Research / Project Portfolio

Independent Research & Industry Projects

Olivia Chatterjee | February 2026

Undergraduate Thesis

Benchmarking OCR Engines and VLMs for Automated Evaluation of Handwritten Scanned Answer Scripts

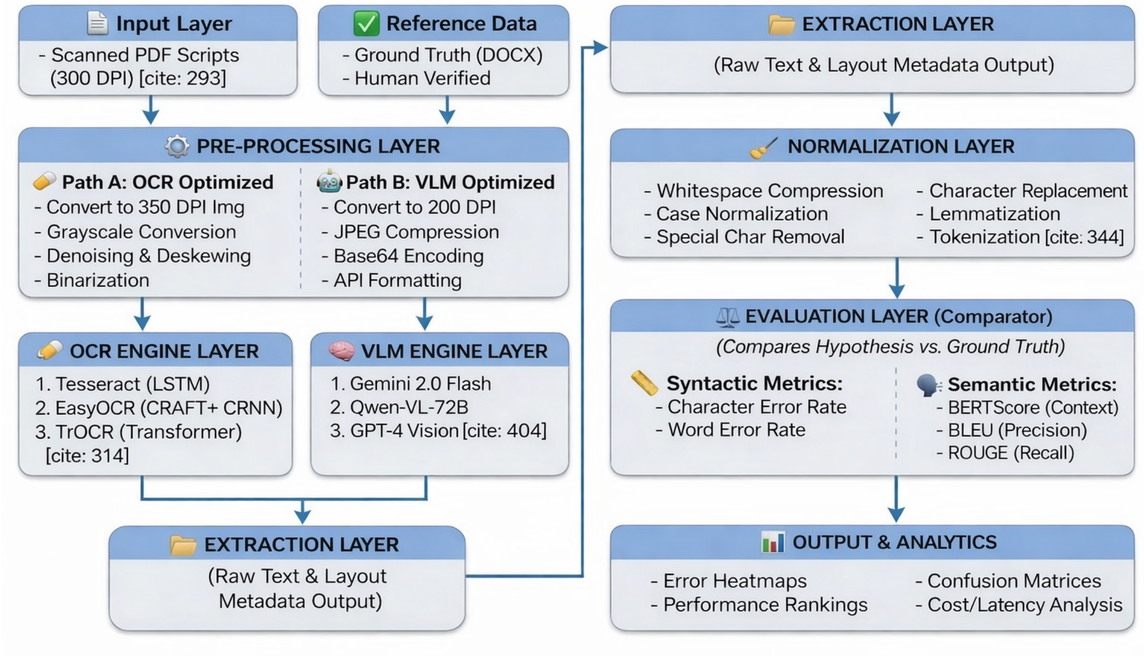
Supervisor; Dr. Uttam Kumar Sarkar, Associate Professor, CSE(AIML) Dept., Techno Main Salt Lake, Kolkata

VIEW REPORT

Built a reproducible benchmarking pipeline that compares traditional OCR, transformer OCR, and vision-language models on real multi-page handwritten university answer scripts using dual-path preprocessing and unified transcription + semantic metrics to assess readiness for automated evaluation.

Research contributions

- Created a curated, anonymized dataset of full handwritten exam scripts with strict human-verified ground truths across modalities (paragraphs, equations, tables, and labelled figures) to enable controlled evaluation.
- Designed a uniform evaluation framework that jointly measures literal transcription fidelity and semantic fidelity using CER/WER alongside BLEU, ROUGE-L, and BERTScore, with per-page and document-level aggregation.
- Implemented and validated a modular benchmarking architecture with model-specific preprocessing pathways (OCR-optimized vs VLM-optimized) and systematic error-mode analysis to compare engines fairly under the same protocol.



Outcome

Traditional OCR engines (Tesseract/EasyOCR) fail on real handwritten answer scripts with extremely high error rates, TrOCR improves but remains unreliable, while VLMs perform substantially better overall, with OpenAI GPT-4 Vision showing the strongest accuracy and semantic alignment across most samples.

