# NISHCHAL MARUR

+1 (240) 438-1916 | nishchal.marur@gmail.com | nishchal-mn.com | LinkedIn | GitHub | US Work Auth (OPT/STEM till 2029)

## SUMMARY

Machine Learning Engineer with 4+ years specializing in Computer Vision and Generative AI. Delivered 96% TPR authentication pipelines at 5% FPR, real-time recommendation systems, and 10x-compressed on-device models from research through production deployment.

## EXPERIENCE

**AI Engineer Intern | Connyct Inc | New York** — **Nov 2025 - Jan 2026**

- Built a Two-Tower hybrid recommendation engine combining sentence transformers and Elasticsearch with multi-signal user profiling, achieving Recall@10 of 0.85 for personalized events at <40 ms latency.
- Implemented MCP tools for the LLM orchestrator with location-aware reranking, composite relevance scoring, and Redis caching, to continuously optimize CTR with A/B testing framework.

**Research Assistant | PRG Lab (PI: Cornelia Fermuller) | Maryland** — **Jun 2025 - Aug 2025**

- Developed a cross-modal generation pipeline converting performer pose sequences to violin audio using DDSP, Transformer encoders, and autoencoder-based MIDI synthesis, enabling audio generation directly from motion capture.

**Machine Learning Engineer II | Entrupy Inc | Bangalore** — **Aug 2021 - Aug 2024**

- Built a deep learning authentication pipeline achieving 96% TPR at 5% FPR for 50K+ luxury items monthly, automating 95% of the verification volume, reducing manual expert reviews.
- Led the end-to-end R&D of a 3D document unwarping system that outperformed SOTA methods, resulting in 0.84 SSIM score and 23% increase in OCR accuracy, trained exclusively on synthetic data.
- Improved on-device segmentation IoU from 87% to 96% using SAM-HQ and trained a compact EfficientNet for CoreML deployment, achieving 10x model compression with FP16 quantization for real-time auto-capture on iOS.
- Optimized macro fingerprinting pipeline for luxury goods tracking using patch embedding similarity, boosting TPR by 15% and reducing authentication latency from 25s to under 15s with Ray parallel processing.
- Designed synthetic data pipelines using Blender Python and Stable Diffusion inpainting to simulate camera intrinsics, lighting, and material textures, reducing time-to-production for new authentication models from weeks to days.

**MLOps Intern | IBM | Bangalore** — **Jan 2021 - Jul 2021**

- Reduced inference latency by 15ms in IBM Watson Cloud deployments by optimizing batch prediction pipelines using Go concurrency and chunked downloads on Kubernetes.
- Benchmarked TensorFlow, PyTorch, and ONNX runtimes to evaluate performance trade-offs, and contributed to the design of an internal model serving architecture shift.

**Software Engineering Intern | SLK Software | Bangalore** — **May 2020 - Jul 2020**

- Saved developers 10+ hours weekly in debugging time by building a centralized log aggregation system using ELK Stack, Filebeat, Node.js across 5+ distributed components.

## TECHNICAL SKILLS

**Languages**: Python, C++, SQL, Go, Scala, Node.js
**ML Frameworks:** PyTorch, TensorFlow, Keras, HuggingFace Transformers, LangChain, LangGraph, OpenCV, CLIP, LoRA, CoreML, ONNX
**Infra & Data:** AWS, Azure, Kubernetes, Ray, Elasticsearch, Redis, FAISS, Pinecone, PostgreSQL, MongoDB, MLflow, WandB, Docker, Airflow, Triton, CUDA, Apache Spark

## PROJECTS

**CAFBrain: Multimodal LLM Platform (Agentic RAG | LLM)**

- Built an agentic RAG system using LangGraph with multi-step tool routing and autonomous query decomposition across 5000+ multimodal documents, reducing Capital Area Food Bank's grant proposal creation from hours to under a minute.

**Temporal Change Retrieval (Computer Vision | Multimodal)**

- Designed a dual-encoder architecture with a custom difference-attention module and LoRA-adapted RemoteCLIP for temporal vision-language alignment, achieving 64% Recall@10 on natural-language change queries.

**Scalable DBaaS for RideShare (Distributed Systems | Cloud Infra)**

- Designed a fault-tolerant Database-as-a-Service on AWS supporting 2000+ RPS with Raft-based leader election, automatic failover, and read/write routing via RabbitMQ RPC, achieving zero-downtime recovery under node failure.

## EDUCATION

**Master of Science (M.S) in Machine Learning** — **Aug 2024 - May 2026**
*University of Maryland, College Park* — *GPA: 3.8/4*

**Bachelor of Technology (B.Tech) in Computer Science** — **Aug 2017 - May 2021**
*PES University, Bangalore* — *GPA: 3.6/4*