International Journal of Communication and Information Technology

E-ISSN: 2707-6628
P-ISSN: 2707-661X
Impact Factor (RJIF): 5.56
www.computersciencejournals.com/ijcit
IJCIT 2025; 6(2): 164-171
Received: 20-10-2025
Accepted: 22-11-2025

Nadia Ibrahim Nife
College of Computer Science and Information Technology, University of Kirkuk, Kirkuk, Iraq

Hoger K. Omar
College of Computer Science and Information Technology, University of Kirkuk, Kirkuk, Iraq

Hero O. Ahmad
College of Computer Science and Information Technology, University of Kirkuk, Kirkuk, Iraq

Corresponding Author:
Nadia Ibrahim Nife
College of Computer Science and Information Technology, University of Kirkuk, Kirkuk, Iraq

AI-driven document forgery detection and similarity analysis using OCR and Levenshtein techniques

Nadia Ibrahim Nife, Hoger K Omar and Hero O Ahmad

DOI: https://doi.org/10.33545/2707661X.2025.v6.i2b.158

Abstract
As the world moves to digital and paper documents in day-to-day, fake document detection tools are increasingly required. This work proposes a novel system for document forgery detection by using image processing technology with artificial intelligence and optical character recognition (OCR) techniques to extract and analyze text. It works by using the Tesseract OCR library to read text from digital documents or images and compares the similarity by a method called Levenshtein Distance. Levenshtein calculates the distance between texts to identify potential changes. Diff Analysis is also used to identify textual differences and presents them in a visual format that allows clearer judgment of the extent of change between documents. The system provides an accurate method to detect distorted content. Experimental results showed that the system can correctly identify even small alterations and forgeries inside documents. The goal of this system is to supply a reliable tool that detects content manipulation and preserves the integrity of information and digital records in fields such as legal auditing and governance in any form of digital documentation.

Keywords: Levenshtein Distance, Tesseract OCR, Data Extraction, Diff Analysis, Document Forgery detection

1. Introduction
The growth of digital information became a major concern for organizations and individuals. Document forgery stood out as a serious issue because some altered texts were very subtle. Traditional methods failed to detect many of these cases. The use of digital documents in legal and administrative work also rose [1]. The need for intelligent systems that can find illegal alterations and verify document authenticity is growing rapidly [2]. Much work relies on Optical Character Recognition (OCR) techniques to extract texts from the documents [3]. Where Tesseract OCR is used as an effective tool to convert image texts into processable data [4]. After extracting the texts, the Levenshtein Distance algorithm is applied to measure the similarity between different texts and identify the differences between the original and modified versions [5]. Diff Analysis is also used to accurately identify text alterations, allowing for a deeper understanding of any changes made to the document [6].
In addition, there are many systems that work to improve document auditing and verification processes in multiple fields, such as law, administration, and cybersecurity, as content tampering can be easily and effectively detected [7]. Due to the lack of an effective anti-forging mechanism, academic certificates are vulnerable to fraud, forgery, and imitation [8]. It also enhances the genuineness of the digital document. Thus reducing the chances of forgery. Therefore, it becomes a vital scheme for every party, depending on the precision and quality of information they handle frequently. In this work, we explain the functionality of the system, outline detection techniques for forgeries and discuss its performance (accuracy and efficiency) for the comparison text and the text modification detector [9]. With the increasing reliance on digital documents in official transactions [10], attempts to forge documents have increased [11]. Therefore, there is a need to develop systems capable of verifying the authenticity of documents effectively and quickly [12]. Texts are extracted from images and digital documents using OCR (Optical Character Recognition) technology [13]. Then compared the original texts to measure similarity and hypothesize differences. In contrast, this research provides a technique, which is applicable by text processing and similarity analysis methods, for detecting forged documents. Yet there have been some significant scientific and technical advances to do with forgery detection in text documents, specifically:

- Designing a unified intelligent architecture incorporating OCR technology and text similarity analysis algorithms for discrimination of forgery in digital documents. The influence of the quality of inputs (images or PDF files) on the performance of the system is also analyzed, and it is shown that an acceptable level of precision can still be achieved for degraded document quality.
- Handling various forgery scenarios and simple alterations (such as changing dates or names) and complex alterations (such as deleting paragraphs or replacing official entities), with a detection accuracy of up to 98%.
- Visualizing performance metrics such as accuracy and inference time. However, building the qualitative analysis of textual differences even when texts are very similar to each other. This creates a flexible and scalable system for artificial intelligence and multilingual technologies.
- This article is structured as follows: in section two, the literature Review for this article is provided. Section three describes the system tools, section four describes the Methodology. Section five shows the results, and finally, the conclusion is presented in section six.

