Arpita Saha

arpitasaha1996@gmail.com | 9374776531 | LinkedIn | Github | Scholar | Website

RESEARCH INTERESTS

Data Management, Databases, Data Mining, Information Retrieval, Big Data Analytics, Machine Learning, Computational Biology, Health Informatics.

EDUCATION

Master of Science in Computer Science and Engineering 2023
Department of Computer Science and Engineering *The Ohio State University*

- Thesis: Deep Phenotyping of COVID-19 Patients Using a Multi-Layered GRU Model on Large-Scale EHR Data
- CGPA: 3.81/4.0

Bachelor of Science in Computer Science and Engineering 2021
Department of Computer Science and Engineering *Bangladesh University of Engineering & Tech.*

- CGPA: 3.82/4.0

EXPERIENCE

Research Associate, Brandeis University, USA. *October 2023-Present*

- Hypothesized the problem of automatic knob tuning in NoSQL databases (LSM engine) for dynamic workload optimized database design and robust configuration selection. Designed a tentative Machine Learning solution and submitted a grant proposal to Amazon Research Awards (current project)
- Designed and presented a poster entitled: Toward Workload-Aware Self-Designing LSM-Engines at NEDB Day 2024.
- Working in a project that aims to achieve LSM memory profiling for different data structure implementations of the memtable. Studying and implementing memtable data structures in CASSANDRA and ROCKSDB
- Paper entitled “KVBench: A Key-Value Benchmarking Suite” accepted for publication in DBTest 2024. This paper introduces a workload generator tool used to stress test NoSQL data systems.: https://dl.acm.org/doi/10.1145/3662165.3662765

Graduate Research Assistant, Ohio State University, USA. *January-August 2023*

- Piloted the project for Covid-19 Mortality Prediction and Patient Phenotyping from large-scale EHR data.
- Published first-authored paper entitled "A Multi-Layered GRU Model for COVID-19 Patient Representation and Phenotyping from Large-Scale EHR Data" accepted at ACM-BCB 2023 (acceptance rate 29%): link
- Developed a GRU-based time-series deep learning model (only 11k parameters) to predict COVID-19 patient mortality outcome with an ROC AUC of 97% that outperforms all baselines (having around 700k parameters).
- Uncovered 4 distinct phenotypes by clustering strong patient representation embeddings and analyzed trends across phenotypes to identify risk factors related to mortality for efficient resource allocation during pandemic.
- Built an interactive desktop application to visualize time-series patient data using PyQT5 python module.

Graduate Teaching Assistant, Ohio State University, USA. *Summer 2023*

- Served as the Teaching Assistant for the OSU BMI Summer 2023 Internship Program
- Assisted in conducting workshops for Python, Statistical Modeling, Data Science, AI, Scientific Writing.
- Mentored students and managed student projects.

Graduate Teaching Assistant, Ohio State University, USA. *January-December 2022*

- Communicated effectively and built good rapport with instructor and students, enabling smooth class conduction.
- Contributed to the development of exam materials for appropriate and timely student evaluation.
- Mentored and managed a class of 100 students, and maintained their course roster and grade sheets.

Undergraduate Researcher, Bangladesh University of Engineering & Tech, BD. *2019 - 2021*

- Designed a novel semi-supervised variational auto-encoder deep learning model to impute missing taxa into gene trees as a member of a 3-person team, published the work as a co-author in RECOMB 2022: link.
- Used NLP techniques such as masked language modeling and positional encoding to improve performance.
- Utilized Python (Numpy, Pandas,Tensorflow) to code an end-to-end analysis pipeline and conduct all experiments.

PUBLICATIONS

1. KVBench: A Key-Value Benchmarking Suite; Zichen Zhu, **Arpita Saha**, Manos Athanassoulis, Subhadeep Sarkar; DBTest 2024: Proceedings of the Tenth International Workshop on Testing Database Systems.

