CONFERENCE PROCEEDINGS, JANUARY 2025
Published by the Faculty of Science (FSC), FULafia
Print ISSN: 2354–3388 DOI: https://doi.org/10.62050/fscp2024.462

Document Classification in Higher Education Institutions Using Deep Learning: A CNN, RNN, and Hybrid CNN-RNN Approach

Abdulkarim Abdullahi✉, John K. Alhassan & Sulaimon A. Bashir
Department of Computer Science, Federal University of Technology, Minna Nigeria
✉abdullahi.abdulkarim@futminna.edu.ng

Abstract: Higher Education Institutions (HEIs) are increasingly confronted with the complexities of evolving rules and requirements, necessitating innovative technology solutions to streamline document handling processes. Traditional paperwork methods are often inefficient and error-prone, leading to potential non-compliance. This research addresses these challenges by developing an AI-powered electronic document management system designed to automate compliance checks and simplify document handling as HEIs grow. The primary objective is to create a document classification model utilizing deep learning techniques, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and a hybrid CNN-RNN approach, to enhance document accuracy and compliance. The study involves collecting and preprocessing a substantial dataset of documents, designing and evaluating various deep learning models, and optimizing hyperparameters. Performance comparisons among the models indicate that the hybrid CNN-RNN architecture outperforms individual models, achieving superior accuracy, recall, and F1-score, alongside a significantly lower mean squared error (MSE). Initial evaluations revealed the CNN, RNN, and CNN-RNN models achieved accuracies of 73, 44, and 27%, respectively, on the raw dataset. However, with an upgraded dataset, these models improved to 76, 48, and 79% accuracy, respectively, highlighting the hybrid model's enhanced capability in accurately classifying documents. The findings revealed the effectiveness of integrating advanced deep learning techniques to improve document verification processes in HEIs, ultimately facilitating better compliance and operational efficiency.

Keywords: Higher education institutions, convolutional neural networks, recurrent neural networks, classification, deep learning

Introduction

Higher Education Institutions (HEIs) are increasingly confronted with the complexities of evolving rules and requirements, necessitating innovative technology solutions to streamline document handling processes. Traditional paperwork methods are often inefficient and error-prone, leading to potential non-compliance [1]. Research by Abang *et al*. [2] highlights how manual document handling becomes cumbersome as institutions expand, necessitating automated systems to maintain efficiency and accuracy. Jayoma *et al*. [3] and Tanuraharja *et al*. [4] emphasize the inefficiencies of paper-based systems, further justifying the adoption of AI-driven solutions.Artificial intelligence (AI) has shown significant promise in document management, particularly through deep learning models that improve classification accuracy and compliance verification. Mittal & Mittal [5] explore how AI-powered document management systems enhance compliance by automating indexing and archiving processes. [6] demonstrate that computer-based document management significantly reduces inefficiencies, thereby streamlining workflows in administrative settings. Nautiyal *et al.* further assert that AI integration in document verification enhances operational efficiency, reducing the risk of human error [7].

Bhatlawande *et al.* present a document classification model using CNN-based hyperparameter tuning to effectively categorize documents, addressing the issue of mixed-up industrial documentation [8]. Their approach achieved an 81% accuracy score and demonstrated the potential for scaling to large-scale document classification scenarios. Moreover, their research highlights instances where traditional machine learning techniques can outperform deep learning models, achieving a remarkable 94% accuracy. This study reinforces the importance of intelligent classification models in preventing document mismanagement and reducing operational costs. Renjith *et al.* [9] propose a hybrid CNN-RNN model for sign language recognition, demonstrating how combining CNNs for spatial feature extraction with RNNs for temporal connections can improve classification performance. Their model achieved a 98.2% accuracy rate, showcasing the effectiveness of hybrid deep learning architectures in tasks requiring both spatial and sequential processing. The success of this approach in sign language recognition indicates its potential applicability in other domains, such as document classification and compliance verification in HEIs.

