Grenze International Journal of Engineering and Technology, January Issue

Automated Document Verification for Official Documents

Niteen Dhutraj[1], Sujal Sonawane[2], Mayur Patil[3], Lalit Patil[4] and Jayesh Patil[5]
[1-5]Department of Information Technology, SVKM's Institute of Technology, Dhule, India
Email: niteen125@gmail.com, sujalsonawane2414, mayurpatil0702, lp.lalitpatil, patiljayesh1242@gmail.com

***Abstract*— The system is an automated approach for document verification using modern technology such as optical character recognition (OCR), natural language processing (NLP), and machine learning techniques. It aims to enhance the precision, security, and efficacy of authenticating official documents, including certificates. The project entails the creation of an automated system for verifying official documents, using sophisticated image processing techniques. Automating the verification process enables companies to substantially decrease the time and expenses linked to human inspections while enhancing accuracy in fraud detection. The system is used to separate the delicate information, draw real safety properties such as hologram and stamps and create the completeness of the document through the consistency of the structure and content. This method is conducted deeply for automatic documents. Demonstrates the deep learning forms to specify and analyses documents such as passports, certificates and documents issued by the government while maintaining accuracy, safety and efficiency. The presented method helps reduce human analysis errors and improve analysis through the sector that requires document inspection.**

***Index Terms*— Feature extraction, document authentication, image processing, automated document verification, deep learning, optical character recognition (OCR), convolutional neural networks (CNNs), and natural language processing (NLP).**

I. INTRODUCTION

Document verification including certification, it is an important requirement in many businesses. Routine manual testing is labour intensive. Vulnerable to errors and fraud, the evolution of image processing technology may provide automated mechanisms for quickly analysing and validaWting documents. Therefore, these problems can be overcome. This document describes a step-by-step approach to building an image processing system that can verify document authenticity. The system includes pattern recognition, feature extraction and integrity testing. This research proposed new methods. For text mining with deep learning methods like optical character recognition (OCR), convolutional neural networks (CNNs). The large amount of document exchange in the digital age increases the risk of forgery and fraud. Automated document review tools attempt to detect fraud by analysing document structure and content. Rule-based approaches are susceptible to fluctuations in false positives. Deep learning models, however, can improve accuracy and adaptability. This paper presents a comprehensive solution with deep learning for document analysis focusing on textual and image feature extractionEach major section begins with a Heading in 10 point Times New Roman font centered within the

Grenze ID: 01.GIJET.12.1.511

column and numbered using Roman numerals (except for ACKNOWLEDGEMENT and REFERENCES), followed by a period, two spaces, and the title using

II. Literature Review

Document verification technology has made considerable evolution- the rule-based field detection to the most advanced deep learning and multimodal based document verification technology. In this part, the existing works are discussed in four broad areas, including traditional approaches, deep learning based approaches to textual verification, visual-based systems, and cross-modality frameworks.

Early methods completely depended on template matching, a fixed-layout zoning, and heuristics. These systems were affected by the change of layout or resolution. Indicatively, Nwanze et al. [2] suggested a certificate verification technique that uses neural networks but their system was still subjected to manual zoning rules, downsizing the level of scalability and filing against layout changes.

These systems were also unable to manage semantic inconsistencies, or identify clever forgeries like reprinted seals or altered dates.

Recent work has exploited CNN and OCR methods to pull text and confirm it. Baviskar et al. [3] CNNs were introduced to use in automating the academic certificate classification and extraction of text. In their turn, Toprak & Turan [4] proposed a transformer pipeline to check on the semantic integrity of the documents concerning finance so that the information could be checked across fields (e.g. perform a check on the currency format, address matching, and so on). Such models were however only devoted to textual integrity and unable to notice visual attack like blurred stamps, displaced logo, tampered signatures.

Similar works have been done on detection of visual manipulation by analyzing individual pixels and the recognition of image patterns. Sarode, Sharmi, and Mirza [1] integrated the blockchain and machine learning to identify abuse in digital signatures. Castelblanco et al. [5] have applied SVM and CNN classifiers in detecting seals and holograms of identity with a high value, or detection rate, notably, in noisy conditions. Though effective, these systems tend to not consider textual consistency and are not able to confirm semantic relationships between fields, so they become vulnerable to false positives in the real world.

Saputra & Nurhaida [6] applied an EfficientNetV2M model to work with signature verification and trained it using a varied set of real and fake instances. It was good in printed pages but since it was not flexible to various signature styles and required real time validation; it was still difficult. The methodology of other systems [8], including feature descriptors and augmentation, was to train document verification pipelines, but without combining the unified framework between text and image modality.

