Abheek Pradhan

469-875-9205 | abheek.prad@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 2026*

- Activities and Societies: Vice President - ACM , Member: Google Developers Club , IEEE

Publications

(1st author) Dualstream Kalman Transformer – *Published* (Link) | Agentic Pervasive Computing – *Upcoming 05/26* (Link)

Work Experience

Software Engineering Intern *05/2025 – 08/2025*

Toshiba International Corporation

- Developed and mocked STM32 FreeRTOS firmware; added and tested features for touchscreen interface on Toshiba MVDs
- Built Jenkins CI/CD DevOps pipeline validating 10,000+ params via unit, integration and HIL tests to cut QA time by 60%
- 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++)
- Engineered RAG based AI agent via Microsoft Copilot, Azure ML and OCR to detect defects in CAD PDFs with 94% precision

Research Assistant *08/2024 – Present*

Texas State University, NSF Funded ***Github***

- Designed cross-modal knowledge distillation pipeline (video to sensor) and trained custom multimodal transformer architectures for real-time time series classification under Dr. Anne Ngu, achieving 91–95+% F1 on several datasets
- Deployed PyTorch / TensorFlow models to edge devices via LiteRT and ONNX using INT8 quantization and mixed-precision training, resulting in 3 to 5× battery life improvement with sub 1% accuracy loss for on-device inference with NPU / GPU .
- Built distributed training infrastructure on Slurm with Ray Serve, achieving 12× inference throughput; integrated attention algorithms (+4% F1), DSP and Kalman-based sensor fusion (+5% F1) during real-time inference to optimize F1 score .
- Created agentic supervised learning framework generating 10,000+ samples for text-to-motion diffusion; automated multimodal alignment + validation of 15,000+ sensor/video files via agentic pipelines, NLP , DSP and computer vision
- Built Python full-stack low-latency automated testing app (React, REST API, MongoDB, GraphQL) with Kafka/ Apache Spark for parallel multicore inference of 8 concurrent ML models in real time w/ Android (Java/Kotlin) wearable data

Machine Learning Engineer *12/2024 – 09/2025*

Texas State University, Center of Analytics and Data Science ***Huggingface*** ***Github***

- Fine-tuned Vision Transformer and Mask R-CNN models on distributed Red Hat Slurm cluster w/ NVIDIA A100 GPUs; created custom dataset w/ 98% precision for defect detection. Built production REST API (FastAPI, SQL , Docker) on Hugging Face
- Led and shipped MVP receiving 50k+ in funding. Built CV labeling pipeline using Detectron2, CVAT and vision LLMs with MLOps ; created active learning loop reducing manual labeling hours 80%. Deployed ML serving infrastructure on huggingface

Projects

Open Apply – Scalable Distributed Application Platform (Rust, Java, Spring Boot, TypeScript, PostgreSQL, Redis, NLP , RAG) **Live**

- Architected production distributed systems platform with Rust worker pool, Java Spring Boot microservices, and fault-tolerant message queues processing 1,000+ postings with idempotent APIs, load balancing, sub-2s P50 latency, and horizontal auto-scaling across containerized (Docker/Kubernetes) services deployed on AWS
- Engineered 3-tier RAG pipeline with NLP , vector similarity search, semantic memory store, and GCP + Vertex AI hosted LLM fallback achieving 94% first-pass field resolution accuracy over 8+ ATS platforms and 500+ field types .
- Designed resilient ML infrastructure with circuit breakers, dead letter queues, exponential-backoff retry with jitter, idempotency-key deduplication, and real-time Prometheus observability across PostgreSQL connection pools and Redis-backed task queues serving 10K+ daily inference requests

Textbook2Video – 2nd Place Antler X NVIDIA Hackathon (MCP, LangChain, Retrieval-Augmented Generation) **Huggingface – Live**

- Deployed agentic (LangChain) multimodal pipeline automating animated educational video gen from PDFs w/ ElevenLabs TTS, OCR and (SFT) fine-tuning Llama LLM via LoRA + 4-bit quantization. Hosted with Docker container on HuggingFace

FPGA Optimized Facial Recognition (C , C++14 , Embedded Linux, Yocto, Vivado, PyTorch) **YouTube — Github**

- Fine-tuned CV models and deployed onto AMD SoC achieving 99.47% accuracy; engineered zero-copy DMA architecture with hardware-accelerated GStreamer pipeline, Vitis AI / Vivado toolchain and INT8 quantization (16× size reduction) .
- Engineered C++ 14 thread pool for parallel DPU batch inference scaling throughput 30–60× over CPU/Python baseline to native webcam rate; cross-compiled via Docker for Embedded Linux (Yocto / PetaLinux) with zero CPU-side memory copies

Distributed Chess Engine - Autonomous online chess bot (C++, C , NodeJS, Docker, Selenium , HTML , CSS) **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: AWS, Go, TypeScript, GCP, JavaScript, Angular, Bash, Spark, SQL, C#, REST APIs, React, Git, Docker, MySQL, Selenium, Jira, Agile , Unix, Kubernetes, NoSQL, Ruby , Hadoop, Node.js, Golang, IoT, Cloud Computing , scripting , shell , bash , version control , HTML
Deep Learning: LLM, CNN, Computer Vision, MLOps, RLHF, scikit-learn , NumPy, Pandas, NLP, ONNX, TensorRT, GPRO , Quantization, Edge Deployment, Model Compression , Airflow , DSP , Tensorflow Keras , Pytorch , NPU / GPU inference , Azure ML , Vertex AI , Sagemaker
Embedded Systems : FreeRTOS, JTAG, Linux, I2C, SPI, UART, RTOS, Device Drivers, DMA, SoC Deployment, VLSI , CUDA