Shivaraj Senthil Rajan

shse1502@colorado.edu | (720)-260-6977 | Boulder, CO (Open To Relocate)
https://github.com/Shiva250503ss | https://linkedin.com/in/shivaraj-senthil-rajan-2b8898227 |
https://shiva250503ss.github.io/shivaraj-portfolio/

Education

University of Colorado Boulder Boulder, CO
Master of Science in Data Science; GPA: 3.9/4.0 *Aug 2024 - May 2026*

Anna University Chennai, Tamil Nadu
Bachelor of Science in Artificial Intelligence and Machine Learning; GPA: 8.52/10 *Aug 2020 - May 2024*

Technical Skills

Programming: Python, SQL, R, PySpark, Java, C++, CUDA, Scala, JavaScript, Shell Scripting, Go, Julia, MATLAB

ML/AI & NLP: PyTorch, TensorFlow, scikit-learn, XGBoost, LightGBM, CatBoost, Transformers, LLMs, RAG, BERT, GPT-4, LangChain, Hugging Face

Data Science & Analytics: Pandas, NumPy, SciPy, Matplotlib, Seaborn, Plotly, A/B Testing, Time Series Analysis, Feature Engineering, Statistical Modeling

Big Data & Streaming: Apache Spark, Kafka, Airflow, Flink, Hadoop, Hive, Presto, dbt, Delta Lake, Iceberg, Databricks

Cloud & Infrastructure: AWS (SageMaker, EC2, S3, EMR, Glue, Redshift, Lambda), GCP (Vertex AI, BigQuery, Dataflow), Azure (ML, Synapse, Data Factory)

Databases & Storage: PostgreSQL, MongoDB, Cassandra, Redis, Elasticsearch, Snowflake, Pinecone, ChromaDB, FAISS

BI & Visualization: Tableau, Power BI, Looker Studio, Excel (Power Query/Pivot/VBA), Streamlit, DAX, SSRS

MLOps & DevOps: Docker, Kubernetes, MLflow, CI/CD, GitHub Actions, Terraform, Great Expectations, TorchServe, DVC

Business & Collaboration: JIRA, Confluence, Agile/Scrum, Stakeholder Management, Requirements Gathering, ERD Design, Data Modeling

Experience

Software Engineering Intern - AI/ML May 2025 - Nov 2025
PM Accelerator *Florida, US*

- Architected end-to-end **RAG pipelines** using **LangChain**, **Pinecone**, and **OpenAI GPT-4o** on **Databricks MLflow**, accelerating retrieval by **40%** across **10,000+ documents** monthly while maintaining **98% semantic accuracy**.
- Orchestrated distributed **Apache Spark** pipelines processing **50TB+ daily workloads**, implementing **partition optimization** and **broadcast joins** that slashed latency from **8.1 seconds to 2.3 seconds** (**70% reduction**).
- Designed self-serve **KPI dashboards** in **Tableau** and **Power BI** featuring **DAX measures** and **row-level security** for **750+ stakeholders**, catalyzing **3.5x** engagement surge while curtailing ad-hoc requests by **42%**.
- Integrated **GPT-4o NLP models** serving active user base through advanced **prompt engineering** and **fine-tuning**, achieving **98% classification accuracy** validated via **k-fold cross-validation**.
- Streamlined **CI/CD deployment** with **Docker**, **Kubernetes**, and **GitHub Actions**, expediting model releases by **35%** while maintaining **99.9% uptime**.
- Pioneered **real-time streaming architecture** using **Kafka** and **Flink**, ingesting **100K+ events/second** with **exactly-once semantics** and **sub-second latency**.

Consultant Intern Feb 2023 - May 2024
Sandron Impex Private Limited *Bengaluru, India*

- Engineered **anomaly detection** system using **Isolation Forest** and **autoencoders**, uncovering **92%** of fraudulent transactions across **1,000+ monthly invoices** with **minimal false positives**.
- Assembled scalable **ETL pipelines** with **PySpark** and **Databricks**, transforming **1B+ records monthly** at **99.5% data accuracy** through **schema validation** and **Great Expectations** checks.
- Spearheaded **cloud migration** initiative coordinating 6-person team, transitioning legacy systems to **AWS S3 data lake** with **Delta Lake**, quadrupling throughput to **500GB/hour**.
- Forecasted demand patterns utilizing **ARIMA** and **Prophet** time series models, trimming processing cycles by **15%** through predictive optimization.
- Established **automated dashboards** in **Tableau** tracking **$200K monthly marketing expenditure**, enabling **ROI optimization** that boosted conversions by **12% MoM**.

