# Saurabh Singh

Graduate Student | Computer Science
6172021415 ssmail@bu.edu Portfolio
LinkedIn/saurabh-sks
GitHub/Saurabh-Singh-55

## Education

**Boston University** **05/23/2023 – 01/17/2025**
*Masters in computer science* *Boston, MA, USA*
*GPA: 3.56/4*

**Tezpur University** **08/01/2018 – 07/28/2022**
*Bachelor of Technology in Computer Science and Engineering* *Tezpur, Assam, India*
*CGPA: 7.48/10*

## PROFESSIONAL EXPERIENCE - Red Hat (Boston, MA)

**Software Engineer, Data & AI** **July 2025– Present**

- Building full-stack AI applications on top of data systems like Snowflake, Atlan, Gitlab via MCP servers to drive Red Hat's data initiatives.
- Led development of an AI assistant to streamline data access via MCP servers on 100+ databases in snowflake via AI generated SQL.
- Created a natural language query AI interface for Snowflake that empowers business users to self-serve data with over 90% accuracy.

**Data Engineer** **Jan 2025– July 2025**

- Designed and optimized data pipelines using DBT, Snowflake, and advanced SQL, improving data transformation processes and streamlining ETL workflows for efficient and scalable data operations across multiple business units.
- Collaborated in delivering source-aligned data products and aggregating data from multiple sources, ensuring consistency and accuracy to support business intelligence and decision-making processes across teams.

**Data and AI Intern** **May 2024 – Dec 2024**

- Developed RHContract.AI, a contract discovery tool using LLMs, reducing contract document processing time by 90% by implementing contract filtering, extraction, and signature detection on 300k PDFs (156GB), leading to improved efficiency for Red Hat's Sales team.
- Achieved over 90% accuracy in attribute extraction and up to 95% in signature detection by leveraging multimodal LLMs.

## AWARDS & RECOGNITION

**Dataverse Business Impact Award (*Best Engineering Innovation in AI*) – Red Hat (Q2 2025)**

- Developed an LLM-powered conversational assistant enabling leaders to extract business insights from large enterprise databases.

## CONFERENCE PROCEEDING

**Adaptive early classification of time series using Deep Learning**
*Conference: International Conference on Neural Information Processing (ICONIP 2022)*

- Developed the RCRL model, leveraging RNN and CNN methods, to pioneer adaptive early time series classification. Achieved remarkable accuracy improvements while enabling timely predictions, making high impact in fields like medical diagnosis.

**Driving behavior analysis using Deep Learning on GPS data**
*Conference: International Conference on Emerging Global Trends in Engineering and Technology (EGTET 2022)*

- Applied advanced statistical techniques and feature extraction methods to assess driving behavior from GPS trajectory data.

## PROJECTS

**Offline Chatbot Development with RAG on PDFs and Webpages**

- Designed and implemented an offline chatbot integrating PDFs and webpages using vector embeddings. Enabled fast and accurate query resolution via similarity search and Retrieval-Augmented Generation (RAG) with local LLMs.)

**Advanced MNIST Classification and Dimensionality Reduction Analysis**

- Analyzed MNIST using PCA, SOM, RBM, and VAE for dimensionality reduction. Developed a custom 2D-VAE model achieving 94% accuracy, significantly outperforming other methods across classification metrics.

**Advanced Query Processing with Spark and OpenAI**

- Built a RAG system on 28GB of arXiv PDFs using Spark and OpenAI. Generated vector embeddings with LangChain and FAISS, deployed a search interface using a custom Streamlit app on AWS EMR.

**Transfer Learning on CIFAR10: Benchmarking Inceptionv2, ResNet50, and VGG19**

- Fine-tuned InceptionV3, ResNet50, and VGG19 on CIFAR-10 using advanced training strategies. InceptionV3 achieved the best performance at 95.44% accuracy and 0.95 F1 score.

**Kafka-Airflow Pipeline Implementation for Real-Time Data Streaming**

- Developed a real-time pipeline streaming data from an external API using Kafka and Python. Scheduled daily Airflow DAGs for automated ingestion and processing of user data.

**Optimal Gameplay using Reinforcement Learning**

**11 different LLMs (3B, 7B, 13B, 40B, 70B) Inference time and Resource Usage Analysis**

## Technical Skills

**AI Skills:** MCP servers, AI agents, Prompt Engineering, LangChain, Vector Embeddings, NL-to-SQL, RAG chatbots.
**Data Engineering:** Snowflake, dbt, fivetran, Apache Kafka, Apache Airflow, Apache Spark, ETL/ELT.