

NetLay: Layout Classification Dataset for Enhancing Layout Analysis*

Sharva Gogawale
Tel Aviv University
sharvag@mail.tau.ac.il

Berat Kurar-Barakat
Tel Aviv University
berat@tauex.tau.ac.il

Daniel Stökl Ben Ezra
École Pratique des Hautes Études
Daniel.Stoekl@ephe.psl.eu

Luigi Bambaci
École Pratique des Hautes Études
luigi.bambaci@ephe.psl.eu

Daria Vasyutinsky Shapira
Tel Aviv University
dariashap@tauex.tau.ac.il

Nachum Dershowitz
Tel Aviv University
nachum@tau.ac.il

Abstract

Within the domain of historical document image analysis, the process of identifying the spatial structure of a document image is an essential step in many document processing tasks, such as optical character recognition and information extraction. Advancements in layout analysis promise to enhance efficiency and accuracy using specialized models tailored to distinct layouts.

We introduce NetLay, a new dataset for benchmarking layout classification algorithms for historical works. It consists of over 1300 images of pages of printed Hebrew (or Hebrew-character) books in a variety of styles, categorized into four different classes based on their layout (the number of text columns and regions). Ground truth was crafted manually at the page level. Furthermore, we conduct an in-depth performance evaluation of various layout classification algorithms, which are based on deep-learning models that learn to extract spatial features from images. We evaluate our algorithms on NetLay and achieve state-of-the-art results on the task of layout classification for historical books. We explore dual classification methodologies: one by assigning a single label and the other by assigning multiple labels, representing each of the four classes as a vector of length five, where each dimension could be common among the four classes. This approach is advantageous as there are some overlapping attributes between classes.

Our results demonstrate the potential of NetLay to be used as a benchmark for evaluating layout classification algorithms and training new classification models. NetLay aims to enhance research in layout classification for historical books driven by the imperative to streamline layout segmentation, particularly for books featuring more than three columns. This initiative seeks to alleviate the segmentation process for a substantial corpus of books.

Keywords: historical document analysis, layout classification, convolutional neural networks, deep learning

*This research was funded in part by the European Union (ERC, MiDRASH, Project No. 101071829). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

1 Introduction

Numerous institutions and libraries worldwide are digitizing their archives to democratize access and safeguard them from physical deterioration. This calls for an ability to perform primary processing of numerous texts automatically. In the field of document image processing,¹ benchmark datasets with corresponding ground truth are essential for evaluating, developing, and comparing algorithms, as they also drive the creation of new approaches to address emerging challenges. Recent advancements in image analysis and computer vision have automated most of the tasks in the pipeline for automatic document analysis. Document layout analysis acts as a crucial preliminary step for various document image analysis tasks. Advancements in this field hold immense potential for boosting efficiency and accuracy through the development of specialized models tailored to diverse document layouts. Document image processing encompasses classical machine learning techniques, requiring meticulous feature selection, and deep neural network-based approaches where features are inherently learned within the network. While both techniques play a role, recent breakthroughs in image classification have been primarily driven by deep-learning methods.

A key advantage deep learning offers over traditional approaches lies in its inherent ability to extract features directly from the data. This not only liberates paleographers from spending weeks or months on feature selection but also empowers neural networks to uncover novel and intricate features that might evade even the most discerning human expert. A critical aspect of this endeavor is addressing the challenges inherent in ancient and medieval handwriting studies, necessitating the training of specialized models tailored to distinct layouts. However, the scarcity of diverse stylistic representations poses challenges for developing multi-domain general layout analysis, compounded by the predominance of datasets containing Latin script.

Addressing these disparities is imperative for advancing historical document analysis research and development, particularly in historical document layout analysis. However, the current landscape of available datasets suffers from two major limitations that hinder progress in historical document analysis. Firstly, the lack of stylistic diversity can significantly hamper the development of general layout analysis methods capable of functioning effectively across multiple domains. Secondly, the vast majority of existing datasets primarily cater to documents in English, neglecting the inherent differences in text features present in other languages. This disparity can lead to problems when applying these methods to languages like Hebrew, highlighting a critical gap in resources dedicated to historical document layout analysis datasets. While significant strides have been made in the domain of modern documents, addressing this discrepancy is paramount to propelling research and development forward in the field of historical document analysis.

