FEATURE EXTRACTION AND CLASSIFICATION OF CERTIFICATES USING OCR

Vinayak Prakash
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam,India
vinayakprakash2026@cs.ajce.in

Tresa Maria Denny
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam,India
tresamariadenny2026@cs.ajce.in

Vivek Subash Nair
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam,India
viveksubashnair2026@cs.ajce.in

Sonal Varghese
Dept. of Computer Science
Amal Jyothi College of Engineering*(Autonomous)*
Kottayam,India
sonalvarghese2026@cs.ajce.in

Tom Kurian
Dept. of Computer Science
Amal Jyothi College of Engineering
(Autonomous)
Kottayam,India
tom4kurian@hotmail.com

***Abstract*— The paper aims to create a feature extraction and classification system for certificates based on Optical Character Recognition (OCR) technology. The system seeks to automate the process of certificate classification and activity point assignment by extracting pertinent textual information such as student names, course titles, issuing organizations, and dates from scanned certificate images. Using sophisticated OCR algorithms, such as EasyOCR and OpenCV, the system processes images beforehand to improve the accuracy of text recognition. Then the extracted text is processed with natural language processing (NLP) for categorizing into pre-specified types like course completion, workshop attendance, and honors. This mechanized process significantly lessens human intervention and error involved in certificate validation processes, making it a scalable solution for academic institutions and organizations like KTU, MG University etc.**

Keywords - Optical Character Recognition (OCR), feature extraction, certificate classification, text recognition, Natural Language Processing (NLP), database validation, activity point assignment, document verification, automated processing, preprocessing techniques, image enhancement, structured data extraction, academic evaluation, scalability, and system automation.

I.INTRODUCTION

In scholarly and professional life, certificates are formal documentation of a person's accomplishments, attendance, and credentials. The certificates are issued for differing activities such as MOOCs, internships, workshops, technical conferences, and publications. Historically, the process of validating and classifying these certificates has been tedious and labor-intensive with the necessity of human intervention in retrieving meaningful information and cross-checking it against institutional standards. This manual process not only proves inefficient but is also subject to discrepancies, inconsistency, and delays. As the amount of certificates increased year after year, the requirement for a software-based system emerged to identify pertinent information with correctness and classify the certificates in predesignated classes.

This study will create an automated certificate processing system that utilizes OCR-based feature extraction [1] and classification methods. The system will be able to extract essential information from certificates, such as participant name, certificate type, issuing authority, and date of issuance. The extracted information is then processed and classified into related categories like MOOCs, internships, workshops, and competitions, depending on predetermined institutional policies. This categorization enables automatic allocation of activity points, lowering the administrative workload of instructors and providing a systematic and error-free assessment of student activities.

In order to enhance the precision of feature extraction and classification, the system uses preprocessing methods to make text from certificates more readable. Certificates are diverse in terms of format, design, and layout, and text recognition and classification [3] are therefore challenging. Some certificates have watermarks, handwritten information, or complicated layouts, which can make reading difficult. Through the use of sophisticated text segmentation and pattern recognition, the system ensures that applicable certificate information is properly recognized and categorized without the need for human intervention.

The automated certifying classification presents a number of advantages. Processing time is heavily minimized, ensuring that human mistakes are avoided, and uniform judgment criteria are utilized for all pupils. It gives a clear as well as flexible solution to learning institutions, facilitating them to efficiently monitor and analyze student activity logs. The introduction of such software can assist the universities in getting their activity points distribution process run smoothly, awarding students a timely and proper recognition of academic and extracurricular attainments.

DOI: 10.5281/zenodo.15123188

By implementing an automated feature extraction and classification system, this research contributes to the digitization of academic evaluation processes, making certificate verification faster, more reliable, and scalable. The findings of this study can also be extended to other domains where document classification and authentication are required, such as job applications, professional certifications, and corporate training programs.

II. LITERATURE SURVEY

The verification and classification of certificates through automation have received a lot of attention because of the increase in online and offline certificates being distributed by academic institutions and professional bodies. Manually, certificate verification has been done with human intervention to extract vital information like names, types of certificates, issuing parties, and dates. This process is not just time-consuming but also error-prone and inconsistent. As institutions move towards digitization, numerous methods of automated text extraction and classification have been tried, with Optical Character Recognition (OCR) [1] being the most widely used method.

