

# Sarthak Singh

+1 (631)-355-8065 | [sarthaksingh1211@gmail.com](mailto:sarthaksingh1211@gmail.com) | [linkedin.com/in/sarthaksingh1211](https://www.linkedin.com/in/sarthaksingh1211) | [github.com/sarthaksingh1211](https://github.com/sarthaksingh1211)

## EDUCATION

### Stony Brook University

Master of Science in Computer Science

### NMIMS MPSTME

Bachelor of Technology in Computer Engineering, GPA: 3.6/4.0

Stony Brook, NY

Aug. 2024 – May 2026

Mumbai, India

Aug. 2020 – May 2024

## EXPERIENCE

### Cloud Computing Intern

Vodafone Idea Ltd.

May 2023 – Aug. 2023

Mumbai, India

- Architected cloud-native ML infrastructure using Docker and AWS EC2, cutting operational costs by \$45K annually (15%) through automated resource provisioning and containerization strategies
- Orchestrated migration of 12 production applications to AWS cloud platforms, achieving 99.8% uptime and eliminating 20% downtime incidents while maintaining zero data loss across 500GB+ datasets
- Engineered auto-scaling solutions with CloudWatch and Lambda functions, improving system throughput by 25% and handling 10K+ concurrent requests with 40% faster response times
- Accelerated CI/CD deployment velocity by 35% implementing Jenkins pipelines and Infrastructure-as-Code (Terraform), reducing release cycles from 2 weeks to 3 days for ML model deployments

### Machine Learning Research Engineer

NMIMS University

Dec. 2023 – May 2024

Mumbai, India

- Engineered end-to-end music recommendation system leveraging ConvLSTM and ResNet-50 architectures for real-time facial emotion detection, achieving 89% F1-score across 7 emotion classes on 15K+ annotated images
- Optimized deep learning pipeline with TensorFlow and OpenCV, reducing inference latency from 450ms to 85ms (81% improvement) enabling real-time processing at 30 FPS on CPU-only environments
- Deployed production-ready Flask API serving 2,500+ daily predictions with 97% recommendation relevance score, increasing user session duration by 42% and playlist completion rates by 38% through A/B testing validation

## PROJECTS

### Enterprise RAG System with Multi-Document QA | LangChain, OpenAI GPT-4, Pinecone, FastAPI Sep. 2024 – Dec. 2024

- Built production RAG system processing 10K+ documents achieving 92% accuracy using FAISS vector indexing and GPT-4 for context-aware QA
- Implemented hybrid search (semantic + BM25) achieving 0.87 precision@5 and reducing hallucination rate from 18% to 4% via citation tracking
- Deployed FastAPI microservice handling 1,200+ queries/hour at 240ms latency with Redis caching cutting costs by \$800/month (45%)

### Agent-Based MLOps Platform | Python, Streamlit, XGBoost, SHAP, LIME, PyDantic

Oct. 2024 – Dec. 2024

- Architected autonomous ML framework with 5 AI agents executing 50+ Optuna trials achieving 92% AUROC on 200K+ healthcare samples
- Integrated explainable AI (SHAP/LIME) dashboards validating fairness metrics and reducing algorithmic bias by 15% for compliance
- Delivered Streamlit application processing 20+ datasets with automated EDA, cutting setup time by 60% tracking 500+ MLflow experiments

### Neural Radiance Fields (NeRF) 3D Reconstruction System | Python, PyTorch, JAX, CUDA, NumPy Jan. 2024 – May 2024

- Engineered GPU-accelerated NeRF pipeline with volumetric ray marching achieving 28.5 dB PSNR and 0.92 SSIM using FP16 mixed-precision
- Optimized 3D reconstruction with COLMAP SfM and Poisson algorithms supporting PLY/OBJ/FBX, improving throughput 40% (21 FPS)
- Implemented camera pose estimation processing 100+ multi-view images at 95% accuracy using SIFT/RANSAC for AR/VR synthesis

## PUBLICATIONS

Movie Recommendation Using Voice Analysis – AIP Conference Proceedings, 2024 (DOI: 10.1063/5.0234460)

Music Recommendation Using Facial Emotion Recognition – Book Chapter, 2024 (ISBN: 978-93-91535-66-7)

## TECHNICAL SKILLS

**Languages:** Python, C++, Go, SQL, R, JavaScript, HTML/CSS, Bash/Shell Scripting

**ML/DL Frameworks:** PyTorch, TensorFlow, Keras, JAX, Hugging Face Transformers, scikit-learn, XGBoost, LightGBM

**LLM & Gen AI:** LangChain, LlamaIndex, OpenAI API, RAG, Fine-tuning (LoRA/QLoRA), Prompt Engineering, Vector Databases (FAISS, Pinecone, ChromaDB)

**Computer Vision & NLP:** OpenCV, YOLO, ResNet, ViT, BERT, GPT, LSTM, Transformers, spaCy, NLTK, Librosa

**MLOps & Development:** Docker, Kubernetes, Git, Jenkins, MLflow, Weights & Biases, DVC, Airflow, FastAPI, Flask, Streamlit, React, Node.js

**Cloud & Databases:** AWS, GCP, PostgreSQL, MongoDB, Redis, Spark, pandas, NumPy

**Software Engineering:** REST APIs, Microservices, CI/CD Pipelines, System Design, Distributed Systems, Algorithms, Data Structures