Deep Learning for Effective Classification and Information Extraction of Financial Documents

Valentin-Adrian Serbanescu[a] and Maruf Dhali[b]
Department of Artificial Intelligence, Bernoulli Institute, University of Groningen, 9747 AG Groningen, The Netherlands

Keywords: Deep Learning, Classification, Information Extraction, Financial Documents, Optical Character Recognition, CNN, RoBERTa, LayoutLMv3, GraphDoc, Document Processing.

Abstract: The financial and accounting sectors are encountering increased demands to effectively manage large volumes of documents in today's digital environment. Meeting this demand is crucial for accurate archiving, maintaining efficiency and competitiveness, and ensuring operational excellence in the industry. This study proposes and analyzes machine learning-based pipelines to effectively classify and extract information from scanned and photographed financial documents, such as invoices, receipts, bank statements, etc. It also addresses the challenges associated with financial document processing using deep learning techniques. This research explores several models, including LeNet5, VGG19, and MobileNetV2 for document classification and RoBERTa, LayoutLMv3, and GraphDoc for information extraction. The models are trained and tested on financial documents from previously available benchmark datasets and a new dataset with financial documents in Romanian. Results show MobileNetV2 excels in classification tasks (with accuracies of 99.24% with data augmentation and 93.33% without augmentation), while RoBERTa and LayoutLMv3 lead in extraction tasks (with F1-scores of 0.7761 and 0.7426, respectively). Despite the challenges posed by the imbalanced dataset and cross-language documents, the proposed pipeline shows potential for automating the processing of financial documents in the relevant sectors.

1 INTRODUCTION

In today's digital era, the rapid growth of data highlights the need for efficient processing of financial documents — especially in finance and accounting, where manually handling invoices and receipts can be labor-intensive and prone to errors. This article aims to develop a machine-learning pipeline that classifies and extracts critical information from scanned financial documents. It will utilize advanced techniques such as Convolutional Neural Networks (CNNs), Graph Attention Networks (GATs), Transformers, and Optical Character Recognition (OCR).

The proposed system addresses the complexities of various document formats, enhancing the accuracy and speed of financial data processing. With the growing demand for automation in large organizations like banks and accounting firms, this article seeks to optimize financial operations.

This work presents a machine learning pipeline capable of accurately classifying and extracting valuable information from diverse financial documents, laying the groundwork for future industrial applications. All materials and scripts are available on GitHub[1].

1.1 Related Works

Document classification in the financial sector has become crucial for processing large volumes of documents like invoices and bank statements. Initially reliant solely on Optical Character Recognition (OCR), the field has evolved with machine learning advancements, mainly through Convolutional Neural Networks (CNNs). In the study by (Chen and Blostein, 2007), the authors emphasized the limitations of traditional classification methods relying solely on OCR. Later, (Kang et al., 2014) demonstrated the effectiveness of CNNs in handling the structural hierarchies and spatial relationships of document images, showing a significant improvement over earlier techniques. (Rusiñol et al., 2014) introduced a multimodal ap-

[a] https://orcid.org/0009-0003-6678-6536
[b] https://orcid.org/0000-0002-7548-3858

[1] GitHub Repository (private - available upon request): DL-financial-doc-classification-extraction

Serbanescu, V.-A. and Dhali, M.
Deep Learning for Effective Classification and Information Extraction of Financial Documents.
DOI: 10.5220/0013261000003905
In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 749-756
ISBN: 978-989-758-730-6; ISSN: 2184-4313

proach combining CNNs for visual analysis and OCR for text extraction, advancing multi-page document classification. (Harley et al., 2015) further explored deep CNNs for document classification, underscoring their capability to manage high visual variability in document categories. More recent research by (Dong and Li, 2020) applied a simplified CNN architecture specifically to financial documents, like invoices and bank receipts, demonstrating practical applications in the finance sector. On the other hand, (Lehtonen et al., 2020) explored the k-Nearest Neighbors (kNN) algorithm for document classification, pointing out the potential of simpler models, while (Ömer Arslan and Uymaz, 2022) provided insights into CNN-based architectures like LeNet-5, VGG-19, and MobileNetV2, highlighting techniques like padding and data augmentation for enhancing classification accuracy.

