YASH KHURANA

+1(872) 218-7570 | yashkhurana2024@u.northwestern.edu | Github | Linkedin

EDUCATION

Masters in Artificial Intelligence, Northwestern University, Evanston, Illinois, USA *Sept 2023 – Dec 2024*
Coursework: ML, Deep Learning, Data Science, Pattern Recognition, NLP, Reinforcement Learning GPA: 3.75/4.0

Bachelor of Technology, Computer Science, Vellore Institute of Technology, Vellore, India *July 2019 – June 2023*
Coursework: Machine learning, Artificial Intelligence, NLP, Image processing GPA: 8.79/10

PROFESSIONAL EXPERIENCE

AI Research Engineer, Arch Systems, Remote *Nov 2025 – Present*

- Developed existing HDC architectures for efficient edge object detection across multimodal inputs, including thermal, vision, and sensor data.

AI Systems and Research Engineer, Intel Labs, OR *Jan 2025 – Oct 2025*

- Primary engineer on a **federally funded research initiative, "Ultra-Efficient In-Hardware Learning Using HDC,"** in collaboration with **US Air Force Research Labs (AFRL)**, consistently delivering milestones ahead of schedule.
- Developing an **unsupervised online learning pipeline** for edge deployment, designing an **adaptive encoder** based on Vector Symbolic Architecture (VSA) to convert YOLOv10 features into **bipolar representations** for efficient **prototype learning**, **continual adaptation** and **lifelong learning** on FPGA and ASIC (sub-0.5MB).
- Established automated benchmarking tools — aligned with **MLPerf BKM standards** (best known methods) — to evaluate classification and detection accuracy, inference latency, and power consumption across CPU, GPU, NPU, and FPGA backends.

Graduate AI Research Intern, Intel Labs, Remote *June 2024 – Aug 2024*

- Optimized **YOLOv10** architecture for edge inference through custom **Structural Pruning** and **Quantization Aware Training (QAT)**, achieving **90% inference time reduction** on CPU and **5x compression rate in file size** with **<5% loss in mAP@0.50**
- Implemented a **lightweight Hyperdimensional Computing Model** to replace the classification head, achieving **20% additional file size reduction** while significantly **enhancing noise tolerance** and resilience against various types of adversarial attacks.

RESEARCH EXPERIENCE

Research Assistant, Radiation Oncology Lab, Feinberg School of Medicine, Chicago, IL *March 2024 – May 2024*

- Developed a chatbot-based application using **RAG framework** to provide a centralized Q/A platform for patients FAQs with a mechanism to generate **citations** from original documents to enhance factual correctness and an automated pipeline for data updates.
- Created an LLM-powered study guide for medical students using RadOncTables, **finetuning Llama3** with **QLoRA**, **TRL**, **PEFT**.
- Implemented **automatic metrics** for assessing LLM output quality, and demonstrated their strong correlation with human judgements.

Graduate Research Assistant, IVP Lab, Northwestern University, Evanston, IL *Oct 2023 – April 2024*

- Developed a robust self-supervised method for **Denoising X-Ray Fluorescence (XRF) Data** from Cultural Heritage Paintings, using pixel-downsampling and a re-visible loss to overcome information loss in blindspot-driven denoising methods.
- Developed a **Reinforcement Learning (RL) agent to optimize the scanning path of X-ray fluorescence (XRF) scanners**, aiming to reduce overall data capture time by employing a **Q-Learning** approach, akin to game models, to innovate XRF mapping.

RESEARCH PUBLICATIONS

- RoBInNet: A Multimodal Speech Emotion and Speaker Recognition System *(IEEE Transactions on Computational Social Systems)*
- Marine Debris Detection using Multi-feature Pyramid Network *(Multi-Feature Pyramid Network, Resnet, Normalization)*
- ULCLU classification of multi-temporal Imagery of Delhi-NCR Region *(RBF Kernel, Maximum Likelihood Classification)*

PERSONAL PROJECTS

DocEase (Founder & Lead Developer) Went from zero web dev experience to shipping a full-stack AI document formatting platform leveraging *vibe-coding* (inspired by Karpathy). Built the entire product, from a Chrome Extension to the web app, growing it to 1,000+ downloads and attracting a community of 500+ active users.

[Patented] A Machine Learning based Intelligent Model for an Autonomous Vehicle System for Data Selection on board: Implements ML based systems for autonomous vehicle driving to pick onboard data.

VisionForBlind: Developed an app for edge devices to assist visually impaired individuals. Supports various **V-LLMs** for real-time scene understanding, integrated **bidirectional speech-text conversion**, and designed adaptive UIs for low-latency, power-efficient wearable use.

SKILLS

Languages and Platforms: Python, C++, C, SQL, Bash, R, CUDA, Linux, GCP, AWS, Azure, SageMaker, ECS, EKS, S3, EC2
Frameworks: Scikit Learn, TensorFlow, Pytorch, Langchain, OpenCV, Codex, LlamaIndex, TensorRT, DeepStream, Claude Code
Domains: Object Detection, GenAI, Data Analytics, 3D Reconstruction, GPT, ChatGPT, Midjourney, DevOps, MLOps, Generative Modelling, Retrieval Augmented Generation (RAG), Product Sense, A/B Testing, LLMOps, Time Series Analysis, Predictive Modelling
Algorithms: Decision Trees, GLMs, Random Forests, Neural Networks, KNN, Kmeans, DBScan, Regression, ARIMA, Transformers