Grenze International Journal of Engineering and Technology, June Issue

An Integrated OCR–NLP Framework for Printed and Handwritten Text Recognition

Ravi P[1], Y. H. Sharath Kumar[2], Thejashwini M A[3], Thanushree S R[4], Sonashree M.S[5] and Nuthan Mourya N[6]
[1-6]Department of Information Science and Engineering, Maharaja Institute of Technology Mysore, Belawadi Mandya, Karnataka, India
Email: ravipbympr@gmail.com, sharathyhk@gmail.com, thejashwinima03@gmail.com, thanusrira654@gmail.com, mssonashree@gmail.com, nuthanmouryan2004@gmail.com

***Abstract*—Optical Character Recognition (OCR) systems often encounter significant challenges when converting handwritten and degraded texts into digital format, primarily due to noise, inconsistent writing styles, and low-quality inputs. To overcome these issues, this research work introduces an integrated OCR–NLP framework that integrates OCR processing model and language model post-processing and correction. This proposed system applied DenseNet-121 for text-type classification, Tesseract OCR for printed text extraction, and OCR Space for handwritten text extraction. This extracted text is further refined by using PHI-3 language method for grammar correction, structural and semantic inconsistencies. Experimental result analysis on a mixed dataset of printed text and handwritten demonstrate an 85% reduction in grammatical errors and enhanced text readability from 69.2 to 91.4 on the Flesch scale. These results confirm that the proposed integrated OCR–NLP framework significantly improves recognition accuracy and produces contextually accurate output suitable for real-world text digitization applications.**

***Keywords*— OCR, NLP, DenseNet-121, PHI-3 , Post-Processing, Text Digitization.**

I. INTRODUCTION

Optical character recognition (OCR) is a technology to extract the text from the images such as scanned documents, handwritten images. OCR is very much helpful and benefited because the text documents requires less space than the image files, another important benefit is that text documents are easily copied altered and manipulated. OCR is one of the important application of AI importantly when it comes to Natural Language Processing (NLP). OCR has many applications most common application are home utility meter readings, smart security and surveillance systems, license plate recognition, and digitalization of older printed documents. the other benefits of OCR includes making the things more secure and reducing crime using automated text recognition. OCR works based on the analyzing the shape and structure of characters in cubes of images to identify the text. OCR particularly well in identifying the printed text especially in license plate and documents, but it still struggles with the handwritten text and performance vary the accuracy compared to printed documents handwritten documents has less accuracy as the handwritten style varies from person to person which creates the problem for the NLP applications to do text summarization, part of speech, sentence boundary detection , topic modeling, and named entity recognition(NER), including sorting all kinds of names, groups and places. The lack of accuracy in OCR may even allow distinct entities to be misidentified as same entity.

Grenze ID: 01.GIJET.12.2.288_21

So to overcome this problem is NLP as a post processing approach will improve the accuracy of the OCR which reduces the error raised by the OCR models. The eventual aspiration is to establish a complete pipeline that includes OCR and post-processing with NLP to improve text recency across the variety of applications.

