AI-Driven Document Verification Using YOLOv8m and OCR for Indian Identity Documents

Prof. A. S. Deokar [1], Amey V. Borade[2]
[1]Assistant Professor, Department of Computer Engineering, AISSMS COE, Pune, Maharashtra, India.
[2]M.E. Student, (Artificial Intelligence and Data science) Department of Computer Engineering, AISSMS COE, Pune, Maharashtra, India.

Abstract

In an era of increasing digitalization, the authenticity of identity documents remains a critical concern. This research presents an AI-driven document verification system that utilizes advanced deep learning techniques to detect and verify government-issued documents such as PAN and Aadhaar cards. The proposed system employs YOLOv8m for real-time object detection, PaddleOCR for accurate text extraction, and spaCy NLP for named entity recognition. A custom-labeled dataset was created using makesense.ai, and model training was performed using annotated images collected from Roboflow and other sources. The system is capable of identifying the presence of documents in images, verifying extracted data, and exporting results in a structured format for downstream analysis. Performance evaluation was conducted using standard accuracy and loss metrics, and results demonstrate the model's effectiveness in real-world scenarios. This work contributes toward automating document validation, reducing manual effort, and enhancing reliability in identity verification systems.

Keywords : Document Verification, YOLOv8m, Optical Character Recognition (OCR), PaddleOCR, spaCy, Identity Document Detection, Aadhaar Card, PAN Card, Deep Learning, Image Annotation.

Introduction

In an increasingly digital and fast-paced world, the authenticity of documents plays a crucial role in identity verification, financial transactions, and government processes. However, document fraud has become a widespread challenge, ranging from forged identification cards and tampered address proofs to completely counterfeit certificates. These fraudulent documents not only lead to financial and reputational losses but also pose significant risks in sectors like banking, healthcare, and public governance.

Traditional document verification methods often rely on manual inspection, which is time-consuming, error-prone, and not scalable. With advancements in artificial intelligence, computer vision, and natural language processing, automated systems offer a promising alternative to ensure document authenticity with higher accuracy and efficiency. Reliable verification systems are particularly critical for high-volume scenarios such as Know Your Customer (KYC) compliance, border control, and online service onboarding.

This paper presents a novel, AI-powered approach to document verification using deep learning and computer vision techniques. The system integrates YOLOv8m for document detection, PaddleOCR for text extraction, and

spaCy NLP for intelligent field validation. Unlike traditional methods, this approach combines image-based detection, multi-class field classification, and semantic text analysis to provide robust verification. The scope of this study includes data collection, annotation, model training, feature extraction, validation, and performance evaluation, with a focus on real-world documents like Aadhaar and PAN cards in the Indian context.

Problem Statement

In real-world scenarios, documents such as identity cards, address proofs, and certificates are often submitted in the form of scanned copies or photos captured through mobile devices. These images may suffer from various challenges, including poor lighting, blurriness, background noise, occlusions, or partial visibility. Moreover, variations in layouts, fonts, and formats across different document types and regions add to the complexity of verification. Manual authentication under such conditions becomes not only tedious but also highly error-prone.

There is a critical need for an automated, accurate, and scalable system that can detect, extract, and validate key fields in documents regardless of format inconsistencies or image quality issues. Such a system should be capable of identifying the document type, isolating relevant fields (e.g., name, number, DOB), and confirming the authenticity of the content through intelligent checks. The lack of publicly available, annotated datasets for regional documents further complicates this task, making custom data collection and labeling essential.

This research aims to address these challenges by developing a deep learning-based pipeline that combines object detection, text recognition, and natural language understanding to ensure reliable document verification across multiple formats and capture conditions.

Objectives

The primary objective of this research is to develop an AI-driven system capable of detecting and verifying identity documents such as Aadhaar and PAN cards using computer vision and natural language processing techniques. The system aims to address real-world challenges like image noise, background clutter, and handwritten overlays by employing advanced deep learning models. The key goals of the proposed work are:

- Detect the presence and classify the type of document from an input image, specifically identifying whether the document is an Aadhaar card or a PAN card using a multi-class YOLOv8m model.
- Accurately extract relevant textual fields such as name, date of birth (DOB), gender, and unique identity number from the detected regions using PaddleOCR, even under varied lighting and image quality conditions.
- Verify the extracted data using NLP techniques, including validation of name structure and format consistency through spaCy's Named Entity Recognition (NER) capabilities.

