Transaction on Informatics and Data Science
Vol. 1, No. 2, 2024, 97-106
ISSN: 3064-1772 (online)
DOI: https://doi.org/10.24090/tids.v1i2.12281

Validation of New Student Registration Documents at Nurul Jadid University Using Convolutional Neural Network

Fathorazi Nur Fajri [*1], Gulpi Qorik Oktagalu Pratamasunu [1], Kamil Malik [1]

[1] Information Systems, Engineering Faculty, Nurul Jadid University, Paiton Probolinggo, Indonesia, fathorazi@unuja.ac.id

Article Information

Submitted March 12, 2024
Accepted June 8, 2024
Published October 1, 2024

Abstract

Every year, Nurul Jadid University admits new students by registering them using the website. Each prospective new student can fill in data independently and upload documents such as Deeds, Family Register, Identity Cards, Diplomas, and SKHU. Often, prospective new students need clarification in uploading documents; for example, the place for uploading ID cards is filled with uploading diplomas and vice versa. It causes the uploaded data not to match the place or group. Today, no document validation technique can match these types of documents. Therefore, a way is needed to overcome this problem. One way to recognize the document type is by its visual form or image. There are several methods for identifying an image, namely deep learning and neural network models. Where the convolutional neural network is known to be fast in processing data in images, this research aims to validate documents on new student registration data with a deep learning method, namely convolutional neural network (CNN). The experimental results show that the proposed method can classify the Nurul Jadid University new student registration documents with an accuracy rate of 0.91, such as the birth certificate at 0.97, diploma documents at 0.88, Family card documents at 0.88, identity cards at 0.84, exam result certificate with an accuracy 0.94.

Keywords: Convolutional neural network, document validation, image classification.

1. Introduction

Nurul Jadid University (UNUJA) is one of the pesantren-based universities in Probolinggo, Indonesia. It was established by three institutions, STT Nurul Jadid, STIKES Nurul Jadid, and IAI Nurul Jadid, under the auspices of the Nurul Jadid Islamic Boarding School Foundation. On October 29, 2017, UNUJA was established by the Minister of Research, Technology, and Higher Education, Mohamad Nasir.

Every year, UNUJA organizes a selection of new student registrations online, accessed through http://pmb.unuja.ac.id, as a new student admission information system that facilitates applicants/ prospective new UNUJA students to obtain information, register and manage data independently, and see the announcement of election results online.

The selection step for new student admissions at Nurul Jadid University includes the administrative selection stage of new student registration documents. New student registration documents (PMB) are one of the documents that must be collected as a condition of registering for prospective new university students. Each university should have its own PMB documents. Diplomas, Certificate of Examination Results, family cards, and birth certificates are categories of new student registration documents generally required by the university, including Nurul Jadid University. In admitting new students, Nurul Jadid University assist by the Bureau of General

* **Author Correspondence**: Fathorazi Nur Fajri: Nurul Jadid University, Karanganyar, Paiton Probolinggo, East Java – Indonesia. Email: fathorazi@unuja.ac.id.

Academic and Financial Administration (BAUAK).

The General Administration Bureau of Academic and Finance/ BAUAK of Nurul Jadid University is an institution to improve the quality of new student admission services at Nurul Jadid University. In the new student registration service, BAUAK classifies and selects PMB documents manually by downloading documents from the UNUJA PMB information system and then checking whether the documents sent are correct. The more the number of prospective new students who register, the longer the document grouping classification process. The problem occurs when new prospective students are mistaken when uploading the document. As is the case, the identity card document fills the place to upload the deed. It is due to the absence of document validation in the new student admission application. Therefore, an algorithm is needed that can validate documents automatically.

We can use several methods to validate image documents [1], [2], one of which is deep learning with the convolutional neural network method [3], [4], [5]. This method helps validate PMB files into the correct category. It is crucial to refer to some previous research so that it follows the study to be carried out; for that, the research must be related to prior research.

In 2017, space learning using a convolutional neural network developed by [6] to face recognition in real-time face recognition. CNN detects faces with the OpenCV library and MTech 5MP webcam. The data is collected from face images divided into two groups: external faces (good lighting). Tests using a CNN model with a depth of 7 layers with input from a spatial binary subtraction model followed by a radius of one and a neighbor of 15 showed that faces recognized in two paper frames had an accuracy of more than 0.89. The following research indicates that CNN models can be used for high accuracy in real-time face recognition [6].

