Rithesh **Kumar**

516 4th Avenue, San Francisco, CA, USA

(+1) 831-266-7201 | rkumar45@ucsc.edu | rithesh17.github.io | Rithesh17

Summary

Founding Engineer building cloud infrastructure for AI agents, with expertise in production ML systems handling 1M+ requests monthly. Master's in Computer Science focused on medical imaging and AI explainability. Experience at Goldman Sachs and Accretional in building reliable systems at scale, with particular interest in bridging the gap between AI capabilities and real-world adoption through explainability and trust.

Work Experience

Founding Engineer - Infrastructure & Platform

Aug. 2025 - Present

Accretional — San Francisco, CA

- Built end-to-end website publishing system by automating domain registration, DNS, and CDN configuration with retry logic, resulting in one-click deployment that eliminates infrastructure complexity for users
- Implemented semantic search in data layer by combining vector similarity with full-text search, enabling AI agents to find code by meaning and improving query accuracy by 40% over keyword-only search
- Redesigned workflow system using file-based persistence, making prompts version-controllable and shareable, resulting in 3x faster team collaboration on AI-assisted coding workflows
- Developed verification system using hCaptcha and signed tokens, reducing authentication friction by 60% while maintaining security standards
- Owned full-stack systems from database internals to deployment pipelines, reducing deployment time by 50% through infrastructure optimization

AI Engineering Intern

Jun. 2024 - Aug. 2024

Accretional — San Francisco, CA

- Built RAG pipeline by retrieving documentation and code examples before generation, reducing hallucinations by 35% compared to base model outputs
- Applied LoRA fine-tuning to adapt LLMs for code generation, improving code accuracy by 25% on domain-specific tasks without full model retraining
- Rebuilt prompt system in VS Code extension and integrated deployment flow, enabling users to deploy generated APIs to cloud functions in under 30 seconds

Machine Learning Scientist

Jul. 2020 - Jul. 2023

Goldman Sachs — Bengaluru, India

- Led real-time loan processing pipeline by implementing caching and circuit breakers, handling 1M+ requests per month with 99.9% uptime and graceful degradation under load
- Developed LSTM models for loan repayment forecasting and optimized inference pipelines, achieving 15% improvement in prediction accuracy and reducing latency by 60% from 200ms to 80ms through model quantization and caching
- Introduced consistent CI/CD processes across ML team and built A/B testing framework, reducing deployment time by 70% and enabling data-driven model rollouts that reduced risk by 80%
- Built simulation models to assess COVID-19 impact on lending risk, enabling business to make data-driven decisions during unprecedented scenarios when historical patterns were unreliable

Machine Learning Research Intern

May 2019 - Jul. 2019

Sprinklr — Gurgaon, India

- Built sentiment analysis pipelines for social media content by processing multilingual and fragmented text, achieving 85% accuracy on noisy real-world data
- Applied model optimization techniques including pruning and quantization, reducing model size by 70% while maintaining 95% of original accuracy for production deployment

Education

Master of Science in Computer Science

Sep. 2023 - Jun. 2025

University of California, Santa Cruz — Santa Cruz, CA

- GPA: 3.9/4.0
- Focused on machine learning and AI, particularly medical imaging applications with emphasis on clinical trust and explainability
- Research in CT scan denoising and AI explainability, developing methods that improve image quality by 30% while maintaining diagnostic accuracy
- Explored attention mechanisms and visualization techniques, creating interpretable models that increase confidence by 60%

Bachelor of Technology in Computer Science Aug. 2016 - May 2020

National Institute of Technology Karnataka, Surathkal Surathkal, India

- GPA: 9.4/10.0, Graduated with Honors
- Focused on ML and AI, with research in medical image analysis resulting in published work on automated cancer grading
- Published work on automated prostate cancer grading achieving 92.38% accuracy, contributing to Springer proceedings
- Collaborated with pathologists to identify key diagnostic features, designing models that achieved 85% agreement with expert assessments

Projects

Go, C++, CGO, OpenVINO Jan. 2025

OpenVINO Go Bindings GitHub

- Built idiomatic Go bindings for Intel's OpenVINO Runtime using three-layer architecture, enabling Go developers to leverage OpenVINO without C++ expertise
- Implemented async inference with state management for recurrent models, reducing inference latency by 40% compared to synchronous implementations
- Delivered production-ready library with 90%+ test coverage, automated CI pipeline, and comprehensive documentation, resulting in 50+ GitHub stars

Python, LLMs, Conversational AI, NLP May 2024

Whispers of the Heart: AI Therapy Assistant GitHub

- Built conversational AI system for journaling and therapeutic assistance using LLMs, processing 1000+ user sessions with 90% user satisfaction
- Designed journaling component with guiding questions and pattern identification, helping users identify emotional patterns with 75% accuracy
- Implemented end-to-end encryption and secure conversation handling, ensuring HIPAA-compliant data protection and user privacy controls

Scala, Hardware Design, DNA Sequencing Dec. 2023

GeneWeaver GitHub

- Created parametric hardware generator for DNA sequence alignment using Scala, generating optimized designs 10x faster than manual RTL coding
- Built flexible system generating optimized FPGA or ASIC designs, reducing hardware development time by 60% through automated design space exploration
- Designed modular architecture enabling extension for new alignment algorithms, supporting 5+ different algorithms with single codebase

Publications

Machine Learning, Image Processing, Network Security and Data Sciences 2023

Prostate Cancer Grading Using Multistage Deep Neural Networks Springer

- Proposed three-stage ensemble deep learning method for automatic prostate cancer grading, achieving 92.38% classification accuracy on 10,000 histopathological images
- Developed pipeline using UNet segmentation followed by ensemble of Xception, ResNet-50, and EfficientNet-b7, outperforming single-model baselines by 8%, and integrated GradCAM visualizations for model explainability
- Trained on PANDA challenge dataset from Karolinska and Radboud, validating method on diverse histopathological image sources

Security in Computing and Communications (SSCC 2018) 2019

Network Anomaly Detection Using ANNs Optimised with PSO-DE Hybrid Springer

- Developed network anomaly detection system using ANNs optimized with hybrid PSO-DE algorithm, achieving 94% detection accuracy on network intrusion datasets
- Addressed network monitoring needs in ubiquitous computing by reducing false positive rate by 30% compared to traditional signature-based methods
- Proposed optimization approach combining PSO and DE algorithms, improving ANN training convergence by 25% and detection performance by 12%

Skills

Programming Languages	Python, Go, C++, TypeScript, JavaScript, Scala
Machine Learning & AI	PyTorch, LLMs, Computer Vision, NLP, RAG, Model Optimization, LoRA/PEFT, Vector Search
Cloud & Infrastructure	AWS, Docker, CI/CD, Google Cloud, Cloudflare, gRPC, Envoy
Web Development	SvelteKit, React, HTML/CSS, Three.js
Data & Storage	SQLite, PostgreSQL, Redis, NumPy/Pandas