Collaborative Research

Core Data Science Team Size 5

Predictive Modeling for Country Vulnerability Assessment

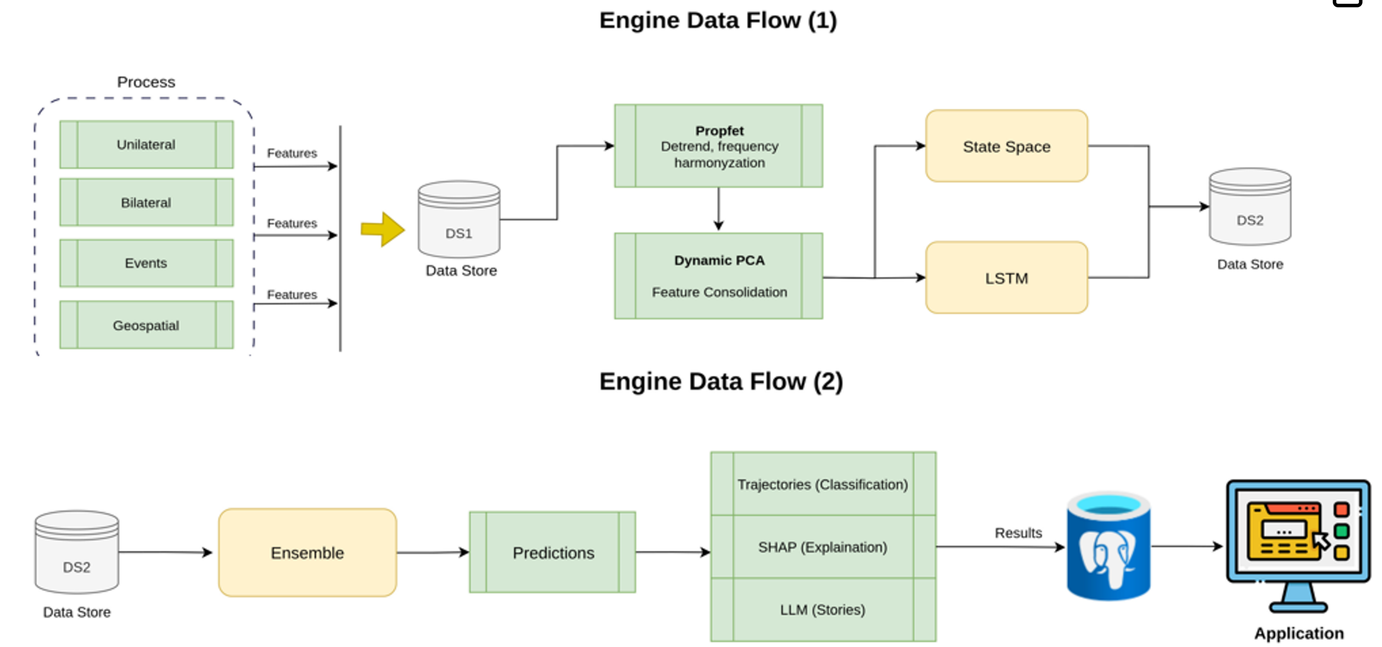
Supervisor; Prof. Sanjeev Khagram, Distinguished Visiting Fellow, Hoover Institution, Stanford University

LIVE ACCESS

Building multi-horizon predictive models to estimate country-level risks across political, economic, environmental, and social domains using high-dimensional data sources.

Research contributions

- Design and implement predictive models using machine learning and deep learning (e.g., LSTM) for multi-horizon vulnerability estimation.
- Collaborate with interdisciplinary teams to integrate diverse country-specific data streams into robust modeling pipelines.
- Validate and refine models through historical analysis to improve accuracy, robustness, and actionable insights.



Outcome

Although the project is ongoing, preliminary models have shown strong potential in enhancing early warning capabilities and laying the groundwork for data-driven resilience planning and integrated geopolitical risk analysis.

