OCR-Guided Occlusion Sensitivity for Auditing Document Classifiers

Valerio Caravani[1,2,*,†], Riccardo De Cesaris[2,†] and Paolo Merialdo[2,†]

[1]myBiros.com, Rome, Italy
[2]Department of Engineering of Roma Tre University, Rome, Italy

Abstract

Document classifiers are increasingly used in settings where predictions must be inspected and audited. However, common perturbation-based explanation methods operate on arbitrary image regions, which poorly reflect the textual structure through which documents convey information. We introduce OCR-guided occlusion sensitivity, a black-box compatible protocol that uses OCR-detected text segments as interpretable perturbation units. Each segment is associated with recognised text and is removed consistently from every input modality through which it is represented. We conduct an exploratory evaluation on a stratified subset of the RVL-CDIP test set using the vision-only DiT and the multimodal LayoutLMv3. The results show that removing the highest-ranked OCR segments produces larger cumulative drop in the score for the predicted class than removing the same number of randomly selected segments. Inspection of the content associated to the regions with the highest importance scores reveals that the predictions of both models are partly driven by corpus-specific provenance cues that are not intrinsically informative for the document classification task itself. We also observe an association between classification errors and instances whose predictions rely on only a few highly important segments, suggesting a potential indicator of prediction fragility. Overall, the study illustrates how OCR-guided occlusion can support the black-box auditing of document classifiers while producing explanations linked to readable document content.

Keywords

Explainable AI, Document Classification, Perturbation-based Explanations, OCR, Model Auditing

1. Introduction

Automated document classification is increasingly used in domains such as financial services, public administration, and legal document management, where predictions may need to be inspected and audited. High classification accuracy alone is insufficient in these settings: understanding whether a model relies on semantically meaningful content or on incidental dataset-specific cues is also important.

Post-hoc explanation methods provide a way to inspect model decisions, but most were originally developed for natural images. Pixel-level saliency and fixed-grid perturbations do not reflect the textual structure of documents: they may split words, merge unrelated fields, or remove mostly whitespace, making the resulting explanations difficult to interpret [1, 2, 3, 4, 5, 6]. Document-specific methods such as DocXplain [7] improve spatial alignment through segmentation-based perturbations. For model auditing, however, it is also useful to associate influential regions with readable text and, for multimodal classifiers, to remove the same information consistently from all input modalities.

We introduce OCR-guided occlusion sensitivity, a black-box compatible protocol that uses OCR-detected text segments as perturbation units. The importance of a segment is estimated from the change in the predicted-class score when that segment is removed. Since each segment is linked to recognised text, the resulting attribution can be inspected both spatially and semantically. Given its model agnostic nature, the protocol requires only query access to model output scores. It does, however, require control over the input representations consumed by the classifier, in order to adapt the perturbations to the modalities considered by the model itself.

XAI.it 2026 - 6th Workshop on Explainable AI, Generative and Agentic Systems: Trust, Transparency, and Human Oversight, co-located with AixiA 2026, October 6-9, 2026, Perugia, Italy

*Corresponding author.
†These authors contributed equally.
valerio.caravani@mybiros.com (V. Caravani); riccardo.decesaris@uniroma3.it (R. De Cesaris); paolo.merialdo@uniroma3.it (P. Merialdo)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-10-30

CEUR-WS.org/Vol-4308/paper3.pdf

We conduct an exploratory evaluation on a stratified subset of RVL-CDIP [8], considering the vision-only DiT [9] and the multimodal LayoutLMv3 [10] as classifiers. We examine whether removing segments according to their estimated importance produces larger cumulative drops in the predicted-class score than random segment removal, whether inspection of the associated text reveals reliance on spurious cues, and whether classification errors are associated with predictions that depend on only few highly important units.

For both models, removing the highest-ranked OCR segments causes larger confidence drops than random segment removal. Inspection of the associated text identifies archive codes among the most influential segments, providing attribution-level evidence consistent with reliance on corpus-specific provenance cues previously discussed for RVL-CDIP [11]. We also observe an association between strong single-segment sensitivity and misclassification.

