Shadman Kaif

647-677-5811 | kaifshadman@gmail.com | in/shadman-kaif | shadman-kaif | shadman-kaif.github.io

Professional Experience

IBM June 2025 – Present
Software Engineer II *Markham, ON*
- Engineered a scalable **TPC-DS workload pipeline** on **watsonx.data** using **Spark**, **MinIO**, and **Presto Java** with **Spectrum Scale** storage backend; optimized MinIO write path, featured at **IBM TechXchange 2025**.
- Optimized **vLLM** performance on state-of-the-art **Spyre off-chip accelerator** on Power11, automating model precision with client data and **LLM benchmarks**, driving increased client partnerships for **Spyre** adoption.

IBM May 2023 – June 2025
Software Engineer *Markham, ON*
- Enhanced query processing speed on Power10 systems by optimizing **PyTorch multithreading** in large language models, resulting in a **42% increase** in inferencing throughput and achieving **sub-second** inferencing latency.
- Developed and migrated foundation models from Python Anaconda environments to Cloud Pak for Data, leading to **65% reduction** in ETL requests and **8x faster** access to distributed data across cloud.

IBM May 2021 – Aug. 2022
Back End Developer Intern *Markham, ON*
- Developed **99% accurate LSTM** models using **Tensorflow**, validating the cross-platform portability between x86 and Power10 systems, yielding a **58% lower cost solution** and a **2.4x per-core** performance advantage.
- Created **Bash** scripts to evaluate concurrent SQL query performance with Python UDFs for ML in DB2.

Ontario Treasury Board Secretariat May 2020 – Aug. 2020
Software Engineer Intern *Toronto, ON (Remote)*
- Spearheaded the division's data modernization initiative, migrating from a restricted MS Access platform to **Azure Cloud** using **Python** and **Pandas** to optimize digital data infrastructure, resulting in **90%** improvement.
- Built web scrapers using **Scrapy, Selenium** and **REST APIs** and analyzed the data using **NLP** and **PowerBI**.

Education

University of Toronto Sept. 2018 – Apr. 2023
BASc in Computer Engineering, Minor in Artificial Intelligence *Toronto, ON*
- Relevant coursework: Data Structures & Algorithms, Operating Systems, AI Fundamentals, Computer Security, Machine Learning, C++ Fundamentals, Computer Networks, Databases, Control Systems, Digital Electronics
- University of Toronto Scholars Award (Feb. 2018): $7,500 & Edward S. Rogers Scholarship (May 2018): $7,500

Awards & Innovation

- *Patent Pending*: Using Lightweight Generative AI Models to Prevent Cyber-attacks Jan. 2025
- IBM Entrepreneur Award: Built AI accuracy automation for client use, integrating new off-chip accelerator. 2Q 2025
- IBM Employee Equity Award July 2024 & July 2025
- MakeUofT 2020 Winner : Facial and optical recognition program using Xilinx's PYNQ-Z1 board. Feb. 2020

Projects

Ensemble ML Fraud Detection | *Jupyter Notebook, TensorFlow, NumPy, Pandas* Aug. 2022 – Apr. 2023
- Leveraged the **AdaBoost ensemble ML algorithm** with **decision tree classifiers** as base learners to minimize false positives from **60%** to **4.1%** on a credit card fraud analytics workload with over **24 million** transactions.

AlexNet Waste Classification CNN | *Jupyter Notebook, PyTorch, NumPy, Matplotlib* Oct. 2022 – Dec. 2022
- Created a transfer learning model based on **AlexNet**, achieving an accuracy rate of **94.1%**, marking an **8.6%** enhancement over a conventional **CNN** when applied to a waste segregation image dataset sourced from Kaggle.
- Accelerated the training process by a factor of **6x** through the utilization of transfer learning techniques.

Technical Skills

Languages: Python, C/C++, SQL (Postgres), Bash, JavaScript, HTML/CSS, MatLab, Verilog, ARM Assembly
Tools & Frameworks: Git, Linux, Kubernetes, Selenium, Scrapy, Node.js, Jira, DB2, Presto/Hive, Kafka, Flask
Libraries: C++ STL, NumPy, Pandas, Tensorflow, Onnx, PyTorch, Scikit-Learn, Matplotlib, Plotly, OpenCV
Microcontrollers: Arduino, DE1-SoC, Raspberry Pi, PYNQ-Z1/Z2