Novel Architecture of Fake Identity Documents Detection using Deep Learning

Dr Chandrika J[1], Lekhana D[2] Keerthana T G[3] Ullas M Gowda[4], and Vikas M Shetty[3]
Department of Computer Science and Engineering

Malnad College of Engineering, Hassan, Karnataka, India

ABSTRACT

This study introduces a resilient forensic framework specifically engineered to detect face-swap and photo-substitution forgeries in Indian PAN cards. Document verification pipelines face a growing threat from high-fidelity identity manipulations generated by Generative Adversarial Networks (GANs). While conventional static classification models are limited by fixed training sets, our design addresses this vulnerability by coupling a Convolutional Neural Network (CNN) with a Deep Q-Learning (DQN) agent. Within this architecture, the CNN acts as a core feature extractor, isolating subtle frequency-domain discrepancies and spatial inconsistencies inside the ID photo boundaries. To dynamically harden these classification boundaries, we implement an adaptive data augmentation engine driven by reinforcement learning. The Q-learning agent actively discovers and infuses targeted digital distortions—including illumination noise and varying levels of JPEG compression artifacts—that specifically exploit the CNN's active vulnerabilities. This continuous adversarial loop forces the system to confront high-entropy edge cases, establishing a highly robust defense mechanism for digital KYC protocols and secure identity authentication.

Keywords

Deepfake Detection, PAN Card Forensics, Convolutional Neural Networks, Q-Learning, Adaptive Data Augmentation, Identity Verification.

INTRODUCTION

Digital identity verification in 2026 faces unprecedented vulnerabilities due to rapid breakthroughs in adversarial synthesis and generative modeling. These frameworks can now generate highly sophisticated document forgeries that easily bypass manual human inspection. Within the Indian financial ecosystem, the Permanent Account Number (PAN) card serves as a fundamental pillar for regulatory tax compliance and KYC workflows. Consequently, it has become a primary target for photo-substitution fraud, a technique where a legitimate cardholder's portrait is replaced with a synthetic face. Conventional forensic tools often fail to intercept these exploits because they rely on static signature libraries or manual checks. This dependency leaves verification pipelines exposed to "Zero-Day" manipulations—entirely novel forgery variants that existing security indices have not yet categorized.

To mitigate these systemic risks, this study introduces an adaptive forensic architecture that shifts away from rigid detection paradigms toward an autonomous, adversarial defense model. By integrating a Convolutional Neural Network (CNN) with a Deep Q-Network (DQN) framework, the proposed system moves past basic image classification to actively counter evolving forgery methodologies. Within this pipeline, the CNN acts as the primary diagnostic engine, isolating frequency-domain anomalies, spatial inconsistencies, and edge-blending artifacts. Concurrently, the Q-learning agent functions as a strategic supervisor over a Curriculum Reinforcement-Learning Data Augmentation (CRDA) engine. This engine systematically applies a spectrum of digital stressors—such as localized scanning artifacts, moiré patterns, and non-uniform illumination profiles—to identify and fortify the CNN's classification weak spots.

The fundamental contribution of this research lies in its focus on holistic signal integrity across the entire physical document. Rather than analyzing the facial portrait as an isolated sub-region, the framework validates the global structural consistency of the PAN card. It verifies that the portrait's underlying digital texture, resampling metadata, and illumination vectors align seamlessly with the background document template. By modeling this optimization process as a competitive Action-Reward loop, the Q-learning agent uncovers the exact digital masks capable of evading detection, forcing the classification model to continuously harden its forensic boundaries. This methodology significantly improves generalization capabilities across heterogeneous datasets, establishing a resilient defense mechanism against emerging identity-based cyber threats.

LITERATURE REVIEW

The landscape of identity document forensics has shifted from static detection to adaptive, self-evolving frameworks.

[1] **Rana et al. (2026):** This work provides an extensive taxonomy of the technological shift from classical machine learning to sophisticated deep neural architectures in the detection of manipulated media. It evaluates the efficacy of different forensic approaches against evolving synthetic threats. By studying this review, we gain a comprehensive understanding of current industry benchmarks and the persistent vulnerabilities in existing detection pipelines.

