Multimodal Approaches for Visually-Rich Document Type Classification: A Comparative Analysis

Catyana Heyne

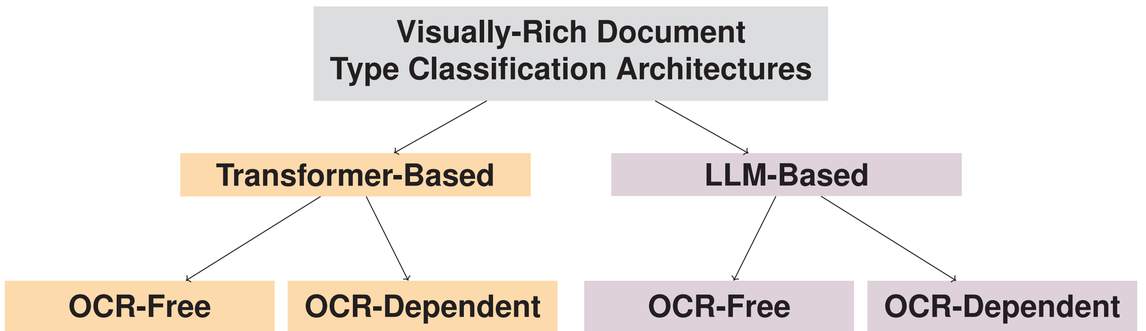
INTRODUCTION

Automated document type classification is a key component of modern **Document AI** systems and enables business process automation, intelligent document routing, and efficient information processing. While traditional linear documents mainly consist of plain text, enterprise documents such as invoices, forms, reports, and letters represent **Visually-Rich Documents** that combine textual, structural, and visual information.

To process these heterogeneous documents, recent research introduced diverse **multimodal Transformer- and LLM-based architectures** that differ in architecture, modality integration, and Optical Character Recognition (OCR) utilization. As a result, consistent evaluation and direct comparison across approaches is challenging.

METHODOLOGY

Representative multimodal architectures were evaluated on the RVL-CDIP benchmark dataset under a unified experimental setup, comparing Transformer- and LLM-based approaches as well as OCR-free and OCR-dependent model designs.



Model	*Donut*	*LayoutLMv3*	*Qwen3-VL*	*Qwen3*
Modalities	Image	Image + Layout[1] + Text[1]	Image	Text[1]
Architecture	Encoder-Decoder Transformer	Encoder-Only Transformer	Vision-Language LLM	LLM
Specialization	Fine-Tuned on RVL–CDIP	Fine-Tuned on RVL–CDIP	General-Purpose Model	General-Purpose Model
Parameters	143M	133M	32B	32B

[1] External OCR integration based on Tesseract.

RESEARCH OBJECTIVES

A systematic comparison of multimodal architectures for visually-rich document type classification.

- Compare representative Transformer- and LLM-based models
- Analyze multimodal feature integration with particular focus on OCR utilization
- Identify strengths and limitations of current multimodal approaches

RVL-CDIP DATASET

The RVL-CDIP benchmark dataset contains visually-rich document images annotated with 16 document types. Experiments were conducted on the official test split comprising 40,000 samples.

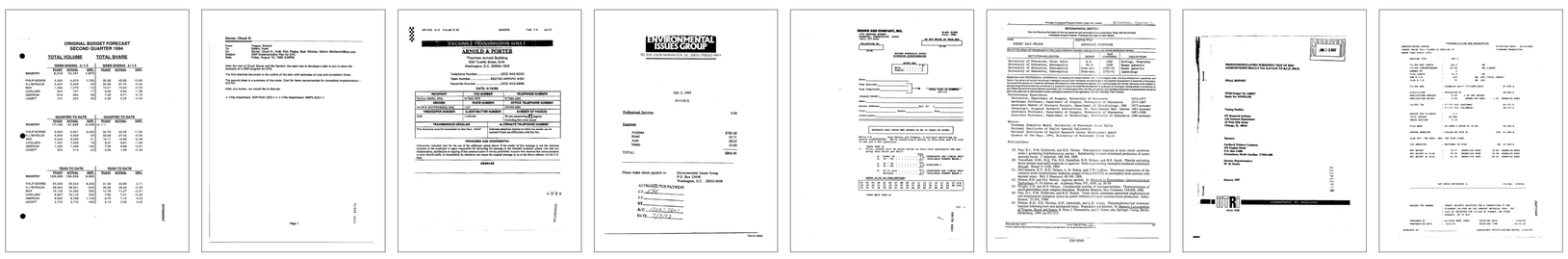


Figure: Representative VRD samples from the RVL-CDIP benchmark dataset (https://huggingface.co/datasets/chainyo/rvl-cdip).

RESULTS

Classification Accuracy

Fine-tuned Transformers achieve higher accuracy than general-purpose LLMs.

