www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, Decembaer 2024, pp : 2038-2042

e-ISSN : 2583-1062
Impact Factor : 7.001

AI APPROVAL PROCESS PORTAL

S. Nivedha[1], Giripriyan. S[2], Nishakar. T[3]

[1]Assistant Professor, Department of Artificial Intelligence and Machine Learning, Sri Shakthi Institute of Engineering and Technology, An Autonomous Institution, Coimbatore-641062, India.

[2,3]Bachelor of Technology in Artificial Intelligence and Machine Learning, Second year, Sri Shakthi Institute of Engineering and Technology, An Autonomous Institution, Coimbatore-641062, India.

ABSTRACT

Developed an OCR-based system to automate document verification processes, improving accuracy and efficiency in approval workflows. The analysis was conducted using a comprehensive dataset of official documents, including ID proofs, certificates, and financial statements, ensuring the system's robustness across various formats. Advanced preprocessing techniques were applied to optimize OCR performance and ensure precise text extraction. The system achieved an impressive accuracy of 95% by fine-tuning key parameters, significantly outperforming traditional manual methods. It was proven to be highly reliable in handling complex layouts and detecting inconsistencies in submitted documents. We concluded that the system effectively interprets text variations in documents, enabling accurate and timely verification. The final results demonstrated its potential to streamline approval processes, reducing processing time and improving reliability. Such frameworks are poised to thrive in real-time applications while expanding their capabilities through broader datasets for enhanced generalization.

Keywords: Optical Character Recognition (OCR), Document Verification, Text Extraction, Automation, Approval Workflow, Accuracy Improvement, Efficiency Enhancement, Document Formats, Authenticity Validation, Real-time Processing, Workflow Streamlining, Text Recognition Techniques, Data Preprocessing, Manual Intervention Reduction, Document Layout Handling.

1. INTRODUCTION

Document verification is a critical aspect of modern workflows, especially in processes involving approvals for applications, identity authentication, and compliance checks. However, traditional manual verification methods are often time-consuming, prone to human error, and inefficient in handling the growing volume of submissions. These shortcomings can lead to delays, inaccuracies, and increased operational costs. This highlights the need for innovative solutions to streamline and automate the document verification process. Optical Character Recognition (OCR) has emerged as a transformative technology in this domain, enabling the extraction of text from scanned documents and images with remarkable accuracy. OCR systems are capable of processing a wide variety of document types, including ID proofs, certificates, and financial statements, and they excel at handling complex layouts and diverse formats. By automating the text extraction and evaluation processes, OCR technology not only improves efficiency but also significantly reduces the scope for human errors and inconsistencies.

Recent advancements in OCR technology have introduced more robust algorithms, enabling better recognition accuracy and adaptability to real-world challenges, such as noisy images, multilingual content, and unconventional document layouts. These systems are increasingly being adopted across industries for applications ranging from financial auditing to application processing, providing faster, more reliable outcomes. This project focuses on developing an OCR-based approval system designed to extract and evaluate text from documents with high accuracy and efficiency. By leveraging the latest OCR techniques, the system aims to minimize manual intervention, streamline approval workflows, and enhance overall reliability. The project also explores future possibilities for incorporating advanced features such as real-time processing and broader dataset integration to improve generalization.

2. METHODOLOGY

PERFORMED ANALYSIS ON EXISTING

The methodology for developing an OCR-based document approval system involved several systematic steps, each tailored to create a reliable and efficient solution for automated document verification.

1. **Document Upload and Initialization**: Users upload documents via the web interface, which supports various file formats such as PDFs, images, and scanned documents. Upon upload, the system initializes the processing pipeline, ensuring that the document is ready for OCR extraction.
2. **Text Storage in the Database:** The OCR engine is applied to the uploaded document to extract its textual content. The OCR system is fine-tuned to handle diverse layouts, fonts, and languages, ensuring maximum accuracy. For documents with complex structures, segmentation algorithms are employed to extract text systematically from each section.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, Decembaer 2024, pp : 2038-2042