Long-standing efforts have been devoted to creating layout analysis datasets, with the huge dataset PubLayoutNet ([Zhong et al., 2019](#)) for contemporary documents emerging recently. However, existing datasets tailored for historical documents remain limited in scope. The majority of openly available historical document layout datasets mostly address more popular scripts and languages. The Europeana Newspapers Project (ENP) ([Clausner et al., 2015](#)) contains common European languages like Dutch, English, German, etc., from the 17th century onward, and contains 500 page images. The PRIMa Layout Analysis Dataset ([Antonacopoulos et al., 2009](#)) places emphasis on magazines and technical/scientific publications, the majority in Latin script. Addressing these disparities and incorporating the representation of less common and older languages—like Hebrew—

¹We use the term *document* in its general sense, ranging from literary works to personal notes, from full-length books to individual pages.



Figure 1: Document samples from NetLay: (a) no text; (b) single column; (c) two column; and (d) complex layout.

in datasets are imperative for advancing historical document analysis research and development.

Before we address the more complicated question of Hebrew *manuscript* layout, we must solve the problem of automatic layout classification for *printed* Hebrew books. Hebrew books often have nonstandard layouts, multiple languages (Hebrew/Aramaic; Hebrew/Yiddish, etc.) per page written in the same script and alphabet, and different script type-modes per page (Ashkenazi square plus Oriental semi-cursive [“Rashi”]). Sometimes, different text fields are not clearly distinguishable.

To address these challenges, we present NetLay, a dataset containing 1352 pages, taken from books with diverse layouts sourced from the collection of the National Library of Israel (NLI). In addition, we propose several benchmark techniques to perform layout classification. We implement various deep-learning models. We also propose a multi-label encoding scheme based on the spatial and global interdependencies of distinct layout elements.

The remainder of this paper is organized as follows: Section 2 is a short survey of the related literature. Section 3 explains the properties of the dataset proposed. Section 4 describes various methods used for layout classification. In Section 5 we evaluate several deep-learning classifiers and present our results.

2 Related Work

Understanding the layout of a document serves as a preliminary step for various document image processing tasks. These tasks include information retrieval, page segmentation, word spotting, and optical character recognition (OCR), which aims to extract meaningful textual information from these images. Breuel (2003) proposed novel algorithms and statistical methods for flexible page layout analysis, combining globally optimal geometric algorithms with robust statistical models and meticulous engineering techniques. Page segmentation algorithms typically fall into two categories: bottom-up and top-down. Bottom-up algorithms work in a hierarchical manner to group elements such as pixels, patches, or connected components into progressively larger regions. In contrast, top-down algorithms divide the entire page into regions in a single step. Many of the early page layout analysis methods often relied on assumptions about document structure and employed a top-down approach, particularly for well-formatted, modern binary documents. Typically, these methods rely on mathematical morphology and connected components (Alarcón Arenas et al., 2018), Voronoi diagrams (Kise et al., 1998), or run length smearing algorithms (Wong et al., 1982).

There are, however, also numerous other techniques that don't fit neatly into one of the above categories. These so-called mixed or hybrid approaches aim to merge the efficiency of top-down methods with the robustness of bottom-up ones. Corbelli et al. (2016) proposed a hybrid layout analysis pipeline, integrating both top-down and bottom-up approaches. They employ the X-Y-cut algorithm and a support vector machine (SVM) classifier for illustration detection, coupled with a convolutional neural network (CNN) and random forest classifier for content classification identifying different classes of layout entities. Pixel classification approaches have also been explored for page segmentation. Wei et al. (2013) framed the problem as pixel classification, where each pixel is represented as a feature vector based on the image's color. They employed techniques like Gaussian mixture models (GMM), multi-layer perceptrons (MLP), and SVM to classify pixels into categories such as decoration, background, periphery, and text pixels. Chen et al. (2014) subsequently improved upon this work by incorporating more comprehensive features encompassing texture and color properties like smoothness, Laplacian, Gabor dominant orientation histograms, local binary patterns, and color variance.

