Sarthak Singh

+1 (631)-355-8065 | sarthaksingh1211@gmail.com | linkedin.com/in/sarthaksingh1211 | github.com/SarthakSingh-96

Education

Stony Brook University — Stony Brook, NY
Master of Science in Computer Science — *Aug. 2024 – May 2026*
- **Coursework:** Distributed Systems, Database Systems, Machine Learning | **Teaching Assistant:** Object-Oriented Programming

NMIMS MPSTME — Mumbai, India
Bachelor of Technology in Computer Engineering, GPA: 3.6/4.0 — *Aug. 2020 – May 2024*

Experience

Cloud Computing Intern — May 2023 – Aug. 2023
Vodafone Idea Ltd. — *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 — Dec. 2023 – May 2024
NMIMS University — *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

CodeWeb - VS Code Extension | *TypeScript, VS Code API, Node.js* | Marketplace — Feb. 2026
- Architected real-time dependency analysis engine processing 1,000+ code references per second with interactive graph-based UI visualizing cross-file impact
- Engineered automated type classification system achieving 95% accuracy categorizing code components using VS Code Language Server Protocol
- Deployed click-to-navigate visualization panel reducing code navigation time by 40% and accelerating developer productivity

Technical Skills

Languages: Python, C++, Go, TypeScript, JavaScript, SQL, R, 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, VS Code Extension Development, System Design, Distributed Systems, Algorithms, Data Structures