Stain Busters

Goal: Fast, objective, and non-invasive AI-based method for automatic stain detection.

Why Stain Detection?

Abstract: AI-based method that uses simple mobile phone photos to quickly identify harmful stains on historical documents and help archivists decide which items need professional restoration.
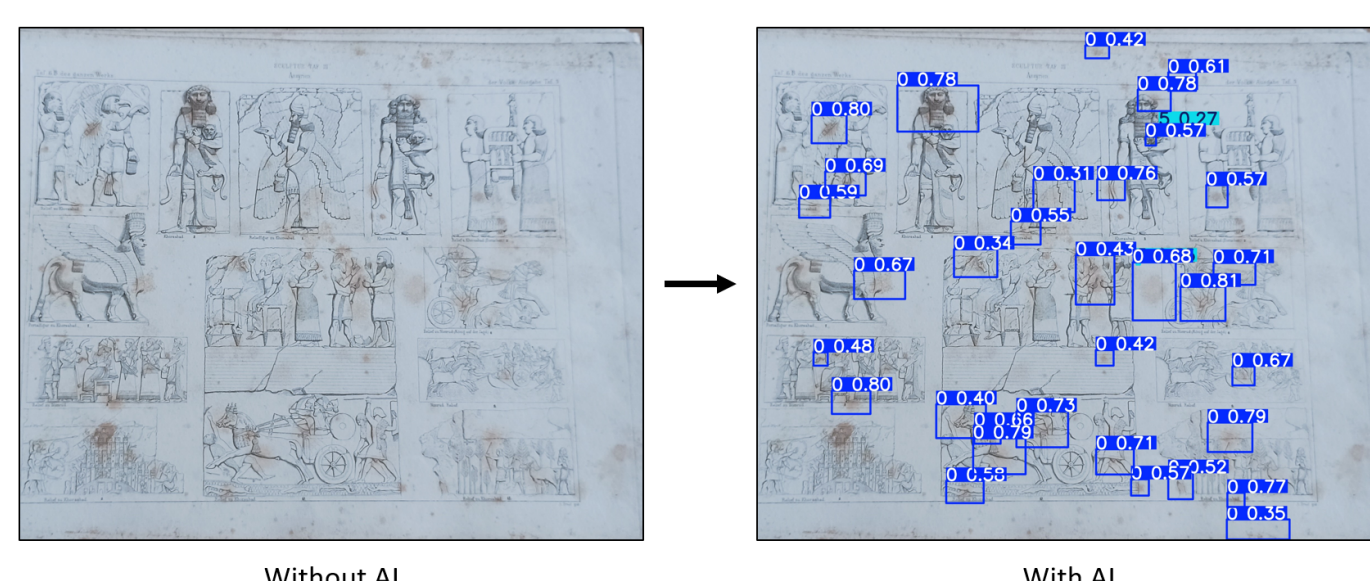- *Fungi, bacteria, insects and foxing*.

Without AI → With AI

Figure 1. Without vs. with AI-assisted detection.

Real Dataset

Semi-automatic annotation.
218 annotated images, 11 387 bounding boxes.
- Automated preprocessing using HSV thresholding – initial BBs.
- Manual correction – *Roboflow*.

Figure 2. Example of annotation process in Roboflow.

Synthetic Dataset

- 651 images, 2 classes (*foxing*, *feces*)
- Stains extracted using SAM
- Generated on undamaged papers

Original Degraded Document with Foxing — Semi Automated Segmentation using SAM — Segmanted Samples — Historical Document — Synthetic Data

Figure 4. Workflow for generating synthetic data.

Example of Synthetic Data

Figure 3. Example of syn. data classes foxing and feces.

Results

YOLOv8 architectures: *n, s, m, s_param*.
- 3 dataset variants:
  - Synthetic dataset.
  - Real dataset.
  - Combined dataset.
- Evaluation metrics: Precision, Recall, mAP@50, mAP@50–95.

Dataset	Precision	Recall	mAP@50	mAP@50-95
Synthetic	0.117	0.132	0.0597	0.0137
Real	**0.386**	**0.343**	**0.297**	**0.101**
Combined	0.323	0.316	0.2690	0.0897

Table 1. Best model results across datasets.

