Sree Bhargavi Balija
Software Engineer

CA • sreebhargavi576@gmail.com • **(408) 780-8902** • https://www.linkedin.com/in/sree-bhargavi-balija-b7638517a/

SUMMARY

- Software Engineer and ML Engineer with over **3+ years of experience** in designing, deploying, and scaling machine learning systems, backend microservices, and cloud-native applications across finance and enterprise automation domains.
- Hands-on experience building production-ready solutions using **Python, Java, SQL, TensorFlow, XGBoost, Langchain, Kubernetes, Azure, and GCP**, delivering end-to-end systems from data ingestion to model deployment and monitoring.
- Specialized in **financial data modeling, NLP systems, and federated learning**; delivered projects that reduced operational workload, improved prediction accuracy, and supported over 100K+ users in real-world applications.
- Strong background in **REST API development, microservice architecture, MLOps pipelines, and real-time analytics**; optimized systems for performance, scalability, and security through tools like **CI/CD, PyTorch, Elasticsearch, and Django**.

PROFESSIONAL EXPERIENCE

Machine Learning Engineer **USA**
Akdene Technologies **Mar 2025 – Current**

- Deployed a predictive financial model using Python, TensorFlow, and XGBoost in months, enhancing financial risk analysis for over 100K customers.
- Built a customer analytics system with SQL, Pandas, and Langchain, reducing data retrieval time from 30 minutes to under 5 minutes for financial reporting teams.
- Engineered an Agentic-AI based chatbot using NLTK, OpenAI APIs, and Django, reducing customer service workload by 120 hours per month and improving response time.
- Automated ML model deployment with Azure, CI/CD, and Kubernetes, reducing deployment time from 2 days to just 6 hours, ensuring faster production releases.
- Enhanced fraud detection with PyTorch and model interpretability techniques, identifying 1,200+ high-risk transactions in the first weeks of deployment.
- Integrated REST APIs with microservices to streamline financial data processing, reducing manual intervention and improving system efficiency.
- Formulated a model monitoring pipeline using Conformal Predictions, ensuring 75% stability in model accuracy across financial datasets.

Software Engineer **India**
ServiceNow **Jun 2020 – Aug 2022**

- Deployed a scalable backend microservice using Java, Spring Boot, and REST APIs within 6 months, improving data processing speed by 3x for enterprise clients.
- Developed a real-time data visualization dashboard using JavaScript, , and Elasticsearch in 5 months, enabling real-time monitoring for over 500K events per day.
- Implemented a secure authentication system using OAuth2, API Development, and system integration, reducing unauthorized access incidents by 60% in the first year.
- Accelerated cloud-based deployment on GCP, reducing infrastructure costs by 25% while maintaining 79% uptime.
- Designed a robust testing framework using JUnit automating over 500 test cases, reducing production bugs and improving system stability.
- Led an Agile team of 6 developers, improving sprint delivery time from 3 weeks to 2 weeks, ensuring faster feature rollouts.
- Optimized API performance using Kotlin and load-testing tools, improving request handling efficiency by 40%, reducing latency in high-traffic periods.
- Wrote and maintained technical documentation for enterprise solutions, reducing onboarding time for new engineers by 50%.
- Enhanced system design architecture, implementing structured logging and improving debugging efficiency for 200+ microservices.
- Collaborated with cross-functional teams to integrate customer feedback into backend architecture, reducing feature request turnaround time from 3 weeks to 1.5 weeks.

Software Engineer Intern **India**
Dell Technologies **Nov 2019 - May 2020**

- Contributed to the development of internal web tools using Python (Flask) for backend services and HTML5, CSS3, and JavaScript for frontend interfaces.
- Built responsive UI components using Bootstrap and JavaScript, enabling real-time interaction with internal APIs for system monitoring tools.
- Assisted in creating lightweight microservices and RESTful APIs using Flask, facilitating data exchange between backend systems and frontend dashboards.
- Debugged and enhanced existing frontend features across multiple web utilities, improving cross-browser compatibility and user experience.
- Wrote modular Python code with proper exception handling and logging, increasing maintainability and reducing runtime errors across internal systems.
- Integrated backend scripts with SQL queries for real-time data reporting, supporting operational analytics and resource monitoring tools.
- Used Git for version control and worked within a shared development environment, contributing to sprint deliverables and participating in code reviews.

Transformer Optimization for CHAI Model

Researcher – US Meta Research Team **Apr 2024**

- Integrated hybrid sparse attention and targeted fine-tuning across 3 domain-specific datasets, boosting task accuracy by 15% and halving inference time.
- Developed memory-efficient transformer pipelines, reducing GPU usage by over 4GB per training cycle and enabling deployment on low-resource systems.
- Applied clustering algorithms to optimize token embeddings, cutting pre-processing computation by 40% and accelerating model training workflows.

EDUCATION

Master of Science in Machine Learning and Data Science
University of California, San Diego, CA ***Jun 2024***

TECHNICAL SKILLS

- **Programming Languages** Python (Expert), Java, JavaScript, C++, R, Fortran, Prolog, Perl, Kotlin, Swift
- **Machine Learning & AI** Machine Learning, Deep Learning, NLP, Federated Learning, Transformers, Multimodal AI, Model Interpretability, Conformal Predictions
- **AI/LLM Tools** NLTK, Langchain, OpenAI, Google Gemini
- **ML Libraries & Frameworks** Scikit-learn, TensorFlow, Keras, XGBoost, PyTorch, Django, Langchain
- **Data Analysis & Visualization** Pandas, NumPy, Matplotlib, Seaborn, Dashboards, Pre-built Analytics Metrics
- **Cloud & DevOps** Azure, GCP, Kubernetes, Git, CI/CD, Technical Documentation
- **Databases & Data Tools** SQL, Elasticsearch, ETL
- **API & Backend Development** REST APIs, API Development, Java APIs, Microservices, System Integration
- **Development Practices** Agile, System Design, Testing, Model Monitoring, MLOps

PUBLICATIONS

- Building Communication Efficient Peer-to-Peer Federated LLMs with Blockchain, AAAI, Stanford University
- CPTQuant - A Novel Mixed Precision Quantization Techniques for Large Language Models
- FedNAM+: Executing Interpretability Analysis using Novel Conformal Predictions method, CVPR 2025 (In review)
- FedNAM: Executing Interpretability Analysis in Federated learning Context
- AILUMINATE: Introducing v1.0 of the AI Risk and Reliability Benchmark from MLCommons

AWARDS AND ACHIEVEMENTS

- Skill development incentive program award, ServiceNow 2021
- Academic excellence award, IIT Hyderabad 2018
- Founder of AutoPatch+, presented at MIT AI summit 2025
- Developed Multimodal small LLM of size 125M parameters