Moolchandani and colleagues [10] proposed a two-branch architecture where they use one to handle a textual model based on LSTMs, and another one that handles an image-based using CNNs. Although they are promising, their methodology features no single decision fusion unit and does not attend to semantics-visual inconsistencies (e.g. conflict between stamp authenticity and date of issue). Rajapashe et al. [9] tried a multi-format verification engine with focus on classification of document layout mostly, than verification itself.

Our identified system is unique, as it aims to integrate the following aspects:

- Structural feature extraction using Deep CNNs (e.g. ResNet),
- Context-aware text validation by means of transformer-based NLP,
- A cross-verification layer with correlation of semantic and visual distractors (e.g. verifying that a recent date appears on an issue and an old-looking stamp),
- Adaptive signature verification both in geometric and in texture descriptors,
- Performance in real-time through model optimization and an expandable administrator interface to keep learning continuously.

III. Research Methodology

The automated document verification system architecture consists of different elements that collaborate to provide quick, secure, and accurate verification of documents. The system provides the advantage of control to administrators, through which they are able to monitor the process of document verification, analyses results, and adjust settings. The uploaded documents are converted into uniform formats, i.e., images or text, and then fed through pre-processing steps. The system reads text information from document images and converts them into machine-processable formats. The system uses models like Tesseract and other tools like internal optical character recognition (OCR). The technique utilizes convolutional neural networks (CNNs) to extract relevant information from document images, such as logos, watermarks, signatures, and graphic structure. The system

utilizes Natural Language Processing (NLP) models to extract and authenticate relevant textual items such as names, dates, and serial numbers. They are cross-referenced with a reference database if necessary. The system uses a deep learning model that has already been trained, like CNN, to find changes in the structure, texture, or patterns of the data that could mean the document has been tampered with. To authenticate the document, the system juxtaposes the retrieved textual and visual attributes against a reference dataset, such as governmental databases or organisational records.

- The user will upload documents.
- After that, the system will select features from the document, like:
  - Name of user
  - Document number
  - Date of issue
- The user will be able to examine all the features on screen after selection.
- The user will click an verify tab then document will be transferred to office portal
- The document will appear here on the office portal.
- The office portal will then scan the document, extract its features, and compare them with those in the databases.
- After examining features, if the office portal finds that the document is legit, then the user will get a token that your document is verified.

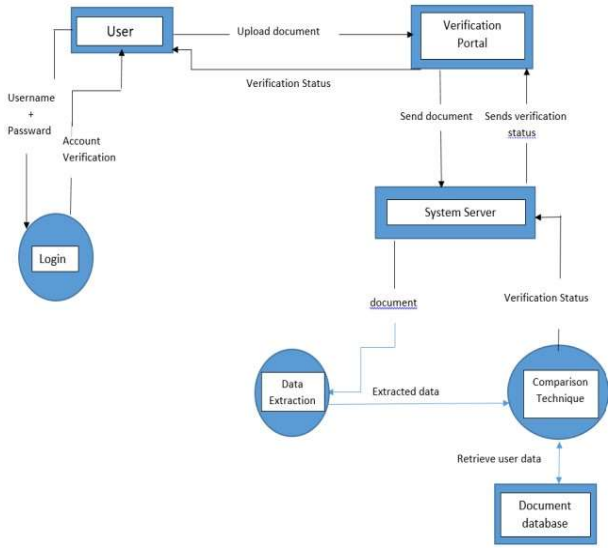

Fig. 1. System Architecture of Proposed System

A. Implementation process

Input: A dataset comprising various official documents (e.g., certificates) was used. Each document underwent pre-processing and labelling to identify authentic and tampered features.

Pre-processing: Image enhancement techniques such as noise reduction, thresholding, and segmentation. Filters such as Gaussian blur are applied to remove noise and improve image clarity. We use adaptive thresholding techniques to convert documents to binary format, which facilitates easier feature extraction. Canny or Sobel edge detectors help to identify the boundaries of text blocks, logos, or seals.

Feature Extraction: Use CNN for extracting key document features like seals, logos, text patterns, and security elements. We use Optical Character Recognition (OCR) to extract and digitise text from document images. Advanced image processing techniques like Fourier transforms are applied to detect specific features of watermarks and holograms, ensuring their authenticity. We apply geometric and texture-based methods to detect and verify handwritten signatures.

Optical Character Recognition (OCR): Implementing OCR to extract textual content from documents for comparison with known valid information. OCR excels in image classification tasks, specifically in differentiating official document types. OCR, when paired with CNN and Long Short-Term Memory (LSTM) networks, can verify sequences such as ID numbers or other standard formatting.

