HARSHIT JAIN

(930)-904-1020 | jainharshitanil@gmail.com | linkedin.com/in/harshitaniijain | github.com/harshitjain0302

EXPERIENCE

Data Scientist **Jan 2026 – Present**
Faculty Assistance in Data Science (FADS), Indiana University Bloomington, IN

- Building scalable preprocessing and ETL pipelines integrating 10K+ multimodal records (images, audio, structured metadata), transforming high-dimensional biomedical data into statistically validated analytical datasets for predictive modeling.
- Standardizing heterogeneous data sources through schema harmonization, automated validation checks, and feature engineering workflows to improve reliability, reproducibility, and downstream model stability.
- Developing regression, classification, and CNN-based models while performing subgroup variance analysis, simulation-based robustness testing, and error profiling to evaluate model generalization across cohorts.
- Producing reproducible metrics dashboards and publication-ready reports comparing model outputs against epigenetic benchmarks, translating statistical insights into actionable recommendations for cross-functional research teams.

Data Scientist - Research **Jun 2025 – Present**
Data Science and AI Lab (DSAIL), Indiana University Bloomington, IN

- Led development of an end-to-end distributed data ingestion and validation pipeline processing metadata from 2.2M+ Hugging Face models, integrating structured/unstructured sources while enforcing data quality controls and schema consistency checks.
- Engineered scalable batch-processing workflows analyzing 100K–200K models per run, generating reliability metrics, anomaly detection summaries, and integrity simulations to monitor ecosystem-wide trends.
- Designed automated ETL processes for extracting and transforming tens of thousands of GitHub repositories, enabling high-dimensional vulnerability pattern analysis across large-scale AI systems.
- Delivered cross-functional insights supporting NSF Safe-OSE initiatives by quantifying systemic risk signals across Jetstream2 and Exosphere infrastructure, influencing engineering and security decision-making.

Machine Learning Engineer (Analytics) **Jun 2023 – Aug 2023**
Feynn Labs Remote, India

- Implemented scalable SQL and Python-based analytics workflows processing 100K+ operational records, defining performance metrics and monitoring statistical shifts across customer engagement datasets.
- Developed production-grade dashboards in Power BI and automated reporting pipelines to support executive-level operational performance tracking and anomaly detection.
- Applied regression modeling and multivariate analysis to evaluate feature performance, translating statistical findings into product optimization recommendations.
- Optimized structured data flows using validation scripts and automated reconciliation checks to enhance reporting accuracy and maintain data integrity.

PROJECTS

Security Threat Intelligence Platform — Python, Spark, Postgres, MLflow
- Architected an end-to-end ML platform orchestrating Spark ETL pipelines to process **257K+** network security events into analytics-ready tables, enabling MLflow-tracked attack classification models achieving **84%** AUC.

Cardiovascular Health Analysis — SQL, BigQuery, Looker Studio
- Engineered a cloud-based analytics pipeline analyzing **68K+** patient records and cohort-level risk drivers, delivering interactive dashboards that surfaced high-risk segments for clinical stakeholders.

Neural Network-Based Image Analysis — Python, TensorFlow, CNN, PCA
- Trained convolutional neural networks on **10K+ labeled images**, improving classification accuracy by **~25%** and applying PCA and t-SNE to explain embedding behavior and clustering structure.

SKILLS

Programming & Engineering: Python, SQL, R, Spark, Pandas, NumPy
Data Infrastructure: ETL Design, Distributed Processing, PostgreSQL, BigQuery
Machine Learning: Regression, Classification, CNNs, Dimensionality Reduction, Model Evaluation
Statistics: Hypothesis Testing, Variance Analysis, Simulation, Causal Analysis
Visualization & Reporting: Tableau, Power BI, Looker Studio

EDUCATION

Indiana University Bloomington, Master of Science in Data Science **Aug 2024 – May 2026**
University of Mumbai, Bachelor of Technology in Electronics & Telecommunication Engineering **Dec 2020 – Jun 2024**