2. Literature Review

R. Sukhija *et al*. 2025 [14] reviewed key strategies used in document forensics to determine the source printers of digital documents and to identify authors of scanned handwritten material. The study surveyed a wide range of machine learning and deep learning techniques applied in this domain and examined approaches for detecting printer types and models through textual pattern analysis and noise-based features. The work also outlined major advances in classification and feature extraction methods, offering a clear overview of current state-of-the-art tools used in modern forensic document examination. S. Carta *et al*. 2024 [15] proposed a method that builds on recent development advances in document understanding for the task of identity document recognition, which is the automatic interpretation and extraction of fields of ID cards. Upon employing a two-stage design, our method first fine-tunes a pre-trained model with synthetic datasets and then a real dataset to make it more robust over different types of documents. The authors used the terms realized on a dedicated synthetic data creation tool created specifically to enhance and reinforce the training. Extensive experiments demonstrate that the proposed method achieves competitive performance, which proves the effectiveness of the proposed strategy for the state-of-the-art ID recognition applications. M. Abdulqader *et al*. 2023 [16] investigated a worrisome trend of increasingly vascular and damaging threats to the security of digital images, particularly in domains that are heavily reliant on photographic evidence, including forensic research and public image forensics. In their analysis, they reviewed how digital imaging has provided new avenues for forensic inquiry, but at the same time, widely available editing tools have made image manipulation easier than ever. The authors presented an overview of typical forgery types, focusing on copy-move and splicing, and how such manipulations undermine the validity of images as evidence. It also recorded the tools and approaches that have been used previously for detection, hence representing a summarized overview of present methodologies in digital image forensics. Md. Abdur Rahman *et al*. 2023 [17] proposed FOIL: a new unsupervised ensemble based framework to automate information extraction and verification from documents. It combines novel language and image processing techniques while utilizing a majority voting scheme over four state-of-the-art OCR engines to maximize recognition accuracy. A transformer model is used to aid contextual understanding and improve the structural format for text extraction with a confidence-based post-process mechanism. A high performance accuracy of 95% + was achieved while testing for the extraction of name, ID, and marks from exam scripts datasets. It also reported very strong precision, recall and F1-score results on the ICDAR2013 benchmark over a range of additional experiments that continue to establish the effectiveness of the framework. A. Amjed *et al*. 2022 [18] proposed an overview of the most prominent streams of research in digital forensics, with 3 main threads: investigation of image-based forensic analysis, video-based conformity, and detection based on the bread and butter of spectroscopy methods. In their review, the authors discussed recent strategies for the detection of forged and counterfeit documents and underscored the role of these strategies in supporting future research for authenticating these documents. It also provides an organized overview of the state of the art in scientific analysis trends and a well-structured guide for the new generation of researchers in the digital forensics area. M. Rajapashe *et al*. 2020 [19] presented a system for document authentication by the relationship of different formats and digital signatures with blockchain to make it more trusted and secure. In their approach, the content of documents is first extracted and an electronic signature is generated using Optical Character Recognition (OCR) methods and embedded into the document in the form of a 2D barcode. It enables extraction and signature validation through an OCR based comparison using a barcode embedded in the document. As an additional step, the authors layered forgery detection methods on top of the existing systems. Another value of this is that by having all signed documents in a blockchain backed decentralized infrastructure, it becomes more resilient to malicious changes by writing copies more immune to them. S. Abramova *et al*. 2016 [20] investigated the task of finding and localizing the copy-move forgeries in scanned text documents, where repeated and visually similar glyphs can make analysis challenging. In their research, they assessed multiple block-based feature representations often used for image forensics to find out how effective they are when it comes to identifying duplicated areas and when it comes to minimizing false detections in scanned media. Extensive explorations on a benchmark dataset of scanned documents found that block-based techniques only yield marginal performance gains unless carefully calibrated with individual thresholds and parameters. They additionally discussed candidates for adapting these methods to the properties of textual document images and proposed ideas towards the copy-move forgery detection in textual documents.