This short paper makes two contributions. First, it defines an OCR-guided perturbation protocol that produces text-linked explanations and removes each segment consistently from the input modalities consumed by the classifier. Second, it provides an exploratory evaluation showing how these explanations can support model auditing by concentrating predictive evidence, characterising corpus-specific cues, and highlighting potentially fragile predictions. The study establishes the protocol and motivates a broader evaluation across models, explanation methods, and document datasets.

2. Related Work

Post-hoc attribution methods include gradient-based approaches, such as Grad-CAM [1] and Integrated Gradients [4], and model-agnostic approaches based on input perturbations, such as LIME [2], SHAP [5], and occlusion sensitivity [3]. These methods require only query access to model output, but the resulting explanations are highly sensitive to how input instances are segmented into sets of interpretable units. In document images, fixed-grid partitions may split words, combine unrelated elements, or cover mostly whitespace, making their attributions difficult to interpret [6].

Among document-specific approaches, DocXplain [7] first separates foreground from background and then applies heuristic image processing techniques operating directly on the document pixels to segment foreground elements into explanatory units. Explanations are subsequently generated by estimating an attribution score for each unit. Our protocol differs primarily in how explanatory units are constructed and in the form of the resulting explanations.

It uses OCR-detected lines as explanatory units, linking each attribution to recognised text that can be inspected during model auditing. Moreover, for multimodal classifiers, the corresponding evidence is removed consistently from the image, textual tokens, and layout representation. This interpretability comes with a dependence on OCR coverage and segmentation quality.

Our auditing perspective is also related to analyses of the RVL-CDIP dataset [8]. Previous studies have documented label noise, ambiguous samples, train–test overlap, and possible reliance on corpus-specific artefacts [11]. While these studies identify limitations at the benchmark level, we investigate whether text-linked attributions can localise and characterise the evidence used in individual predictions.

3. OCR-Guided Occlusion Protocol

Let x denote a document and let $\mathcal{S}(x) = \{s_1, \dots, s_n\}$ be its OCR-detected text segments. Let c be the class predicted by a classifier f for the input x, and let $f_c(x)$ denote the scalar score assigned to class c, such as the predicted probability. Given a segment $s \in \mathcal{S}(x)$, let $x \setminus s$ denote the perturbed input obtained by occluding that segment. When the classifier operates on multiple input modalities, the evidence associated with s is removed consistently from every modality in which it is represented. Following the standard occlusion sensitivity [3] principle, we define the importance of s as

$$\mathrm{imp}_x(s) = f_c(x) - f_c(x \setminus s). \tag{1}$$

The target class c is kept fixed when evaluating the perturbed input. A positive value indicates that the segment supports the original prediction, whereas a negative value indicates that its removal increases the score assigned to c. The resulting importance score is a *leave-one-out* estimate: it measures the effect of removing the s while leaving all other segments unchanged. It therefore characterises the contribution of s in the context of the complete document, but does not capture how that contribution changes when other segments are removed simultaneously.

Segments are ranked in decreasing order of $\mathrm{imp}_x(s)$. Let $\mathcal{S}_K(x) \subseteq \mathcal{S}(x)$ denote the set of the K highest-ranked segments. We measure the effect of their joint removal through

$$\mathrm{Drop@K}(x) = f_c(x) - f_c(x \setminus \mathcal{S}_K(x)). \tag{2}$$

Whereas the importance scores are obtained from individual ablations, $\mathrm{Drop@K}(x)$ measures the variation in the classifier's confidence under joint removals of the highest-ranked segments. Examining its values as K increases therefore provides a behavioural assessment of whether the individual importance scores, and the ranking they induce, remain informative about the evidence that drives the classifier's prediction for x.

As a reference, $\mathrm{RandomDrop@K}(x)$ measures the confidence drop obtained by removing K uniformly sampled OCR segments. Comparing the two measures assesses whether the ranking induced by importance scores concentrates predictive evidence more effectively than arbitrary segment selections. Since confidence scores may be calibrated differently across classifiers, drops are interpreted within each model rather than used for direct cross-model comparison.

