Zain Taufique

Turku, Finland
zain.t.316@gmail.com
+358 414877718
Website LinkedIn Google Scholar Orcid

Profile

AI researcher with over seven years of experience in performance analysis and optimization of AI inference on heterogeneous CPU, GPU, and NPU platforms. Focused on real-time systems and the efficient execution of modern AI workloads, including multi-DNN pipelines and large language models. Expertise includes improving latency, throughput, and energy efficiency of AI inference across diverse computing architectures.

Technical Skills

Programming: C++, C, Python
Performance Computing: CUDA, multi-threading, inference on CPU/GPU/NPU platforms
AI Frameworks: PyTorch, TensorFlow, TensorRT, ONNX
Systems: Linux, Git
Platforms: NVIDIA Jetson series, Odroid XU3/XU4, Raspberry Pi 3/4/5, ASUS Tinker Edge R

Education

Doctoral Program in Technology (DPT): 2021 – 2026

Institution: University of Turku, Turku, Finland
Thesis topic: Efficient Run-time Systems for Edge AI Inference
Doctoral project: Approximate Computing for Power and Energy Optimisation (APROPOS)
Awarded grant: European Union's Horizon 2020 (H2020) Marie Sklodowska-Curie Innovative Training Networks H2020-MSCA-ITN-2020 call, under the Grant Agreement no 956090
Teaching courses:

- Computer Architecture and Operating Systems
- Embedded Systems Programming

Research output:

- 11 publications
- 153 citations
- 5 peer reviews conducted (2 conference papers, 3 journal articles)
- Contributions in funding applications for grants of the Academy of Finland, European Research Council (ERC), PoDoCo, Nessling Foundation, and Kordelin Foundation

Masters in Electronics and Embedded Systems: 2018 – 2021

Institution: Lahore University of Management and Sciences (LUMS), Lahore, Pakistan
Thesis topic: Early Detection of Chronic Neurological Disorders using Fast Fourier Transform Accelerator
Major subjects: Deep Learning, Embedded Systems, Digital Logic Design, Smart Grids

Bachelor of Science (Electrical Engineering: 2011 – 2015

Institution: University of Engineering and Technology (UET), Lahore, Pakistan
Major subjects: Programming in C, Data structures, Digital System Design, Signals and Systems

Professional Experience

PhD Researcher, University of Turku — Dec 2021 – May 2026

- Research on optimization of multi-DNN and transformer inference on heterogeneous platforms
- Designed runtime frameworks for dynamic model and data partitioning, precision control, cluster-level task mapping, and DVFS control
- Implemented scheduling and optimization strategies in C++ and Python with a focus on minimizing synchronization overhead and memory bottlenecks
- Conducted systematic profiling and benchmarking on multiple platforms with CPU, GPU, and NPU architectures
- Supervised a team of four resources in the computing lab that conducted the roles of teaching assistants and early-stage researchers.
- Conducted lectures and supervised lab sessions for multiple courses.
- Reviewing papers of multiple journals, including Microprocessors and Microsystems, and Journal of Systems Architecture

Industrial Research Secondment, Bosch Sensortec — Jun 2024 – Nov 2024

- Analyzed low-power firmware optimization techniques in C and C++
- Studied memory access behavior and compute efficiency tradeoffs on MEMS

Founding AI Engineer & Product Lead, Ferd.AI — Jan 2025 – Present

- Built and deployed AI-driven generative systems using Python backend and AWS
- Designed model selection pipelines integrating multiple LLM providers

Embedded Systems Engineer, Motive — Jan-2021 – Dec-2021

- Testing of Motive's ELD (Electronic Logging Device), an in-vehicle electronics system
- Implementing Agile methodologies, SCRUM, and Agile Testing Life Cycles

Research Assistant, LUMS — Jun-2019 – May-2020

- Research on low-power and small-area digital back-end hardware for wearables detecting neurological disorders using Verilog HDL Xilinx

Embedded Systems Engineer, Powersoft19 — June 2018 – Dec 2021

- Firmware development in C for NXP, Aurix, and STM microcontrollers
- Implemented communication protocols, including UART, SPI, and I2C
- Worked on hardware level performance validation and system testing

Utilities Engineer, US Denim Mills — Feb 2016 – May 2018

- Planning and erection of electrical projects for the industrial power plant

List of Publications

First author publications

Papers on AI

[1] **Zain Taufique**, Aman Vyas, Anotnio Miele, Pasi Liljeberg, and Anil Kanduri, "Twill: Scheduling Compound AI Systems on Heterogeneous Mobile Edge Platforms", In Proc. of ACM/IEEE International Conference on Computer Aided Design (ICCAD), 2025,
Pre-print available at: https://arxiv.org/abs/2507.00491

[2] **Zain Taufique**, Aman Vyas, Anotnio Miele, Pasi Liljeberg, and Anil Kanduri, "HiDP: Hierarchical DNN Partitioning for Distributed Inference on Heterogeneous Edge Platforms", In Proc. of IEEE Design Automation and Test in Europe Conference and Exhibition (DATE), 2025
Full text available at: https://ieeexplore.ieee.org/document/10992692

[3] **Zain Taufique**, Anil Kanduri, Anotnio Miele, Christia Bolchini, Nikhil Dutt, Amir Rehmani, and Pasi Liljeberg. Exploiting for Run-time Resource Management of Embedded HMPs. ACM Transactions on Embedded Computing Systems (ACM-TECS), 2025
Full text available at: https://dl.acm.org/doi/10.1145/3723357

[4] **Zain Taufique**, Aman Vyas, Anotnio Miele, Pasi Liljeberg, and Anil Kanduri, "TANGO: Low Latency Multi-DNN Inference on Heterogeneous Edge Platforms", In Proc. of IEEE International Conference on Computer Design (ICCD), 2024
Full text available at: https://ieeexplore.ieee.org/document/10817997

[5] **Zain Taufique**, Anotnio Miele, Pasi Liljeberg, and Anil Kanduri, "Adaptive workload distribution for accuracy-aware DNN inference on collaborative edge platformsæ, 29th Asia and South Pacific Design Automation Conference (ASP-DAC), 2024.
Full text available at: https://dl.acm.org/doi/10.1109/ASP-DAC58780.2024.10473987

Papers on health technology

[6] **Zain Taufique**, Awais Bin Altaf, Pasi Liljeberg, and Anil Kanduri, "Approximate Feature Extraction for Low Power Epileptic Seizure Prediction in Wearable Devices", IEEE Nordic Circuits and Systems Conference (NORCAS), 2021
Full text available at: https://ieeexplore.ieee.org/abstract/document/9599870

[7] **Zain Taufique**, Bingzhao Zhu, Gianluca Coppola, Mahsa Shoaran, Muhammad Awais Bin Altaf, "A low power multi-class migraine detection processor based on somatosensory evoked potentials", IEEE Transactions on Circuits and Systems II: Express Briefs (TCAS-II), 2021
Full text available at: https://ieeexplore.ieee.org/abstract/document/9380681

[8] **Zain Taufique**, Bingzhao Zhu, Gianluca Coppola, Mahsa Shoaran, Muhammad Awais Bin Altaf, "An 8.7 uJ/class. FFT accelerator and DNN-based configurable SoC for Multi-Class Chronic Neurological Disorder Detection", 2021 IEEE Asian Solid-State Circuits Conference (A-SSCC), 1-3
Full text available at: https://ieeexplore.ieee.org/abstract/document/9634763