II. Related Work

The paper addresses the process of converting both printed and handwritten text into editable digital form by applying the deep learning models as CNN,LSTM and NLP models. The study tells that result in better accuracy when evaluated against different types of dataset and can be evaluated from different domains such as education, healthcare, and finance[1]. This paper addresses the three phase OCR and NLP system investigated recovers data from old academic cards by using AI models such as Chatgpt as a error correction model. In this study the model resulted CER of 2.15% and WER of 7.05%, which is a very good result[2]. This system recognizes the Tamil character using the methods HOG and VGG19 models. This study shows that VGG19 is the best method but also improving recognition of ancient language scripts. This will help in digitalization and preservation of the old handwritten manuscripts[3]. This paper addresses the security of the OCR system on protecting sensitive text data regions of the scanned documents by encrypting the texts. This approach uses MSER text recognition detection system, AES encryption system, and NLP to secure the secret data. This reduces the processing burden while balancing data privacy and security[4].OCR and NLP are used together for analyzing power grid project documents. This automation extracts useful insights from textual and scanned data for better decisions. It improves efficiency, identifies risks, and enhances evaluation accuracy[5].This survey highlights the lack of Pashto language resources in NLP research. It reviews existing tools and calls for collaborative efforts to develop more Pashto datasets. The goal is to digitally preserve the language and promote cultural heritage[6]. Introduces a deep learning approach using RoBERT a for correcting OCR errors in medical reports. Achieves 81% accuracy without needing large domain-specific datasets. It simplifies document management and reduces manual healthcare workload[7]. Develops a unified system for text extraction and plagiarism detection using OCR and NLP. Combines preprocessing, TF-IDF, n-grams, and BERT models for accurate results. The framework enhances scalability and precision in handling image-based text[8]. Proposes an automated system to evaluate handwritten answer sheets using OCR and NLP. It compares student responses with model answers via semantic similarity. The system ensures fair grading and integrates easily with learning platforms[9]. Applies OCR and NLP to extract quality metrics from colonoscopy reports automatically. The approach reduces manual work and speeds up medical data analysis. It improves patient outcomes and supports large-scale healthcare research[10]. It concentrates on developing an AI-based system that automatically analyzes handwritten exam answers using OCR, deep learning, and NLP. The system turns handwriting to digital text, it reads and checks what students written, helping teachers save time by , keep grading fair, and give useful feedback even though messy handwriting and tough answers can still cause problems[11]. It focuses on Collecting information from medical report images. Images were converted to a standard format and checked for quality, then text was extracted using OCR tools. Key details were Checked by hand, and Named Entity Recognition Arranged the data. AI techniques, including a genetic algorithm and a neural network, were applied to improve OCR performance, creating an smart system for more reliable information extraction[12]. This study presents a method for Automated license plate identification. A specially created dataset of 300 images from Moroccan websites was created and Hand-labeled. YOLO v7 detected and extracted license plates, which were cleaned with image cleaning techniques to make the characters easier to read OCR tools—EasyOCR for Latin and Arabic OCR for Arabic—were used to identify the text. The system achieved high performance, of about 99% accuracy and a 98% F1-score in character recognition[13]. It addresses on correcting noisy text from OCR systems, especially in ancient or damaged records, by introducing a post-OCR pipeline for error detection, classification, and correction, analyzing manual, semi-automatic and automatic approaches, surveying relevant datasets and testing metrics like CER, WER, precision, recall, and F1-score, and showing recent trends such as neural models like BERT and NMT, encouraging standardization and collaboration to advance post-OCR research[14].This document provides a detailed pipeline to enhance OCR text correction using NLP. The datasets used to validate the PP-OCR and TrOCR approaches were Born-Digital Images, Scene Text, License Plates, and Handwritten Text, where the handwriting and printing were studied in isolation to address plausibly and evaluate the correctness of recognition, applied NLP models (ByT5, BART, Alpaca-LORA) to correct OCR errors based on their respective approaches with decreased costs (CER and WER). The method increases the text quality and reliability , it shows that it works well for different kinds of real life data overall it is an efficient approach for OCR post processing across different types of text [15]. This method extracts characters using OCR and then applies NLP based post processing step for Indonesian ID cards(KTP).There are three OCR libraries(PYOCR , Pytesseract , and

TesseractOCR) ,these were tested and among them Pytesseract performed the best on both scanned and photos. The NLP technique then corrected the extracted text, resulting in an F-score of 0.78 and an average processing time of 4510 ms per ID card [16].This study purposes a method for extracting text from unstructured sources. The IAM and A-Z datasets were used for training. The model combines OCR and NLP ,using a CNN for the OCR component and a t5 model for text correction .The t5 model was trained using the JFLEG datasets Combining OCR and NLP decreases mistakes and spelling mistakes, generating cleaner and more accurate text output. This approach has applications in converting historical documents into digital format[17]. This paper discusses a method to extract essential medical data from scanned health records. OCR and machine learning models, including Clinical BERT, were used to analyze the text. The focus was on sleep study reports, identifying Apnea-Hypopnea Index (AHI) and oxygen saturation (SaO_2). Accuracy improved when layout information was used. It shows that combining image processing with smart language models enhances data extraction from scanned medical documents[18].

III. Proposed Approaches

The proposed model describes a complete Optical Character Recognition (OCR) and Natural Language Processing (NLP) post-processing framework that can accurately pull out and improve text from both handwritten and printed images. The proposed system does away with the need for segmentation by treating the whole document as one image input. Figure 1 shows the overall structure.