2. A Multi-Layered GRU Model for COVID-19 Patient Representation and Phenotyping from Large-Scale EHR Data; **Arpita Saha**, Maggie Samaan, Bo Peng, Xia Ning; ACM BCB '23: Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics.

3. DePCoM: Deep Phenotyping of COVID-19 Patients Using a Multi-Layered GRU Model on Large-Scale EHR Data; **Arpita Saha**; Master's Thesis from The Ohio State University.

4. PRIEST: predicting viral mutations with immune escape capability of SARS- CoV-2 using temporal evolutionary information; Gourab Saha, Shashata Sawmya, **Arpita Saha**, Md Ajwad Akil, Sadia Tasnim, Md Saifur Rahman, M Sohel Rahman; Briefings in Bioinformatics.

5. Quartet based gene tree imputation using deep learning improves phylogenomic analyses despite missing data; Sazan Mahbub, Shashata Sawmya, **Arpita Saha**, Rezwana Reaz, M Sohel Rahman, Md Shamsuzzoha Bayzid; Journal of Computational Biology.

6. QT-GILD: Quartet based gene tree imputation using deep learning improves phylogenomic analyses despite missing data; Sazan Mahbub, Shashata Sawmya, **Arpita Saha**, Rezwana Reaz, M Sohel Rahman, Md Shamsuzzoha Bayzid; RECOMB '22: International Conference on Research in Computational Molecular Biology.

POSTER

- Toward Workload-Aware Self-Designing LSM-Engines; **Arpita Saha**, Alexander Ott, Subhadeep Sarkar: NEDB Day 2024.

SKILLS & INTERESTS

- Languages: Python, Java Script, C, C++, Java, MATLAB, SQL, HTML, CSS, SHELL
- Frameworks/Libraries: Django, Java Swing, PyQT5, MySQL, SQLite, PyTorch, Pandas, Matplotlib, Express
- Tools/Infrastructure: Git, SLURM, Linux, UNIX, Java Unit Testing, Agile, Scrum.

SELECTED PROJECTS

Studying the effect of Sparsification and Quantization on Large Language Models *June 2024*
- Sparsified TinyLlama-1.1B with sparse-gpt and quantized to 8 bits: link
- Studied the effect on accuracy of token predictions for datasets such as hellaswag, arch_challenge, mmlu, gsm8k, TruthfulQA, Winogrande

RESTful API for data exchange about Products and Order *June 2024*
- Built a RESTful API using Node.js and Express: link
- Stateless data exchange in JSON in a client-server architecture about products and orders
- GET, POST, PUT, DELETE, PATCH endpoints supported.

O-H-I-O Pose Detection from Live Video Input (Computer Vision) *December 2022*
- Built a desktop app using Python for collecting images to curate a dataset by collecting live video feed using webcam at different lighting and background conditions.
- Leveraged the frames from videos to build MEI, MHI images and calculate similitude moments, which were used as features for the KNN classifier that detects the correct pose.

Rating Software for Alpha Credit Rating Company *March 2021*
- Built a software to calculate transition probability from one rating to another in a year based on past data of companies using Java Swing and MySQL.

- Planned and executed **full-stack** development of the Software, including relational database design.

Tour Planner Website (Software Engineering Project) *January 2019*
- Built a website using Django and SQLite for planning a tour given destination and time budget: link.
- Used Traveling Salesman Problem as backend algorithm; incorporated search and admin privileges.

PERSONAL ACHIEVEMENTS

- Anita B. Org Scholarship for Attending GHC 2022 *January 2022*
- BUET Dean's List Award *2016-2021*
- BUET Merit Scholarship for top 10% in Computer Science *2016-2021*

LEADERSHIP ROLES

- Vice President and Co-founder, IEEE Computer Society Student Branch, BUET.
- Treasurer, Bangladeshi Women in CSE, BUET.
- Senior host and organizer, BSADD, BUET.