With the onset of deep learning, many authors have addressed the problem of layout segmentation and analysis using different deep neural network configurations. Borges Oliveira and Viana (2017) introduced a novel one-dimensional CNN approach for rapid automatic layout detection of structured documents. Barakat and El-Sana (2018) presented a binarization-free method for page layout analysis of historical Arabic manuscripts, training an FCN to predict the class of each pixel and segmenting main text and side text regions. Kosaraju et al. (2019) proposed DoT-Net, a texture-based CNN for document layout analysis that can capture textural variations among the multiclass regions of documents. Alaasam et al. (2019) proposed a Siamese network-based layout analysis method tailored for challenging historical Arabic manuscripts. Da et al. (2023) introduced a two-stream vision grid transformer for layout analysis, conducting visual pre-training in two stages utilizing 2D token-level and segment-level understanding.

Although layout analysis and segmentation have been extensively explored, layout classification remains relatively understudied. This process involves categorizing documents based on their spatial arrangement, aiming to comprehend the overall layout of content within a document. This understanding serves as a cornerstone for the development of advanced algorithms for segmentation and OCR. Hu et al. (1999) introduced interval encoding, a novel feature set for capturing layout information. They utilize this encoding within an HMM framework for fast document image



Figure 2: Examples of complex layouts.

classification based solely on spatial layout.

3 Dataset

There is a critical necessity of implementing a layout classifier to augment the efficacy of dedicated models used in transcription systems like eScriptorium (Kiessling et al., 2019). To that end, we sourced images from the digital collections of NLI, tailored specifically for this task. High resolution images of pages in the NetLayout dataset were curated from a random selection of printed Hebrew books at NLI. From each book, one page image was carefully chosen for inclusion in the dataset. The dataset includes a total of 1352 images of single pages or facing pages. It is balanced and comprises the following classes: no text (“empty”), single column, two columns (occasionally on facing pages), and complex layout (three or more regions, or regions with insets), with 300, 442, 300, and 310 samples, respectively, for each class.

Facing pages, each containing one column, are usually one continuous work, but may also be

two related works, one on even numbered pages and the other on odd ones. Two-column text may be read across both columns (as in poetry, for example), or column by column, or they may be two works side by side—in the same language or in two (perhaps a translation or commentary), in the same font or not. Complex layouts often contain separate, but related, works by different authors.

Figure 1 contains illustrative examples of document images representing each class within the dataset. The dataset is publicly available.²

4 Methods

The challenge of layout class identification presents itself as an image classification task, where the goal is to assign a specific class to a given document image. Given the complexity and variability of layouts, employing deep-learning models emerges as the most effective strategy for image classification tasks. Therefore, our approach uses deep-learning-based models to accurately categorize document images into distinct layout structures. In this section, we outline the experimental setup, including model architecture, training methodology, and evaluation procedures. We adopt state-of-the-art deep-learning models tailored for image classification tasks. To assess the performance of our approach, we conduct several benchmark experiments. These experiments aim to evaluate the efficacy of the proposed deep-learning models in accurately classifying layout structures. To ensure a robust evaluation, we divided our dataset into three distinct subsets: training (80%), validation (10%), and testing (10%). This split allows for effective model training, hyperparameter tuning, and unbiased performance evaluation. All the experiments for training the deep-learning models were conducted on a machine equipped with an NVIDIA Titan T4 GPU with 15 GB of memory.

Predictions are evaluated based on four standard performance metrics: accuracy, precision, recall, and F1-score.

We employ two methods for the task of document image layout classification.

