Shaibal Saha

+1-248-710-7404 | shaibalsaha@oakland.edu | Portfolio | LinkedIn | GitHub | Google Scholar

Auburn Hills, Michigan, USA

Summary

PhD Candidate in Computer Science at Oakland University conducting research on efficient AI systems for computer vision and large language model (LLM) applications, with a focus on connected health and autonomous driving. Research includes improving deep learning inference efficiency through model quantization, developing retrieval-augmented generation (RAG) frameworks, and investigating computational efficiency in resource-constrained environments. Experience includes optimizing CNNTransformer models, developing LLM-based reasoning systems, and evaluating AI performance on edge devices. Published research in efficient AI, computer vision, and edge computing.

Education

■ **Oakland University** *September 2022 - August 2028 (Expected)*

PhD Candidate, Department of Computer Science and Informatics Rochester Hills, USA

Advisor: Dr. Lanyu Xu

■ **Oakland University** *September 2022 - December 2024*

Master of Science, Department of Computer Science and Informatics Rochester Hills, USA

■ **North South University** *January 2014-August 2018*

Bachelor of Science in Computer Science and Engineering Dhaka, Bangladesh

Completed Courses (MS and PhD): Embedded System Design Using FPGAs, Introduction to EdgeAI, Machine Learning, Computer Architecture, Advanced Data Structures and Algorithms, Cloud Computing, Database Systems, Discrete Methods I, Component-based Software Engineering.

Publications

B= Book Chapter, C= Conference, J= Journal

[C] **Shaibal Saha**, Fan Li, Yunge Li, Arun Iyengar, Lucas Alves, Lanyu Xu. KD-Judge: A Knowledge-Driven Automated Judge Framework for Functional Fitness Movements on Edge Devices, Accepted by the Eleventh IEEE/ACM Conference on Connected Health: Applications, Systems, and Engineering Technologies (CHASE) 2026. arXiv/DOI

[C] Yunge Li, Zhaodong Zhou, **Shaibal Saha**, Ali Irshayyid, Keer Chen, Lanyu Xu, Jun Chen. LightAD: A Lightweight Vision-Based Autonomous Driving System, Accepted by the Fourth IEEE International Conference on Mobility: Operations, Services, and Technologies 2026. DOI

[B] **Shaibal Saha**, Qiren Wang, Lichen Xia, Yongtao Yao. Chapter 4: Towards edge intelligence. In *Edge Computing: Systems and Applications*, edited by Lanyu Xu and Weisong Shi. Wiley Press, 2025.

[J] **Shaibal Saha**, Lanyu Xu. Vision transformers on the edge: A comprehensive survey of model compression and acceleration strategies. *Neurocomputing*, 643 (2025), 130417. DOI

[C] **Shaibal Saha**, Lanyu Xu. EfficientQuant: An Efficient Post-Training Quantization for CNN-Transformer Hybrid Models on Edge Devices. *T4V-CVPR Workshop*, 2025. arXiv/DOI

[C] Yunge Li, **Shaibal Saha**, Lanyu Xu. The architectural implications of multi-modal detection models for autonomous driving systems. *IEEE MOST*, 2024. DOI

[C] Ronald Tudu, **Shaibal Saha**, Prasun Nandy Pritam, Rajesh Palit. Performance analysis of supervised machine learning approaches for Bengali text categorization. *IEEE APWC on CSE*, 2018. DOI

Work Experience

■ **Oakland University**

Graduate Research Assistant *September 2022 – Present*

– Research Topic: Generative AI and Efficient LLM Systems

– Proposed a resource-efficient RAG-based pipeline for domain-specific automated judgment with optimized deployment on edge devices, resulting in a peer-reviewed conference manuscript (CHASE 2026). This project proposed an efficient RAG system with prompt engineering techniques and later proposed a deterministic judging system for the connected health domain.

– Research Topic: Efficient AI Systems on the Edge

– Conducted a comprehensive study of efficiency and compression strategies for vision transformer models on resource-constrained edge platforms, resulting in a peer-reviewed journal publication (Neurocomputing 2025).

– Proposed EfficientQuant, a post-training quantization technique for CNN–Transformer hybrid models, enabling resource-efficient deployment on edge devices, resulting in a peer-reviewed CVPR Workshop (2025).

– Research Topic: Perception, Safety, and System for Autonomous Driving

– Systematically retrofitted a semi-rugged outdoor vehicle to enable fundamental autonomous driving capabilities, resulting in one peer-reviewed conference publication (IEEE MOST 2026) and a 3rd-place award at the Intelligent Ground Vehicle Competition (IGVC) 2025.
– Deployed and analyzed multimodal, multi-task perception models for autonomous driving across heterogeneous computing platforms, resulting in a peer-reviewed conference publication (IEEE MOST 2024).

