Deep Vivek Sheth

(213) 681-2323 | deepsheth3@gmail.com | linkedin.com/in/deepsheth3 | github.com/deepsheth3

Education

University of Southern California Los Angeles, CA
Master of Science in Computer Science Jan 2024 – Dec 2025

- Coursework: Algorithms, Machine Learning, Parallel and Distributed Systems, Modern Database Architectures

Vishwakarma Institute of Technology Pune, India
Bachelor of Technology in Computer Engineering Aug 2018 – May 2022

- Coursework: Data Structures and Algorithms, Operating Systems, Computer Networks, Cloud Computing, Data Mining

Technical Skills

Languages: Python, C++, C, SQL, Bash, MATLAB, Cypher, Shell Scripting, R, Java, CUDA (Parallel Programming)
ML/AI Frameworks: PyTorch, TensorFlow, Keras, Transformers, scikit-learn, XGBoost, Generative AI, RAG, OpenCV
Data & MLOps: AWS (SageMaker, S3, Lambda, API Gateway), MLflow, ChromaDB, FAISS, Neo4j, MongoDB, Airflow
Developer Tools: Git, Docker, Kubernetes, Linux, REST APIs, FastAPI, VS Code, JupyterLab, Vim, Postman, CI/CD

Experience

Cadence Design Systems San Jose, CA
Machine Learning Engineer Intern Jun 2025 – Aug 2025

- Engineered a **GPU-accelerated** compute kernel pipeline, optimizing high-throughput data flows to boost analysis speed
- Designed **scalable log templating algorithms**, improving pattern extraction accuracy, enabling faster root-cause analysis
- Architected a **RAG** system with **ChromaDB** and **LangChain**, optimizing **retrieval evaluation** metrics by 35%
- Performed **LLM Optimization** on internal models using 10M+ logs, achieving high accuracy for real-time queries

Persistent Systems Pune, India
Software Engineer Jul 2022 – Nov 2023

- Engineered scalable **backend systems** on **AWS SageMaker** managing real-time traffic for 220K+ users with 99.9% uptime
- Architected **REST APIs** and **CI/CD pipelines** using **AWS Lambda**, reducing deployment cycles by 40% for ML services
- Built a deep learning **CNN** for lung cancer detection on 9K images, cutting false positives by 62% and diagnosis time by 70%
- Integrated multilingual **BERT** with **Elasticsearch** and **Kibana** to streamline search across multi-region data pipelines
- Built **Linear Regression** models to detect shipping logistics trends and predict customer retention, improving accuracy 80%

Persistent Systems Pune, India
Software Engineer Intern Feb 2022 – Jul 2022

- Implemented **autoencoder-based** anomaly detection achieving F1 score of 0.9, strengthening network threat identification
- Co-authored and published a comparative **ML** study on intrusion detection, enhancing model clarity and overall performance
- Optimized **SQL** schema design and **backend logic**, streamlining data flow and reducing query execution latency by 40%

Centre for Development of Advance Computing Delhi, India
Software Engineer Intern Jun 2021 – Jan 2022

- Optimized a large-scale pipeline using **BERT** and **Neo4j**, reaching 92% accuracy on COVID-19 **knowledge graph** queries
- Improved data throughput by parallelizing **PySpark ETL** jobs, reducing preprocessing time by 35% across large corpora
- Extended the **NLP** pipeline with **Transformers** and fuzzy logic for healthcare, boosting accuracy and efficiency by 13%

Projects

Oddesey: AI-Powered Travel Itinerary & Route Optimizer Jan 2026 – Present
Python, FastAPI, Next.js, OpenAI GPT-4, Google Maps API, Pinecone

- Built a production-grade planner using **Advanced Prompt Engineering** and **LLM Orchestration** for reliable outputs
- Integrated **Google Places API** and **Distance Matrix** to fetch location metrics, ensuring geographically viable schedules
- Designed multi-step **AI Agent workflows** (ReAct pattern) to orchestrate complex tool use for real-world trip planning tasks

Bloom Filter-Aided Hash Join Optimization Feb 2025 – Mar 2025
C++, DuckDB, TPC-H, Bloom Filters

- Implemented a multilevel **Bloom filter** within **DuckDB's** hash join operator to prefilter 100K+ probe-side keys per query
- Reduced redundant lookups by 45% while improving join selectivity using adaptive probabilistic data structures
- Benchmarked the optimized operator on **TPC-H** workloads and large real datasets for reproducible performance validation
- Achieved up to 2.5x query speedup on selective joins through advanced memory-aware precomputation and hash partitioning

Parallel PageRank Algorithm Optimization Oct 2024 – Dec 2024
C++, CUDA, NVIDIA A100 GPU

- Optimized large-scale PageRank compute on **NVIDIA A100 GPU** using data-parallel execution and **kernel acceleration**
- Improved kernel time from 9.3s to 1.3s through tuned **parallel thread scheduling** & optimized **shared-memory**
- Increased compute throughput by 2x using advanced **memory tiling** methods and **parallel edge traversal** optimization
- Profiled and refined **compute workloads** to maintain stable performance across diverse **graph structures** and depths