

VARUN SHARMA

Karnal, Haryana, India | +91 8376865125 | sharma.varun1213@gmail.com
[linkedin.com/in/varun-sharma97](https://www.linkedin.com/in/varun-sharma97) | github.com/vshar102

SUMMARY

AI/ML engineer with an M.S. in Data Science from Drexel University (GPA 3.91/4.0) and two years of production software engineering. Builds applied AI systems for documents and medical imaging — retrieval-augmented generation, LLM fine-tuning and computer vision — with a consistent focus on evaluating whether they actually work. Currently building multimodal video-processing pipelines at Modern Streaming. Based in India and available for globally remote roles, working client hours.

TECHNICAL SKILLS

Languages & Backend: Python, TypeScript, C#/.NET, SQL (PostgreSQL, MySQL, SQL Server, SQLite), FastAPI, AsyncIO, REST APIs

AI & ML: RAG (LangChain, LlamaIndex), vector stores (pgvector, Pinecone, Chroma), sentence-transformers, LLM fine-tuning (LoRA/PEFT, Hugging Face TRL), PyTorch, scikit-learn, OpenCV, YOLOv8, ViT/Swin/CLIP, OCR (Tesseract, PaddleOCR)

Data & Cloud: Pandas, MLflow, GCP (Cloud Run), AWS (S3, Lambda), Docker, n8n

Practices: Git, CI/CD (GitHub Actions), Agile/Scrum, HIPAA & 21 CFR Part 11 controls

PROFESSIONAL EXPERIENCE

Modern Streaming — Software Developer

Dec 2025 – Present

Karnal, India | Full-time

- Build multimodal processing pipelines using YOLOv8, OpenCV and LLM-based components to analyze and index unstructured video telemetry, enabling searchable retrieval for auditing workflows.
- Implemented a multi-threaded frame-extraction pipeline with frame-skipping logic to improve throughput on high-resolution video streams.
- Applied KV caching and quantization to reduce inference latency on multimodal query workloads.
- Built validation workflows combining deep-learning detection with classical computer vision (HSV thresholding) for automated compliance checks.

Triveni IT — Software Developer

Aug 2025 – Dec 2025

Remote (New Jersey, US) | Part-time

- Worked with real-world datasets on collection, cleaning and analysis using Python and SQL.
- Built analytical workflows and validation checks to assess and improve data quality.
- Developed predictive models and presented findings in a structured format to support decision-making.

Smart Rewards Inc. — AI/ML Engineer

Mar 2025 – Aug 2025

Remote (New York, US) | Internship

- Built a retrieval-augmented internal assistant using LangChain orchestration over a Pinecone vector index, reducing manual document search time for support agents.
- Stood up an evaluation harness for ongoing retrieval-quality monitoring, and improved adoption by adding source citations, confidence scores and a user feedback loop.
- Deployed on GCP Cloud Run with Docker and GitHub Actions CI/CD; instrumented with OpenTelemetry tracing and Grafana dashboards.
- Integrated clinical trial data across EDC, CTMS, LIMS and ePRO systems into automated reporting pipelines, building reusable Power BI templates and standardized Python/SQL Server ETL under HIPAA and 21 CFR Part 11 controls.
- Authored runbooks and led internal training sessions for onboarding users.

Drexel University, College of Computing & Informatics — Research Assistant

Oct 2024 – Mar 2025

Philadelphia, PA | Remote

- Directed dataset curation and model development on 10,000+ mammograms (CBIS-DDSM, VinDr-Mammo) for a Drexel-Thomas Jefferson Hospital collaboration under full HIPAA governance — de-identification, data use agreements and audit logging.
- Engineered a dual-view fusion architecture for CC/MLO mammography views, fine-tuning ViT, SWIN-T and CLIP backbones.
- Established patient-level train/validation/test splits, augmentation and hyperparameter sweeps with GitHub versioning and reproducibility controls; mentored a junior researcher on preprocessing and annotation standards.

Drexel University, College of Computing & Informatics — Research Intern

Sep 2023 – Dec 2023

Philadelphia, PA | Remote

- Modeled coronary heart disease risk stratification on longitudinal EHR/claims data (~250K patients, 8 years, 120 features), benchmarking logistic regression, random forest and XGBoost against deep generative models (MAF, RealNVP).
- Built reproducible data and feature pipelines in Python/SQL on a secure Linux research cluster; harmonized Epic extracts inside a de-identified, IRB/HIPAA-controlled enclave.
- Engineered domain features (rolling labs, comorbidity index, refill gaps), handled class imbalance with weighting and Platt scaling, and reviewed SHAP explanations with clinicians.

Cognizant — Junior Software Engineer

Jul 2020 – Jul 2022

Bengaluru, India | Full-time

- Full-stack .NET engineer owning feature delivery for three regulated pharmaceutical systems (SFC, JPUBS, ReCAP3) across .NET Framework, C#, SQL Server, JavaScript and Azure DevOps.
- Scaled field-representative workflows and upgraded validation and exception handling to reduce support tickets and improve reliability.
- Modernized publication approval flows and resolved high-severity defects to accelerate submissions.
- Implemented repeatable testing and debugging practices in partnership with QA, and automated build and deployment pipelines in Azure DevOps.

EDUCATION

Drexel University — M.S., Data Science | GPA 3.91/4.0

Sep 2022 – Jun 2024

Philadelphia, PA

Coursework: Applied Machine Learning, Data Science Capstone (A+), Healthcare Informatics, Information Retrieval Systems, Applied Cloud Computing. Inducted into Upsilon Pi Epsilon, the international honor society for computing, Apr 2024.

SRM University — B.Tech, Computer Science & Engineering

Jul 2016 – May 2020

Chennai, India

SELECTED PROJECTS (github.com/vshar102)

- **Rocathon Hybrid Search Engine** — TypeScript, Postgres/pgvector, local MiniLM embeddings. Two-stage creator search: vector retrieval followed by hybrid re-ranking across semantic relevance, commercial performance and brand-audience fit, with a categorical zero-GMV veto. Type-checked in CI.
- **Resume-ATS** — Streamlit screening application scoring candidates across six weighted dimensions, combining sentence-transformer similarity with rule-based extraction; concurrent PDF parsing via a thread pool, SQLite persistence and Plotly reporting.
- **Document-Backed Q&A Bot** — LangChain RAG over user-supplied PDFs with a Groq-hosted Llama 3.1 model, surfacing the retrieved source chunks alongside every answer so responses can be verified against the document.

CERTIFICATIONS & EXTERNSHIPS

- Advanced Data Engineering with Snowflake — Snowflake (Oct 2025)
- Outamation Document Intelligence Externship — OCR and RAG pipelines for document classification and extraction (Tesseract, PaddleOCR, LlamaIndex)
- Hydroficient IoT Cyber Defense Externship — threat modeling, MQTT security, TLS/mTLS, ML-based anomaly detection
- Extern Healthcare Operations & Data Analytics Program

PUBLICATION

Classification of Malignant Melanoma using BPN and ABCD Rule. International Journal of Psychosocial Rehabilitation, 2020.