4.1 Single Label Classification

Single-label classification involves assigning one class label to each instance from a predefined set of classes. In the context of document layout classification, our objective is to categorize layouts into four distinct classes: no text, single column, double column, or complex.

Below, we explore various architectures and propose methods employed for this task.

EfficientNetV2. We utilize EfficientNetV2 (Tan and Le, 2021) for spatial feature extraction, pre-trained on the ImageNet dataset. The core architecture employs the mobile inverted bottleneck (MBConv) (Sandler et al., 2018), with squeeze and excitation optimization.

In the EfficientNet family, comprising models from EfficientNet B0 to B7 (Tan and Le, 2019) which employs mobile inverted bottleneck convolution (MBConv) with squeeze and excitation optimization. The variations can be seen in MBConv block count, width, depth, resolution, and overall size of the model. EfficientNetV2 introduces enhancements like fused-MBConv blocks alongside regular MBConv blocks, which lead to higher accuracies with fewer parameters. EfficientNetsV1s demonstrate adaptability through transfer learning, excelling when trained on diverse datasets. However, challenges such as slow training with large image sizes and inefficiencies in early layers due to depthwise convolutions are evident. Addressing these concerns, EfficientNetV2 introduced

²At https://github.com/TAU-CH/midrash_layout_classification_using_multilabel_vgg/tree/main/data.

Class	Page width text line	Half page width text time	Page height vertical separator	Half page height vertical separator	Multiple fonts
Empty	0	0	0	0	0
Single column	1	0	0	0	0
Two column	0	1	1	0	0
Complex layout	1	1	0	1	1

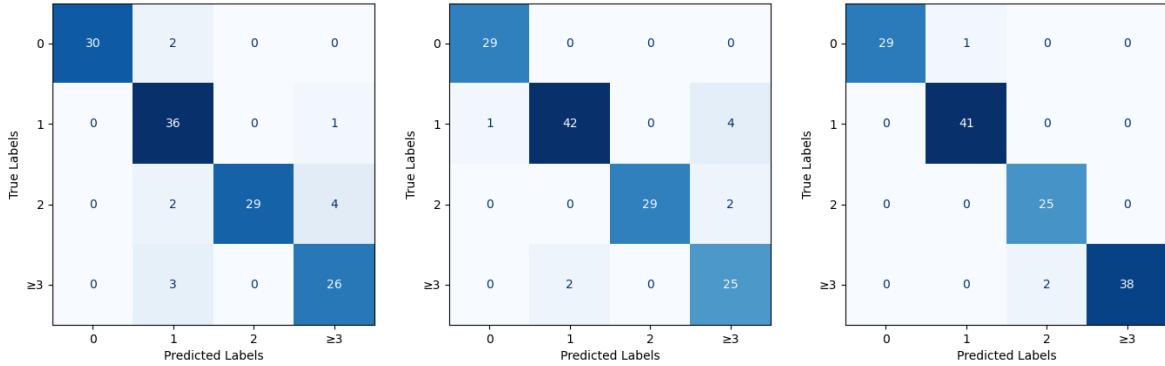
Table 1: The multi-label encoding scheme.

Method	Class	Accuracy	Precision	Recall	F1 Score
Efficient-Net	0 (Empty)	98.50%	0.94	1.00	0.97
	1 (Single column)	93.98%	0.97	0.84	0.90
	2 (Two columns)	95.49%	0.83	1.00	0.91
	3 (Complex layout)	93.98%	0.90	0.84	0.87
ViT	0 (Empty)	99.25%	1.00	0.97	0.98
	1 (Single column)	94.78%	0.89	0.95	0.92
	2 (Two columns)	98.51%	0.94	1.00	0.97
	3 (Complex layout)	94.03%	0.93	0.81	0.86
VGG16	0 (Empty)	99.26%	0.97	1.00	0.98
	1 (Single column)	99.26%	1.00	0.98	0.99
	2 (Two columns)	98.53%	1.00	0.93	0.96
	3 (Complex layout)	98.53%	0.95	1.00	0.97

Table 2: Performance metrics (accuracy, precision, recall, and F1), for each class, using Efficient-NetV2, ViT, and VGG16 with multi-label encoding.

novel design elements and employs training-aware neural architecture search and scaling strategies to jointly optimize model accuracy, training speed, and parameter size.