4. Experimental Setup

Dataset and models. We initially sample 125 documents per class from the official RVL-CDIP test split [8], for a total of 2,000 documents. Ten documents with no valid OCR-detected segments are excluded, yielding a final evaluation set of 1,990 documents. We consider two classifiers with different input modalities: the vision-only DiT [9][1] and the multimodal LayoutLMv3 [10][2]. On the selected subset, DiT achieves 93.2% accuracy and LayoutLMv3 94.7%.

OCR and perturbations. Text segments and bounding boxes are extracted at line level using AWS Textract[3]. We retain at most 100 segments per document. The resulting sample contains an average of 43.3 segments per document and a median of 38, while 9% of documents reach the imposed limit. For DiT, a segment is removed by replacing the pixels within its bounding box with white pixels. For LayoutLMv3, the same visual region is masked, and the corresponding OCR tokens and layout coordinates are removed.

Evaluation protocol. Segment importance is computed independently for each OCR segment. We evaluate progressive occlusion for $K \in \{1, 3, 5, 10, 20, 50\}$. For a document containing n segments, the effective number of removed segments is $K' = \min(K, n)$, for both ranked and random removal. At each value of K, we compare Drop@K with RandomDrop@K, obtained by uniformly sampling the same number of segments. Results are aggregated over the full subset and, where relevant, analysed by document class.

5. Results

5.1. OCR-guided ranking concentrates predictive evidence

Figure 1 shows the mean score assigned to the originally predicted class as progressively more OCR segments are removed. For both models, confidence decreases more rapidly when segments are removed

[1] https://huggingface.co/microsoft/dit-base-finetuned-rvlcdip
[2] https://huggingface.co/gordonlim/layoutlmv3-base-finetuned-rvlcdip
[3] https://aws.amazon.com/it/textract/

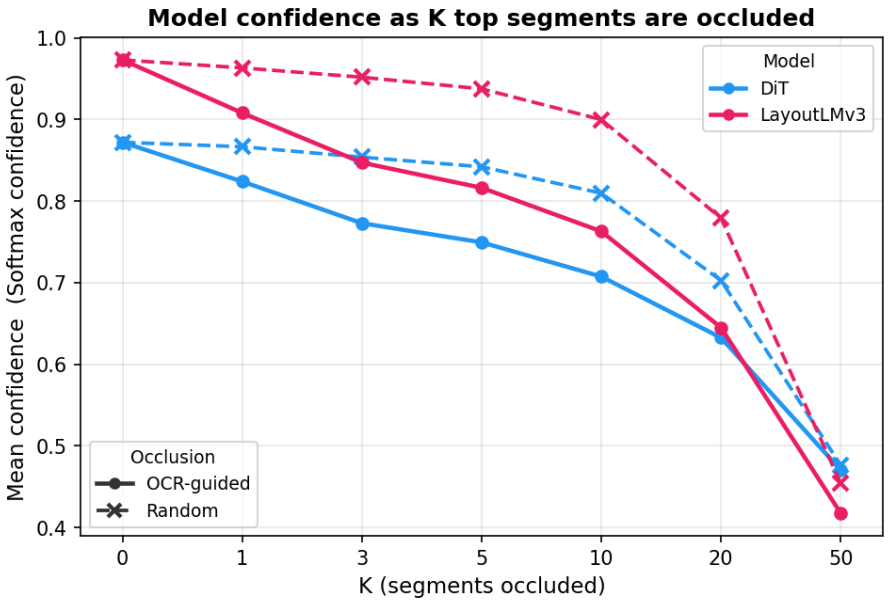


Figure 1: Mean predicted-class confidence under progressive segment removal. Solid lines show the confidence remaining after removing the highest-ranked OCR segments, while dashed lines show the removal of an equal number of randomly selected segments. The larger separation at small values of K indicates that the proposed ranking concentrates influential predictive evidence. Confidence values are interpreted within each model and are not used for direct cross-model comparison.

according to the ranking induced by the importance estimates than when the same number of segments is selected at random. The difference is most pronounced at small values of K, where the objective is to identify a limited set of influential segments.

At $K = 3$, ranked removal produces a confidence drop 5.4× larger than random removal for DiT and 5.9× larger for LayoutLMv3. As larger portions of the document are removed, the ranked and random curves progressively approach each other because the ordering of the removed segments becomes less influential.