This research addresses these challenges by developing an AI-powered electronic document management system designed to automate compliance checks and simplify document handling as HEIs grow. The primary objective is to create a document classification model utilizing deep learning techniques, including Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and a hybrid CNN-RNN approach, to enhance document accuracy and compliance.The significance of this research lies in its potential to improve accuracy, reduce manual efforts, and ensure that documents meet institutional compliance standards. The study builds upon relevant literature that underscores the inefficiencies of traditional document management systems and explores how AI-driven solutions have improved compliance and accuracy in various domains [7]. Jayoma *et al.* [10] and Tanuraharja *et al.* [4] highlight the shortcomings of conventional document management approaches, further

supporting the need for automated solutions. This research contributes to the growing body of work on intelligent document management by demonstrating the effectiveness of hybrid deep learning models in classifying and verifying academic documents.

Materials and Methods

The dataset for this study was sourced from the Information Technology Services (ITS) of the Federal University of Technology, Minna. It comprises birth certificates, indigene letters, O' level results, and other institutional documents. The dataset consists of 287 O' level result images, 288 referee images, 257 birth certificates, and 259 JAMB certificate images, summing up to a total of 1,091 raw images. Ethical considerations were adhered to, ensuring all collected data were anonymized to protect user privacy.Given that 1,091 images are relatively small for training a deep learning model, the study incorporated data augmentation techniques to expand the dataset. Six augmentation operationsrotation, flipping, contrast adjustments, scaling, noise addition, and brightness variationswere applied to each raw sample image. This augmentation process increased the dataset size significantly, generating a total of 7,637 images from the initial 1,091 samples. The expanded dataset ensures improved model generalization and performance. To standardize the data, all document images were resized to a uniform dimension of 200x200 pixels and converted to grayscale to enhance feature extraction and reduce computational complexity. Textual data were extracted from images using Tesseract Optical Character Recognition (OCR) to facilitate document classification beyond image-based features. The dataset was then partitioned into 80% training and 20% testing sets to ensure a balanced evaluation and improve model reliability.

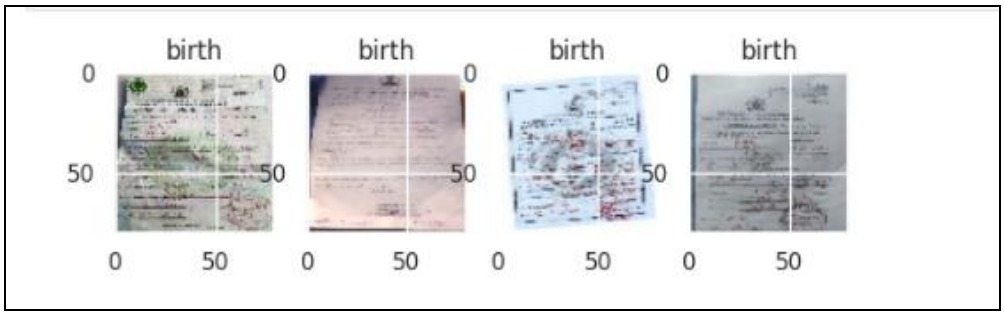

Figure 1: Birth cert. image sample plot

Figure 2: O'level image sample plot

A hybrid deep learning model combining CNN and RNN architectures was developed for document classification. The CNN component extracted spatial features from document images, while the RNN component captured sequential dependencies in textual data extracted via OCR. The model was implemented using Tensor Flow 2.16.1, with the CNN consisting of multiple convolutional layers followed by max-pooling layers to extract hierarchical feature representations, and the RNN using Long Short-Term Memory (LSTM) units to analyze sequential text data. Hyperparameter tuning was performed to optimize batch size, learning rate, and dropout rates, while the training process employed Stochastic Gradient Descent (SGD) with adaptive learning rates.

Model performance was evaluated based on multiple metrics, including accuracy, precision, recall, and F1-score, to assess classification reliability and prediction consistency. A five-fold cross-validation technique was applied to evaluate model robustness and prevent overfitting. Additionally, a paired t-test was used to compare the performance of CNN, RNN, and hybrid CNN-RNN models, with a significance threshold of $p < 0.05$. The results demonstrated that the hybrid CNN-RNN model outperformed standalone models in terms of accuracy and reliability.By leveraging AI-driven methodologies, this research contributes to enhancing document classification efficiency, reducing manual workload, and improving institutional compliance verification.