Financial Time Series Analysis with Data-Science Techniques

Prof. Indranil SenGupta, Professor of Mathematics and Statistics, Hunter College, City University of New York

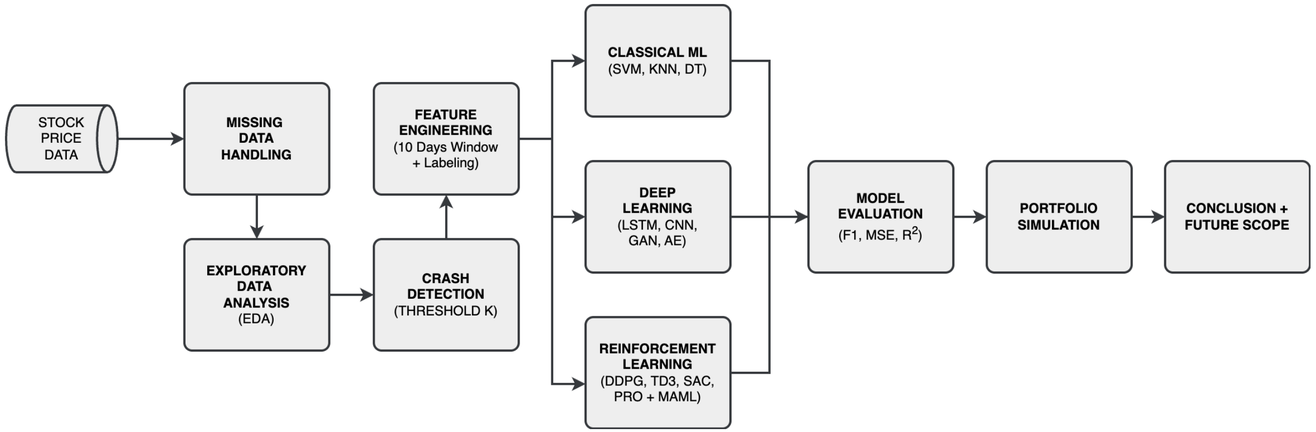
DOWNLOAD PAPER

Analyzed a decade-long stock price dataset to develop machine learning models capable of identifying early indicators of market crashes through pattern detection and classification techniques.

Research contributions

- Conducted exploratory data analysis to identify key trends and patterns.
- Handled missing data, ensuring the integrity and accuracy of the analysis and created a feature matrix using consecutive Close prices to build a predictive model for identifying potential market crashes.
- Implemented and compared various supervised learning algorithms to classify market conditions as crash-prone or stable, optimizing the model by experimenting with different parameters.
- Evaluated the performance of classification algorithms using metrics such as confusion matrices and classification reports, iterating on model parameters to improve prediction accuracy.

O. Chatterjee, A. Kole, I. SenGupta and A. Majumdar. "Forecasting Stock Market Crashes Using Deep Learning" Preprints. DOI:10.20944/preprints202510.1781.v1, Oct. 2025. Accepted at ICDMAI 2026 (January 2026, peer-reviewed, Springer–indexed series) - www.icdmai.org



Outcome

Successfully demonstrated the feasibility of crash prediction using time-series transformations and supervised learning. This research laid the foundation for a paper (under review) and contributed valuable insights into early warning systems for financial market anomalies.

Classification of Tabular Data Using CNN for Cancer Detection

Prof. Treena Basu, Associate Professor, Department of Mathematics, Occidental College, California, USA

VIEW REPORT

This research explores transforming structured tabular data into image-like representations to enable CNNs to classify complex biological datasets for cancer detection.

Research contributions

- Conducted an in-depth analysis of existing research and methodologies in the classification of tabular data using CNNs, with a focus on identifying challenges in achieving the required accuracy, particularly in cancer detection.
- Developed and implemented a novel method to convert tabular data into image-like structures for CNN processing, addressed key limitations observed in prior studies.

Conclusion

This internship successfully demonstrated the application of Tabular Convolution (TAC) for transforming non-image gene expression data into image representations suitable for CNNs. The performance of CNN-based TAC was benchmarked against traditional machine learning pipelines using PCA followed by classification with algorithms such as KNN, Logistic Regression, Decision Trees, SVM, Naive Bayes, and Random Forest.

