Fake ID / Document Detection Detect Forged ID Using OCR and AI

Ameer Althaf F[1], Ms.Abinaya S[2]
[1]*Junior Researcher Department of Information Technology Sri Krishna Adithya College of Arts andScience*
[2]*Assistant Professor Department of Information Technology Sri Krishna Adithya College of Arts andScience*

***Abstract*—The rapid digitization of identity verification processes has created both opportunities and challenges in ensuring document authenticity. Forged identification documents remain a significant threat to security, financial institutions, and governance systems worldwide. This paper presents an AI-driven framework for detecting counterfeit IDs by integrating Optical Character Recognition (OCR) with advanced machine learning techniques.**

OCR is employed to extract textual and structural features from identity documents, while deep learning models analyze inconsistencies in fonts, layouts, and embedded security elements. The proposed system leverages image prepossessing, feature engineering, and anomaly detection to identify subtle manipulations that are often overlooked by manual inspection. Experimental results demonstrate that the hybrid approach achieves high accuracy in distinguishing genuine documents from forged ones, outperforming traditional rule-based methods.

By automating forgery detection, this research contributes to strengthening digital trust, reducing fraud, and enhancing security in identity verification systems. This version balances technical detail with readability, making it suitable for publication.

***Index Terms*—Finding fake IDs and finding forged documents OCR, or Optical Character Recognition AI, or artificial intelligence, Machine learning, deep learning, image processing, identity verification, and fraud detection are all parts of this field.**

I. INTRODUCTION

In the modern digital world, identification verification plays an important role in many systems. These systems may include banking systems, online services, travel systems, and government systems. In recent times, with the rapid increase in digital transaction processes and online identification verification, the use of identification documents such as Aadhaar cards, passports, and driving licenses has become very common. However, with the increased use of identification documents, there are also increased instances of identity fraud, in which forged identification documents are used for unauthorized access to these systems. The detection of such forged identification documents is a major challenge for many organizations. In most organizations, identification verification through traditional methods is considered to be time-consuming and prone to human error.

To overcome these problems, technologies such as Optical Character Recognition and Artificial Intelligence are being widely employed. The forgery that can take place in the identity document may be in the form of changes to the personal details, the image, the fonts, and the layout of the document. The system that is proposed in the paper utilizes the combination of OCR and machine learning/deep learning techniques to provide an effective solution for the detection of fake documents. The system not only checks the details present in the document but also checks the visual features to identify the patterns and thus determines whether the document is fake or real by comparing the details with the standard formats and rules.

II. PROBLEM MOTIVATION WITH REAL WORLD STATISTICS

Fake IDs have turned into a big issue lately, especially with everything going online more and more. Services like banking or government stuff now check identities from afar, which opens the door for fraud easier than before. People use these fakes for things like stealing identities, setting up illegal accounts, or just getting into places they should not. It feels like we really need better ways to verify documents automatically, something reliable that does not miss much.

In India, the Reserve Bank reports show banking frauds shooting up, with losses in the thousands of crores from messed up documents and identity tricks. Globally, the Federal Trade Commission keeps track of identity theft complaints, millions every year, and a lot involve altered or fake papers. That is pretty alarming, I think.

When it comes to KYC processes, they rely so much on those ID submissions. But doing it manually takes forever and people make mistakes all the time. Jumio had this report saying about one in twenty verifications has fraud signs, like forged docs or tweaked data. It shows how common this stuff is getting.

Advancements in photo editing software make it simple for anyone to whip up realistic fakes. You do not even need to be an expert, just change a photo or text a bit, and it looks good. Traditional checks struggle with that, particularly when there are tons of documents to go through fast. Some people might think it is not that bad, but it seems like it is.
This pushes for smarter automated systems to spot fakes. Using OCR along with AI, you can dig into the text and images closely, finding odd patterns or inconsistencies humans might overlook. The idea here is to build something that makes verification quicker and more accurate.