B. Training Dataset Process

Input: Threshold Th, activation functions [], training dataset [].
Output: Determined Features A completed training module must include certain features.
Steps:

- Define the data input array d[], select an appropriate activation function, and determine the epoch size for training.
- Features_data_list ← Feature-Extraction data (d[])
- Features-Dataset [] ← optimized (Features_data_list)
- Return train_rule_list []

C. Testing Process

Input: Features extracted from the testing instance set, Data [i...n], are derived according to the training rules, PSet [1]...T[n].
Output: document verify (Yes or No).

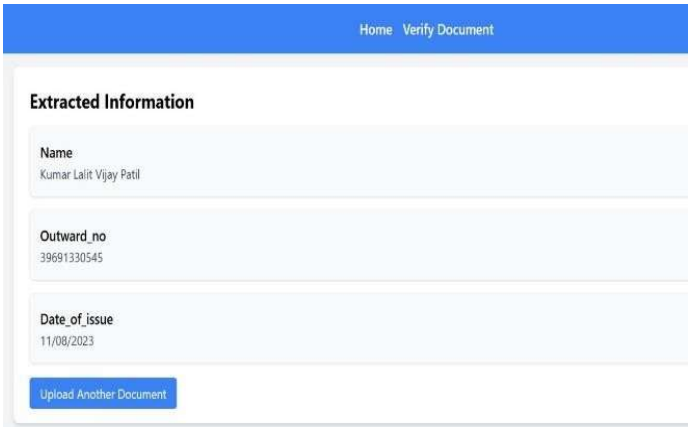

Fig. 2. Feature Extraction

- For each Data[i] within the dataset, select n attributes from Data[i] using the formula provided below.

$$\text{TestsDB(j)} = \sum_{\text{j=1}}^{\text{n}} \text{attribute (D[i] ….. D}[n]) \quad (1)$$

Proof:
Let D[i] be a document with n attributes A_1, A_2,…, A_n.
The function attribute(D[i]) extracts the relevant attributes from the document.
Thus, we can express it as: TestsDB(j) = A_1+ A_2+…A_n+ = $\sum_{j=1}^{n} A_j$
This summation represents the aggregation of attributes for document D[i].

- For each (PSet [i] from PSet),

$$\text{TrainDB[t]} = \sum_{\text{t=1}}^{\text{n}} \text{attribute (T[i] ….. T[n])} \quad (2)$$

Proof:
Let T[i] be a training document with n attributes B_1,B_2,…, B_n.
The function attribute(T[i]) extracts the relevant attributes from the training document.
Thus, we can express it as: TrainDB[t] = B_1+ B_2+…+ B_n = $\sum_{t=1}^{n} B_t$
This summation aggregates the attributes for the training documents.

- Evaluate the training and testing examples using the formula given below

$$\text{TestDataset[k].weight} = \text{similarity(TestsDS[k]} \sum_{\text{m=1}}^{\text{n}} \text{TrainDS [m]} \quad (3)$$

Proof:
Let TestsDS[k] be the k-th test document.
The aggregated training dataset can be represented as: TrainAgg = $\sum_{m=1}^{n} TrainDS[m]$

The similarity function similarity(A,B)similarity(A,B) can be defined defined (e.g., cosine similarity) as:
similarity(A,B) = $\frac{A.B}{||A||\,||B||}$
Thus, the weight can be expressed as: TestDataset[k].weight = similarity(TestsDS[k], TrainAgg)

- If (TestDatalist[k]: score >= Th),

TestData_list[k].class ← TrainDB[m]: (4)

Let score be the computed similarity score from the previous step.
The condition checks if: score ≥ Th
If true, we classify the test document based on the training data: TestData_list[k].class = TrainDB[m]
This indicates that the test document is classified under the same category as the training document m.

