TASNIA SULTANA

Phone: 253.357.3143 **Email:** sultanatasniaeee92@gmail.com / tsulta@uw.edu

Link 1: LinkedIn, **Link 2:** My Portfolio, **Link 3:** GitHub, **Link 4:** Google Scholar

Machine Learning | Scalable AI | Vision Transformer | Model Compression | Multimodal Learning | Model Evaluation

Technical Skills

Programming Languages: *Python, JavaScript, TypeScript, SQL, R, Java, C++, GoLang, HTML/CSS, MATLAB*

ML Frameworks: *PyTorch, PyTorch Lightning, TensorFlow, Scikit-learn, Hugging Face, OpenCV*

Tools: *Weights & Biases, TensorBoard, Tableau, Power BI, Azure, AWS, Git, JIRA, Kubernetes, SQL Server, Google Colab*

Education

MS in Computer Science & Systems | University of Washington, Tacoma, WA | Expected: June 2026

- Research: Multimodal LLM compression (pruning, quantization, distillation) for resource-constrained AI deployment
- Relevant Courses: Algorithm, Distributed Systems, Machine Learning, Multimodal LLM, Query Optimization

BS in Electrical Engineering | Chittagong University of Engineering and Technology, Bangladesh

- Awarded HUAWEI AI Research Training Grant

Work Experience

Data Intern | IB Analytica, USA Mar 2025 – May 2025

- Processed large datasets using Python/SQL; built dashboards for model insights.
- Investigated data issues impacting user-facing ML outputs; collaborated on feature design.

Research Assistant | University of Washington Tacoma Sept 2023 - Present

- Conducted research on scalable AI, token pruning and quantization of LLaVA-NeXT for multimodal reasoning.
- Reduced inference latency and FLOPs by over 60% via layer pruning and token merging.
- Applied Grad-CAM, LIME, and SHAP for visual model interpretability.
- Benchmarked models using LMM-Eval and ONNX deployment for edge devices.
- Designed student-teacher framework using GPT-4 knowledge distillation.

Junior Software Engineer Intern | Sourcetop, Inc Bangladesh Jul 2022–Dec 2022

- Built frontend/backend features and automated deployments with CI/CD.
- Participated in team planning, debugging, and cross-functional reviews.

Research & Project Experience

MiniMed Assistant: A Scalable Multimodal Medical Assistant Model for Resource Constrained Devices

- Developed a lightweight multimodal assistant by compressing vision-language models (LLaVA-NeXT) using token and layer pruning, quantization, and distillation for real-time edge deployment.

A Computer Vision Approach for Detecting Discrepancies in Map Textual Labels, Journal Published

- Trained Faster R-CNN/DETR models to detect map label discrepancies across different map providers; integrated OCR.

Fashion Product Recommendation System

- Built a multimodal recommendation engine combining image and text embeddings (using CLIP).
- Extracted visual features and product descriptions to compute cosine similarity.
- Delivered top-5 fashion product suggestions, improving user-item match relevance by 30%.
- Applied to Kaggle's Fashion Product Images dataset to simulate real-world e-commerce behavior.

Quantization-Aware Dynamic Task Scheduling for Resource-Constrained Devices

- Developed a compiler-aware task scheduling framework for ML models under edge constraints.
- Reduced latency and energy consumption via dynamic graph partitioning and quantization.
- Integrated heterogeneous parallelism to optimize runtime on mobile-class hardware.

Leadership & Activities

Program Coordinator, BSO, University of Washington (Jan 2025–Mar 2025): Organized cultural and educational events to build community and support student well-being across diverse backgrounds.

General Secretary & Python Instructor, ASRRO, CUET, Bangladesh (Dec 2019–Mar 2022): Successfully led a team of 15 members to organize coding hackathons, robotics competitions, webinars, and taught Python libraries to peers.