

Shreyans Jain

shreyanssjain.com • LinkedIn.com/in/shreyans-s-jain • GitHub.com/shreyanse081
shreyanssjain@outlook.com • +91 87670 87603

Sr. Data Scientist with 7+ years architecting production ML systems for chemical process and mining operations — from Azure ML MLOps pipelines and real-time scoring services to predictive maintenance and process optimization models running on hundreds of industrial sensor streams. Combines causal inference, computer vision, and time-series modeling with end-to-end ownership: data pipeline architecture, model training and validation, and cloud deployment on AWS, Azure, and Kubernetes. Track record of turning noisy plant-floor data into models that directly inform process control, equipment reliability, and operational decisions.

Experience

PROCESS POINT TECHNOLOGIES | SENIOR DATA SCIENTIST (CONTRACTED TO THE MOSAIC COMPANY) | OCT 2024 – PRESENT

Project 1 - : LIBS Spectral Calibration MLOps Pipeline — The Mosaic Company, Brazil (Jul 2026 – Present)

- Architected an Azure ML (SDK v2) pipeline to automate biweekly retraining of a LIBS spectral calibration model (SFFS + ElasticNet), replacing manual notebook execution.
- Designed a versioned Azure SQL schema and coefficient store to track and audit model versions across sites and sample types.
- Designed and built a real-time scoring service (Azure Function, scheduled inference) integrated with the plant's PI process information system.
- Designed tiered hot/cold storage (Azure SQL + Blob/Parquet archival) with sizing analysis to control long-term data growth.
- Resolved cross-team architecture trade-offs with OT/PI engineering, simplifying the system and removing redundant components.

Project 2 - : Computer Vision Proof-of-Concepts — The Mosaic Company, Brazil (May 2026 – Jun 2026)

- **Sulfur Pile Density Monitoring:** Led a POC to determine whether existing plant CCTV footage could support automated sulfur pile fill-level classification (low/medium/high), building a SAM-based segmentation pipeline seeded from color/saturation cues and a zoom-invariant sizing metric to handle a pan/zoom camera; validated across 5 calibrated clips, confirming that classical CV segmentation reliably isolates the pile across camera views. Low/medium/high threshold calibration remains in progress pending more varied footage.
- **Truck-Tipping Detection:** Led a POC evaluating truck detection and state monitoring (loaded/raised/tipped) during unloading via existing CCTV, comparing a fine-tuned RT-DETRv2 object detector (10,900+ images from 4 public datasets, 0.87 mAP@0.5 in-domain) against off-the-shelf Vision-Language Models (Qwen2.5-VL, SmolVLM) queried directly for truck state.
- Diagnosed a ~6x confidence drop when transferring the in-domain detector to real plant footage and partially closed the gap with domain-randomization training, isolating camera geometry — not model choice — as the root cause, confirmed across both detection and VLM approaches and multiple model families/sizes.
- Concluded the approach is not currently feasible with the existing camera; paused the POC pending better camera placement or in-domain labeled data, redirecting the ask from “train a better model” to “fix the camera.”

Project 3 - : Phosphate Beneficiation Process Optimization — Modelling — The Mosaic Company, Brazil (Jan 2026 – Jun 2026)

- Built a multi-stage data processing pipeline for 440+ PI Vision sensor tags and 263 Dynamox vibration monitoring points on Azure Blob Storage — chunk merging, Parquet conversion, metadata-driven range filtering, parallel fetching with thread-safe API quota management (150K requests/day), and checkpoint-based resume — with corruption detection and memory-efficient iterative processing on a 164GB VM.
- Designed physically grounded stability labelling algorithms for the UF filter process — detecting pump current drop events, post-restart settling periods, and throughput surge events — producing regime-agnostic stable/transient labels across 5 years of 1-minute resolution data (~3M rows).
- Trained and benchmarked throughput and concentrate P2O5 grade prediction models (ElasticNet, XGBoost) on 16 process features with TimeSeriesSplit CV; systematic experimentation with SHAP analysis and partial dependence plots identified feature gap as the primary bottleneck, guiding the team to prioritise upstream feed assay and reagent data collection.
- Built a tailings P2O5 mass-balance calculation pipeline using a process-engineer-specified formula, validated against lab measurements with Spearman correlation, OLS regression, and year-by-year regime analysis.
- Engineered upstream P2O5 concentration and throughput features using pump-current signals, flow measurements, stockpile characteristics, and mineralogical variables; characterized upstream-to-downstream process lag via cross-correlation analysis, and reintroduced raw stockpile composition features as an alternative to PCA-reduced scores.