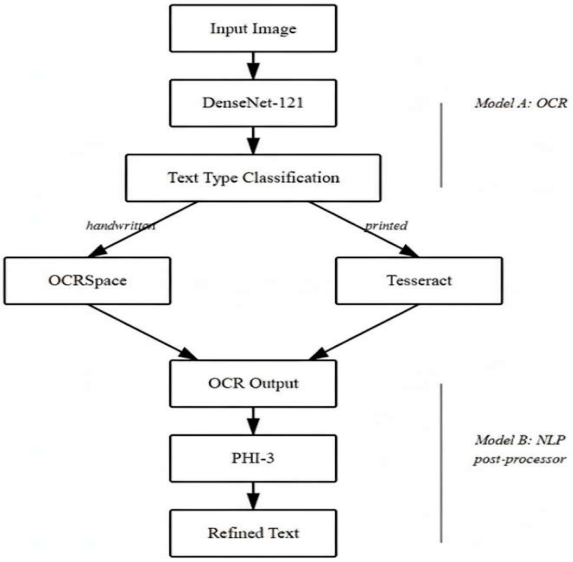

Fig 1: Workflow of the proposed integrated OCR–NLP Framework

The first approach incorporates the use of two OCR engines for the extraction of text from images. The selection of the engine is dependent on the predicted output of a text type classifier. The process of processing an image containing handwritten or printed text contains steps to convert the image to grayscale and then normalized to improve contrast and reduce background noise. These movements will result in better functioning of both successive classifiers and the OCR engines regardless of lighting and quality. DenseNet-121 is utilized to classify between text types, whether handwritten or printed. The architecture contains 121 densely connected layers, which helps maintain strong gradient flow, improves feature reuse, and keeps the overall number of parameters low. In the final stage, the network's softmax layer produces two probability scores —one for handwritten text and one for printed text— which allows the system to direct the image to the appropriate OCR engine. The process of recognizing text have two different paths depending on the output predicted by the classifier. The handwritten text is processed using OCRSpace. The OCRSpace is a cloud-based OCR engine that has the capacity to read variety of handwriting styles, ink densities, and background textures. The printed text is processed using Tesseract. Tesseract is an OCR engine which is known for its strength in printed text recognition.After recognition, the second step focuses on cleaning and improving the OCR output. This involves

using a grammar-aware NLP model to make the text clearer and more consistent. The PHI-3 model verifies the grammar correction,it also checks punctuation and spelling mistakes, this is shown in OCR generated text. Sentence restructuring refers to changing cluttered text into clear, readable sentences. Context preservation ensures that even after these corrections, the original meaning and intent of the document remain intact. To quantitatively evaluate an OCR model's performance, common metrics such as Character Error Rate (CER) and Word Error Rate (WER) are used . The CER calculates a ratio of the number of incorrect characters to the total number of characters in the ground truth text.It is defined as:

$$CER=(S+D+I)/N$$

Where
S: Denotes the number of character substitutions.
D: Denotes deletions.
I: Denotes insertions.
N: Denotes the total number of characters in the reference text.
WER estimates how many words have been recognized incorrectly with respect to the total words of the ground truth text and is given by:

$$WER=(S+D+I)/N_w$$

Where
N_w: indicates the total number of words in the ground truth.
Smaller values of both CER and WER correspond to higher OCR accuracy. These two metrics together describe the character and word-level reliability of the extracted text. The performance of the system is checked by using different datasets, including: Handwritten Datasets that include scanned handwritten notes, cursive and non-cursive scripts, and various writing styles. The scanned printed datasets include book pages, printed reports, and administrative forms representing variations of resolution and noise. This varied dataset selection ensures the model will generalize well across a variety of real-world scenarios.

IV. EXPERIMENTATION

A. The proposed integrated OCR–NLP Framework for printed image

The workflow of the proposed system for printed text recognition and enhancement is presented in Figure 2. The whole process starts by taking an image with printed text as input, followed by passing it through two main stages OCR and (NLP) post-correction.

