

Mohammad Alhan Bellary

MLOps & AI Deployment Engineer

Bengaluru, India | +91 8867634490 | alhanbellary123@gmail.com | linkedin.com/in/alhan-bellary | github.com/alhaannn | alhan.me

PROFESSIONAL SUMMARY

MLOps & Platform Engineer pairing algorithmic foundations (NeetCode 150) with automated CI/CD, Kubernetes orchestration, and model observability. Built drift-monitored ML pipelines and inference services with MLflow, DVC, Prometheus, and Grafana.

TECHNICAL SKILLS

Containers & Orchestration: Docker, Kubernetes (Deployments, Services, HPA), Linux/Bash

MLOps Tools & Registries: MLflow (Tracking, Model Registry), DVC, Evidently AI, ONNX Runtime

CI/CD & Testing: GitHub Actions, Jenkins, Ansible, Automated Rollbacks, PyTest, Linting Gates

Monitoring & Serving: Prometheus, Grafana, Alertmanager, Python 3.13, FastAPI, Redis, Celery, PostgreSQL

CORE PROJECTS

Production MLOps Anomaly Detection & Deployment Pipeline | *DVC, MLflow, Docker, Kubernetes, Prometheus, Grafana, Evidently AI*

- Automated model deployment, cutting turnaround from 3 days to under 12 minutes at 99.9% uptime, by building a DVC + MLflow Registry pipeline deployed as multi-stage Docker images on Kubernetes with HPA and zero-downtime rolling updates.
- Instrumented drift governance, as measured by Prometheus/Grafana dashboards, by integrating Evidently AI to track feature distribution drift and trigger automated retraining alerts.

High-Throughput AI Gateway & Model Routing Engine | *FastAPI, Celery, Redis, PostgreSQL, GitHub Actions*

- Sustained 250+ concurrent inference requests at sub-50ms queue latency, by offloading workloads to Celery/Redis queues with dynamic fallback routing and retry policies.
- Slashed redundant model compute by 28%, as measured by cache hit rates, by implementing Redis semantic caching and token-bucket rate limiting, gated by PyTest coverage above 80% and GitHub Actions.

Quantized Computer Vision Edge Deployment System | *PyTorch, YOLOv8, ONNX Runtime, TensorFlow Lite, FastAPI, Docker*

- Cut container footprint by 62% and reached 42 FPS on CPU-only hardware at 94.2% mAP, by quantizing YOLO checkpoints with ONNX Runtime and TensorFlow Lite in a containerized FastAPI service.
- Automated deployment verification, as validated by inference latency SLAs, by adding GitHub Actions gates before production promotion.

PROFESSIONAL EXPERIENCE & TRAINING

AI Technocrat Industry Training (Mandatory OJT, 120 Hours) | IIT DHARWAD | Dharwad | Sep 27 – Oct 22, 2027

- Engineered Jenkins and GitHub Actions CI/CD pipelines with automated rollback triggered by anomalous deployment health telemetry.
- Containerized Python inference services with Docker and orchestrated multi-pod Kubernetes clusters with HPA.
- Configured Prometheus scrapers and Grafana dashboards monitoring p50/p95/p99 latency, error rates, and CPU/memory utilization.

Junior Software Developer (Internship), Heagl Technologies Pvt Ltd | Hubballi | Feb – May 2026

- Delivered 5 features across 2 production Django codebases, cutting feature turnaround by ~25%, by collaborating with senior developers.
- Resolved 15+ bugs, reducing debugging time by ~30%, by diagnosing defects with AI coding tools (Claude Code, Cursor).

EDUCATION & CERTIFICATIONS

AI Technocrat, NSQF Level 5.0 Professional Certification | IIT DHARWAD | Oct 22, 2026 – Oct 22, 2027

BCA, Information Technology, IBMR Business School | Hubballi | Sep 2023 – Jul 2026 | CGPA 7.66/10

Certifications: Oracle Certified Data Science Professional, Oracle, 2025 | AI Literacy, IBM, 2025 | Certified Product Manager (CPM), GeeksforGeeks

Awards: 1st Place, CheTech Hackathon (Chetan College), 2026 | Empathy Encryption Hackathon (UnsaidTalks), 2025