Ishrak Jahan Ratul

Ph.D. Candidate · Department of Computer Science · Texas State University

ishrakratul@txstate.edu | github.com/ishrakratul | linkedin.com/in/ishrakratul | Google Scholar

Research interest:

Machine Learning Systems · Parallel and GPU Computing · Real-Time Systems · Inference Optimization · Deep Learning · AI for Healthcare

Education

Ph.D. in Computer Science — San Marcos, TX

Texas State University *Advisor: Dr. Kecheng Yang* — *All But Dissertation (ABD)*

B.Sc. in Electrical and Electronic Engineering — Gazipur, Bangladesh

Islamic University of Technology (IUT-OIC) — *Completed*

- Undergraduate thesis in biomedical machine learning; multiple first-author publications.

Research Experience

Texas State University — San Marcos, TX

Graduate Research Assistant *Advisor: Dr. Kecheng Yang* — *Aug 2023 – Present*

- Accepted to **IEEE RTSS 2026** (artifact evaluation accepted): proved **accuracy anomalies**, where adding a stage to a classifier cascade can lower end-to-end accuracy, and designed two optimization algorithms evaluated with six classifiers on an **NVIDIA Jetson AGX Orin** (with Prof. Sanjoy Baruah, Washington University in St. Louis).
- Cut energy per request by about **30%** versus parallel execution, with under **0.3 points** of accuracy loss, by stopping the losing branches of parallel IDK cascades early on an **NVIDIA RTX A4000**; also evaluated on Jetson Orin, artifact released (*MLSys 2027, under review*).
- *(In progress)* Developing **parallel speculative decoding** for transformer **LLM inference** on a shared GPU, targeting lower decode latency.
- Achieved **27% latency reduction** at **89.35% top-1 accuracy** with **IDK classifier cascades** that route uncertain inputs to higher-capacity models, extended to parallel cascades for real-time edge deployment (IEEE COMPSAC 2025, 2026; *J. Systems Architecture* 2026).
- Conducted the first systematic benchmark of PyTorch, ONNX Runtime, TensorRT, **Apache TVM**, and JAX on Jetson AGX Orin across inference time, throughput, memory, and power; TensorRT ran ResNet152 **about 4× faster** than PyTorch (*Electronics* 2025).
- Built **distributed fine-tuning pipelines across multi-GPU systems**; studied accuracy–latency–energy trade-offs for edge deployment.

AIMS Lab, United International University — Dhaka, Bangladesh

Research Engineer *Supervisor: Dr. Khondaker A. Mamun* — *Aug 2022 – Apr 2023*

- Co-developed a **national Digital Health Surveillance System for COVID-19** in Bangladesh, integrating real-time data pipelines and predictive models (*Frontiers in Digital Health* 2023).

Islamic University of Technology — Gazipur, Bangladesh

Undergraduate Researcher *Advisor: Md. Mirza Muntasir Nishat* — *Jan 2021 – May 2022*

- Developed ML classifiers for heart failure, HSCT survival, and cervical cancer prediction; the HSCT model retained **94.73% accuracy** using only the top 11 of 59 features selected by Chi-square testing.

Publications

See in Google Scholar.

Conference Papers

1. Sanjoy Baruah, **I. J. Ratul** and K. Yang. "**Accuracy Anomalies in Classifier Cascades**." *47th IEEE Symposium on Real-Time Systems (**RTSS**)*, **2026**. **(accepted; artifact evaluation accepted)**
2. **I. J. Ratul** and K. Yang. "Early Stopping of Speculative Classifier Cascades." *Conference on Machine Learning and Systems (MLSys)*, **2027**. **(under review)**
3. **I. J. Ratul** and K. Yang. "Fast and Accurate Classification with Parallel IDK Classifier Cascades." *IEEE 50th Annual Computers, Software, and Applications Conference (COMPSAC)*, **2026**.
4. **I. J. Ratul**, Z. Guo, and K. Yang. "Cascading IDK Classifiers to Accelerate Object Recognition While Preserving Accuracy." *IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC)*, **2025**.
5. **I. J. Ratul**, A. Al-Monsur, B. Tabassum, et al. "Early Risk Prediction of Cervical Cancer: A Machine Learning Approach." *19th Int'l Conf. on Electrical Engineering/Electronics, Computer, Telecommunications and IT (ECTI-CON)*, **2022**.

