6. Testing and Validation Artefacts

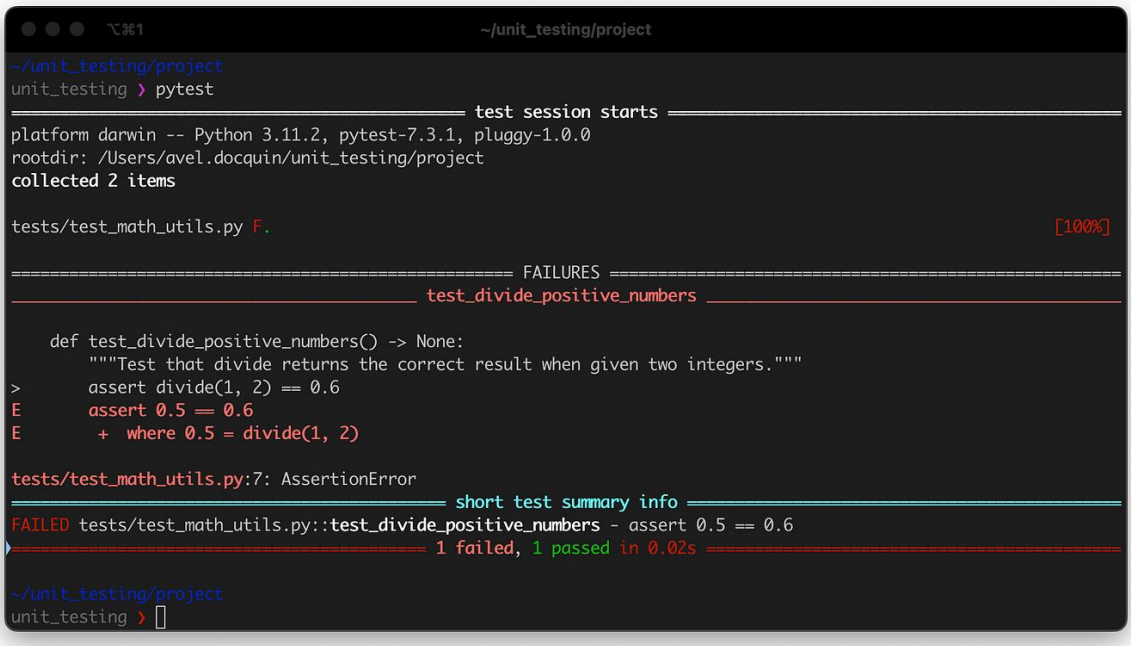
Comprehensive testing and validation are essential for demonstrating reliability, safety, and performance in a regulated medical AI system. All tests use only public or fully anonymised datasets (no real patient data), aligning with UK GDPR, MHRA GMLP, and NHS digital standards. Artefacts are placed in /tests/ and /docs/validation/ for assessor review, with commit history showing iterative improvements.

Key Components

- **Testing Framework**: pytest with coverage (pytest-cov) – unit tests for individual modules (anonymisation, ViT inference, LLM drafting), integration tests for pipeline end-to-end.
- **Datasets Used**:
    - Chest X-rays: MIMIC-CXR-JPG (PhysioNet), PadChest, NIH ChestX-ray14.
    - General/Other: TCIA collections (e.g., LIDC-IDRI for lung nodules).
    - Reports: IU X-ray or MIMIC-CXR paired reports for ground truth.
- **Metrics**:
    - ViT Analysis: Accuracy, Sensitivity, Specificity, AUC-ROC (per-class and macro-averaged).
    - Report Drafting: BLEU-4, ROUGE-L, METEOR (compared to reference reports).
- **Pilot Simulations**: Timed runs on sample cases (manual transcription vs AI-assisted editing) using public images.
- **Error Handling**: Structured logging (Python logging module) capturing invalid inputs, model failures, timeouts.

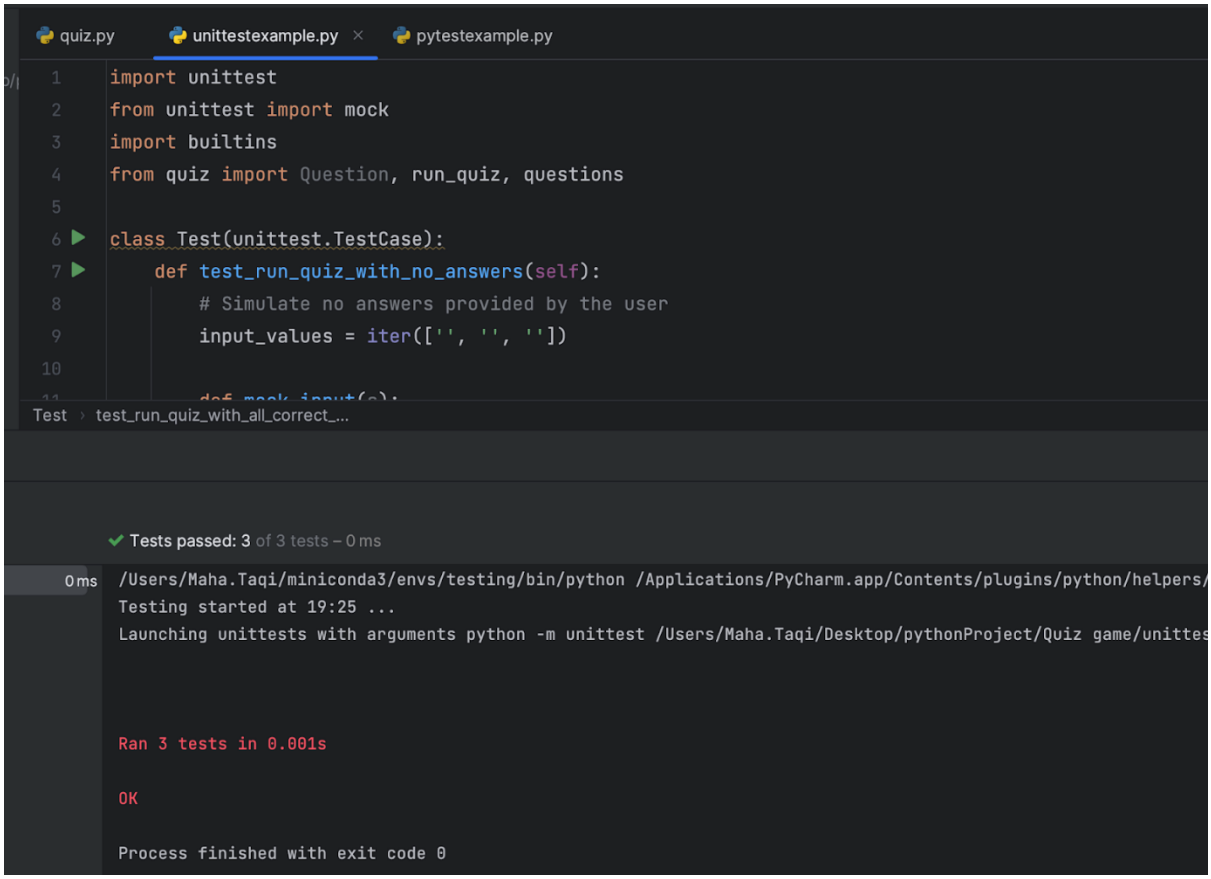
Industry-Level Example Diagrams and Visuals

