MUNISH PERSAUD

Orlando, FL | 407-867-7739 | munishpersaud.mp@gmail.com
linkedin.com/in/munish-persaud | github.com/munishbp | munishbp.com

SUMMARY

Computer Science and Biomedical Sciences graduate from UCF (B.S., May 2026), building AI for healthcare across industry, research, and startups. Sole AI engineer on an agentic CRM that cut manual data entry by 63%, lead engineer on a HIPAA-compliant clinical app shipped to a 50-patient pilot, and first author of AI safety research submitted to NeurIPS 2026 that beat the prior state of the art by 28%. Strong in Python, PyTorch, TypeScript, and cloud-deployed ML systems. Seeking software engineering, machine learning, and AI research roles.

EDUCATION

University of Central Florida | Orlando, FL | Graduated May 2026
Bachelor of Science in Computer Science & Biomedical Sciences | *Minor in Non-Profit Management* | *GPA: 3.8/4.0*
Relevant Coursework: Machine Learning (CAP 5610), Algorithms for Machine Learning, Software Engineering (CEN 5016), Robot Vision, Robotic Systems, Quantitative Biological Methods, Advanced Data Structures and Algorithms, Matrix and Linear Algebra

HONORS & AWARDS

- 1st Place, UCF Innovate Startup Pitch Tournament ($2,000), November 2025
- Volunteer of the Year, Shepherd's Hope, 2024

TECHNICAL SKILLS

Languages: Python, Java, C, C++, TypeScript, Rust, JavaScript, SQL, Swift
Frameworks & Libraries: PyTorch, TensorFlow, Keras, React, React Native, Flutter, Node.js, Express.js, Flask
ML & AI: RAG, Light-RAG, LoRA / QLoRA fine-tuning, vLLM, LangChain, LangGraph, LlamaIndex, CUDA, ROCm, High-Performance Computing (HPC) Clusters, Diffusion Models, Qwen, Gemma
Cloud & Infrastructure: Google Cloud Services, Azure, FastAPI, REST APIs, gRPC APIs, Linux, Git, Docker, Kubernetes, Neo4j, PostgreSQL, Virtual Machines
Bioinformatics: MACS2, HOMER, UCSC Genome Browser, BLAST
Tools & Methodologies: Jira, Confluence, Agile, Linear, Huly, Claude Code, Cursor

WORK AND RESEARCH EXPERIENCE

AI Solutions Software Engineer Intern, Command Post Technologies | May 2026 – August 2026

- Project lead and sole AI engineer on an agentic CRM, owning AI, backend, and frontend; cut manual data entry by 63%, time from lead to pursuit by 22%, and review meetings by 19%.
- Interviewed the business development team to scope pain points, then built automated data entry, opportunity ranking on internal metrics, and shared pursuit tracking.
- Built the backend on a company hosted VM with Node.js and PostgreSQL, a Neo4j Graph-RAG knowledge base, and a vector database matching incoming opportunities to company capabilities.
- Deployed two LLMs on an NVIDIA DGX Spark with vLLM and Docker: Gemma 4 26B-A4B for automation and OCR, GPT-OSS 120B for document summarization and reasoning.
- Added document preprocessing for consistent model I/O and built guardrails that keep agent behavior inside company policy.

Research Assistant, Dr. Bedi’s SAFERR AI Lab, UCF | January 2025 – June 2026

- Led safety testing for Evo2, a 7B-parameter generative biology foundation model, and designed the evaluation workflows and protocols.
- Deployed and optimized 3 models on UCF's HPC cluster across up to 4 linked H100 GPUs, with containerized environments for reproducible inference.
- Analyzed 100+ test scenarios to surface safety vulnerabilities and measure performance.
- First-authored the resulting paper, submitted to NeurIPS 2026 and under review, proposing safety improvements for bio-foundation models.

Undergraduate Research Assistant, NSF REU at UCF Center for Research in Computer Vision | May 2025 – August 2025

- Developed a method that bypassed Evo2's safety protocols with a 28% higher success rate than the prior state of the art.

- Designed and built the evaluation framework testing model vulnerabilities and safety gaps across generative biology use cases.
- Produced the evidence for stronger safety protocols that fed into policy recommendations for generative biology.

