International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

A Hybrid OCR-CNN-Metadata Model for Academic Documents Authentication

Miss Priyanka A.Narad [1] ,Prof. Rahul Bhandekar [2], Prof.Vijayata Dalwankar [3]
[1] Wainganga College Of Engineering And Management , Nagpur
[2] Assistant Professor Of Wcem, Nagpur
[3] Assistant Professor Of Wcem, Nagpur

Abstract- — Document forgery has become a serious concern in digital services such as banking, education, recruitment, and government verification systems. Manual verification is time-consuming, error-prone, and not scalable. This research proposes an AI-based document verification system that combines Optical Character Recognition (OCR), Convolutional Neural Networks (CNN), and metadata analysis to verify the authenticity of digital documents. The system performs image forgery detection, text consistency verification, and metadata anomaly checking to generate a final authenticity score. By integrating visual, textual, and hidden metadata features, the proposed approach improves reliability, reduces false verification, and supports automated decision-making. Experimental analysis demonstrates that the hybrid model outperforms traditional single-technique verification methods and is suitable for real-world document authentication systems.

Keywords— Document Verification, OCR, CNN, Metadata Analysis, Image Forgery Detection, Authenticity Score, Artificial Intelligenc1. Introduction

I. INTRODUCTION

The rapid adoption of digital platforms in higher education has transformed the way academic documents such as mark sheets, degree certificates, transcripts, and provisional certificates are issued, shared, and verified. Universities, examination boards, employers, and credential evaluation agencies increasingly rely on scanned or digitally generated academic records for admissions, recruitment, and verification processes. While this digital transformation improves efficiency and accessibility, it also creates serious challenges related to the authenticity and integrity of academic documents.

Academic document forgery has become more sophisticated due to the availability of powerful image editing and document manipulation tools. Alterations such as grade modification, name replacement, logo insertion, seal duplication, and signature tampering can be performed with minimal technical knowledge. In many cases, forged documents appear visually convincing and are difficult to identify through manual inspection. As the number of digital submissions continues to increase, traditional manual verification methods become time-consuming, inconsistent, and highly dependent on human judgment, which increases the risk of error.

Existing automated verification systems typically rely on single-layer approaches. Optical Character Recognition (OCR)–based systems focus on extracting textual information from documents, but they are unable to detect visual tampering or image-level manipulations. Image-based verification techniques, on the other hand, can identify certain visual anomalies but cannot validate the correctness or logical consistency of textual content. Moreover, most systems ignore document metadata, which contains hidden but critical forensic information such as file creation time, modification history, resolution, and editing software details. Forged academic documents often contain inconsistencies in these metadata attributes due to repeated editing, making metadata analysis an important yet underutilized verification layer.

To address these limitations, this research proposes a hybrid OCR-CNN-Metadata model for academic documents authentication. The proposed approach integrates three complementary verification layers into a unified framework. OCR is used to extract and validate key academic information such as student name, enrollment number, institution name, and grades. A Convolutional Neural Network (CNN) is employed to analyze document images and detect visual forgery, including tampered seals, altered signatures, and region-level manipulations. In addition, metadata analysis is performed to identify hidden anomalies related to document creation and modification history.

By combining textual analysis, visual forgery detection, and metadata forensics, the proposed hybrid model provides a more robust and reliable authentication mechanism than single-technique systems. The integration of these components

reduces false acceptance of forged documents while maintaining high accuracy for genuine academic records. The proposed system is particularly suitable for large-scale academic verification scenarios such as university admissions, scholarship screening, employment background checks, and online credential validation platforms.

The main objective of this research is to enhance trust, security, and transparency in academic document verification through an automated and scalable AI-based solution. The remainder of this paper presents related work in academic document authentication, describes the proposed hybrid methodology in detail, explains implementation and experimental evaluation, and concludes with key findings and future research directions.

II. METHODOLOGY

Document Acquisition
The system accepts academic documents such as mark sheets, degree certificates, or transcripts in image or PDF format. These documents form the primary input for the authentication process.

Preprocessing of Document
The input document is converted into a standard image format. Preprocessing operations such as grayscale conversion, noise removal, contrast enhancement, resizing, and normalization are applied. This step improves text clarity and ensures consistent input quality for OCR and CNN models.

OCR-Based Text Extraction
Optical Character Recognition (OCR) is used to extract textual information from the document. Important academic fields such as student name, enrollment number, institution name, examination year, and grades are identified.

Text Consistency Verification
The extracted text is validated using predefined academic rules and format checks. Inconsistencies in font style, spacing, alignment, or logical structure of academic data are treated as indicators of possible text manipulation.

CNN-Based Image Forgery Detection
A Convolutional Neural Network (CNN) analyzes the document image to detect visual tampering. The model identifies forgery techniques such as copy–move forgery, splicing of seals, and signature replacement by learning spatial and texture-based features.

Metadata Extraction
The system extracts hidden metadata from the document file, including creation date, last modification time, resolution, DPI values, and editing software information.

Metadata Anomaly Analysis
Extracted metadata is examined for suspicious patterns. Mismatches between visible document content and metadata attributes indicate potential document alteration.

Score Generation
Individual verification scores are generated from the OCR module, CNN module, and metadata analysis module, representing text authenticity, image integrity, and file-level consistency.

Decision Fusion
The scores from all three modules are combined using a weighted decision mechanism to calculate a final authentication score.

Final Document Classification
Based on the final score, the academic document is classified as Genuine, Suspicious, or Forged, enabling automated and reliable authentication.