- Return class

IV. RESULTS

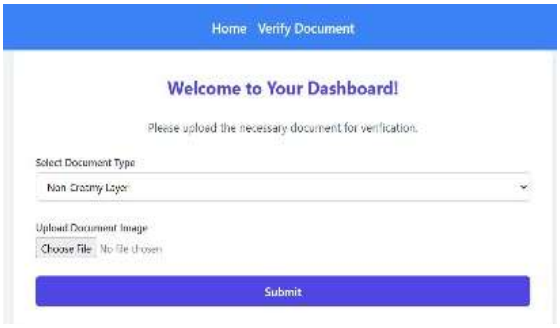

Fig 3

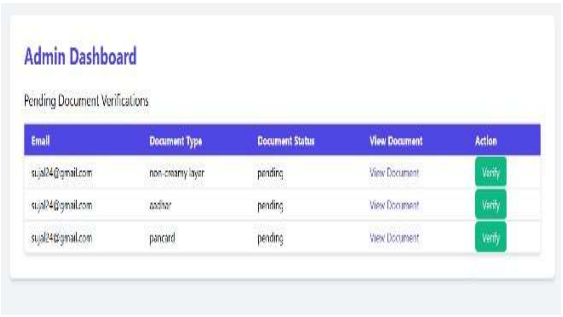

Fig 4

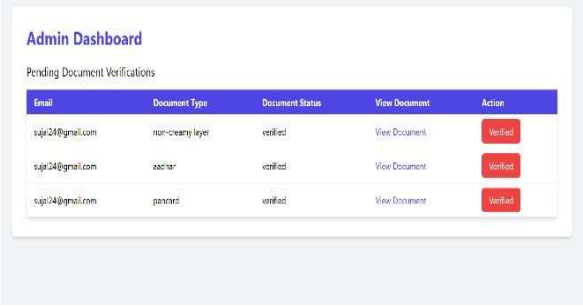

Fig 5

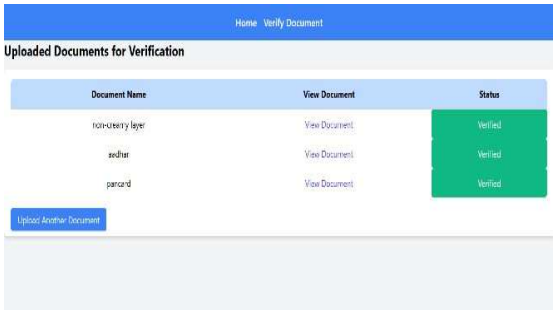

Fig 6

A. Result Analysis

TABLE I. ACCURACY TABLE

Document Verification Metrics	Percentage (%)
Document Type Recognition	100%
Text Extraction Accuracy	96%
Field Extraction Accuracy	97%
Image Quality Assessment	60%
Overall Document Verification	95%

TABLE II. PERFORMANCE METRICS REPORT

Metric	Value	Description
Accuracy	95%	Percentage of correct document Verifications (true positives + true negatives).
Precision	96%	The proportion of relevant results (true positives) among all identified documents.
Recall	94%	The proportion of relevant documents that were correctly identified.
F1-Score	95%	The harmonic mean of Precision and Recall.

The graph represents the accuracy performance of a document verification system across different metrics.
Document Type Recognition (100%) The system perfectly recognizes document types with 100% accuracy. This suggests zero errors in classifying documents into predefined categories.
Text Extraction Accuracy (96%) The system extracts text from documents with a high accuracy of 96%. This means that most extracted text is correct, with minor errors or omissions.
Field Extraction Accuracy (97%) The system correctly extracts specific fields (e.g., names, dates,
IDs) with 97% accuracy. This shows strong performance in identifying and retrieving structured data from documents.
Image Quality Assessment (60%) The lowest accuracy (60%) indicates room for improvement in assessing document image quality. This suggests that the system may struggle with blurry, low-resolution, or poorly scanned documents.
Overall Document Verification (95%) The system achieves an overall verification accuracy of 95%, combining multiple factors. This means that 95 out of 100 document verifications are correct.

