IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, November 2025
Impact Factor: 7.67

Document Forgery Detection using OCR and Deep Learning

Dr. Smita Nirkhi Singh, Ms. Sayali Kakde, Ms. Shivani Pimpalshende, Ms. Shivani Govindwar, Ms. Shravani Bhandari, Ms. Tanushree Raut
Department of Artificial Intelligence
G. H Raisoni College of Engineering and Management Nagpur, India

Abstract: *The integrity of textual information is paramount in domains such as banking, legal services, and digital identity management, where document forgery has emerged as a critical security concern. Traditional methods of manual verification are slow, prone to mistakes, and can't keep up with the increasingly complex nature of forgeries. Incorporating Optical Character Recognition (OCR) and Deep Learning models, this study introduces an automated framework for detecting document forgeries. Optical character recognition (OCR) is used to extract structural and textual information from digital or scanned documents, and deep learning algorithms look for visual and textual discrepancies to detect possible manipulation. To detect patterns of forgery in space, language, and context, the suggested method uses architectures based on convolution and sequences. Experiments on benchmark datasets show excellent accuracy, precision, and recall, proving the method's resilience to a variety of forgery techniques, such as text replacement, insertion, and style manipulation. The system offers a dependable way to improve document authentication procedures in financial institutions, e-governance, and forensic investigations. It is also scalable and flexible enough to accommodate real-world applications.*

Keywords: Document Forgery Detection, Optical Character Recognition (OCR), Deep Learning, Convolution Neural Network (CNN), Text Extraction, Image Processing, Computer Vision

I. INTRODUCTION

In fields like finance, legal systems, government services, and education, where even small changes can result in large financial losses or legal repercussions, document authenticity is vital. Document forgery, which started out as basic text edits but has since progressed to more complex manipulations using modern editing tools, has grown in sophistication alongside the fast expansion of digitization. When faced with large-scale and complicated forgery attempts, traditional manual verification methods fall short due to their inefficiency and human error.

Recent years have seen the development of computational approaches that automate document verification.

When presented with a variety of forgery techniques, traditional image-processing and rule-based methods frequently fail to provide adequate protection. This is what prompted researchers to combine Deep Learning with Optical Character Recognition (OCR) for better, more scalable solutions. Optical character recognition (OCR) allows for the efficient extraction of textual information from digital or scanned documents, and deep learning models, especially CNNs and hybrid architectures, offer strong capabilities in learning spatial and semantic patterns that differentiate real documents from fakes.

Using optical character recognition (OCR) text extraction and deep learning (DL) classification, this study aims to build and test an automated system for document forgery detection. human inspection are the goals of the suggested method, which makes use of both visual and textual features.

II. RELATED WORK

1. Conventional Methods for Identifying Document Forgeries

Heuristic-based methods and manual inspection were the mainstays of early document forgery detection research. Watermark analysis, signature authentication, and texture-based printed text verification were common methods.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

Although these techniques worked well for identifying simple changes, they were labor-intensive, largely reliant on human judgment, and not scalable for big datasets. Additionally, these methods were unable to withstand sophisticated digital manipulations, which frequently left no observable artifacts that could be examined by humans.

2. Methods Based on Optical Character Recognition (OCR)

Automated document analysis underwent a dramatic change with the advent of optical character recognition (OCR). OCR made it possible to compare fonts, character spacing, and text consistency by extracting text from scanned or image-based documents. OCR was used by researchers for tasks like alignment detection, font shape comparison, and textual content semantic verification. However, issues like noisy images, different scan quality, and multilingual text made OCR-based techniques insufficiently reliable. These restrictions made it harder for them to identify complex or subtle forgeries.

3. Deep Learning Methods for Identifying Forgeries

The field of forgery detection made significant strides with the advent of deep learning. Convolutional Neural Networks (CNNs) have shown great promise in identifying manipulations like splicing, copy-move, and character replacement by extracting deep visual features from document images. Long Short-Term Memory (LSTM) models and Recurrent Neural Networks (RNNs) were also investigated for capturing contextual and sequential dependencies in document text. Transformer-based models and attention mechanisms, which provide enhanced semantic understanding and resilience against noisy or deteriorated documents, have been used more recently. Despite these developments, the majority of deep learning techniques mostly concentrate on features at the image level.

