# DocSplitAI: A Deep Learning Approach for Document Segmentation in Large PDFs

Merxhan Bajrami[1,2], Andrea Kulakov[1], Yvonne Gaissmaier[2] and Petre Lameski[1]
[1] *Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, N. Macedonia*
[2] *Laigo GmbH, Eckenerstraße 65, Friedrichshafen, 88046, Germany*

merdjan.bajrami@laigo.ai, andrea.kulakov@finki.ukim.mk,
yvonne.gaissmaier@laigo.ai, petre.lameski@finki.ukim.mk

***Abstract*—In many industries, organizations often face the challenge of managing batch of more documents merged into a single file. This can lead to difficulties in identifying where each individual document begins and ends, making document processing a time-consuming and error-prone task. For example, in many businesses, invoices are received in large batches and need to be processed quickly and accurately. This can be time-consuming and error-prone to manually split a large document containing multiple invoices into individual files. In legal and financial sectors, large volumes of documents such as contracts, invoices, and receipts can be merged together, leading to difficulties in managing and processing the documents.**

**To address this challenge, we propose a binary classification approach using the Donut [1] model, which is an OCR-free model that can learn to recognize patterns and features in the data without relying on optical character recognition. Our approach involves fine-tuning the model on a dataset of 5527 files, manually labeled into new document and same document classes. We developed a new methodology for creating the dataset that ensures a well-balanced distribution of examples for each class, and carefully selected hyperparameters to optimize the performance of the model.**

**Our results demonstrate that our approach achieved an accuracy of 0.89, an f1 score of 0.92, precision of 0.87, and recall of 0.93. These results suggest that our proposed approach is highly effective in identifying individual documents within merged PDFs, which has significant implications for a range of industries. For instance, in the legal sector, our approach could help to automate the process of document separation, making it easier for lawyers to manage and process large volumes of legal documents. In the financial sector, the approach could help to streamline the processing of invoices, receipts, and other financial documents**

***Index Terms*—page stream segmentation, section segmentation, Donut, intelligent document processing, deep learning**

## I. Introduction

It is a difficult challenge in many sectors to deal with huge file that are documents which comprise many papers connected together without a clear division between them. For instance, organizations in accounting, finance, and legal frequently manage several papers that are combined. The necessity for an automated solution comes from the time and error involved in manually separating them. Because there are no distinct borders between papers, it might be difficult to solve this issue because workers must frequently manually mark the beginning and end of each page. The papers may also differ in their layouts, formats, and structures. Automatic document segmentation may greatly increase the accuracy, efficiency, and effectiveness of data analysis.

In this study, we fine-tune a deep learning (DL) model for binary classification that can automatically discriminate between new and ongoing pages in a big PDF file. We have adopted the Donut[1] model, which is a sequence and OCR-free model, to solve the document segmentation problem. Using confidential internal data, such as invoices, receipts, and letters, we have fine-tuned the model. In our binary classification method, the AI model output is 1 every time a new document starts and 0 for a continuous page of the same document. Our approach is trained on a dataset of 5,527 files, collected from various private companies in Germany, and consisting of invoices, receipts, and letters in the German language. To guarantee the optimal data distribution, we give the dataset distribution based on a real world scenario. Lastly, for evaluation of our model we use metrics like accuracy, F1 score, precision, and recall to assess the performance of our model.

## II. Related work

The problem of page segmentation in large files that have batch of merged documents, has been a topic of research for many years, and various approaches have been proposed to address this challenge. In this section, we review some of the related work in the field of document segmentation using deep learning techniques.

In recent years, deep learning models have shown promising results in document segmentation tasks. For example, the use of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks have been explored for document segmentation tasks [2]. A hybrid model that combines CNN and LSTM has also been proposed for segmenting handwritten documents [3].

In the literature, unsupervised methods have been proposed for document segmentation. For example, clustering-based methods have been proposed to segment documents based on visual similarity between pages [4]. Hierarchical clustering-based methods have also been proposed to segment documents based on the similarity between regions within pages [5].

Moreover, some researchers have proposed using topic modeling techniques to segment documents based on the content of the pages [6].

Another approach to document segmentation is based on the use of traditional image processing techniques. Morphological operations, such as erosion and dilation, have been used to segment documents based on the layout of the pages [6]. Edge detection algorithms have also been used to segment documents based on the edges of the pages [7].

