

# Sri Akash Kadali

8417 48th Ave, College Park, MD, 20740

Availability: June 1st, 2026

240-726-9356 | [kadali18@umd.edu](mailto:kadali18@umd.edu) | <https://www.linkedin.com/in/sri-akash-kadali/> | <https://github.com/Akash-Kadali>

## EDUCATION

### University of Maryland, College Park, United States

CGPA: 3.55/4

Master of Science in Applied Machine Learning

August 2024 - May 2026

- **Relevant Coursework:** Machine Learning, Deep Learning, Software Engineering

### Indian Institute of Information Technology, Vadodara, India

CGPA: 8.78/10

Bachelor of Technology in Computer Science and Engineering

December 2020- June 2024

- **Relevant Coursework:** Performance Analysis, Algorithms, GPU Programming

## SKILLS

<b>Programming:</b>	Python
<b>Deep Learning Techniques:</b>	Deep Learning, PyTorch, TensorRT, Generative AI, Tensor Parallelism, Machine Learning Frameworks
<b>Deployment Tools:</b>	CUDA, Performance Analysis, Algorithms, NVIDIA SDKs, HuggingFace, TRT-LLM, TRT Model Optimizer, GPU, Inference, Automated Deployment Solutions, Model Sharding
<b>Software Development Skills:</b>	Sequence Parallelism, Efficient Attention Kernels, KV-Caching, Problem Solving, Debugging, Software Design, Data Structures, CUTLASS, Triton, GPU Architecture, End-to-End Performance, Model Optimization, English (professional)

## EXPERIENCE

### Machine Learning Intern

May 2023 – December 2023

Indian Institute of Technology, Indore

Indore, India

- Leveraged supervised contrastive learning to enhance feature representation, resulting in an 8% increase in classification accuracy on the IHSate and IHC datasets, optimizing model performance.
- Designed and implemented a DeBERTa-based architecture using attention mechanisms for implicit hate speech detection, achieving a 5% improvement in F1-score, enhancing inference efficiency.
- Developed emotion synthesis pipelines incorporating sentiment features, contributing to a 6% boost in model precision, aligning with generative AI model optimization.

### Machine Learning Intern

January 2024 – June 2024

National Institute of Technology, Jaipur

Jaipur, India

- Engineered Cascaded Deformable Transformer Layers (CDTL) to optimize AI workflows, improving feature dependency modeling by 20% for enhanced inference efficiency.
- Developed classification pipelines for breast tumor analysis, achieving a 15% reduction in misclassification rates, supporting accurate clinical decision-making in medical imaging.
- Designed and deployed MaxViT-based models for histopathological image classification, achieving a 92% classification accuracy on large-scale medical datasets, enhancing model performance.

### Machine Learning Intern

July 2024 – December 2024

Indian Institute of Technology, Indore

Remote, USA

- Engineered a graph-based framework for user-level feature modeling, enhancing contextual embeddings and improving model interpretability.
- Implemented automated inference solutions using TensorRT and TRT-LLM, optimizing deployment efficiency and reducing latency by 30%.
- Collaborated on model optimization strategies, achieving a 10% increase in inference speed across various generative AI applications, including diffusion models.

### Machine Learning Engineer

May 2025 – August 2025

Ayar Labs

Santa Clara, CA

- Developed and deployed high-performance inference solutions using YOLOv8 and Transformer ensembles, achieving 99% accuracy and 96% recall on critical defect classification tasks.
- Engineered a serverless GPU deployment with FastAPI endpoints for automated inference, optimizing cold-start and concurrency for enhanced model performance.
- Analyzed and profiled model performance, implementing techniques like focal loss and class weighting to improve minority class precision and recall in imbalanced datasets, contributing to automated deployment solutions.

## ACHIEVEMENTS AND LEADERSHIP

Published "CaDT-Net: Cascaded Deformable Transformer for Breast Cancer" at ICONIP 2024, achieving 92% accuracy in image classification using **Neural Networks**.