Koushik Sivarama Krishnan

+1 (236) 882-7140 koushik.nov01@gmail.com Canada

LinkedIn Github Portfolio

Highlights

- Master's graduate in Applied Data Science at the University of Victoria with 3+ years of strong foundation in **Machine Learning** (ML) with expertise in MLOps pipelines, CI/CD, and model lifecycle management.
- Strong technical foundation in **end-to-end Machine Learning workflows**, complemented by proven cross-functional collaboration skills, including guiding non-technical stakeholders, product managers, and teams to implement and deploy AI features aligned with business objectives effectively.
- Adept in designing robust solutions across both structured and unstructured data sources, leveraging Deep Learning, Natural Language Processing (NLP), and **Retrieval Augmented Generation** (RAG) techniques to drive actionable outcomes and improve system intelligence in real-world applications.

Skills

Programming Languages	Python, SQL, Bash, C++, CUDA
Machine Learning	PyTorch, TensorFlow, Scikit-learn, XGBoost, OpenCV, MLflow, Weights & Biases, Data Version Control (DVC)
Natural Language Processing	LangChain, LangGraph, CrewAI, Prompt Engineering, OpenAI API, Retrieval Augmented Generation (RAG), Agentic AI, Model Context Protocol (MCP), HuggingFace
API Development	Flask, FastAPI, REST APIs, ML Model Serving, PostgreSQL
Cloud Computing	Amazon Web Services (AWS), Google Cloud Platform (GCP), Docker, Kubernetes, GPU Clusters, Distributed Computing

Experience

Co-op Data Scientist — *09/2024 - 08/2025*

Insurance Corporation of British Columbia (ICBC) — *Vancouver, BC*

- Led development and deployment of ICBC's in-house agentic AI assistant using Retrieval Augmented Generation (RAG) with LangChain, Milvus, and MCP servers—reducing knowledge retrieval time by ∼**60%** for **100+ daily users** via Azure Kubernetes Service.
- Designed and implemented systematic evaluation frameworks to **benchmark LLM performance** and robustness, using metrics like accuracy, faithfulness, response consistency, and bias, significantly improving model reliability in production environments.
- Built a CatBoost **risk-prioritization model** on structured case features; implemented class-imbalance handling, k-fold CV, and probability calibration, with **SHAP reports** for case-level explanations consumed by the CMA team.

Research Engine Developer — *05/2024 - 02/2025*

Justice Data and Design Lab — Victoria, BC

- Collaborated with MITACS to leverage data analysis and machine learning, aiming to enhance access to justice through evidence-based opportunities.
- **Doubled** the performance and efficiency of the research engine chatbot using **Retrieval Augmented Generation (RAG)** to better identify and address unmet legal needs.
- Conducted advanced data analysis on large-scale legal text datasets from Reddit and People's Law School using **Topic Modeling** techniques and **Power BI** visualizations to extract key insights and refine legal service strategies.

Founding Machine Learning Engineer — *04/2023 - 07/2023*

SeiSei.ai — *India*

- Fine-tuned FreeVC voice conversion model with 15% performance improvement on Hindi audio through advanced hyperparameter optimization and custom audio preprocessing pipelines.
- Architected automated audio processing systems handling 5000+ hours of diverse audio content, implementing feature extraction, audio quality assessment, and batch processing workflows.
- **Directed and supervised** a team of interns in the deployment of machine learning models using Truefoundry, enhancing project delivery and meeting key performance targets.

Computer Vision Intern *09/2021 - 02/2022*

Drive Analytics *India*

- Engineered a Major League Baseball video analytics system and deployed it on **AWS using Celery, RabbitMQ, and Django**.
- Led a team in implementing image-to-3D model rendering with advanced vision techniques, **boosting profits by 10%** through innovation and teamwork.
- Orchestrated comparative analysis of various **object detection** models for glove and baseball detection.

Deep Learning Intern *12/2020 - 06/2021*

MURF.ai India

- Developed and deployed a voice cloning application using the FastSpeech-2 algorithm, implementing advanced audio processing pipelines with librosa and custom preprocessing techniques for high-quality speech synthesis.
- Fine-tuned Speech-to-Text models using Azure Speech Studio with focus on acoustic feature extraction and audio signal processing for improved transcription accuracy.
- Built end-to-end audio processing pipelines handling diverse audio formats, implementing STFT, mel-spectrograms, and other audio representations for model training.

Education

Master of Engineering in Applied Data Science, *University of Victoria* **Victoria, BC** *2023-2025*

Notable Coursework: Optimization for Machine Learning, Music Information Retrieval, Data Privacy

Bachelor of Engineering in Computer Science and Engineering, *Anna University* **India** *2019-2023*

Notable Coursework: Artificial Intelligence, Cloud Computing, Data Structures, Data Warehousing, and Data Mining

Projects

GenZ AI Therapist *06/2025*

- Built a non-clinical GenZ-friendly mental wellness AI Assistant with **advanced prompt engineering techniques** to provide a safe, empathetic space for users to vent and reflect—offering sentiment/intent analysis, emotional trend tracking, and real-time resource suggestions without giving clinical advice.
- Implemented **Multi-Agent architecture** using CrewAI, combining custom LLM agents for session summarization, mood visualization (Plotly), guardrails, and Google-integrated resource search (Serper.dev), all deployed through an interactive Streamlit interface.
- Designed **guardrails agents** and session filters to detect crisis signals and moderate unsafe or irrelevant content, ensuring respectful and focused conversations aligned with mental wellness goals.

ClotSense — Stroke Blood Clot Origin Identification *04/2023*

- Built an automated pathology AI to classify the **origin of ischemic stroke clots** (Cardioembolic vs. Large Artery Atherosclerosis) from whole-slide digital pathology images (WSIs) supporting faster, targeted secondary prevention decisions in the care pathway.
- Designed a novel **two-stage pipeline**: a MobileNetV3 background/tissue detector to tile WSIs, followed by fine-tuned classifiers aggregated to slide-level predictions; released pretrained weights and a Dockerized web app for reproducible local validation.
- Drove the model to **F1 = 96.3** on a held-out data over the best PoolFormer baseline and **published a research work on it**. [arXiv]

Publications

ORCID ID:0000-0003-0525-0677 | **Google Scholar**

- "Vision transformer based COVID-19 detection using chest X-rays"
- "SwiftSRGAN–Rethinking Super-Resolution for Efficient and Real-time Inference"
- "Benchmarking Conventional Vision Models on Neuromorphic Fall Detection and Action Recognition Dataset"
- "Efficient Super-Resolution For Chest X-rays"
- "MFAAN: Unveiling Audio Deepfakes with a Multi-Feature Authenticity Network"

Peer-reviewed research spans medical imaging for disease detection, real-time model optimization, and multi-modal healthcare data processing.