Vision transformer. We also experiment with the vision transformer (ViT) architecture ([Dosovitskiy et al., 2021](#)), which transforms image processing by dividing input images into fixed-sized patches, departing from the conventional pixel-based evaluation of CNNs. ViT encapsulates each patch into a latent representation while retaining positional information, forwarding them through a transformer encoder. The input image, denoted $x \in \mathbb{R}^{H \times W \times C}$, undergoes transformation into a sequence of flattened 2D patches $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where $N = W \cdot H / P^2$ signifies the resulting number of patches of size $P \times P$, and $H \times W$ is the resolution of the image. With C representing the channels, typically 3 for RGB images, our model embraces a patch size of 16×16 pixels. This architecture facilitates the breakdown of images into manageable patches, subsequently processed through transformer layers adept at capturing both local and global dependencies. Our methodology aligns with the ViT paradigm, expanding the adaptability of transformers to encompass image classification tasks.



(a) Confusion matrix for Effi-clientNetV2. (b) Confusion matrix for ViT. (c) Confusion matrix for multi-label encoding with VGG16.

Figure 3: Confusion matrices for the different classifiers.

4.2 Multi-label Classification

Multi-label classification involves the assignment of multiple labels to each instance simultaneously. It involves predicting multiple categories or classes for a given input, making it a more complex problem compared to traditional single-label classification. To address potential overlap in class characteristics, we also employ a multi-label classification approach. Each of the four classes is encoded as a five-dimensional vector, allowing for shared attributes across classes (refer to Table 1 for details). This method offers distinct advantages, particularly in handling overlapping attributes among certain classes. Furthermore, the extraction of spatial document image features for layout classification is facilitated through the utilization of a VGG16 ([Simonyan and Zisserman, 2015](#)) based backbone.

Complex layout classification. We delve deeper into understanding the complexities of layout structures. Figure 2 showcases various examples from the dataset, highlighting the variability in spatial arrangements of text columns within the complex layout structure. For instance, Figure 2(a) exhibits a C type structure, while Figure 2(b) displays an L type arrangement. Moreover, Figures 2(c) and (d) portray complex spatial configurations bearing resemblance to an O and a U, respectively. We identified seven distinct subcategories within the complex layout arrangement, as depicted in Figure 4. These subcategories are characterized by different spatial configurations of text columns, including variations such as C, L, U, and O shapes, along with their corresponding reflected counterparts—C2, L2, and U2. Each of these subcategories captures unique layout features, contributing to the complexity of the overall structure, and poses different challenges for accurate classification. Through training an end-to-end CNN-based classifier, we aimed to comprehend these features and effectively capture the nuanced spatial relationships within the complex layout structures. Our experiments yielded a classification accuracy of 60%, indicating the model’s ability to distinguish these spatial features significantly better than random guessing.