4. Hybrid Methods

Deep Learning and OCR Together Researchers have recently suggested hybrid approaches that combine OCR with deep learning models in order to get around the drawbacks of individual techniques. In these systems, textual content is first extracted using OCR, and then semantic and visual consistency are confirmed by deep learning models. For example, OCR text is examined for linguistic irregularities or unusual alterations, while CNNs are used to identify image-based manipulations. To increase detection accuracy, some studies have also cross-checked OCR output with external knowledge bases or metadata.

5. Motivation and Research Gap

According to the reviewed literature, deep learning improves visual analysis and traditional methods offer basic verification, but neither is adequate on its own for thorough forgery detection. Despite being helpful for semantic analysis, OCR-based systems have drawbacks when dealing with complex or noisy document formats. Although they still have issues with scalability, multilingual adaptability, and the detection of subtle content-level forgeries, hybrid OCR-deep learning techniques show promise. These shortcomings show how important it is to have a strong framework that combines deep learning and OCR's advantages to guarantee accurate and trustworthy document authentication.

6. Opportunities and Difficulties

The process of identifying document forgeries still faces a number of difficulties despite tremendous progress. Low-quality or damaged scans are a significant problem since they make it difficult to extract trustworthy features for deep learning and OCR models. The existence of multilingual documents and a variety of scripts presents another challenge, as current systems still struggle to maintain consistent recognition accuracy. Furthermore, detection is made more difficult by the ability of contemporary editing tools to make subtle changes to layout, text, and images without leaving obvious traces.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

Summary
Traditional manual and heuristic-based techniques for document forgery detection, like watermark analysis and signature verification, have given way to automated techniques that use optical character recognition (OCR) and, more recently, deep learning. OCR makes it possible to extract and verify text, but it has trouble with documents that are noisy, multilingual, or of poor quality. Although visual forgery detection has been enhanced by deep learning techniques, especially CNNs and transformer-based models, semantic- level inconsistencies are frequently overlooked. By utilizing both textual and visual cues, hybrid frameworks that combine OCR and deep learning show promise. However, they still have issues with handling complex digital edits, multilingual adaptability, and high computational costs. These gaps point to areas that warrant further investigation, such as cross-modal verification methods, multilingual OCR integration, and lightweight deep learning models.

II. METHODOLOGY

1. System Design
With an emphasis on accuracy, scalability, and modularity, the system is a multi-layered pipeline that combines web, computer vision, and natural language processing technologies. Its preprocessing layer uses adaptive thresholding, skew correction, resizing, noise reduction, and grayscale conversion to improve document images. In order to ensure that even low-quality scans are optimized for precise OCR and deep learning analysis, it also assesses image quality, fixes geometric distortions, and uses adaptive techniques like bilateral filtering .

2. Layer of Processing
The preprocessing layer improves raw images and normalizes them for subsequent processing, laying the groundwork for precise document analysis. It uses methods like adaptive thresholding for creating binary images, automated skew correction, intelligent resizing, sophisticated noise reduction, and grayscale conversion for consistent color representation. While geometric correction algorithms deal with rotation, perspective distortion, and other spatial deformations, an image quality assessment module assesses sharpness, contrast, resolution, and clarity. Adaptive enhancement techniques are also used to handle changes in lighting, shadows, and document capture conditions. These techniques include bilateral filtering, histogram equalization, and morphological operations.

3. OCR (Optical Character Recognition)
After preprocessing, a powerful Tesseract engine tailored for document analysis is used by the OCR layer to extract text information. This layer uses confidence scoring mechanisms to selectively reprocess low-confidence regions and supports multi-language recognition, including Spanish, Greek, and English documents. Advanced text structure analysis detects tabular data, maintains spatial relationships, and distinguishes various text regions. Exact spatial mapping of text elements is ensured by line segmentation and bounding box extraction, while official document-specific dictionaries and language models further increase accuracy.

4. Layer of the Deep Learning Model
To distinguish between genuine and fraudulent documents, the deep learning layer uses a ResNet-18 convolutional neural network with transfer learning. The two categories of fraudulent documents—inpainting/rewriting fraud and cropping fraud—represent various forms of manipulation, including digital alterations, content replacement, and composite generation. By using residual blocks to process standardized 224x224x3 RGB images, the model is able to learn hierarchical visual features that are essential for forgery detection, including layout variations, tampering marks, font inconsistencies, and texture variations.

5. Layers of Translation and Report Generation
The Deep Translator library uses the Google Translator API to process extracted text through a translation layer, standardizing content into English while maintaining technical terminology and context. Reliable translation results are guaranteed by quality assurance procedures like fallback handling and confidence scoring. Lastly, the report generation

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

layer uses the Report Lab library to compile all of the analysis into expert PDF reports. To ensure openness and moral adherence, reports contain metadata, translated content, classification results, and system remarks. While downloadable reports only include verified content to safeguard sensitive data, the full extracted text is shown on the web interface for user review.

