

GeoAITest

Automated Data & Image Extraction from Technical Reports

Project Overview

This section outlines the main goal and overall approach of the GeoAITest project. It aims to provide a high-level understanding of how GeoAITest leverages AI technologies for technical document analysis.

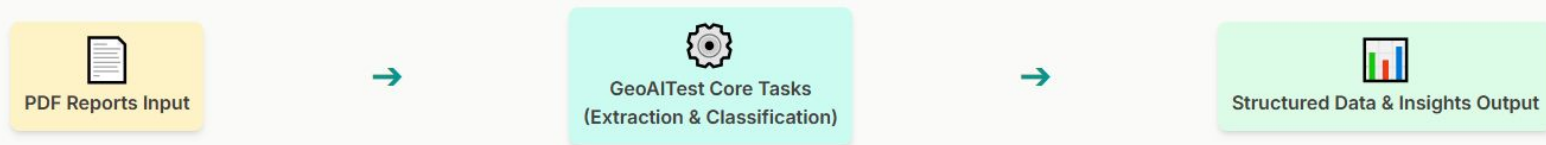
Goal

Automate extraction, classification, and structuring of data from technical PDF reports (financials, images, tables, geological maps).

Approach


Combine computer vision, NLP, and OCR for robust, scalable analysis.

GeoAITest Process Flow



This diagram illustrates the high-level workflow of GeoAITest, from initial PDF document ingestion through automated processing to the final structured data output.

Core Tasks

A horizontal bar with a teal segment on the left and an orange segment on the right.

GeoAITest is comprised of several key tasks, each focused on a specific aspect of data extraction and analysis. Click on each task to learn more about its approach, tools used, challenges faced, and planned improvements.

- Task 1: Table Extraction with AI
- Task 2: Image Extraction & Classification
- Task 3: Geospatial & Technical Info Extraction

Task 1: Table Extraction with AI

Approach:

- Extract key financial/operational tables, even as images.
- Use OpenAI GPT-4o (multimodal) for table detection & structuring.

Tools/Models:

pdfplumber, pdf2image, OpenAI API, pandas.

Challenges:

Challenge: Table images, inconsistent layouts.

Result: Excel tables, robust to scanned docs.

Results:

The process extracted 100 % of the robust_cash_flow figures for 2023 and the third quarter of 2024; while highly accurate, it can still be improved.ables, even a

Improvement:

Fine-tune prompts, add table structure validation; Implement a validation step to measure extraction accuracy; consider cost-effective LLMs like Gemini, Claude or Llama 3 (or local open-source models for privacy/cost); and use Camelot, Tabula or PyPDF2 for text-based tables (noting they may struggle with images).

Metric	Q1 2023	Q2 2023	Q3 2023	Q4 2023	Q1 2024	Q2 2024	Q3 2024
Net cash generated from operating activities (M\$)	\$47.0	\$89.6	\$44.2	\$120.0	\$79.8	\$97.4	\$149.5
Free cash flow (M\$)	(\$54.0)	(\$37.4)	(\$69.7)	(\$24.3)	(\$49.1)	(\$62.3)	(\$0.7)
Media Luna Project capex (M\$)	\$66.4	\$77.2	\$98.7	\$124.0	\$126.4	\$108.2	\$113.9
Free cash flow prior to Media Luna Project (M\$)	\$12.4	\$39.8	\$29.0	\$99.7	\$77.3	\$45.9	\$113.2
Gold sold (koz)	118.5	105.7	81.8	138.8	111.6	113.5	122.1
Total cash costs (\$/oz)	\$709	\$848	\$1,086	\$885	\$918	\$1,014	\$926
All-in sustaining costs (\$/oz)	\$1,079	\$1,308	\$1,450	\$1,073	\$1,202	\$1,239	\$1,101
Average realized gold price (\$/oz)	\$1,899	\$1,960	\$1,944	\$1,995	\$2,023	\$2,193	\$2,313

Task 2: Image Extraction & Classification

Approach:

- Extract all images (embedded, rendered tables, photos).
- Classify as map, table, or picture using CLIP.

Tools/Models:

PyMuPDF, Pillow, OpenCV, CLIP (HuggingFace), torch.

Challenges:

Challenge: Mixed content, small/low-quality images.

Result: Labeled images, CSV summary, ZIP export.

Results:

In terms of accuracy, the algorithm correctly classified 10 map images and 30 images of tables and statistical information (graphs), demonstrating that this open-source approach delivers a high level of precision.

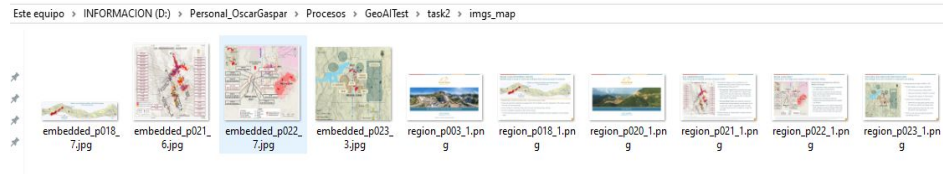


Improvement:

Improve table detection, add more classes.

```

♦ Classification results:
label
map      10
picture  50
table    30
dtype: int64
♦ Created imgs_raw.zip
PS D:\Personal_OscarGaspar\Procesos\GeoAITest\task2>
  
```



Task 3: Geospatial & Technical Info Extraction

Approach:

- Detect/extract coordinates (lat/lon, UTM, DMS) using OCR/regex.
- For geological plans: extract drill hole names, intervals, references.

Tools/Models:

Pytesseract, OpenCV, regex, pandas..

Challenges:

Challenge: Lack of explicit coordinates, varied formats.

Result: Cropped segments, marked visualizations, CSV.

Results:

Task 3 was originally designed to detect and extract explicit geographic coordinates (lat/long, UTM, DMS) from maps using OCR and advanced regular expressions. When applied to real geological plans, however, we found those coordinates are often absent; instead, it's common to encounter internal references (well numbers, drilling intervals, and other technical data typical of mineral exploration)..



Improvement:

Although the full process was not completed due to time constraints, this adaptation is proposed as a suggested improvement. The workflow would use technologies such as OCR (Tesseract + pytesseract) for digitizing visual content, regular expressions (re) to identify technical patterns (e.g., well names, drilling intervals), Pillow (PIL) for image loading and manipulation, and standard Python libraries for processing and structuring the results. It's worth noting that the original code already performs well when traditional coordinate formats like latitude/longitude are present.

filename	coordinates_found	segments_extracted	candidates
region_p003_1.png	0	0	[]
region_p018_1.png	0	0	[]
region_p020_1.png	0	0	[]
region_p021_1.png	0	0	[]
region_p022_1.png	0	0	[]

Summary, Impact & Future Improvements

GeoAITest represents a key advancement in automating the analysis of complex technical documents, integrating advanced AI and computer vision tools to drive digital transformation in technical data processing.

Impact:

- Scalable, automated technical analysis
- Integration of state-of-the-art AI and vision tools
- Boosting digital transformation in mining and geotechnics

Common Challenges:

- Variable PDF/image quality
- Inconsistent layouts and OCR errors
- Diversity of document formats

Future Improvements:

- More robust error handling
- Custom model fine-tuning
- User-friendly interface for review and correction.
- Integration with databases