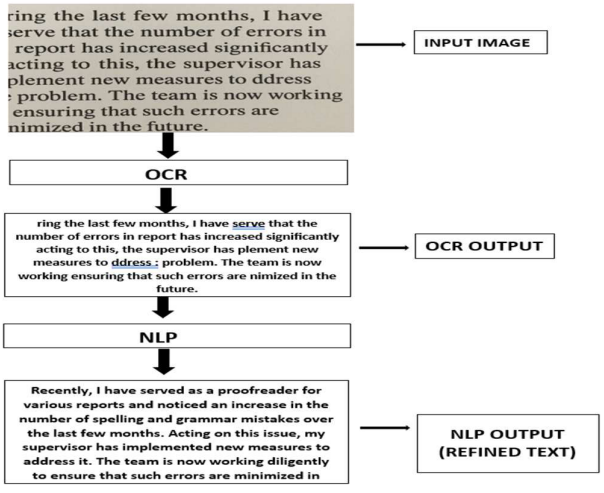

Fig 2: Example of the proposed integrated OCR–NLP Framework for printed image

Figure 2 illustrates the working process of the integrated OCR–NLP framework designed for printed text. In the first phase, the scanned document is carried out using the Tesseract OCR engine, which reads the visual forms of characters with the help of its LSTM-driven identification model. This permits Tesseract to convert the image

data into digital text. even though the initial output is usually close to the original printed text, it may still include small errors including spelling mistakes, missing characters, or incorrect spacing. During the later phase, the text produced by the OCR system is enhanced using the PHI-3 Natural Language Processing model. This language-processing unit automatically improves grammar, inserts the missing words, modifies sentence arrangement when necessary. The integrated process of OCR and NLP confirms that the final output is not only highly accurate in identifying the printed text but also more refined, readable, and more natural-sounding in the final output. Figure 2 shows that the output of the NLP stage is fluent and free from errors, thus accurately replicating the printed text.

B. The proposed integrated OCR–NLP Framework for handwritten image

The figure 3 shows summarize the complete process for the recognition and enhancement of handwritten text with our proposed system, step by step. The process starts with the input image, which is a photo of handwritten text. The image is processed through the OCR engine, which is an OCR model based on the cloud, specifically made for detecting and converting handwritten text to machine-readable text. In the previous step, the system has produced the raw output of the OCR portion, which means it processed the handwritten text image, extracted the text, and produced and saved the first draft OCR output. However, as we see in the example given, even the first draft has spelling mistakes, missing words or just incomplete phrases—this is when the OCR engine is affected by the challenges of response to varying handwriting styles, differences in human abilities, and image quality. To overcome the limitations observed in the OCR phase , the output generated by the recognition systems is passed to the Phi-3 Mini NLP model for further enhancement. This model evaluates the extracted text, detects grammatical errors and transforms incomplete or unclear sentences into more coherent ones, more meaningful language. By correcting swapped words, reorganizing sentence structures, and making the context clear again. the NLP layer produces text that is substantially more readable and closer to the original human-written content. Upon evaluation using a mixed dataset containing both printed and handwritten samples, the integrated OCR–NLP pipeline showed a clear improvement in accuracy.

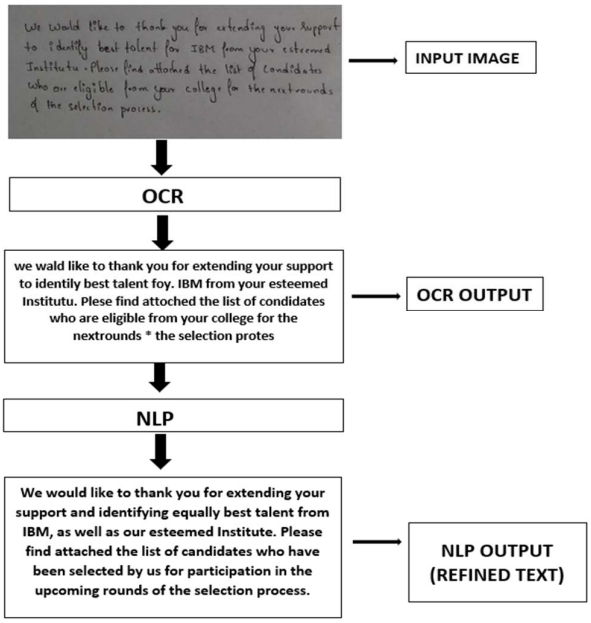

Fig 3: Example of the proposed integrated OCR–NLP Framework for handwritten image

