KISHORE SAMPATH

(857) 891-4581 sampath.ki@northeastern.edu linkedin.com/in/s-kishore/ kishore-s-15.github.io

Education

Northeastern University **Sep 2023 – Expected May 2025**

Master of Science in Artificial Intelligence *Boston, MA*

Coursework: Machine Learning, Unsupervised Learning & Data Mining, Deep Learning, MLOps, Bias/Ethics Implications in AI, Masters Project.

Gradute Teaching Assistantship: Foundations of Artificial Intelligence

Professional Experience

Graduate Research Assistant — Advisor: Dr. Resmi Ramachandranpillai **Boston, MA**

Khoury College of Computer Sciences, Northeastern University *Mar 2024 - Present*

- Researched and developed a **multimodal, multi-label fair** classifier for clinical decision-making using **EHR** data.
- Built a radiology report pipeline with a custom **Faster R-CNN** and **GPT-2**, improving coherence by **10%** over SOTA.
- Developed a multimodal architecture with **MedBERT** and self-attention, achieving an **F1 score** of **0.93**.
- Designed a **heterogeneous bagging** algorithm mitigating intersectional bias, improving demographic parity by **6%** while preserving model utility. The work is currently submitted and under review in **IJCAI 2025**.

Full Stack Data Scientist **Chennai, India**

Indian Institute of Technology, Madras *May 2021 - Jul 2023*

- Developed and deployed an **end-to-end NLP-based Customer Research Analysis tool** featuring automated sentence tagging, enhancing operational efficiency and scalability on client's **HPC**.
- Designed a **Context-Aware Sentence Tagging module** leveraging **Zero-Shot Text Classification** with the **BART Large MNLI model**, ensuring accurate tagging of sentences based on user-defined keywords.
- Performed **A/B testing** to evaluate and refine tagging strategies, leading to an **18% increase** in user satisfaction scores.
- **Reduced** transcript processing time by **80%** and achieved **annual cost savings** of approximately **$10,000** by automating the manual tagging process.
- Built another **data-driven application** to optimize induction furnace operations using **advanced analytics**, collaborating with key stakeholders to ensure alignment with operational goals.
- Developed an Insights module for **exploratory data analysis**, based on a **Tableau POC** and an **Analytics** module for **training machine learning** models on furnace operational data, resulting in a **66% productivity gain** for the client.

Projects

Curriculum Compass: A RAG Chatbot for Personalized Course Guidance | (Link)

- Engineered a **Hybrid RAG** system utilizing dual retrieval strategies with cross-encoder re-ranking and **multi-agentic workflow** for retrieving course and review data, for assisting Northeastern University students' course registration.
- Supervised fine-tuned **Qwen2.5 0.5B** using **QLoRA** with **knowledge distillation** on 2.5K **synthetic** instructions generated via **Llama3 7B**, achieving notable improvements in n-gram evaluation and LLM-as-a-judge metrics.
- Automated **ETL**, model training and drift detection pipelines with **Apache Airflow** on **GCP** through **Cloud Composer** and **BigQuery**. Implemented LLM tracing with **Weights & Biases** and data/model registry on **Huggingface Hub**.

Pencil2Pixel: Gamma-Corrected GANs for Refined Forensic Sketch Generation | (Link)

- Designed a **Conditional Pix2Pix GAN** with **grid search** for forensic sketch enhancement resulting in a 31% increase in SSIM and a 35% boost in PSNR.
- Evaluated the performance of **Gamma-Corrected GANs** using a **ResNet-50** classifier on a custom celebrity dataset, resulting in a **10% increase in classification accuracy** with **Gamma-inverted** sketches.

ML Olympiad - Autism Prediction Challenge | (Link)

- Developed a high-performing **two-level stacking ensemble classifier** combining **XGBoost**, **LightGBM**, **CatBoost**, and **Random Forest** classifiers as base estimators with a final **Logistic Regression** model.
- Resolved the **unbalanced data problem** using **SMOTE** and trained the model with **stratified k-fold cross-validation**, optimizing hyperparameters via a **Bayesian tuning framework** to achieve a test set **AUC-ROC score of 0.943**.

Technical Skills

Languages & Databases: Python, C, C++, Java, JavaScript, SQL, NoSQL, R, ChromaDB
ML Frameworks: PyTorch, Scikit-Learn, Tensorflow, OpenCV, NLTK, Huggingface Transformers, LangChain, PowerBI
MLOps Frameworks : AWS, GCP, Vertex AI, Amazon Sagemaker, MLFlow, Wandb, Databricks, Apache Airflow, Docker

Publications and Contributions

- The Multimodal Paradox: How Added and Missing Modalities Shape Bias and Performance in Multimodal AI, **CVPR, Responsible GenAI Workshop 2025.**
- Fairness at Every Intersection: Uncovering and Mitigating Intersectional Biases in Multimodal Clinical Predictions. (Arxiv)
- Contributed to the **Deepchecks** (PR links) and **Hugging Face Transformers** (PR links) **open source** libraries.