CNN is also used for traffic sign recognition, such as M. Akbar's study of traffic signal recognition using CNN [7]. Visalini has published research about traffic sign recognition using convolutional neural networks at the International Conference on Innovative Mechanisms for Industry Applications [8]. The data is taken directly from geolocation using the Android application. The study does not mention the amount of data used, but the accuracy of using CNN to identify or recognize traffic signs is 0.85 to 0.90, with a 3-layer convolution method [8].

One of the most critical problems in computer vision is recognizing objects given computer images to identify and use several recognition algorithms [9]. The primary purpose of identification is to examine items in the image. In computer vision, there is machine learning, one of the neural network techniques that can classify patterns in known patterns that can be used to test images (known image patterns and patterns, not a picture) or identify photographs. One method that can be used for pattern recognition is deep learning methods, such as CNN [9].

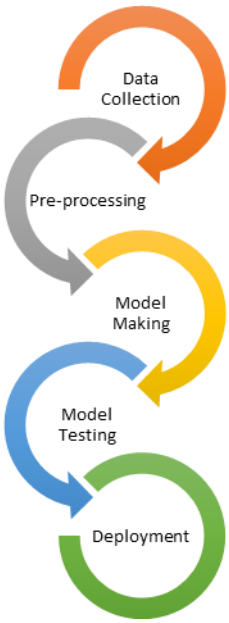
CNN can be used to validate new student registration data documents from PMB image patterns by looking at PMB image patterns with unique characteristics of each text group. With this research, CNN will be able to analyze new student document information accurately so that it can validate documents to help the process of completing the administration of new student registration at the Nurul Jadid University BAUAK Office in disseminating information.

2. Method

The research method for validating new student registration documents at Nurul Jadid University using a Convolutional Neural Network (CNN) can be systematically described by utilizing the structure as shown in Figure 1. Figure 1 illustrates a multi-stage, iterative process involving five interconnected circular phases, each representing a distinct step in the

research methodology. These stages can be aligned with typical machine learning workflow stages, ranging from data collection to model evaluation.

Figure 1.
Research method



The first and second phases in the methodology is data collection and preprocessing. The first step is data collection, where the data used in the following study is an image dataset from the New Student Registration (PMB) file taken directly from the General Administration and Finance Agency (BAUAK) of Nurul Jadid University. The next step is preprocessing to convert raw data into quality data. In preprocessing, the process of removing background and resizing is carried out. This aims to make the data more focused on the object, namely Identity Card (KTP), Family Card (KK), Diploma, Birth Certificate (akta), and Exams Result Certificate (SKHU), and to equalize the size of each data.

Preprocessing techniques are applied to standardize the format of these documents, such as resizing, normalizing image quality, and ensuring uniformity in document type and resolution. This step is crucial for preparing the dataset that will be used to train the CNN model.

As shown in Figure 1, model design and training in third stage, a Convolutional Neural Network (CNN) is constructed. The CNN architecture is designed to recognize patterns within the documents through layers of convolution, pooling, and activation functions. Training is conducted using the labeled dataset, where the model iteratively adjusts its parameters to minimize classification errors. Backpropagation and optimization algorithms, such as stochastic gradient descent, are applied to refine the model's ability to accurately classify new student registration documents.

The fourth phase is model validation and testing. After the CNN model has been trained, it undergoes a validation process using a separate portion of the dataset that was not involved in the training. The performance of the model is evaluated based on its ability to accurately classify unseen documents, and metrics accuracy is used to assess its effectiveness. Testing is conducted to ensure the model generalizes well to real-world data and can handle document variations that were not explicitly seen during training.

Furthermore, the modeling uses the CNN method because CNN has advantages in processing fast and accurate image data [10]

[11]. Then, the measuring instrument in model testing uses a confusion matrix. With the following formula.