Considerable progress has been achieved in document information extraction with the advancement of pre-trained models. (Liu et al., 2019) introduced RoBERTa, an enhanced version of BERT, which showed superior performance on various benchmarks by optimizing training methods and data. (Majumder et al., 2020) proposed a representation learning approach for form-like documents, significantly improving extraction performance across multiple domains by generating extraction candidates based on target fields. Another critical study by (Oral et al., 2020) tackled the challenge of extracting information from visually complex banking documents using deep learning algorithms and neural word representations such as FastText, ELMo, and BERT. The authors noted a significant 10% improvement in named entity recognition and relation extraction. (Huang et al., 2022) introduced LayoutLMv3, a multimodal pre-trained model for Document AI, combining text and image masking to align modalities, leading to state-of-the-art performance across various document understanding tasks without the need for CNN-based image feature extraction. (Ha and Horák, 2022) developed the OCRMiner system, designed specifically for extracting information from scanned documents like invoices, showing success rates of over 88% for Czech and English invoices. Finally, (Štěpán Šimsa et al., 2023) presented the DocILE benchmark, using RoBERTa and LayoutLMv3 for essential information extraction tasks, further contributing to advancements in document AI.

Chargrids: The introduction of chargrids by (Katti et al., 2018) added a new dimension to document understanding, converting documents into a 2D grid of characters for processing with CNNs. This approach preserves the spatial layout of the document and improves semantic segmentation and information extraction, particularly from layout-rich documents like invoices (see Figure 1). By representing documents in this format, the chargrid model captures both textual and spatial information, which proves beneficial for information extraction, particularly in layout-rich documents like invoices. In their experiments, the authors demonstrated that chargrid significantly outperforms traditional text-only or image-based methods in extracting structured information from documents with complex layouts. Based on these findings, we incorporate chargrid into our model training process for information extraction, aiming to leverage its ability to improve spatial and layout understanding in document processing tasks.

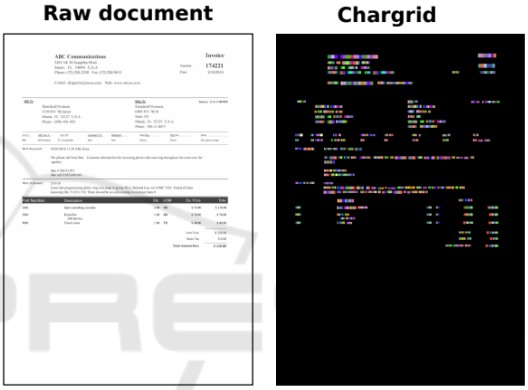

Figure 1: Example of chargrid created on the basis of a financial document. Figure adapted from the study of (Katti et al., 2018).

Data Augmentation: Data augmentation techniques play a vital role in improving model performance. (Perez and Wang, 2017) explored various augmentation methods such as cropping, rotating, and GAN-based style transfer. Similarly, (Mikołajczyk-Bareła and Grochowski, 2018) proposed blending content and style through image style transfer, which proved particularly effective in tasks like medical imaging. (Shorten and Khoshgoftaar, 2019) categorized augmentation methods into data warping and oversampling techniques, such as SMOTE and GANs, while (Yang et al., 2023) provided a comprehensive taxonomy of augmentation methods, highlighting the benefits of AutoAugment for optimizing policies. These techniques, particularly traditional ones, are essential for improving the performance of models trained on financial documents, as demonstrated by previous research. We will only use traditional augmentation techniques to simulate realistic variations of financial documents based on conclusions from related works.

2 METHODS

2.1 Classification

In the classification task, this study uses two datasets: a new proposed dataset with Romanian Financial Documents (RFD) comprising seven classes (invoice, receipt, accompanying notice of wares, bank statement, return receipt, payment disposition, and collection disposition) and the RVL-CDIP Dataset (Harley et al., 2015). The RFD dataset contains 896 samples distributed across these classes (for example, see Figure 6 in the Appendix), revealing notable class imbalances, where certain document classes are underrepresented compared to others, as illustrated in Figure 2. To address these discrepancies, traditional data augmentation techniques, including flipping, rotating, translation, scaling, random cropping, brightness adjustment, and Gaussian noise, are applied to increase diversity within each class. These augmentation techniques lead to a more balanced distribution of samples, enhancing the dataset's diversity, as seen in Figure 3, and ultimately improving the model's generalization ability.