3. Tools and techniques for detecting forgery

Specialized tools and analytical techniques are required to detect forgery in digital documents, as even the most minute changes will be evident, and these changes need to be detected. Recent developments in image processing and machine learning [21], as well as document forensics, have resulted in increased accuracy of manipulation and inconsistency detection. Recent approaches integrate state of the art OCR technologies and feature extraction methods

with statistical models that ascertain the authenticity of documents. A brief Overview of tools and techniques to detect forgery in various document types has been outlined in this section.

3.1 Tesseract OCR
The Tesseract OCR is used for image-based documents. Tesseract OCR is one of the most popular and accurate optical character recognition (OCR) engines [22]. It supports many languages, including Arabic [23]. Supporting more than 100 languages and widely used in text digitization applications [24]. In a recent study published in 2024, many researchers developed a combined approach that combines machine learning techniques and the Tesseract OCR engine to enhance the accuracy of text recognition in these documents [25]. In a recent study, researchers analyzed the performance of five OCR libraries across multiple languages such as English, Hindi, Arabic, Tamil, and Malayalam [26].

3.2 Levenshtein Distance algorithm
Levenshtein Distance is an algorithm that indicates how similar two texts are and how dissimilar they are; it measures the edit distance (i.e., the smallest number of operations needed) needed to convert one text to another (addition, removal, or substitution) [27, 28]. Using the Levenshtein Distance algorithm to evaluate how similar the text that OCR extracts from the original text is significantly helped increase the percentage of correctly spelled words [29]. For text recognition in natural scenes [30], LevOCR outperforms the traditional methods. Implementation of this algorithm in spelling error correction systems results in less amount of misprints and becomes more accurate for text matching [31].

3.3 Diff analysis
Difflib is particularly advantageous for the detection of minor changes, including changing dates, amounts of money, distorting legal clauses, and the like [32]. Due to the effectiveness of spelling correction systems in analyzing textual differences, this algorithm in spelling correction systems reduces the error rate and increases the accuracy of matching between texts [33]. The algorithm is very good in discriminating the original text from the extracted one, thus it adds more robustness to the reliability of the system in identity confirmation and image credibility processes [34]. We use pytesseract, pdfplumber, and difflib which are key libraries for our work, e.g., pytesseract for image-based text extraction [35], pdfplumber for direct PDF text extraction [36], and difflib for highlighting the difference in the text [37].

4. Methodology
The proposed methodology consists of three stages in document mismatch detection and forgery detection, text extraction, and Similarity identification. Optical character recognition (OCR) is a technology that recognizes text that is embedded within scanned images or documents, and the quality of these documents or images will affect the accuracy of the OCR. Sometimes texts written in multiple languages or with unsupported fonts make it hard for the system to recognize. Blurred background or blurry text can also relate to the low accuracy of the Extractor. When summarized, those stages can be expressed in the following subsections:

4.1 Text Extraction
A Python library (PDFPlumber): This is a Python library that extracts text accurately from PDF files, whether the text is located within tables or laid out in multiple columns. After extraction, the detected text is cleaned and formatted. It's a powerful supplement to Tesseract OCR for working with digital PDFs that don't need OCR. With PDFPlumber, the structural layout of each page is also preserved and thus the logical reading also preserved when extracting content. This prevents sentences from being truncated or tabular data from becoming misaligned. This makes downstream processing much easier as the text should be cleaner and well organised. This sets the stage for subsequent similarity analyses and comparison tasks. It offers the opportunity of solidifying the fidelity of the whole forgery detection pipeline.

4.2 Text Comparison
The text is extracted from documents (this could be based on Tesseract OCR or from PDF based extraction tools like pdf-miner) then the next thing to do is analyze the text for possible signs of modification or forgery. The Levenshtein Distance is one of the classical methods applied in this analysis to determine the distance between two texts. In this approach, it counts how many insertions, deletions or substitutions are needed for transforming your extract version to a reference version. However, this allows it to detect even the smallest adjustments that may point to a forgery attempt. Naturally, this distance, as it goes up, means there is a greater chance of tampering and provides a measure we can turn to tell if a document has been modified or not. The output from the algorithm is a similarity score that gives us a score, and if it is less than 1.00 it is a potential forgery and must needs to be inspected further. This comparison step adds a validation layer as it provides an accurate numeric reading of how reliable a piece of text is.