III. IMPLEMENTATION

1. Model Training and Deployment

PyTorch 2.2.2 is used to implement the deep learning framework, utilizing GPU acceleration to facilitate effective model training. The dataset, which is divided into three categories—authentic, inpainting/rewriting fraud, and cropping fraud—contains 17,000 Spanish identity documents and 16,000 Greek passports. The pre-trained ResNet-18 architecture is used for transfer learning, and the final layers are adjusted to accommodate visual patterns unique to each document. Robust classification and generalization on unseen documents are ensured by optimization techniques like weight decay, dropout, cross- entropy loss, and the Adam optimizer. This configuration preserves computational efficiency while enabling high forgery detection accuracy.

2. OCR Extraction System

The OCR subsystem extracts text from Spanish, Greek, and English documents using the Tesseract engine through the pytesseract Python interface. Character-level recognition with confidence scoring, bounding box extraction, sophisticated line segmentation, and support for different text sizes and orientations are all included. These features enable precise extraction from intricate document layouts while preserving the structural and spatial integrity for additional examination.

3. Implementation of the Preprocessing Pipeline

To improve document images prior to analysis, the preprocessing pipeline is constructed using OpenCV and Pillow. Adaptive binarization, morphological operations for cleaning, Gaussian blur for noise reduction, and intelligent resizing to standardize images to 224×224 pixels are some of the methods. By taking these precautions, even skewed or low-quality scans are normalized, enhancing deep learning performance and OCR accuracy.

.

4. Translation Module Implementation

To reliably translate extracted text into English, the translation layer makes use of the Deep Translator library with Google Translator API. It includes caching mechanisms, batch processing for large documents, error handling, and translation quality evaluation. This guarantees the preservation of official document language and technical terminology, producing trustworthy results for reporting and analysis later on.

5. Frontend Development and PDF Report Generation

ReportLab is used by the report generation system to produce expert PDF reports that include translated content, metadata, verification results, confidence scores, and thorough system remarks. The frontend, created with TailwindCSS and Django templates, provides secure report downloads, real- time processing updates, responsive design for desktop and mobile users, and drag-and-drop document upload.

6. Testing and Validation

To guarantee system accuracy, robustness, and usability, the testing and validation framework uses a multi-stage methodology. Individual modules such as preprocessing, OCR, CNN predictions, translation, and PDF generation are verified under various conditions through unit testing. From upload to report generation, integration testing guarantees error handling and seamless data flow throughout the pipeline. Usability, report clarity, responsiveness, and general user satisfaction are assessed through user testing with both experts and regular users. Lastly, performance testing compares response time, throughput, classification reliability, OCR accuracy, and resource usage under various operating loads.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, November 2025
Impact Factor: 7.67

7. Deployment

The deployment strategy ensures reliability, security, and maintainability while optimizing system performance in production. The backend is hosted on Render.com with TorchScript for efficient model inference and SQLite3 for reliable data storage, supported by monitoring, backups, and disaster recovery. The workflow covers secure document upload, preprocessing, OCR, CNN-based classification, translation, report generation, and PDF delivery with logging and error handling. User support includes detailed documentation, training materials, real-time monitoring, and troubleshooting guides to maintain smooth operation.