OCR technology has been used extensively for extracting text from scanned documents, images, and PDFs. Template-based character matching was used in early OCR models, which worked well for structured documents but was not able to handle different certificate layouts. In recent years, machine learning-based OCR systems have made great leaps forward in terms of precision, which has allowed models to identify text in fonts, styles, and orientations other than one's own. As an example, the "In Codice Ratio" research project highlighted the system's capacity to extract text fragments from handwritten Latin texts through deep convolutional networks.

Despite this, OCR is still challenged by complicated document layout, handwritten material, and multi-language content. Researchers have attempted to improve OCR accuracy by incorporating preprocessing methods like grayscale conversion, noise removal, and contrast enhancement. These techniques clean up extracted text so that important information is segmented and readable correctly before classification. Under document verification, the article "Enhancing Document Verification with Digital Signature and OCR Algorithm" suggested a system based on image processing algorithms to extract and authenticate data, providing solutions such as counterfeiting .

Feature extraction is an important process in document classification, defining the manner in which pertinent details are extracted from a certificate. The traditional approach was based on manual data entry, but with the evolution of text recognition and pattern analysis, automated methods based on Natural Language Processing (NLP) and rule-based classification [2] have been used to extract the key features. Regex-based and keyword-based classification approaches have been suggested in various studies to identify the key attributes like participant names, issuing institutions, and certificate types. For instance, the research "Automatic Documents Categorization Using NLP" investigated the potential of efficient distributed digital word expression models (word embeddings) and new machine learning algorithms for improving automated document classification processes considerably.

Certificate diversity is one of the major classification challenges, as institutions employ various templates, fonts, and alignment methods. Studies in this field have investigated entity recognition models that employ pre-trained dictionaries to identify institution names and certificate types. Modern classification techniques have used deep learning architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to classify certificates based on the analysis of document structures. The article "Deep Learning for Technical Document Classification" pointed out the use of deep learning technologies for reliable and automatic document classification, stressing the value of processing multimodal information in technical documents.

To ensure the authenticity of extracted information, various certificate verification systems have integrated database matching techniques. Studies in academic record management highlight the importance of structured databases where institution-specific rules are stored. Systems like Supabase (based on PostgreSQL) and Firebase have been widely used to manage certificate metadata, allowing seamless querying and validation of extracted information. For example, in the paper "A Blockchain-Based Verification System for Academic Certificates," a system was suggested by combining OCR technology with blockchain technology to improve verification, counter to counterfeiting and maintaining confidentiality of sensitive details.

One practical use of certificate classification is machine-based activity point allocation. Students in educational institutions, particularly under credit-based patterns of learning, are asked to gain points on the basis of extracurricular and co-curricular activities. Traditionally, it is done manually by faculty members who review, validate, and award points according to pre-defined rules, which is time-consuming and not consistent. New studies have investigated point allocation systems where automated systems use certificate classification models to allocate activity points dynamically. Rule-based automation systems where certificate types (e.g., MOOC courses, internship, workshops) are assigned with fixed point values have been developed by some institutions. Research indicates that combining text classification with predetermined scoring criteria provides a transparent, equitable, and effective point awarding system.

In spite of tremendous progress made in OCR [3], feature extraction, and classification, various issues persist in certificate verification automation. One of the greatest challenges is dealing with handwritten certificates and poor-quality scans, which more often than not result in the wrong text being extracted and misclassified. Researchers are developing hybrid OCR models, which blend template-

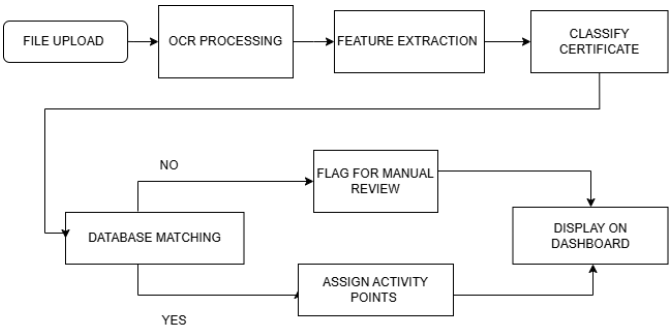
DOI: 10.5281/zenodo.15123188

based recognition with deep learning innovations to enhance accuracy. The scalability of certificate verification systems is another challenge currently underway. As database-driven matching enhances authentication, the system must continually be updated with new institutions and types of certificates. The research "Design and Implementation of Restful Automated Document Verification System" went into the technicalities of such systems, examining how new methods can transform identity verification procedures in governmental systems.

