IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14445

An Advanced Explainable Deep Learning Approach with Grad-CAM and Post-Hoc Analysis for Secure QR Code Threat Detection

Y Roshni[1] and Dr. Golda Dilip[2]

Student, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India-600026[1]

Guide, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India-600026[2]

Abstract: The rapid adoption of QR codes in digital payments, authentication, and information sharing has increased cybersecurity risks, particularly QR-code-based phishing attacks known as quishing. Traditional machine learning methods rely on handcrafted features and often lack interpretability, limiting their effectiveness against evolving threats. This paper proposes an explainable deep learning–based framework for malicious QR code detection using a convolutional neural network (CNN) to classify QR code images as benign or malicious. To improve transparency, Grad-CAM is applied to highlight important regions influencing model decisions, while a post-hoc URL analysis module examines protocol usage, domain age, and suspicious lexical patterns to validate predictions. Experimental results demonstrate high detection accuracy along with meaningful visual and analytical explanations, making the proposed approach suitable for real-world cybersecurity applications.

Keywords: QR Code Security, Malicious QR Detection, Deep Learning, CNN, Grad-CAM, Cybersecurity, URL Analysis, Image Classification

I. INTRODUCTION

Quick Response (QR) codes are two-dimensional matrix barcodes designed to store information in a compact, machine-readable format. Their ability to encode URLs, text, contact details, payment information, authentication tokens, and location data has made them highly popular across both consumer and enterprise environments. Today, QR codes are widely used in mobile payments, digital menus, ticketing systems, attendance systems, healthcare access, login authentication, public services, and marketing campaigns.

Despite their convenience, QR codes also introduce significant cybersecurity concerns. A QR code can encode a malicious URL that redirects users to phishing websites, malware download pages, fake payment portals, or credential harvesting platforms. Because users cannot visually inspect the actual destination encoded in a QR image before scanning, QR-based attacks are especially deceptive. This threat has become more prominent with the increase in mobile-based transactions and contactless digital interactions.

Traditional URL-based threat detection methods usually operate after a QR code has already been scanned and decoded. However, this still exposes the user to risk at the point of interaction. A more proactive approach is to analyze the QR code image itself and predict whether it belongs to a malicious or benign class before the user trusts or acts on it.

In this work, a deep learning-based QR security analysis system is proposed to classify QR codes as malicious or benign using image-based features extracted through a Convolutional Neural Network (CNN). In addition, a domain-aware validation mechanism is integrated to improve practical reliability by analyzing the decoded URL and associated domain characteristics. This combined approach improves not only classification performance but also real-world interpretability and security usefulness.The main objective of this research is to develop an intelligent, explainable, and security-focused QR code classification system that can support safer digital interaction.

II. PROBLEM STATEMENT

The increasing use of QR codes in digital transactions and information access has also increased their misuse. Malicious QR codes can silently redirect users to phishing websites, scam pages, fake login portals, or malware-hosting URLs. Since the content encoded in a QR code cannot be directly interpreted by the human eye, users often scan QR codes without verifying their safety. Existing detection approaches are mostly URL-based and operate only after scanning.

DOI: 10.17148/IJIREEICE.2026.14445

Therefore, there is a need for an intelligent system that can automatically classify QR codes as benign or malicious using image-based learning and security-aware validation before the user is exposed to harm.

III. OBJECTIVES

The objectives of this study are to develop a reliable dataset of benign and malicious QR code images, preprocess and augment the data for robust learning, design a CNN-based model for QR threat classification, interpret model decisions using Grad-CAM, perform post-hoc domain-aware analysis on extracted URLs, and evaluate the overall system using standard performance metrics in order to build an accurate, explainable, and security-oriented QR code threat detection framework.

IV. LITERATURE REVIEW

QR code research has gained significant attention in recent years due to the growing reliance on QR-based communication in digital systems. Existing studies in QR-related research generally fall into several categories: QR extraction under complex imaging conditions, counterfeit or authenticity verification, malicious URL validation, and deep learning-based pattern recognition. Prior studies have shown that image preprocessing and neural-network-based learning can improve QR analysis under challenging backgrounds, varying illumination, and structural distortions.