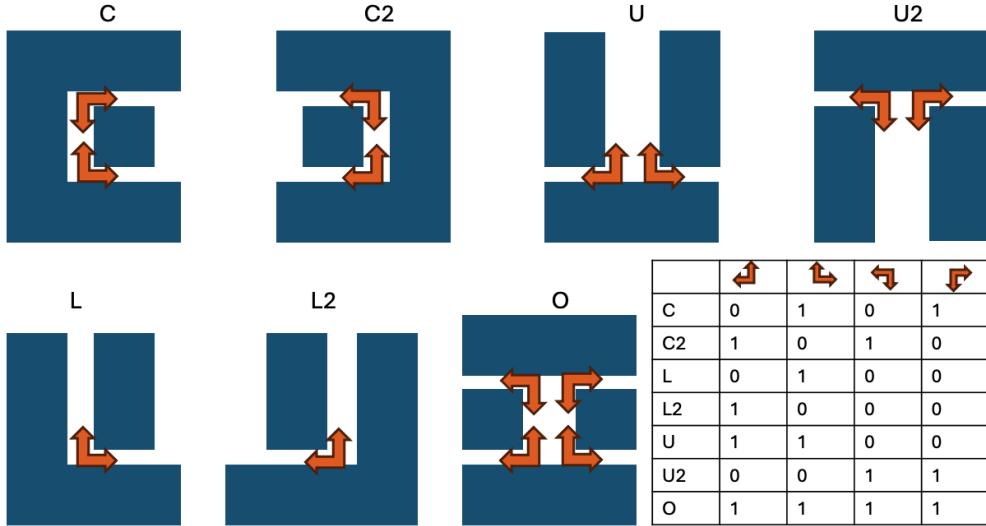


Figure 4: Examples of complex layout structures with corresponding spatial arrangement features.

5 Results

In this section, we present the outcomes obtained from various deep-learning classifiers, which serve as foundational benchmarks for future comparative analyses. The aim was to assess the effectiveness of the proposed features and methods introduced here for facilitating efficient document layout classification. We achieved competitive performance on the document classification task. Figures 3a, 3b, and 3c showcase the confusion matrices corresponding to the trained models. The evaluation metrics, including accuracy, precision, recall, and F1, are utilized to assess the models' performance across different classes, as showcased in Table 2. Notably, employing EfficientNetV2 yielded an impressive overall accuracy of 90.98%, while the ViT model achieved an even higher accuracy of 93.28%. Furthermore, leveraging the multi-label encoding approach with VGG16 resulted in the highest accuracy of 97.79%. To elucidate the influential features guiding the model's final prediction, we employ the gradient-weighted class activation mapping (Grad-CAM) technique (Selvaraju et al., 2017). This approach leverages the gradients of a target class flowing into the underlying CNN architecture, specifically VGG16 in our study, to generate a coarse localization map, thereby accentuating pivotal regions crucial for predicting the target class. Figure 5 depicts the salient features relevant to the classification of layout structures.

6 Conclusions and Future Work

Conducting layout analysis on simple layouts, containing one or two columns of text, is relatively straightforward, but analyzing complex layouts that feature text columns in structures diverging from the standard one or two columns, such as L, U, O, and X shapes, alongside other complexities, presents significant challenges. Therefore, layout classification is vital for distinguishing between simple and complex layouts. This distinction allows for the application of existing layout analysis algorithms on simple layout document images but specialized analysis methods for complex layout document images.