Visual Results – YOLOv8s

Comparison of predictions from models trained on different datasets:
- Green – Ground Truth.
- Blue – Combined dataset.
- Violet – Synthetic dataset.
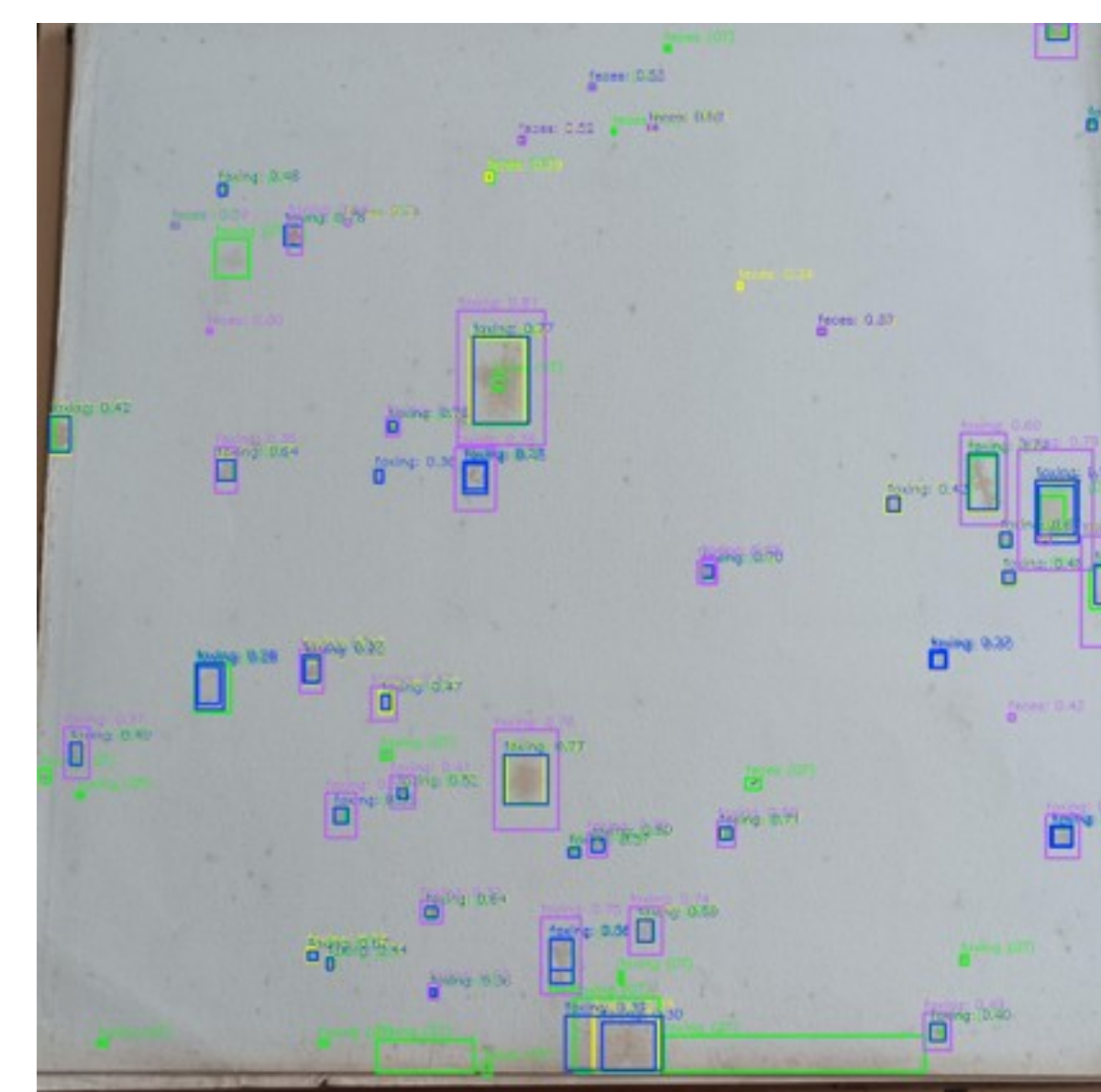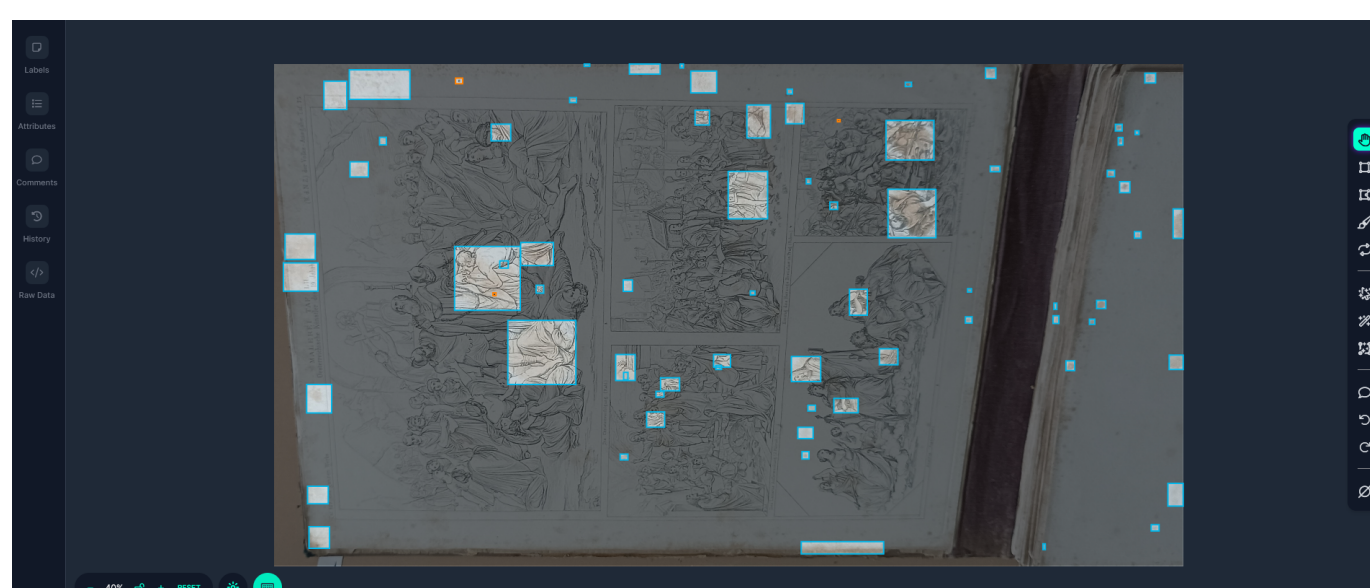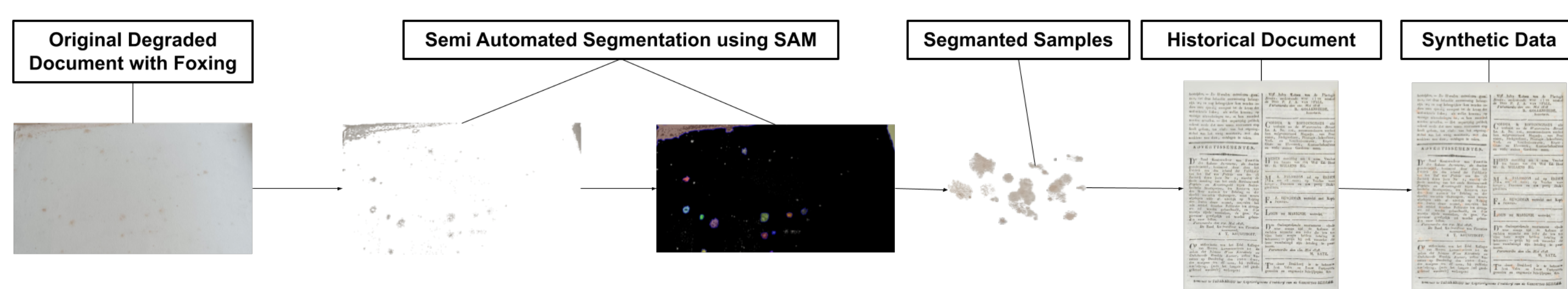- Yellow – Real dataset.

Figure 5. Visual comparison of predictions across dataset variants (YOLOv8s model).

Future Work

- Expand real dataset.
- Explore various synthetic data generation methods (*GANs, style transfer*).
- Generate and analyze artificially aged or damaged documents to simulate real deterioration.
- Test advanced architectures: *Faster R-CNN, RT-DETR, EfficientDet*.

References

[1] Yuxian Lin, Chenqi Xu, and Shuqiang Lyu. Disease regions recognition on mural hyperspectral images combined by mnf and bp neural network. In *Journal of Physics: Conference Series*, volume 1325, page 012095. IOP Publishing, 2019.

[2] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY
Univerzita Komenského
v Bratislave

Miriam Grznárová, Zuzana Berger Haladová, Jiří Hladůvka, Viktor Kocur, Zuzana Černeková
Department of Applied Informatics, Faculty of Mathematics, Physics and Informatics, Comenius University, Slovakia

Robert Bohdal, Andrej Ferko
Department of Algebra and Geometry, Faculty of Mathematics, Physics and Informatics, Comenius University, Slovakia

MATFYZ CONNECTIONS