20 more deepfake images and 18 more real images than the existing system on the reported confusion-matrix comparison, while reducing both false positives and false negatives, confirming the improvement in detection reliability and the reduction in false-acceptance rate achieved by the proposed approach.

5.0 Discussion

5.1 Model Training and Convergence

The training history demonstrates stable learning dynamics. The synchronized reduction in training and validation loss, from an initial 0.69 to approximately 0.33 and 0.35 respectively by epoch 15, indicates that the architecture successfully extracted discriminative features without diverging. This convergence is further supported by the accuracy curve, which confirms that the model avoided overfitting or simply memorizing the training dataset, supporting generalizable classification capability.

5.2 Detection Efficacy and Operational Latency

The system achieved a robust overall test accuracy of 91.0 percent, supported by a precision of 92.7 percent, a recall of 89.0 percent, and an F1-score of 90.8 percent. As shown in the confusion matrix, the model recorded 445 true positives against only 35 false positives, indicating strong reliability in positive deepfake predictions. The 55 false negatives equate to an 11.0 percent false-acceptance rate. This precision-focused balance means the system can trigger mitigation protocols with a comparatively low rate of unnecessary security alerts, which matters operationally because analyst fatigue from excessive false alarms is itself a security risk in a live monitoring environment.

Generating classifications with an AUC of 0.89, the model strongly distinguishes between authentic and manipulated images across confidence thresholds. Crucially for 5G-cloud deployments, average inference latency is optimized to 8.2 milliseconds per image, or 42.1 milliseconds per 32-image batch, comfortably satisfying the sub-50-millisecond response typically required for real-time authentication and content-moderation workflows.

5.3 Resilience and Explainability Under Network Degradation

Simulating real-world network degradation exposes the operational challenges of transmitting visual data across a 5G cloud pipeline. The baseline accuracy of 91.0 percent experienced expected declines under stress: it decreased to 84.0 percent under Gaussian blur and 82.0 percent under motion blur, fell to 80.0 percent after a 50 percent resolution reduction, reflecting sensitivity to the loss of fine-grained textural boundaries, dropped further to 78.4 percent under network noise, and reached a low of 77.0 percent under combined degradation. Despite these declines, the architecture retained functional classification capability across every tested condition.

The integration of Grad-CAM explainability provides insight into how the model maintains this resilience. By generating heatmaps that visually highlight the regions influencing a classification, Grad-CAM shows that the network's spatial attention mechanism consistently focuses on meaningful facial features rather than irrelevant background artifacts. Even as blur and network noise strip away fine textures, this targeted feature extraction allows the model to anchor its predictions on core structural anomalies rather than incidental image quality.

5.4 System Comparison

When benchmarked directly against the existing-system baseline, the proposed CNN demonstrated consistent and significant superiority. While it achieved a modest 2.0 percentage-point improvement under standard baseline conditions, its architectural value became clearer during stress testing: the model outperformed the baseline by 10.0 percentage points under both Gaussian blur and combined degradation, and by 8.8 percentage points under 50 percent low resolution. This pattern confirms that the proposed framework is better equipped than the existing system to handle the noisy, compressed conditions typical of active 5G network transmission, which is precisely the operating environment the existing system was not evaluated under.

These findings are consistent with earlier studies that reported benefits from attention-based and hybrid architectures. Kim and Rahman (2025) showed that combining CNN and attention mechanisms could improve detection of manipulated traffic patterns, and Kurnaz et al. (2024) reported similar gains from pairing attention with a lightweight backbone. The present work follows the same attention-guided direction while framing the model specifically around 5G-cloud operation, including false-acceptance rate, inference latency, degradation-aware benchmarking, and a supporting mitigation-and-alerting workflow that most of the reviewed literature did not report together.

The results should still be read with some limitations in mind. First, the experiment used a Kaggle-sourced dataset of 5,000 images with a 90/10 training-testing split, so the findings may not fully represent every deepfake generation technique or real-world 5G traffic pattern. Second, the evaluation was performed in an experimental software environment rather than on physical edge hardware, so the reported latency values should be read as model-inference measurements rather than total 5G end-to-end latency guarantees. Third, an 11.0 percent false-acceptance rate, while an improvement over the existing-system baseline, is not zero, and the current evaluation does not yet include adversarial-evasion testing of the kind demonstrated against other detectors by Meng et al. (2023).

6.0 CONCLUSION

The study successfully developed and implemented a deepfake detection and mitigation system for securing 5G cloud networks using a Convolutional Neural Network with a spatial attention mechanism. The framework was designed using Object-Oriented Design principles and implemented in Python using TensorFlow 2.16, Scikit-learn, and Scikit-image, combining image preprocessing, low-level feature extraction, high-level artifact detection, spatial attention, and final classification, with Grad-CAM incorporated to generate visual explanations of the model's predictions.

