Arsh Verma

Phone: +1-(412)-909-7601 **E-mail:** arsh@cmu.edu **LinkedIn:** https://www.linkedin.com/in/arshverma **Google Scholar:** Google Scholar

Machine Learning Engineer with 3+ years of industry experience owning end-to-end, production-grade ML and GenAI systems, from model development to real-world deployment. Skilled at aligning technical decisions with product goals for real-world impact.

Work Experience

Associate Machine Learning Scientist - 2 at Wadhwani AI — Aug 2021 – Aug 2024

- Deployed a production-grade deep learning system for detecting 17 chest X-ray abnormalities, improving mean AUROC to 90% ($\sim$3% improvement over SotA) and serving 2M+ patients annually on India's national digital diagnosis platform.
- Built and deployed a CXR-based TB screening model for the National Health Mission, aiding diagnosis in over 3M presumptive TB cases per year; achieved 90% sensitivity with 80% specificity on low-quality CXR photographs, exceeding WHO screening tool benchmarks.
- Enabled smartphone-based inference of CXR photos through 90% model compression via pruning and quantization.
- Led deployment readiness in collaboration with radiologists and public health officials, aligning model behavior with clinical workflows.
- Earned a fast-tracked promotion for end-to-end project ownership, technical leadership, and mentoring 7 junior engineers.

Education

M.S. in Machine Learning and Robotics @ Carnegie Mellon University — CGPA: 4.25/4.00 — Aug 2024 – Apr 2026

Research: Learning-based planning and decision-making for real-time robotic systems.
Coursework: Generative AI, DL Systems and Algorithms, Multi-Modal ML, Advanced Computer Vision, Reinforcement Learning.

B.Tech. in Computer Science @ IIIT Delhi — CGPA: 9.04/10 — Aug 2017 – Jun 2021

Coursework: Deep Learning, Advanced Machine Learning, Natural Language Processing, Probability & Statistics, Linear Algebra.

Research

Planning for Real-time Robot Decision Making in Unstructured Environments — Aug 2024 – Apr 2026

Advisor: Dr. Jeff Schneider (Research Professor at Carnegie Mellon University (CMU))

- Designed waypoint and path-integral planning strategies for efficient multi-agent target search under real-time constraints.
- Achieved 97% reduction in on-robot inference time by training GNN-based policies using behavior cloning and reinforcement learning.
- Analyzed generalization across maps (shapes, sizes, orientations, target counts/densities).
- Deployed and evaluated policies on real robots, achieving 55% reduction in search time over baselines.

Publications

Efficient Active Search via Amortized Path-Integral Policies **Verma A.**[*], Gupta T.[*], Schneider J. at **ICRA 2026**

Generalized Cross-domain Multi-label Few-shot Learning for Chest X-rays Aimen A., **Verma A.**, Tapaswi M., Krishnan N. C., at **ISBI 2025**

Using LLMs in Software Requirements Specifications: An Empirical Evaluation Krishna M., Gaur B., **Verma A.**, Jalote P. at **RE Conference, 2024**

Towards long-tailed, multi-label disease classification from chest X-ray Holste, G. et al, in In **Medical Image Analysis, 2024**

How Can We Tame the Long-Tail of Chest X-ray Datasets? **Verma A.**, at **ICCV-W: CVAMD, 2023**

Can we Adopt Self-supervised Pretraining for Chest X-Rays? **Verma A.** and Tapaswi, M., at **ML4H 2022**

Projects

- **RoPE + GQA for Transformers**: Implemented Rotary Positional Embeddings and Grouped-Query Attention in a GPT-style Transformer, with Key-Value caching and causal masking. Benchmarked attention latency across sequence lengths and head configurations, reducing validation loss by $>$10% over a minGPT baseline and improving throughput/memory behavior.
- **LoRA for Efficient LLM Fine-Tuning**: Implemented LoRA for GPT-2 by injecting trainable low-rank adapters into attention layers while freezing base model weights. Achieved $\sim$2% higher accuracy than full fine-tuning while updating $<$5% of parameters, benchmarking convergence and performance across multiple LoRA ranks.
- **Needle: A Minimal Deep Learning Framework (Mini-PyTorch)**: Built a PyTorch-like deep learning framework from scratch, implementing dynamic computation graphs and reverse-mode automatic differentiation. Supported core neural architectures (CNNs, RNNs, Transformers), optimizers, and GPU acceleration for end-to-end training.
- **Agentic System for Multi-Step Decision Support**: Built an LLM agent that decomposes user requests into sequential steps and invokes external tools via structured function calls. Implemented execution control, state tracking, and error recovery to improve reliability in multi-step workflows.
- **Diffusion Models (DDPM)**: Built a DDPM with cosine noise scheduling and a U-Net–based noise predictor, implementing both forward noising and reverse denoising processes. Trained and sampled images via iterative reverse diffusion and achieving FID $<$160 after $\sim$2 hours of training on a T4 GPU.

Interests & Skills

Research Interests: Deep Learning for Computer Vision, Language Modelling, Reinforcement Learning, Multi-modal Learning, AI for Healthcare.
Skills and tools: Python, PyTorch, Generative AI, Large Language Models (LLMs), Distributed Training, Parallelization, Docker, TorchServe, MongoDB.

Positions of Responsibility

Reviewer, Machine Learning for Health Conference [2023, 2024]
Head Teaching Assistant for the graduate-level ML course. Managed a team of 7 TAs for the course taken by 150 UG, Masters and PhDs. [2020, 2021]
Member, Chess Club, Math Club and Poker Club @ IIITD and CMU [2017-2025]