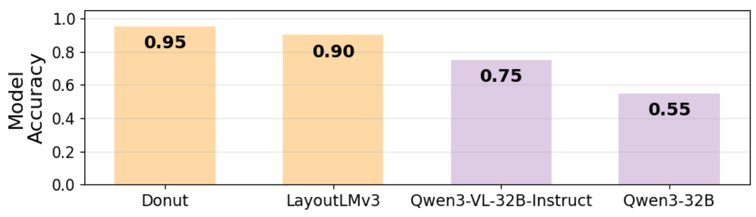


Figure: Accuracy results.

Multimodality in Visually-Rich Documents

Multimodal approaches outperform text-only LLMs, emphasizing the importance of visual representations for visually-rich document understanding.

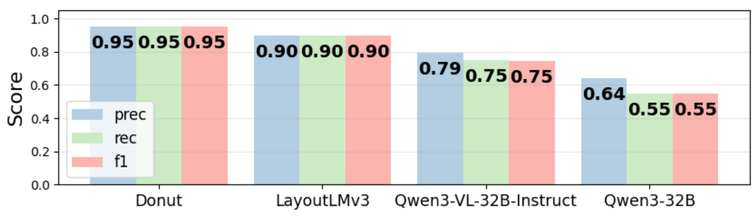


Figure: Macro-averaged precision, recall and f1-score.

Multimodality in Linear Documents

Textual information is highly informative for linear document types (e.g., emails), while multimodal architectures improve classification robustness.

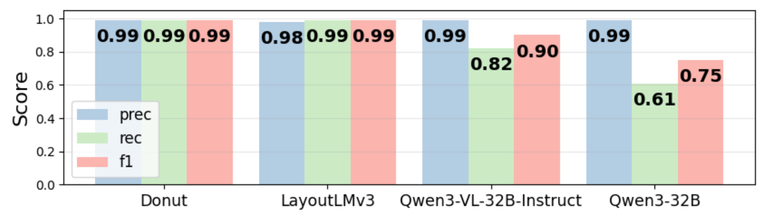


Figure: Classification results for email documents.

Inference Time

OCR-free architectures achieve significantly lower average inference times.

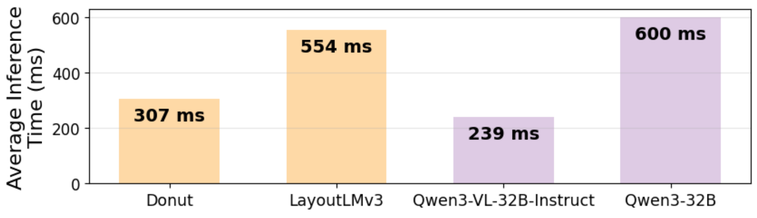


Figure: Inference time results per document.

Keyword-Dependent Performance

Document types characterized by distinctive textual patterns (e.g., resumes) can be classified reliably, even using text-only LLMs.

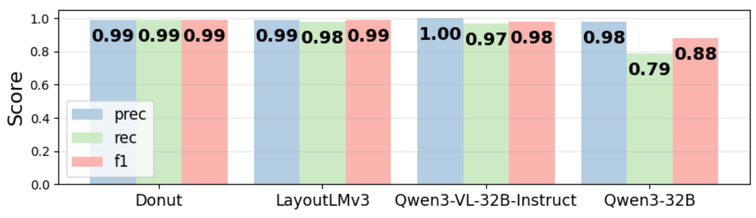


Figure: Classification results for resume documents.

Well-Defined Label Sets

Classification performance depends on clearly distinguishable categories, while document types with inconsistent layouts and ambiguous structures (e.g., forms) lead to increased misclassifications.

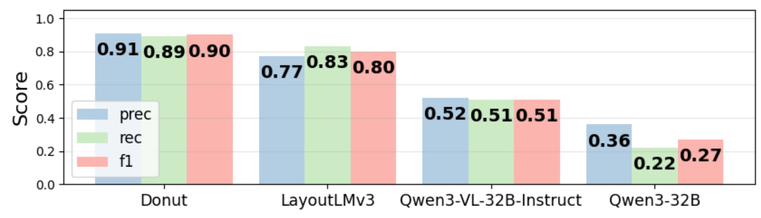


Figure: Classification results for form documents.

BAICERT PROJECT

This analysis was conducted within the BAICert project, which develops **AI-driven solutions for automated document processing** in certification bodies, public authorities, and other high-volume document environments.

ACKNOWLEDGEMENTS

This research is supported by the **Bavarian Collaborative Research Programme BayVFP** of the Bavarian State Ministry for Economic Affairs, Regional Development and Energy, funding line “Digitalization”, funding area “Information and Communication Technology”.

CONTACT

Laboratory Imaging and Data Science
Prof. Dr. Jürgen Frikel
Prof. Dr. Filippo Riccio
Catyana Heyne

Collaboration Partner:
OmniCert Umweltgutachter GmbH