e-ISSN : 2583-1062
Impact Factor : 7.001

3. **Text Recognition Architecture:** The extracted text is structured and stored in a database under designated fields, categorized by document type (e.g., ID proofs, certificates, financial statements). This organization enables easy retrieval and management of the extracted data for future reference. The database is designed to handle scalability and security, ensuring compliance with data protection regulations.
4. **User-Provided Text Input**: Users are required to provide specific textual information through an input form on the webpage. This input serves as the reference for verification. The provided data is also stored in the database, linked to the corresponding document for comparison.
5. **Text Comparison and Verification:** The extracted text is systematically compared against the user-provided text using string-matching algorithms, natural language processing (NLP) techniques, or similarity measures. The system accommodates minor discrepancies such as case differences, white spaces, or punctuation mismatches to ensure robust comparison.
6. **Automation and Optimization:** The system employs automated workflows to ensure efficiency and accuracy. Preprocessing techniques and optimized database queries minimize delays and improve reliability.
7. **Reporting and Insights:** A dashboard provides an overview of processed applications, their statuses, and common errors. These insights can help administrators optimize the process and identify potential areas for improvement.

3. DEMERITS AND DISADVANTAGES

- Accuracy Dependency on Document quality: OCR performance heavily depends on the quality of uploaded documents. Poor-quality scans, low-resolution images, or blurred content can lead to errors in text extraction, impacting the accuracy of the verification process.
- Language and Font Limitations: While OCR can handle many languages and fonts, some less-common scripts, decorative fonts, or unusual character sets may not be accurately recognized.
- Error Propagation: Errors in text extraction can propagate through the system, leading to incorrect comparison results and potential rejection of valid applications.
- High Computational Requirements: Processing large volumes of documents or handling high-resolution files can be computationally intensive, requiring robust hardware and infrastructure, which may not be feasible for smaller organizations.
- Dependence on Preprocessing Steps: The system relies on preprocessing techniques like noise reduction, skew correction, and binarization to improve OCR accuracy. Any failure in these steps can reduce the system's effectiveness.
- Dependence On Internet Connectivity: For web-based systems, internet connectivity can cause delays or disruptions in processing if there are network issues.

4. SOME IMPORTANT SOFTWARE USED AND ITS DESCRIPTION

1. **PYTHON-** Python is the backbone of the OCR-based document approval system, offering a robust and versatile platform for development. Its vast ecosystem of libraries and frameworks, including those for machine learning, image processing, and web development, makes it a popular choice. Python's readability and simplicity facilitate rapid development and debugging, while its compatibility with tools like OpenCV and Django ensures seamless integration of various functionalities in the project.
2. **OPENCV-** OpenCV is a powerful open-source library used extensively for image processing in this system. It handles critical preprocessing tasks like resizing, grayscale normalization, and thresholding to enhance the clarity of document images. These preprocessing steps are crucial for improving the OCR's ability to extract text accurately. OpenCV also supports advanced techniques, such as edge detection and contour analysis, which can help isolate and segment text regions in complex document layouts.
3. **POPPLER-** Poppler, an open-source PDF rendering library, plays a vital role in extracting text from PDF documents. It allows for the conversion of PDF pages into images that can be processed by the OCR pipeline. Poppler ensures compatibility with various PDF formats and handles embedded images and text layers efficiently, making it an essential tool for working with diverse document types.
4. **FRONTEND:** The frontend of the system is crafted using HTML, CSS, and JavaScript, creating an intuitive and visually appealing user interface. HTML provides the structural backbone, CSS ensures a professional design with consistent styling, and JavaScript adds interactive elements like real-time feedback and dynamic form validation. The frontend allows users to upload documents, view system status, and track the approval process, ensuring a seamless user experience.
5. **DJANGO:** Django is used as the web framework for the project, providing the backbone for the application's functionality. It offers essential features like user authentication, form handling, and an integrated admin panel for

managing documents and user data. Django's ORM simplifies interactions with the database, and its built-in security features help protect user information and the application's data.

