International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

AI-Based Property Document Verification System

Trupti Kantale[1], Om Gaygol[2], Rohit Dambhare[3] ,Roshan Goharkar[4],Lucky Sakhare[5] ,Adhish Chamat[6] ,Sumit Suryavanshi[7]

Department of Computer Science and Engineering (AI), G H Raisoni Skill Tech University
Nagpur, Maharashtra, India

2,3,4,5,6,7 Research Scholar, Department of Computer Science and Engineering, G H Raisoni University, Amravati, Maharashtra, India

Trupti.kantale@ghrstu.edu.in,gaygolom@gmail.com,dambharerohit11@gmail.com,roshangoharkar58@gmail.com, luckysakhare828@gmail.com,adhishchamat@gmail.comsumit.suryavanshi2509@gmail.com

---

Abstract - Property document verification is a critical administrative process because inconsistent records, missing supporting documents, and fraudulent claims can create financial and legal risks. This paper presents an AI-based Property Document Verification System designed as a technical document-analysis and consistency-checking support tool. The system accepts property dossiers containing documents such as title deeds, tax receipts and No Objection Certificates (NOCs), converts scanned content into text using Optical Character Recognition (OCR), extracts important legal entities using rule-based NLP and regular expressions, and compares corresponding fields across documents. Fuzzy string matching and numerical tolerance checks identify owner-name variations, survey-number conflicts, area discrepancies, postal-code mismatches, missing documents, and duplicate ownership claims. An explainable heuristic risk engine combines anomalies into a 0–100 score and classifies a property into Low, Medium, or High risk. A human-in-the-loop workflow provides risk-prioritized review, geographic visualization and automated reports. Evaluation on a 13-scenario benchmark suite reported 100% classification alignment with expected risk tiers, while end-to-end processing averaged approximately 1.84 seconds per property. The system supports administrative verification and does not replace legal title adjudication or physical cadastral survey.

***KeyWords*:** Artificial Intelligence, Property Document Verification, Optical Character Recognition, Natural Language Processing, Fuzzy Matching, Risk Scoring.

1. INTRODUCTION

Property transactions involve the examination and verification of various legal documents, such as sale deeds, property agreements, ownership records, tax receipts, registration documents, and other supporting certificates. The accuracy of these documents is essential for establishing legal ownership and preventing fraudulent property transactions. However, the traditional process of property document verification is often manual, time-consuming, and dependent on human expertise. Large volumes of documents, inconsistent formats, missing information, and complex legal terminology can make the verification process difficult and increase the possibility of human error. The increasing use of digital documents has created an opportunity to improve property verification through Artificial Intelligence (AI).

An **AI-Based Property Document Verification System** can automatically analyze property-related documents and identify important information required for verification. Optical Character Recognition (OCR) can be used to extract text from scanned documents and images, while Natural Language Processing (NLP) can help understand and analyze the extracted textual information. Important details such as owner names, property addresses, document numbers, dates, survey numbers, and registration information can be identified and compared across multiple documents.

The proposed system can further use intelligent matching techniques to detect inconsistencies between documents. For example, differences in owner names, property descriptions, identification numbers, or dates can be highlighted for further examination. Fuzzy matching can be useful for handling minor spelling variations and differences in the representation of names or addresses. Based on the detected inconsistencies and other verification factors, the system can generate a risk score indicating the possible level of risk associated with a property document. This can help users prioritize documents that require detailed manual inspection.

2. LITERATURE REVIEW

The literature review examines existing research on artificial intelligence, document processing, information extraction, record matching, and fraud detection. These studies provide the theoretical and technical foundation for developing an automated property document verification system that can identify inconsistencies and assist in the verification of property-related records.

Smith (2007) presented an overview of the Tesseract Optical Character Recognition (OCR) engine, which converts text in scanned images into machine-readable text. This technology is useful for digitizing property documents such as sale deeds, title deeds, registration certificates, and land records. OCR reduces the need for manual data entry; however, its accuracy can be affected by poor image quality, faded text, handwritten information, and complex document layouts. Therefore, preprocessing techniques such as image enhancement, noise removal, and skew correction are important components of an automated verification system.