Accuracy = (TP+TN) / (TP+FP+TN+FN) (1)

Precision=TP/ (TP+FP) (2)

Recall=TP/ (TP+FN) (3)

where

- True Positive (TP) is the object case to be positive, and true
- True Negative (TN) is the object case to be negative, and true
- False Positive (FP) is the object case to be positive, and false
- False Negative (FN) is the object to be negative, and false

The final phase, deployment and continuous improvement, involves the implementation of the CNN model into the actual document validation system at Nurul Jadid University. Once deployed, the model operates in real-time, assisting the administration by automatically validating student registration documents. Feedback loops are integrated into the system to allow for continuous learning and improvement. New data can be added to retrain the model, ensuring it adapts to evolving document standards and improves accuracy over time. This cyclical process of monitoring and updating the model ensures its long-term reliability in automating the validation of registration documents. After getting good model measurement results, the next step is implementing or deploying using Python and Flask programming languages at the deployment stage.

3. Result and Analysis

Data Collection

Several techniques can be used during data collection, such as web scraping or public data [12]. The dataset used in the CNN method is image data, as shown in Figure 2. The CNN model will work well when using a large amount of image data. So, the model will learn about the image.

Figure 2.
Sample Dataset AKTA, KK, KTP, Diploma

The data used in the following research is from images collected from the new student admission data archive at the Nurul Jadid University General Administration and Finance Agency (BAUAK). The image data used is the Nurul Jadid University New Student Registration document. The initial data at this stage is 100 images for the PMB document class of Birth Certificate, 100 images of PMB diploma documents, 100 images of PMB SKHU documents, 100 images of PMB KK documents, and 100 images of PMB KTP documents. Then, the initial data will be divided for training and validation with a ratio of 8:2.

Pre-processing Data

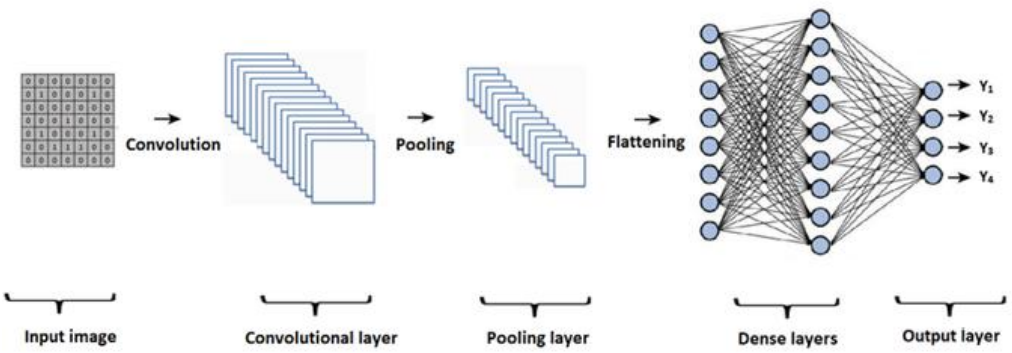
Pre-processing is the process of converting raw data into quality data by removing the background and resizing the image [13]. Then, divide the collected data into three parts, namely training data (data used for the training process), validation data (validation data from training data), and testing data (data used for testing).

Figure 3.
Background Remove Crop Process

Modeling

The CNN comprises the input, convolution, pooling, dense, and output layers. The architecture in this research looks like Figure 4, with five input layers, 32 convolution layers, 64 convolution layers, 64 convolution layers, pooling with a drop out of 0.3, 64 dense layers, and five output layers.

Figure 4.
Convolution Neural Network Architecture



From various experiments that have been carried out, at batch size 32, steps per epoch 13, epochs 50, validation steps 10, has a high level of accuracy and is balanced with the accuracy of validation data. In Figure 5, a graph of the accuracy value of the PMB document training data, it can be seen from the figure that the accuracy value at the initial iteration is low, and the longer the iteration, the higher the accuracy level. Figure 6 shows that the loss rate is high for the initial iteration, while the loss rate decreases along with the number of iterations. Based on the model built, the highest accuracy of the training data is 100%.

Figure 5.
CNN model accuracy graph for each iteration.

