Vraj Shaileshbhai Patel

Jersey City, NJ • +1 (605) 212-0452 • pvraj0019@gmail.com
linkedin.com/in/vraj-s-patel • github.com/Vraj2712 • Portfolio

Computer Science M.S. candidate (May 2026) building and deploying **machine learning on large-scale clinical and financial data** end-to-end SQL feature engineering, gradient boosting / ensembles, leak-free evaluation (walk-forward, AUC-ROC, calibration), and AWS deployment - plus **LLM / GraphRAG systems** (LangChain, Neo4j, vector + graph retrieval). Strong in Python and SQL with reproducible workflows and clear communication to technical and non-technical stakeholders.

Education

New Jersey Institute of Technology, Newark, NJ — Sep 2024 – May 2026
M.S. in Computer Science | GPA: 3.80 / 4.0

Gujarat Technological University (SAL Institute), Ahmedabad, India — Sep 2019 – May 2023
B.E. in Computer Engineering | GPA: 8.13 / 10

Experience

New Jersey Institute of Technology — Graduate Researcher — Sep 2025 – Present
Clinical Data Analysis (PLCO Cancer-Screening Trial) | Advisor: Prof. Arashdeep Kaur

- Analyze the large-scale PLCO cancer-screening dataset (older-adult participants) in Python (Pandas), running EDA and statistical testing to rank candidate predictors and define the feature set used for downstream risk modeling.
- Process **confidential patient health and medical-imaging data** under strict privacy and governance standards, keeping sensitive records secure and analysis-ready without exposing protected health information.
- Develop preprocessing and feature-engineering pipelines, cutting redundant variables by 30% via data-driven selection; evaluate outputs with AUC, F1, and Brier score and communicate findings through a standardized, reproducible workflow usable by technical and non-technical members.

Mind Inventory, Ahmedabad, India — Jan 2023 – Apr 2023
Intern — Web Systems Design & Development

- Built and maintained scalable, database-driven web systems (PHP, MySQL/XAMPP) in an agile team, delivering responsive UI across 3+ platforms (∼30% faster page loads) and contributing to QA / bug-fix cycles.

EliteEvince Technologies, Ahmedabad, India — Jun 2022 – Jul 2022
Web Developer Intern (PHP & MySQL)

- Translated real-world client requirements into deliverables under tight timelines, resolving UI and documentation issues across 2 internal projects to improve maintainability and usability.

Projects

GraphRAG — Knowledge-Graph Q&A over Cancer-Prediction Research — GitHub • Live Demo
Python, Neo4j (Cypher), LangChain, Claude, Gemini embeddings, FastAPI, React, TypeScript, Vercel, Render

- Built a GraphRAG question-answering system over 16 peer-reviewed cancer-prediction papers (prostate, lung, colorectal, ovarian), extracting entities and relationships into a Neo4j knowledge graph with Claude via LangChain's `LLMGraphTransformer`.
- Combined vector search (Gemini embeddings + Neo4j vector index) with one-hop graph traversal to answer multi-hop questions plain RAG can't, surfacing the exact Entity → RELATIONSHIP → Entity facts behind every answer for full explainability.
- Shipped an end-to-end full-stack app — FastAPI REST backend and a React + TypeScript + Vite + Tailwind chat UI — deployed across Render, Vercel, and Neo4j Aura; migrated embeddings from a local model to the Gemini API to fit a 512 MB target without re-running costly LLM extraction.

Lung Cancer Risk Prediction - Data Efficiency & Stability Analysis — GitHub
Python, scikit-learn, XGBoost, LightGBM, CatBoost, pandas, Matplotlib

- Designed a super-stacking ensemble combining six base learners (Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost) with a logistic-regression meta-learner to predict lung cancer risk from 15 clinically interpretable variables in the PLCO Cancer Screening Trial.
- Ran a data-efficiency study training on progressively smaller fractions (100% → 10%) against a fixed, stratified 20% test set with 5 random seeds per fraction, reporting all metrics as mean ± std for reproducibility.
- Demonstrated strong robustness under data scarcity (AUC-ROC held near 0.836, dropping only ∼0.005 at 10% training data), while mitigating severe class imbalance and tuning the decision threshold on validation only (F1-max) to avoid test-set leakage.

VolGuard - End-to-End Stock Volatility Forecasting — GitHub • Live Dashboard
Python, scikit-learn, XGBoost, SQL (SQLite), AWS (Lambda, S3), Terraform, Tableau

- Built an end-to-end ML pipeline forecasting 5-day realized volatility for five large-cap stocks: raw prices → SQL feature engineering → validated model → AWS Lambda → a live Tableau dashboard.
- Engineered features entirely in SQL (daily returns, 5/10/30-day rolling volatility, forward-5-day target with no look-ahead) and enforced strict chronological splits to prevent data leakage.
- Validated with walk-forward testing across 4 out-of-sample folds, benchmarking Linear Regression, Random Forest, XGBoost, and GARCH(1,1) against a naive baseline; shipped the simplest model (RMSE ∼20% below baseline) to AWS via Terraform (S3 + Lambda, API-key auth) feeding the public dashboard.

Skills

ML & GenAI	scikit-learn, XGBoost, LightGBM, CatBoost, ensemble / stacking, ROC-AUC / model evaluation, feature selection, LangChain, RAG / GraphRAG, embeddings, Claude / Gemini APIs
Data & Healthcare	Python (Pandas, NumPy), EDA, statistical analysis, Matplotlib; clinical dataset & confidential health-data handling, medical-imaging research
Databases	SQL, PL/SQL, MySQL, PostgreSQL, MongoDB, Neo4j (Cypher)
Web & Cloud	FastAPI, React, TypeScript, AWS (Lambda, S3), Terraform, Vercel, Render, Docker
Tools & Languages	Git / GitHub, VS Code, Jupyter, Linux CLI, Tableau, Excel | English, Hindi, Gujarati