# Lakshmi Kavya Kalyanam

**ML Engineer | Model Optimization | Neural Networks | Inference Optimization**
linkedin.com/in/lakshmikavya-kalyanam-a88633131 | github.com/kavyakl | kavyakalyanamk@gmail.com

## Summary

• Machine Learning Engineer with 5+ years of experience in model compression, pruning, quantization, and embedded AI deployment.

• Author of 9 peer-reviewed publications and 3 patents in hardware-aware and compiler-integrated ML optimization.

• Skilled in neural network sparsification, ONNX graph transformations, and deployment across microcontrollers, GPUs, and FPGAs.

• Experienced in analyzing compute–memory trade-offs and building optimization pipelines for efficient real-time, on-device inference.

• Passionate about making ML systems smaller, faster, and reliable for production edge and cloud platforms.

**Languages:** C++17, Python, Embedded C, Shell
**GPU & Hardware:** CUDA 12.3, OpenCL, GPU Memory Management, NVIDIA GPUs
**ML & Optimization:** Structured/Unstructured Pruning, Quantization, Sparse Training, ONNX Graph Rewrites, Model Compression
**Frameworks & Tools:** PyTorch 2.6.0+cu124, TensorFlow, ONNX Runtime, TensorRT, TFLite
**Embedded & Edge Systems:** PYNQ-Z1 AP-SoC, FPGA (Virtex-7), Arduino MKR1000
**Math & Compiler Tools:** Linear Algebra, Fast Math Libraries, CSR/COO Formats, Algorithm Optimization, MLIR-inspired IR Transformations
**Inference & Runtime:** On-device inference, latency/memory profiling, performance benchmarking
**Development Tools:** Git, Linux CLI, Vivado, pytest, Jupyter
**Cloud & ML Workflows:** Docker, containerized ML workflows, cloud-based GPU inference (AWS/GCP)

## Research Experience

**Lead Research Engineer & PhD Fellow — University of South Florida, Jan 2019 – Dec 2025**

- Developed multi-stage compression and pruning pipelines for CNNs using PyTorch and ONNX Runtime, achieving up to 98% sparsity and 60% model compression with minimal accuracy loss.
- Designed activation-aware pruning for MLPs, achieving 82% sparsity and 37% hardware savings (<1.5% accuracy drop) and deployed on ARM Cortex-M4 microcontrollers with a 95% success rate.
- Built a distributed real-time object detection system using PYNQ-Z1 AP-SoCs and BNNs, achieving 19.23 FPS across a 3-node edge network with minimal communication latency.
- Developed ONNX graph rewrites simulating compiler IR transformations, enabling 4.33× inference speedup through structure-preserving sparsity optimizations.
- Engineered a RAG-based literature analysis system using local LLM inference and FAISS-based vector search, reducing manual review time by 80% across 200+ scientific papers.
- Modeled compute–memory trade-offs and implemented optimization techniques for real-time embedded inference and performance scalability, using Python scripting, debugging ML workflows, performance profiling.
- Ran optimized models on cloud-based GPU platforms using ONNX Runtime and TensorRT.

## Education

**Ph.D. in Computer Science and Engineering** *University of South Florida* (December 2025)

- Focus: Sparse model optimization, embedded ML, compiler-aware inference
- 9 peer-reviewed publications, 3 patents filed, 2 Best Paper Awards

**M.S., Computer Science and Engineering** *University of South Florida* (2020)

- Thesis: Real-time object detection using BNNs on PYNQ-Z1 (19.23 FPS) - IEEE iSES 2020 Best Paper
- Focus: Embedded Deep Learning, Edge Inference Systems, Distributed Computing

**B.Tech., Electronics & Communication Engineering** — *GITAM University* (2017)

## Patents

- "Layer-Wise Filter Thresholding Based CNN Pruning for Efficient IoT Edge Implementations" Inventors: Srinivas Katkoori, Lakshmi Kavya Kalyanam. Application No.: 63/552,084. Filed: 2024-02-09. USF Ref: 24T085US. Provisional patent for a thresholding-based CNN pruning method for efficient IoT edge deployment.

- "Unstructured Pruning for Multi-Layer Perceptrons with Tanh Activation" Inventors: Srinivas Katkoori, Lakshmi Kavya Kalyanam. Invention ID: USF23/00331. Tech ID: 24T063. Patent for unstructured pruning techniques for MLPs with tanh activation.
- "Range-Based Hardware Optimization of Multi-Layer Perceptrons with ReLUs" Inventors: Srinivas Katkoori, Lakshmi Kavya Kalyanam. USF Ref: 23T078US. Q&B Ref: 173738.02709. Patent for range-based hardware optimization of MLPs with ReLU activation.

### Awards & Honors

- Best Paper Award, IEEE iSES 2023 — Unstructured Pruning for Multi-Layer Perceptrons with Tanh Activation
- Best Paper Award, IEEE iSES 2020 — Distributed Real-Time Object Detection with IoT Edge Nodes
- Judge, USF Virtual Graduate Research Symposium — 2022, 2023