IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155251

AI-Based Intelligent Document a Verification System Using OCR and Machine Learning

Vrushali Gavali[1], Rutuja Shelar[2], Pranita Sandim[3]

Student, Department of Computer Engineering, BSCOER Narhe, Pune, India[1-3]

Abstract: In the modern digital era, document verification has become an essential requirement across various sectors such as education, banking, government, and corporate industries. Traditional verification methods rely heavily on manual processes, which are time-consuming, error-prone, and inefficient in handling large volumes of data. This paper presents a Document Verification System that automates the process of validating documents using advanced technologies. The system aims to reduce human intervention, improve verification accuracy, and provide faster results. The proposed system allows users to upload documents through a web-based interface developed using ReactJS, with state management implemented using Redux/Context API and UI styling using TailwindCSS/Bootstrap. The backend is developed using Python, which processes the uploaded documents and verifies them against predefined data stored in MongoDB/MySQL databases. The system incorporates TensorFlow-based machine learning models and Optical Character Recognition (OCR) techniques to extract and analyze document content for authenticity.

Keywords: Document Verification, OCR, Machine Learning, Deep Learning, Fraud Detection, TensorFlow, ReactJS, Artificial Intelligence, Authentication System.

I. INTRODUCTION

In the rapidly evolving digital landscape, the need for secure and efficient document verification has become increasingly important across multiple domains such as education, banking, government services, and corporate sectors. With the growing adoption of digital documents, traditional verification methods that rely on manual inspection are no longer sufficient.

This research proposes a Document Verification System that leverages modern web technologies and machine learning techniques to automate the verification process. The system enables users to upload documents through a user-friendly interface developed using ReactJS, with state management handled by Redux or Context API, and responsive design achieved using TailwindCSS or Bootstrap. The backend is implemented using Python (Django/Flask), which processes the uploaded documents and interacts with the database (MongoDB/MySQL) for data storage and retrieval.

To enhance verification accuracy, the system integrates TensorFlow-based machine learning models along with Optical Character Recognition (OCR) techniques to extract and analyze textual information from documents. The extracted data is then compared with predefined or stored data to determine the authenticity of the document.

II. PROBLEM STATEMENT

Traditional document verification systems are inefficient in handling large-scale digital verification tasks due to their dependency on manual inspection and basic validation methods. These systems are vulnerable to human error, delays, document forgery, and data in consistency. Existing OCR-based systems often suffer from reduced accuracy when processing low-quality or tampered documents. Furthermore, many current verification systems lack intelligent fraud detection mechanisms and real-time automation capabilities. Therefore, there is a need for an intelligent, automated, and scalable document verification framework capable of accurately detecting forged documents while minimizing processing time and human intervention.

III. LITERATURE SURVEY

Document verification has gained significant attention in recent years due to the rapid increase in digital data and the need for secure authentication systems. Various researchers have proposed different techniques combining machine learning, image processing, and blockchain technologies to enhance verification accuracy and reliability.

DOI: 10.17148/IJARCCE.2026.155251

1. Shende and Mullapudi (2024) presented a comprehensive study on modern document verification techniques, including signature verification, stamp detection, and image-based validation using machine learning algorithms. Their research highlights the effectiveness of combining multiple verification techniques to improve accuracy. However, the system requires large datasets and high computational resources for training.
2. Mohana Priya et al. (2025) developed an Automated Document Verification System (CADVS) that utilizes Optical Character Recognition (OCR) and machine learning models to extract and validate textual information from official documents. The system demonstrated improved efficiency and reduced manual workload. However, its performance depends heavily on image quality and OCR accuracy.
3. Kumar and Meena (2020) introduced a secure framework combining machine learning and blockchain for document verification. Their system focuses on enhancing security and preventing unauthorized access. Although the framework improves trust and reliability, it may not be suitable for real-time applications due to latency issues.

Recent advancements in deep learning frameworks such as TensorFlow have enabled more accurate feature extraction and anomaly detection in document verification systems. Integration with web technologies like ReactJS and Python-based backends has further improved usability and scalability.

IV. EXISTING IDEA

The existing document verification systems primarily rely on manual verification processes or semi-automated methods. In these systems, users submit physical or digital documents, which are then verified by authorized personnel through visual inspection or basic software tools.

The verification process often involves checking document details against stored records or databases without the use of advanced machine learning techniques.
Although some systems use basic digital tools for storage and retrieval, they lack intelligent automation, real-time processing, and fraud detection capabilities. As highlighted in the reference presentation, the absence of a centralized and automated system makes the process inefficient and difficult to manage at scale.

