# SAUMIT PAUL

+1 (984) 335-8991 | Raleigh, NC | sspaul3@ncsu.edu | saumitpaul | 5aumit | 5aumit.github.io

## SUMMARY

Electrical Engineering M.S. candidate (**May 2027 Grad**) with 3 years of experience building, validating, deploying, and monitoring production machine learning systems. Experienced in computer vision, medical imaging, deep learning, model compression, automated evaluation, data pipelines, regulatory validation, and LLM-based agent systems.

## EDUCATION

**North Carolina State University** 08/2025 - **05/2027**
Master of Science, Electrical Engineering GPA: 4.0
Relevant Coursework: Generative AI For Systems, Data Science for Signal Processing, Neural Networks

**Manipal Institute of Technology** 07/2018 - 08/2022
Bachelor of Technology, Electrical and Electronics Engineering WES GPA: 3.8

## PROFESSIONAL EXPERIENCE

**Data Scientist**, DeepTek.ai 06/2022 - 07/2025

- Developed and validated PyTorch and TensorFlow/Keras models for chest X-ray classification and segmentation, including pleural effusion detection with over 90% AUROC on held-out test data.
- Owned model lifecycle from data collection and preparation through training, validation, ONNX deployment, production monitoring, and failure analysis.
- Built an automated model evaluation and debugging platform with batch inference, MLflow experiment tracking, metric computation, and visual error analysis, reducing evaluation time from several hours to under 5 minutes.
- Maintained and validated a MongoDB-backed annotation and metadata system covering more than 1.8 million imaging studies, resolving schema and data-quality issues across heterogeneous clinical datasets.
- Applied INT8 quantization-aware training with TensorFlow Model Optimization and measured the resulting accuracy and deployment-efficiency tradeoffs for a pleural effusion model.
- Collaborated with engineering, clinical, regulatory, product, and customer-facing teams to investigate model failures, explain behavior, support demonstrations, and translate requirements into model changes.

## PROJECTS

**Floki: MLflow Experiment Analysis Agent** (GitHub) 12/2025 - Present

- Built LangChain-based CLI ReAct agent to query MLflow experiments, compare runs, retriev metrics and parameters, and summarize performance and failure patterns through structured tool calls.
- Integrated Langfuse tracing, Pydantic validation, and Rich TUI to improve observability and reliability across multi-step agent workflows.

**LLM-Based Cache Replacement Policy Optimization** 08/2025 - 12/2025

- Built an autonomous LLM harness to generate, compile, benchmark, and refine C++ cache replacement policies in ChampSim using SPEC CPU 2006 IPC feedback, reaching an average IPC of 0.414.
- Added graph-structured vectorless RAG over prior policies and early rejection using representative workloads, reducing evaluation time from roughly 40-50 minutes to about 30 minutes.

**Model Compression and Accuracy Analysis for Medical AI** 01/2026 - 04/2026

- Applied unstructured L1 pruning to a pretrained DenseNet-121 chest X-ray model and tested sparsity levels from 0% to 97% across 2,797 unique patients.
- Measured AUROC, recall, and specificity across four clinical findings, finding stable performance through approximately 80% sparsity and sharp degradation above 85-90%.

## PUBLICATIONS

- A Comprehensive Evaluation of DeepTek CXR Analyzer in Detecting and Localising Suspicious Findings in Chest X-rays (doi.org/10.26044/ecr2024/C-23322)
- Artificial Intelligence as a Proficient Tool in Detecting Pulmonary Tuberculosis in Massive Population Screening Programmes: A Case Study in Chennai, India (doi.org/10.2185/jrm.2024-015)

## SKILLS

| | |
|---|---|
| **Programming** | Python, SQL, C++, R, Bash |
| **Machine Learning** | PyTorch, TensorFlow/Keras, Scikit-Learn, Hugging Face Transformers |
| **Computer Vision** | OpenCV, DICOM, TorchXRayVision, Voxel FiftyOne, CVAT |
| **GenAI** | LangChain, LangGraph, Langfuse, OpenAI, Ollama, Groq |
| **Model Optimization** | Quantization-Aware Training, Model Pruning |
| **MLOps and Data** | MLflow, DVC, ONNX, Git, Linux, MongoDB, SQL, Pandas, NumPy, Azure Data Factory |