Figure 43: Sample pytest unit and integration test results with coverage report (Standard output showing passed/failed tests.)



Python Unit Tests with pytest Guide | Avel Docquin | Medium

Suggested filename: fig43-pytest-results-example.pdf



Pytest vs. Unittest: Which Is Better? | The PyCharm Blog

Suggested filename: fig44-pytest-verbose-output.pdf

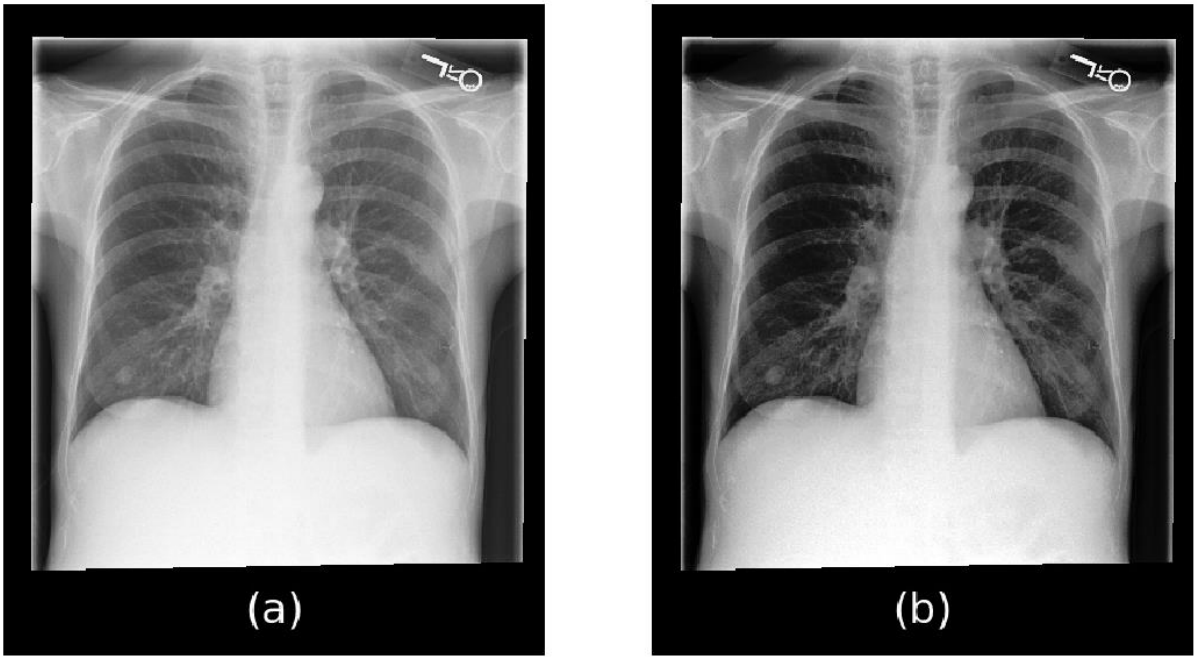
```
~/PycharmProjects/Personal/pytest-code-coverage-example  master
$ coverage report -m
Name                             Stmts   Miss  Cover   Missing
--------------------------------------------------------------
bank_app/__init__.py                 0      0   100%
bank_app/core.py                    41      0   100%
tests/__init__.py                    0      0   100%
tests/unit/__init__.py               0      0   100%
tests/unit/conftest.py              32      0   100%
tests/unit/test_bank_app.py         30      0   100%
--------------------------------------------------------------
TOTAL                              103      0   100%

~/PycharmProjects/Personal/pytest-code-coverage-example  master
$
```

How To Generate Beautiful & Comprehensive Pytest Code Coverage ...