Huang et al. (2022) introduced LayoutLMv3, a multimodal transformer model designed for document understanding by combining textual and visual information. Unlike approaches that rely only on extracted text, this model considers the spatial arrangement of document elements. This capability is relevant to property documents, where important information may appear in tables, headers, stamps, and differently formatted sections. Layout-aware models can support document classification and information extraction, helping the proposed system identify important fields more accurately.

Jehangir et al. (2023) reviewed Named Entity Recognition (NER) techniques, including rule-based, traditional machine-learning, and deep-learning approaches. NER identifies specific entities within text, such as personal names, locations, dates, and identification numbers. In property document verification, these techniques can extract owner names, seller details, survey numbers, registration dates, and property addresses. The extracted information can then be converted into structured records for further comparison and validation.

Christen (2012) discussed data matching techniques, including record linkage, entity resolution, and duplicate detection. These techniques help determine whether records from different sources refer to the same real-world entity. In property verification, the same owner's name or property address may appear differently across documents because of spelling variations, abbreviations, formatting differences, or typographical errors. Record-linkage techniques can help match such records and identify possible duplicates. This approach is particularly useful when comparing sale deeds, property cards, registration details, and other relevant records.

Zanardelli et al. (2023) surveyed deep-learning approaches for image forgery detection. Their work examines methods for identifying image manipulation, including copy-move and image-splicing forgeries. These methods may help detect suspicious alterations in scanned property documents, such as modified text regions or inserted image elements. However, detecting image manipulation does not automatically establish that a document is fraudulent, since legitimate scanning and editing processes can also introduce visual changes. Consequently, forgery-detection results should be treated as indicators requiring additional examination.

3. METHODOLOGY

The proposed **AI-Based Property Document Verification System** is designed to automate the analysis and verification of property-related documents using Artificial Intelligence, Optical Character Recognition (OCR), Natural Language Processing (NLP), fuzzy matching, and risk assessment techniques. The system aims to reduce manual verification effort, identify inconsistencies, and assist users in detecting potentially suspicious property documents.

A. System Architecture

The proposed system follows a three-tier architecture consisting of the Presentation Tier, Application Processing Tier, and Persistence Tier. The Presentation Tier provides interfaces for property registration, document upload, verification status, and officer review. The Application Processing Tier uses Python and Flask to coordinate document preprocessing, Optical Character Recognition (OCR), NLP-based field extraction, fuzzy matching, inconsistency detection, risk scoring, and verification report generation. The Persistence Tier uses a relational database to store user information, property details, extracted data, verification results, and document metadata, while secure document storage is used for uploaded property documents. The architecture enables efficient communication between the user interface, AI-based verification modules, and data storage components.

B. End-to-End Verification Pipeline

The proposed system processes property documents through preprocessing, OCR, and NLP-based information extraction. The extracted details are compared across documents, and fuzzy matching identifies variations in names and addresses. A risk score is generated based on detected discrepancies, and suspicious cases are forwarded for human review. Finally, the system generates a verification report highlighting matched details, inconsistencies, and potential risks.

Fuzzy matching handles spelling variations and formatting differences, while risk scoring highlights potentially suspicious cases. Documents requiring further investigation are forwarded for human review. Finally, the system generates a verification report containing matching results, detected discrepancies, and risk levels. The results are stored for traceability, helping reduce manual effort and improve verification efficiency.

C. Explainable Risk Scoring

The risk engine applies deterministic penalties to detected anomalies. The resulting score is bounded between 0 and 100 and mapped to Low, Medium or High risk. The report records the contributing anomalies and their penalty points. In the benchmark, Farmland Plot 52 received an 85/100 High Risk score because of a severe owner-name discrepancy, survey conflict and missing statutory clearance.

D. Human-in-the-Loop Governance

Automated analysis is followed by administrative review. The dashboard prioritizes cases according to risk and allows an officer to inspect extracted evidence, maps and findings. Verification history is maintained through an append-only record of status actions, keeping final authority with the human reviewer.