4.3 Difference Visualization
The difflib helps to mark the changes between documents in the Text (Added words, removed words, or modified words) which we can extract easily in the Form of Possible changes of a Document. The differences are represented visually which assists a reviewer like a school teacher to decide whether an edit is suspicious or not. This utility determines modified portions and assists in verifying the document and offers a clear comparison between documents. We decided to go with Python as the main development environment since it provides numerous libraries for OCR, PDF processing, text comparison, and by using Python we can easily access MultiProcessing and Threading paradigms for parallel processing too. Its ecosystem allows extraction, analysis, and visualization components to work together easily. Fig. 1 shows the overall workflow that consists of an integrated process for image processing, text extraction, content analysis and similarity comparison. Through this multi-staged design, visual and text evidence can be thoroughly examined. In conclusion, by integrating difflib and using a pipeline based on Python, we are making the forgery detection system more accurate and reliable.

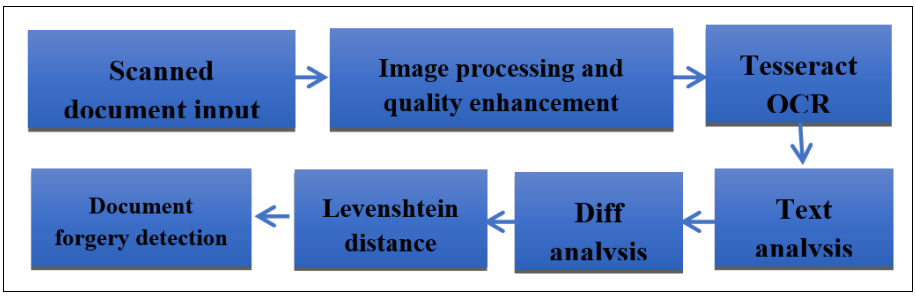

Fig 1: The proposed model Steps

5. Results

The model was used to implement forgery detection in scholarly documents based on earlier reads through the combination of optical character recognition (OCR) methods with using the Levenshtein Distance Algorithm for text-similarity measurement. This technique allowed the system to distinguish between the original and manipulated versions of the document, which led to accurate and consistent tampering identification. Two PDF documents are tested by the system: a PDF document (1) provided by the University of Kirkuk with an official title. Forged document, an exact copy of the original document, but the header has been tampered with, and the university name has been replaced with Tikrit University. Fig. 2 and 3 represent the first page of each document and clearly show that all elements except the university name match.


University of Kirkuk
College of Computer Science and
Information Technology
Software Department

Subject: Microprocessors and Computer Architecture
Time: 3 Hours
Class: 2nd Year
Date: 2/1/2024
Score: 50 Marks

Final Examination, First Attempt, First Term, Academic Year (2024-2025)

الاجابة تكون بقلم الرصاص

NOTE:

Answer only Five questions; Q5 and Q6 must be included.

Each question carries 10 Marks.

Q1/ Explain the main purposes of following: (10marks)

Fig 2: Original document


University of Tikrit
College of Computer Science and
Information Technology
Software Department

Subject: Microprocessors and Computer Architecture
Time: 3 Hours
Class: 2nd Year
Date: 2/1/2024
Score: 50 Marks

Final Examination, First Attempt, First Term, Academic Year (2024-2025)

الاجابة تكون بقلم الرصاص

NOTE:

Answer only Five questions; Q5 and Q6 must be included.

Each question carries 10 Marks.

Q1/ Explain the main purposes of following: (10marks)

Fig3: Forged document

Texts from both documents were converted into an analyzable format using the pdfplumber library. Then cleaned and formatted before content comparison. Comparison using the Levenshtein Distance algorithm showed that the similarity score between the two documents reached 100% (1.00). Despite this complete numerical match, the diff analysis revealed only one change: the replacement of the university name in the page header from: University of Kirkuk → University of Tikrit. No additional differences were detected, confirming the high accuracy of the falsification to achieve the target through limited selection. The effectiveness of the model in detecting documents in Fig. 2 and 3 was tested in different types of forgery, and the detection rate of each type was evaluated. Fig. 4 illustrates the minuscule alterations found between the two documents (to be original and forged). The only thing manipulated was the name of the university.

