SHAMIT SAVANT | ML Engineer

+1 (352) 709-6370 | savantshamit@gmail.com | Gainesville, FL | LinkedIn | Github | Portfolio

EDUCATION

University of Florida | *Master's in Electrical and Computer Engineering* **GPA: 4.00/4.00 | May 2026**
Coursework: Deep Learning in Medical Image Analysis, Database Management Systems, Computational Photography

Veermata Jijabai Technological Institute | *Bachelor's in Electrical Engineering* **GPA: 8.99/10.0 | May 2022**
Coursework: Fundamentals of Machine Learning, Python Programming, Applied Linear Algebra, Probability and Statistics

EXPERIENCE

Machine Learning Engineer Co-op | UF Intelligent Clinical Care Center (IC3) **Jan 2026 – present**

- Engineered a production image registration and feature extraction system processing **200M+ pixel images** across an 8-slide cohort, extracting features across **820K+ cells per slide** on HiPerGator HPC via affine registration, memory-aware coordinate transforms, and optimized rasterization (manuscript in preparation, UF Medicine 2026)
- Reduced pipeline peak memory by **~70%** on previously failing large-scale workloads by diagnosing three concurrent OOM failure modes and re-architecting data flow with adaptive downsampling, selective morphological operations, and backward-compatible transform serialization
- Recovered **1.6M annotations** and **3+ hours** of compute after a production job failure with zero reprocessing, by designing a checkpoint-and-replay mechanism that reconstructed the full processing pipeline from intermediate outputs already persisted to the server

Research Assistant — Machine Learning | CMIL, UF Medicine **Jan 2025 – Jan 2026**

- Unblocked deep learning pipeline operations across **3 partner sites** by diagnosing a GPU driver incompatibility introduced by a cluster upgrade, directly enabling the lab's full framework migration to the current GPU stack
- Reduced team data operations overhead by an estimated **4+ hours/week** by developing and open-sourcing a Python automation toolkit for bidirectional HPC-to-cloud transfer, annotation pipeline management, and cloud sync, adopted as standard lab tooling on GitHub

Technology consultant | PricewaterhouseCoopers | Mumbai, India **Jul 2022 – Aug 2024**

- Reduced manual document processing time by **~75%** for a client handling 12,000+ compliance documents monthly by engineering a GPT-4 extraction pipeline with domain-specific prompt engineering and custom chunking, achieving **91% field-extraction** accuracy on a held-out test set
- Reduced deployment time by **50%** and maintained p95 latency under **200ms** through 200% peak traffic surges by migrating client services to containerized microservices on AWS ECS with ALB and auto-scaling, sustaining **99.9% uptime** across production deployments

PROJECTS AND RESEARCH

Prostate MRI Through-Plane Resolution Enhancement via Deep Learning - arXiv preprint, 2026 | [Link]

- Benchmarked 5 deep learning architectures (CNN, U-Net, GAN, Attention-GAN, Diffusion) on 46K prostate MRI slices (TCIA, 58 patients), finding problem formulation had greater impact than architecture choice, with best model achieving 30.08 dB PSNR and 0.898 SSIM, 10.1% above linear interpolation baseline

Multimodal Vision-Language Model| [Link]

- Improved image-text alignment by ~18% over a CLIP-ViT-B/32 baseline on a held-out validation set by designing a dual-stream Transformer encoder + causal decoder with rotary positional embeddings and patch-based image tokenization; reduced hallucination rate 23% → 14% on a 1,000-sample human eval
- Reduced peak GPU memory 38GB → 26GB and improved training throughput by ~20% via gradient checkpointing, bf16 mixed-precision training, and fused attention kernels; enabling full model training on a single A100

GPT-4 Vision Multimodal RAG Pipeline| [Link]

- Improved document retrieval relevance by ~25% over a text-only BM25 baseline by designing a hybrid retrieval system (dense vector search + keyword ranking) using ChromaDB with parallel text/table/image processing, achieving 90%+ QA accuracy on a 300-question held-out benchmark
- Reduced end-to-end query latency from ~1.5s to under 800ms by implementing optimized chunking strategies, modality-specific embedding pipelines, and efficient approximate nearest-neighbor indexing for large financial document collections

SKILLS

ML & Deep Learning: PyTorch, TensorFlow, HuggingFace Transformers, Scikit-Learn, OpenCV · CNNs, Vision Transformers, Transfer Learning, MONAI
LLMs & Agents: LangChain, LangGraph, LangSmith, OpenAI GPT-4/Vision · RAG, FAISS, ChromaDB, Pinecone, Multi-Agent Systems
Infrastructure: Python, C++, SQL · CUDA, PyTorch DDP, HiPerGator HPC, AWS (ECS/EC2), Docker, Git, CI/CD