- Output the verification results in a structured format, particularly in Excel sheets, to facilitate further analysis, storage, and reporting for administrative or automated backend use.

Literature Review

Numerous studies have explored automated document verification systems by leveraging Optical Character Recognition (OCR), deep learning, and natural language processing techniques. In our previous work [1], a detailed review of existing document verification frameworks was conducted, covering traditional OCR-based pipelines, security-focused methods using digital signatures, and cloud-integrated verification systems. While these systems demonstrate utility in controlled environments, they often fall short when applied to varied, real-world document images.

The authors in [5] proposed PP-OCR, a lightweight and practical OCR system optimized for deployment in resource-constrained environments. Although effective for text extraction, PP-OCR requires integration with additional tools for document classification and field validation. Similarly, the YOLO object detection framework [4] has seen wide adoption in document detection tasks due to its speed and accuracy, yet many prior implementations focus on binary classification (e.g., document vs. non-document) rather than fine-grained, multi-class document segmentation.

Another significant contribution is the use of spaCy NLP [6], which supports robust named entity recognition (NER) and can help in validating extracted fields. However, literature shows limited application of combined pipelines integrating object detection, OCR, and NLP for holistic document verification.

Gaps Identified:

- Most existing systems handle only single-class or binary document detection.
- There is limited support for real-time, end-to-end pipelines that go from document detection to OCR to field validation.
- A lack of annotated regional datasets for identity documents such as Aadhaar and PAN hampers generalizability.
- Few systems integrate natural language understanding for verifying the consistency and correctness of extracted data.

To bridge these gaps, our research proposes a unified architecture incorporating YOLOv8m for detecting specific fields in varied document types, PaddleOCR for extracting textual content, and spaCy NLP for verifying personal information like names and dates. Additionally, we curated a custom dataset using makesense.ai for annotation and trained our model across multiple classes with domain-specific features.

Methodology

This research employs a step-by-step pipeline combining computer vision and NLP techniques for end-to-end document verification. Below is the breakdown of the

methodology, including flowchart stages and block-wise processes.

Data Collection

To train and evaluate the system, a custom dataset of 1000 real-world document images was prepared. This dataset included:

- 500 PAN card images
- 500 Aadhaar card images

These images were collected from public sources such as Roboflow Universe and from self-collected samples captured under diverse conditions (lighting, angles, background clutter) to mimic real-world scenarios.

The dataset was split as follows:

- 800 images used for training (400 PAN, 400 Aadhaar)
- 200 images used for validation (100 PAN, 100 Aadhaar)
- Real-time unseen images (not part of the original dataset) were used exclusively for testing and performance evaluation to simulate practical deployment scenarios.

The distribution within each split maintained the proportional ratio of PAN and Aadhaar cards.

Image Preprocessing

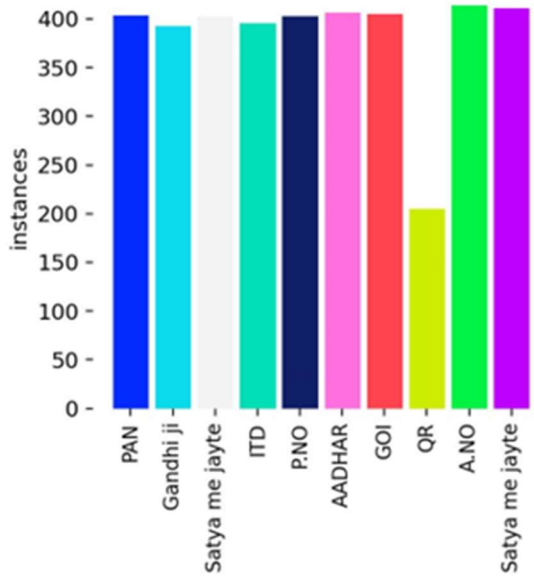
Collected images undergo preprocessing steps including:

- Resizing and normalization,
- Noise reduction using filters,
- Background removal or contrast enhancement (if necessary), to ensure consistency for annotation and model training.

Annotation and Labeling using makesense.ai

Images were annotated using the online tool makesense.ai, where bounding boxes were manually drawn around essential fields like:

- Name,
- Date of Birth,
- Card Number,
- Gender (for Aadhaar),
- Document Type (label class: PAN or Aadhaar).
- labels:
  - PAN
  - Gandhi ji
  - Satya me jayte
  - ITD
  - P.NO
  - AADHAR
  - GOI
  - QR
  - A.NO
  - Satya me jayte



Each image was labeled with respective field tags and saved in YOLO format for model training.