- Standardized ETL pipelines and flotation-stage features across linear and non-linear model architectures (ElasticNet, XGBoost, Random Forest) to improve dataset consistency and reproducibility; retrained and validated models against updated LIBS online-analyzer datasets and prepared models for downstream optimizer integration.
- Documented XGBoost and Random Forest training workflows and registered a tree-based simulation model to support knowledge transfer and model governance.

Project 4 - : Predictive Pump-Failure Analytics Platform — The Mosaic Company, Brazil (Oct 2025 – Dec 2025)

- Built an end-to-end RUL prediction system for 12 industrial pump systems at Mosaic Brazil, using ensemble stacking regressors on PI Vision process sensors and Dynamox vibration data to predict hours until failure across four severity levels.
- Designed data extraction pipelines for PI Vision (WebAPI) and Dynamox (REST API), unifying 68+ process attributes and 94+ vibration sensors into a single configuration-driven schema; reduced PI API call overhead by ~50%.
- Architected an equipment-agnostic deployment framework with interface-based OOP design, enabling zero-code onboarding of new equipment through config files alone — deployed as Azure ML online endpoints with Non-Prod/Prod separation.
- Identified and fixed a critical scaler bug in the legacy training pipeline where MinMaxScaler was serialised before fitting; rebuilt feature alignment logic, improving preprocessing data retention from 8.8% to 99.6%.
- Developed dynamic asset-specific model selection logic and refactored legacy predictive-maintenance code into a modular, scalable ML architecture with consistent preprocessing between training and inference.
- Harmonized industrial maintenance reports with sensor telemetry to support model training and real-time scoring, enabling proactive asset management in collaboration with plant operations and technical teams.

Project 5: Industrial AI & Process Optimisation — Internal (Oct 2024 – Sept 2025)

- Conducted causal analysis of Urea and ANP fertiliser plants using clustering, anomaly detection, and time-lagged causality (Granger, TFE, CCM); identified root causes for ammonia loss and dust buildup, leading to improved operational consistency and reduced emissions.
- Reduced demurrage costs at shipping yards by building a Mixed-Integer Programming model that optimised vessel scheduling using integrated data from production, procurement, and inventory systems.
- Developed Alntropy and Alnalyzer platforms for industrial AI workflows, implementing modular pipelines for ingestion, AutoML, and visualisation; reduced project cycle time by 60%.
- Built Vision-Language Model (VLM)-based prototypes for PPE detection and unsafe ground detection in mining environments; demonstrated real-time feasibility on annotated CCTV footage with high detection accuracy.
- Conducted plant-wide process assessments across DAP, PAP, SAP, DM Water, and Power units; mapped control logic and material flows, delivering AI/ML opportunity roadmap with ROI-focused recommendations.

DRISHYA AI LABS

Drishya AI Labs is a deep-tech company specialising in AI-powered digital twin solutions for the energy and oil & gas sectors, optimising industrial operations with real-time insights.

SENIOR DATA SCIENTIST | JAN 2024 – SEPT 2024

- Enhanced OCR performance on P&ID diagrams by experimenting with LLM's and designing an ensemble pipeline using Google Vision and EasyOCR with a voting-based strategy and advanced text grouping logic, resulting in a 10% increase in extraction performance.
- Replaced clustering-based symbol-text association with a supervised ANN model, increasing mapping accuracy by 25% and improving reliability of downstream component tagging.
- Reduced model training costs by 70% by implementing automated training on AWS Spot Instances with fallback handling, without compromising training stability.