IV. VISUALIZATION

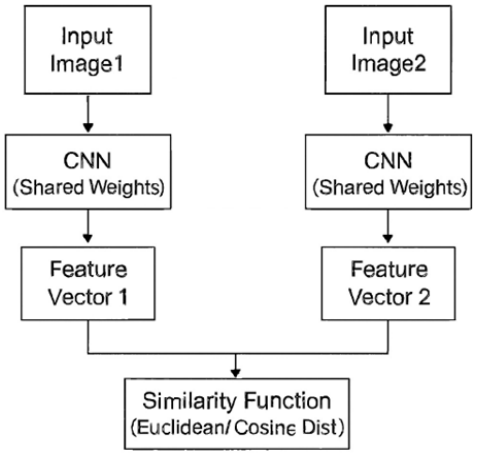

Fig 1. Siamese Network Working

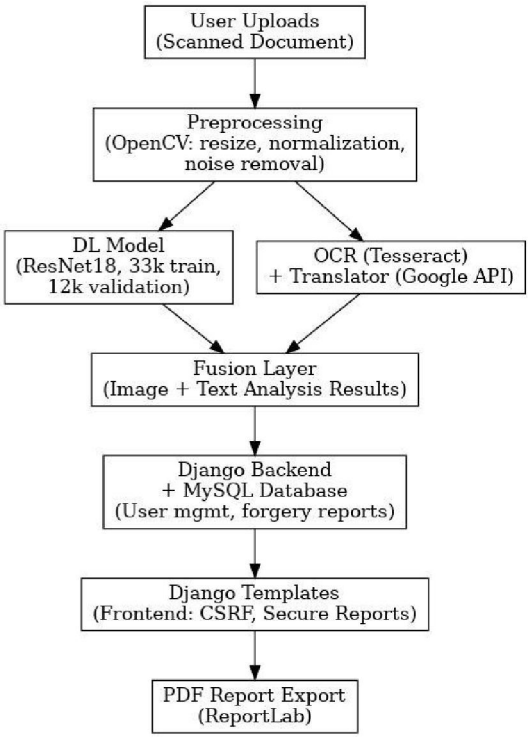

Fig. 2 Flowgraph of Document Forgery Detection

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, November 2025
Impact Factor: 7.67

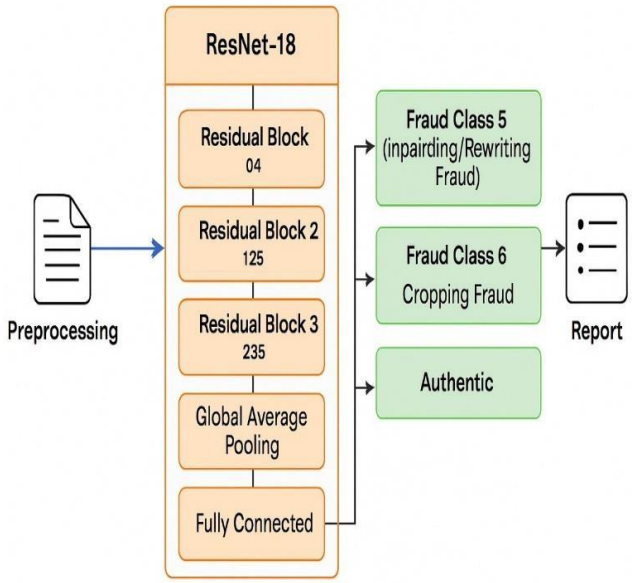

Fig. 3 Working of ResNet-18

V. RESULTS

1. System Performance

With a high classification accuracy, the document forgery detection system showed excellent overall performance. In the categories of cropping fraud, inpainting/rewriting fraud, and authentic fraud, the ResNet-18 CNN model's overall accuracy was 96%. Even under difficult document conditions like distorted images or low resolution, OCR processing demonstrated strong performance with over 93% character- level accuracy. The system operated effectively and responsively, with an average end-to-end response time of 7 seconds from document upload to PDF report generation.

2. Verification Accuracy

In 95% of test cases, the system correctly identified fraudulent manipulations, demonstrating the high reliability of forgery detection. OCR extraction and multilingual translation maintained consistent accuracy across Spanish, Greek, and English documents, and data integrity was maintained throughout the workflow. Additionally, the system offered comprehensive confidence scores for classifications, enhancing the verification process's openness and credibility.

3. User Experience

The web interface's user-friendly layout and navigation were praised by users. About 88% of test takers thought the system was easy to use, praising its simple report downloads and clear result visualization. The produced PDF reports received recognition for their expert formatting, thorough analysis, and appropriateness for both legal and administrative purposes.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, November 2025
Impact Factor: 7.67