Results and Discussion

The results demonstrate that the hybrid CNN-RNN model outperforms standalone CNN and RNN models in document classification. Initial evaluations indicate that the CNN, RNN, and CNN-RNN models achieved accuracy rates of 73, 44, and 27%, respectively, on the raw dataset. However, after applying data augmentation techniques, the CNN accuracy increased to 76%, RNN to 48%, and the CNN-RNN model achieved the highest accuracy at 79%. These improvements validate the effectiveness of data preprocessing techniques in enhancing classification performance. Figs 3 and 4 show the performance of the standard alone models while Figs 5 and 6 show the performance of the hybrid model.

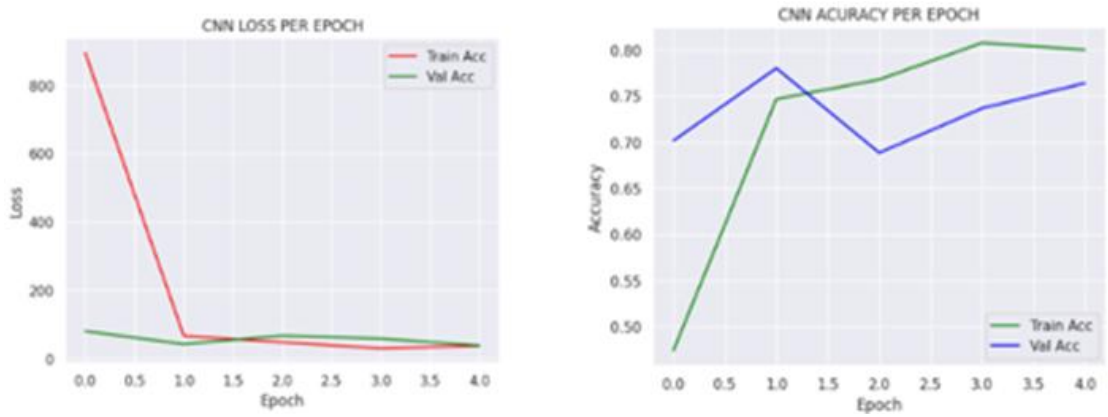

Figure 2: CNN augmented dataset

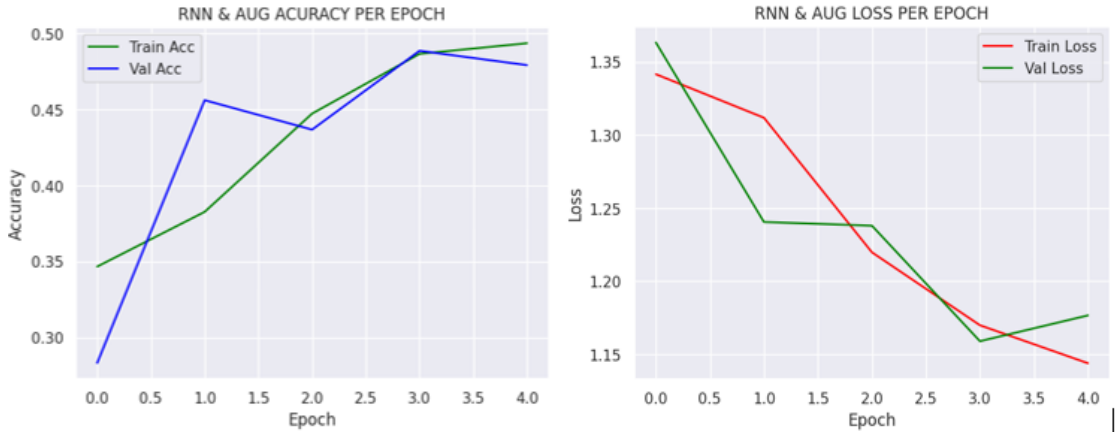

Figure 4: RNN augmented dataset

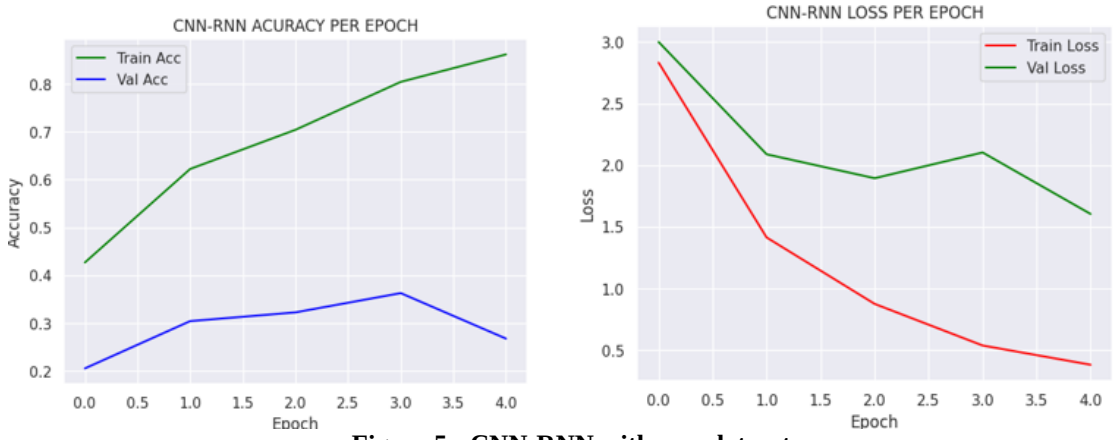

Figure 5: CNN-RNN with raw dataset

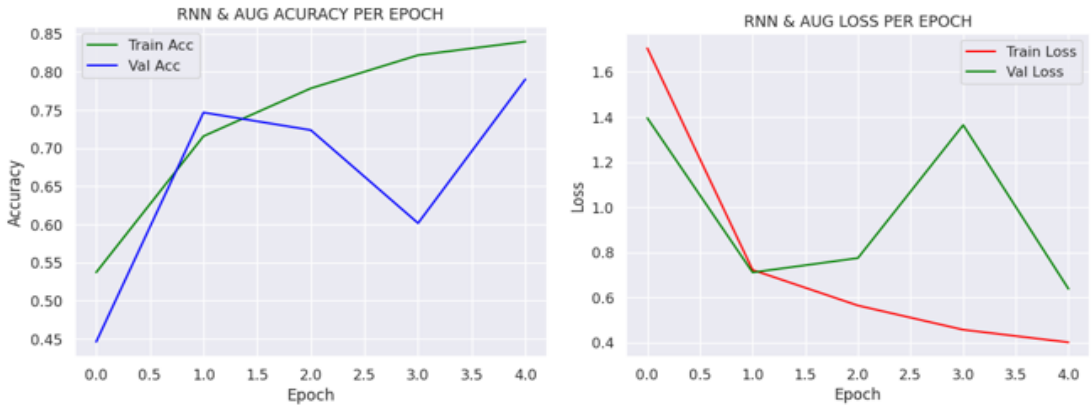

Figure 6: CNN-RNN with augmented dataset