Graduate Teaching Assistant *May 2025 – Present*

– CSI4350: Programming Languages - Summer 2026
– CSI3450: Database Design and Implementation - Summer 2026
– CSI5240: Cloud Computing - Winter 2026
– CSI2310: Data Structures - Fall 2025, Winter 2026, Fall 2026
– CSI3660: System Administration - Fall 2025
– CSI3620: Data Structures and Algorithms - Summer 2025
– CSI3480: Security and Privacy in Computing - Summer 2025
– CSI2460: Fundamentals of Cybersecurity - Fall 2026

■ **Code Aristos** *January 2019 - June 2022*
Software Developer Dhaka, Bangladesh

– Started as an intern (Jan 2019 - Apr 2019) and transitioned to a full-time role due to strong performance.
– Worked on Hospital Management ERP for large-scale data handling for both web and mobile apps.
– Provided real-time client support for bug fixes for live systems all across the world.
– Designed and implemented robust backend functions for multi-level user access and customization based on client needs.
– **Frequently used tools:** Django, Laravel, MySQL, MongoDB, Azure, REST APIs, Docker, Redis, Nginx, GitHub Actions.

SKILLS

- **Programming Languages: Advanced:** Python | **Proficient:** C, C++, PHP, JavaScript, VHDL
- **ML Frameworks & Libraries:** PyTorch, TensorFlow, Hugging Face, NumPy, Pandas, Scikit-learn
- **ML Systems & Edge Deployment:** CUDA, TensorRT, ONNX Runtime, NVIDIA Triton, TVM, XFormers, vLLM
- **LLM & RAG Infrastructure:** LangChain, LlamaIndex, FAISS, RedisStack, LangSmith
- **Hardware Platforms:** NVIDIA GPUs, Jetson Nano, Jetson AGX Xavier, Xilinx ZCU104 MPSoC
- **Systems, MLOps & Tooling:** Docker, MLflow, CI/CD (GitHub Actions), Git, Jira, NVIDIA Nsight, Vaitrace, AWS, Azure
- **Web & Backend Development:** Django, Laravel, REST APIs, Nginx
- **Databases:** MySQL, PostgreSQL, MongoDB, Redis

GRANTS AND RESEARCH FUNDING

- **NSF ACCESS Explore Research Allocation (Principal Investigator)**, 2026. Project CIS262056; 200,000 ACCESS credits; 100,000 credits exchanged for Jetstream2 GPU resources (NSF-estimated value: $21,400).
- Oakland University Provost Graduate Student Research Award, Winter 2026 and Spring 2023.

HONORS AND AWARDS

- Participated in the 32nd Intelligent Ground Vehicle Competition (IGVC) 2025 in the self-driving category, *Position: 3rd (Grand award and overall).*
- NSF Student Travel Awards: IEEE MOST 2024, IEEE/ACM SEC 2023, IEEE/ACM CHASE 2023.
- Bangladesh-Sweden Trust Fund (BSTF) Travel Grant 2023.
- Selected for university special scholarship as an undergraduate for maintaining good results.

PROFESSIONAL SERVICES

- **Journal Reviewer:** IEEE Transactions on Circuits and Systems for Video Technology [2026–Present]; IEEE Transactions on Emerging Topics in Computing [2025–Present]; Elsevier Neurocomputing [2026–Present]; Elsevier Expert Systems with Applications [2025–Present]; Elsevier Journal of Systems Architecture [2026–Present]; Elsevier Smart Health [2025–Present]; Computers, Materials & Continua [2026–Present]; Recent Advances in Computer Science and Communications [2026–Present]; Journal of Ambient Intelligence and Smart Environments [2023–Present].
- **Conference & Workshop Reviewer:** AAMAS [2027], SIGCSE TS [2027], EIML@ICML [2026], DravidianLangTech@ACL [2026], LT-EDI@ACL [2026], MRL@EMNLP [2024, 2025].

LEADERSHIP EXPERIENCE

■ **ACM Student Chapter** *September 2024 - Present*
Oakland University [President]

REFERENCES

Lanyu Xu, Ph.D
Assistant Professor, Computer Science and Engineering
Oakland University
Email: lxu@oakland.edu
Relationship: PhD Advisor

Arun Iyengar, Ph.D
Computer Scientist
Intelligent Data Management and Analytics, LLC
Email: aki@akiyengar.com
Relationship: Mentor