DATA SCIENTIST | MAR 2022 – DEC 2023

- Developed server-less micro-services using AWS Lambda and Step Functions for modular processing of engineering drawings, improving throughput and pipeline scalability.
- Enhanced text detection from P&ID diagrams using grouped OCR post-processing and a rolling window extraction strategy, improving text coverage by 10% and achieving 74% classification accuracy.
- Implemented a custom distance-based K-Means clustering approach for symbol-text mapping, replacing fragile heuristics and improving visual mapping consistency.
- Created a model comparator tool for evaluating modules like text extraction and symbol-text mapping with spatial and semantic validation, enabling real-time confidence scoring.
- Developed an LLM-based document understanding and extraction of structured data module using GPT-4 for extracting insights from plant sops, process manuals, and control narratives; applied prompt engineering, retrieval-augmented generation (RAG) and voting-based ensemble methods to support operator queries and incident diagnostics.

DELOITTE SHARED SERVICES INDIA LLP | SENIOR EXECUTIVE | JUN 2019 – MAR 2022

- Developed an Employee Attrition Prediction system using XGBoost and custom loss functions on 300+ employee-level features. Achieved 92% prediction accuracy and delivered SHAP/LIME-based insights for HR interventions.
- Automated Salesforce-to-PMS data integration via Azure Logic Apps, including full data migration and scheduled ETL processes, reducing manual effort and eliminating human errors.
- Built ensemble ML models (XGBoost + Logistic Regression) to predict opportunity pricing and probability of success, improving pricing decisions with 89% and 81% model accuracies respectively.
- Designed and implemented scalable model deployment infrastructure using Docker and Azure Kubernetes Service (AKS) within Virtual Network constraints, enabling secure firm-wide ML deployments.
- Led development of a policy Q&A chatbot using Machine Reading Comprehension (MRC), trained on internal policy documents; deployed with a PHP frontend and integrated issue redirection. Achieved 79% accuracy and scaled across 17 Deloitte offices.

Personal Projects

n8n-Based Outlook Email Agent

- Built an end-to-end n8n workflow for an intelligent Outlook email agent that monitors incoming mail in real time and uses AI to analyze and classify messages by importance.
- Implemented automated, content- and priority-based folder organization with instant notifications for emails needing immediate attention, eliminating manual inbox triage.
- Extending the pipeline for standalone deployment on a Raspberry Pi, to run continuously independent of a personal laptop.

Skills

- **Languages & Tools:** Python, Git, Bash, Qt5, DearPyGUI, n8n
- **Machine Learning:** XGBoost, Scikit-learn, Clustering, AutoML, Causal Inference (Granger, CCM, TFE), Anomaly Detection
- **Deep Learning & CV:** PyTorch, TensorFlow, Vision-Language Models, OCR
- **NLP & LLMs:** GPT-4, Prompt Engineering, Retrieval-Augmented Generation (RAG), Document Understanding, Machine Reading Comprehension (MRC)
- **Optimisation & Modelling:** Mixed-Integer Programming (MIP), Time Series Analysis, Process Simulation
- **Cloud & Deployment:** AWS (S3, Lambda, Step Functions, EC2), Azure (Logic Apps, ML Service, AKS), Docker, Kubernetes
- **MLOps & Infra:** Micro-services, Server-less Pipelines, Model Deployment, Custom Evaluators
- **Domains:** Chemical Process Industries, Mining Safety, Engineering Drawing Automation (P&ID)

Education

- Masters of Technology in Computer Engineering | Bennett University | 2017-2019
- Bachelors of Engineering in Computer Engineering | MCT Rajiv Gandhi Insti. of Technology (Mumbai University) | 2014-2017
- Diploma in Computer Engineering | Thakur Polytechnic (MSBTE) | 2011-2014

Certifications

- Adobe Certified Associate in Photoshop CC
- Coursera - Developing AI Applications on Azure
- Deeplearning.ai - ChatGPT Prompt Engineering for Developers