Advantages of Existing System

Simple and easy to implement without the need for complex technologies
Low initial setup cost as it does not require advanced infrastructure
Human judgment allows flexibility in handling complex or unclear cases
Widely adopted in traditional systems across organizations.

V. SYSTEM ARCHITECTURE

The proposed Document Verification System follows a modular multi-tier architecture consisting of frontend, backend, machine learning, OCR, and database modules.

Frontend Layer

The frontend is developed using ReactJS and TailwindCSS/Bootstrap to provide a responsive and user-friendly interface.
The frontend supports:
Document upload
User authentication
Dashboard visualization
Verification result display

Backend Layer

The backend is implemented using Python with Flask/Django frameworks. The backend:
Handles API requests
Processes uploaded documents
Integrates OCR and TensorFlow models
Communicates with the database
OCR and Machine Learning Layer

OCR extracts textual information from uploaded documents. TensorFlow-based CNN models analyze extracted features and detect anomalies or tampered content.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155251

Database Layer

MongoDB/MySQL databases store:
User details
Verification records
Document metadata
Authentication logs

Workflow of the System

The overall workflow of the system is as follows:

1. User accesses the dashboard and fills the form.
2. User uploads the document.
3. Frontend sends data to backend via API.
4. Backend processes the request.OCR extracts document data.
5. TensorFlow model verifies authenticity.
6. Data is compared with database records.
7. Verification result is generated.

VI. DATASET DESCRIPTION

The experimental dataset used in this research consists of genuine and tampered documents collected for verification testing.

Document Type	Genuine	Tampered
Aadhaar Cards	200	80
PAN Cards	150	60
Certificates	100	40
Passports	50	20

Total dataset size: 700 documents.

The dataset includes:

PDF documents
Scanned images
JPEG and PNG formats
Low-resolution and high-resolution samples

The dataset was divided into:

Training Set: 80%
Testing Set: 20%

Dashboard Interface: Displays the document verification form.
Form Handling: Allows users to input required details.
Document Upload Module: Supports multiple formats (PDF, JPG, PNG, DOCX).

VII. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed system was evaluated based on verification accuracy, fraud detection capability, and processing speed.

Metric	Existing System	Proposed System
Verification Accuracy	84%	96%
Fraud Detection Rate	72%	93%
Average Processing Time	12 sec	3 sec
Error Rate	11%	3%

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026
DOI: 10.17148/IJARCCE.2026.155251

The integration of OCR and CNN-based anomaly detection significantly improved document verification performance. The system successfully identified forged and tampered documents with high accuracy while reducing verification time.

The ReactJS frontend provided smooth user interaction, while the backend efficiently handled multiple verification requests simultaneously. Experimental results demonstrate that the proposed system is suitable for real-time document verification applications.

VIII. CONCLUSION

This research presents an AI-Based Intelligent Document Verification System that integrates OCR, machine learning, and database validation to automate document authentication. The proposed framework addresses the limitations of traditional manual verification systems by improving accuracy, reducing processing time, and enhancing fraud detection capability.

The integration of TensorFlow-based CNN models enables intelligent anomaly detection and efficient verification of digital documents. Experimental analysis demonstrates that the proposed system achieves higher verification accuracy and improved scalability compared to existing methods.

The proposed system provides a reliable, secure, and scalable solution for modern document verification applications in educational institutions, banking systems, government services, and corporate organizations.

IX. FUTURE SCOPE

The proposed system can be further enhanced using advanced technologies such as:

Blockchain-based tamper-proof storage
Multi-language OCR support
Mobile application integration
Cloud-based deployment
Transformer-based AI models
Facial recognition verification
QR-code and digital signature validation
Real-time government database integration

REFERENCES

[1]. Marella, V., & Vijayan, A. “Document Verification Using Blockchain for Trusted CV Information,” 2020.
[2]. Shende, A., & Mullapudi, M. “Enhancing Document Verification Systems Using Machine Learning Techniques,” 2024.
[3]. Mohana Priya, N., et al. “Automated Document Verification System (CADVS),” IJRASET, 2025.
[4]. Kumar, P., & Meena, P. “A Secure Framework for Document Verification Using Blockchain and Machine Learning,” IJCET, 2020.
[5]. TensorFlow Documentation
[6]. ReactJS Documentation
[7]. MongoDB Documentation
[8]. MySQL Documentation
[9]. OpenCV Documentation
[10]. Tesseract OCR Documentation