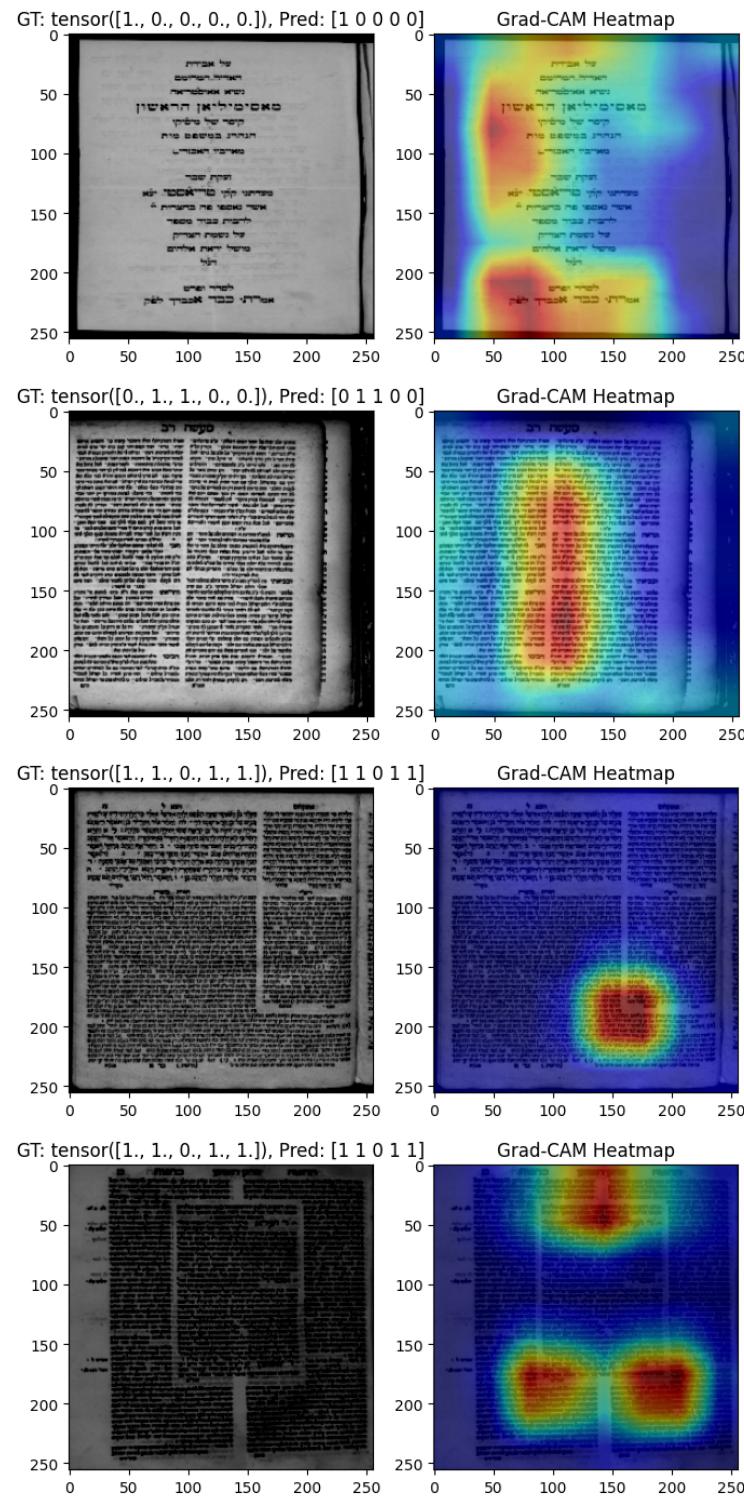


Figure 5: Visualization of important features for classification using Grad-CAM.

We have introduced a dataset designed for benchmarking layout classification methods, along with a single-label multi-classification algorithm and a multi-label multi-classification algorithm to address the layout classification challenge. Our findings indicate that multi-label encoding leads to a more separable feature space, thereby enhancing accuracy. The visualization of classifiers further supports this conclusion, revealing that the classifiers indeed focus on features employed to encode the multi-labels for each class.

Future work includes further improving results for complex layout classification in a variety of languages and scripts, considering pages with marginal and inter-textual comments, considering books with changes of script size and/or language within paragraphs, and pages from incunabula and other early printed books with unusual nonstandard layouts. This will be combined with reading-direction recognition, language, and script detection to achieve complex page analysis. These algorithms would serve as a solid base for efficient automatic processing of printed books. At the same time, the automatic classification of page layouts for printed books is an important preparatory step for the more challenging task of page layout analysis of handwritten manuscripts.

