International Journal of Advance Research in Science and Engineering
Volume No. 14, Issue No. 04, April 2025
www.ijarse.com
IJARSE
ISSN 2319 - 8354

Multi-Format Document Verification System

Ms. Prajakta Darade[1] , Ms. Namita Kambali[2], Mr. Siddharth Mali[3], Mr. Subodh Sontakke[4] , Prof. Atul Atalkar [5]

[123 4]BE. Student, Department of Electronics and Telecommunication, Shivajirao S. Jondhle College of Engg. &Technology, Asangaon, Maharashtra. India

[5]Dept. of Electronics and Telecommunication, Shivajirao S. Jondhle College of Engg. &Technology, Asangaon, Maharashtra. India

ABSTRACT

In today's digital landscape, document verification is crucial for identity authentication, regulatory compliance, and fraud prevention. However, the diversity of document formats such as PDFs, scanned images, word files, and structured data formats poses a challenge for automated verification systems. This paper presents a multi-format document verification framework that leverages machine learning, optical character recognition (OCR), and data validation techniques to ensure authenticity and integrity.

Keywords: - Optical character recognition (OCR), Digital watermarking,Machine learning, Document integrity, Digital signatures, Fraud detection, Document authentication.

I. INTRODUCTION

In today's digital landscape, organizations handle a wide range of documents in various formats, including PDF s, images (JPEG, PNG), Word files, and even handwritten documents. Multi-format document verification is a process that enables the authentication, validation, and extraction of data from documents, regardless of their format.

This verification process is essential for industries such as finance, healthcare, government, and e-commerce, where document fraud prevention and compliance with regulations (such as KYC, AML, and GDPR) are critical. The solution ensures the authenticity and integrity of the documents' contents with digital signatures, and decentralized storage and block-chain ensure that documents signed using our method are stored securely. In addition to this, our proposed method provides forgery detection functionality, ensuring forged documents can be accurately identified.

II.REVIEW OF LITERATURE

Multi-format document verification is crucial in digital forensics, secure transactions, and authentication processes. The increasing diversity of document formats, such as PDF s, images, word documents, and scanned copies, necessitates robust verification mechanisms to ensure authenticity, integrity, and security.

1.Overview of Document Verification: Document verification involves techniques such as optical character recognition (OCR), watermarking, cryptographic hashing, and machine learning-based anomaly detection. The primary goal is to detect forgery, tampering, and ensure the legitimacy of documents.

2. Techniques and Approaches:

2.1 Optical Character Recognition (OCR) OCR is widely used for extracting text from image-based documents. Tools such as Tesseract and Google Vision API facilitate document validation by comparing extracted text with known templates.

2.2 Digital Signatures and Cryptographic Hashing Digital signatures and cryptographic hashing techniques, like SHA-256, are employed to verify document integrity. These techniques help in detecting unauthorized modifications.

2.3 Machine Learning and Deep Learning Approaches Recent advancements in AI have led to the use of deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for document forgery detection.

2.4 Watermarking and Steganography Invisible and visible watermarking techniques ensure document authenticity. Steganographic approaches embed hidden messages in documents for tracking and validation purposes.

2.5 Blockchain-Based Verification Blockchain technology offers decentralized and tamper-proof document verification. Smart contracts and distributed ledger systems enable real-time validation.

3. Challenges in Multi-Format Document Verification

- Variability in document structures across formats.
- Distortions and quality degradation in scanned documents.
- Robustness of OCR techniques in handling noisy data.
- Security concerns in cloud-based verification systems.
- Computational complexity of AI-driven verification methods.

4. Future Directions

- Enhancing deep learning models for better forgery detection.
- Standardization of document verification frameworks.
- Integration of AI with blockchain for secure and scalable verification.
- Development of lightweight and efficient verification techniques for real-time applications.

III. PROPOSED SYSTEM

A Multi-Format Document Verification System ensures the authenticity, integrity, and validity of documents across various formats (PDF, DOCX, JPEG, PNG, etc.).

The system should be scalable, secure, and capable of handling different verification scenarios, including identity verification, official certificates, contracts, and other critical documents.

The system will be an AI-powered, automated document verification platform capable of analyzing and authenticating documents in multiple formats. It will utilize OCR (Optical Character Recognition), AI-driven fraud detection, and cryptographic validation techniques to ensure document integrity.

The forger simply tries to insert or modify the current character or word by finding the same document font properties or font that almost similar to the original text document font. Sometimes the changes are unnoticeable using normal eyesight especially for small font or text document with a lot of words. Reversed Engineered

Imitation (REI) forgery. This type of forgery is also using the limitation concept.