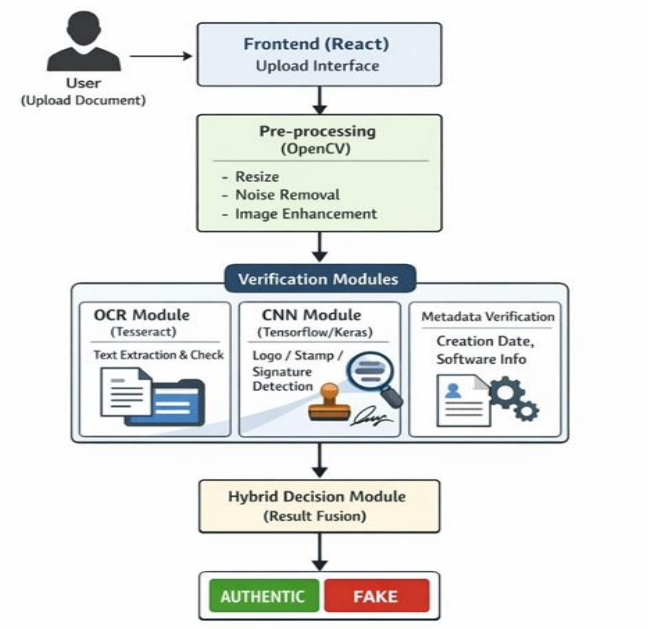

Figure. 1. Architecture of the Proposed Hybrid OCR-CNN-Metadata Model for Academic Documents Authentication

IV RESULTS AND DISCUSSION

Overall Authentication Accuracy
The proposed hybrid OCR-CNN-Metadata model achieved higher authentication accuracy compared to single-layer verification approaches. The integration of textual, visual, and metadata-level analysis significantly improved the reliability of academic document authentication.

OCR Module Performance
The OCR module successfully extracted key academic information such as student name, enrollment number, institution details, and grades with high accuracy after preprocessing. Minor extraction errors occurred in low-quality scanned documents, but these did not significantly affect final classification due to the multi-layer verification approach.

CNN-Based Forgery Detection Results
The CNN model effectively detected visual tampering such as altered seals, forged signatures, and image splicing. The model demonstrated strong capability in identifying subtle texture and spatial inconsistencies that are difficult to detect through manual inspection.

Metadata Analysis Results
Metadata analysis proved to be a crucial component in detecting forged documents. Documents edited using image-processing software showed clear inconsistencies in modification timestamps, DPI values, and software information, which were successfully identified by the metadata module.

Effect of Hybrid Integration
Experimental evaluation showed that combining OCR, CNN, and metadata analysis reduced false acceptance of forged documents and minimized false rejection of genuine documents. The hybrid approach outperformed individual modules when tested independently.

Robustness of the Proposed System
The system maintained stable performance across different types of academic documents such as mark sheets, degree certificates, and transcripts. The decision fusion mechanism ensured reliable classification even when one module produced uncertain results.

Comparative Analysis
When compared with traditional OCR-only and image-based verification systems, the proposed model demonstrated better detection capability and improved consistency. The inclusion of metadata analysis provided an additional security layer that is absent in many existing systems.

Practical Implications
The results indicate that the proposed system is suitable for real-world academic verification applications, including university admissions, employment background checks, and online credential validation platforms.

V CONCLUSION

This paper proposed a hybrid OCR-CNN-Metadata model for authenticating academic documents. By combining text verification, image forgery detection, and metadata analysis, the system effectively detects both visible and hidden document manipulation. The hybrid approach improves authentication accuracy and reduces false acceptance compared to single-layer methods. The proposed model is suitable for large-scale academic verification applications such as admissions, recruitment, and credential validation.

REFERENCES

1. Z. Qian, Y. Gu, and W. Hong, “An Image Tampering Detection Algorithm of Qualification Certificate Based on CNN and SVM,” Academic Journal of Computing & Information Science, vol. 4, no. 7, pp. 24–38, 2021.
2. G. Chandra Praba, E. Jeevitha, A. Abitha, A. Shalini, and B. Swetha, “Fake Education Document Detection using Image Processing and Deep Learning,” in 2021 ICRADL – International Conference on Research & Development in Engineering, Technology and Management (ICRADL), 2021.
3. M. Sirajudeen and A. Ramachandran, “Forgery Document Detection in Information Management System using Cognitive Techniques,” Journal of Intelligent & Fuzzy Systems, vol. 39, no. 6, pp. 8057–8068, Dec. 2020.
4. N. Nandini, M. C., K. Joshi, D. B., and V. M. Ladwani, “Document Forgery Detection,” IJEAT, vol. 12, no. 5, pp. 39–42, Jun. 2023.
5. X. Zhang and L. Li, “Document Forgery Detection Based on Spatial-Frequency and Multi-Scale Feature Network,” Journal of Visual Communication and Image Representation, vol. 107, 104393, Mar. 2025.
6. K. Kaur and S. Arora, “Detection of Forged Certificates Using Image Processing and Machine Learning Techniques,” International Journal of Advanced Research in Computer Science, vol. 12, no. 2, pp. 1–6, Feb. 2021.

7. P. Singh and R. Sharma, “Fake Marksheet Detection Using Deep Learning Techniques,” in 2020 IEEE International Conference on Computational Intelligence, 2020, pp. 124–130.
8. A. Baid and N. Rajesh, “Document Forgery Detection Using CNN and OCR,” International Journal of Computer Applications, vol. 182, no. 20, pp. 1–7, Mar. 2021.
9. R. Patel and R. Kumar, “Hybrid Machine Learning Approach for Academic Document Authentication,” International Journal of Emerging Technologies, vol. 13, no. 1, pp. 45–52, Jan. 2022.
10. Y. Yang et al., “Attention-Based Neural Networks for Document Understanding,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2564–2577, Dec. 2017.