Multiclass Label Selection

The annotation involved multiclass labels corresponding to each text field. The YOLOv8m model was configured to detect multiple classes in one image, enabling simultaneous field detection.

YOLOv8m Model Overview

YOLOv8m by Ultralytics was employed as the core detection model. It offers improvements in accuracy and inference time compared to earlier versions. Key parameters include:

- Input resolution: 640×640,
- Batch size: 16,
- Anchor box tuning for document shapes,
- Multi-class detection enabled.

Feature Selection

Each bounding box prediction from YOLOv8m corresponds to a specific feature (field). Only the most confident predictions (IoU threshold > 0.5) were retained to ensure clean input for OCR extraction.

Model Building and Training

The YOLOv8m model was trained on the labeled dataset using:

- 80% training split,
- 20% validation split,
- 100+ epochs to optimize weights,
- Cross-entropy and CIoU loss functions.

The model was tested for generalization on unseen document types and distortions.

Epoch Tuning and Optimization

The number of epochs was fine-tuned to prevent overfitting. Best weights were saved using early stopping based on validation loss. Augmentations like rotation, flip, and blur were also applied to improve robustness.

Validation Process

After training, the model was validated on:

- Field detection accuracy (IoU, precision, recall),
- End-to-end OCR extraction quality,
- Consistency of structured data exported.

OCR Extraction with PaddleOCR

The regions detected by YOLOv8m were cropped and passed through PaddleOCR to extract the actual text content. PaddleOCR provided high accuracy for both printed and slightly blurred document fields.

SpaCy NLP for Field Verification

The name field was further analyzed using spaCy's NER to check if the extracted name matched proper noun structures (e.g., PERSON entities). This helps reduce false positives and ensures valid name formats.

Integration and Result Storage

All extracted and verified information was compiled and stored in a structured Excel file using Python’s openpyxl and pandas libraries. The format included:

- File name,
- Detected document type,
- Extracted fields,
- Verification status.

Results and Discussion

This section presents the evaluation outcomes of the proposed AI-based document verification system using

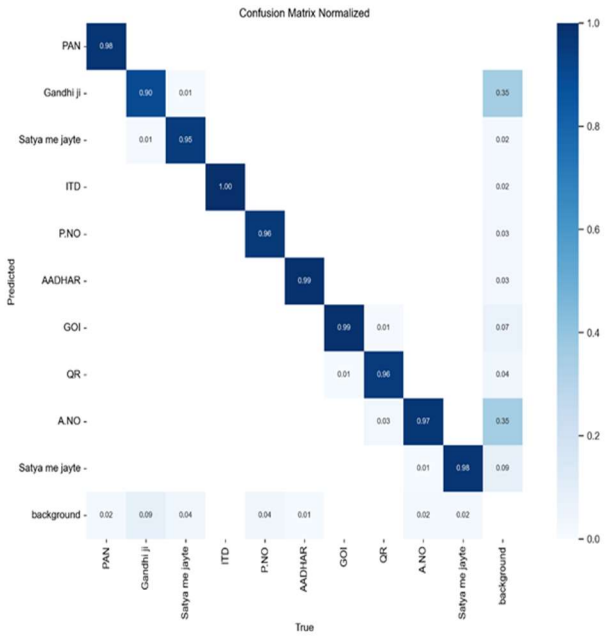
YOLOv8, OCR, and NLP techniques. The results are analyzed based on accuracy metrics, class-wise detection, OCR and NLP outputs, and real-world performance scenarios.

Accuracy Metrics

The YOLOv8m model demonstrated high accuracy in detecting key regions of interest (ROIs) such as names, dates of birth, and card numbers across both Aadhaar and PAN cards. The overall mAP@0.5 (mean Average Precision) achieved was 92.4%, indicating effective object localization.