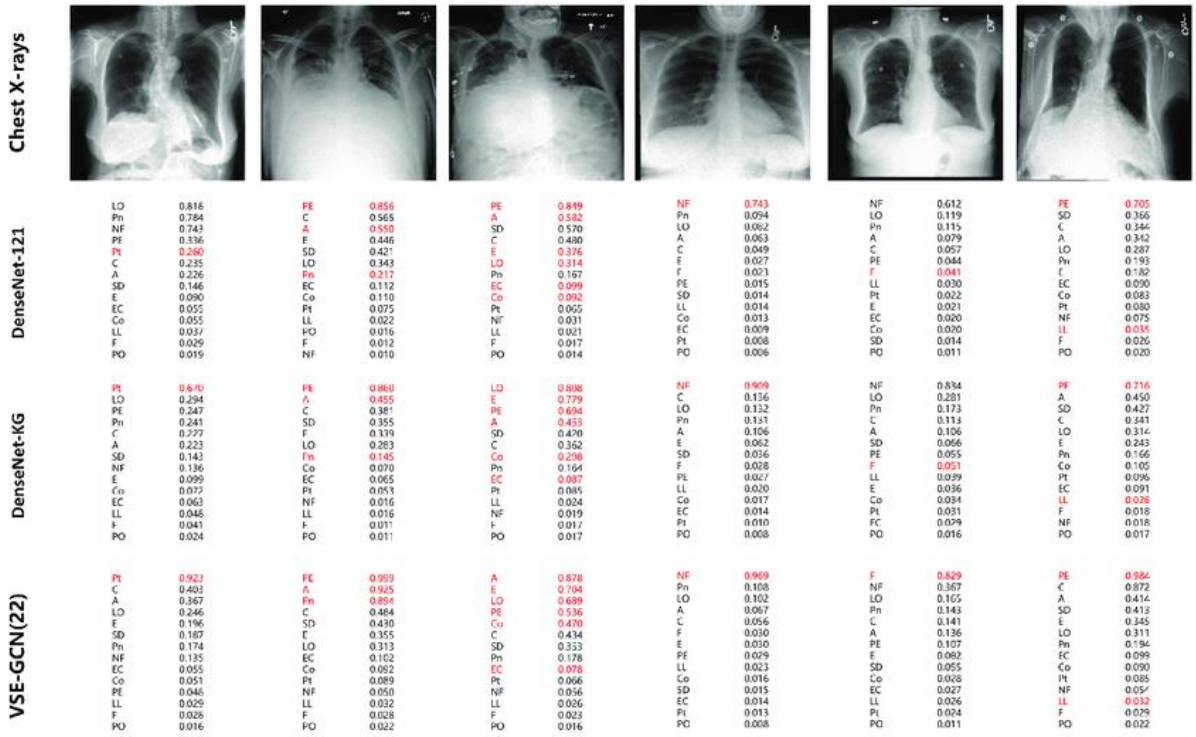
Suggested filename: fig45-pytest-coverage-report.pdf

Figure 46: Sample chest X-ray images from MIMIC-CXR dataset used for validation (Public anonymised examples for testing.)



MIMIC-CXR-JPG, a large publicly available database of labeled ...

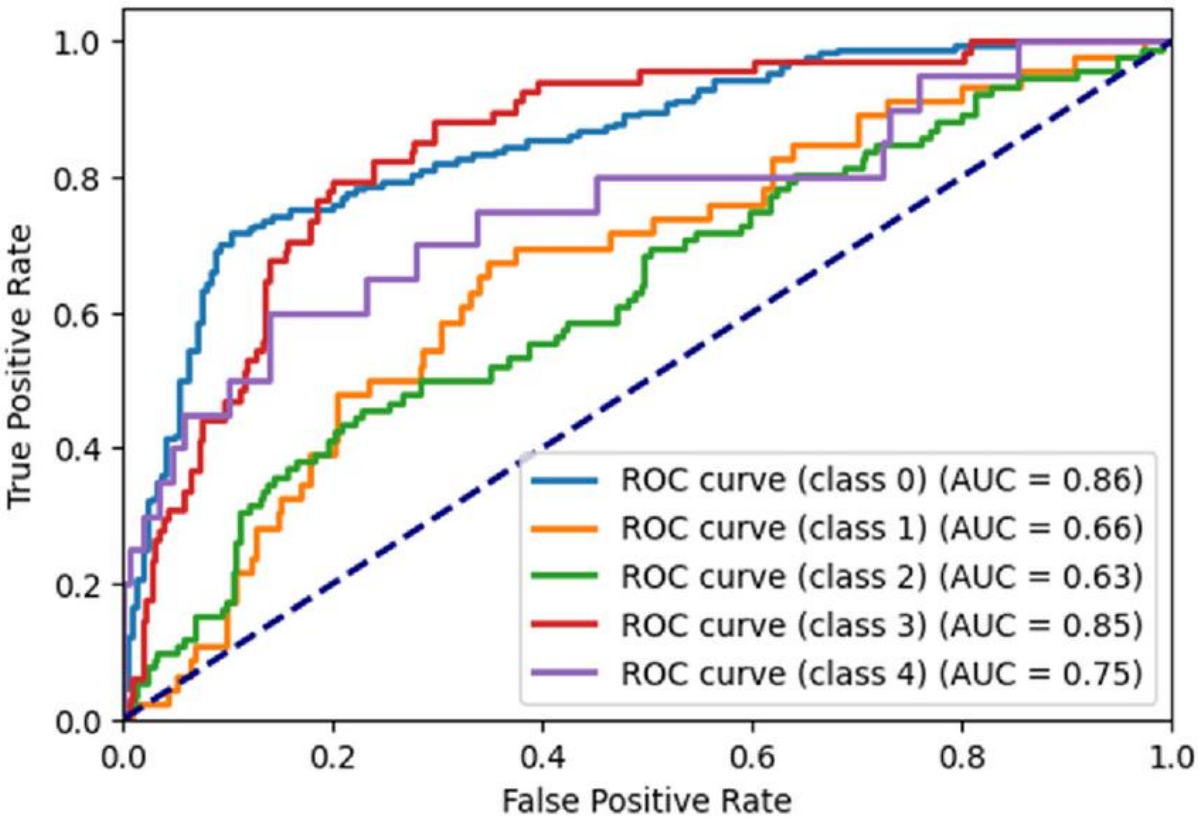
Suggested filename: fig46-mimic-cxr-samples.pdf



Diagnostic results of six samples from MIMIC-CXR test set. Three ...