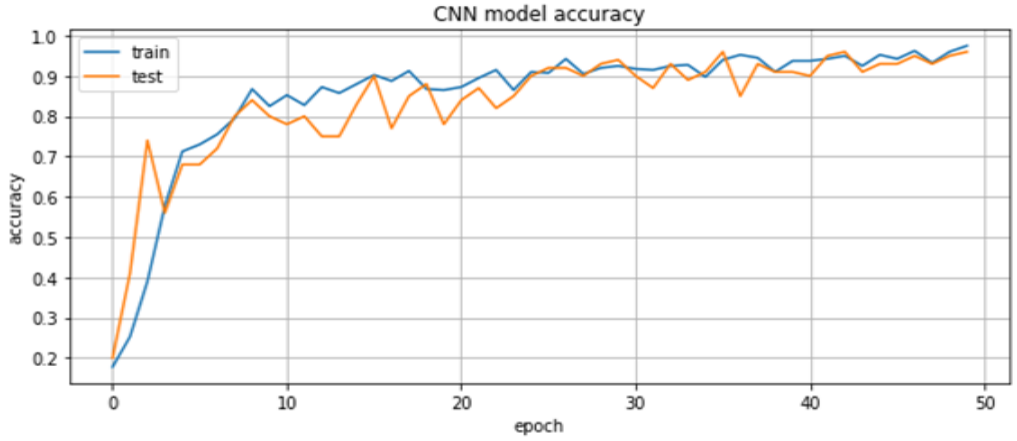
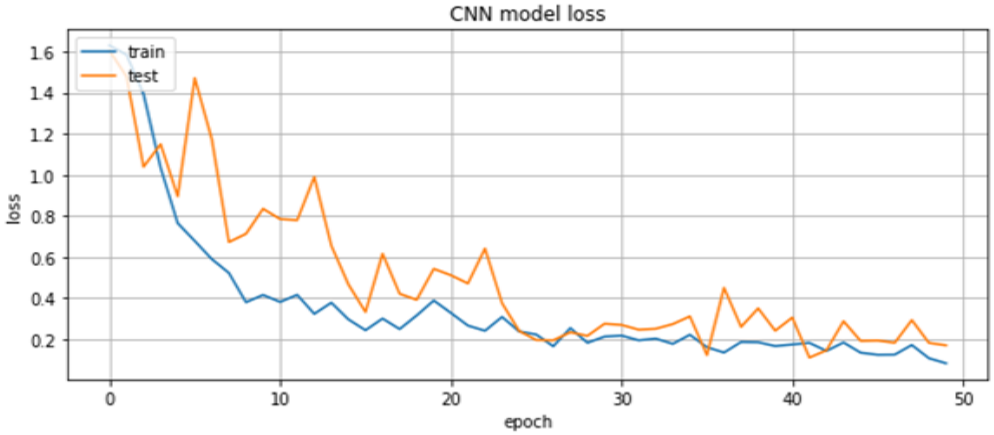


Figure 6.
CNN model loss graph for each iteration



Model Testing

Table 1 is the result of testing the model built. The amount of test data is 170 images of new student registration documents (PMB), which include 34 images of Birth Certificate documents, 34 images of Diploma documents, 34 images of Family Card documents, 34 images of KTP documents, and 34 images of Exam Result Certificate documents. Based on Table 1, fifteen PMB document image data and the accuracy obtained from the test data have been classified.

Table 1.
Accuracy per-class

No	Lable	Amount	True	False	Accuracy
1	Birth Certificate	34	33	1	0.97
2	Diploma	34	30	4	0.88
3	Family Card	34	30	4	0.88
4	Identity Card	34	30	4	0.88
5	Exam Result Certificate	34	32	2	0.94
	Total	170	155	15	

The results obtained from the classification of PMB UNUJA documents have an accuracy of 93.5%, precision of 92.9%, and recall of 90.5% using confusion matrix calculations.

$$Accuracy = (144+15)/(144+11+15+0)$$
$$= 159/(170)$$
$$= 0{,}935 \times 100\%$$
$$= 93{,}5\%$$
$$Precision = (144)/(144+11)$$
$$= (144)/(155)$$
$$= 0{,}929 \times 100\%$$
$$= 92{,}9\%$$