Key findings include:

- *PCA Effectiveness: Increasing PCA components improved model accuracy, with performance stabilizing around 40–50 components.*
- *Best Traditional Models: KNN and Logistic Regression achieved the highest accuracy (99.17%) on PCA (50) datasets with minimal computational cost.*
- *CNN TAC Performance: The CNN model trained on TAC-transformed data achieved competitive accuracy, validating the feasibility of applying deep learning to tabular gene expression datasets.*

Outcome

Achieved around 75% classification accuracy on cancer datasets using transformed tabular genomic data with CNNs—demonstrating the feasibility of applying deep learning to structured biomedical data. This work opens new directions in spatial reasoning for genomics, offers a generalizable pipeline for the disease detection, and lays foundational insight for using CNNs beyond traditional image-based domains.

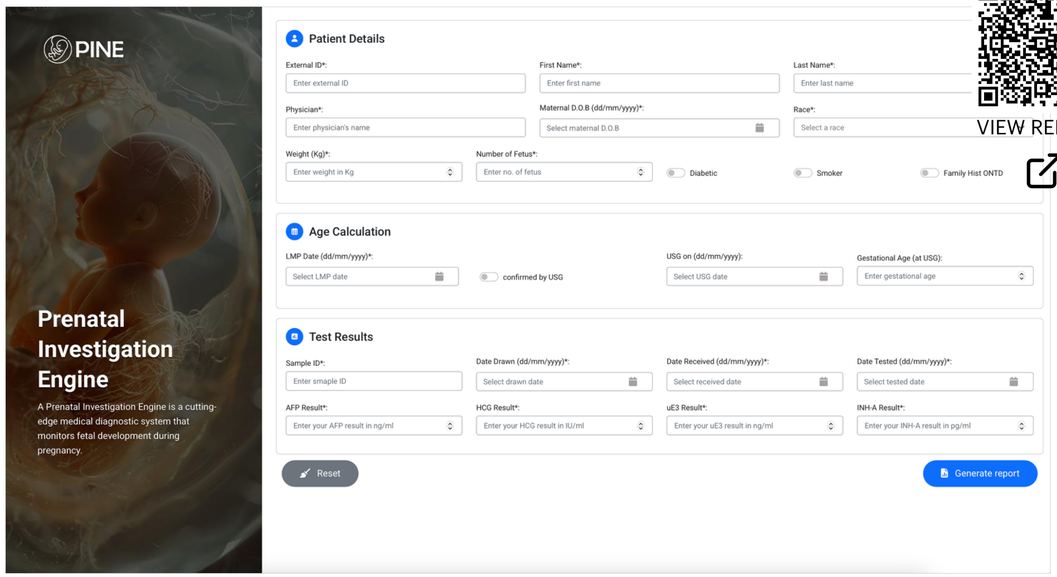
Industry Project - Live in Production

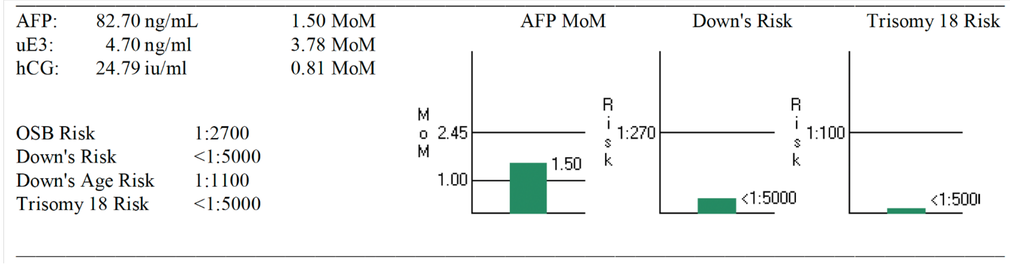
Engine Development Team Size 3

Pre-Natal Risk Assessment Engine

Tosoh India Pvt. Ltd., subsidiary of Tosoh Corporation, Japan

VIEW REPORT





Duties Involved

- Data De-identification and Privatization using Differential Privacy.
- Training, implement and fine-tune artificial neural network models, regression models, and ensemble methods like XGBoost and Random Forest for accurate risk prediction of congenital disorders.
- Development of the Engine API for the middleware.

