Shriyansh Singh

+1-930-333-5141 | **shriyansh.singh24@gmail.com** | **LinkedIn**

SUMMARY

ML Engineer specializing in **high-throughput systems** and **optimization algorithms**, passionate about leveraging **machine learning** to revolutionize **autonomous vehicle** operations

EDUCATION

Indiana University Bloomington Aug 2023 – May 2025
Master of Science in Data Science *Indiana, United States*

PROFESSIONAL EXPERIENCE

Machine Learning Systems Engineer Intern *April 2024 - Dec 2024*
Hyphenova AI *Los Angeles, CA*

- **Pioneered** a **high-throughput inference system** that processed 10,000+ concurrent requests by implementing **optimization algorithms** for resource allocation, resulting in 40% latency reduction
- **Constructed** an end-to-end **predictive model pipeline** leveraging **distributed computing** techniques in **Go** and C++, achieving 99.9% reliability for critical production services
- **Formulated** novel **batch processing algorithms** that dynamically optimized compute resources based on traffic patterns, boosting system throughput by 30% without hardware upgrades
- **Orchestrated** a scalable deployment infrastructure using Kubernetes that reduced time-to-production from days to hours while maintaining consistent testing environments
- **Designed** a comprehensive real-time monitoring system that enabled data-driven decisions for proactive **optimization** of system performance across geographically distributed services

Machine Learning Infrastructure Intern *May 2022 - Oct 2022*
Enterprise Business Technologies Pvt Ltd *Mumbai, India*

- **Architected** a **real-time optimization engine** for **autonomous resource allocation** that reduced processing latency by 65% through intelligent workload distribution algorithms
- **Implemented** adaptive **machine learning inference systems** for resource-constrained environments, enabling efficient operation with minimal computational overhead
- **Developed** a priority-based **scheduler algorithm** in **Python** and **Go** that maximized throughput by dynamically allocating resources based on business-critical metrics
- **Established** asynchronous processing patterns for machine learning pipelines that maintained strict SLAs while increasing overall system capacity by 40%

PROJECTS

Autonomous Vehicle Routing Optimization System | *Go, Python, Apache Spark, TensorFlow* *Jan 2024 – Apr 2024*

- **Engineered** a machine learning platform for **autonomous vehicle** route **optimization** that reduced idle time by 27% through predictive demand modeling
- **Devised** a multi-objective **optimization algorithm** balancing vehicle availability, energy consumption, and user wait times across geographically distributed resources
- **Implemented** a real-time decision engine processing 1000+ requests/second with sub-100ms latency using **Go** and advanced in-memory computational techniques
- **Integrated** live traffic and environmental data streams to dynamically adjust routing decisions, improving overall fleet efficiency by 18%

High-Throughput Prediction System for Time-Sensitive Applications | *Apache Spark, Python Sep 2023 – Dec 2023*

- **Created** a **distributed prediction system** for time-sensitive applications that maintained 99.95% availability while serving 8000+ daily inference requests
- **Optimized** inference latency for complex neural network models through custom parallelization techniques, achieving 65ms p95 response times
- **Designed** an adaptive batching algorithm that maximized computational resource utilization while meeting strict SLA requirements
- **Built** comprehensive A/B testing infrastructure to rapidly iterate on model improvements without service disruption

SKILLS & CERTIFICATIONS

Programming: Go, **Python**, C/C++, Shell Scripting, SQL
Systems: High-Throughput Computing, **Optimization Algorithms**, **Distributed Systems**, **Parallel Processing**
ML & Data: Machine Learning Pipelines, **Predictive Modeling**, **Large-Scale Data Processing**, **Apache Spark**
Technologies: Kubernetes, Docker, **Real-time Systems**, **Monitoring Tools**, Cloud Computing