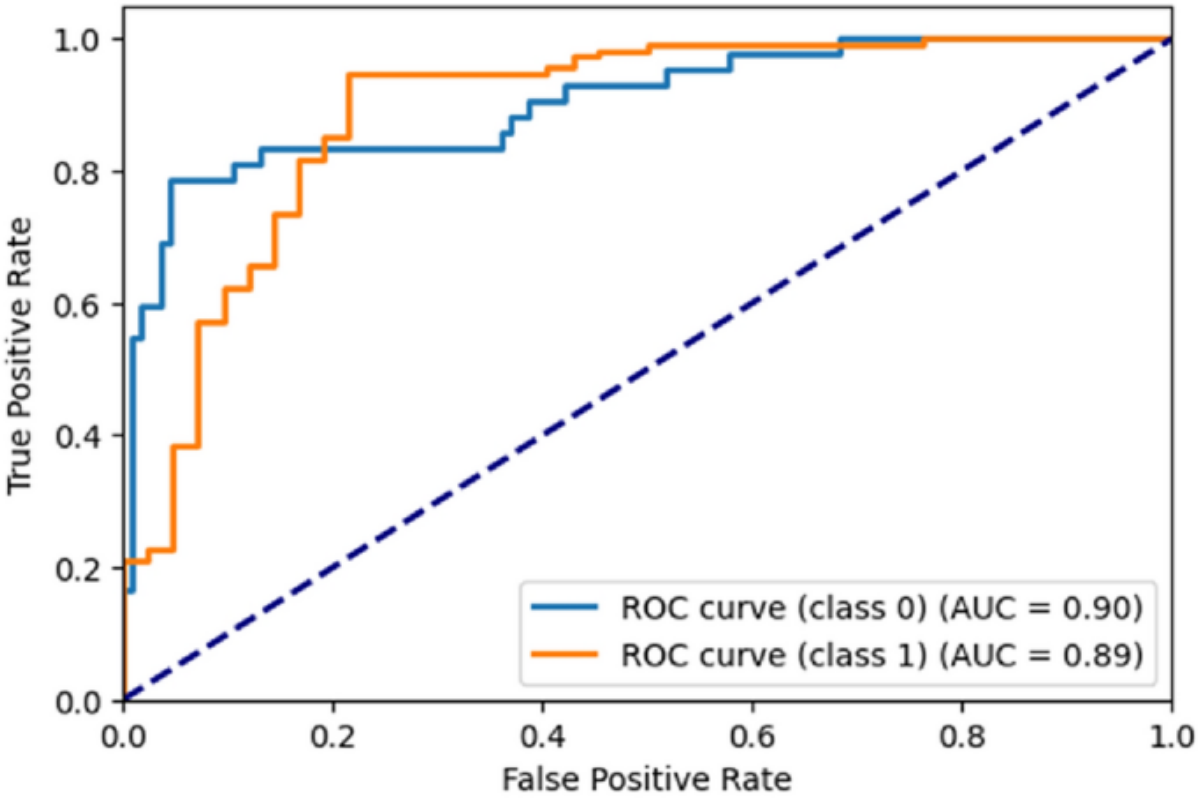
Suggested filename: fig47-mimic-cxr-diagnostic-samples.pdf

Figure 48: ROC curves with AUC values for Vision Transformer models in medical imaging (Performance metrics for classification tasks.)



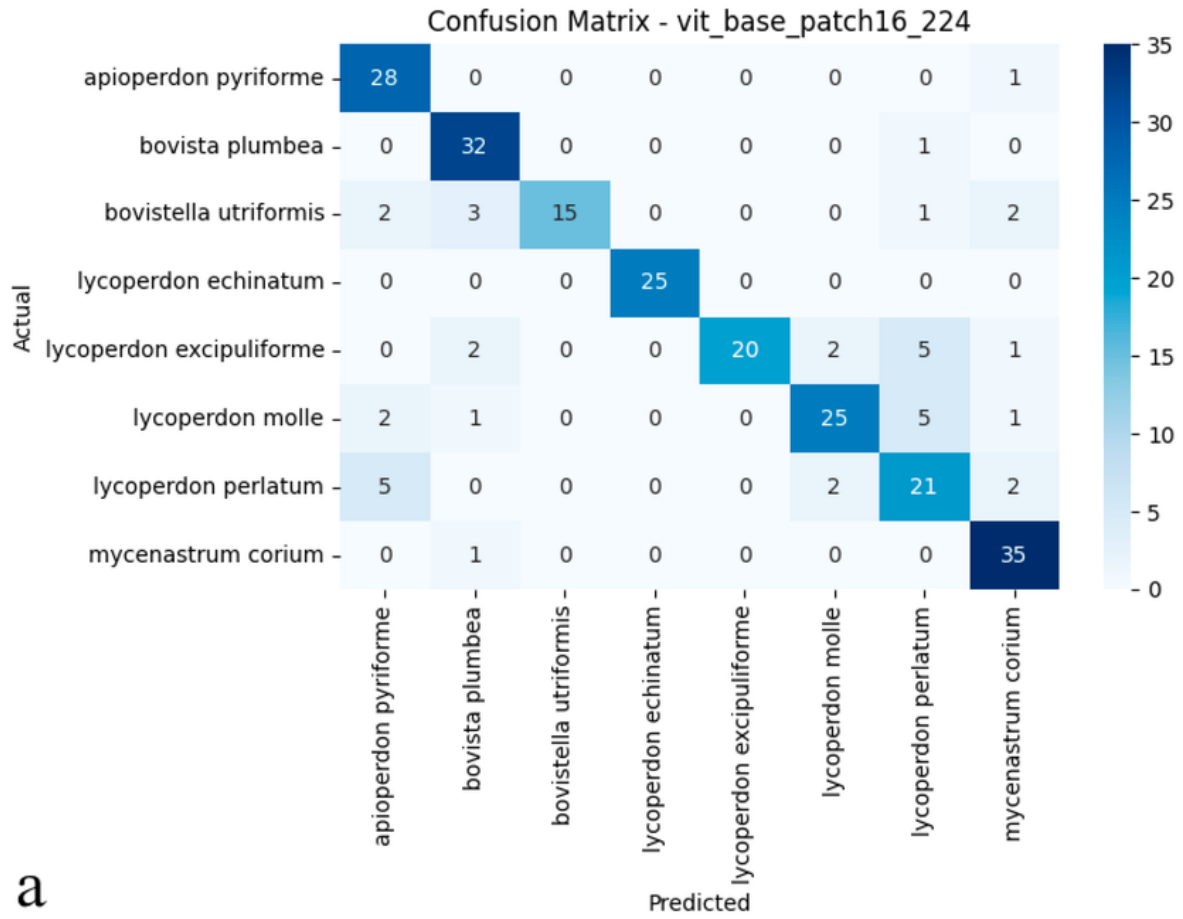
ROC curve generated by the proposed ViT model for RetinaMNIST ...

