Don Kurian Dennis

PhD Candidate, Machine Learning Department
Carnegie Mellon University, Pittsburgh, PA, USA

dondennis@cmu.edu | donkdennis@gmail.com
Webpage : https://metastableb.github.io
Github : www.github.com/metastableB

Research Interests

I aim to understand, design, and implement theoretically sound end-to-end machine learning (ML) systems that bridge low-level hardware, algorithms, and high-level applications. My research focuses on resource-efficient ML, on-device ML, and optimization. More recently, I have been exploring multi-tier ML training, cascaded LLM inference, and resource scheduling and allocation across heterogeneous cloud clusters.

Key areas: Resource-efficient machine learning inference and training, systems for ML and on-device (edge) ML, convex optimization, and online learning/optimization in dynamic environments.

Software Experience

Languages — C/C++, Python. ML workflow — PyTorch, TensorFlow, JAX, MLflow. Optimization — Gurobi, CvxPy. Distributed computing — Ray, Slurm. On-device ML — Arduino, embedded-C tool chain, TFLite.

Education

Carnegie Mellon University
PhD in Machine Learning — *August '19 - May '25 (expected)*
Thesis title: *Adaptive Machine Learning in Dynamic Environments*
Advisor: Prof. Virginia Smith
Committee: Prof. Pradeep Ravikumar, Prof. Greg Ganger, Dr. Sanjiv Kumar (Google)

Indian Institute of Technology Patna
Bachelor of Technology, Computer Science and Engineering — *July '13 - May '17*

Work Experience

Microsoft Applied Sciences Group, Redmond, *Research Intern* — *May '21 - Dec '21*
Advisors: Dr. Kazuhito Koishida

Optimized the compute footprint of on-device noise suppression models using Shallow RNNs. Developed B-Distil, a novel method combining model ensembles and distillation to create elastic models for adaptive inference, achieving up to 10× reduction in inference cost for vision and language tasks while maintaining high accuracy metrics.

Microsoft Research India, *Research Fellow* — *July '17 - July '19*
Advisors: Dr. Prateek Jain & Dr. Harsha Simhadri

Developed and applied efficient ML algorithms (EMI-RNN, Shallow RNN) for resource-constrained edge devices, including IoT sensors and embedded systems. Designed real-world applications, such as GesturePod, a smart cane performing on-device gesture recognition with only 32kB RAM and 256kB flash memory on an ARM Cortex-M0+. Also developed an audio keyword spotting system optimized for a Cortex-M4, achieving real-time, low-latency performance.

ChironX, India, *Consultant* — *Mar '17 - July '17*

Consulted for ChironX, an early-stage startup developing AI-driven medical image analysis tools. Designed and implemented scalable infrastructure to support a computer vision-based retinal diagnostics engine, enabling efficient image processing and analysis for clinical applications.

Center for Smart Systems, SUTD/NUS Singapore, *Research Intern* — *May '16 - Aug '16*
Advisors: Dr. Vishram Mishra & Prof. Hock Lim Beng

Developed a prototype universal IoT Gateway to enable seamless device communication across diverse protocols, including WiFi, Bluetooth 3.0, BLE, and Zigbee. The gateway improved interoperability in IoT systems, simplifying device integration for smart environments.

Indraprastha Institute of Information Technology, Delhi, *Research Intern* — *May '15 - Aug '15*
Advisors: Prof. Debajyoti Bera

Explored a multi-point initialization Breadth First Search algorithm to improve throughput efficiency on Hadoop's distributed map-reduce framework. Investigated scalable ear-decomposition algorithms for graph processing to evaluate their performance in distributed computing environments.

Google Summer of Code '15, *Software Engineering Intern* *May '15 - Jul '15*
Advisors: David Anders (Intel) & Tom King (Minnowboard)

Developed the first complete simulation of the Harwell WITCH, a Dekatron-based (vacuum tube) computer used at the Atomic Energy Research Establishment, Oxfordshire, UK, in the early 1950s. Implemented in C++, the simulator was meticulously constructed using scarce schematics declassified in the late 2000s, preserving the historical accuracy.

Conference Publications

Progressive Ensemble Distillation: Building Ensembles for Efficient Inference
Don Dennis, Abhishek Shetty, Anish Sevekari, Kazuhito Koishida, Virginia Smith
Advances in Neural Information Processing Systems (NeurIPS), 2023.

Bitrate-Constrained DRO: Beyond Worst Case Robustness to Unknown Group Shifts
Amrith Setlur, Don Dennis, B Eysenbach, Aditi Raghunathan, Chelsea Finn, Virginia Smith, Sergey Levine
International Conference on Learning Representations (ICLR), 2023.

Heterogeneity For the Win: One-Shot Federated Clustering
Don Dennis, Tian Li, Virginia Smith
International Conference on Machine Learning (ICML), 2021.

Shallow RNN: Accurate Time-series Classification on Resource Constrained Devices
Don Dennis, Durmus Alp Emre Acar, Venkatesh Saligrama, Prateek Jain
Advances in Neural Information Processing Systems (NeurIPS), 2019.