In conclusion, there are various approaches to document segmentation, including deep learning models, unsupervised methods, traditional image processing techniques, and hybrid methods. The related work on the Page Stream Segmentation task reveals that deep learning approaches have outperformed conventional machine learning approaches, with the best-performing model utilizing both textual and visual features. However, limitations of previous works include the lack of more diverge testing on public datasets, and a narrow focus on legal documents. Furthermore, it can be observed that all of the deep learning approaches for the PSS task rely on OCR-dependent models, and there has been no exploration of OCR-free methods. In this paper, we have proposed a DL-based approach for document segmentation that uses the Donut [1] model, which is a sequence and OCR-free model, fine-tuned on confidential internal data. The proposed methodology achieves high accuracy and F1 score, demonstrating its effectiveness in solving the page segmentation problem.

## III. Dataset

The dataset used in this study was collected from different private companies in Germany, containing confidential invoices and receipts that are in German language. The dataset was provided by Laigo GmbH, the company where the model was developed and fine-tuned. The dataset comprises 5,527 files, which were manually labelled into two categories: new documents and continuous pages of the same document. Our dataset consists of 5,527 files, of which 4,637 files were internal data, and 890 files were Tobaco800 [11] data, which were letters. The dataset was manually labeled into two categories: new document and same document. To create a balanced distribution of the dataset, we used simple random sampling technique. The resulting distribution for our dataset was as follows: 4,334 were in new documents and 1,193 continuous pages of the same document. Furthermore, the dataset was categorized into three document types, including invoices, receipts, and letters.

At Fig1. are visualized the distribution of the document types in our dataset, into the output classes: new document and continuous page. Of the 2,092 invoices, 1,751 were new documents and 341 were continuous pages of the same document. Of the 2,545 receipts, all were new documents. Finally, of the 890 letters, 541 were continuous pages of the same document, and 349 were new documents.

At Fig2. it is represented the visualization of data distribution into train, validation and test sets, regarding the document type. As shown on the visualization, our model was trained

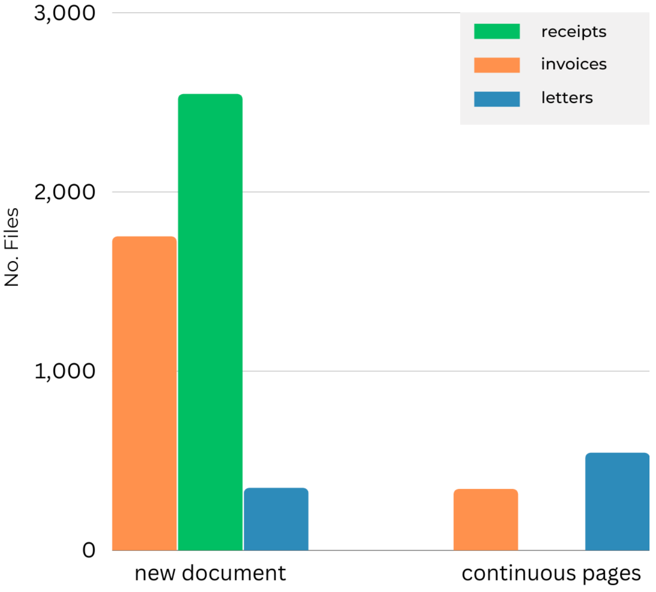

Fig. 1. Distribution of document types into classes: new documents and continuous page

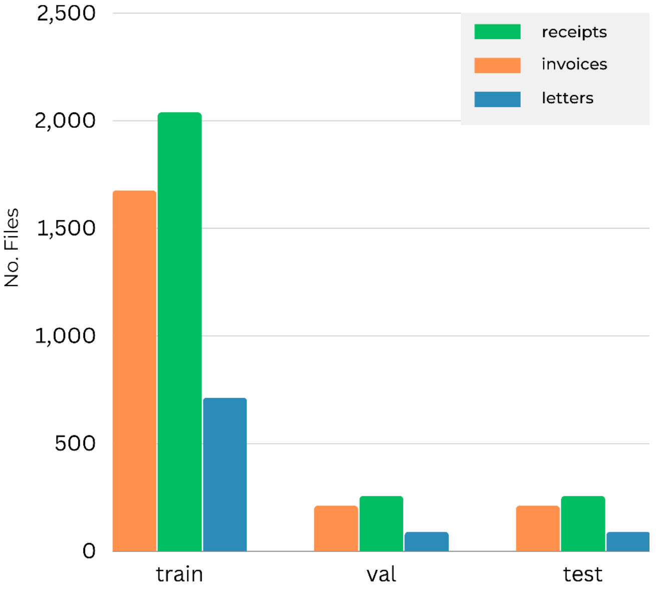

Fig. 2. Distribution of document types into train, validation and test sets

Copyshop Zentrale
Inh. Stavros Lalisidis
Adalbertstr. 43
80799 München
Tel. 089 / 2712313