Overall, the results show that the ranking induced by importance scores identifies segments whose cumulative removal produces larger reductions in the predicted class score than random removal of the same number of OCR segments.

5.2. Text-linked attributions reveal archive identifiers

Since each attribution is linked to recognised text, we can inspect the content of the segments that most influence a prediction. For both models, the highest-ranked segment is frequently an archive identifier: a string, such as `TIMN 0171637` or `TIFL 0526958`. Figure 2 illustrates a case in which removing such an identifier produces a substantial decrease in the predicted-class score.

Candidate archive identifiers account for approximately 4.5% of all detected OCR segments, but represent approximately 26% of the highest-ranked segments for DiT and 21% for LayoutLMv3. They are therefore over-represented among top-ranked segments by factors of approximately 5.8 and 4.7, respectively. This provides attribution-level evidence consistent with both classifiers relying in part on corpus-specific provenance cues, in line with previously reported limitations of RVL-CDIP [11].

These findings show how text-linked attributions can help characterise corpus-specific evidence used by document classifiers.

5.3. Single-segment sensitivity is associated with errors

We finally examine whether strong dependence on a single OCR segment is associated with classification errors. For each model, documents are ranked by Drop@1, and the misclassification rate among the

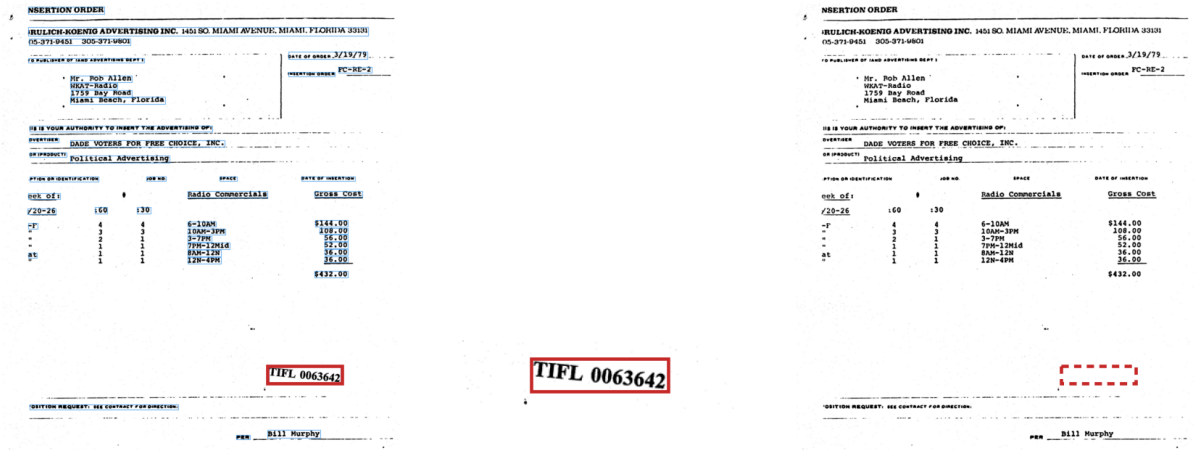


Figure 2: Example of a text-linked attribution (Model: DiT, RVL-CDIP's instance of the *budget* class). Left to right: the page with its 45 OCR line segments, the top-ranked one outlined in red; that segment magnified, the archival stamp `TIFL 0063642`; the page with the segment occluded. Occluding it decreases the score of the originally predicted class from 0.908 to 0.569, despite the segment containing corpus provenance metadata rather than document-type information.

10% of documents with the largest values is compared with the rate over the complete evaluation set.

For both models, the highest-sensitivity decile exhibits a misclassification rate approximately three times larger than the overall rate. This association suggests that predictions whose confidence depends strongly on one segment may be more fragile.

Since Drop@1 can be computed without access to ground-truth labels at inference time, it may provide a signal for prioritising predictions for human inspection. This result is exploratory, however, and does not establish Drop@1 as an error-detection criterion: assessing its precision, recall, and practical utility would require a dedicated evaluation.

6. Discussion and Limitations

The main advantage of OCR-guided occlusion is that influential regions can be inspected as readable document content rather than only as spatial highlights.