Journal Articles

1. **I. J. Ratul**, Z. Guo, and K. Yang. "Accuracy-Aware IDK Cascades for Real-Time Object Classification at the Edge." *Journal of Systems Architecture*, vol. 175, 103753, **2026**.
2. **I. J. Ratul**, Y. Zhou, and K. Yang. "Accelerating Deep Learning Inference: A Comparative Analysis of Modern Acceleration Frameworks." *Electronics*, vol. 14, no. 15, 2977, **2025**.
3. **I. J. Ratul**, M. M. Nishat, F. Faisal, S. Sultana, A. Ahmed, and M. A. Mamun. "Analyzing Perceived Psychological and Social Stress of University Students: A Machine Learning Approach." *Heliyon*, vol. 9, no. 6, e17307, **2023**.
4. F. Sarker, M. H. Chowdhury, **I. J. Ratul**, S. Islam, and K. A. Mamun. "An Interactive National Digital Surveillance System to Fight COVID-19 in Bangladesh." *Frontiers in Digital Health*, vol. 5, 1059446, **2023**.
5. **I. J. Ratul**, U. H. Wani, M. M. Nishat, et al. "Survival Prediction of Children Undergoing Hematopoietic Stem Cell Transplantation Using Different Machine Learning Classifiers by Performing Chi-Square Test and Hyperparameter Optimization: A Retrospective Analysis." *Computational and Mathematical Methods in Medicine*, 9391136, **2022**.
6. M. M. Nishat, F. Faisal, **I. J. Ratul**, et al. "A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset." *Scientific Programming*, 3649406, **2022**.

Presentations and Talks

"Fast and Accurate Classification with Parallel IDK Classifier Cascades," IEEE 50th Annual Computers, Software, and Applications Conference (COMPSAC), **2026**.

"Cascading IDK Classifiers to Accelerate Object Recognition While Preserving Accuracy," IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC), **2025**.

"Fast and Accurate Classification with Parallel IDK Classifier Cascades," Poster Presentation, MVAPICH User Group (MUG) Conference, **2026**.

Teaching Experience

Doctoral Instructional Assistant, Department of Computer Science — San Marcos, TX
Texas State University — *Aug 2023 – Present*

- **CS 3360 — Computing System Fundamentals** (Undergraduate): Supported lectures, graded assessments and exams; held regular office hours and helped students.

Teaching Interests: Computer Architecture; Parallel and GPU Computing; Operating Systems; Real-Time Systems; Machine Learning; Deep Learning; Data Structures and Algorithms.

Technical Skills

Programming Languages: Python, C/C++, CUDA C/C++, MATLAB, SQL

AI-Assisted Development: Claude Code, Codex

ML & Inference: PyTorch, TensorFlow, JAX, ONNX Runtime, TensorRT, Apache TVM, scikit-learn

Parallel & GPU: CUDA, MPI, OpenMP, Pthreads, HIP; multi-GPU; hybrid parallel programming (MPI + CPU threading + GPU)

Performance & Profiling: Nsight Compute/Systems, perf; GPU tuning (occupancy, memory behavior, atomics); throughput/speedup/efficiency analysis

HPC & Systems: Linux, Git, Slurm; HPC cluster workflows (TACC Lonestar6, Frontera); irregular workloads (graph analytics, BFS/SSSP, partitioning and load balancing)

Hardware: NVIDIA Jetson (AGX Orin), NVIDIA discrete GPUs, FPGA, microcontrollers

Service, Awards, and Memberships

Peer Review: 24 completed reviews for 17 journals, including IEEE Transactions on Services Computing, Scientific Reports, and BMC Medical Informatics and Decision Making.

Awards and Fellowships:

- Spoonamore Doctoral Merit Fellowship, Texas State University, 2023
- Travel Grant, MVAPICH User Group (MUG) Conference, 2026

Memberships: Student Member, IEEE · Student Member, IEEE Computer Society · Student Member, ACM