Abheek Pradhan

469-875-9205 | appradhann@gmail.com | linkedin.com/in/abheekpradhan | github.com/lolout1 | abheekp.dev

Education

Texas State University San Marcos, TX
Bachelor of Science in Computer Science (Computer Engineering Concentration) *Expected May 2026*

- Activities and Societies: Vice President - ACM, IEEE

Upcoming Publications: *Multimodal HAR using Cross Modal Learning ETA: 03/2026 (Link)*

Work Experience

Toshiba - Software Engineering Intern *May 2025 - August 2025*
Toshiba International Corporation

- Developed STM32H7 FreeRTOS firmware for Toshiba medium voltage drives; built comprehensive testing infrastructure from zero with unit/integration/HIL tests **(Python, Bash, Ruby)**. Reduced QA team workload by approx. 60% by creating production CI/CD pipeline validating 10,000+ parameters , running all unit, integration, and HIL tests in around 10 minutes.
- Optimized RTOS task priorities and DMA scheduling, reducing TouchGFX CPU overhead from 41% to 18% , achieving 25% total system CPU reduction , and eliminating all timing violations to correct issues revealed by my new HIL tests. **(C, C++)** .
- Developed TouchGFX UI screens with inter-task communication using CMSIS osMessageQueue for telemetry transfer between tasks. Double-buffered 30 FPS rendering with sub 80µs UI execution time ensures high-priority operations arent blocked.
- Engineered a RAG based AI agent using Azure Copilot and OCR for detecting defects in CAD drawnings with 94% precision.

Research Assistant (Texas State University , NSF Funded) *August 2024-Present* ***Github***

- Developed cross-modal distillation pipeline (video to IMU sensors) and personally trained custom multimodal transformers for fall detection under Dr. Anne Ngu. Deployed to edge devices (phones , wearables) with 92% F1 score in real time
- Built custom **CUDA** preprocessing kernels in C++ (windowing, normalization, FIR) w/ 11x speedup vs numpy, parallelized training via PyTorch **DDP**, tested DSP algorithms to clean dataset and reduce noise, improving F1 score by 10%
- Refactored multimodal **PyTorch** / **TensorFlow** transformer models to edge via TFlite and ONNX using INT8 quantization + mixed-precision training, achieving 2-3x battery life w/ sub-1% accuracy loss allowing for on device inference (Python) .
- Refactored full-stack **Android** app (**Java / Kotlin**) to support **ONNX** , **AWS S3** , and **DynamoDB** for sensor or user analytics. Shipped features like dynamic input support / pre-processing configs, async, to reduce noise and improve reliability

Machine Learning Engineer (Texas State University) *Dec 2024 - Sep 2025* ***Huggingface*** ***Github***

- Collaborated with research team funded by Texas State C.A.D.S to fine-tune Vision Transformer and MASK R-CNN models on Red Hat Linux SLURM cluster with Nvidia A100 GPUs. Created custom dataset achieving **98%** precision for defect detection
- Built production REST API using Python **FastAPI**, **PostgreSQL**, and **Docker** for deployment on **Huggingface**; implemented server side async request handling and batch processing to handle concurrent requests from React Native mobile app
- Accelerated inference via layer fusion and ONNX to TensorRT engine conversion; reducing latency and cloud costs by **40%**

Projects

Textbook2Video - 2nd Place Antler X Nvidia Hackathon (Python, GenAI, React, Langchain) **Huggingface - Live deployment**

- Deployed agentic (LangChain) multimodal pipeline automating animated educational video gen from PDFs using ElevenLabs TTS, Deepseek OCR, + fine-tuned Llama 3 via LoRA + 4b quantization. Hosted containerized model on HuggingFace

FPGA Optimized Facial Recognition (C, C++, Embedded Linux, Yocto, Ubuntu, PyTorch) **YouTube — Github**

- Developed facial recognition on AMD Kria KV260 SoC achieving 99.47% accuracy with ensemble architecture for face detection, recognition, and landmark extraction using Docker containerization for cross compilation on ARM64
- Engineered zero-copy DMA architecture with hardware-accelerated GStreamer pipeline, Vitis AI / Vivado toolchain optimizations, and INT8 quantization (16x size reduction, sub 0.5% accuracy loss), reducing memory bandwidth by 60%
- Delivered 100x CPU speedup and 300-800% industry improvement through multi-threaded processing and parallel DPU inference with custom Vitis kernels, improving throughput from 0.5-5 FPS to 30-500+ FPS

Sortify - Full-Stack Document Management (Python , PyTorch , RAG, PostgreSQL, React , FastAPI , Supabase) **Bitbucket**

- Built RAG pipeline using NLP and sentence transformers to generate vector embeddings for semantic PDF search with PostgreSQL pgvector database, hitting 94% retrieval accuracy through custom chunking algorithms

Distributed Chess Engine - Autonomous online chess bot (C++, C , NodeJS, Docker, Selenium) **YouTube — Github**

- Architected multi-process system with C++ TCP server, IPC communication, and browser automation pipeline using JSON messaging and containerized deployment ; Optimized for accuracy achieving 100% winrate in live chess games.

Skills

Full-Stack: ECS , C++, Java, AWS, GCP, Javascript , Apache, command line scripting , Go, Ada, SQL, C#, Restful API, Kubernetes, Maven, Gradle, Git, Docker, NoSQL, MySQL, CI/CD, Selenium, HTTP, Qt, Jira, Agile, SCRUM, Unix
AI / ML: LLM, CNN, Computer Vision, Streamlit, MCP, MLOps, Spark, Power BI, Tableu, attention , NLP , Kafka
Embedded Systems: FreeRTOS, VHDL, DSP, JTAG, Linux, I2C, SPI, UART, Drivers, Operating systems , VLSI