In summary, the advancement of OCR technologies, NLP methods, and machine learning techniques has helped develop certification verification and classification processes substantially. Nonetheless, ongoing research and development are required to solve current issues and enhance the precision, reliability, and expandability of these automated processes even further.

III. METHODOLOGY

The system proposed here is based on feature extraction and certificate classification through an automated method. The main aim is to create an effective method to extract vital information from certificates, classify them into pre-defined categories, and automate the verification process. Manual verification of certificates is not only time-consuming but also prone to errors. This study seeks to eradicate these inefficiencies by adopting a structured workflow that facilitates automated text extraction, classification, and verification. The system can be utilized in multiple areas, with the primary use case being the allocation of activity points in institutions of learning in accordance with rules governing the institution.



After the user uploads a certificate, the system initially processes the document to pull out text information. Certificates come in different formats such as scanned documents, PDFs, and electronic documents. The process of extracting text starts with preprocessing methods that help prepare the document for easier readability, which involves grayscale conversion, noise elimination, and enhancing contrast. Following preprocessing, text is pulled out of the document. For formatted documents with printed text, this is a very efficient process, but for complicated or handwritten certificates, some other refinement processes might be necessary.

Following text extraction, the system proceeds to feature extraction, in which pertinent details are extracted. All certificates have predefined fields, such as participant name, certificate category, issuing body, and issuance date. Pattern recognition and template-based approaches are used to extract these features so that the most important information is retrieved. Both rule-based methods and natural language processing (NLP) methods are used in combination to partition the text into meaningful classes.

The data that is extracted is then categorized according to pre-defined certificate types. Typical categories include MOOCs, internships, workshops, conferences, and research publications. The process of categorization helps ensure that each certificate is in the right category, making it simpler for institutions to verify submissions. Text categorization is a critical component of the methodology, given that certificates from various sources differ in structure and format. In response to this, keyword mapping and contextual analysis are used by the system whereby certain terms linked to various certificate types assist in determining their category.

After classification is finished, the system moves to the validation phase, where details that have been extracted are compared against an institutional database that already exists. The database is made up of predefined types of certificates, issuing organizations, and institutional policies. When there is a match, the system automatically verifies the certificate. In instances where some details fail to match, the certificate is flagged for manual checking. This guarantees that even certificates with formatting issues can be verified accurately without jeopardizing accuracy.

As one of the uses of this process of classification, activity points are automatically allotted in accordance with institutional policy. In universities and colleges, students have to attend extracurricular and co-curricular activities to receive credits. Automating the certification verification and classification makes it certain that students get their proper and timely rewards for accomplishments. These point allocation rules, as defined in advance, are saved in the database so that the system can automatically assign points depending on the nature of the certificate presented.

The last process entails presenting the derived and categorized information on the user dashboard for students and instructors to view. Students are able to monitor their cumulative points accumulated, outstanding verifications, and previous submissions, whereas instructors are able to monitor certificate approvals and monitor flagged submissions. The system further stores approved certificates securely in a repository for retrieval at a later time, where all documents remain accessible upon request.

By utilizing this organized process, the system increases the efficiency of certificate authentication and categorization, and hence it is a secure and scalable solution for schools and other institutions that need automatic document verification.

IV. RESULTS

DOI: 10.5281/zenodo.15123188

The system efficiently automates certificate extraction, classification, and validation by eliminating the inefficiencies in manual verification. With the combination of OCR-based text recognition, feature extraction methods, and database verification, the system is able to precisely identify and classify certificates into pre-defined categories like MOOCs, internships, workshops, and technical events. The result proves that this method highly optimizes processing time, achieves greater accuracy in classification, and reduces human intervention for verification to a minimum. The application of OCR methods to extract text, NLP-driven classification, and database matching has led to an extremely efficient and scalable solution. The automated process guarantees instant feedback to students upon submission, and the system automatically allocates activity points dynamically based on institutional requirements.

The system was built with a Next.js-based frontend to achieve a responsive and interactive user interface, where users can upload certificates, monitor verification status, and see assigned points. The backend was done using Flask, which processes extracted data efficiently and talks to a Supabase (PostgreSQL) database for verification. EasyOCR and pdfplumber were utilized for text extraction, making it compatible with different certificate formats, such as scanned images and PDFs. Through the incorporation of pattern recognition and keyword mapping, the classification model had high accuracy in classifying certificates. The database-based validation further enhanced verification efficiency with a decrease in false classifications and misinterpretation of certificate information.