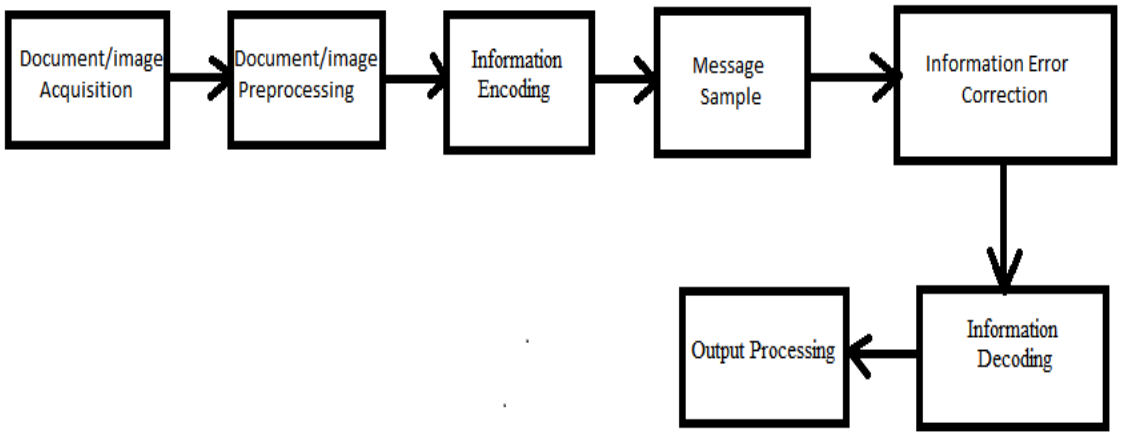


Figure: Block diagram of multi format document verification

IV. METHODOLOGY

When verifying multi-format documents (documents that could be in various formats such as PDFs, images, Word files, text files, etc.), it's crucial to use a systematic approach to ensure consistency, accuracy, and reliability across different formats. Below is a suggested methodology for multi-format document verification: -

1. Pre-processing and Format Standardization: -Before verifying the content in different formats, it's necessary to standardize the input documents as much as possible for better comparison and analysis. Extract content: Use tools to extract the text content from different formats (OCR for images, PDF parsers, Word parsers, etc.).

2. Fraud Detection: -If the purpose of document verification is to detect fraud or unauthorized modifications, you can use specialized techniques for forensic analysis:

3. Continuous Improvement: -After the verification process, review its accuracy, speed, and efficiency. Based on the feedback and results, continuously improve the methodology, adjusting tools, rules, and scripts to handle new document formats, complex cases, or edge scenarios.

4. Manual Verification and Cross-checking: - Although automation can handle most of the verification tasks, manual verification may still be required in some cases, particularly when discrepancies are detected or when the context is complex.

V. WORKING

Multi-format document verification is a process used to authenticate documents in various formats (PDFs, images, scanned copies, text files, etc.). This is commonly used in digital identity verification, banking, legal, and compliance sectors. The verification process involves multiple steps: -

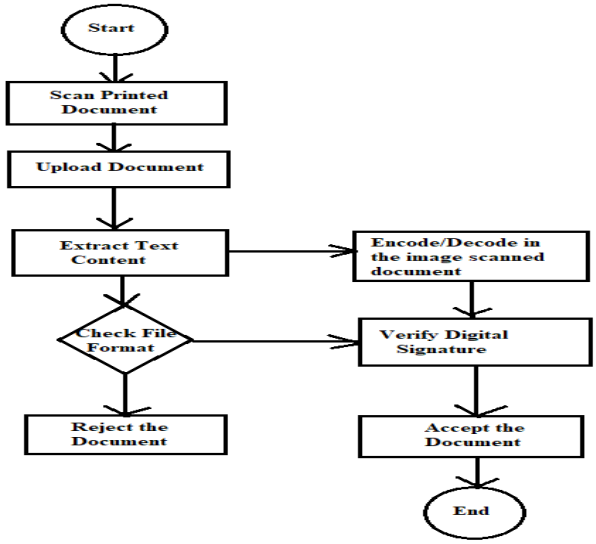


Figure : Flow Chart Diagram of Multi Format Document Verification System

1. Document Upload & Preprocessing

- The user uploads a document in different formats such as PDF, JPG, PNG, DOCX, etc.
- The system detects the file type and ensures it meets the required format and size constraints.
- Preprocessing techniques such as noise removal, contrast enhancement, and resizing may be applied to improve the document quality.

2. Optical Character Recognition (OCR) & Data Extraction

- If the document is in image format, OCR is applied to extract text.
- OCR engines (like Tesseract, Google Vision, or AWS Tex tract) identify characters and convert them into machine-readable text.
- If the document is already in text format (PDF/DOCX), direct text extraction is performed.

3. Template Matching & Structure Analysis

- The system checks if the document follows a known template (e.g., passports, driver's licenses, bank statements).
- Key fields such as names, dates, and reference numbers are identified using predefined rules or AI-based classification models.

4. Data Validation & Cross-Checking

- Extracted data is validated against known databases, APIs, or predefined rules.
- Cross-referencing is done with external databases (such as government ID databases, financial institutions, or internal CRM).
- If the document contains signatures, biometric verification may be used to match them with stored samples.