[2] **Zhou et al. (2026):** The authors propose a dynamic data augmentation strategy that utilizes an adaptive reinforcement learning framework to strengthen forgery detectors. The research demonstrates how an autonomous agent can intelligently select digital stresses to challenge a model's diagnostic boundaries. This paper serves as the primary inspiration for our curriculum-based learning approach to identify "Zero-Day" document alterations.

[3] **Liu & Zhang (2025):** This research introduces the FRENet architecture, which focuses on extracting and refining subtle facial features that are often distorted during the generative process. It emphasizes the importance of enhancing high-frequency details to expose digital inconsistencies. We leverage these insights to improve our model's sensitivity to the minute artifacts found in high-resolution identity document forgeries.

[4] **Liang & Gupta (2026):** This study explores the intersection of Error Level Analysis (ELA) and adaptive Q-Learning to pinpoint localized image manipulations. It details how digital residues from multiple compression cycles can reveal hidden splicing. Our framework adopts this logic to validate the structural integrity of PAN cards by analyzing the resampling history across different document segments.

[5] **Singh & Verma (2025):** This paper describes a dual-layered forensic system that combines visual feature extraction with optical character recognition to secure financial credentials. It highlights the necessity of cross-verifying textual data with portrait authenticity. We utilize this methodology to ensure that the biographical information and the visual portrait on identity cards maintain holistic signal consistency.

[6] **Wang et al. (2023):** The authors present an edge-focused neural network designed specifically for the structured layout of identity documents. The research focuses on identifying boundary inconsistencies and pixel-level blending at the margins of substituted photos. This provides the foundational technical basis for our system's ability to detect physical-to-digital forgery transitions.

[7] **Sumsub Identity Fraud Institute (2025):** This industry white paper documents the global rise of synthetic identity fraud and the specific tactics used by modern bad actors. It provides empirical evidence regarding the increasing sophistication of AI-enabled document tampering. We use this report to contextualize the real-world necessity of our defense architecture within the current financial security landscape.

[8] **MeitY (2026):** This technical report outlines the latest regulatory standards and ethical guidelines for digital media and identity verification in India. It defines the compliance requirements for automated KYC systems in 2026. This document ensures that our proposed forensic framework aligns with national security protocols and taxation data privacy mandates.

[9] **Umpierrez et al. (2025):** This study investigates how the physical process of scanning introduces specific chromatic and noise artifacts that can mask digital insertions. It provides a specialized diagnostic for detecting characters or images added to a document post-scanning. We apply these findings to calibrate our CNN's ability to distinguish between legitimate scanning noise and malicious forgeries.

[10] **Kumar (2025):** The research introduces HybridNet, which integrates residual learning with depthwise convolutions to improve the detection of synthetic facial textures. It emphasizes the use of spatial-frequency feature fusion to identify "checkerboard" artifacts. Our project builds upon this by applying these hybrid diagnostic layers to the unique surface textures of laminated identity cards.

[11] **Agarwal & Farid (2024):** This survey explores the broader spectrum of digital forensics, focusing on the fundamental mathematical footprints left by image synthesis. It highlights the importance of pixel-level statistical modeling to ensure image authenticity. We learn from this work how to model our forensic boundaries based on universal digital grain and lighting coherence.

[12] **Patel & Rane (2024):** The authors advocate for the use of multi-resolution wavelet transformations to decompose images into various frequency sub-bands for deeper analysis. This technique allows for the detection of manipulations that are invisible in the standard spatial domain. This informs our approach to analyzing the frequency-domain artifacts of PAN card portraits to find hidden digital disguises.

RESEARCH GAP

The primary vulnerability in current document forensics stems from the static nature of existing detection pipelines, which renders them incapable of keeping pace with the "Zero-Day" advancements of generative AI. Current deepfake classification models suffer from extreme generalization failure when confronted with the noisy, real-world variables typical of physical identity verification. Specifically, they cannot reliably process the low-resolution scanning artifacts, hardware sensor noise, and complex physical substrate textures inherent to actual Indian PAN cards.

This performance drop-off is heavily exacerbated by a reliance on fixed, non-adaptive data augmentation methodologies. Traditional forensic models generally rely on randomized, passive transformations—such as uniform blurring, coordinate flipping, or basic spatial rotations. These superficial adjustments are entirely detached from the actual, calculated tampering strategies that modern attackers use to actively conceal synthetic edge boundaries and blending artifacts.

