International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Scalable ML Framework for Insurance Document Indexing

Ramprakash Kalapala

Riverside, CA

Abstract: ***We present a scalable ML pipeline leveraging multimodal OCR, transformer - based architectures (e. g., LayoutLMv3, Donut), and human - in - the - loop active learning for fine - grained classification and structured information extraction across heterogeneous insurance document corpora. The platform incorporates end - to - end data ingestion, annotation, model versioning, drift detection, and continuous retraining to ensure robustness and operational scalability. This approach significantly enhances throughput, reduces manual annotation overhead, and supports compliance in large - scale insurance ecosystems.***

Keywords: Multimodal OCR; Transformer architectures; LayoutLMv3; Donut model; Human - in - the - loop learning; Active learning

1. Introduction

Insurance enterprises handle a diverse array of semi - structured and unstructured documents, posing significant challenges for automation using rule - based systems. The variability in layout, language, and modality across forms such as claims, policies, and legal documents necessitates robust machine learning (ML) approaches.

This paper introduces a scalable ML framework that leverages layout - aware transformer models (e. g., LayoutLMv3, Donut) combined with OCR and active learning to automate classification and indexing of 15–20 insurance document types. The system supports the entire ML lifecycle including ingestion, annotation, training, deployment, drift detection, and retraining, enabling reliable document processing at scale.

CCS Concepts
Computing methodologies → Machine learning → Machine learning applications.
Computing methodologies → Information extraction.
Applied computing → Document management and text processing.
Applied computing → Insurance.

2. Document Types and Use Case

Insurance operations involve a wide spectrum of documents that vary in structure, semantics, and source. Accurate classification and indexing of these documents are critical for downstream tasks such as claim adjudication, risk assessment, fraud detection, regulatory compliance, and customer service automation.

This paper focuses on 15–20 core document types commonly encountered across insurance workflows, including:
- Claims - related: Claim forms, explanation of benefits (EOB), incident reports, medical bills
- Policy - related: Application forms, declarations, endorsements, cancellation notices
- Identity and Legal: Government - issued IDs, POA (proof of address), court notices, affidavits
- Medical and Financial: Diagnostic reports, prescriptions, bank statements, invoices

Each document type presents unique challenges due to variable formats (structured, semi - structured, unstructured), visual noise (scans, handwriting), and differing information density. For example, a claims form requires key - value extraction of policyholder and incident details, while a medical bill involves tabular data parsing and cost validation.

The ML pipeline supports multiple use cases:
Automated classification and routing to the appropriate processing queues
Metadata extraction (e. g., policy number, date of loss, claimant details)
Validation and compliance checks against business rules and regulatory constraints
Workflow prioritization based on document content and urgency indicators
Human - in - the - loop review for edge cases and ambiguous classifications
By automating document understanding at scale, the proposed system enhances operational throughput, reduces manual handling costs, and increases accuracy across high - volume insurance environments.

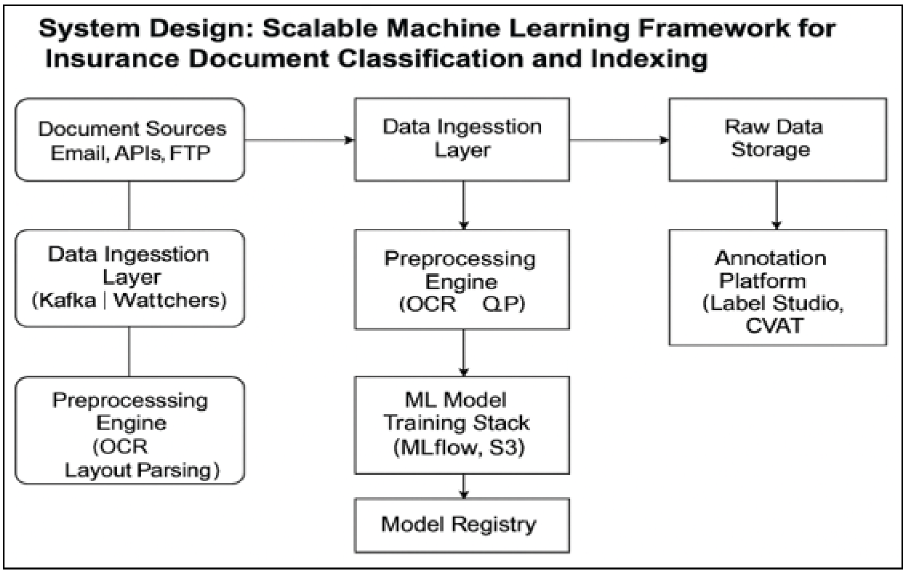
3. ML Platform Lifecycle

A robust machine learning (ML) platform for document indexing in insurance must support the full lifecycle of model development, deployment, and continuous improvement. This ensures scalability, traceability, and adaptability to evolving document distributions and business needs.