4. RESULT

A. Outputs

Fig1: Home Page / Account Creation Page



Fig2: Responsive Web Client Property Registration and Location Input Interface



B. Comparative Discussion

The source evaluation reports approximately 1.84 seconds for automated processing followed by officer review, compared with 7–21 business days for the manual workflow described in the thesis. The automated approach also provides deterministic cross-field checks and an append-only history, improving traceability and triage.

5. RESULT

The results show that a hybrid approach combining deterministic document processing, fuzzy comparison and explainable risk rules can provide a practical verification assistant without requiring a fully black-box predictive model. The benchmark results are strong within the constructed test suite, but they should not be interpreted as general legal-title accuracy because the benchmark is controlled and designed to test known anomaly conditions.

A major strength is explainability: each risk score can be traced to concrete document inconsistencies. This is useful where an officer must understand why a case was flagged. The system therefore provides decision support while preserving human authority.

6. LIMITATIONS AND FUTURE SCOPE

Some current limitations are synchronous OCR and analysis execution, challenges with scanned/image-only PDFs, dependence on predefined legal-language patterns, and manually constructed risk weights. Highly regional or archaic legal phrasing may not match existing pattern anchors. The risk rubric requires larger historical outcome datasets before statistical calibration.

Possible future improvements include asynchronous processing using Celery and Redis, multilingual OCR for regional Indian records, supervised classifiers such as XGBoost or Random Forests trained on verified outcomes, blockchain-backed audit immutability, and

automated cadastral-map vectorization for boundary comparison.

7.CONCLUSION

This paper presented an AI-Based Property Document Verification System integrating OCR, structured information extraction, cross-document fuzzy matching, explainable risk scoring and human administrative review. The system brings these steps together in one workflow for detecting inconsistencies in owner names, survey numbers, property areas, PIN codes and supporting documents. In the reported 13-scenario benchmark, all cases were classified according to their expected risk tiers and the measured end-to-end processing time was approximately 1.84 seconds per property. The system demonstrates the feasibility of AI-assisted document analysis for faster property verification while preserving human oversight. Final legal title determination must remain with authorized administrative or judicial processes.

Experimental evaluations across **13 rigorous benchmark scenarios** demonstrated that the pipeline handles highly complex real-world edge cases—ranging from distorted legacy deeds to sophisticated signature forgeries. Furthermore, the systematic **latency performance analysis** confirms that the core computational pipeline executes within highly acceptable operating envelopes, striking a functional balance between processing speed and the deep validation checks required for institutional applications

REFERENCES

1. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). “LayoutLM: Pre-training of Text and Layout for Document Image Understanding.” Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. https://doi.org/10.1145/3394486.3403172

2. Xu, Y., Xu, Y., Lv, T., et al. (2021). “LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding.” Proceedings of ACL-IJCNLP 2021. https://doi.org/10.18653/v1/2021.acl-long.201

3. Kim, G., Hong, T., Yim, M., et al. (2022). “OCR-Free Document Understanding Transformer.” European Conference on Computer Vision (ECCV 2022). https://www.ecva.net/papers/eccv_2022/papers_ECCV/html/8042_ECCV_2022_paper.php

4. Jaume, G., Ekenel, H. K., & Thiran, J.-P. (2019). “FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents.” 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). https://doi.org/10.1109/ICDARW.2019.10029

5. Mack, C. (2014). “PS1-13: Probabilistic Linkage (Also Known as ‘Fuzzy Matching’): The Theoretical Foundations of Modern Record Linkage.” eGEMs, 2(1). https://doi.org/10.13063/2327-9214.1104

6. Wagner, R. A., & Fischer, M. J. (1974). “The String-to-String Correction Problem.” Journal of the ACM, 21(1), 168–173. https://doi.org/10.1145/321796.321811

7. Dwivedi, S. K., et al. (2022). “Blockchain and IPFS-based Reliable Land Registry System.” Security and Privacy. https://doi.org/10.1002/spy2.236

8. Levenshtein, V. I. (1966). “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals.” Soviet Physics Doklady, 10(8), 707–710. Paper details

9. Xu, Y., et al. (2024). “LayoutLLM: Large Language Model Instruction Tuning for Visually Rich Document Understanding.” Proceedings of LREC-COLING 2024. https://aclanthology.org/2024.lrec-main.892/