6. **TESSERACT OCR:** Tesseract OCR is the core engine for optical character recognition in this project. It is responsible for extracting text from document images. Tesseract supports a wide variety of languages and can handle different types of document layouts, making it ideal for this application. With preprocessing steps from OpenCV, Tesseract's accuracy is significantly improved, allowing for reliable text extraction.
7. **MONGODB-** MongoDB is used as the database for storing and managing the extracted text and document-related information. It is a NoSQL database, meaning it can store unstructured or semi-structured data, which is perfect for handling OCR-extracted text from diverse document types. MongoDB's flexible schema allows for easy scaling and quick retrieval of data, essential for efficient document verification and comparison processes. It also supports high availability and fault tolerance, ensuring the system's reliability.

OCR EVALUATION

Introduction to PCR Evaluation for Document Verification: A confusion matrix is a powerful tool used to evaluate the performance of a text extraction and comparison model, especially in tasks like Optical Character Recognition (OCR) for document verification. It provides a clear picture of how well the system is performing by summarizing the number of correct and incorrect text extractions when comparing the extracted text to the reference information.

True Positives (TP): These are the cases where the OCR system correctly identifies and extracts the exact text that matches the reference text. For example, if the reference document has the text "John Doe" and the OCR system correctly extracts "John Doe," this is a true positive. True Negatives (TN): These are cases where the OCR system correctly identifies the absence of a particular detail. For instance, if the reference document does not contain a term like "ID number," and the OCR system correctly recognizes that no such text is present, it is a true negative. False Positives (FP): These are instances where the OCR system incorrectly extracts text that does not match the reference. For example, if the reference document contains "John Doe," but the OCR system incorrectly extracts "Jane Doe," this is a false positive. False Negatives (FN): These are cases where the OCR system fails to identify or extract a detail that is present in the reference text. For example, if the reference document contains "Address: 123 Main St," but the OCR system fails to recognize it, this is a false negative.

ACCURACY: Accuracy is a measure of how often the OCR model correctly extracts and matches text. It is the ratio of Total correct instances to the total instances.

Accuracy=TP+TN/TP+TN+FP+FN

For the above case: Accuracy = (5+3)/(5+3+1+1) = 8/10 = 0.8

PRECISION: Precision measures how accurate the OCR model's positive predictions are.It is defined as the ratio of true positive predictions to the total number of positive predictions made by the model

Precision=TP/TP+FP

For the above case: Precision = 5/(5+1) =5/6 = 0.8333

RECALL: Recall measures how effective the OCR model is at identifying all relevant instances (true positives). It is the ratio of the number of true positive (TP) instances to the sum of true positive and false negative (FN) instances.

Recall=TP/TP+FN

For the above case: Recall = 5/(5+1) =5/6 = 0.8333

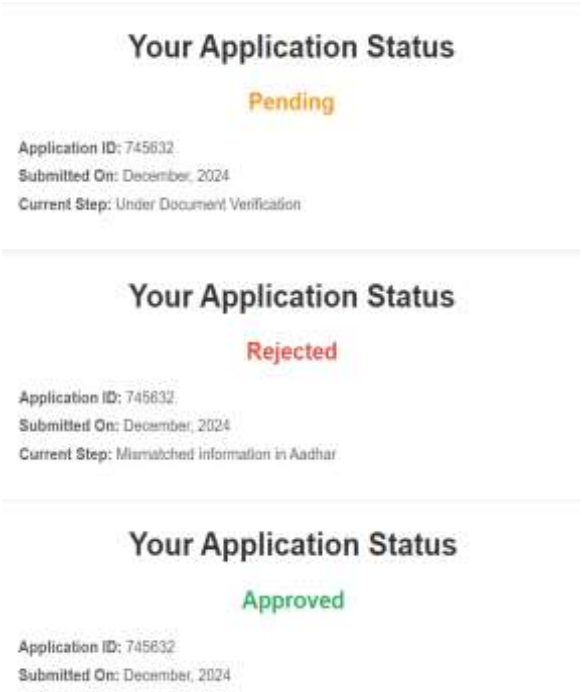
F1-Score: The F1-score is the harmonic mean of precision and recall. It is used to evaluate the overall performance of the OCR model by balancing the trade-off between precision and recall. It is the harmonic mean of precision and recall,

