

SUR VAGHASIYA

Stony Brook, NY | +1 (631) 513-5446 | survaghasiya@gmail.com | [LinkedIn](#) | [GitHub](#)
AI/ML Engineer · Large Language Models · RAG Systems · Agentic Workflows · Neural Networks · Deep Learning · Scalable ML · MLOps

EDUCATION

Stony Brook University

Master of Science in Data Science

Relevant Coursework: Machine Learning, Statistical Computing, Probability & Statistics, Computer Science Fundamentals, Data Science Fundamentals, Data Analysis, Data Management

New York, United States

Expected May 2027

Indus University

Bachelor of Technology in Computer Science Engineering | GPA: 3.4 / 4.0

Relevant Coursework: Machine Learning, Natural Language Processing, Deep Learning, Data Structures & Algorithms, Design Analysis & Algorithms, Operating System, Database Management System, Cloud Computing, IoT, Data Science, Web Technology

Ahmedabad, India

May 2025

SKILLS

Languages: Python (NumPy, Pandas, Matplotlib, Seaborn, SciPy, Prophet, Statsmodels), R (ggplot2, data.table, glmnet, caret, tidyverse), SQL (MySQL, PostgreSQL), TypeScript, JavaScript, Solidity, C++, Java, Shell Scripting

Technical Skills: TensorFlow, PyTorch, Scikit-learn, Keras, XGBoost, ARIMA, YOLOv5, OpenCV, LangChain, FAISS, HuggingFace Transformers (BERT, GPT), Apache Spark, PySpark, MLflow, Groq, OpenAI API (GPT-4o), Claude (Haiku, Opus 4.7), DALL-E 3, Retell AI, boto3, MCP, FastMCP, CopilotKit, AG-UI/A2UI, Pydantic v2, NetworkX, GeoPandas, osmnx, yfinance, SHAP, Plotly, Recharts

Core Competencies: LLMs, Generative AI (Gen AI), RAG Systems, Agentic Workflows, Multimodal AI, Voice AI (TTS/STT), Transformer Architecture, Deep Learning (CNN/RNN/LSTM), Model Fine-Tuning, Hyperparameter Tuning, Model Evaluation & Optimization, Model Compression (Pruning, Quantization), Computer Vision, Natural Language Processing (NLP), Prompt Engineering, MLOps, Real-Time ML, Time-Series & Ensemble Forecasting, Feature Engineering, Data Pre-processing & Cleaning, EDA, Data Visualization, Clustering, Geospatial & Graph Analytics, A/B Testing, Drift Detection, SHAP Explainability, Cross-Validation, Vector Search & Embeddings, Statistical Inference, Predictive Modeling, Hypothesis Testing, Changepoint Detection, Online/Streaming ML, Regression & Classification, Cloud Model Deployment, Overfitting Mitigation, Big Data Analytics, Supervised & Unsupervised Learning, Neural Networks, Data Pipelines, Transfer Learning, Object Detection, Data Structures & Algorithms

Tools & Technologies: FastAPI, Flask, Streamlit, Docker, AWS (EC2, S3, Lambda, SageMaker), PostgreSQL, Redis (Pub/Sub), SQLAlchemy, WebSocket, RESTful APIs, React, Git, CI/CD, Next.js, Tailwind CSS, Framer Motion, MinIO, Cloudflare Workers, Upstash Redis, Hono.js, Vercel, Railway, ethers.js, MetaMask

EXPERIENCE

Adani Enterprise Ltd.

Data Science & AI Intern | Python, Flask, Scikit-learn, AWS (EC2, S3, Lambda), Pandas, NumPy, Matplotlib

Jan 2025 – Jul 2025

Ahmedabad, India

- Engineered AI-powered sales & revenue forecasting engine (Python, Flask, Scikit-learn) achieving **78% accuracy**; deployed on AWS dashboards reducing analysis time **34%** for 10+ product teams.
- Automated model retraining pipelines, cutting operational overhead **37%**.
- Conducted EDA and data pre-processing (imputation, normalization, encoding, outlier detection) on enterprise sales datasets; applied hyperparameter tuning (GridSearchCV) and model evaluation protocols (MAE, RMSE, cross-validation, A/B testing) to validate predictive accuracy and drive business insight generation.
- Administered AWS cloud infrastructure (EC2, S3, Lambda) with CI/CD pipeline integration, cutting maintenance tasks **23%** and deployment time **44%** while sustaining production-grade model reliability.

