KUMAR SELVAKUMARAN

857-396-6078 | Boston, MA | kumar.selvak.27@gmail.com | kumar-selva | kumar-selvakumaran | Website

EDUCATION

Northeastern University 09/2023 - 12/2025
Master of Science in Artificial Intelligence - GPA: 3.76/4.0 Boston, MA
Relevant coursework: Verifiable Machine Learning, Pattern Recognition and Computer Vision, Unsupervised Machine Learning

TECHNICAL SKILLS AND FRAMEWORKS

Programming Languages:	Python, C++, C, SQL
Frameworks and Databases:	PyTorch, TensorFlow, HuggingFace, LangChain, Ray, NVIDIA Rapids, NVIDIA DALI, FastAPI, Git Apache Spark, Airflow, MLflow, Docker, Uvicorn, PostgreSQL, Databricks, DataRobot, Qdrant
Operating Systems and Cloud:	Linux, WSL2, GCP: Cloud Run, Kubernetes Engine, AWS: SageMaker, Lambda, EC2, App Runner

PROFESSIONAL EXPERIENCE

Generative AI Intern 01/2025 - 04/2025
Norfolk Southern Atlanta, Georgia
Developed an agentic RAG solution and engineered GPU-accelerated, high-throughput computer vision and Big Data pipelines.

- Implemented a **multimodal RAG** workflow on the **Qdrant vector database** for large-scale PDF-based question answering.
- Developed a large-scale PDF retrieval pipeline using the **ColPali** model, improving top-5 recall by **8%** over OCR solutions.
- Built a **multi-agent** system that divides the task into sub-tasks, leverages a document retriever, a metadata crawler, and feedback loops to **maximize cited content** and diagram inclusions in the generated answer.
- Implemented a **system prompt refinement** module using **Mutual Information Maximization** on LLM-generated candidates which demonstrated an average reduction of 6.2 steps for task completion, and a **14%** increase in task completion rate.
- Accelerated data manipulation with **NVIDIA RAPIDS** and **Spark**, saving **23%** in cost and reducing total job time by **30%**.
- Migrated the data loader pipeline to **NVIDIA DALI** and **Ray**, boosting throughput of large-scale model evaluation by **2.1** times.

Artificial Intelligence Intern 05/2024 - 08/2024
Inflohealth Atlanta, Georgia
Built a language model pipeline that processes millions of radiology reports to curate datasets for downstream healthcare analytics.

- Fine-tuned the **CXR-BERT** biomedical transformer language model on 830,000 radiology reports across four T4 GPUs using **Distributed Data Parallel** and adapted it for **knowledge-graph creation** using the RadGraph-XL dataset.
- Built an **explainability** module that highlights text spans relevant to specific nodes by visualizing BERT's attention maps.
- Achieved a retrieval F1 score of **0.81**, on par with **LLaMA 3 8B**, while additionally supporting **evidence selection**.
- Containerized the retrieval system as a **serverless FastAPI** application and deployed it on **AWS App Runner**.
- Developed an **XGBoost** classification pipeline on radiology claims (such as encounter, diagnosis, and procedure codes) to predict follow-up likelihood, achieving **0.72 AUC** and improving recall of missed visits by **18%**.

Artificial Intelligence Intern 03/2023 - 07/2023
Sentient.io Singapore (remote)
Customized open-source computer vision models with application-specific optimizations for AI **microservices**.

- Implemented a **video action-recognition** pipeline, **quantizing** the model with **TensorRT** to reduce latency by **12%**.
- Reduced video object detection misclassification rate by **7%** using a **Kalman filter** motion model.
- Built a general-purpose auto-labeler using **SAM** and **GroundingDINO**, which was applied to three datasets (15,000+ images).

Artificial Intelligence Engineer Intern 04/2021 - 09/2022
Juhomi Chennai, India
Co-developed an AI-powered, microservice-based retail analytics platform from the ground up; it was contracted by four MNCs.

- Managed a crowdsourced data annotation job for **object detection** of 252 product classes across 4750 images in Amazon Mechanical Turk, and developed a **YOLOv5**-based auto-labeler in **Amazon SageMaker** to extend the dataset to 21,000 images.
- Integrated **image super-resolution** and optical character recognition (**OCR**) improving mAP by **0.2** over naive object detection.
- Demonstrated the combined detection pipeline's ability to perform **zero-shot** object detection of **previously unseen classes.**
- Upgraded the API architecture to support asynchronous communication using **FastAPI** with **Uvicorn**, enabling an API throughput increase of **341%** (i.e., from 12 requests per second to 53 requests per second).
- Developed **continuous training and monitoring** workflows using **Apache Airflow**, **Data Version Control**, and **MLflow**, enabling class-specific model updates through model-guided data selection (**active learning**).

PROJECTS

HuskyGuide: Database Interface Agent (sponsored) | Project: *Link* 07/2025

- Spearheaded the development of a **multi-agent** system leveraging Northeastern University's knowledge base for policy navigation.
- Developed a **SQL agent** that generates complex queries through **multi-hop reasoning** using execution plans and database contents.
- Implemented it with **LangChain** and packaged it as a **FastAPI** app that supports integration into concurrent multi-agent systems.

RobAnn: Exploring Neural Network Robustness | Project: *Link* 10/2024

- Developed an algorithm to quantify deep neural networks' resistance to **adversarial attacks** and noisy perturbations.
- Performed **feature-wise robustness analysis** to expose adversarially susceptible neurons by targeted weight perturbation.
- Illustrated the algorithm's behavior and efficacy through comprehensive visualizations.

ProdSeek: Semantic Product Catalog Search | Project: *Link* 05/2024

- Developed a recommender that takes product selections from images and suggests similar products from a product image corpus.
- Built an adaptive **vector search** using **YOLOv3**; embeds live product selections and performs real-time semantic retrieval.
- Conducted ablation studies demonstrating higher product specificity and qualitative semantic capacity than **ResNet** embeddings.

PUBLICATIONS

Transformers for Browse Node Classification with Class Imbalance | CISES 2023: *Link* 04/2023

- Fine-tuned the **DeBERTa** transformer model for e-commerce classification on 10M+ records across 250 browse node categories.
- Applied **Focal Loss** to mitigate severe class imbalance by down-weighting the gradient updates of dominant classes over time.
- Achieved an increase of **2%** in validation accuracy with faster convergence compared to vanilla DeBERTa and other BERT variants.

Safety surveillance using Explainable Object Detection | SmartCOM 2023: *Link* 06/2023

- Built an automatic object detection explainer that visualizes the top three salient activation regions influencing predictions.
- Implemented a novel **model-agnostic pipeline** that automatically finds salient layers to bypass manual exploration.
- Demonstrated the pipeline on an artificially biased dataset simulating effects of irresponsible data collection practices.
- Augmented the object detection pipeline with **Sobel features** to improve **generalizability** and **reduce bias**.

AR-enabled textbooks | ICESC 2023: *Link* 07/2023

- Built an Augmented Reality mobile application that scans QR codes embedded in textbooks to render animated 3D models.
- Secured TNSCST funding to deploy the AR pipeline in 9th-grade textbooks for the Tamil Nadu state curriculum.

OPEN SOURCE CONTRIBUTION

- Contributed to **PyTorch/xla** by suggesting a fix in the PyTorch's environment configuration for Google Colab: Link