KAUSTUBH SONAWANE

Raleigh, NC | ksonawa@ncsu.edu | +1-352-679-2771 | linkedin.com/in/iamkaustubhs | kaustubhas.github.io

EDUCATION

North Carolina State University, Raleigh, NC — Aug 2024 – May 2026
Master of Computer Science — GPA: 3.7/4.0

University of Mumbai, Mumbai, India — Aug 2020 – Jun 2024
Bachelor's in Computer Engineering, Minor: Data Science — CGPA: 8.83/10

Coursework: Data Structures and Algorithms, Software Engineering, Computer Architecture, ML, AI, NLP, Operating Systems, Deep Learning, Statistics, Design and Analysis of Algorithms, Database Management Systems, Optimal transport in ML, Generative AI

TECHNICAL SKILLS

Programming Languages: Python (Matplotlib, Seaborn), SQL, R, Kotlin, C++

Frameworks: PyTorch, TensorFlow, PySpark, Apache Spark, FastAPI, LangChain, Scikit-learn, Hugging Face, CUDA, Keras

Tools & Cloud: PowerBI, Tableau, Jupyter, AWS, Azure, PostgreSQL, SQLite, SQLAlchemy, Git, GitHub, Docker, Excel

AI & ML: DNNs, LightGBM, Statistical Models, Hypothesis Testing, Data Pipelines, Transformers, LLM, Computer Vision, RAG

Certifications: Accelerating End-End Data Science Workflows (NVIDIA), DevOps, Student Development Program, (SIESGST)

WORK EXPERIENCE

Data Science Intern, *NetApp*, Raleigh, NC, USA — May 2025 – Aug 2025

- Built two end-to-end forecasting pipelines (AzureML, AbacusAI) by analyzing 3M+ sales records to predict renewal and tech-refresh bookings, translating ambiguous requirements into scalable solution, increased quarterly revenue forecasting accuracy by 15%.
- Designed SQL and Python ETL pipelines to process 3M+ operational records and engineer 20+ features for customer lifecycle modeling, reducing risk estimate variance by 9% and manual processing time by 40% while accelerating stakeholder alignment.

Data Visualization Intern, *SIES GST*, Navi Mumbai, India — Jul 2023 – Aug 2023

- Designed Tableau and Power BI dashboards (Crypto, Covid-19, Spotify trends) tracking KPIs via time-series analysis; engineered automated pipelines that instantly ingest standardized new data to update metrics, eliminating 90% of manual dashboard rebuilds.

TECHNICAL PROJECTS

Robust Partial Wasserstein (RPW) Analysis | *Sep 2025 – Present*

- Engineered Optimal Transport models for high-dimensional anomaly detection, enhancing models for 15+ distributed FedAF clients
- Derived $O(n^{-0.4})$ convergence rates via multiscale heuristics, maximizing statistical efficiency over standard grid-based transport
- Validated p=infinity bounds, proving RPW rejects outliers to achieve 65.4% accuracy where classical Wasserstein fails

Code Smell Classification And Refactoring using LLMs | *Aug 2025 – Nov 2025*

- Deployed a 2-agent LLM-based technical debt automation system utilizing GitHub Actions and an MCP-enabled VS Code workflow.
- Fine-tuned GPT-4o-mini on 140K+ real-world samples to detect 20+ code smells (70-80% F1-score), achieving $\sim$70% safe refactors and enabling end-to-end remediation in $\sim$10s via MCP and $\sim$2 min via GitHub PRs.

LLM-Powered Clinical QA System | *Mar 2025 – Aug 2025*

- Built a biomedical RAG-based QA assistant utilizing Knowledge Graphs and LLMs, achieving >95% reliable responses across 10K+ structured biomedical entities to significantly enhance clinical decision making workflows and reduce manual diagnosis time.
- Engineered a distributed PySpark preprocessing pipeline to ingest and structure 1M+ medical records, implementing a robust fallback system for edge cases to ensure seamless data retrieval and processing within highly scalable cloud environments.

Comparative Analysis of Real-Time Object Recognition | *Sep 2024 – Dec 2024*

- Designed a PyTorch/CUDA benchmark on the PASCAL dataset (9,600+ images) to quantify speed-accuracy trade-offs, demonstrating YOLO's speed (135 FPS) against SSD's precision (91.7%) for optimal edge deployment.

Customer Behavior Analysis | *Jul 2023 – May 2024*

- Analyzed large-scale customer behavior data (Amazon Food Reviews: 568K reviews, 74K products, 256K users) utilizing SQL-driven feature extraction and statistical analysis to uncover actionable insights into purchasing trends and customer lifecycle.
- Built recommendation systems comparing classical ML (Random Forest, SVD, PMF) and deep learning (CNNs, Stacked LSTM, Autoencoders) via TensorFlow and PyTorch to improve personalization accuracy and drive increased user engagement and retention.

VOLUNTEER EXPERIENCE

Cross-Functional Team Lead, *Girija Welfare Association*, India — Jul 2022 – Jul 2024

- Led cross-functional teams for community events and fundraising, leveraging management skills to execute engagement initiatives.