The Drop@K trends provide an additional view of the classifier's behaviour as progressively larger portions of the most influential document content are removed. The observed score degradation as (K) increases is consistent with the expected progressive loss of relevant evidence and suggests that the ranking captures regions that contribute cumulatively to the prediction.

In our experiments, the attribution scores are often concentrated on a small number of segments, indicating that some predictions depend heavily on only a few document elements.

The association between large Drop@1 values and classification errors also suggests a possible operational use of the protocol. Predictions that depend strongly on a single segment may be more fragile and could be prioritised for inspection. This observation remains exploratory and does not establish segment sensitivity as a reliable error-detection criterion.

Several limitations should be considered. Filling an occluded region with white pixels may introduce an input pattern that differs from the model's training distribution. Consequently, the observed score change may reflect both the removal of document information and the visual artefact introduced by masking. The protocol also depends on OCR coverage and segmentation quality, while the archive-identifier analysis relies on a pattern-matching heuristic.

The computational cost is another practical limitation. Segment-level occlusion requires an additional model inference for each analysed region, with further evaluations needed to compute Drop@K under different removal configurations. The total explanation time therefore grows with the number of OCR

segments and may become substantial for dense documents or large-scale analyses.

The study is limited to two classifiers and a subset of one benchmark, with LayoutLMv3 evaluated using a community-provided Hugging Face checkpoint rather than an official release. It also uses random removal as the main reference and estimates segment importance independently, without modelling interactions among regions. Broader evaluation across datasets, architectures, checkpoints, masking strategies, and explanation baselines is therefore required.

7. Conclusion

We introduced OCR-guided occlusion sensitivity as a black-box approach for connecting model attributions with the readable content of document regions. The experiments on RVL-CDIP show that OCR segments provide meaningful perturbation units for investigating which document elements support a classifier's predictions and for exposing cases of concentrated or potentially undesirable reliance.

Future work will evaluate the approach across additional datasets, architectures, and explanation baselines, and investigate alternative masking strategies and methods that account for interactions among segments. Further research should also explore more efficient attribution procedures for dense documents and large-scale deployment.

8. Acknowledgments

This work was supported by myBiros S.r.l. Valerio Caravani's PhD fellowship was co-funded by myBiros S.r.l. and the Lazio Region under the Public Notice issued pursuant to Decreto Dirigenziale No. G06899 of 8 June 2021 (Grant No. F83C22001250005). Riccardo De Cesaris's PhD fellowship was funded by Seeweb S.r.l. and the European Union under the NextGenerationEU programme, Mission 4, Component 2, "From Research to Business," Investment 3.3, pursuant to Italian Ministerial Decree No. 630 of 24 April 2024 (Grant No. F81J24000370004).

Declaration on Generative AI

During the preparation of this work, the authors used Claude by Anthropic in order to: grammar and spelling check, paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[2] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[3] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.

[4] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International conference on machine learning, PMLR, 2017, pp. 3319–3328.

[5] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[6] S. Agne, A. Dengel, S. Ahmed, et al., The reality of high performing deep learning models: A case study on document image classification, IEEE Access 12 (2024) 103537–103564.

[7] S. Saifullah, S. Agne, A. Dengel, S. Ahmed, DocXplain: a novel model-agnostic explainability method for document image classification, in: International Conference on Document Analysis and Recognition, Springer, 2024, pp. 103–123.

[8] A. W. Harley, A. Ufkes, K. G. Derpanis, Evaluation of deep convolutional nets for document image classification and retrieval, in: 2015 13th international conference on document analysis and recognition (ICDAR), IEEE, 2015, pp. 991–995.

[9] J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, F. Wei, DiT: Self-supervised pre-training for document image transformer, in: Proceedings of the 30th ACM international conference on multimedia, 2022, pp. 3530–3539.

[10] Y. Huang, T. Lv, L. Cui, Y. Lu, F. Wei, LayoutLMv3: Pre-training for document ai with unified text and image masking, in: Proceedings of the 30th ACM international conference on multimedia, 2022, pp. 4083–4091.

[11] S. Larson, G. Lim, K. Leach, On evaluation of document classification with RVL-CDIP, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 2665–2678.