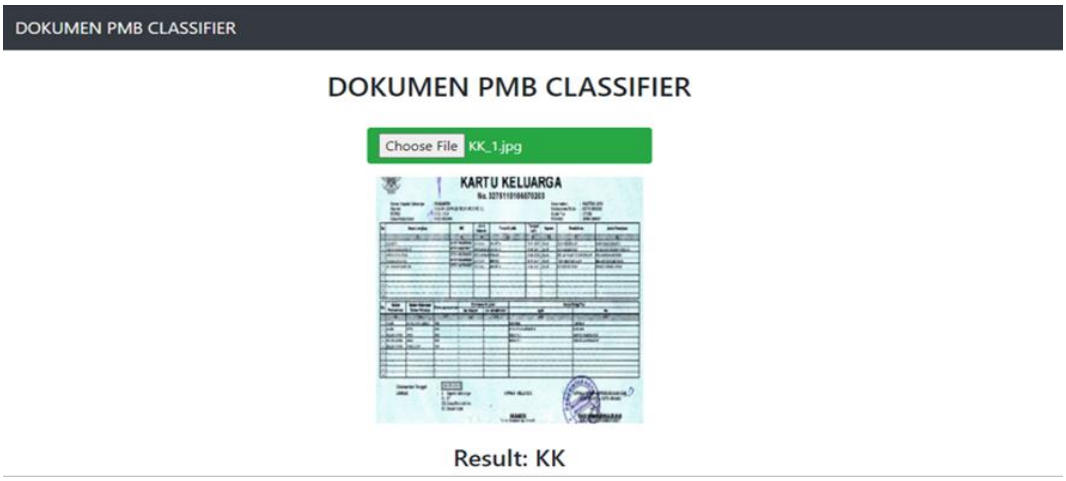
After analyzing and finding the desired results, the developed model is stored as H.5 to be processed for the web deployment step.

Deploy Model

The model H.5 obtained from the CNN implementation is then developed into a prototype web deployment that can be used to predict the classification of PMB document images at Nurul Jadid University. A front-end application can be integrated directly with the CNN model to perform deployment. Flask is a web framework that uses the Python programming language and is classified as a type of micro-framework.

The results of the deployment are shown in Figure 7, where there is a select file feature to upload the file further, then the application can validate the file type in each category, such as Family Cards, ID cards, Diploma, Birth certificate, and Certificate of Examination Results.

Figure 7.
Deploy on website.



By integrating this CNN model with an Application Program Interface (API) opens the possibility of scalability and flexibility [14]. By enabling API integration, the model can be embedded within various university systems, including admission portals, student databases, and even external systems used by other institutions. This capability can standardize the validation process across multiple platforms, enhancing the interoperability of systems within the university's digital infrastructure.

This research has the potential to influence other universities and institutions that manage large volumes of document verification tasks. The model could be adapted and applied to different use cases, such as verifying diplomas, transcripts, and other administrative documents. In doing so, it could contribute to improving the overall security and authenticity of academic documentation in higher education.

Furthermore, the architecture of the CNN model, with its carefully designed layers and

dropout strategy, suggests that similar machine learning models could be developed for other document-heavy sectors such as government services, financial institutions, and healthcare. The relatively high accuracy achieved with moderate computational resources (e.g., batch size of 32 and epoch 50) also implies that the model could be used efficiently in environments where computational power is limited.

4. Conclusion

This research can conclude that the validation of the Nurul Jadid University New Student Registration Document (PMB) using the CNN method can be applied. The model architecture used is five input layers, 32 convolution layers, 64 convolution layers, 64 convolution layers, 64 convolution layers, pooling with drop out 0.3, 64 dense layers, and five output layers using parametric epoch 50 with a combination of using batch size 32, steps per epoch 13 and validation steps 10. With the accuracy results from the trials that have been carried out, the accuracy value is 93.5%, precision is 92.9%, and recall is 90.5%, which shows that the accuracy level is perfect for validating documents.

Furthermore, the results of this research can be integrated with an Application Program Interface (API) so that it is easier to incorporate into several systems or applications.

The success of this project presents an opportunity for future research. Additional optimization techniques, such as hyperparameter tuning, could be explored to further improve the model's performance. Moreover, future studies could investigate the application of other deep learning architectures, such as recurrent neural networks (RNNs) or transformer-based models, to compare their effectiveness in document validation tasks.