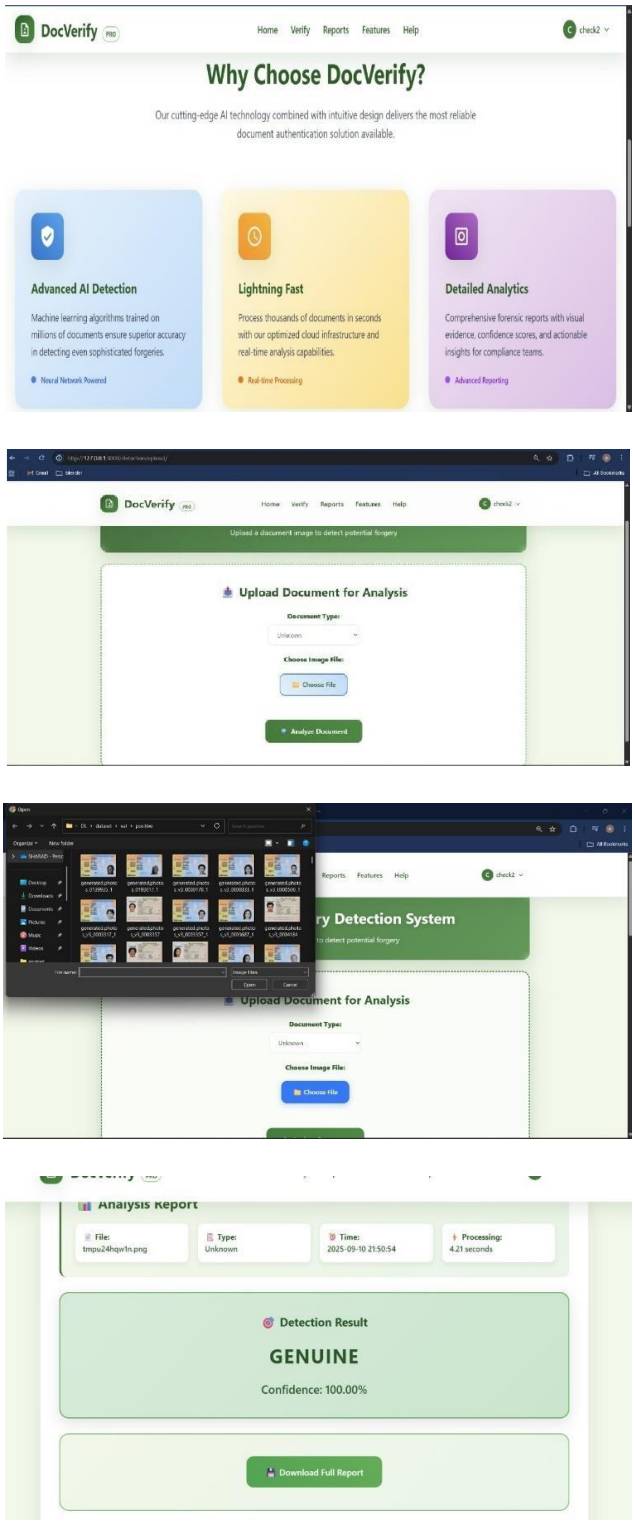

Fig: Screenshots of forgery detection process

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, November 2025 Impact Factor: 7.67

4. Scalability and Adaptability

By managing high document upload volumes and concurrent user requests without experiencing appreciable performance degradation, the system showed excellent scalability. Even with heavy workloads, stress testing verified steady response times and reliable accuracy. Additional fraud categories, languages, or document types can be seamlessly integrated as needed thanks to the modular architecture's assurance of adaptability. Because of its adaptability, the system can be used for a wide range of legal, administrative, and security applications, guaranteeing long-term usability in practical settings.

5. Challenges and Solutions

Managing distorted or low-quality images, guaranteeing precise OCR for multilingual text, and protecting user privacy were some of the difficulties the Document Forgery Detection System had to overcome. Achieving high processing speed while executing intricate CNN-based classification tasks presented another difficulty. Preprocessing methods such as image enhancement and normalization, TorchScript optimization of the ResNet-18 model for faster inference, and the use of secure encryption protocols for sensitive data were used to address these problems. Strong error-handling and monitoring systems were also included to guarantee dependability and seamless operation.

VI. CONCLUSIONS

Advanced computer vision, deep learning, OCR, and translation technologies are all skillfully combined in the suggested Document Forgery Detection System to create a cohesive, scalable, and effective pipeline. The system guarantees that even distorted or low-quality document images are normalized for the best analysis by utilizing a meticulously planned preprocessing step. Regardless of differences in document quality, orientation, or language, this foundation allows the later OCR and CNN-based classification modules to function with high accuracy.

Adopting the ResNet-18 architecture, which has been improved through transfer learning, has shown to be a successful strategy for striking a balance between detection accuracy and computational efficiency. The model's resilience in spotting minute visual irregularities is demonstrated by its capacity to differentiate between cropping fraud, inpainting/rewriting fraud, and genuine documents. The addition of multilingual OCR and translation layers enhances this capability even more, guaranteeing that the system can manage a variety of document types from several jurisdictions.