Outcome

Achieved over 99% model accuracy even after training on privatized data, validating the feasibility of privacy-preserving machine learning in healthcare. The successful deployment enabled secure and cost-effective prenatal risk screening, reducing dependence on expensive confirmatory diagnostics while safeguarding sensitive patient data. The project was deployed in production and operational now.

Industry Project - Live in Production

Algorithm Development Team Size 3

Report Private - available on request.

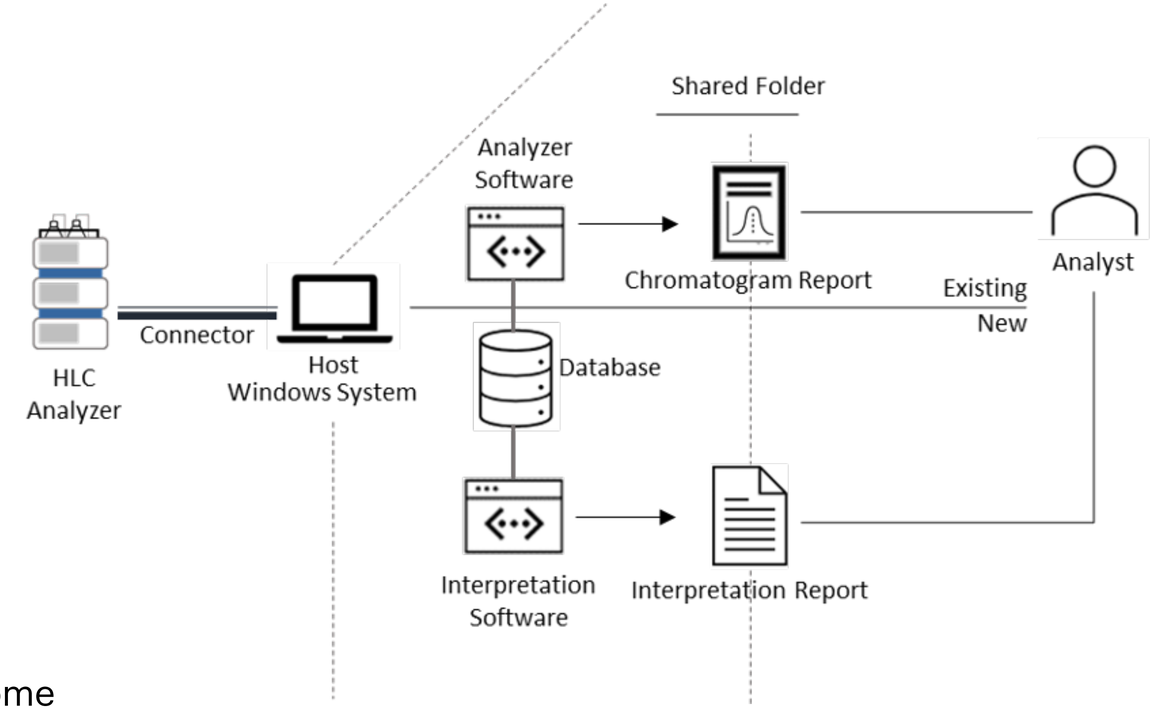
Chromatography Data Analysis

Dr. Arnab Majumdar, Chief Data Scientist, Entiovi Technologies Private Limited, India.

Built a logic-based diagnostic platform to automate interpretation of liquid chromatogram data for HbA1c analysis, enabling rapid classification of haemoglobin variants and reducing manual diagnostic effort.

Duties Involved

- Engineered classification logic using **Prolog** to detect hemoglobinopathies based on retention times and peak structures, automating the detection of known and unknown peaks.
- Developed and applied encryption techniques using Python to secure sensitive medical data, ensuring the privacy and protection of patient information throughout the diagnostic process.



Outcome

The platform has been successfully deployed in HPLC machines across hospitals and diagnostic labs in India and is now running in live production. It has significantly reduced diagnostic turnaround times and established a scalable, privacy-preserving solution for HPLC-based clinical diagnostics.

Thank You

Olivia Chatterjee

olivia.chatterjee@ieee.org | +91-85850-48450
olivia-chatterjee.github.io