The RVL-CDIP Dataset, a subset of the IIT-CDIP Test Collection, contains 400,000 grayscale images in 16 document categories (for example, see Figure 7 in the Appendix). From these categories, seven similar financial and business document types are selected to align with the CD, providing a consistent foundation for evaluating classification performance across various document types. This alignment allows a more comprehensive comparison of model effectiveness on different document distributions and class structures across the two datasets.

The models used for classification include various CNN architectures:

- CNN: Three convolutional layers followed by fully connected layers, based on the works of (Dong and Li, 2020), (Harley et al., 2015), and (Kang et al., 2014).
- LeNet5: A classic two-layer convolutional architecture from (Lecun et al., 1998).
- VGG19: A modified version of the VGG19 network, as outlined by (Simonyan and Zisserman, 2015), with single-channel input.
- MobileNetV2: Based on (Sandler et al., 2019), using depthwise separable convolutions for reduced complexity.

Experiments. Four experiments are conducted: two using the RFD dataset and two with a subset of RVL-CDIP. Each experiment employed 5-fold cross-validation, with an 80/20 split for training+validation and testing and another 80/20 split for training and validation. Results were generated from validation and testing, with testing accuracy calculated using models trained on combined training and validation sets. Evaluation metrics included mean accuracy, standard deviation, and testing accuracy across folds. The experiments aimed to assess the impact of data augmentation on the RFD dataset, evaluate pre-trained models on RVL-CDIP when applied to the augmented RFD dataset, measure model performance on the RVL-CDIP subset, and test the performance of models pre-trained on the augmented RFD dataset when applied to RVL-CDIP.

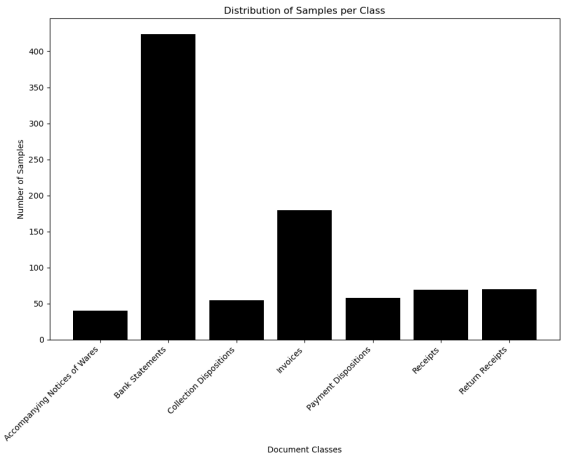

Figure 2: Number of samples per financial document class in RVL-CDIP dataset.

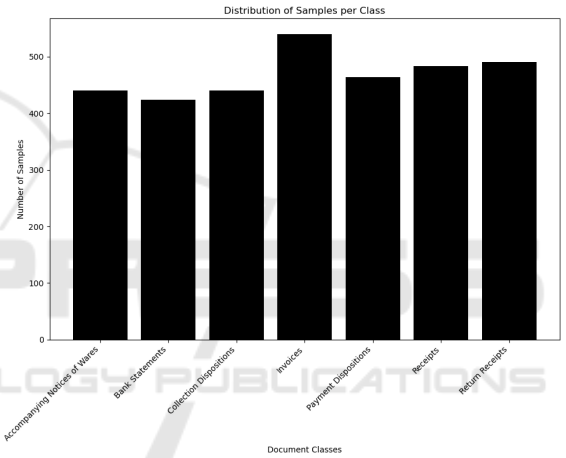

Figure 3: Number of samples per financial document class after data augmentation.