In the wider cybersecurity literature, malicious URL detection has often relied on blacklist systems, lexical analysis, domain-based heuristics, and machine learning classifiers. However, many such systems operate only after the QR code has already been decoded. Comparatively fewer approaches focus on QR image-level threat prediction before the user interacts with the encoded destination.

Another important gap is explainability. Many security models provide a final prediction without showing why that decision was reached. In QR security applications, interpretability is especially valuable because users and evaluators benefit from understanding which visual regions or security signals influenced the decision.

The present work addresses these limitations by combining image-based CNN classification, Grad-CAM explainability, and post-hoc domain-aware validation within a single framework. This makes the proposed approach both technically meaningful and practically relevant.

V. PROPOSED METHODOLOGY

The proposed framework consists of multiple stages including QR image acquisition, preprocessing, dataset management, CNN-based classification, Grad-CAM interpretation, QR decoding, post-hoc domain-aware security analysis, decision fusion, and final user reporting.

A. Dataset Collection

A binary dataset was prepared for classification into two categories: benign QR codes and malicious QR codes. Benign QR codes correspond to safe or non-malicious encoded content, while malicious QR codes correspond to suspicious, deceptive, or harmful destinations. The dataset was organized into training and testing subsets to enable supervised model development and performance evaluation.

B. Image Preprocessing

To ensure consistency during training and testing, all QR code images were resized to 224 × 224 pixels. This size offers a suitable balance between preserving QR structural information and maintaining computational efficiency. Input images were normalized to improve optimization stability and model convergence.

C. Data Augmentation

To enhance generalization and reduce overfitting, augmentation techniques were applied to the training data. These included controlled transformations such as rotation, translation, scaling, and brightness variation. Such augmentation improves robustness under practical scanning conditions, including camera tilt, mild blur, and changes in lighting.

D. CNN-Based Classification

A Convolutional Neural Network was used as the primary visual classification engine of the framework. The CNN automatically learns discriminative structural patterns from QR code images and predicts whether the input belongs to

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14445

the benign or malicious category. Convolutional layers perform feature extraction, while pooling and dense layers support classification.

E. Grad-CAM-Based Explainability

To improve model transparency, Grad-CAM was incorporated as an explainability mechanism. Grad-CAM highlights the regions of the QR image that contribute most strongly to the final prediction. This helps verify that the model is responding to meaningful structural cues rather than irrelevant artifacts, thereby improving confidence in the model's decisions.

F. Post-Hoc Domain-Aware Analysis

After classification, the QR code is decoded and the embedded content is extracted. If a URL is detected, the system performs post-hoc domain-aware analysis. This stage evaluates multiple security-related signals including blacklist status, suspicious lexical indicators, suspicious TLDs, entropy score, IP-based URL usage, and domain intelligence features. These signals provide an additional validation layer to support or question the primary CNN prediction.

G. Decision Fusion

The final verdict is generated by combining the CNN prediction with the post-hoc domain-aware analysis. This hybrid design reduces over-reliance on image classification alone and enables a more security-aware final decision.

VI. SYSTEM ARCHITECTURE

The system architecture of the proposed secure QR threat detection framework is shown in Figure 1. The workflow begins with QR image acquisition and preprocessing, followed by dataset preparation and CNN-based classification. The trained model is evaluated using performance metrics and subsequently linked with QR decoding and URL extraction. The extracted URL is then processed through a post-hoc domain validation module, and the final verdict is generated through a decision fusion engine. This architecture supports both predictive intelligence and post-prediction security reasoning, making the framework more reliable for practical deployment.