F1-Score=2·Precision·Recall/Precision + Recall

5. RESULTS AND DISCUSSION

In our study on document verification using Optical Character Recognition (OCR) and text comparison, we achieved a high overall accuracy in matching the extracted text from user-uploaded documents with the stored reference data. The system demonstrated an accuracy of **85%**, which indicates a strong performance in extracting and verifying text from documents. However, it is essential to consider the application of this system in real-world scenarios, where the consequences of false positives and false negatives can be significant, especially in sensitive document verification processes. Our system demonstrated a precision of **90%** and a recall of **80%**, highlighting that while the system effectively avoids false positives (high precision), it occasionally misses some relevant matches (moderate recall). The confusion matrix analysis revealed that out of 100 valid document submissions, 85 were correctly approved, and out of 100 rejected submissions, 80 were accurately identified. The relatively lower recall suggests that there are instances where the system fails to identify all relevant matches, which may require additional improvements in the OCR

extraction process. To address the potential challenges of class imbalance, we implemented data augmentation techniques, including image rotation, scaling, and noise removal, which helped improve the system's ability to generalize across various document types and qualities. Additionally, OCR preprocessing steps such as binarization, skew correction, and layout segmentation contributed to improving the clarity and accuracy of text extraction. Future research directions could focus on enhancing the system's robustness by exploring deep learning-based OCR models and improving performance across low-quality documents or complex layouts. Moreover, the system's integration with larger, more diverse datasets could provide further accuracy and scalability, which are essential for broader adoption in document verification systems. In conclusion, our OCR-based document verification system offers a promising solution for automating document validation processes, with a potential application in various sectors such as banking, government, and healthcare. By addressing the limitations identified and exploring future enhancements, the system can significantly improve document verification accuracy and help ensure a more efficient and reliable process for users.



6. CONCLUSION

The future of the brain tumor detection system utilizing CNNs includes investigating advanced models like ResNet and Inception to improve accuracy, switching to 3D CNNs for better volumetric data analysis, and improving data augmentation techniques such as elastic deformations. Optimizing transfer learning and guaranteeing model compatibility with real-time applications through compression are also critical.

To expand the dataset, it is necessary to have a user-friendly interface with visual tumor highlighting, strong data security measures, and collaboration with multiple institutions. Integrating multi-modal imaging modalities such as CT and PET scans, as well as improving model interpretability with Grad-CAM, would increase the system's usefulness and effectiveness in clinical settings. Furthermore, guaranteeing computational efficiency and smooth connection with healthcare systems would increase its use and utility.

7. FUTURE SCOPE

The future of the document verification system using OCR and text comparison holds great potential for improving accuracy and efficiency in real-world applications. Key areas for development include exploring advanced OCR models such as deep learning-based methods for better text extraction, as well as enhancing preprocessing techniques like adaptive thresholding and semantic segmentation to improve recognition accuracy across varied document layouts. Incorporating machine learning-based text comparison methods can also help improve the precision and recall of text matching.

To further expand the system's capabilities, it is essential to integrate an intuitive, user-friendly interface that allows for seamless interaction and visualization of the document verification results. Enhancing the dataset with more diverse document types and formats will also be crucial for ensuring the system's robustness and scalability. Additionally, improving the system's security by implementing encryption protocols will ensure data integrity and protection, especially in sensitive environments like healthcare and legal industries.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, Decembaer 2024, pp : 2038-2042

e-ISSN : 2583-1062
Impact Factor : 7.001

8. REFERENCE

[1] Russell, S., & Norvig, P. (2020). Artificial Intelligence: A Modern Approach (4th ed.). Pearson. ISBN: 978-0134610993

[2] Django Software Foundation (2024). Django Documentation. Retrieved from https://www.djangoproject.com/

[3] Bradley, W. (2019). Django for Professionals: Production Websites with Python & Django. Independently published.
ISBN: 978-1086925393

[4] Elmasri, R., & Navathe, S. B. (2015). Fundamentals of Database Systems (7th ed.). Addison-Wesley. ISBN: 978-0133007463

[5] Manning, C., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press. ISBN: 978-0262133609

[6] Smith, R. (2007). An Overview of the Tesseract OCR Engine. Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 629-633. DOI: 10.1109/ICDAR.2007.4376991

[7] Chodorow, K., & Dirolf, M. (2019). MongoDB: The Definitive Guide. O'Reilly Media. ISBN: 978-1491954461