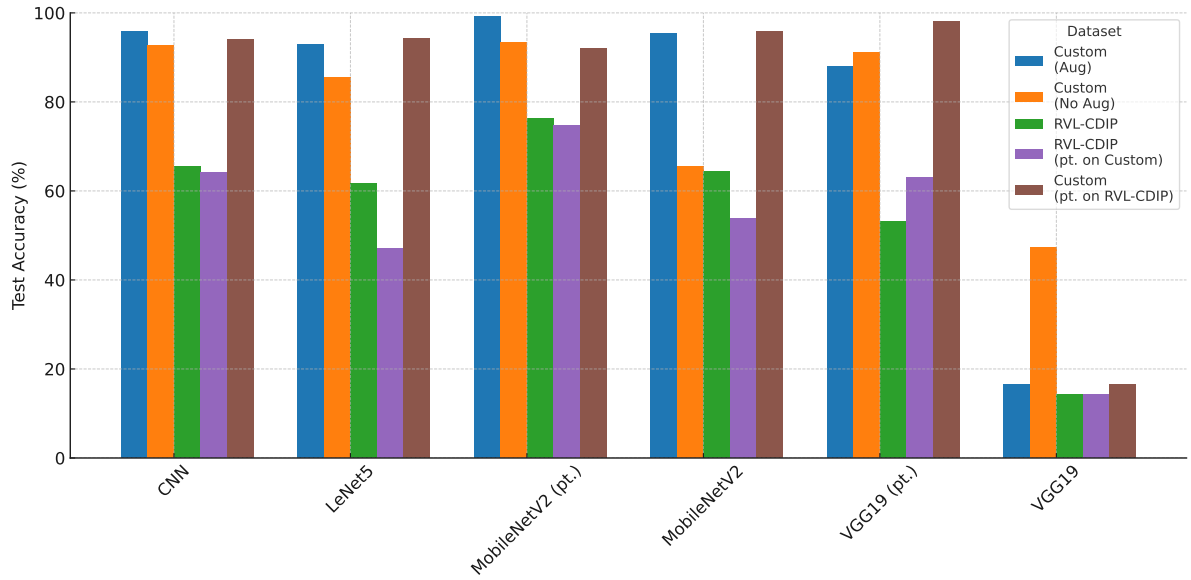

Figure 4: Test accuracy of various models on different datasets for the classification task (see also Tables 1, 2, 3, and 4).

Table 1: Classification accuracy using the RFD dataset with and without data augmentation.

Models	Test (Aug)	Val Mean (Aug)	Val Std (Aug)	Test (No Aug)	Val Mean (No Aug)	Val Std (No Aug)
CNN	**95.89**	95.16	0.98	**92.78**	92.87	1.01
LeNet5	92.94	89.25	2.43	85.56	81.28	4.37
MobileNetV2 (pretrained)	**99.24**	93.10	8.78	**93.33**	94.41	4.12
MobileNetV2	**95.28**	95.38	1.59	65.56	86.17	7.80
VGG19 (pretrained)	87.98	73.17	31.52	**91.11**	89.52	2.39
VGG19	16.44	14.51	1.40	47.22	47.33	4.56

2.2 Information Extraction

The information extraction task focuses on Key Information Localization and Extraction (KILE) and Line Item Recognition (LIR). Two datasets are used: the RFD dataset of manually annotated financial documents (35 documents across seven classes) and the DocILE Benchmark Dataset (Štěpán Šimsa et al., 2023). The DocILE dataset comprises 6.7k annotated business documents, 100k synthetic documents, and 1 million unlabeled documents. The RFD dataset includes fields like invoice numbers, item descriptions, and total values. These datasets test the ability to extract essential information from financial documents.

The models used for information extraction include:

- RoBERTa: A transformer model for text-based tasks, as described by (Liu et al., 2019).
- LayoutLMv3: A multimodal model handling both text and layout, based on the work of (Huang et al., 2022).
- GraphDoc: A graph-based model that integrates text, layout, and visual features, as described by (Wang et al., 2023).

Experiments: Two experiments were conducted: one on the DocILE dataset and the other on the RFD dataset. The first experiment used 5-fold cross-validation on DocILE, with models evaluated on F1-score, average precision (AP), precision, and recall. This experiment aimed to compare results with those of (Štěpán Šimsa et al., 2023) and to assess model robustness by combining KILE and LIR tasks for overall performance, unlike the separate evaluations in the original study.

In the second experiment, models trained on English documents were tested on a manually annotated RFD dataset, translated from Romanian to English. This experiment used F1-score, AP, precision, and recall metrics for both KILE and LIR tasks to evaluate model performance on novel, non-English data.