Suggested filename: fig48-vit-roc-curve.pdf



Implementing vision transformer for classifying 2D biomedical ...

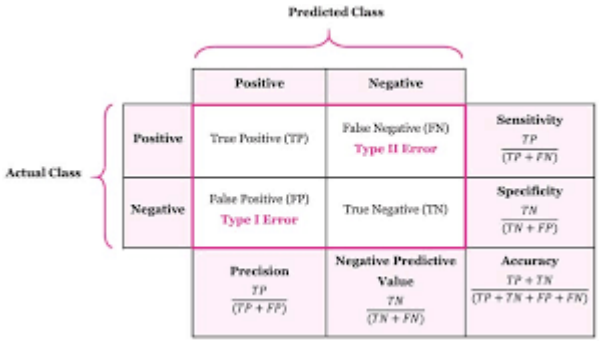
Suggested filename: fig49-vit-biomedical-roc.pdf



a) Confusion matrix and (b) ROC curves with AUC values of the ...

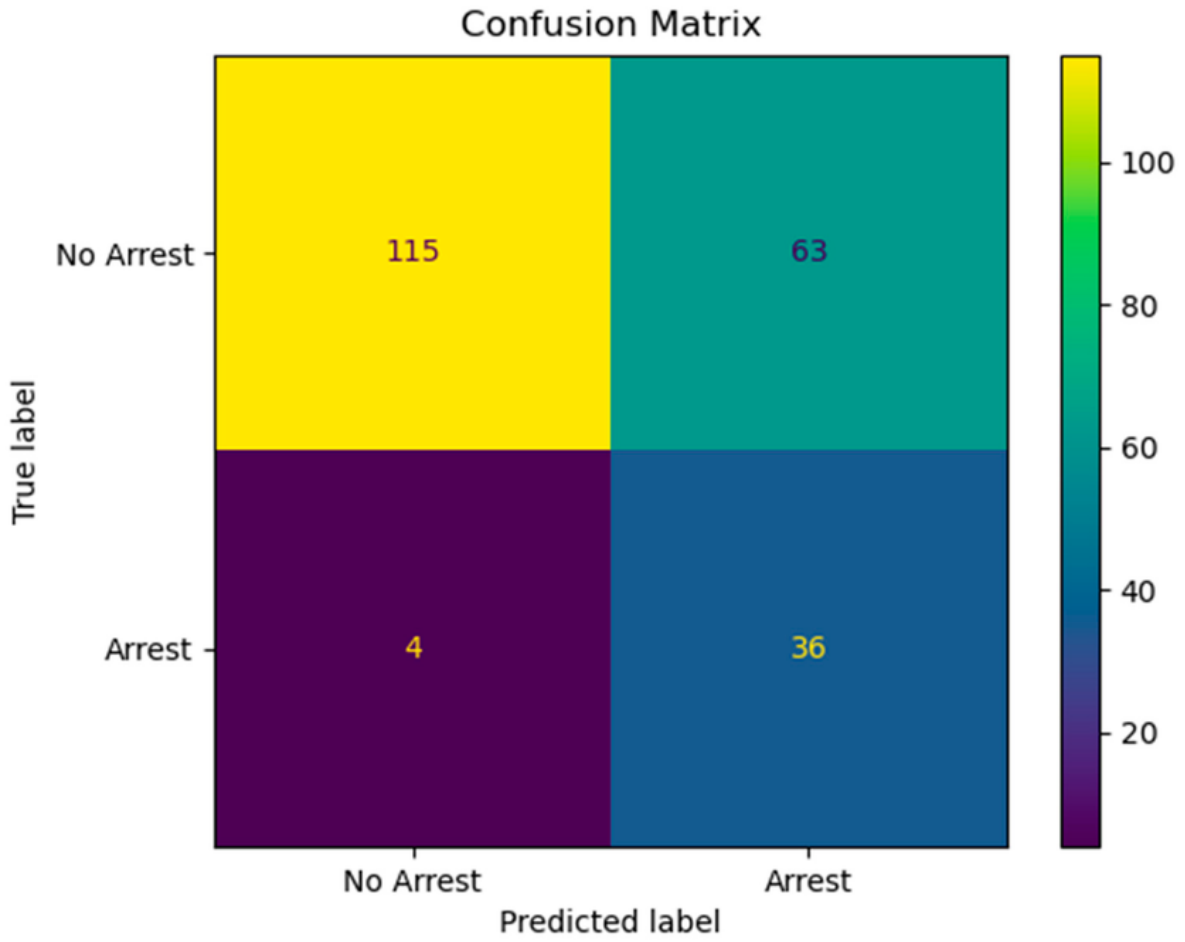
Suggested filename: fig50-vit-confusion-roc.pdf

Figure 51: Confusion matrix showing sensitivity and specificity in medical AI evaluation (Diagnostic performance breakdown.)



terminology - What is the best way to remember the difference ...

Suggested filename: fig51-confusion-matrix-sens-spec.pdf



Confusion matrix: the model had a sensitivity of 90% and a ...

Suggested filename: fig52-medical-confusion-matrix.pdf

Figure 53: BLEU and ROUGE score tables for LLM medical report generation evaluation (NLG metrics for draft quality.)

Metric	Value Mean (std)	Interpretation
BLEU Score	0.82 (0.19)	High n-gram overlap with the reference text, indicating strong **word- and phrase-level** similarity.
ROUGE-1 F-measure	0.94 (0.08)	Excellent unigram recall, showing that most individual words match the reference text (**word-level**).
ROUGE-2 F-measure	0.87 (0.17)	Strong bigram overlap, reflecting the model's ability to capture **phrase-level** coherence.
ROUGE-L F-measure	0.92 (0.11)	High similarity in the longest common subsequence, suggesting well-preserved **sentence-level** structure.
METEOR	0.92 (0.11)	Incorporates **synonymy and word order alignment**, indicating semantically accurate and fluent descriptions.
BERTScore_F1	0.99 (0.02)	Extremely high **semantic similarity** based on contextual embeddings, showing alignment in meaning beyond surface-level text.