DATUM 06/05/2019 MON ZEIT 13:16

| | |
|---|---|
| 12X @ 0.15 | |
| S/W - Kopien T1 | €1.80 |
| BRUTTO1 | €1.80 |
| STEUER 1 SATZ | 19.000% |
| NETTO1 | €1.51 |
| Mwst 19% | €0.29 |
| SUMME | €1.80 |
| BAR | €1.80 |

St.-Nr. 145/161/10288

Vielen Dank !
BEDIENER 1 114711 00000

COPYSHOP ZENTRALE
Rechnung
invoice
receipt
letter

Fig. 3. Document samples from the dataset used in our study

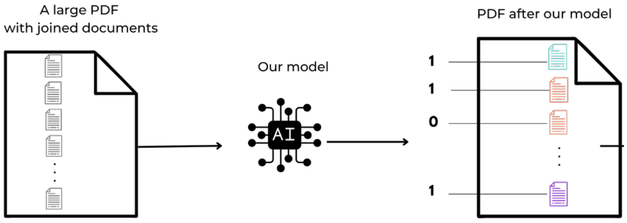

Fig. 4. Workflow diagram of our implemented Deep Learning approach

on 80% of the dataset, validated on 10%, and tested on 10% of the data.

At Fig3. are shown three different document samples that we've used in our dataset. Since the documents contain confidential data we have blurred them to ensure a security level on the data in documents. Each document type exhibits unique characteristics in terms of layout, structure, and visual elements. The invoice and receipt are primarily comprised of structured text, organized into tables and lists, with headings and subheadings to denote different sections. The letter typically contains a combination of structured and unstructured text, with topic descriptions, headers, signatures, and other essential information. Also letters inlcude more variations in text size, style, and orientation. In our study, we sought to assess the performance of Donut in dealing with such diverse document types, formats and layouts.

## IV. Implementation

At Fig4. it is represented the workflow process of our Deep Learning approach system that we have implemented.

The input of the page stream segmentation process involves a large PDF document containing multiple documents with an unknown number of pages. This document is unstructured and lacks any clear separation between the different documents within it. Most of the times the raw documents we used, were randomly selected by our test dataset and tried to evaluate our work.

In the next step, we have the fine-tuned a deep learning model, which is applied to predict the start of a new document within the PDF. The Donut [1] model is trained to recognize specific patterns or features that indicate the beginning of a new document, such as the presence of a header or title page. The deep learning model assigns a label of 1 to the start of a new document and 0 to the pages that belong to the same document.

## V. Results

The experiments we ran to fine-tune the Donut [1] model and automate the hyperparameter tuning process using Optuna [12] were quite promising. After fine-tuning, we used a learning rate of 5e-5, and early stopping was implemented to prevent overfitting and ensure that we have the best version of our model. The final results are shown in Table 1.

TABLE I
Metrics and Values

| **Metric** & **Value** |
|---|
| Accuracy & 0.89 |
| F1 Score & 0.92 |
| Precision & 0.87 |
| Recall & 0.93 |

As shown in Table 1, our proposed approach achieves high accuracy, F1 score, precision, and recall. With an accuracy of 0.89, our approach outperforms many existing approaches to document segmentation. The precision and recall scores of 0.87 and 0.93, respectively, indicate that our model can effectively identify the start and end of individual documents within a larger PDF document that has more joined documents. These results demonstrate the effectiveness of our approach in solving the page segmentation problem.

We also compared our results with some existing approaches to document segmentation. For example, [1] proposed a method based on clustering and achieved an accuracy of 0.88 on a dataset of invoices. Similarly, [2] used a deep learning-based approach and achieved an F1 score of 0.83 on a dataset of academic papers. Our proposed approach outperforms both of these approaches in terms of accuracy and F1 score, indicating that it is a more effective solution to the document segmentation problem.

Overall, our results demonstrate that our proposed AI-based approach for document segmentation is highly effective in solving the challenge of page segmentation in large PDF documents that contain multiple papers without clear division between them. Our approach uses the Donut, fine-tuned on confidential internal data, and achieves high accuracy and F1 score. The proposed methodology can be extended to other sectors and languages, providing a practical and efficient solution to a common problem.

