

SUR VAGHASIYA

Stony Brook, NY | +1 (631) 513-5446 | survaghasiya@gmail.com | [LinkedIn](#) | [GitHub](#)

Software Engineer · AI/ML Systems · Full-Stack Development · LLMs & RAG · Agentic Workflows · Web3 & Smart Contracts · Real-Time Systems · MLOps

EDUCATION

Stony Brook University

New York, United States

Master of Science in Data Science

Expected May 2027

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

Indus University

Ahmedabad, India

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

May 2025

Relevant Coursework: Machine Learning, NLP, Deep Learning, Data Structures & Algorithms, Design Analysis & Algorithms, Operating System, Database Management System, Cloud Computing, Web Technology

SKILLS

Languages: Python (NumPy, Pandas, Matplotlib, Seaborn, SciPy, Prophet, Statsmodels), TypeScript, JavaScript, SQL (MySQL, PostgreSQL), Solidity, C++, Java, Shell Scripting

AI/ML Frameworks & Libraries: TensorFlow, PyTorch, Scikit-learn, Keras, XGBoost, ARIMA, LangChain, FAISS, HuggingFace Transformers, MLflow, Groq, OpenAI API (GPT-4o), Claude (Haiku, Opus 4.7), DALL-E 3, Retell AI, FastMCP, CopilotKit, AG-UI/A2UI, Pydantic v2, Recharts

Core Competencies: LLMs & RAG Systems, Agentic Workflows, Multimodal AI, Voice AI (TTS/STT), MLOps, Real-Time ML Systems, Time-Series & Ensemble Forecasting, Feature Engineering, Vector Search & Embeddings, Prompt Engineering, Distributed Systems, Statistical Inference, Clustering & Classification

Backend & Infrastructure: FastAPI, Flask, Next.js (App Router), React, Redis (Pub/Sub), Docker, AWS (EC2, S3, Lambda), PostgreSQL, SQLAlchemy, WebSocket, RESTful APIs, Git, CI/CD, Cloudflare Workers, Upstash Redis, Hono.js, Vercel, Railway, ethers.js, MetaMask, Hardhat

EXPERIENCE

Adani Enterprise Ltd.

Jan 2025 – Jul 2025

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

Ahmedabad, India

- Engineered AI-powered sales & revenue forecasting engine (Python, Flask, Scikit-learn) achieving **78% accuracy**; deployed 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

Jun 2024 – Jul 2024

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

Ahmedabad, India

- Developed end-to-end ML pipelines (data pre-processing, cleaning, normalization, encoding, feature engineering); EDA via correlation matrices & 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

Jun 2023 – Jul 2023

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

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

SAARTHI | Python, FastAPI, Flask, Groq, Claude Haiku/Opus 4.7, Retell AI, Scikit-learn, PostgreSQL, Next.js, FastMCP, Vercel | 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)→Personal RF (StratifiedKFold+F1)→Z-score override ($|Z| \geq 2/3$); Welford O(1) baselines; Multi-LLM routing (Groq+Claude Opus 4.7); FastMCP ENS server; 4-table PostgreSQL (SQLAlchemy+Pydantic v2); Vercel+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, Docker, MLflow, MCP, AG-UI/A2UI | Feb 2026

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; War Games crisis simulator (4 scenarios); live DB connectors (MySQL, PostgreSQL, Google Sheets) + yfinance; 7-service Docker stack.

EchoMindAI | Python, Streamlit, LangChain, OpenAI GPT-4o, FAISS, HuggingFace, Groq, DALL-E 3, Docker, FastMCP | Jan 2026

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 **7.5×** (1,500→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%**, zero crashes; FastMCP server; LRU cache; tenacity retry; **300% engagement lift**.

Decentralized Escrow Platform | Solidity, React, TypeScript, ethers.js, MetaMask, Hardhat, Hono.js, Cloudflare Workers, Upstash Redis | Oct 2024

- Engineered and deployed a Solidity escrow smart contract on Ethereum Sepolia testnet enforcing dual-party approval, deterministic fund locking, and refund-safe resolution via strict access controls; eliminating intermediaries for fully trustless peer-to-peer transactions with zero counterparty risk.
- Built React + Tailwind CSS frontend with MetaMask and ethers.js; Hono.js + Cloudflare Workers serverless backend with Upstash Redis powering "my contracts" views; Hardhat test suite validates dual-party approval, fund locking, and refund-path correctness across all contract states.

CERTIFICATIONS & AWARDS

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