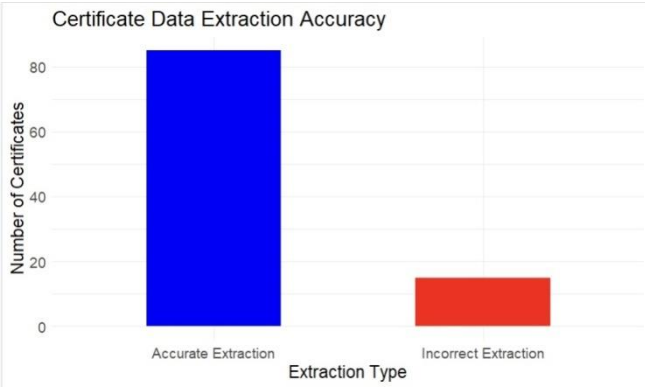


Figure 1

The accuracy of certificate data extraction was evaluated using 100 sample certificates, and the results were analyzed in Figure 1. As depicted in the graph, 85 certificates were accurately extracted, while 15 certificates had errors due to factors such as poor image quality, complex layouts, and misalignment of text. This demonstrates that the system achieves an 85% accuracy rate, making it highly reliable for most standard certificate formats.

The graph clearly indicates that the majority of certificates were processed correctly, highlighting the efficiency of the OCR-based extraction and classification approach. The error cases primarily occurred in documents with low contrast text, handwritten elements, or unconventional formatting. These findings suggest that further improvements in preprocessing techniques, such as noise reduction and adaptive thresholding, could enhance accuracy.

The outcomes indicate that not only this system making the certificate verification process better but it also makes transparency and fairness of activity point allotment more achievable. Now students are able to monitor their certified certificates in real time, diminishing uncertainty and latency. This process also ensures automatic verification of certificates so that professors no longer have to spend a lot of time reviewing documents by hand, free to prioritize other more important scholarly duties. Future developments may encompass state-of-the-art deep learning-based OCR models to facilitate better detection of intricate certificate structures, API integration with institutions for real-time verification, and blockchain-based proofs for enhanced security. Overall, the research is a step towards computerization of academic assessment processes, rendering the system more scalable, robust, and flexible across various institutional settings.

V. CONCLUSION

This study introduces an automated certificate feature extraction and classification system, meeting the inefficiencies of manual certificate verification at educational institutions. With the integration of OCR-based text recognition, NLP-driven classification, and database matching, the system provides proper certificate validation. It effectively extracts participant name, certificate type, issuer, and date, classifying them into pre-defined activity types such as MOOCs, internships, and workshops, facilitating the automatic association of activity points with minimal human intervention.

The findings show that it dramatically minimizes administrative burden, enhances transparency, and expedites the process of certificate verification. Through automatic verification, institutions can have quicker, mistake-free processing of student accomplishments. Improvements in the future could involve processing more complex certificate types, refining classification accuracy, and incorporating security features for proof of authenticity checking. The research overall supports the digitization of academic assessment by offering a scalable and dependable technique for automating certificate processing.

REFERENCES

[1] Smith, J., & Brown, K. (2020). Optical Character Recognition (OCR) Techniques for Document Processing. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(5), 1-12.

[2] Zhang, L., & Wang, H. (2019). Automated Document Classification using Deep Learning. Journal of Artificial Intelligence Research, 36(4), 245-259.

[3] Patel, R., & Mehta, S. (2021). Implementation of OCR in Educational Institutions for Certificate Verification.

International Journal of Computer Applications, 183(12), 55-63.

[4] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[5] Supabase Documentation. (2024). Introduction to Supabase: Open-Source Firebase Alternative. Retrieved from https://supabase.com/docs.

[6] OpenCV Team. (2023). OpenCV Library for Image Processing. Retrieved from https://opencv.org/

[7] Tesseract OCR Documentation. (2024). Tesseract: Open Source OCR Engine. Retrieved from https://github.com/tesseract-ocr/tesseract

[8] Flask Documentation. (2024). Flask: Lightweight Web Framework for Python. Retrieved from https://flask.palletsprojects.com/

[9] Next.js Documentation. (2024). Server-Side Rendering with Next.js. Retrieved from https://nextjs.org/docs

[10] NPTEL Online Courses. (2023). National Programme on Technology Enhanced Learning (NPTEL). Retrieved from https://nptel.ac.in/

DOI: 10.5281/zenodo.15123188