Future work based on this research could focus on optimizing the CNN model through hyperparameter tuning and exploring alternative architectures like ResNet or transformers to enhance performance. Expanding the dataset with more diverse and imperfect document types would improve the model's robustness in real-world applications. Integration of Natural Language Processing (NLP) could allow for the verification of both visual and textual content in documents, while deploying the model in real-time using cloud or edge computing could increase processing efficiency. Additionally, future studies could explore fraud detection by training the model to recognize counterfeit or tampered documents, enhancing its security features. Seamless integration with other university systems, such as enrollment or financial aid, could automate document verification across departments. Researchers may also explore alternative deep learning models like RNNs or hybrid CNN-RNN models for more complex document types. Ensuring ethical compliance, particularly with data privacy regulations, could be another area of focus, possibly incorporating secure methods like federated learning. Cross-institutional collaboration to develop a generalized API for document validation could benefit other educational institutions, and future development could also enhance user interfaces to provide real-time feedback, improving overall usability. By pursuing these areas, the document validation system could be made more accurate, secure, and adaptable, benefitting a broader range of applications beyond just the university.

5. References

[1] Y. Jin *et al.*, "Image matching across wide baselines: From paper to practice," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 517–547, 2021.

[2] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "Layoutlm: Pre-training of text and layout for document image understanding," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 1192–1200.

[3] N. Audebert, C. Herold, K. Slimani, and C. Vidal, "Multimodal Deep Networks for Text and Image-Based Document Classification BT - Machine Learning and Knowledge Discovery in Databases," 2020, pp. 427–443.

[4] N. Ghanmi, C. Nabli, and A.-M. Awal, "CheckSim: A Reference-Based Identity Document Verification by Image Similarity

Measure BT - Document Analysis and Recognition – ICDAR 2021 Workshops," 2021, pp. 422–436.

[5] T. M. Ghazal, "Convolutional neural network based intelligent handwritten document recognition," *Comput. Mater. Contin.*, vol. 70, no. 3, pp. 4563–4581, 2022.

[6] M. Zufar and S. Budi, "Convolutional Neural Networks Untuk Pengenalan Wajah Secara Real-time," *J. Sains dan Seni ITS*, vol. 5, no. 2, pp. 2337–3520, 2016.

[7] M. Akbar, "Traffic sign recognition using convolutional neural networks," *J. Teknol. dan Sist. Komput.*, vol. 9, no. 2, pp. 120–125, Apr. 2021, doi: 10.14710/jtsiskom.2021.13959.

[8] S. Visalini and R. Kanagavalli, "A Comprehensive Survey of Pneumonia Diagnosis: Image Processing and Deep Learning Advancements," in *2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Dec. 2023, pp. 734–742, doi: 10.1109/ICIMIA60377.2023.10426403.

[9] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[10] A. W. Salehi *et al.*, "A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope," *Sustainability*, vol. 15, no. 7, p. 5930, Mar. 2023, doi: 10.3390/su15075930.

[11] Y. Liu, H. Pu, Q. Li, and D.-W. Sun, "Discrimination of Pericarpium Citri Reticulatae in different years using Terahertz Time-Domain spectroscopy combined with convolutional neural network," *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.*, vol. 286, p. 122035, Feb. 2023, doi: 10.1016/j.saa.2022.122035.

[12] M. E. Laily, F. N. Fajri, and G. Q. O. Pratamasunu, "Deteksi Penggunaan Alat Pelindung Diri (APD) Untuk Keselamatan dan Kesehatan Kerja Menggunakan Metode Mask Region Convolutional Neural Network (Mask R-CNN)," *J. Komput. Terap.*, vol. 8, no. 2, pp. 279–288, Dec. 2022, doi: 10.35143/jkt.v8i2.5732.

[13] N. ŞENGÖZ, T. YİĞİT, Ö. ÖZMEN, and A. H. ISIK, "Importance of Preprocessing in Histopathology Image Classification Using Deep Convolutional Neural Network," *Adv. Artif. Intell. Res.*, vol. 2, no. 1, pp. 1–6, Feb. 2022, doi: 10.54569/aair.1016544.

[14] M. Gheisari *et al.*, "Deep learning: Applications, architectures, models, tools, and frameworks: A comprehensive survey," *CAAI Trans. Intell. Technol.*, vol. 8, no. 3, pp. 581–606, Sep. 2023, doi: https://doi.org/10.1049/cit2.12180.

This page is intentionally left blank