Reducing manual work could help a lot in areas like banking or online services. The goal is to cut down fraud and make security stronger overall, though it is not totally clear how perfect it can get yet.

III. LITERATURE REVIEW & REVIEW OF RECENT RELATED STUDIES

Detecting forged identity documents is a big deal these days, especially with all the digital systems popping up everywhere. I think it's gotten more attention because people are relying on quick verifications online or something. Researchers have tried all sorts of ways to tackle this, from old school image stuff to fancy AI and deep learning.

But on its own, OCR misses the visual tricks, like if someone tampered with the image. So, they came up with hybrids, mixing OCR and AI. One study had this setup with CNNs to classify real versus fake, and it hit about 93.8 percent accuracy. That sounds pretty good, but I wonder how it holds up in real life.
Deep learning has taken over for the visual side too. CNNs and GANs pick up on tiny changes, like splicing parts of images or copy move forgeries, even deepfakes. Humans might not see those, but the models do. It's kind of amazing how they catch what we miss.

Some work now combines everything, text and images and structure. A review on AI for identity fraud says integrating multiple techniques boosts things, handles complicated cases better. That makes sense, since one method alone falls short sometimes.
On the forensic end, there's this 2026 study using spectroscopy with machine learning to find chemical changes in docs. Accuracy was high, 97 to 100 percent. But it needs special gear, not great for quick checks.
Presentation attack detection is another area, spotting fakes in remote verifications. Reviews show a move from basic CNNs to feature analysis and bigger models. Still, real world data is scarce, and generalization is iffy. That part gets messy, like models trained on one thing don't work elsewhere.
Lots of current stuff has limits though. Many use fake or small datasets that don't match reality. And with AI making super advanced forgeries, its hard to keep up.

The field has shifted from rule based manual ways to smarter AI systems. OCR plus deep learning seems effective for fake IDs. But we need better ones that scale, work fast, deal with all kinds of docs and new tricks. Its not fully there yet.

IV. DATASET DESCRIPTION

Multi-dimensional records with multiple attributes that are necessary for processing skyline queries are included in the dataset used in this study. Values like price, rating, distance, and other pertinent factors that aid in determining the best outcomes are included in each record.[8] To guarantee data security and privacy, all attribute values are encrypted prior to the dataset being stored on the cloud server. Skyline query operations are then carried out using the encrypted dataset without disclosing the original data values.[9] This dataset aids in assessing the suggested skyline query processing method's effectiveness and capacity to protect privacy.

V. EXISTING SYSTEM

1. Limited Forgery Detection
Most systems currently rely on the use of OCR to obtain text information from an identity document. Although OCR can be very efficient in reading text from an image, it does not help in detecting forgeries. For instance, in an ID card, the image may be swapped, or the font may be changed. In addition, the layout may be changed to include a signature. This may not affect the text in any way, and the OCR system may still be able to read it as legitimate.

2. Low Accuracy with Poor-Quality Images
In a real-world scenario, images of documents are often acquired in sub-optimal conditions. For instance, if a document is scanned, it might have a low resolution, and if a photograph is taken with a camera, it might have varied lighting, shadows, or even reflections and blur. OCR is very sensitive to image quality, and any slight change in an image might cause incorrect or incomplete output during text extraction. For example, if an image of an ID number is slightly blurred or if the date of birth is partially occluded, incorrect output might be generated.

3. Human Dependency
The existing systems that verify the details may still need human intervention for resolving ambiguity and for checking suspicious patterns. This makes the process slow, particularly when dealing with a high volume of documents such as during online account registration and banking transactions. Moreover, humans may get tired and become inconsistent in the process.

4. Inability to Detect Advanced Forgeries
With technology rapidly advancing, it has become much simpler for a person attempting to commit fraud to use image editing software or artificial intelligence-based tools to generate extremely realistic forged documents. Traditional approaches, such as OCR-based or rule-based approaches, cannot identify such sophisticated document forgery. This makes them more prone to identity fraud. Detection of such sophisticated document forgery can only be done through advanced AI models.