Comparing the hybrid CNN-RNN model to previous studies, Bhatlawande *et al*. reported an 81% accuracy using CNN-based classification models [8], whereas Renjith *et al*. demonstrated superior performance with hybrid deep learning architectures [9]. The results of this study align with their findings, reinforcing the notion that integrating CNN and RNN components improves classification accuracy for sequential data, such as text extracted from images.

Table 1: Experimental result summary

S/N	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	CNN & Raw Dataset	73.0	79.5	72.0	68.2
2	CNN & Aug Dataset	76.0	84.0	76.0	77.0
3	RNN & Raw Dataset	44.0	43.0	42.0	42.0
4	RNN & Aug Dataset	48.0	53.2	47.2	47.2
5	CNN-RNN & Raw Dataset	27.0	56.7	30.0	21.0
6	CNN-RNN & Aug Dataset	79.0	79.2	78.7	79.2

The statistical analysis using a paired t-test confirmed that the CNN-RNN model's improvements were statistically significant ($p < 0.05$) compared to standalone CNN and RNN models. Additionally, the precision, recall, and F1-score metrics indicated a balanced classification performance, with the hybrid model outperforming individual models in document categorization. Figures 3 and 4 illustrate the performance of the standalone CNN and RNN models, while Figs 5 and 6 depict the enhanced classification capabilities of the hybrid CNN-RNN model.