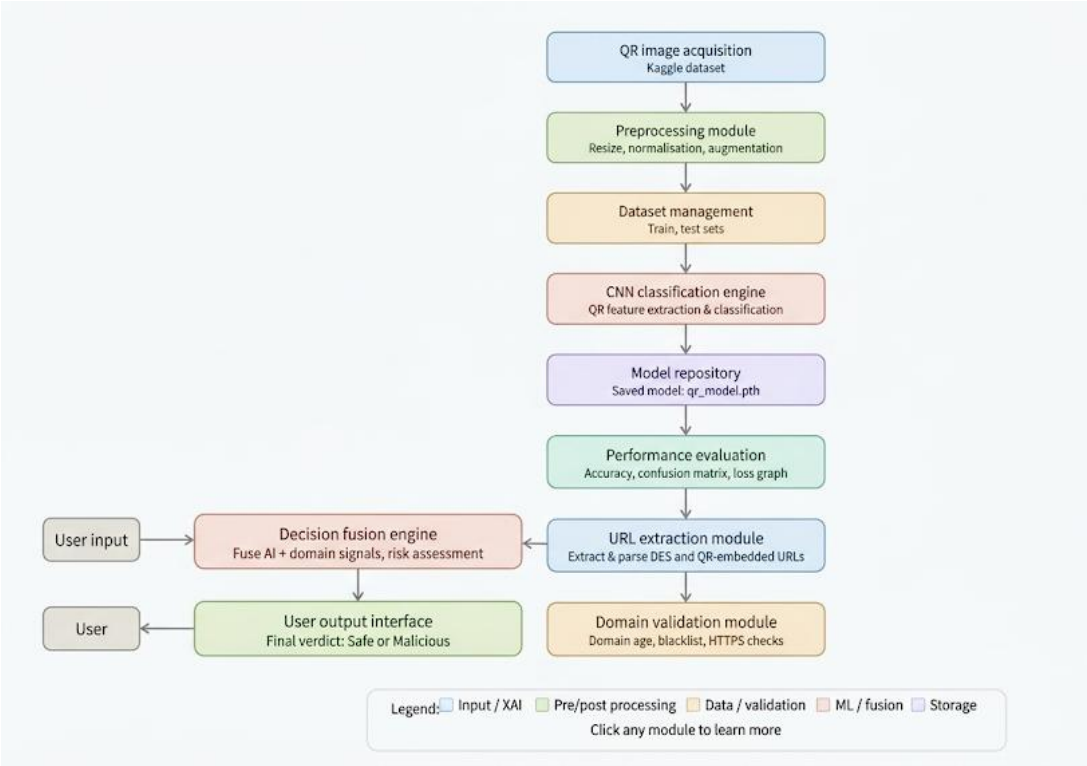

Figure 1. Proposed System Architecture for Secure QR Code Threat Detection

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14445

VII. RESULTS AND DISCUSSION

The proposed framework was evaluated using standard binary classification metrics. In addition to numerical evaluation, the study uses the confusion matrix, training loss curve, Grad-CAM interpretation, and final QR security report to analyse model behaviour and practical relevance.

A. Confusion Matrix Analysis

The confusion matrix shown in Figure 2 summarizes the classification performance of the proposed model across the benign and malicious classes. The matrix values indicate 18,923 true negatives, 1,077 false positives, 743 false negatives, and 19,257 true positives. These results demonstrate that the classifier successfully distinguishes between benign and malicious QR code images with high effectiveness. In cybersecurity applications, false negatives are especially critical because they represent malicious QR codes incorrectly classified as benign. Therefore, the relatively low false negative count supports the security relevance of the model.