VI. PROPOSED SYSTEM

In order to overcome the limitations of the existing methods for identity verification, the present paper proposes a hybrid AI-based system that incorporates Optical Character
Recognition (OCR) and Artificial Intelligence (AI) and Deep Learning for the effective detection of fake and forged ID documents. The proposed system is aimed at analyzing the textual and visual characteristics of the documents for high accuracy and speed.

A. Document Image Acquisition
The system can take scanned document images or photos of identity documents such as Aadhaar cards, passports, or driving licenses. Image preprocessing operations are applied for enhancing the quality of the image, removing noise, enhancing contrast, correcting skew, resizing the image, etc

B. Text Extraction Using OCR
OCR is used to extract text data such as name, date of birth, ID number, and address from the document. The text data is then cleaned and standardized to remove unwanted characters or distortions resulting from poor image quality.

C. Text Analysis & Verification
The system verifies the extracted information with respect to various formats, databases, and constraints. It may also check for inconsistencies like incorrect patterns in ID numbers, incorrect date formats, and

incorrect fields like gender and age.

D. Visual Forgery Detection UsingAI

The document image is analyzed using deep learning techniques, specifically Convolutional Neural Networks (CNNs). The system is able to identify various types manipulation,including changed photos, incorrect usage of fonts, modified logos, and even changed layouts.The system to concentrate on areas that are more likely to be forged.

E. Multimodal Verification & Decision Making

Results from textual as well as visual analysis are integrated, and a decision algorithm decides whether the document is real or fake. If inconsistencies are found, the document is flagged. Otherwise, the document is accepted.

VII. RESEARCH DESIGN METHODOLOGY

The objective of this research is to create a robust system that can identify any form of fake or forged identity documents using a combination of OCR, AI, and deep learning algorithms. The research design is based on a systematic approach that incorporates both text and image analysis for the accurate and automated verification of documents. The methodology has been divided into several major phases that address specific challenges associated with other systems.

1. Research Approach

The research approach employed in this study is quantitative and experimental,

which emphasizes accuracy, precision, and efficiency. Performance metrics such as accuracy, precision, recall, and F1-score are used to validate the system. Datasets containing identity documents like Aadhaar cards, passports, driving licenses, etc., are used to test the system. Both real and artificial data are used. Genuine as well as forged documents are used.

2. Data Collection

Document Types: Aadhaar cards, passports, driving licenses, voter IDs, and other government-issued ID cards. Source: Existing datasets, synthetic forgery datasets, and scanned images for research purposes.

Data Preprocessing: All the images are resized, denoised, and enhanced for better image quality for the process of OCR and AI.

3. Image Preprocessing

Noise Reduction: Filters are applied to reduce background noise, stains, and scanning artifacts.

Contrast Enhancement: Enhances text readability and feature detection.

Skew and Orientation Correction: Corrects text and layout alignment.

Resizing: Normalizes image dimensions for consistent input to OCR and deep learning models.

4. Text Extraction Using OCR

OCR is applied to obtain important text fields such as Name, Date of Birth, ID Number, Address, and other personal details.

Text is pre-processed to avoid errors resulting from image noise or skew.

Textual validation is carried out using rule-based and database validation to identify any text-related errors or invalid formats.

VIII. MODEL COMPARISON

In fake document detection, various AI and machine learning models are compared in relation to their performance in text, image, and layout manipulation. Each model has its strengths and weaknesses:

1. Convolutional Neural Networks (CNNs)
CNNs are mainly used for image processing. They are capable of automatically detecting image features in documents, such as edges, textures, logos, and signatures. They are very effective in detecting photo tampering, logo tampering, and layout inconsistencies.

2. Recurrent Neural Networks (RNN) / LSTM
RNNs, especially LSTM networks, are mainly used for sequential data. They are very effective in detecting anomalies in ID numbers, dates, and names. Their strength lies in detecting inconsistencies in text.