Dataset used in pretrained model uses the wide range of variety of images of both text-images and handwritten images dataset to train the model. It covers the different types of fonts and document structure to ensure the robust in text recognition. The training process relies on synthetic and real world datasets, including high-resolution scan of books , newspapers, administrative documents and printed corpora. This diverse combination of clean and degraded printed datasets enables Tesseract to achieve high accuracy on structured, font-based textual images. Dataset used for testing the model two distinct types of datasets is used such as printed samples

and handwritten samples. The printed dataset consists of document-style images containing structured fonts, capitalized text, punctuation, and line-based formatting, including both clean page scans and moderately degraded text with misspellings, spacing inconsistencies, and minor distortion. The handwritten testing dataset included images of manually written sentences exhibiting natural variations in handwriting style, stroke width, spacing, and alignment—conditions known to produce higher OCR errors.

The DenseNet-121 classifier attained an accuracy of 94.8% across 2,000 test images. After implementing the NLP correction phase, the system recorded a significant decrease in both Character Error Rate (CER) and Word Error Rate (WER). CER dropped from 0.092 to 0.039, while WER decreased from 0.183 to 0.082, indicating an improvement of over 50% in both metrics In addition, the Phi-3 model reduced grammatical issues by nearly 85%, and the overall readability score increased from 6.92 to 9.14, showing a marked enhancement in the readability of the text.

TABLE I: COMPARISON SUMMARY WITH STATE-OF-THE-ART OCR MODELS

System	Performance Summary	Limitation	Advantage of Proposed Framework
Google Vision, Azure OCR, Textract, PaddleOCR, EasyOCR	Good accuracy on printed text	Struggle with handwritten and noisy inputs	Lower CER/WER and better readability after PHI-3 post-processing
Proposed OCR–NLP Framework	High accuracy on both printed and handwritten text	less	Robust output with improved grammar, structure, and context

V. CONCLUSION

The proposed framework effectively enhances text recognition for both printed and handwritten documents. By using DenseNet-121 with Tesseract and OCRSpace to classify the text and extraction, this system achieves reliable multi-format recognition. The PHI-3 NLP model significantly reduces grammatical errors and improves sentence structure and ensures context accuracy. Experimental results show performance gains, with CER reduced from 0.092 to 0.039, WER from 0.183 to 0.082, and readability increasing from 6.92 to 9.14. These improvements confirm that the integrated OCR–NLP framework produces clearer and more readable text. Overall, this approach offers a practical and efficient solution for real-world digitization, archival work and automated document processing. Future extensions may include multilingual support, real-time processing, and integration with advanced transformer-based vision models to further improve performance across larger document domains.

REFERENCES

[1] Adenekan, T. K., "Advancing Text Digitization: A Comprehensive System and Method for Optical Character Recognition," June 2024. Available: https://www.researchgate.net/publication/387271086

[2] Casas-Huamanta, E. R.; Pinedo, L.; Barbachán-Ruales, E. A.; Cárdenas-García, Á.; Rossel-Bernedo, L. A.; and Seijas-Díaz, J. G., "Optical character recognition system with natural language processing for data recovery on scanned old academic card reports," Acta Scientiarum. Technology, vol. 47, e69814, 2025, doi: 10.4025/actascitechnol.v47i1.69814.

[3] Devendiran, S.; Jyothiratnam; Raman, R.; and Nedungadi, P., "Handwritten Text Recognition using VGG19 and HOG Feature Descriptors," 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), 2024, pp. 1–6, doi: 10.1109/IATMSI60426.2024.10502872.

[4] Ghafoor, A.; Kayani, M.; and Riaz, M. M., "Multi-modal text recognition and encryption in scanned document images," Journal of Electronic Imaging, Springer, 2022. Accepted: 24 Sept. 2022; Published online: 9 Dec. 2022.

[5] Huang, J.; Jin, L.; Wang, Y.; and Wang, Y., "Intelligent analysis and application of NLP and OCR technologies in power grid project evaluation," Int. J. New Dev. Eng. Soc., vol. 7, no. 9, pp. 8–12, 2023, doi: 10.25236/IJNDES.2023.070902.