The lifecycle comprises the following key stages:

Stage	Description	Key Components
Data Ingestion	Collects and streams raw documents from sources (e. g., email, SFTP, APIs) into the processing pipeline.	Kafka, AWS S3, file watchers
Preprocessing & OCR	Applies layout - preserving OCR (e. g., Tesseract, Amazon Textract) and document normalization.	Text extraction, layout parsing
Annotation & Labeling	Human - in - the - loop labeling using tools like CVAT or Label Studio for supervised learning datasets.	Active learning, feedback loop
Model Training	Fine - tunes multimodal transformer models (e. g., LayoutLMv3, Donut) for classification and key - value extraction.	PyTorch, Hugging Face, distributed training
Evaluation & Validation	Runs automated and manual evaluations across precision, recall, and document - type confusion matrices.	CI/CD for ML, MLflow
Model Deployment	Containerizes and deploys models as REST APIs or gRPC services using orchestrators like Kubernetes.	Docker, KServe, TorchServe
Inference Pipeline	Routes incoming documents through classification, extraction, and post - processing workflows.	Real - time and batch processing
Monitoring & Drift Detection	Monitors input features, prediction distributions, and system latency for drift and anomalies.	Prometheus, Evidently AI
Continuous Retraining	Triggers periodic or event - based retraining with updated data and feedback for model refresh.	Scheduled pipelines (Airflow, Kubeflow)

4. High - Level Architecture



5. Benefits of ML - Based Insurance Document Indexing

5.1 Accelerated Document Processing

Automation of OCR and classification speeds up handling large volumes of insurance documents.

```
# OCR using Tesseract to extract text quickly from scanned documents
import pytesseract
from PIL import Image
def extract_text (image_path):
 img = Image. open (image_path)
 return pytesseract. image_to_string (img)
text = extract_text ('insurance_claim_scan. png')
print ("Extracted Text: ", text)
```

5.2 Enhanced Accuracy and Consistency

Using ML models for document classification ensures consistent categorization and reduces manual errors.

```
from transformers import pipeline
# Load pretrained classifier (fine - tuned on insurance docs for best results)
classifier = pipeline ('text - classification', model='distilbert - base - uncased - finetuned - sst - 2 - english')
document_text = "Insurance claim for vehicle accident filed on 2025 - 05 - 20. "
result = classifier (document_text)
print (f"Document type: {result [0] ['label']} with confidence {result [0] ['score']: .2f}")
```

5.3 Operational Cost Reduction

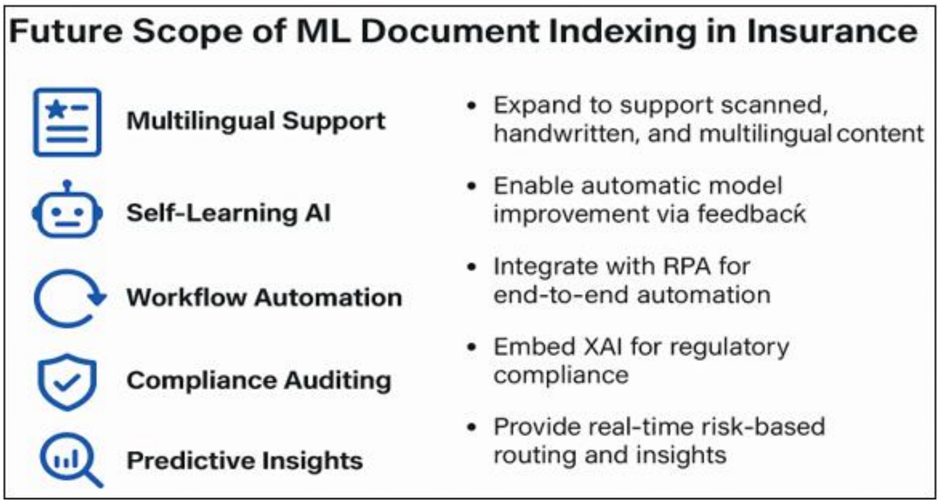
Indexing extracted metadata enables fast retrieval and reduces manual search efforts.

```
from elasticsearch import Elasticsearch
es = Elasticsearch ("http: //localhost: 9200")
doc = {
 "claim_id": "12345",
 "policy_number": "POL98765",
 "claimant": "John Doe",
```

```
 "date_filed": "2025 - 05 - 20",
 "document_type": "insurance_claim",
 "full_text": document_text
}
response = es. index (index="insurance_documents", id=doc ['claim_id'], document=doc)
print ("Indexed Document: ", response ['result'])
```

6. Future Scope of ML - Based Insurance Document Indexing

As technology and business needs evolve, ML - driven document indexing platforms in insurance will expand into the following areas:



7. Conclusion

Implementing an ML - based document indexing platform empowers insurers to automate processing, reduce errors, and accelerate claim handling. It enhances efficiency, improves customer experience, and lays the groundwork for scalable, intelligent operations. As the technology evolves, it will become a strategic asset in driving digital transformation and regulatory readiness across the insurance sector.

References

[1] H. Zhang and R. Binns, "Insights from Insurance for Fair Machine Learning, " in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp.1556–1567, 2024. [Online]. Available: https: //dl. acm. org/doi/10.1145/3630106.3658914

[2] M. M. Abid and A. Khan, "Digitization of Health Insurance Documents for the Cashless Claim Process, " in *Procedia Computer Science*, vol.226, pp.1123–1131, 2024. [Online]. Available: https: //dl. acm. org/doi/10.1016/j. procs.2024.04.125

[3] Y. Li, C. Lin, and F. Zhang, "A Machine Learning Based Anti - Fraud Approach for Life Insurance, " in *Proceedings of the 2023 International Conference on Artificial Intelligence, Automation and Industrial Engineering*, ACM, pp.43–48, 2023. [Online]. Available: https: //dl. acm. org/doi/abs/10.1145/3708360.3708389

[4] K. Anand and A. Mishra, "Intelligent Document Analysis in the Insurance Market, " in *Proceedings of the 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, ACM, 2021. [Online]. Available: https: //dl. acm. org/doi/abs/10.1145/3478905.3479000