Removed (From Original)	Added (In Forged)
Kirkuk	Tikrit
↓ Word removed from original document	↑ Word added to forged document

Fig 4: Removed vs Added (detailed)

The robustness of the proposed forgery detection system against the document quality is tested by comparing the original scanned exam paper, at three different image resolutions that emulate the normal scanning conditions, as high quality (300dpi), medium quality (150 dpi), and low quality (72 dpi) in Table 1.

Table 1: OCR performance results

Image Quality	Observations	Word Count Extracted via OCR	Image Quality
High (300 dpi)	Excellent clarity and near-perfect extraction	236 words	High (300 dpi)
Medium (150 dpi)	Minor degradation, acceptable for detection	228 words	Medium (150 dpi)
Low (72 dpi)	Noticeable loss in text fidelity, but still analyzable	190 words	Low (72 dpi)

Fig. 5 conducted a comparison between the number of words extracted through OCR from original and forgery files of various resolutions. A better resolution (300 dpi for instance) permits more accurate text recognition and an increased amount of words extracted. For low-quality images, we lose more words. This study was conducted in order to assess how successful the system was in detecting different types of forgery and four simulated forgery cases based on the original case were simulated in order to assess this. Each forged form corresponds to a common form of manipulation that we have studied in cases of academic and administrative fraud. Detection performance of spoofing types in terms of forgery type is shown in Table 2. The similarity scores approaching 1 show only small changes, but the diff analysis was enough for the system to flag the suspicious ones.

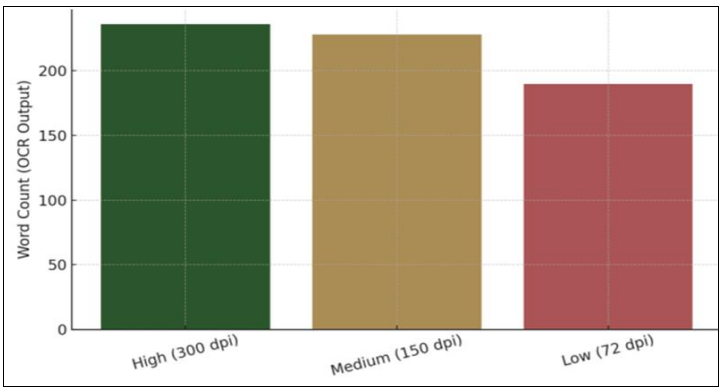

Fig 5: Impact of image quality on OCR text extraction

Table 2: Forgery Scenarios and Detection Results

Forgery Type	Description	Similarity Score	Words Removed	Words Added
Forged Signature	Replaced examiner and department head names	0.981	2	2
Forged Date	Altered the date from 2/1/2024 to 2/1/2023	0.994	1	1
Forged University	Changed “University of Kirkuk” to “University of Basra”	0.987	1	1
Deleted Question (Q4)	Removed the fourth question entirely	0.956	~38	0

The results of the experiments demonstrated the system's high efficiency in detecting various types of forgery, as follows:

- The average similarity score between the original and forged copies ranged from 0.956 to 0.994, where:

a) The date change (from 2/1/2024 to 2/1/2023) was detected with an accuracy of 0.994.
b) The university name fraud was detected with an accuracy of 0.987.
c) Signature change detected with (0.981) accuracy.
d) Deleting an entire question (question 4) reduced the similarity to 0.956 it indicates the system's ability to distinguish between minor and radical modifications.

- When testing the impact of image quality on system performance, the system demonstrated consistent performance:

a) At 300 dpi, 236 words were extracted.
b) At 150 dpi, 228 words were extracted.
c) Even at a low resolution of 72 dpi, 190 words were extracted with acceptable accuracy, demonstrating the system's tolerance for visual noise and poor quality.

Fig. 6 shows similarity scores for each forgery type. The system proves high sensitivity especially for significant alterations such as deleted sections.