Tech Micra

Python Machine Learning Intern | Python, Scikit-learn, Pandas, NumPy, Git

Jun 2024 – Jul 2024

Ahmedabad, India

- Developed end-to-end ML pipelines (data pre-processing, cleaning, normalization, encoding, feature engineering); EDA via correlation matrices and ANOVA boosted interpretability **42%**.
- Grid Search CV reduced overfitting **35%** and improved generalization **27%** across 43+ Agile commits.
- Applied LASSO/Ridge regularization and systematic feature selection for model selection & refinement; maintained full reproducibility via Git version control in an Agile five-member team.

Shaligram Infotech

Web Development Intern | JavaScript, HTML5, CSS, MySQL, WordPress

Jun 2023 – Jul 2023

Ahmedabad, India

- Designed 5 responsive cross-browser web pages (HTML/CSS/JS) improving load efficiency **48%** and 2 MySQL-backed WordPress client sites, boosting delivery efficiency and client satisfaction **45%**.

PROJECTS

LI Traffic Accident Cluster & Severity Analytics | R, Python, DBSCAN, FAMD, XGBoost, SHAP, GeoPandas, ggplot2 |

Apr 2026

- Pinpointed 12 crash hotspot clusters (DBSCAN + Haversine) and 4 risk archetypes (FAMD + k-Means, silhouette=0.4268) from 33,228 Long Island records.
- Validated k-Prototypes (k=3 to 8) via bootstrap (Jaccard ≥ 0.40 , ARI ≥ 0.65); applied SHAP per-cluster attribution via Random Forest.
- Trained XGBoost severity classifier (AUC **0.7185**, 5-fold CV 0.7164 ± 0.0042 , Acc **71%**) on 27 engineered features including sin/cos temporal, weather dummies, night-rain and rush-junction interactions.
- Applied LASSO regression baseline; KDE heatmaps and osmnx + NetworkX road-graph overlays support logistics risk modeling and last-mile delivery routing optimization.

SAARTHI | Python, FastAPI, Groq, Claude (Haiku/Opus 4.7), Retell AI, Scikit-learn, PostgreSQL, Next.js, FastMCP |

Mar 2026

- Architected voice-first ENS rare-disease monitor (HackRare 2026): Retell AI automates daily patient calls; Claude Haiku extracts 7 validated clinical scores with ENS bias correction (prosody signals, temporal context) and keyword-heuristic fallback: zero patient form-filling.
- Orchestrated tri-level ML prediction: Global RF (n=2,000) to Personal RF (StratifiedKfold + F1) to Z-score override (Z-score threshold ≥ 2 or ≥ 3); Welford O(1) streaming baselines.
- Multi-LLM routing (Groq + Claude Opus 4.7), FastMCP ENS server, and 4-table PostgreSQL schema (SQLAlchemy + Pydantic v2); deployed on Vercel and Railway.
- Built Next.js doctor dashboard (Recharts trend charts, CopilotKit AI copilot, jsPDF clinical reports); dual-path transcript delivery (Retell webhook + polling fallback with deduplication guard) ensuring zero transcript data loss across all patient monitoring sessions.

InsightX | Python, FastAPI, Next.js, Prophet, XGBoost, ARIMA, Redis, PostgreSQL, MLflow, Docker |

Feb 2026

- Achieved **33% accuracy gain** via Prophet+XGBoost+ARIMA ensemble (40/40/20%, STL decomposition); P10/P90 confidence intervals, K-S drift detection, CI Width Ratio reliability scoring, rolling MAE/RMSE backtests, and MLflow artifact tracking for full model lineage.
- Eliminated HTTP timeouts on 5-30s ML inference via Redis Pub/Sub + WebSocket streaming; A2UI/MCP AI Copilot renders React charts in chat with a War Games crisis simulator (4 scenarios).
- Built live DB connectors for MySQL, PostgreSQL, and Google Sheets with yfinance integration; containerized as a 7-service Docker stack.

EchoMindAI | Python, Streamlit, LangChain, OpenAI GPT-4o, FAISS, HuggingFace, Groq, DALL-E 3, FastMCP |

Jan 2026

- Deployed multimodal RAG system (FAISS IndexFlatL2 k=4, LangChain ReAct, HuggingFace Sentence Transformers) with 16 live agentic tools: vision (GPT-4o), DALL-E 3, stocks, news, travel, 100+ language translation, PDF/DOCX/CSV/PPTX ingestion; DuckDuckGo fallback on no local match.
- Accelerated STT latency **7.5x** (1,500 to 200 ms) via Groq LPU; CSS Injection Engine (29 KB) + HTML Bridge renders 60fps glassmorphism UI from LLM token streams.
- 4-layer heuristic recovery cut failures **90%** with zero crashes; FastMCP server, LRU cache, and tenacity retry yielded **300% engagement lift**.

CERTIFICATIONS & AWARDS

[AWS Academy Cloud Foundations](#) | [AWS Academy ML Foundations](#) | [Python Bootcamp](#) | [HackRare 2026](#)