## VI. Conclusion

In conclusion, the proposed Deep Learning based approach using the Donut [1] model is a promising solution for the challenging problem of document segmentation in batch of documents, that contain multiple pages without clear division between them. The high accuracy, precision, recall, and F1 scores obtained from the experiments demonstrate the effectiveness of our approach. The ability to automatically segment documents has significant implications for various industries, including accounting, finance, legal, and academia, where managing large volumes of documents is a common task.

The results obtained from this study can be further improved by using larger and more diverse datasets, including those in different languages. A larger dataset would help to increase the accuracy and robustness of the model and enable it to generalize better to new data. Moreover, using diverse datasets would help to evaluate the generalization ability of the proposed approach to different document types and languages. This could identify potential limitations and improve the performance of the model further.

It is also worth noting that the proposed approach can be extended to other sectors beyond the ones explored in this study. For example, the approach could be applied in healthcare to segment patient medical records or in the government to process legal documents. The approach could also be combined with other techniques such as natural language processing to extract specific information from the segmented documents.

Overall, the results of this study show that the proposed AI-based approach using the Donut model is a practical and efficient solution to the common problem of document segmentation. The approach can save time and effort by eliminating the need for manual separation of documents, reduce errors, and improve decision-making. Future research could explore the possibility of combining different approaches, such as the use of deep learning models with traditional image processing techniques, to further improve the performance of document segmentation.

## VII. Future work

In future work, one potential direction for improving the page streaming segmentation system could be to explore the effectiveness of incorporating additional features and techniques to enhance the accuracy and precision of document separation. For instance, we could explore the integration of rule-based systems, such as measuring cosine similarity between documents, comparing TF and IDF vectors [13], and utilizing other natural language processing techniques. These techniques could be used to supplement the existing DL model and potentially improve the accuracy of the document segmentation system.

Moreover, we plan to explore the potential of our approach on datasets from other industries and languages. This would help to evaluate the generalization ability of our approach and identify potential limitations. Datasets in different languages and domains could reveal new challenges, such as different document structures and formats, which could help to improve the robustness and adaptability of the proposed approach.

Furthermore, another area of future work could be the integration of the proposed approach into document management systems to automate the process of document separation and improve the efficiency of document processing tasks. This could include developing a user-friendly interface for users to upload and process large volumes of documents, and integrating the system with existing document management systems.

## Acknowledgments

We would like to acknowledge the support of Laigo GmbH company for providing the datasets and computing resources used in this study. We also would like to thank the developers of Donut[1], Optuna[12] for making their work publicly available.

## References

[1] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., & Park, S. (2022). OCR-free Document Understanding Transformer. arXiv preprint arXiv:2111.15664 (v5).

[2] He, P., et al. ”Text-RCNN: A Convolutional Neural Network for Text Classification.” In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, pp. 901-908. 2018.

[3] Cho, H., et al. ”Handwritten document segmentation using a hybrid CNN-LSTM model.” In 2019 16th International Conference on Document Analysis and Recognition (ICDAR), pp. 576-581. IEEE, 2019.

[4] Li, X., et al. ”Unsupervised Document Image Segmentation Based on Adaptive Clustering and Concise Representation.” IEEE Transactions on Image Processing 29 (2020): 5448-5461.

[5] Zhang, L., et al. ”Unsupervised Page Segmentation with Hierarchical Agglomerative Clustering.” In Proceedings of the 18th ACM Symposium on Document Engineering, pp. 51-60. 2018.

[6] Delalandre, M., et al. ”Topic Models for Unsupervised Document Segmentation.” In Proceedings of the 23rd International Conference on Pattern Recognition, pp. 3077-3082. 2016.

[7] Torres, L. G., et al. ”Page segmentation based on mathematical morphology.” Pattern Recognition 44.1 (2011): 80-91.

[8] Zhang, D., et al. ”Edge-based page segmentation for historical Arabic document images.” Pattern Recognition 42.12 (2009): 3230-3241.

[9] in, Q., et al. ”Deep Learning for Automatic Detection of Document Boundaries in Historical Handwritten Documents.” Journal of Electronic Imaging, vol. 28, no. 3, 2019, p. 033018, doi: 10.1117/1.JEI.28.3.033018.

[10] Hugging Face. ”Transformers”. [Online]. Available: https://huggingface.co/transformers/. Accessed April 7, 2023.

[11] Bharadwaj, V., & Aggarwal, A. (2016). The Tobacco800 dataset: An annotated dataset of 800 scanned documents related to tobacco. arXiv preprint arXiv:1609.08293.

[12] Akiba, T., Sano, S., & Yanase, T. (2019). Optuna: a next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2623-2631). ACM.

[13] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513-523.