*Digital Health and Informatics Innovations for Sustainable Health Care Systems*
*J. Mantas et al. (Eds.)*

*doi:10.3233/SHTI240382*

# Unsupervised Extraction of Body-Text from Clinical PDF Documents

Adel BENSAHLA[a,b,1], Jamil ZAGHIR[a,b], Christophe GAUDET-BLAVIGNAC[a,b] and Christian LOVIS [a,b]
[a] *Division of Medical Information Sciences, Geneva University Hospitals, Switzerland*
[b] *Department of Radiology and Medical Informatics, University of Geneva, Switzerland*

**Abstract.** Automatic extraction of body-text within clinical PDF documents is necessary to enhance downstream NLP tasks but remains a challenge. This study presents an unsupervised algorithm designed to extract body-text leveraging large volume of data. Using DBSCAN clustering over aggregate pages, our method extracts and organize text blocks using their content and coordinates. Evaluation results demonstrate precision scores ranging from 0.82 to 0.98, recall scores from 0.62 to 0.94, and F1-scores from 0.71 to 0.96 across various medical specialty sources. Future work includes dynamic parameter adjustments for improved accuracy and using larger datasets.

**Keywords.** clinical data, pdf, information extraction, unsupervised, DBSCAN

## 1. Introduction

Extracting body-text information from Electronic Health Records, typically stored in PDFs, is a complex task. Conventional supervised methods face limitations such as the requirement for extensive training data and the variability of document template used across clinical units [1], alongside reliance on GPU resources. Lin. et al.[2] used page-association across neighboring pages within a document, but this approach falls short when different pages have varying headers and footers. To address these challenges, we propose an unsupervised method that also leverages the page-association method, using large volume of clinical data. Our approach involves analyzing numerous documents from a single source, understanding their organizational patterns, and categorizing them into templates based on the analysis of their aggregated first pages. We extract clinical PDFs body-text from various medical specialties, including Cardiology, Endoscopy, and Radiology reports. Notably, the last two present unique challenges due to the presence of text in left margins, which can complicate the examination process.

## 2. Methods

*Preprocessing*: The clinical PDFs analyzed had processable text, enabling direct extraction. We used MuPDF[3], which extracts text in BBoxs.

[1] Corresponding Author: Adel BENSAHLA; E-mail: adel.bensahla@etu.unige.ch.

*Clustering template*: We grouped documents into templates by analyzing their aggregated first pages. We clustered the BBoxs whose centers were within a predefined margin around the border using DBSCAN. Files within a common cluster were organized into the same template folder.

*Extraction*: We reused the DBSCAN clustering algorithm, applying it separately to the first pages and other pages within each template (since templates often differ between first and subsequent pages). If DBSCAN identified clusters, these BBoxs were considered "constraints"; if deemed noise, they were ignored. We then applied a modified version of the "Largest Empty Rectangle"[4] algorithm among the pages and BBoxs constraints. This process provided coordinates for both the first and subsequent pages for each template, facilitating the extraction of meaningful text.

## 3. Results, Discussion and Conclusions

To evaluate, we compared each combination of file and BBoxs against corresponding values between our algorithm's predictions and our annotated gold standard, achieving a macro-F1 of 91.5% for Cardiology, 77% for Endoscopy and 72% for Radiology (Table 1).

**Table 1.** Performances of our approach. ('relevant' = text we seek, 'irrelevant' = text we aim to exclude.)

| Source | Nb docs | Nb BBoxs | Relevance | Precision | Recall | F1-Score |
|---|---|---|---|---|---|---|
| Cardiology | 80 | 1034 | irrelevant | 0.80 | 0.93 | 0.86 |
| | | 3642 | relevant | 0.98 | 0.94 | 0.96 |
| Endoscopy | 80 | 2220 | irrelevant | 0.81 | 0.86 | 0.83 |
| | | 1404 | relevant | 0.75 | 0.68 | 0.71 |
| Radiology | 80 | 1573 | irrelevant | 0.82 | 0.62 | 0.71 |
| | | 1468 | relevant | 0.68 | 0.85 | 0.75 |

We introduced an unsupervised, resource-efficient algorithm for body-text extraction from PDF documents, beneficial for downstream tasks and ideal for healthcare settings with large data volumes in PDF format. However, we faced challenges, particularly with more complex reports, where the algorithm struggled to identify margin constraints, resulting in lower performance scores (i.e., Radiology).

Acknowledging limitations, such as the fixed number of BBoxs used for clustering files into folders, future work could explore dynamic parameter adjustments to enhance clustering accuracy. We aim to leverage clustering for predictive file organization, particularly as datasets grow beyond 100’000 files per source.

This study, part of the HERO project (CCER 2023-01571), is funded by Mr. Nicolas Pictet's philanthropic fund, for which we are grateful.

## References

[1] Subramani N, et al. A Survey of Deep Learning Approaches for OCR and Document Understanding [Internet]. arXiv; 2021 [cited 2024 Mar 16]. doi: 10.48550/arXiv.2011.13534.

[2] Lin X. Header and footer extraction by page association. Document Recognition and Retrieval X [Internet]. SPIE; 2003 [cited 2024 May 17]. p. 164–171. doi: 10.1117/12.472833.

[3] Mupdf [Internet]. 2024 [cited 2024 May 15]. Available from: https://github.com/ArtifexSoftware/mupdf.

[4] Computing the Largest Empty Rectangle | SIAM Journal on Computing [Internet]. [cited 2024 Mar 15]. Available from: https://epubs.siam.org/doi/10.1137/0215022.