3 RESULTS

This section presents the classification results of the RFD and RVL-CDIP datasets (see Figure 4 for the plot). For the information extraction tasks, focusing on the DocILE benchmark and the RFD dataset, see

Figure 5 in the Appendix for the results plot. Detailed results are discussed in the following subsections.

3.1 Classification Task

RFD Dataset: The results for the classification task on the RFD dataset with and without data augmentation are shown in Table 1. The CNN model achieved the highest accuracy on unseen data (95.89%) with augmentation, while MobileNetV2 (pretrained) achieved the highest accuracy without augmentation (93.33%). Data augmentation significantly improved model performance across most models. Pretrained models, particularly MobileNetV2 and VGG19, consistently outperformed their non-pretrained counterparts.

RVL-CDIP Dataset: The results for the RVL-CDIP dataset are presented in Table 2. MobileNetV2 (pretrained) achieved the highest accuracy (76.34%), while CNN and LeNet5 followed with lower performance. Pretrained models performed significantly better than non-pretrained ones, highlighting the benefit of transfer learning.

RFD Dataset with RVL-CDIP Pretraining: The results for the RFD dataset with models pretrained on the RVL-CDIP dataset are in Table 3. MobileNetV2 (pretrained) and VGG19 (pretrained) achieved high accuracy, showing that models pretrained on RVL-CDIP can generalize well to the RFD dataset.

RVL-CDIP Dataset with RFD Pretraining: Table 4 shows that MobileNetV2 (pretrained) achieved the highest accuracy at 74.65%, followed by VGG19 (pretrained) at 62.98%. The CNN performed well with 64.21% accuracy. LeNet5 and non-pretrained models underperformed, with VGG19 scoring the lowest at 14.29%. Pretrained models generalize better to RVL-CDIP.

3.2 Information Extraction Task

DocILE Dataset: The results for the combined KILE and LIR tasks on the DocILE Benchmark Dataset are presented in Table 5. RoBERTa achieved the highest F1-score (0.7761), while LayoutLMv3 performed slightly lower. GraphDoc showed weaker performance compared to the other models.

RFD Dataset: The results for the RFD dataset for KILE and LIR tasks are presented in Table 6. RoBERTa showed strong performance in the KILE task but weaker in the LIR task. LayoutLMv3 performed well across both functions, while GraphDoc underperformed in comparison.

4 CONCLUSIONS

Modern architectures like MobileNetV2 (pretrained) and CNN, combined with data augmentation, delivered the best results for the classification tasks. Pretrained models consistently outperformed non-pretrained ones, emphasizing the importance of transfer learning. In document information extraction tasks, RoBERTa and LayoutLMv3 proved the most effective, with RoBERTa excelling in token-level predictions. LayoutLMv3 performed well, particularly for structured data extraction like line items. GraphDoc struggled in both tasks, indicating limitations for complex document processing. MobileNetV2 and CNN are best suited for classification, while RoBERTa and LayoutLMv3 excel in information extraction.

Hardware limitations impacted this work, particularly during model training for complex models like RoBERTa and LayoutLMv3. The RFD dataset used for classification was relatively small, which may have affected the models' ability to generalize across various document types. Furthermore, the annotations for information extraction were inadequate, and translation errors in the RFD dataset could have further hindered model performance. By addressing these challenges through the use of larger datasets, improved annotations, and enhanced computational resources, we are likely to achieve more accurate models.

Future research should concentrate on developing a benchmark specifically for classifying financial documents, as existing datasets may not adequately reflect the unique features of these documents. Additionally, creating multilingual models is crucial, particularly for processing languages like Romanian, to enhance the effectiveness of document extraction systems in various linguistic contexts. Expanding datasets in both size and diversity will be essential for improving the generalization and performance of models in classification and information extraction tasks.

In conclusion, this research proposes and analyzes machine learning pipelines for classifying and extracting information from financial documents. It demonstrates promising results in automating document processing and paves the way for further development in industrial applications.

Table 2: Classification accuracy using the RVL-CDIP dataset.