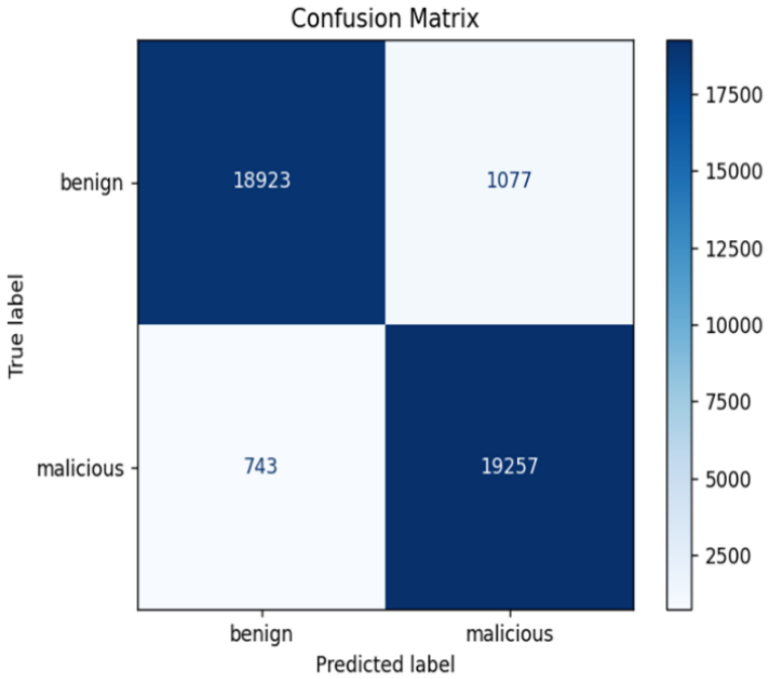

Figure 2. Confusion Matrix of the Proposed Malicious QR Classification Model

B. Performance Metrics

Based on the confusion matrix values, the following performance measures were obtained:

TABLE I PERFORMANCE METRICS OF THE PROPOSED FRAMEWORK

Metric	Value
Accuracy	95.45%
Precision	94.70%
Recall	96.29%
F1-Score	95.49%
Specificity	94.62%

The results indicate balanced and strong classification performance. The recall value is particularly important because it reflects the model's ability to identify malicious QR samples effectively. In a threat-detection setting, strong recall is essential for reducing the risk of unsafe QR codes being misclassified as safe.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14445

C. Training Loss Analysis

The training loss curve shown in Figure 3 illustrates the behavior of the model during learning. The loss drops sharply in the early epochs and gradually stabilizes near zero, indicating effective convergence and successful feature learning.

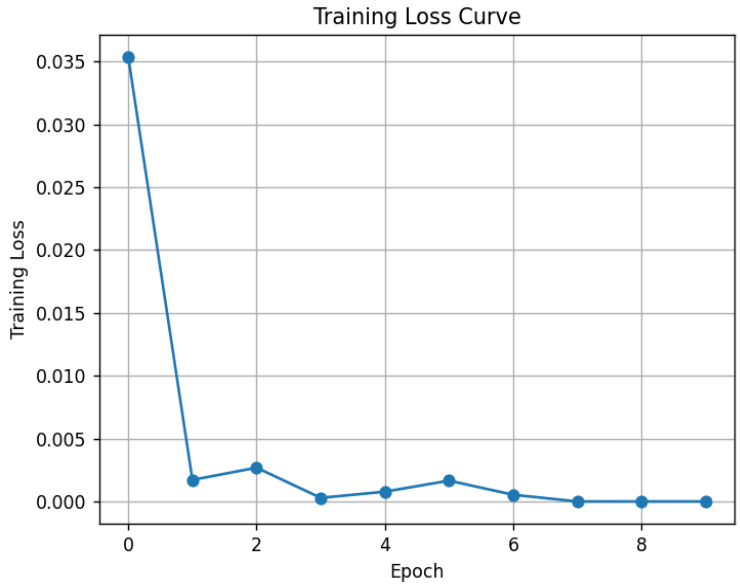

Figure 3. Training Loss Curve of the Proposed CNN Model

D. Grad-CAM Interpretation

The Grad-CAM visualization shown in Figure 4 highlights the QR image regions that most strongly influenced the prediction. This strengthens the explainability of the model and provides evidence that the system is learning meaningful structural patterns. In security-oriented machine learning, such interpretability is valuable because it supports transparency, trust, and analytical understanding of model behavior.

Figure 4. Grad-CAM Visualization of Salient QR Regions Influencing Model Prediction

E. Post-Hoc Security Analysis Report

The final analysis report shown in Figure 5 demonstrates the practical importance of integrating post-hoc domain-aware validation with deep learning classification. In the illustrated example, the CNN predicted the QR code as malicious with 99.87% confidence, but the domain validation module did not observe sufficiently suspicious domain-level evidence and therefore flagged the result as a possible false positive. The report includes extracted URL details and security indicators such as blacklist status, domain age availability, entropy score, phishing word presence, IP-address usage, suspicious TLD detection, and overall risk score.