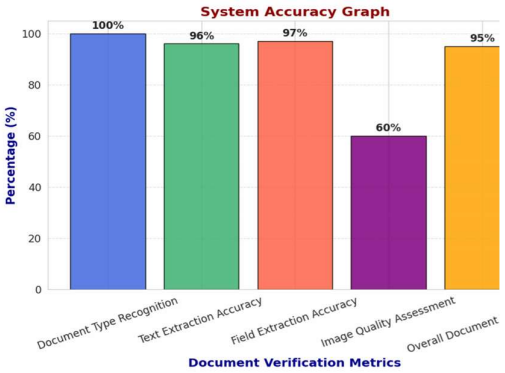

Fig. 7. System Accuracy Graph

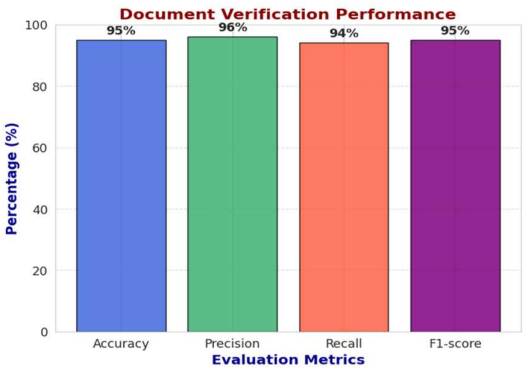

Fig. 4. Performance Metrics

The graph above shows measures of performance of a document verification system. Each bar shows the percentage score of different measures used to quantify the effectiveness of the system.
Key Indicators and What They Mean:
Accuracy (95%) Accuracy means the system's overall accuracy at authenticating documents. 95% accuracy means 95 out of every 100 document authentications were correct.
Precision (96%) Precision refers to the number of confirmed documents that were indeed relevant. 96% precision means that most of the identified documents were correct with minimal false positives.
Recall (94%) Recall is the percentage of relevant documents the system was able to retrieve. 94% recall indicates that the system did miss some relevant documents, but was overall doing a good job.
F1-score (95%) F1-score is a harmonic mean of recall and precision. 95% F1-score indicates that the system has high recall and precision, i.e., it is able to retrieve relevant documents and reduce errors.

V. Conclusion

The application of deep learning-based automated verification systems for documents has been proven to have great potential to significantly enhance the speed, accuracy, and reliability of official document verification procedures. The system proposed can identify counterfeit documents while authenticating genuine documents using advanced deep learning techniques, such as convolutional neural networks (CNNs) for feature extraction and classification procedures. The research proved an automated process with deep learning methods for official document verification. This is achieved by eliminating the limitations with manual verification, such as inefficiency in terms of time, human error, and vulnerability to forgery. The method proposed entails image pre-processing, feature extraction with CNN-based models, and classification algorithms, which result in increased accuracy and security in verifying documents. Empirical results showed that the model was impressive in authenticating counterfeit documents, thus increasing the accuracy and efficiency of the verification process. This technology can potentially provide an automated, efficient, and safer method for institutions dealing in the banking, educational, and government sectors by reducing the application of human operators. Future research will be focused on adding multilingual text recognition, enabling real-time verification, and the application of blockchain technology for increased security and improved traceability.

REFERENCES

[1] Sarode, Shantanu, et al. "Document manipulation detection and authenticity verification using machine learning and blockchain." Int. Res. J. Eng. Technol 7 (2020): 4758-4763.

[2] Nwanze, D. E., O. P. Okechukwu, and C. H. Nnaji."DOCUMENT VERIFICATION SYSTEM FOR FRAUD DETECTION USING MACHINE LEARNING TECHNIQUE." International Journal of Computing, Science and New Technologies (IJCSNT) 1.1 (2023): 1-9.

[3] Baviskar, Dipali, et al. "Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions." IEEE Access 9 (2021): 7289472936.

[4] Toprak, Ahmet, and Metin Turan. "TransformerBased Approach for Automatic Semantic Financial Document Verification." IEEE Access (2024).

[5] Castelblanco, Alejandra, et al. "Machine learning techniques for identity document verification in uncontrolled environments: A case study." Pattern Recognition: 12th Mexican Conference, MCPR 2020, Morelia, Mexico, June 24–27, 2020, Proceedings 12. Springer International Publishing, 2020.

[6] Saputra, Muhammad Azi, and Ida Nurhaida. "Signature Originality Verification Using A Deep Learning Approach." Electronic Journal of Education, Social Economics and Technology 5.1 (2024): 19-29.

[7] Castelblanco, Alejandra, et al. "Machine learning techniques for identity document verification in uncontrolled environments: A case study." Pattern Recognition: 12th Mexican Conference, MCPR 2020, Morelia, Mexico, June 24 – 27, 2020, Proceedings 12. Springer International Publishing, 2020.

[8] Shende, Abhishek, Mahidhar Mullapudi, and Narayana Challa. "ENHANCING DOCUMENT VERIFICATION SYSTEMS: A REVIEW OF TECHNIQUES, CHALLENGES, AND PRACTICAL IMPLEMENTATIONS." Technology (IJCET) 15.1 (2024): 16-25.

[9] Rajapashe, Madura, et al. "Multi-format document verification system." American Scientific Research Journal for Engineering, Technology, and Sciences 74.2 (2020): 48-60.

[10] Moolchandani, Jhankar, Rinki Pakshwa, and Kulvinder Singh. "Machine Learning for Identifying and Validating Document Authenticity." 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 2024.

[11] Bhushan Chaudhari, Rajesh Prasad, Jayashri Prasad, Nihar Ranjan, Rajat Srivastava, “FCM with Spatial Constraint Multi-Kernel Distance-based segmentation and Optimized Deep Learning for flood detection”, International Journal of Image and Graphics, Vol. 22, No. 5, 2023, ISSN: 1793-6756, pp. 2450041-1 to 2450041-35