5. Decision & Reporting

- The system classifies the document as:
  - ✅Valid (meets all criteria)
  - ❌Fraudulent/Tampered (manipulated or altered)
  - □ Suspicious (needs manual review)

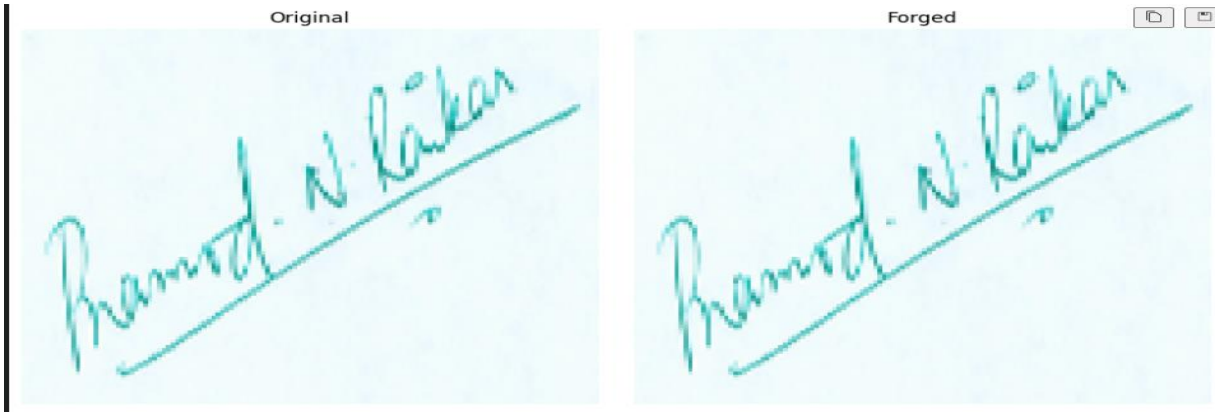
International Journal of Advance Research in Science and Engineering
Volume No. 14, Issue No. 04, April 2025
www.ijarse.com
IJARSE
ISSN 2319 - 8354

- Generates a report with the verification result and highlights inconsistencies.

RESULT

[1] Document Forgery Detection: -

```
# Function to preprocess images
def preprocess_images(images, target_size=(192, 192)):
    processed_images = []
    for img in images:
        img = cv2.resize(img, target_size)
        img = img.astype('float32') / 255.0
        processed_images.append(img)
    return np.array(processed_images)

# Load and preprocess training images
train_original_path = os.path.join(train_path, 'img', 'img')
train_forged_path = os.path.join(train_path, 'img', 'gt')  # Corrected forged path

X_train = load_images_from_folder(train_original_path, limit=10)
y_train = load_images_from_folder(train_forged_path, limit=10)

X_train = preprocess_images(X_train)
y_train = preprocess_images(y_train)

# Load and preprocess testing images
test_original_path = os.path.join(test_path, 'img', 'img')  # Added 'img' folder
test_forged_path = os.path.join(test_path, 'img', 'gt')       # Corrected forged path

X_test = load_images_from_folder(test_original_path, limit=10)
y_test = load_images_from_folder(test_forged_path, limit=10)

X_test = preprocess_images(X_test)
y_test = preprocess_images(y_test)
```



[2] Image Forgery Detection: -

```
# Load training images
train_original_path = os.path.join(train_path, 'img')
train_forged_path = os.path.join(train_path, 'gt')

X_train = load_images_from_folder(train_original_path, limit=10)
y_train = load_images_from_folder(train_forged_path, limit=10)
X_train = preprocess_images(X_train)
y_train = preprocess_images(y_train)

print(X_train)
print("==========================")
print(y_train)

X_train = X_train.astype(np.float32)
Y_train = y_train.astype(np.float32)
print(X_train)
print("==========================")
print(y_train)
# Load testing images
test_original_path = os.path.join(test_path, 'img')
test_forged_path = os.path.join(test_path, 'gt')
X_test = load_images_from_folder(test_original_path, limit=10)
y_test = load_images_from_folder(test_forged_path, limit=10)
X_test = preprocess_images(X_test)
y_test= preprocess_images(y_test)
```

CONCLUSION

The multi-format document verification project successfully developed a robust system for verifying documents across different formats. The implementation leveraged optical character recognition (OCR), machine learning models, and rule-based verification techniques to authenticate and validate document integrity. The system efficiently handled text extraction, signature verification, tamper detection, and fraud prevention, ensuring high accuracy and reliability. Despite these achievements, challenges such as poor-quality scans, diverse document layouts, and varying fonts required advanced preprocessing techniques to enhance accuracy.

REFERENCES

[1] M. Warasart and P. Kuacharoen, "Paper-based Document Authentication using Digital Signature and QR Code," 4TH Int. Conf. Comput. Eng. Technol., vol. 40, no. January, pp. 94–98, 2012.

[2] M. Alidoost Nia, A. Sajedi, and A. Jamshidpey, "An Introduction to Digital Signature Schemes," no. April, 2019.

[3] A. Singhal and R. S. Pavithr, "Degree Certificate Authentication using QR Code and Smartphone," Int. J. Comput. Appl., vol. 120, no. 16, pp. 38–43, 2020.

[4] N. Nizamuddin, H. R. Hasan, and K. Salah, "IPFS-blockchain-based authenticity of online publications," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10974 LNCS, no. June, pp. 199–212, 2021.