Certified Nursing Assistant, AdventHealth Fish Memorial | 2023 – Present

- Delivered direct bedside care (mobility, vitals, daily living support) for 12 to 15 patients per shift across 3 years on a hospital floor.
- Charted patient status and flagged changes to nursing and physician teams for clean handoffs in a high-stakes setting.

PROJECTS

MAIA (Medical Administrative Intelligence Agent) | *Technical Founder*
maiamed.ai | In Development

- Validated the product through 50+ interviews with doctors and healthcare professionals and defined the technical requirements.
- Built an agentic framework (multi-agent orchestration, Light-RAG, LangGraph) that cut prior authorization from a ~20 minute industry average to 7 minutes per request, validated across 3 pilot tests.
- Built a voice note to EHR pipeline that finishes documentation in 11 minutes per encounter against a 16 minute physician average.

AI-PEER (Fall Risk Assessment & Intervention App) | *Project Manager & Lead Engineer*
Senior Design Project in partnership with UCF College of Medicine | Clinic Pilot (~50 Patients)

- Led a 5-person team to ship a HIPAA-compliant React Native app for gait assessment and fall-risk reduction, from first sprint to production IT handoff for a 50-patient clinic pilot.
- Shipped 23 Otago-protocol exercises and assessments (Chair Rise, Timed Up and Go) with on-device MediaPipe pose estimation at 30 FPS and 30+ tuned geometric heuristics for real-time form feedback.
- Fine-tuned Qwen3.5-2B with QLoRA on a 10,800-row geriatric dataset and quantized to 1.2 GB (Q4_K_M) for fully offline on-device chat that keeps patient data on the phone.
- Built a Google Cloud backend (Cloud Run, Cloud Storage with signed URLs, Firestore, Firebase Auth) and a scheduled REDCap sync, with a trilingual UI in English, Spanish, and Haitian Creole.

react-native-mediapipe-pose-plugin | *Author & Maintainer*
github.com/munishbp/react-native-mediapipe-pose-plugin | Published April 2026 (MIT, npm)

- Built a native VisionCamera v4 frame processor for iOS (Swift, Metal) and Android (Kotlin, MediaPipe GPU delegate) exposing 33 pose and 21 hand landmarks per frame at 10 to 20 ms inference.
- Published to npm under MIT with 230+ downloads as a drop-in replacement for YOLO and TFLite pose estimation, with typed peer dependencies for single-command install.

CAP 5610 Video Action Recognition Benchmark | *Course Project, UCF Machine Learning*
github.com/munishbp/action-world-recognition | Benchmarking 10 video understanding models on Something-Something V2 (174 classes, ~220K clips)

- Fine-tuned Qwen3.5-4B with 4-bit QLoRA (3.15M params, 0.12% of weights) on 168,913 clips to 58.19% top-1 and 0.56 weighted F1 over 174 classes, within 6 points of a frozen V-JEPA ViT-L baseline at 1/100 the parameters.
- Reproduced Meta FAIR's V-JEPA ViT-L/16 SSv2 eval at 64.51% top-1 on a single V100-32GB in 72 minutes, tracing the 5-point gap to published results to a hardcoded float16 autocast overriding the BF16 config.
- Trained an 18.6M-parameter PredRNN world model from scratch for 15 epochs, migrating mid-run from RTX 5090 to V100-32GB while preserving optimizer and scheduler state.

AURA (Plastic Surgery Facial Augmentation Tool) | *Founder & Full Stack Developer*

- Led a 4-person team building a plastic surgery outcome preview tool: React web, Swift iOS, and a LoRA fine-tuned DreamOmni2 model.
- Built CLIP-based guardrails restricting model input and output to facial imagery.

LISA (Lossless Image Slide Archival) | *Founder & Full Stack Developer*
Built in 36 hours at SWORL (Startup Weekend Orlando) | Proof of Concept

- Pitched and led a 4-person team building a VQ-VAE compression system for whole slide pathology images, hitting 98.8% size reduction (272 MB to 3.24 MB) with diagnostic quality preserved.
- Validated the clinical need through interviews with 7 medical professionals.