Multiple Instance Learning for Sequential Data Classification on Resource Constrained Devices
Don Dennis, Chirag Pabbaraju, Harsha Simhadri, Prateek Jain
Advances in Neural Information Processing Systems (NeurIPS), 2018.

GesturePod: Programmable Gesture Recognition for Augmenting Assistive Devices
Shishir Patil, Don Dennis, Chirag Pabbaraju, Harsha Simhadri, Manik Varma, Prateek Jain
ACM Symposium on User Interface Software and Technology (UIST), 2019.

Single Cycle RISC-V Micro Architecture Processor and its FPGA Prototype
Don Dennis, Ayushi Priyam, Sukhpreet Virk, Sajal Agrawal, Tanuj, Arijit Mondal, Kailash Ray
International Symposium on Embedded Computing and System Design (ISED), 2017.

Preprints, Workshops & Demos

Agreement-Based Cascading for Efficient Inference
Don Dennis*, Steven Kolawole*, Ameet Talwalkar, Virginia Smith
In submission at *Conference on Machine Learning and Systems (MLSys), 2025*
arXiv preprint 2407.02348

Revisiting Cascaded Ensembles for Efficient Inference
Steven Kolawole, Don Dennis, Ameet Talwalkar, Virginia Smith
ES-FOMO Workshop at International Conference on Learning Representations (ICLR), 2024

Progressive Knowledge Distillation: Balancing Inference Latency and Accuracy at Runtime
Don Dennis, Abhishek Shetty, Anish Sevekari, Kazuhito Koishida, Virginia Smith
ES-FOMO Workshop at International Conference on Machine Learning (ICML), 2023

EdgeML: Demonstration of Low resource Keyword Spotting
Don Dennis, Harsha Simhadri, Prateek Jain
MLPCD2 Workshop at Advances in Neural Information Processing Systems (NeurIPS), 2018

GesturePod: Demonstrating On-Device Gesture Recognition
Shishir Patil, Don Dennis, Chirag Pabbaraju, Harsha Simhadri, Manik Varma, Prateek Jain
Microsoft Booth 203, Advances in Neural Information Processing Systems (NeurIPS), 2018

Talk-Bot: Federated Human Detection for Collaborative Multi-Angle Videography
Don Dennis, Harshit Singh, Karan Jakhar, Prashant Baghel
International Symposium on Embedded Computing and System Design (ISED), 2016

Software

EdgeML: Machine Learning for Edge and End-Point Devices [GitHub]

Open Source ***Microsoft Research***

Core developer and previous maintainer of EdgeML, Microsoft Research India's machine learning library for edge devices. Developed and applied efficient ML algorithms (EMI-RNN, Shallow RNN) for resource-constrained edge devices, including IoT sensors and embedded systems. Designed real-world applications, such as GesturePod, a smart cane performing on-device gesture recognition with only 32kB RAM and 256kB flash memory on an ARM Cortex-M0+, and an audio keyword spotting system optimized for Cortex-M4 devices, achieving real-time, low-latency performance. Also served as an advisor for startups leveraging EdgeML for use cases such as solar panel maintenance and voice-activated interfaces in sensitive environments

★ 1.6k+ ⑂ 350+

RISCV-RV32I-Assembler [GitHub]

Open Source

Developed a simplified assembly language and an assembler for the RV32I integer instruction subset of the RISC-V instruction set architecture (ISA). Used for simulating RV32I hardware design modifications in Verilog, testing FPGA prototypes and as an instructive tool for courses in pipelined architecture design.

Mixxx: Open Source DJ Mixing Software [GitHub]

Open Source

Worked as a contributor for Mixxx, a Music/DJ Mixing Software mainly written in C++/Qt. Worked on various bugs and improving the Auto-DJ feature for next track selection. My improvements were shipped in the 1.12 release.

WITCH On A Board [GitHub]

Open Source ***Google Summer of Code***

Awards & Achievements

2016 Runner up, Grand Challenge, ISED 2016
2016 Certificate of Leadership, National Entrepreneurs Network, India
2015 Bronze Medalist, CodeStorm 2015
2013 All-India-Rank 42 in CUSAT entrance exam amongst 40,000 candidates
2013 Top 0.3 Percentile in JEE Advanced 2013 amongst 150,000 candidates

Services and Positions of Responsibility

2024	Reviewer for ICLR	CMU
2023	Reviewer for NeurIPS, ICML	CMU
2023	Teaching Assistant, Federated and Collaborative Learning	CMU
2022	Teaching Assistant, Graduate Machine Learning	CMU
2021	Reviewer for NeurIPS, ICML	CMU
2020	Reviewer for NeurIPS	CMU
2018	Mentor, Machine Learning Summer School	Microsoft Research
2015	Helped win £50,000 funding for WITCH On A Board	Google Summer of Code
2015-16	Coordinator, NJACK, (Computer Science Club)	IIT Patna
2015-16	Instructor, Lecture Series on Operating Systems, NJACK (Computer Science Club)	IIT Patna
2015-16	Coordinator, Entrepreneurship Club	IIT Patna
2014-15	Sub-coordinator, Anwesha '15, IIT Patna's Annual Techno-cultural Fest	IIT Patna