References

- Reem Alaasam, Berat Kurar, and Jihad El-Sana. Layout analysis on challenging historical Arabic manuscripts using Siamese network. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 738–742, 2019. doi: 10.1109/ICDAR.2019.00123.
- Sebastian Wilde Alarcón Arenas, Yessenia Yari, and Graciela Meza-Lovon. A document layout analysis method based on morphological operators and connected components. In *XLIV Latin American Computer Conference (CLEI)*, pages 622–631, 2018. doi: 10.1109/CLEI.2018.00080.
- Apostolos Antonacopoulos, David Bridson, Christos Papadopoulos, and Stefan Pletschacher. A realistic dataset for performance evaluation of document layout analysis. In *10th International Conference on Document Analysis and Recognition (ICDAR)*, pages 296–300, 2009. doi: 10.1109/ICDAR.2009.271.
- Berat Kurar Barakat and Jihad El-Sana. Binarization free layout analysis for Arabic historical documents using fully convolutional networks. In *IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, pages 151–155, 2018. doi: 10.1109/ASAR.2018.8480333.
- Dário Augusto Borges Oliveira and Matheus Palhares Viana. Fast CNN-based document layout analysis. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1173–1180, 2017. doi: 10.1109/ICCVW.2017.142.
- Thomas Breuel. High performance document layout analysis. Technical report, ResearchGate, May 2003. URL https://www.researchgate.net/publication/2564797_High_Performance_Document_Layout_Analysis.
- Kai Chen, Hao Wei, Jean Hennebert, Rolf Ingold, and Marcus Liwicki. Page segmentation for historical handwritten document images using color and texture features. In *14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 488–493, 2014. doi: 10.1109/ICFHR.2014.88.

Christian Clausner, Christos Papadopoulos, Stefan Pletschacher, and Apostolos Antonacopoulous. The ENP image and ground truth dataset of historical newspapers. In *13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 931–935, 2015. doi: 10.1109/ICDAR.2015.7333898.

Andrea Corbelli, Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Historical document digitization through layout analysis and deep content classification. In *23rd International Conference on Pattern Recognition (ICPR)*, pages 4077–4082, 2016. doi: 10.1109/ICPR.2016.7900272.

Cheng Da, Chuwei Luo, Qi Zheng, and Cong Yao. Vision grid transformer for document layout analysis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19405–19415, 2023. doi: 10.1109/ICCV51070.2023.01783.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>.

Jianying Hu, Ramanujan S. Kashi, and Gordon Wilfong. Document image layout comparison and classification. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR '99)*, pages 285–288, 1999. doi: 10.1109/ICDAR.1999.791780.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. eScriptorium: An open source platform for historical document analysis. In *International Conference on Document Analysis and Recognition Workshops (ICDARW)*. IEEE, September 2019. doi: 10.1109/icdarw.2019.10032. URL <http://dx.doi.org/10.1109/icdarw.2019.10032>.

Koichi Kise, Akinori Sato, and Motoi Iwata. Segmentation of page images using the area Voronoi diagram. *Computer Vision and Image Understanding*, 70(3):370–382, 1998. ISSN 1077-3142. doi: <https://doi.org/10.1006/cviu.1998.0684>. URL <https://www.sciencedirect.com/science/article/pii/S1077314298906841>.

Sai Chandra Kosaraju, Mohammed Masum, Nelson Zange Tsaku, Pritesh Patel, Tanju Bayramoglu, Girish Modgil, and Mingon Kang. DoT-Net: Document layout classification using texture-based CNN. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1029–1034, 2019. doi: 10.1109/ICDAR.2019.00168.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. doi: 10.1109/CVPR.2018.00474.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, May 7–9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1409.1556>.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research (PMLR)*, pages 6105–6114, June 2019. URL <https://proceedings.mlr.press/v97/tan19a.html>.

Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research (PMLR)*, pages 10096–10106, July 2021. URL <https://proceedings.mlr.press/v139/tan21a.html>.

Hao Wei, Micheal Baechler, Fouad Slimane, and Rolf Ingold. Evaluation of SVM, MLP and GMM classifiers for layout analysis of historical documents. In *12th International Conference on Document Analysis and Recognition*, pages 1220–1224, 2013. doi: 10.1109/ICDAR.2013.247.

Kwan Y. Wong, R. G. Casey, and F. M. Wahl. Document analysis system. *IBM Journal of Research and Development*, 26(6):647–656, 1982. doi: 10.1147/rd.266.0647.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. PubLayNet: largest dataset ever for document layout analysis, September 2019. ISSN 1520-5363.