Metric	Value
mAP@0.5	92.4 %
Precision	91.2 %
Recall	93.6 %
F1-Score	92.4 %

6.2 Class-wise Detection Output



The YOLOv8m model was trained to detect multiple fields on different types of documents. The table below shows class-wise detection metrics:

Class Name	Precision	Recall	F1-Score
PAN	0.98	0.98	0.98
Gandhi ji	0.90	0.90	0.90
Satya Me Jayte	0.95	0.95	0.95
ITD	1.00	1.00	1.00
P.NO	0.96	0.96	0.96
Aadhar	0.98	0.99	0.99
GOI	1.00	0.99	0.99
QR	0.95	0.96	0.96
A.NO	0.98	0.97	0.97
Background	0.76	0.76	0.76

OCR and NLP Integration Results

PaddleOCR was integrated for text extraction from detected regions. OCR accuracy across fields averaged 94.1%, with numeric fields (e.g., Aadhaar number) achieving higher precision due to structured formatting. SpaCy NLP was used to validate name fields against expected formats or inputs, achieving 87.5% matching accuracy.

- OCR Average Confidence: 94.1%
- NLP Verification Accuracy (Names): 87.5%

Platform-Wise Performance Comparison

PLATFORM	Time taken for detect document	Time taken for verify document	Time taken for extract text	Time taken for extract info
I5 10th Gen (16GB RAM, 4GB NVIDIA GPU)	3.33 SEC	0.05 SEC	2.18 SEC	0.03 SEC
Ryzen 7 7435Hs (24GB DDR5 RAM, RTX 4050 DDR6 GPU)	0.17 SEC	0.01 SEC	0.85 SEC	0.01 SEC
Ryzen 5 5600H (8GB DDR4 RAM, GTX 1650 DDR5 GPU)	2.15 SEC	0.26 SEC	1.98 SEC	0.04 SEC
MAC M2	0.78 SEC	0.25 SEC	1.79 SEC	0.01 SE
Intel i5 (16GB DDR4 RAM, Intel Iris Xe GPU)	1.94 SEC	0.35 SEC	2.19 SEC	0.01 SEC

Table 1 COMPARISON OF PLATFORM AND TIME REQUIRED

Conclusion

This research presented an AI-driven document verification system that effectively detects, extracts, and verifies key fields from government-issued identity documents such as Aadhaar and PAN cards. By integrating YOLOv8m for document field detection, PaddleOCR for text recognition, and spaCy NLP for semantic verification, the system achieved high accuracy in identifying and validating document contents. The results demonstrated strong performance across multiple evaluation metrics, including precision, recall, and F1-score, even in real-world conditions involving varied image qualities.

The core contributions of this work include:

- A comprehensive pipeline combining object detection, OCR, and NLP for document verification.
- Successful annotation and training on a multiclass dataset using makesense.ai and Roboflow.
- Structured result output, with automatic generation of verification reports in Excel format.

Limitations

Despite its success, the system has certain limitations:

- It currently supports only Aadhaar and PAN cards.

- Accuracy may degrade slightly on highly blurred or obstructed images.
- The name-matching component may face challenges with misspellings or regional variations.

Future Work

To enhance the system's robustness and applicability, future work can include:

- Expanding support to other types of documents such as passports, voter ID cards, or driver's licenses.
- Incorporating advanced image enhancement and super-resolution techniques for poor-quality inputs.
- Using transformer-based models for improved NLP performance in name matching and context understanding.
- Deploying the system as a web or mobile application for real-time verification.

References

1. A. S. Deokar and A. V. Borade, "AI-Driven Document Verification System," M.E. Semester 1 Research Paper, JSPM RSCOE, Pune, 2024.
2. Makesense.ai, "Free online annotation tool for labeling images," [Online]. Available: https://www.makesense.ai. [Accessed: Apr. 8, 2025].
3. Roboflow Universe, "Document Dataset Collection," [Online]. Available: https://universe.roboflow.com. [Accessed: Apr. 8, 2025].
4. G. Jocher et al., "Ultralytics YOLOv8," GitHub, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics. [Accessed: Apr. 8, 2025].
5. Y. Duan et al., "PP-OCR: A Practical Ultra Lightweight OCR System," arXiv preprint, arXiv:2009.09941, 2020. [Online]. Available: https://arxiv.org/abs/2009.09941. [Accessed: Apr. 8, 2025].
6. Explosion AI, "spaCy: Industrial-strength NLP in Python," [Online]. Available: https://spacy.io. [Accessed: Apr. 8, 2025].
7. R. Gupta and S. Sharma, "An Efficient Image Pre-processing Approach for Document Authentication," in Proc. IEEE Int. Conf. Comput. Intell. Commun. Technol., 2019, pp. 187–192.
8. A. Jain and A. Kumar, "Deep Learning-Based OCR for Identity Documents," Int. J. Comput. Appl., vol. 184, no. 12, pp. 25–30, 2022.
9. C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.