Summary of text description evaluation metrics (BLEU, ROUGE ...

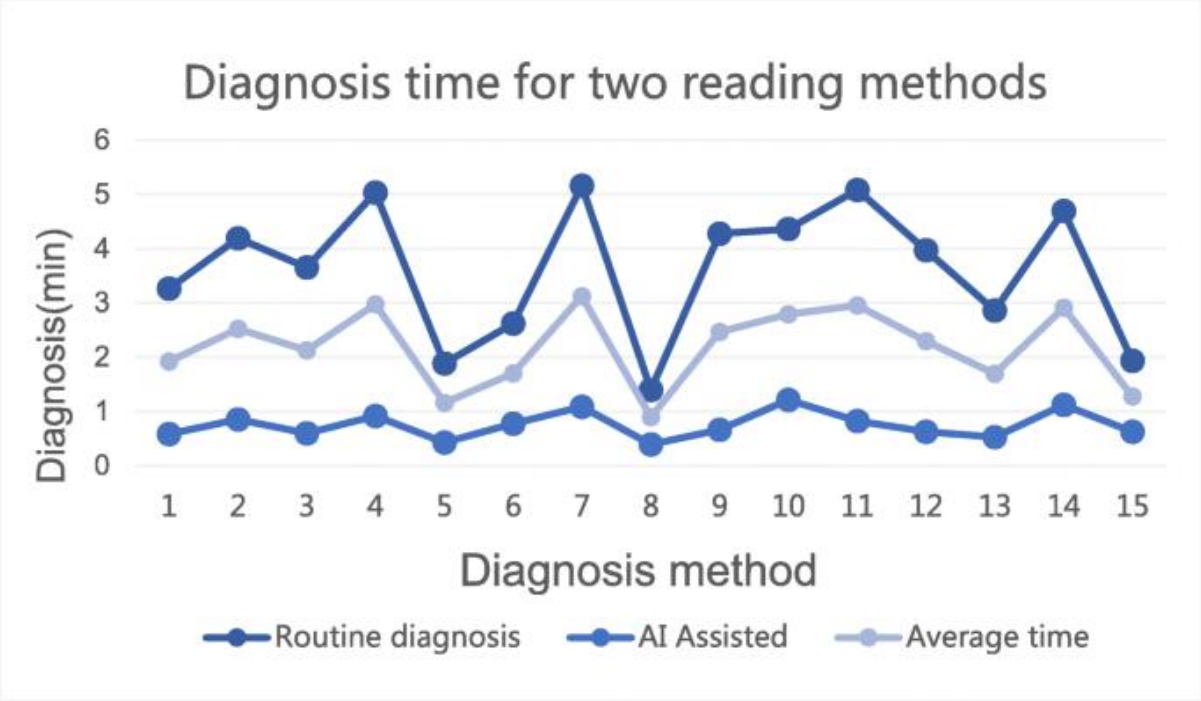
Suggested filename: fig53-bleu-rouge-table.pdf

Healthcare Categories & Subcategories	Dataset	Evaluation Metric	Model					
			Large		Medium		Small	
			GPT-4o	Gemini 1.5 Pro	Llama-3.3-70B-instruct	GPT-4o-mini	Qwen-2.5-7B-instruct	Phi-3.5-mini-instruct
Clinical Decision Support								
Supporting Diagnostic Decisions	MedCalc-Bench *	Exact Match	0.188	0.124	0.112	0.156	0.091	0.01
	CLEAR-AD	Exact Match	0.635	0.558	0.865	0.673	0.856	0.087
Planning Treatments	MTSamples	BertScore-F1	0.722	0.702	0.717	0.709	0.708	0.725
	Medec	MedecFlagAcc	0.536	0.526	0.534	0.531	0.496	0.508
Predicting Patient Risks and Outcomes	EHRShot *	Exact Match	0.467	0.559	0.704	0.902	0.806	0.231
Providing Clinical Knowledge Support	HeadQA *	Exact Match	0.91	0.84	0.848	0.832	0.734	0.679
	Medbullets	Exact Match	0.701	0.477	0.617	0.571	0.406	0.192
	MedAlign *	BertScore-F1	0.78	0.772	0.781	0.779	0.738	0.708
	ADHD-Behavior	Exact Match	0.837	0.761	0.641	0.806	-	-
	ADHD-MedEffects	Exact Match	0.945	0.731	0.925	0.756	-	-
Clinical Note Generation								
Documenting Patient Visits	DischargeMe *	BertScore-F1	0.754	0.762	0.751	0.758	0.751	0.755
	ACI-Bench	BertScore-F1	0.833	0.828	0.827	0.828	0.818	0.806
Recording Procedures	MTSamples Procedures	BertScore-F1	0.713	0.706	0.705	0.706	0.704	0.723
Documenting Diagnostic Reports	MIMIC-RRS *	BertScore-F1	0.77	0.826	0.775	0.787	0.815	0.81
Documenting Care Plans	NoteExtract	BertScore-F1	0.769	0.772	0.782	0.769	0.743	0.761
Patient Communication and Education								
Providing Patient Education Resources	MedicationQA	BertScore-F1	0.756	0.714	0.747	0.755	0.735	0.725
Delivering Personalized Care Instructions	PatientInstruct	BertScore-F1	0.751	0.744	0.758	0.754	0.751	0.755
Patient Provider Messaging	MedDialog *	BertScore-F1	0.755	0.748	0.76	0.754	0.744	0.668
	MedConfInfo *	Exact Match	0.772	0.74	0.789	0.738	-	-
Enhancing Understanding in Health Communication	MEDIQA	BertScore-F1	0.774	0.75	0.769	0.775	0.778	0.777
Facilitating Patient Engagement and Support	MentalHealth	BertScore-F1	0.821	0.797	0.802	0.819	0.787	0.39
Medical Research Assistance								
Conducting Literature Research	PubMedQA *	Exact Match	0.69	0.084	0.786	0.654	0.55	0.476
Analyzing Clinical Research Data	EHR-SQL	EHRSQLReAns	0.11	0.21	0.07	0.13	0	0.05
Recording Research Processes	BMT-Status	Exact Match	0.741	0.782	0.836	0.836	-	-
Ensuring Clinical Research Quality	RaceBias	Exact Match	0.874	0.82	0.599	0.76	0.587	0.144
Managing Research Enrollment	N2C2 CT	Exact Match	0.841	0.318	0.438	0.764	0.461	0.372
Administration and Workflow								
Scheduling Resources and Staff	HospiceReferral *	Exact Match	0.656	0.745	0.768	0.708	-	-
Overseeing Financial Activities	MIMIC-IV Billing Code *	Micro-F1	0.346	0.186	0.197	0.263	0.028	0.091
Organizing Workflow Processes	ClinicReferral	Exact Match	0.896	0.859	0.783	0.899	-	-
Care Coordination and Planning	CDI-QA *	Exact Match	0.589	0.466	0.563	0.579	-	-
	ENT-Referral *	Exact Match	0.611	0.576	0.363	0.594	-	-