Models	Test Accuracy	Val Mean	Val Std
CNN	**65.59**	63.03	0.60
LeNet5	61.75	59.58	0.96
MobileNetV2 (pretrained)	**76.34**	72.64	2.49
MobileNetV2	**64.36**	58.94	1.18
VGG19 (pretrained)	53.15	51.89	20.14
VGG19	14.29	13.12	1.06

Table 3: Classification accuracy using the RFD dataset with models pretrained on the RVL-CDIP dataset.

Models	Test Accuracy	Val Mean	Val Std
CNN	94.06	93.63	0.96
LeNet5	**94.22**	91.15	1.20
MobileNetV2 (pretrained)	91.93	97.21	1.48
MobileNetV2	**95.88**	96.76	1.17
VGG19 (pretrained)	**98.17**	94.62	2.87
VGG19	16.44	16.46	1.91

Table 4: Classification accuracy using the RVL-CDIP dataset with models pretrained on the RFD dataset.

Models	Test Accuracy	Val Mean	Val Std
CNN	**64.21**	62.08	1.21
LeNet5	47.16	48.13	3.22
MobileNetV2 (pretrained)	**74.65**	73.59	1.45
MobileNetV2	**53.76**	55.37	0.73
VGG19 (pretrained)	62.98	60.40	1.13
VGG19	14.29	11.96	0.37

Table 5: Information extraction results (F1-score, average precision, precision, recall) for the 5-fold cross-validation on the DocILE dataset.

Models	F1-score	average precision	precision	recall
RoBERTa	**0.7761 ± 0.0268**	0.8194 ± 0.0204	0.8180 ± 0.0252	0.7650 ± 0.0253
LayoutLMv3	**0.7426 ± 0.0198**	0.7878 ± 0.0198	0.7752 ± 0.0260	0.7324 ± 0.0148
GraphDoc	0.4497 ± 0.0043	0.4974 ± 0.0059	0.5133 ± 0.0051	0.4402 ± 0.0060

Table 6: Information extraction results for KILE and LIR tasks on the RFD dataset, including F1-score, precision, recall, and average precision.

Models	F1-score (KILE)	Avg. precision (KILE)	precision (KILE)	recall (KILE)	F1-score (LIR)	Avg. precision (LIR)	precision (LIR)	recall (LIR)
RoBERTa	**0.5101**	0.0133	0.5101	0.5101	**0.2439**	0.1522	0.2439	0.2439
LayoutLMv3	**0.4973**	0.0152	0.4973	0.4973	**0.3396**	0.2234	0.3396	0.3396
GraphDoc	0.3033	0.0200	0.3033	0.3033	0.2052	0.1532	0.2052	0.2052

ACKNOWLEDGMENTS

We thank Vamadi Contab, the accounting company in Romania (https://vamadi.ro), for providing the financial documents that enabled the practical validation of our models. Finally, we appreciate the Hábrók HPC Cluster for providing computational resources, without which the article would not have reached its current depth and scope.

REFERENCES

Chen, N. and Blostein, D. (2007). A survey of document image classification: problem statement, classifier architecture and performance evaluation. In *International Journal of Document Analysis and Recognition (IJDAR)*, volume 10, pages 1–16.

Dong, J. and Li, X. (2020). An image classification algorithm of financial instruments based on convolutional neural network. In *Traitement du Signal*, volume 37, pages 1055–1060.

Ha, H. and Horák, A. (2022). Information extraction from scanned invoice images using text analysis and layout features. In *Signal Processing: Image Communication*, volume 102, page 116601.

Harley, A. W., Ufkes, A., and Derpanis, K. G. (2015). Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995.

Huang, Y., Lv, T., Cui, L., Lu, Y., and Wei, F. (2022). Layoutlmv3: Pre-training for document ai with unified text and image masking.

Kang, L., Kumar, J., Ye, P., Li, Y., and Doermann, D. S. (2014). Convolutional neural networks for document image classification. In *2014 22nd International Conference on Pattern Recognition*, pages 3168–3172.

Katti, A. R., Reisswig, C., Guder, C., Brarda, S., Bickel, S., Höhne, J., and Faddoul, J. B. (2018). Chargrid: Towards understanding 2d documents. *arXiv preprint*.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324.