This result is particularly important because it shows that the proposed framework does not rely blindly on the CNN output. Instead, it applies a post-hoc reasoning layer that can refine the practical interpretation of a high-confidence prediction. This hybrid behaviour significantly improves the framework's real-world relevance.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14445

```
==================================================
          QR SECURITY ANALYSIS REPORT
==================================================

QR Image Path    : sample_images\sample4.jpg
Model Prediction : MALICIOUS
Confidence       : 99.87%

Extracted URL    : https://29a.ch/photo-forensics/

--- Domain Intelligence ---
Domain           : 29a.ch
Blacklisted      : False
Domain Age (days): None
Entropy Score    : 2.585
Phishing Words   : False
Uses IP Address  : False
Suspicious TLD   : False

Risk Score       : 0

Final Decision:
⚠POSSIBLE FALSE POSITIVE: CNN flagged QR but domain appears benign.
```

Figure 5. Sample Post-Hoc QR Security Analysis Report

VII. ADVANTAGES OF THE PROPOSED SYSTEM

The proposed framework offers several important advantages. It supports automated malicious QR code detection, improves decision reliability through post-hoc domain-aware analysis, provides visual interpretability through Grad-CAM, and reduces overconfidence in standalone image predictions. By combining explainable deep learning with security-aware validation, the framework is more practical than systems based solely on visual classification or only on post-scan URL inspection.

VIII. LIMITATIONS

Despite its promising results, the proposed work has certain limitations. The model's effectiveness depends on the representativeness and diversity of the training dataset. Novel adversarial QR patterns may still challenge the classifier. The post-hoc domain analysis relies on heuristic and available intelligence signals, which may not fully capture all malicious behavior. Additionally, large-scale deployment would require broader real-world testing and periodic updating of threat-detection rules.

IX. FUTURE WORK

Future work may include expanding the dataset with more real-world malicious QR samples, evaluating transfer-learning models such as MobileNet or EfficientNet, integrating external threat-intelligence services, improving adversarial robustness, and designing a real-time mobile QR scanner with built-in warning capability. Another valuable direction would be a multi-layer framework combining image-based detection, lexical URL analysis, webpage reputation, and runtime behavior assessment.

X. CONCLUSION

This paper presented an advanced explainable deep learning framework for secure QR code threat detection. The proposed system combines CNN-based image classification, Grad-CAM interpretability, and post-hoc domain-aware analysis to distinguish benign and malicious QR codes more reliably. Experimental results demonstrated strong performance in terms of accuracy, precision, recall, and F1-score. The integration of explainability and post-hoc validation strengthened both the transparency and trustworthiness of the system.

DOI: 10.17148/IJIREEICE.2026.14445

The study shows that malicious QR detection can be improved by moving beyond pure image prediction and incorporating meaningful security-aware reasoning after classification. Overall, the proposed framework contributes a practical, interpretable, and cybersecurity-focused solution for safer QR-enabled digital interaction.

ACKNOWLEDGMENT

The authors would like to thank the Department of Computer Science and Engineering at SRM Institute of Science and Technology for providing the resources and support needed to carry out this research.

REFERENCES

[1]. Denso Wave. QR Code Essentials.
[2]. I. Goodfellow, Y. Bengio and A. Courville, Deep Learning. MIT Press, 2016.
[3]. S. Haykin, Neural Networks and Learning Machines. Pearson, 2009.
[4]. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” in Proc. IEEE International Conference on Computer Vision, 2017.
[5]. J. Ma, L. K. Saul, S. Savage and G. M. Voelker, “Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs,” in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009.
[6]. A. Khonji, Y. Iraqi and A. Jones, “Phishing Detection: A Literature Survey,” IEEE Communications Surveys and Tutorials, 2013.
[7]. N. Alam, A. S. M. S. Sagar, W. Zhang, T. Jin, A. Dosset, L. M. Dang and H. Moon, “A comprehensive study on enhanced QR extraction techniques with deep learning-based verification,” Applied Intelligence, 2025.
[8]. Benign and Malicious QR Codes Dataset. [Online]. Available: https://www.kaggle.com/datasets/samahsadiq/benign-and-malicious-qr-codes
[9]. Springer Article on QR Deep Learning Verification. [Online]. Available: https://link.springer.com/article/10.1007/s10489-025-06509-y