3. Random Forest (RF)
The Random Forest model is a feature-based machine learning model. It mainly depends on structured features, such as font type, spacing, and alignment. It is very effective in detecting forgeries in documents.

4. Support Vector Machine (SVM)
SVM is another feature-based model. It mainly depends on distinguishing between real and fake documents. It performs very well in detecting forgeries.

5. Autoencoders
Autoencoders are unsupervised deep learning models. They are mainly used for anomaly detection. They are very effective in detecting forgeries in documents.

IX. INTEROPERABILITY AND DATA INTEGRATION

Interoperability refers to the ability of a system to work seamlessly with other systems, even though the two systems may use different platforms, software, or data formats. In the case of the AI and OCR system, interoperability would mean that the system would be able to work seamlessly with other external databases or sources of information. This would allow the system to perform real-time checks on the information gathered, such as the ID number, name, or date of birth.

Data integration refers to the process of integrating data from different sources and formats to create a single data structure. In the case of the AI and OCR system, this would mean integrating the data gathered by the system, such as the information gathered by the OCR system, images of the photo and signature, and any barcode or QR code information. Once the data has been integrated, the AI system would be able to analyze the information in the documents in a comprehensive manner.

Overall, interoperability would mean that the AI and OCR system would be able to work seamlessly with other sources of information, while data integration would mean that the system would be able to integrate the different information gathered in a single document.

X. CONCLUSION

The issue of fake ID and document forgery is becoming a major problem in the banking system, government agencies, and educational institutions. The project has proven that using a combination of OCR technology and AI models can be an efficient solution in the detection process. The process of using AI in detecting document forgery includes using the CNN model for image analysis, the RNN model for text analysis, and the Autoencoder model for anomaly detection.

Based on the model comparison table, there is no single model that can perform all the detection processes for document forgery; however, a combination of all three models using CNN for image analysis, RNN for text analysis, and Autoencoder for anomaly detection can perform all the processes. The use of interoperability and data integration can improve the system by enabling cross-system validation and processing of diverse data.
This system can improve the accuracy and efficiency of document verification systems and provide a robust system for real-time automated detection of document forgery.

REFERENCE

[1] Bhatia, M., & Mehta, N. AI-Based Document Forgery Detection Using OCR and Convolutional Neural Networks. SSRN, 2022.
[2] Boonkrong, S. Design of an Academic

Document Forgery Detection System. International Journal of Information Technology, Springer 2024.

[3] Al-Ameri, M. A. A., Mahmood, B., Ciylan, B., & Amged, A. Unsupervised Forgery Detection of Documents: A Network-Inspired-approach.Electronics 2023.

[4] Automatic Fake Document Identification and Localization using DE-Net and Color-Based Features of Foreign Inks. Journal of Visual Communication and Image Representation (2023).

[5] DFD-SS: Document Forgery Detection using Spectral-Spatial Features for Hyperspectral Images. Journal of Visual Communication and Image Representation (2022).

[6] Deep Feature Extraction for Document Forgery Detection with Convolutional Autoencoders. Computers and Electrical Engineering. (2022).

[7] Praba, G. C., Jeevitha, E., Abitha, A., Shalini, A., & Swetha, B. Fake Education Document Detection using Image Processing and Deep Learning. IJERT. 2021.

[8] Yoosuf, M. S., & Ramachandran, A. Forgery Document Detection in Information Management Systems using Cognitive Techniques. Journal of Intelligent & Fuzzy Systems. (2020) -

[9] Castelblanco, A., Solano, J., Lopez, C., et al. Machine Learning Techniques for Identity Document Verification in Uncontrolled Environments: A Case Study. Springer. 2020.

[10] George, A., & Marcel, S. EdgeDoc: Hybrid CNN-Transformer Model for Accurate Forgery Detection and Localization in ID Documents. arXiv. (2025).