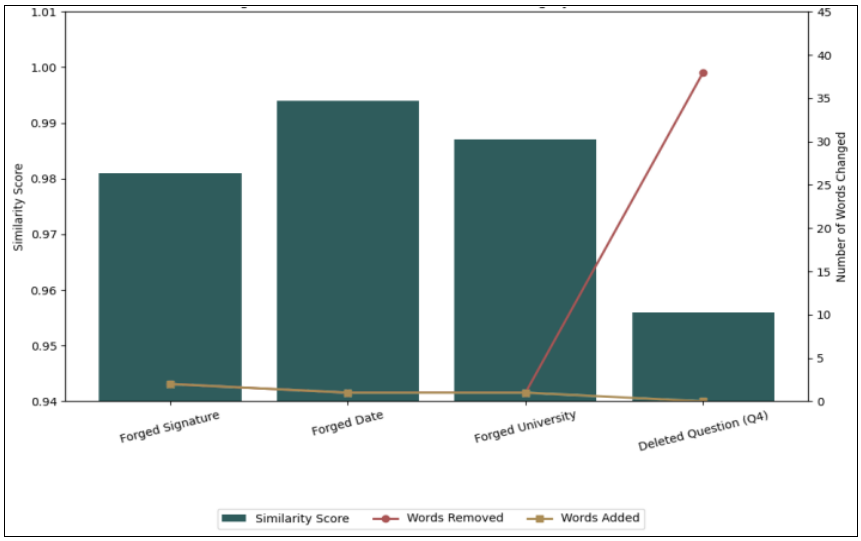

Fig 6: Forgery Detection Scores

After using the Levenshtein algorithm with fine grained variance analysis the results indicated that the method achieved an accuracy of 98% in detecting text forgery. The accuracy in text extraction for high-quality documents was 95-98% however, for low-quality documents it reduced to 80-85% which was partially compensated by the use of a preprocessing technique. Time Efficiency: Per-document processing time was, on average, between 0.5 and 3 seconds. Detection of changes (large = 90% accuracy; small = 75-80% accuracy) was generally more effective in the system. These results showed that the system was able to still reliably flag even the more subtle forgeries in documents that would remain seemingly consistent in text by appearing to only change a part of the text like the university name or date. The results show that the proposed model effectively detects different types of forgery with significant flexibility to handle the visual challenges present in signature images, and has great potential for numerous applications in the academic, administrative, and financial fields.

The fraud detection system efficiency at the three quality levels of document, high, medium, and low is shown in Fig. 7. The next is a chart of each of the metrics of the forgery detection system, where the accuracy of the system in all cases, detection of simple and complex forgeries, and average processing time for all cases are clearly and separately illustrated as follows. The bar charts show the system's overall accuracy in detecting forgeries (green) and how it performed on simple (orange) and complex (red) edits. It can be witnessed in the data as well, accuracy is decreased massively on low-quality documents, with a high accuracy of 98% for high-quality documents, and a low accuracy of 85% for low-quality documents cumulatively. The blue line graph, which utilizes the secondary Y-axis, illustrates average processing duration per document. Naturally, the slower the processing time for small-scale images, escalating from 0.5 seconds with high-quality inputs to 3.0 seconds with low-quality documents. The reverse correlation is likely due to the added processing required by the OCR engine and analysis algorithms to extract text from noisier or degraded histograms. As we can see from the columns, as quality decreases, accuracy also decreases. Overall, quality documents yield the best combinations of accuracy while detecting both trivial and harder to detect changes. Regarding quality, the detection period lengthens because more operation was required to detect if changes were made. The quality of the documents has a strong consequence on the accuracy of detection.

