Ansh Kumar Dev

520-599-5589 | anshkumardev@gmail.com | linkedin.com/in/ansh-kumar-dev/ | github.com/Anshkumardev

EDUCATION

University of Arizona — Tucson, AZ
Master of Science in Data Science — *Aug. 2023 – May 2025*

Gautam Buddha University — Greater Noida, India
B.Tech. in Artificial Intelligence — *Aug. 2019 – May 2023*

EXPERIENCE

AI Engineer Intern — Jun 2025 – Present
Right Skale — *Pleasanton, CA*
- Built and maintained structured **"one source of truth" datasets** by ingesting SEC 10-Q and balance sheet data into standardized schemas using containerized microservices (LlamaParse, LlamaExtract, vector-DB).
- Led development of robust **data pipelines** with built-in **validation, anomaly filtering, and quality checks**, ensuring integrity across 1M+ document chunks.
- Implemented **data governance** controls and CI/CD workflows (CloudWatch, X-Ray, IAM, GitHub Actions) to support SOX-compliant, auditable deployments.
- Collaborated cross-functionally with **data analysts, product teams, and cloud engineers** to resolve data issues, troubleshoot edge cases, and align pipelines to evolving client requirements.
- Recommended and implemented enhancements that **standardized and streamlined data workflows**, reducing processing time by over 30% and increasing delivery accuracy.

Data Scientist (Graduate Research Assistant) — Aug 2024 – Jan 2025
University of Arizona — *Tucson, AZ*
- Designed and deployed a full-stack **ETL pipeline in Python (Pandas, SQL)** to ingest and clean satellite imagery & sensor data, ensuring **data quality and transformation consistency** across systems.
- Validated joins and feature integrity across geospatial and time-series datasets using exploratory analysis and **data transformation testing** in Jupyter and SQL.
- Mentored 15+ students on **data wrangling**, QA best practices, and performance tuning techniques for large-scale model-ready datasets.
- Partnered with environmental researchers and IT staff to **troubleshoot data anomalies**, refine workflows, and improve downstream modeling reliability.

Research Data Scientist — Jun 2022 – Nov 2022
Gautam Buddha University — *Greater Noida, India*
- Conducted **root cause analysis** and anomaly detection in IoT traffic using **C++** and Python, boosting data throughput by **20%** and cutting latency by **25%**.
- Designed and benchmarked **QoS-aware streaming algorithms** with strong emphasis on **data validation and statistical integrity** (Published in e-Prime, Elsevier).
- Developed real-time anomaly detection protocols that improved reliability of ML systems by **40%**; presented at LICTIC 2023.

AI Engineer Intern — Apr 2022 – Jul 2022
Merkletree Technologies — *Delhi, India*
- Optimized model deployment pipelines using **Docker, Kubernetes, Flask**, and MLflow, reducing deployment errors and **streamlining data flow** into production systems.
- Interfaced with Android and backend teams to align **data validation and inference pipelines**, ensuring delivery of accurate, real-time authentication predictions for **50k+ users**.
- Verified integrity of pre-processed features and tested transformation logic prior to production deployment, reducing bugs and improving model performance.

PROJECTS

Healthcare Accessibility & Risk Mapping Platform | *Project Link* — Apr 2025
- Built a distributed **GeoPandas + OSMnx** pipeline unifying Geospatial, socioeconomic & clinical data across 3,000+ U.S. census tracts to identify underserved areas for policy action.
- Developed a **LangChain** multi-agent LLM workflow with **OpenAI API** and **FAISS** RAG and a **Streamlit** choropleth dashboard, enabling real-time Q&A and evidence-based insights.

TECHNICAL SKILLS

Languages/ML: Python (Pandas, NumPy, Scikit-learn, TensorFlow, PyTorch, OpenCV), R, SQL, C++
Models: RoBERTa, SVM, Decision Trees, K-Means, PCA, Forecasting & Statistical Analysis
Cloud & MLOps: AWS (EKS, EC2, S3), Azure (ADF, Synapse, Databricks), GCP, Docker, Kubernetes, Spark, CI/CD, MLflow
Tools & APIs: Flask, REST APIs, TensorBoard, Git, Linux, Power BI, Streamlit