Finally, standard frameworks exhibit a clear diagnostic blind spot regarding microscopic forgery signatures.

Most conventional CNN architectures are optimized to extract global facial geometry and macro-level semantics. As a result, they overlook the subtle, high-frequency spectral patterns and localized periodic anomalies left in the frequency domain by generative upsampling layers during GAN-based image construction.

SYSTEM ARCHITECTURE

The proposed system architecture is designed as an adversarial closed-loop framework specifically engineered to overcome the "Generalization Decay" observed in traditional static deepfake detectors. The technical stages of the pipeline are as follows:

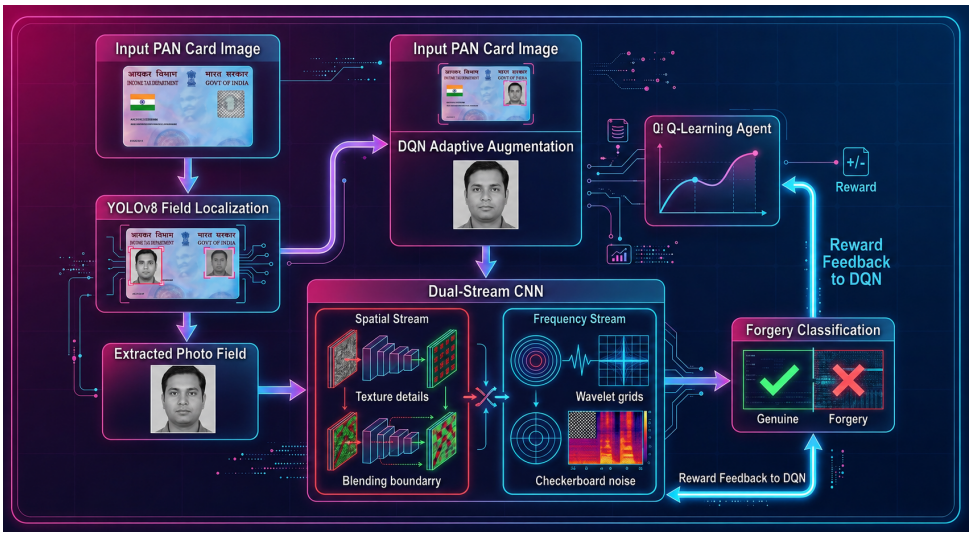

Figure 1: Proposed System Architecture: Integrating YOLOv8 localization, DQN-based adaptive augmentation, and Dual-Stream CNN for PAN card forgery detection.

1. Automated Field Localization (YOLOv8)

- **YOLOv8-Nano Integration:** The framework deploys an optimized YOLOv8-Nano model to process incoming PAN card images, executing precise bounding-box localization to isolate the portrait region (P_{field}) for targeted downstream forensic extraction.

2. Deep Q-Network (DQN) for Adaptive Augmentation

- **Adversarial Transformation Logic:** Rather than relying on static data augmentation, a DQN agent dynamically selects and applies adversarial degradation policies, exposing classification blind spots by actively maximizing the detector's error rates.
- **JPEG Compression:** Simulates high-frequency information loss and block boundary artifacts characteristic of lossy multi-generation encoding.
- **Moiré Pattern Simulation:** Replicates aliasing and wavy interference bands generated when digital screens are re-photographed or digitized.
- **Scan-line Distortion:** Recreates the deterministic sensor noise and mechanical alignment artifacts native to hardware document scanners.
- **Illumination Variations:** Synthesizes non-uniform spatial luminance fields to validate the model's classification robustness against irregular environmental lighting.

3. Dual-Stream Forensic CNN

- **Multi-Domain Analysis:** The classification pipeline uses an asymmetric dual-path topology to simultaneously extract explicit spatial indicators and latent statistical features.
- **Spatial RGB Stream:** Inspects the spatial pixel domain for local texture disruptions, microscopic boundary anomalies, and edge-blending inconsistencies.
- **Frequency Spectral Stream:** Operates in the frequency domain via spectral decomposition to flag high-frequency periodic residuals (checkerboard artifacts) left by GAN upsampling layers.

