VENKATESH SUBRAMANIYAN

458-272-7232 subramav@oregonstate.edu linkedin.com/in/venseven github.com/venseven

WORK EXPERIENCE

Bayer R&D Services LLC **Jan 2025 - Present**
Data Scientist *St Louis, MO*

- Engineered a dynamic hybrid creation algorithm based on user-defined constraints; implemented priority heap sampling and batch processing to handle combinatorial growth. Deployed globally across 21 regions via AWS Lambda to support inbred advancement decisions in field trials.
- Built a semantic layern using CubeJS over commercial seed warehouse databases; developed an AWS Lambda operator to dynamically generate data workers for advancement analytics.

Data Scientist Intern (Remote) **May 2024 - Aug 2024**

- Developed a **Generative AI interface** with LLM, integrated into Bayer's Global NitrogenDB across NA, EMEA, and APAC, to improve customer interactions and automate statistical analysis for identifying top-performing pedigree lines. Built LangChain tools: SQLCodeGen, DBQueryEngine, RecallAgent, PythonCodeGen, PythonDataAnalyst, ReasonerAgent.
- Implemented a **caching mechanism** integrated with RAG for historical chat improvements, utilized AgentExecutor and LangGraph for automated stateful execution and ensured compliance with Bayer's code standards for CodeGen through Masked Language Modelling and connected the reasoning LLM with **LLM Guard**.

Precision Medicine Group **Jul 2023 - Dec 2023**
Data Scientist Intern (Remote) *Corvallis, OR*

- Architectured and developed an enterprise-level question-answering system for biomarker and sample inventory management, utilizing large language models (**LLAMA v2**), **RAG prompt management**, and microservices orchestration on **EC2 with Airflow and Docker**.
- Worked with the business and technical team to design a benchmarking system which utilizes Python for data processing, **Langchain + SQL** for data record generation, and a qualitative verification framework for clinical contexts.

Oregon State University **Aug 2022 – Mar 2024**
Graduate Research Assistant *Corvallis, OR*

- **DARPA's MCS competition**: Led RL algorithm design and development for 16 out of 22 interactive scene types, propelling the team to first place among competitors from MIT and UCB in the MCS competition, under the mentorship of Dr. Alan Paul Fern.
- Achieved 20% performance boost to faster RCNN & UNET for human and soccer ball identification in the DARPA's OpenAI Gym environment.
- Developed computer vision-based reasoning agent for 7 agency scenes; built a 2.5D to 2D homogeneous transformation with C++; YOLOv8 Object Detector with ByteTracker using torch; point cloud mapper using OpenAI Gym, **Ray** and **Shapely**.
- Orchestrated **Airflow DAGs** for Unity-based image data gathering from various 3D scenes, data processing, ML training scheduling and evaluation pipeline.

Saama Technologies, Inc. **May 2020 – Aug 2022**
ML Research engineer I & II *Chennai, India*

- ***COVID-19 Clinical trial optimization with Pfizer:*** Developed a DL solution leveraging **GAN** and **SBERT** to automate clinical data mapping, with dockerized ETL pipelines for data onboarding from **Postgres** server and data transformations with pandas, orchestrated through Airflow.
- Developed an ETL pipeline (PSQL + Python) for transforming a massive volume of IQVIA insurance claims data for ML training.
- ***Preventive care prediction:*** Processed insurance claims data with PSQL and helped the team build a cohort regression model with Graph Convolutional Transformers (**TensorFlow**) to predict Pulmonary Embolism occurrence in Patients with Rheumatoid Arthritis.
- ***Medical coding algorithm:*** Engineered an automated medical coding system using SBERT to convert MeDDRA HLT to LLT, delivering medical coding suggestions for over 10 million verbatim terms.
- Developed an NLP query generator with **PyTorch**, enabling the translation of technical mapping requirements into SQL for efficient data querying.

ML Research intern (Python, PyTorch, R, PostgreSQL/SQL) **May 2019 – May 2020**

- Led **Synthetic data generation** effort with Gaussian Mixture Model, Hidden Markov Model and GAN to preserve temporal dependencies and benchmarked with synthetic databases from multiple clients (Celgene, Pfizer)
- Implemented the synthetic data framework on Celgene and Pfizer's EC2 server and evaluated the reliability on data operations.
- Built a **Named Entity Recognition system** with BiLSTM and CRF(Tensorflow 1.x) to identify and scrap PHI from medical documents.

EDUCATION

Oregon State University **Dec 2024**
Master of Science in Computer Science (GPA: 3.7/4.0) *Corvallis, OR*

Anna University **Mar 2020**
Bachelor of Engineering in Electronics and Communications Engineering *Chennai, India*

PROJECTS

AI Capstone *(Python, PyTorch, Roboflow, YoloV8)* **Sept 2023 - Mar 2024**

- Building a marine tracking for ***National Geographic Research***, utilizing YoloV8 to precisely tag dolphins from aerial drone footage. Built a Byte Tracker utilizing visual cues association and next state prediction with Kalman Filter to track and handle dolphin underwater occlusion.

AVX2 ML acceleration *(C++, Python, AVX, OpenMP, OpenCL, VTune)* **Feb 2024**

- Performed advanced vector optimizations for deep learning, leveraging Python, **C++ intrinsics** with a focus on CPU and GPU enhancements through advanced vectorization register units (**AVX256, AVX512**), OpenMP parallel programming directives, and SIMD and benchmarked the application with **Intel's Vtune** and analyzed the bottlenecks.

GANash - A GAN approach to Steganography *(Python, TensorFlow)* **Jul 2020**

- Trained encoder using **Generative Adversarial Attacks** to improve encoding and prevent data leakage. This technique reduced decoding time by 93% while maintaining the encoding strength. Presented the research findings at a national conference on cybersecurity.

TECHNICAL SKILLS

Languages: Python, C#, C, C++, Embedded C, JS, Next.js, Java, Matlab, Bash, SQL, PostgreSQL, GraphQL.
Development Tools: VS code, Linux/Unix, Vim, Tmux, cron, GitHub, Postman, Spark, GCP(cloudrun, VertexAI), Azure AI, AWS (IAM, S3, EC2, & lambda), MongoDB
Frameworks: TensorFlow, PyTorch, HuggingFace, LangChain, LangGraph, LangSmith, LLamaIndex, REST API (Flask, FastAPI, Django), Airflow, MLFlow, Docker, Singularity, Microservices, Ray, AVX, RoboFlow, Agile, Scrum, JIRA, Tableau, PowerBI, OpenMP, VTune, PySpark, Apache Kafka, Snowflake