[6] Khan, Z. A.; Xia, Y.; Khaliq, F.; Khan, J. A.; and Khan, N., "From Tradition to Technology: A Systematic Survey on Navigating Pashto in Modern NLP," SSRN, Jan. 2024. Available: https://ssrn.com/abstract=5031721. doi: 10.2139/ssrn.5031721.

[7] Karthikeyan, S.; Seco de Herrera, A. G.; Doctor, F.; and Mirza, A., "An OCR Post-correction Approach using Deep Learning for Processing Medical Reports," IEEE Transactions on Circuits and Systems for Video Technology, accepted for publication, 2021, doi: 10.1109/TCSVT.2021.3087641.

[8] Kumar, P. S., and Prasad, K., "A Unified Framework for Text Extraction and Plagiarism Detection in Image-Based Content Using OCR and NLP," Cuestiones de Fisioterapia, vol. 54, no. 1, pp. 132–141, Jan. 2025, doi: 10.48047/CU/54/01/132-141.

[9] Kulkarni, M.; Adhav, G.; Wadile, K.; Deshmukh, V.; and Chavan, R., "Digital Handwritten Answer Sheet Evaluation System," International Journal of Computer Applications, vol. 186, no. 16, pp. 1–5, Apr. 2024.

[10] Laique, S. N.; Hayat, U. H.; Sarvepalli, S. S.; Vaughn, B. V.; Ibrahim, M. I.; McMichael, J. M.; Qaiser, K. N.; Burke, C. B.; Bhatt, A. B.; Rhodes, C. R.; and Rizk, M. K., "Application of optical character recognition with natural language processing for large-scale quality metric data extraction in colonoscopy reports," Gastrointest Endosc., vol. 93, no. 3, pp. 750–757, Mar. 2021, doi: 10.1109/j.gie.2020.08.038

[11] Mahalingam, M. S.; Kumar, N. S.; Reddy, B. S. C.; Reddy, N. V. S. K.; Govardhan, A.; and Bontha, H. S., "Optical Character Recognition based Answer Script Correction using Deep Learning and Language Processing Algorithms," 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), 2024, pp. 1–6, doi: 10.1109/ICUIS64676.2024.10866861.

[12] Malashin, I. M.; Masich, I. M.; Tynchenko, V. T.; Gantimurov, A. G.; Nelyub, V. N.; and Borodulin, A. B., "Image text extraction and natural language processing of unstructured data from medical reports," Machine Learning and Knowledge Extraction, vol. 6, pp. 1361–1377, 2024. doi: 10.3390/make6020064.

[13] Moussaoui, H.; El Akkad, N.; and Benslimane, M., "License plate text recognition using deep learning, NLP, and image processing techniques," Science of Information and Computing, vol. 1, no. 1, pp. 1–10, 2024. Available: https://www.iapress.org/index.php/soic/article/view/1966.

[14] Nguyen, T. T. H.; Jatowt, A.; Coustaty, M.; and Doucet, A., "Survey of Post-OCR Processing Approaches," ACM Computing Surveys, vol. 54, no. 6, Article 124, pp. 1–37, July 2021. Available: https://doi.org/10.1145/3453476.

[15] Rakshit, A.; Mehta, S.; and Dasgupta, A., "A novel pipeline for improving optical character recognition through post-processing using natural language processing," 2023 IEEE Guwahati Subsection Conference (GCON), 2023. doi: 10.1109/GCON58516.2023.10183509.

[16] Rusli, F. M.; Adhiguna, K. A.; and Irawan, H. I., "Indonesian ID card extractor using optical character recognition and natural language post-processing," 2021 9th International Conference on Information and Communication Technology (ICoICT), 2021, pp. 1–6, doi: 10.1109/ICoICT52021.2021.9527510.

[17] Singh, Anuj; Jangra, Suraj; and Aggarwal, Gaurav. “EnvisionText: Enhancing Text Recognition Accuracy through OCR Extraction and NLP-based Correction.” 2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2024, pp. 47–51. doi: 10.1109/CONFLUENCE60223.2024.10463478.

[18] Sultana, R.; Malagaris, I.; Hsu, E.; Kuo, Y.-F.; and Roberts, K., "Deep learning-based NLP data pipeline for EHR-scanned document information extraction," JAMIA Open, vol. 5, no. 2, pp. 1–12, 2022, doi: 10.1093/jamiaopen/ooac045.