Projects

Smart Fields - AI-Powered Precision Agriculture [Link] *PyTorch, ResNet, Flask, Delta Lake, Kafka*

- Constructed **CNN-based disease detection** using **ResNet-9** architecture across **38 disease categories** from **10,000+ leaf images**, attaining **94% F1-score** verified through rigorous **k-fold cross-validation**.
- Formulated **LSTM networks** integrating **weather API data** (temperature, rainfall, humidity), delivering **90% prediction accuracy** via **multi-variable correlation analysis** on **15+ environmental parameters**.
- Developed **Delta Lake** infrastructure processing **100TB+ IoT sensor data**, leveraging **Z-ordering** and **partition pruning** to achieve **90% compression ratio**.
- Implemented **event-driven ingestion** with **Kafka Connect** managing **1M+ sensor readings/minute** with **exactly-once delivery** guarantees and **schema registry**.

DataPilot-AI - Intelligent Data Analysis Platform [Link] *LangChain, OpenAI GPT-4, ChromaDB, Streamlit*

- Devised **RAG-based** conversational AI combining **LangChain**, **GPT-4**, and **ChromaDB vector embeddings**, diminishing query response time by **60%** for analyst workflows.
- Crafted **natural language to SQL** translation employing **LLM agents** with context-aware prompting and dynamic **schema understanding** for multi-database connectivity.
- Delivered interactive **Streamlit** dashboard with **real-time visualizations** and **caching strategies** supporting CSV, Excel, and database connections.

TrustLens AI - Real-Time Toxicity Detection [Link] *BERT, FastAPI, Chrome Extension, Transformers*

- Launched **BERT-based multi-label toxicity classifier** utilizing **Hugging Face Transformers** with **GPU-accelerated inference**, categorizing **6 toxicity types** in real-time.
- Configured **FastAPI async backend** with **Uvicorn ASGI** server handling **5 concurrent workers**, seamlessly integrated with **Chrome Extension** (Manifest V3) for Reddit analysis.
- Synthesized **source credibility verification** pipeline incorporating DNS resolution, SSL validation, and page classification for three-tier trustworthiness scoring.

AIDIY - AI-Powered Financial Literacy Platform [Link] *GPT-4o, MongoDB, Redis, FastAPI, React*

- Deployed **GPT-4o** integrated EdTech platform with **speech-to-text** processing and **personalized learning paths** through adaptive AI recommendations.
- Fashioned **RESTful API architecture** incorporating **JWT authentication**, **MongoDB aggregation pipelines**, and **Redis caching** for real-time synchronization across multi-child profiles.
- Produced **gamification engine** with behavioral analytics dashboard, monitoring engagement patterns and presenting **multi-dimensional assessment scoring**.

Virtual Try-On - Computer Vision Pipeline [Link] *PyTorch, AWS S3, Docker, TorchServe*

- Executed **distributed data pipeline** on **AWS S3** with **batch processing** for **10K+ garment images**, incorporating **data versioning** with **DVC**.
- Activated **containerized inference** using **Docker** and **TorchServe**, reaching **30ms latency** through **model optimization** and **GPU batching**.
- Validated **pose estimation** preprocessing pipeline, ensuring **SSIM improvement of 5-8%** through quality-controlled training data curation.

CU Boulder Campus Safety Dashboard [Link] *Streamlit, Plotly, Power Query, Pandas*

- Generated **executive BI dashboard** with **drill-down capabilities**, **trend analysis**, and **KPI scorecards** for campus leadership decision-making.
- Unified VAWA compliance metrics with enrollment data, computing **15+ safety KPIs** with **year-over-year variance analysis** and **rolling averages**.

Publications

Smart Fields: Enhancing Agriculture with Machine Learning [Link] – IEEE AIMLA 2024