Holistic Evaluation of Large Language Models for Medical ...

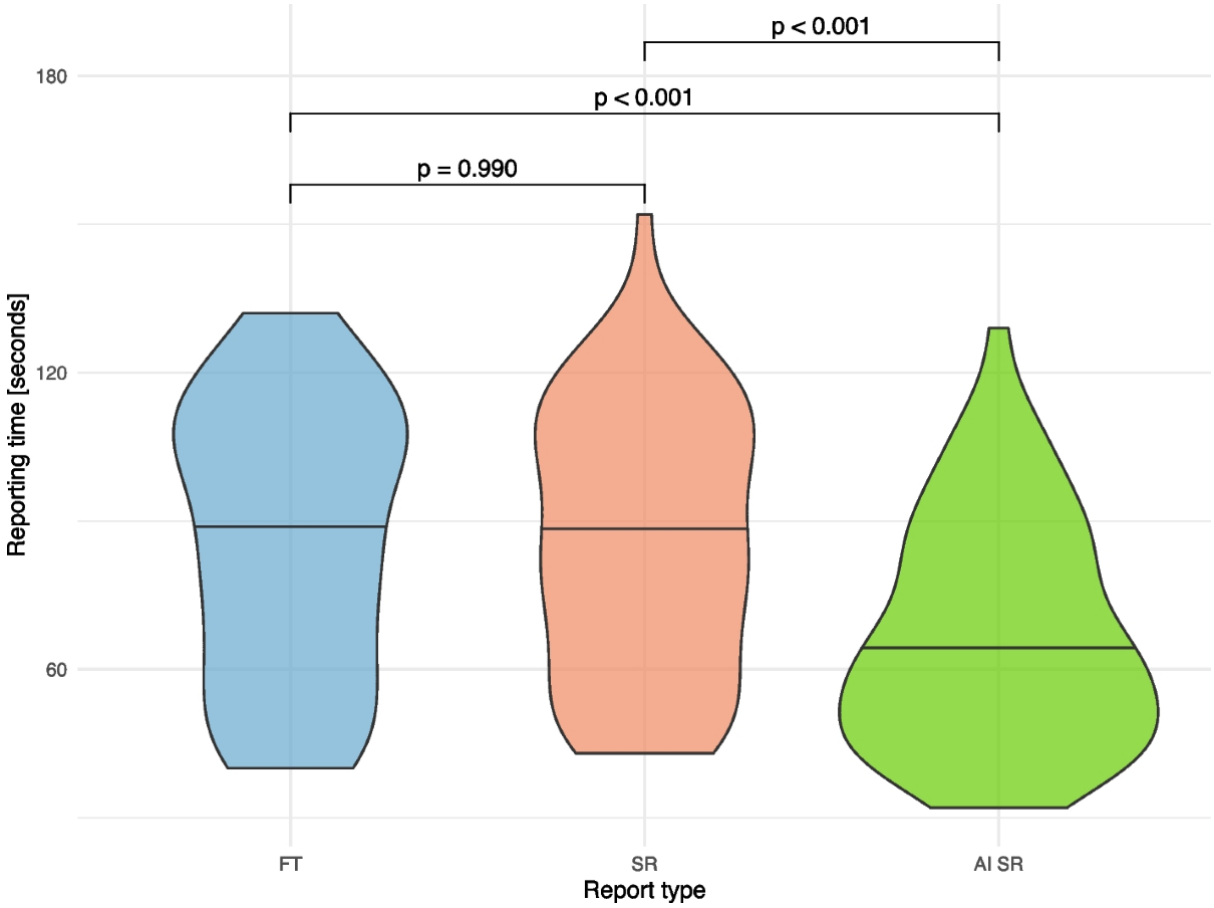
Suggested filename: fig54-llm-medical-metrics.pdf

Figure 55: Time comparison charts for manual vs AI-assisted radiology reporting (Simulated workflow efficiency gains.)