Conclusion

In conclusion, the experimental analysis shows that the Hybrid CNN-RNN certificate classification approach does not work well with small amount of certificate image dataset, but in the present of large dataset (which is achieved via image augmentation) a significant spike in performance can be observed in term of accuracy. The performance has higher chance of improving better if more dataset can be collected. Hence, in the present of smaller sample dataset, the CNN classification model can be use in real life application, while CNN-RNN is advisable if large dataset is available.

Reference

[1] Sagum, J. K. A. (2021). Web-based document management system for PEP squad events and marketing services. *2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management*, 1–5. https://doi.org/10.1109/HNICEM54116.2021.9732033

[2] Abang, K. I. R., Gatmaitan, D. L. V., Manalo, F. R., Torcelino, M. R., Rodriguez, R. L. & Serrano, E. A. (2022). CCT online request of students credentials: A document management system for private HIEs in the Philippines. *2022 2nd International Conference in Information and Computing Research (ICORE)*, 25–29. https://doi.org/10.1109/iCORE58172.2022.00024

[3] Jayoma, J. M., Moyon, E. S. & Morales, E. M. O. (2020). OCR Based Document Archiving and Indexing Using PyTesseract: A Record Management System for DSWD Caraga, Philippines. *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*, 1–6. https://doi.org/10.1109/HNICEM51456.2020.9400000

[4] Tanuraharja, C. D., Tiara, K. A., Wang, G. & Alianto, H. (2022). Applying for E-signature approval with TOGAF framework to improve productivity: Case study SAP document management system. *2022 International Conference on Informatics, Multimedia, Cyber and Information System*, 77–81. https://doi.org/10.1109/ICIMCIS56303.2022.10017845

[5] Mittal, M. & Mittal, M. (2022). An electronic health record management system based on blockchain technology. *2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP)*, 285–290. https://doi.org/10.1109/ICFIRTP56122.2022.10059456

[6] Setianto, F. & Suharjito (2018). Analysis the acceptance of use for document management system using technology acceptance model. *2018 Third International Conference on Informatics and Computing (ICIC)*, 1–5. https://doi.org/10.1109/IAC.2018.8780462

[7] Nautiyal, N., Agarwal, P. & Sharma, S. (2023). Rechain: A secured blockchain-based digital medical health record management system. *2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT)*, 1–6. https://doi.org/10.1109/ICITIIT57246.2023.10068707

[8] Bhatlawande, S., Shilaskar, S., Gupta, D., Dupare, P. & Ghode, R. (2024). Automated Identity Document Classification. In: Sharma, H., Shrivastava, V., Tripathi, A.K., Wang, L. (eds) *Communication and Intelligent Systems*. ICCIS 2023. Lecture Notes in Networks and Systems, vol 969. Springer, Singapore. https://doi.org/10.1007/978-981-97-2082-8_30/

[9] Renjith, S., Manazhy, R. & Suresh, M.S.S. (2024). Recognition of Sign Language Using Hybrid CNN–RNN Model. In: Hassanien, A.E., Anand, S., Jaiswal, A., Kumar, P. (eds) *Innovative Computing and Communications*. ICICC 2024. Lecture Notes in Networks and Systems, vol 1021. Springer, Singapore. https://doi.org/10.1007/978-981-97-3591-4_2/

[10] Jayoma, J. M., Moyon, E. S. & Morales, E. M. O. (2020). OCR based document archiving and indexing using PyTesseract: A record management system for DSWD Caraga, Philippines. *2020 IEEE 12th International Conference on Humanoid, Nanotech., Information Techn., Communication and Control, Env., and Mgt.*, 1–6. https://doi.org/10.1109/HNICEM51456.2020.9400000