SUHANI KHANNA

Backend Software Engineer | Machine Learning Systems
https://github.com/suhanikhanna31 | https://www.linkedin.com/in/suhani-khanna-673936284/ | suhanikhanna31@gmail.com | +91 7428462400

Professional Summary

Backend-focused software engineer with strong foundations in Data Structures & Algorithms and Object-Oriented Programming, experienced in building scalable distributed systems and production ML pipelines. Skilled in REST APIs, microservices architecture, and cloud deployment, with hands-on experience in model deployment, system design, and performance optimization.

Technical Skills

Languages: Python, Java, SQL
Computer Science Fundamentals: Data Structures & Algorithms, Object-Oriented Programming (OOP), Time & Space Complexity, Design Patterns
Mathematics: Probability & Statistics, Linear Algebra, Optimization (applied in ML model evaluation, anomaly detection, and system modelling)
Backend & Systems: REST APIs, Microservices Architecture, Distributed Systems, Concurrency & Multithreading, API Security (JWT, OAuth2), Rate Limiting, Caching (Redis), Load Balancing
ML & AI: Supervised & Unsupervised Learning, Scikit-learn, Feature Engineering, Data Preprocessing, Model Evaluation (Precision, Recall, F1-score, ROC-AUC), Hyperparameter Tuning, Anomaly Detection, NLP, Generative AI, Production ML Systems
Databases: PostgreSQL, Redis, Query Optimization, Indexing
Cloud & DevOps: AWS (EC2, S3, Lambda, API Gateway, CloudWatch), OCI, Docker, Kubernetes (Basic), CI/CD (GitHub Actions), Serverless Architecture
MLOps: Model Deployment, Model Serving, Pipeline Orchestration, Experiment Tracking, Monitoring
Tools: Git, GitHub

Education

Galgotias College of Engineering and Technology (affiliated with Dr. A.P.J. Abdul Kalam Technical University) – Greater Noida, UP
B.Tech., Information Technology | July 2027 | CGPA: 7.82

Projects & Technical Experience

Distributed Rate Limiting System (Java, Spring Boot, Redis)

- Designed a scalable distributed rate limiting system using token bucket algorithm, handling high-throughput API traffic with low latency.
- Implemented centralized caching with Redis and documented rate-limiting policies and API behaviors to ensure consistent integration across microservices
- Designed stateless microservices architecture and defined clear service contracts/interfaces to enable scalable and maintainable integrations
- Integrated adaptive request filtering to detect anomalous traffic patterns and dynamically throttle abusive clients.

DecisionForge – ML-Driven Backend Decision Engine with AI Agents (Python, Scikit-learn)

- Built a production-oriented ML pipeline and exposed well-documented REST APIs for seamless integration of real-time decision services.
- Implemented supervised learning models for churn prediction and anomaly detection with structured feature engineering and preprocessing.
- Applied model evaluation metrics (Precision, Recall, F1-score) and offline validation to optimize decision accuracy.
- Designed a low-latency decision engine and articulated system design trade-offs (latency vs accuracy) for scalable deployment

Transaction Anomaly Notifier (Python, PostgreSQL, REST APIs, AWS, MLOps)

- Developed a real-time fraud detection system and structured API responses for clear interpretation of risk scores by downstream services
- Deployed on AWS (EC2, S3, Lambda, API Gateway) with serverless components for scalable and reliable processing.
- Designed a rule-based alerting engine reducing false positives while maintaining high recall in anomaly detection.
- Optimized PostgreSQL database performance using indexing and query optimization for fast retrieval of audit logs.
- Optimized PostgreSQL database performance using indexing and query optimization for fast retrieval of audit logs.

Certificates

Deloitte: Virtual Work Experience (Coding, Development)
Oracle: Oracle Cloud Infrastructure AI Foundations Associate
IBM SkillsBuild: Artificial Intelligence Fundamentals