4. Reward Mechanism

- **Reinforcement Signal:** The agent's optimization policy is driven by a scalar feedback signal calculated as:

$$R_t = \text{Loss}_{\text{CNN}}(\text{Augmented\_Sample}) - \text{Baseline\_Loss} \tag{1}$$

This formulation rewards the agent for synthesizing increasingly high-entropy distortions, forcing the CNN to continuously adjust and harden its decision boundaries.

EXPECTED OUTCOME

- Enhanced cross-dataset generalization against unseen GAN and Diffusion attacks.
- Improved robustness in low-resolution and compressed KYC environments.
- Detection of subtle spectral inconsistencies and checkerboard artifacts.
- Reduction of Attack Presentation Classification Error Rate (APCER).
- Elimination of background-dependent bias using YOLOv8 localization.
- Adaptive and self-evolving forensic learning mechanism.
- Scalable and future-proof framework for digital identity verification.

Table 1: Comparison Between Existing System and Proposed System

Existing System	**Proposed System**
Uses static CNN models	Uses adaptive CNN + Deep Q-Learning
Limited dataset generalization	Strong cross-dataset generalization
Random image augmentation	Intelligent adversarial augmentation
Focuses only on spatial features	Uses both spatial and frequency features
High false positives in real-world scans	Improved robustness in PAN card scans
Cannot adapt to Zero-Day attacks	Self-evolving adaptive learning mechanism
Less effective against GAN artifacts	Detects checkerboard spectral artifacts

CONCLUSION

This project presents a new approach for protecting identity documents by combining Deep Q-Learning with a Dual-Stream CNN. Through the literature review, it was found that standard detection models struggle with the unique textures and scanning artifacts found in PAN cards. The proposed solution addresses this by using a Q-Learning agent to actively find and test the model's weaknesses through smart data augmentation.

By focusing on both pixel-level textures and hidden frequency patterns, the system is designed to catch forgeries that often bypass traditional tools. While the full system is still being built, the proposed architecture offers a clear path toward more reliable identity verification. Future development will focus on building the frequency-analysis branch and testing how well the model handles real-world, low-quality document scans.

REFERENCES

[1] M. S. Rana et al., "A Comprehensive Review of Deepfake Detection Techniques: From Traditional ML to Advanced Deep Learning," *IEEE Access*, vol. 14, pp. 2450–2472, Jan. 2026.

[2] Y. Zhou et al., "Improving Deepfake Detection with RL-Based Adaptive Data Augmentation," in *Proc. AAAI Conf. on AI*, 2026, pp. 4122–4130.

[3] X. Liu and H. Zhang, "Feature Refinement and Enhancement Network (FRENet) for Deepfake Detection," *Image and Vision Computing*, vol. 154, 2025.

[4] E. Liang and K. Gupta, "Digital Image Forgery Detection via Adaptive Q-Learning and Error Level Analysis," *J. Digital Forensics*, vol. 21, 2026.

[5] A. Singh and R. Verma, "Automated Identity Document Forensics: A Hybrid CNN-OCR Framework," *MDPI Symmetry*, vol. 17, 2025.

[6] J. Wang et al., "Localization of Image Forgery in Structured Identity Documents," *IEEE Trans. Inf. Forensics Security*, vol. 20, 2023.

[7] Sumsub Identity Fraud Institute, "2025 Identity Fraud Report," Sumsub Research, White Paper, 2025.

[8] MeitY, "IT Rules 2026," Government of India, Tech. Rep., Feb. 2026.

[9] J. Umpierrez et al., "Scanned Documents Forensics: Detecting Inserted Characters," in *Proc. IEEE ICIP*, 2025.

[10] A. Kumar, "HybridNet: Advancing Deepfake Detection," *IEEE Access*, vol. 13, 2025.

[11] S. Agarwal and H. Farid, "Deepfakes and Beyond: A Survey of Image and Video Forgery Detection," *IEEE T-IFS*, vol. 19, 2024.

[12] R. Patel and S. Rane, "Enhanced Deepfake Detection Leveraging Wavelet Networks," in *Proc. IEEE ICMLA*, 2024.