Lehtonen, R., Nevalainen, P., and Murtojärvi, M. (2020). Automated classification of receipts and invoices along with document extraction. In *University of Turku*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Majumder, B. P., Potti, N., Tata, S., Wendt, J. B., Zhao, Q., and Najork, M. (2020). Representation learning for information extraction from form-like documents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6504.

Mikołajczyk-Bareła, A. and Grochowski, M. (2018). Data augmentation for improving deep learning in image classification problem. In *IIPHDW*, pages 117–122.

Oral, B., Emekligil, E., Arslan, S., and Eryiğit, G. (2020). Information extraction from text intensive and visually rich banking documents. In *Information Processing and Management*, volume 57, page 102361.

Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning.

Rusiñol, M., Frinken, V., Karatzas, D., Bagdanov, A. D., and Lladós, J. (2014). Multimodal page classification in administrative document image streams. In *International Journal on Document Analysis and Recognition (IJDAR)*, volume 17, pages 331–341.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2019). Mobilenetv2: Inverted residuals and linear bottlenecks.

Shorten, C. and Khoshgoftaar, T. (2019). A survey on image data augmentation for deep learning. In *Journal of Big Data*.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Wang, Y., Du, J., Ma, J., Hu, P., Zhang, Z., and Zhang, J. (2023). Ustc-iflytek at docile: A multi-modal approach using domain-specific graphdoc. In *Conference and Labs of the Evaluation Forum*.

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., and Shen, F. (2023). Image data augmentation for deep learning: A survey.

Ömer Arslan and Uymaz, S. A. (2022). Classification of invoice images by using convolutional neural networks. In *Journal of Advanced Research in Natural and Applied Sciences*, volume 8, pages 8–25. Çanakkale Onsekiz Mart University.

Štěpán Šimsa, Šulc, M., Uřičář, M., Patel, Y., Hamdi, A., Kocián, M., Matas, J., Doucet, A., Coustaty, M., and Karatzas, D. (2023). Docile benchmark for document information localization and extraction.

APPENDIX

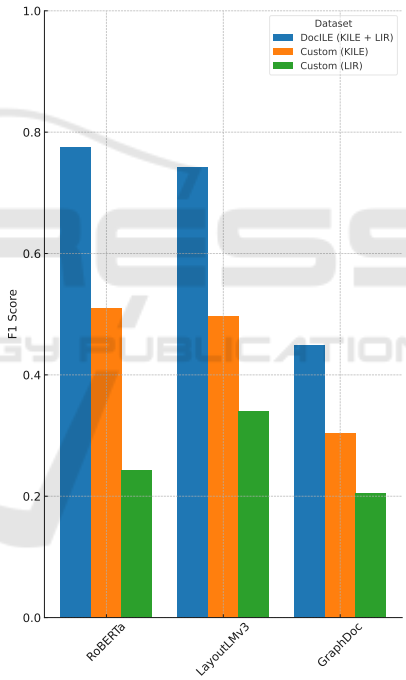

Figure 5: F1-scores of different models tested on different datasets for the information extraction task.

