Harshith Kethavath

+1 (980) 290-8567 | hello@harshithkethavath.com | LinkedIn | GitHub | Website

Education

The University of Georgia — Athens, GA
Master of Science in Computer Science, GPA: 4.0/4.0 — *Aug 2024 – May 2026*

- **Coursework**: Computer Networks, Software Engineering, Algorithms, Database Management, Operating Systems, and Advanced Data Intensive Computing

Indian Institute of Technology Bhubaneswar — Bhubaneswar, India
Bachelor of Technology in Electrical Engineering — *Jul 2019 – May 2023*

Research

Lab for Geoinformatics and AI Modeling — Athens, Georgia
Graduate Student Assistant, ***Director:*** *Dr. Weiming Hu* — *Oct 2025 – Present*

- Developing a deep learning framework using PyTorch and Hugging Face Transformers for early weather prediction by fine-tuning Vision Transformers (ViT) on upward-facing sky images.
- Engineering a data-labeling pipeline to process and align historical ASOS weather station data with image timestamps, generating ground-truth labels for both classification and regression tasks.

UGA Edge/Cloud Research Lab — Athens, Georgia
Student Researcher, ***Director:*** *Dr. In Kee Kim* — *Oct 2024 – Present*

- Developed and deployed a suite of Computer Vision scripts, and a Recommendation Model, onto Jetson Orin Nano, generating a 7 million point dataset to benchmark throughput against thermal stress.
- Discovered critical performance degradation across models when GPU temperatures reached the 60 to 70°C range under full load, including a 36% drop in recommendation system at 75°C.

Work Experience

HCLTech — Pune, India
Cloud Engineer, ***Client:*** *Amazon Lab126* — *Oct 2023 – Jun 2024*

- Automated HPC cluster provisioning using SOCA template and CloudFormation, saving 80 manual configuration hours. Resolved over 120 client tickets related to cluster access, job submission (Slurm), and queue management.
- Created and documented reusable POCs for large-scale AWS systems, which streamlined strategic decision-making, and reduced new hire onboarding time by 75%.

Bhaktivedanta Institute — Remote
Web Development Intern — *May 2022 – Jul 2022*

- Developed 6 new responsive front-end pages using HTML, CSS, and JavaScript to ensure a consistent user experience across all devices and refactored main CSS file to reduce code duplication by 27%.
- Coordinated with back-end developers to integrate REST APIs seamlessly into front-end, performed regular testing, resolved over 30 bugs and updated website content, increasing user engagement by 15%.

Projects

PocketRAG: A Framework for Selection of Quantum Encodings:

- Engineered a modular RAG pipeline on Apple Silicon (M3) using PyTorch MPS and Python, implementing hybrid retrieval (BM25 & FAISS) to benchmark trade-offs between dense/sparse indexing and generation latency.
- Developed a reproducible CLI evaluation toolkit to standardize LLM performance, enabling end-to-end analysis of chunking strategies and embedding models against metrics like Recall@K, MRR, and NDCG.

Q-Select: A Framework for Selection of Quantum Encodings:

- Engineered a QML benchmarking pipeline in Python and Qiskit to profile 6 quantum encoding methods across 6 diverse datasets, then used the resulting 36-configuration performance matrix to train a random forest regressor that predicts classifier accuracy based on 20+ statistical features, achieving an RMSE of 11.90%.
- Discovered that encoder complexity is the top predictor of performance and packaged the final model into a Python CLI tool that analyzes new, unseen datasets to recommend the optimal encoding strategy.

W-Shingling for Wikipedia Document Evolution Study:

- Engineered a Python data pipeline to analyze Wikipedia article evolution, using W-Shingling, MD5 hashing, and MinHash to generate compact document signatures and Jaccard similarity to quantify content decay.
- Discovered that shingle size (w) was the dominant factor in computational cost, while signature size (λ) had a negligible performance impact, proving that accuracy ($\lambda = 64$) could be maximized without a speed trade-off.

Geospatial and Temporal Analysis of Clark County Crash Data:

- Engineered a data pipeline using Pandas and NumPy to clean, transform, and feature-engineer 9+ years of NDOT crash data, creating 12+ dummy variables to classify incidents by impairment, speeding, and adverse weather.
- Executed a geospatial and temporal analysis by calculating incident density within 1, 5, and 10-mile radii of the Super Bowl and Grand Prix, aggregating results by hour and day to identify high-risk periods.

Creo on Amazon AppStream 2.0:

- Designed a solution with SSO access, by integrating IAM with Azure AD, provisioning AppStream desktops, and implementing FSLogix with FSx to enable a scalable user cache of up to 3TB, replacing the 1GB default.
- Implemented automation for provisioning, configuration, and deployment of resources using Terraform for IaC, reducing deployment time by 20%.

3-Tier Web Application:

- Built a 3-tier web architecture on AWS for a student club, using an auto-scaling web tier, and an Application Tier with a secure backend and Bastion host, and Amazon RDS for MySQL for secure inventory management.
- Configured VPC, subnets, route tables, and security groups to enhance security and traffic segregation. Used auto-scaling groups for EC2 and an ALB to optimize costs by 25% while ensuring high availability.

Serverless Sending Application:

- Built a web-based solution enabling users to send emails and SMS via Amazon SES and SNS, integrated with API Gateway and AWS Lambda for processing requests, to achieve an average response time of 300 ms.
- Orchestrated workflows using AWS Step Functions and Amazon S3, resulting in a secure, serverless, event-driven infrastructure with a pay per use model that eliminated all idle infrastructure costs.

Skills

Languages: Python, C, C++, Java, and JavaScript

Data Science: SQL, Numpy, Pandas, Matplotlib, PyTorch and Qiskit

Infrastructure Engineering: AWS, IAM, EC2, VPC, S3, RDS, FSx, Lambda, Terraform, Git, and Linux

Certifications: AWS Certified Cloud Practitioner