

# AI-Based Document Forgery Detection Using OCR and Convolutional Neural Networks

## Authors:

Manasvee Bhatia<sup>1</sup>, Navya Mehta<sup>1</sup>

<sup>1</sup>Department of Computer Science and Engineering, Thapar Institute of Engineering & Technology,  
Patiala, India

Date: May 2022

## Abstract

Document forgery remains a growing challenge in digital transactions, education, and identity verification. Traditional Optical Character Recognition (OCR) systems identify text accurately but fail to detect tampered or falsified content. This paper presents a lightweight hybrid approach that combines OCR-based preprocessing with a Convolutional Neural Network (CNN) to detect document forgery in scanned images. The pipeline first extracts textual and structural regions using Tesseract OCR, followed by CNN-based classification to differentiate between genuine and forged document patches. A synthetic dataset of 1,200 documents containing both real and manipulated versions was created using OpenCV for forgery simulation. The proposed model achieved 93.8 percent accuracy and 89 percent recall, outperforming classical threshold-based image comparison techniques. The study demonstrates how combining text recognition and deep visual analysis enhances the reliability of automated document verification systems.

## 1 Introduction

With the growing digitization of administrative and financial operations, document authenticity has become increasingly critical. Forgeries involving altered text, modified seals, or forged signatures can lead to serious legal and financial repercussions. While OCR systems are effective at extracting text, they lack the capability to detect manipulations in the underlying image structure. This motivates the use of computer vision and machine learning to address document forgery detection through visual intelligence. In this study, we propose a hybrid pipeline that integrates OCR preprocessing with a CNN-based classifier to distinguish genuine from tampered documents. The approach focuses on detecting texture inconsistencies, font mismatches, and signature manipulation patterns that OCR alone cannot capture.

## 2 Literature Review

Previous research on document forgery detection has explored a range of techniques including statistical texture analysis, watermark verification, and histogram-based comparison. OCR-based systems rely heavily on textual consistency but struggle with spatial manipulations. Recent advances in deep learning, particularly CNNs, have demonstrated strong potential in visual anomaly detection and forgery localization. However, most existing works demand large datasets or complex architectures unsuitable for lightweight deployment. Our method aims to balance accuracy with computational efficiency, enabling real-time detection in typical office or institutional settings.

## 3 System Design

The proposed system consists of three main components: OCR preprocessing, CNN-based image classification, and result interpretation. The OCR module uses Tesseract to identify and extract textual regions, while OpenCV is employed for image normalization and noise reduction. The CNN model analyzes document patches to identify irregularities associated with forgery, such as uneven text regions or tampered signatures. The architecture was implemented in Python using TensorFlow and Keras frameworks.

## **4 Methodology**

**Data Preparation:** The dataset comprised scanned and synthetically altered documents, with forgeries created through localized edits such as copied signatures, pixel-level erasure, and text overlays using OpenCV. **Preprocessing:** All images were converted to grayscale and normalized to 128×128 pixels. **Model Training:** The CNN model included three convolutional layers followed by max-pooling and dropout regularization. It was trained for 30 epochs using Adam optimizer and binary cross-entropy loss. **Evaluation Metrics:** Accuracy, precision, recall, and F1-score were used to measure performance.

## **5 Results**

The CNN achieved 93.8 percent accuracy, 91 percent precision, and 89 percent recall on the test dataset. Compared to conventional pixel-difference techniques, the proposed method demonstrated superior resilience to illumination changes and noise. False positives primarily resulted from low-quality scans and uneven background textures. The findings validate that deep-learning-based approaches can effectively enhance document integrity verification systems.

## **6 Discussion**

Integrating OCR and CNN processing yielded a robust framework for detecting forged regions without manual intervention. The hybrid model leveraged OCR to focus on text-rich zones and CNN layers to analyze visual artifacts, improving classification reliability. Despite promising results, limitations include restricted dataset diversity and dependence on scanning quality. Expanding the dataset and incorporating attention-based deep networks could further enhance performance in real-world document validation.

## **7 Conclusion**

This research presents an accessible and efficient AI-based solution for document forgery detection. By integrating OCR text analysis with CNN-based image classification, the system demonstrates strong potential for real-world use in academic, financial, and governmental sectors. Future work will investigate multimodal verification using transformer-based architectures and on-device learning to enable scalable, privacy-preserving document authentication.

## **Acknowledgments**

The authors thank the Department of Computer Science and Engineering at Thapar Institute of Engineering & Technology for their guidance and support during this study.

1. Patel, R., & Singh, V. (2019). OCR-Based Document Authentication Techniques. *IJCA Journal of Computer Applications*, 12(4).
2. Nguyen, T., et al. (2021). CNN Approaches for Image Forgery Detection: A Survey. *Sensors Journal*, 21(15), 4983.
3. Sharma, A., & Kumar, R. (2020). Lightweight CNN Models for Image Tamper Detection. *IEEE Access*, 8, 19482–19493.