Shaibal Saha

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

Auburn Hills, Michigan, USA

SUMMARY

Ph.D. candidate in the Edge Intelligence Systems Lab at Oakland University. Research focuses on efficient AI systems for edge and real-time environments, with applications in LLM-driven intelligence, connected health, and autonomous driving.

EDUCATION

- **Oakland University** — *September 2022 - August 2027 (Expected)*
  PhD Candidate, Department of Computer Science and Informatics — Rochester Hills, USA

- **Oakland University** — *September 2022 - December 2024*
  Masters 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

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 to enable 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

– CSI3660: System Administration - Fall 2025

– CSI3620: Data Structures and Algorithms - Summer 2025

– CSI3480: Security and Privacy in Computing - Summer 2025

■ **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: Expert:** Python, PHP, VHDL | **Proficient:** C, C++, JavaScript.
- **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, Generative AI Integration
- **Hardware Platforms:** NVIDIA GPUs, Jetson Nano, Jetson AGX Xavier, Xilinx ZCU104 MPSoC.
- **Systems, MLOps & Tooling:** Docker, MLflow, CI/CD (GitHub Actions), NVIDIA Nsight, Vaitrace, AWS
- **Web & Backend Development:** Django, Laravel, REST APIs, Nginx.
- **Databases & Collaboration:** MySQL, PostgreSQL, MongoDB, Git, Jira.
- **Completed Courses:** Embedded System Design Using FPGAs, Introduction to EdgeAI, Computer Architecture, Advanced Data Structure and Algorithms, Cloud Computing, Database System, Discrete methods I, Component-based software engineering.

HONORS AND AWARDS

- Oakland University Provost Graduate Student Research Award grant - Winter 2026.
- Participated in the 32nd Intelligent Ground Vehicle Competition (IGVC) 2025 in the self-driving category, *Position: 3rd (Grand award and overall).*
- NSF Student Travel Award for attending IEEE/ACM MOST 2024.
- NSF Student Travel Award for attending IEEE/ACM SEC 2023.
- Bangladesh-Sweden Trust Fund (BSTF)- Travel Grant 2023.
- NSF Student Travel Award for attending IEEE/ACM CHASE 2023.
- Oakland University Provost Graduate Student Research Award grant - Spring 2023.
- Selected for university special scholarship as an undergraduate for maintaining good results.

PROFESSIONAL SERVICES

- Reviewer of Elsevier Journal of Systems Architecture. [2026-Present]
- Reviewer of Elsevier Neurocomputing. [2026-Present]
- Reviewer of Elsevier Expert Systems with Applications. [2025-Present]
- Reviewer of Elsevier Smart Health. [2025-Present]
- Reviewer of IEEE transactions on emerging topics in computing. [2025-Present]
- Reviewer of Journal of Ambient Intelligence and Smart Environment. [2023-Present]
- Reviewer of EIML@ICML [2026], DravidianLangTech@ACL [2026], LT-EDI@ACL [2026], MRL@EMNLP [2024,2025]

LEADERSHIP EXPERIENCE & PROFESSIONAL MEMBERSHIPS

■ **ACM Student Chapter**,President *September 2023 - Present*

REFERENCES

1. **Lanyu Xu, Ph.D**
   Assistant Professor, Computer Science and Engineering
   Oakland University
   Email: lxu@oakland.edu
   Phone: (248)370-4079
   Relationship: PhD Advisor

2. **Arun Iyengar, Ph.D**
   Computer Scientist
   Intelligent Data Management and Analytics, LLC
   Email: aruniyen3@gmail.com
   Relationship: Mentor