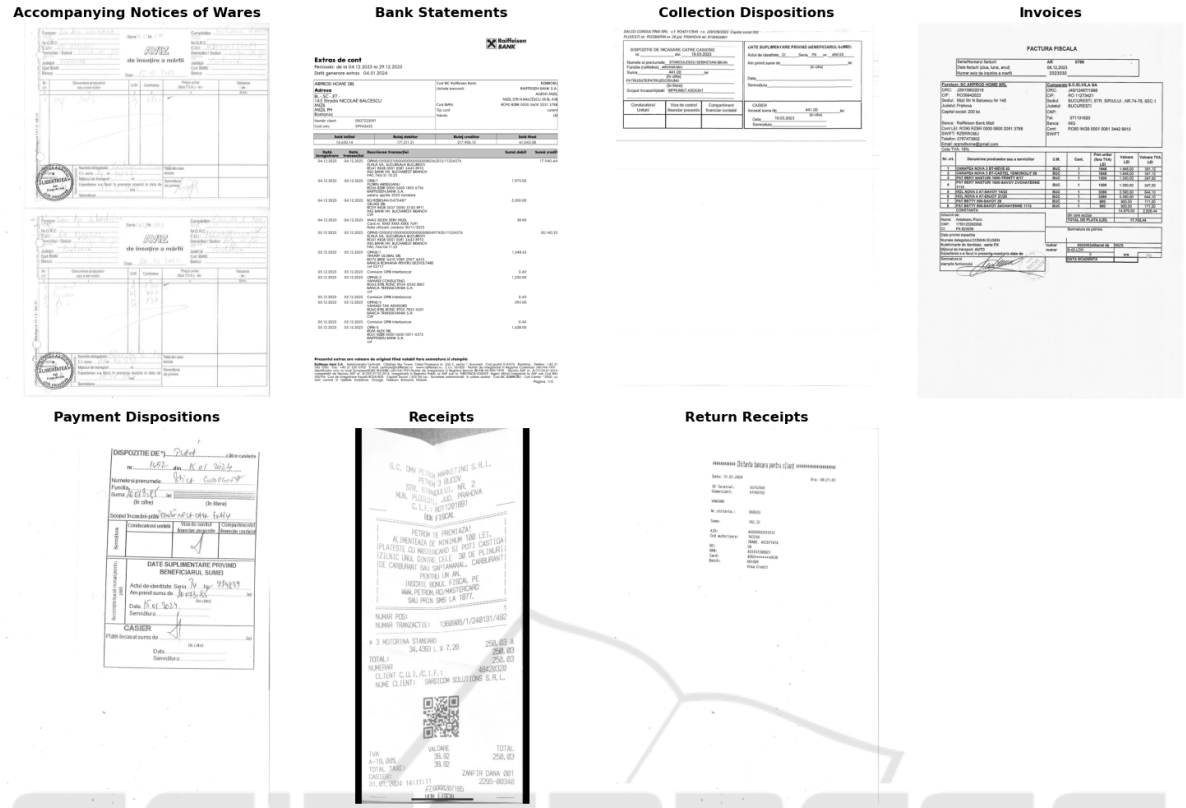

Figure 6: Example documents and classes from RFD dataset used for the image classification task.

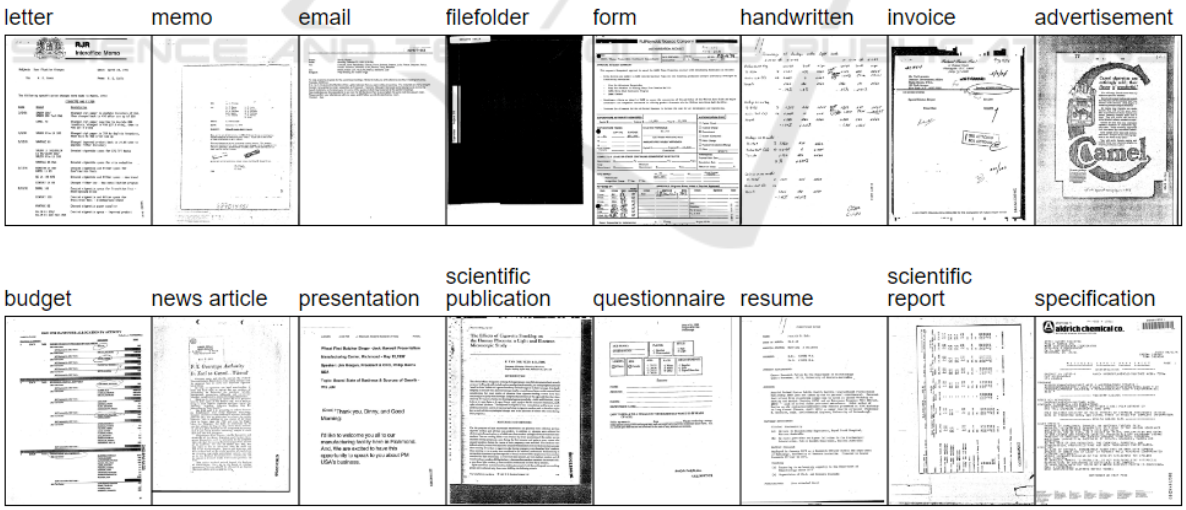

Figure 7: Example documents and classes from the RVL-CDIP Dataset, figures taken from the study of (Harley et al., 2015).