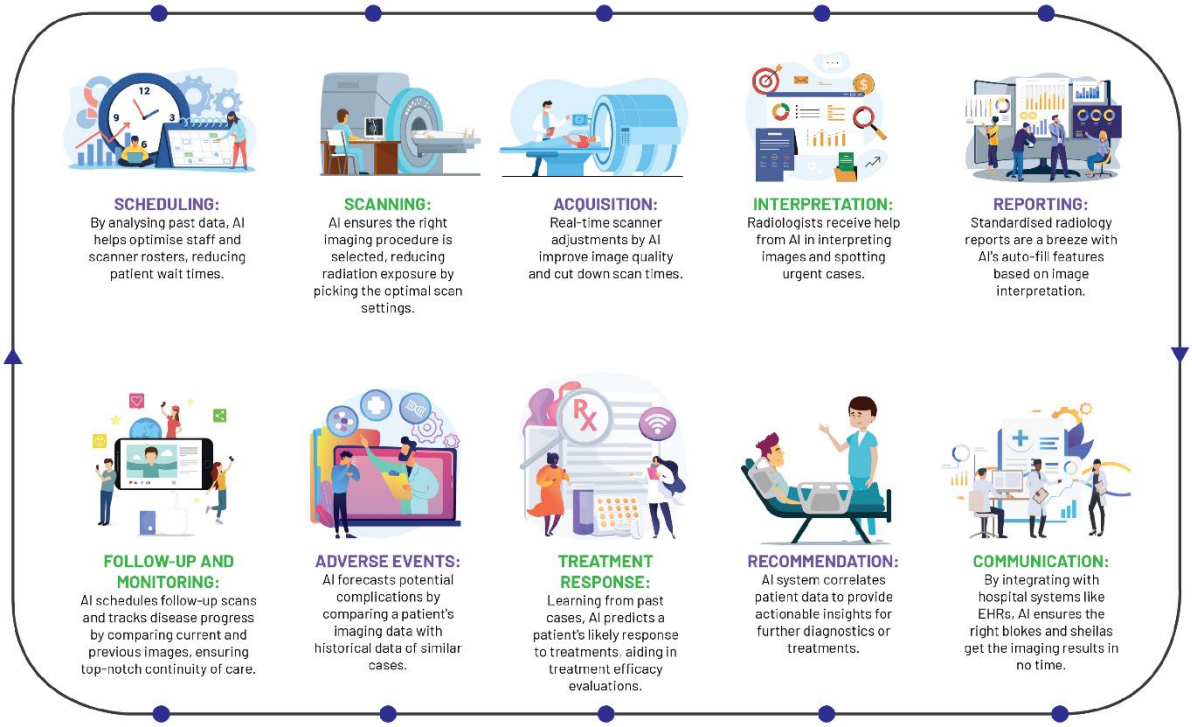


RETRACTED ARTICLE: The value of artificial intelligence and ...

Suggested filename: fig55-ai-reporting-time-comparison.pdf



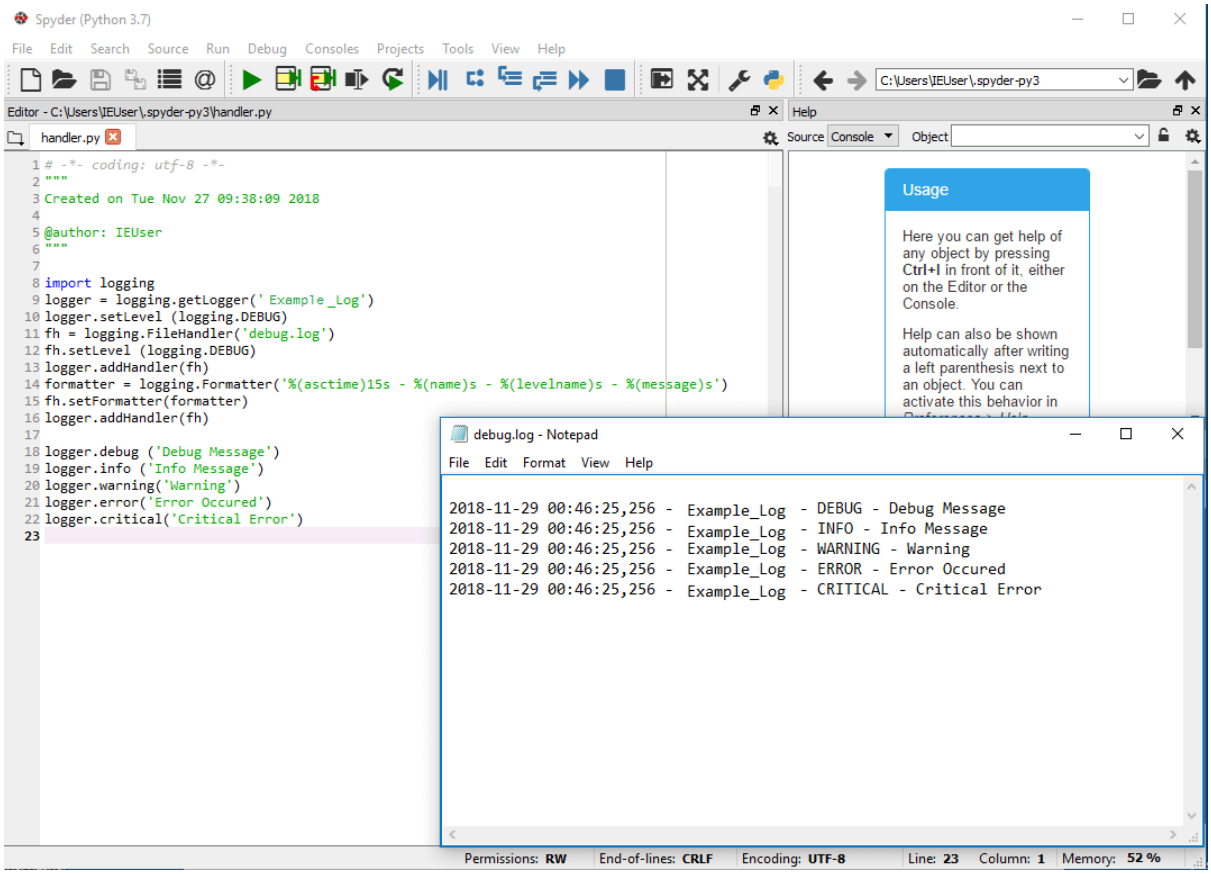
A novel reporting workflow for automated integration of artificial ...



Redefining Radiology: A Review of Artificial Intelligence ...

Suggested filename: fig57-ai-efficiency-bar-chart.pdf

Figure 58: Python logging and error handling examples (Structured logs for failures.)



The Python logging module: How logging to file works - IONOS

Suggested filename: fig58-python-error-log.pdf

Evidence Artefacts for Repository

- /tests/: test_anonymizer.py, test_vit_inference.py, test_report_generator.py, test_pipeline_integration.py.
- Results: test_report.html (pytest-html output), coverage reports.
- Validation docs: Tables of metrics (CSV/Markdown), timed simulation logs.
- Logs: error_log_sample.txt (anonymised entries).

These artefacts provide verifiable evidence of robust development, performance, and error resilience, supporting claims of workflow improvement in pilot simulations.