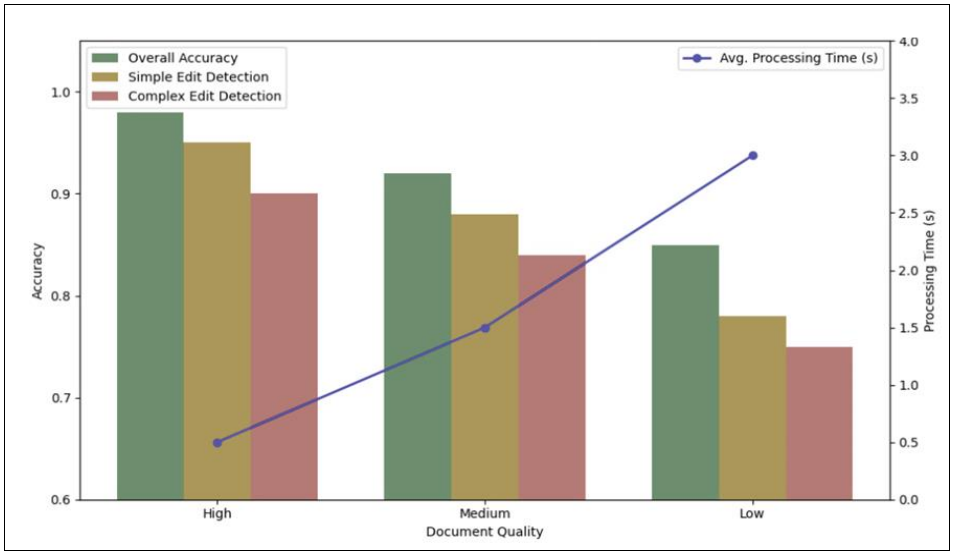

Fig 7: The performance of the forgery detection system between different document quality levels

6. Conclusion

This study proves a robust system for detecting forged text documents by integrating optical character recognition with semantic text similarity analysis. The experimental results show that the model achieves high reliability. It describes the authentic documents from the manipulated ones. Overall detection accuracy has reached 98 percent. while blind testing discovered strong performance even in complex alterations with an accuracy of about 90 percent and simpler forgeries with an accuracy of 75 to 80 percent. The OCR component also exhibited stable behavior by producing text extraction accuracy between 95 and 98 percent for high quality documents and between 80 and 85 percent for documents of lower visual quality. The time evaluation shows that the system operates efficiently with an average processing time with scoring half a second to three seconds per document. Even instances where the modified text appeared close to the original text (e.g., only the name of the institution changed a slightly different signature was added). The technique was able to detect manipulation using fine grained textual differences. This feature is especially important to detect indirect or well hidden attempts at forgery. This work might be extended for future development to support multi lingual and handwritten signatures. Moreover, using deep learning technologies to support semantic text analysis and pattern recognition.

References

1. Abinesh G, Kavitha V, JV P. Signature verification using deep learning and CNN. International Journal of Innovative Science and Research Technology. 2025;10(3):374-381.
2. Moolchandani J, Pakshwa R, Singh K. Machine learning for identifying and validating document authenticity. In: Proceedings of the 15th International Conference on Computing Communication and Networking Technologies (ICCCNT); 2024 Jun; IEEE. p. 1-9.
3. Sirajudeen M, Anitha R. Forgery document detection in information management system using cognitive techniques. Journal of Intelligent & Fuzzy Systems. 2020;39(6):8057-8068.
4. Okamoto Y, Genki O, Yahiro I, Hasegawa R, Zhu P, Kataoka H. Image generation and learning strategy for deep document forgery detection. arXiv preprint arXiv:2311.03650. 2023.
5. Slavin O, Farsobina V, Myshev A. Analyzing the content of business documents recognized with a large number of errors using modified Levenshtein distance. In: Cyber-Physical Systems: Intelligent Models and Algorithms. Cham: Springer International Publishing; 2022. p. 267-279.
6. Pun AK, Javed M, Doermann DS. A survey on change detection techniques in document images. arXiv preprint arXiv:2307.07691. 2023.
7. Sani AI, Olajide AO, Abosede OV, Oyebode DF, Vayyala R, Alfred JG, *et al*. Cybersecurity challenges in digitizing government administration. 2025.
8. Kareem AS, Shakir AC. Verification process of academic certificates using blockchain technology. Kirkuk University Journal for Scientific Studies. 2023;18(1).
9. Qu C, Liu C, Liu Y, Chen X, Peng D, Guo F, *et al*. Towards robust tampered text detection in document image: new dataset and new solution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 5937-5946.
10. Degeneve C, Longhi J, Rossy Q. Analysing the digital transformation of the market for fake documents using a computational linguistic approach. Forensic Science International: Synergy. 2022;5:100287.
11. Goc M, Semków D. Public documents act and its role in preventing document forgery. Studia Iuridica Lublinensia. 2020;29(4):4-85.
12. Boonkrong S. Design of an academic document forgery detection system. International Journal of Information Technology. 2024;1-13.
13. Bae YY, Cho DJ, Jung KH. Visual complexity in Korean documents: toward language-specific datasets for deep learning-based forgery detection. Applied Sciences. 2025;15(8):4319.
14. Sukhija R, Kumar M, Jindal MK. Document forgery detection: a comprehensive review. International Journal of Data Science and Analytics. 2025;1-23.
15. Carta S, Giuliani A, Piano L, Tiddia SG. An end-to-end OCR-free solution for identity document information extraction. Procedia Computer Science. 2024;246:453-462.
16. Amjed A, Mahmood B, Almukhtar KA. Approaches for forgery detection of documents in digital forensics: a review. In: International Conference on Emerging Technology Trends in Internet of Things and Computing. Cham: Springer International Publishing; 2021. p. 335-351.
17. Rahman MA, Hasan MT, Howlader UF, Kader MA, Islam MM, Pham PH, *et al*. uFOIL: an unsupervised fusion of image processing and language understanding. IEEE Access. 2025.
18. Abdulqader MF, Dawod AY, Ablahd AZ. Detection of tamper forgery image in security digital image. Measurement: Sensors. 2023;27:100746.
19. Rajapashe M, Adnan M, Dissanayaka A, Guneratne D, Abeywardena K. Multi-format document verification system. American Scientific Research Journal for Engineering, Technology, and Sciences. 2020;74(2):48-60.
20. Abramova S, Böhme R. Detecting copy-move forgeries in scanned text documents. In: Proceedings of the 1st Workshop on Information Security and Privacy; University of Innsbruck; 2016.
21. Nife NI, Chtourou M. Video objects detection using deep convolutional neural networks. In: Proceedings of the 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP); 2022 May; IEEE. p. 1-6.
22. Ramdhani TW, Budi I, Purwandari B. Optical character recognition engines performance comparison in information extraction. International Journal of Advanced Computer Science and Applications. 2021;12(8).
23. Ramteke R, Al Maamari MRAO. Tesseract OCR recognition based on Arabic. In: Proceedings of the First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022). Springer Nature; 2023. p. 347-356.