The experimental evaluation demonstrated that the proposed CNN achieved a test accuracy of 91.0 percent, precision of 92.7 percent, recall of 89.0 percent, an F1-score of 90.8 percent, and an AUC of 0.89, confirming that the model accurately distinguished genuine facial images from manipulated deepfake images. Robustness testing under simulated 5G network conditions showed the model maintaining 84.0 percent accuracy under Gaussian blur, 82.0 percent under motion blur, 80.0 percent after a 50 percent resolution reduction, 78.4 percent under network noise, and 77.0 percent under combined degradation, outperforming the existing-system

baseline by 2.0 to 10.0 percentage points across every tested condition. The proposed framework achieved an inference latency of 8.2 milliseconds, confirming its suitability for real-time deployment in latency-sensitive 5G cloud networks, while the Object-Oriented Design enhanced modularity, maintainability, and scalability for future extension.

Based on these findings, three directions are recommended for future development. Adaptive preprocessing should adjust degradation-correction strength according to the dominant real-time degradation condition, since network noise, 50 percent resolution reduction, and motion blur produced substantially different accuracy drops of 12.6, 11.0, and 9.0 percentage points respectively, rather than assuming any one condition is always the most damaging. Edge-deployed super-resolution should be evaluated for its ability to recover part of the 11.0-percentage-point loss observed under 50 percent resolution reduction, while keeping additional inference latency within an acceptable range. Ensemble detection methods should be implemented for periods of poor network conditions, training multiple lightweight detector variants on diverse degradations and combining their outputs through majority voting or confidence averaging, since combined degradation was identified as the most challenging overall condition.

The main limitations of this study are its reliance on a Kaggle-sourced dataset with a 90/10 training-testing split rather than field-captured 5G traffic, its evaluation in an experimental software environment rather than on physical edge hardware, and an 11.0 percent false-acceptance rate that, while improved relative to the existing-system baseline, is not yet zero. Future work should validate the framework against a wider range of deepfake generation techniques, test on imagery captured over live 5G infrastructure, and evaluate resistance to adversarial evasion attacks. Overall, the findings show that a spatial-attention CNN approach can strengthen 5G-cloud security by supporting accurate, low-latency, and explainable deepfake detection close to the point of image submission.

REFERENCES

Khan, A., Ahmed, W., & Malik, A. W. (2023). Machine Learning Approaches for Security Threat Detection in 5G IoT Networks. Journal of Information Security and Applications, 67, 103234. https://doi.org/10.1016/j.jisa.2023.103234

Kim, H., & Park, J. (2023). AI-Based Intrusion Detection System for 5G-Enabled IoT Networks. Future Internet, 15(12), 354. https://doi.org/10.3390/fi15120354

Kim, H., Park, J., & Lee, S. (2022). AI-Driven Security for 5G Networks: Hybrid Intrusion Detection Systems. Future Internet, 14(11), 301. https://doi.org/10.3390/fi14110301

Konstantinos Kalodanis, Charalampos Papapavlou & Georgios Feretzakis (2025). Enhancing Security in 5G and Future 6G Networks: Machine Learning Approaches for Adaptive Intrusion Detection and Prevention. Future Internet, 17(7), 312

Alkhateeb, F., Nasser, N., & Al-Ali, A. (2022). Deepfake detection for edge intelligence in 5G networks. Sensors, 22(24), 9957. https://doi.org/10.3390/s22249957

Ahmad, W., Malik, A. W., & Aslam, M. (2024). Intrusion Detection and Deepfake Mitigation in 5G Networks Using Hybrid Deep Learning Approaches. Future Internet, 16(10), 359. https://doi.org/10.3390/fi16100359

Ahmad, W., Malik, A. W., & Aslam, M. (2024). Intrusion Detection and Deepfake Mitigation in 5G Networks Using Hybrid Deep Learning Approaches. Future Internet, 16(10), 359. https://doi.org/10.3390/fi16100359

Ahmad, W., Malik, A. W., & Aslam, M. (2024). Intrusion Detection and Deepfake Mitigation in 5G Networks Using Hybrid Deep Learning Approaches. Future Internet, 16(10), 359. https://doi.org/10.3390/fi16100359

Ahmad, W., Malik, A. W., & Aslam, M. (2024). Intrusion Detection and Deepfake Mitigation in 5G Networks Using Hybrid Deep Learning Approaches. Future Internet, 16(10), 359. https://doi.org/10.3390/fi16100359Kim, H., & Rahman, M. (2025). Hybrid AI model for real-time deepfake threat mitigation in 5G networks.

Yatin, P., & Mayank, J. (2024). Deepfake image detection using machine learning and deep learning. Educational Administration: Theory and Practice, 30(5), 14987–14993.