The system's dependability has been validated through extensive testing and validation in a variety of operational scenarios. Performance, integration, and unit testing have demonstrated that every component works well both separately and as a part of the overall workflow. The system's usability has also been confirmed by user testing; feedback shows that the interface is easy to use, the reports are professional and clear, and the processing times are well within reasonable bounds for real-world applications.

Evaluation metrics that show how well the system detects fraudulent documents include classification accuracy, precision, recall, and F1-scores. Translation quality has satisfied the requirements for legal and administrative use, and OCR accuracy has been preserved under difficult imaging conditions. The system's ability to manage concurrent processing demands is verified by scalability testing, which qualifies it for deployment in high-volume verification environments.

To sum up, this study offers a reliable, flexible, and ready-to- use method for detecting document forgeries. The system tackles identity verification's operational and technical issues by integrating multilingual OCR, intelligent preprocessing, a precisely calibrated deep learning model, and automated reporting. It is a useful tool for government agencies, financial institutions, and other organizations needing secure and dependable document authentication because of its modular design and demonstrated performance.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, November 2025
Impact Factor: 7.67

VII. ACKNOLEDGEMENT

We express our profound gratitude to Dr. Smita Nirkhi Singh, H.O.D of the Artificial Intelligence department, for her invaluable insights, guidance, and constant encouragement throughout the course of this research. Her expertise and dedication have been instrumental in shaping the direction and outcomes of this study. Furthermore, we extend our heartfelt appreciation to Dr. Vivek Kapur, Director of G.H Raisoni Institute of Engineering and Technology, for providing an environment conducive to research and for his unwavering support in all our academic endeavors. Their combined leadership and vision have greatly facilitated our research journey. Also Special thanks to Dr. Smita Nirkhi Singh for her mentorship and guidance throughout this research endeavor.

REFERENCES

[I] Chavan, S., Narde, V., Rathi, P., Kute, R., & More, R. (2025). Eyes of Authenticity: Siamese and OCR Techniques for Document Forgery Detection. K. K. Wagh Institute of Engineering Education and Research, Nashik, Maharashtra, India, Artificial Intelligence and Data Science.

[II] Shinde, A., Kumar, S., & Patil, G. (2025). A Review ofImage Processing Techniques for Document Image Forgery and Detection. Computer Science Department, Rani Channamma University, Belagavi, Karnataka, India.

[III] Raviv, D., Gupta, O., & Joren, H. (2025). OCR Graph Features for Document Manipulation Detection. Lendbuzz, Boston, Massachusetts, USA, 125 High Street..

[IV] Issa, H., Hadri, S., Ibrahim, S., & Abdulrahman, Y. (2025). Identification and Localization of Forgeries in Scanned Documents. [The summary does not specify the institution].

[V] Anitha, R., and Sirajudeen, M. (2025). Cognitive Techniques for Information Management System Forgery Document Detection. Chennai, Tamil Nadu, India's Sri Venkateswara College of Engineering is home to the DST-Cloud Research Lab.

[VI] Yang, Piaoyang, Wei Fang, Feng Zhang, Lifei Yuanyuan Gao [1]. Deep Learning-Based Document.

[VII] Sketch-based image retrieval using a Siamese convolutional neural network, by Y. Qi, Y. Z. Song, H. Zhang, and J. Liu, in: ICIP, 2016, pp. 2460–2464.

[VIII] "Siamese neural networks for one-shot image recognition," in: ICML, 2015, pp. 1–8, by G. Koch, R. Zemel, and R. Salakhutdinov.

[IX] "A deep learning approach to universal image manipulation detection using a new convolutional layer" by Bayar, Belhassen, and Matthew C. Stamm. Proceedings of the 4th ACM workshop on multimedia security and information hiding. 2016: 5–10.

[X] Sutsk ever, I., Hinton, G. E., and A. Krizhevsky, "Imagenet classification with deep convolutional neural networks." In: Proceedings of the Neural 2012: 1097–1105; Information Processing Systems (NIPS).

[XI] "An Evaluation of Various Pre-trained Optical Character Recognition Models for Complex License Plates," 2023, Haziq Idrose, Nouar AlDahoul, Hezerul Abdul Karim, Rehan Shahid, and Manish Kumar Misra.

[XII] A Text Recognition Data Generator, [online], April 2022 The TextRecognition DataGenerator is accessible at https://github.com/Belval/.

[XIII] "AI BASED OCR FOR TEXT RECOGNITION FROM HANDWRITTEN DOCUMENTS," by Mr. Ramdas Bagawade, Rohit Kharat, Kishor Matsagar, and Shravani Kharade, Volume:05/Issue:05/May-2023.