24. Syam Had I, Baihaqi WM, Kinding DPN. Improving Tesseract OCR accuracy using SymSpell algorithm on passport data. Sinkron: Jurnal dan Penelitian Teknik Informatika. 2025;9(1).
25. Fleischhacker D, Goederle W, Kern R. Improving OCR quality in 19th century historical documents using a combined machine learning based approach. arXiv preprint arXiv:2401.07787. 2024.
26. Nazeem M, Anitha R, Navaneeth S, Rajeev RR. Open-source OCR libraries: a comprehensive study for low resource language. In: Proceedings of the 21st International Conference on Natural Language Processing (ICON); 2024. p. 416-421.
27. Celikten T, Onan A. Exploring text similarity in human and AI-generated scientific abstracts: a comprehensive analysis. IEEE Access. 2025.
28. Coates P, Breitinger F. Identifying document similarity using a fast estimation of the Levenshtein distance based on compression and signatures. arXiv preprint arXiv:2307.11496. 2023.
29. Pal A, Mustafi A. Vartani Spellcheck—automatic context-sensitive spelling correction of OCR-generated Hindi text using BERT and Levenshtein distance. arXiv preprint arXiv:2012.07652. 2020.
30. Cheng D, Wang P, Yao C. Levenshtein OCR. arXiv preprint arXiv:2209.03594. 2022.
31. Hu Y, Jing X, Ko Y, Rayz JT. Misspelling correction with pre-trained contextual language model. arXiv preprint arXiv:2101.03204. 2021.
32. Davenport MJ. Enhancing legal document analysis with large language models: a structured approach to accuracy, context preservation, and risk mitigation. Open Journal of Modern Linguistics. 2025;15(2):232-280.
33. Hládek D, Staš J, Pleva M. Survey of automatic spelling correction. Electronics. 2020;9(10):1670.
34. Vukatana K. OCR and Levenshtein distance as a measure of image quality accuracy for identification documents. In: Proceedings of the International Conference on Electrical, Computer and Energy Technologies (ICECET); 2022; IEEE. p. 1-4.
35. Sharmeen A, Agarwal N, Suman A, Agarwal S, Kumar K. Process design to self-extract text from images for similarity check. In: Advances in Intelligent Computing and Communication: Proceedings of ICAC 2021. Singapore: Springer Nature; 2022. p. 109-117.
36. Koning B. Extracting sections from PDF-formatted CTI reports [Bachelor's thesis]. University of Twente; 2022.
37. Jain P, Anand D, Behal V, Kapoor V, Sharma N, Jindal N. Verification of news and certificate, and plagiarism detector. Multimedia Tools and Applications. 2025;1-28.