

Abhishek Srivastava

abhisheksrivastava05102004@gmail.com | +91-8076170946
LinkedIn | GitHub | Portfolio

Education

Indian Institute of Technology Kanpur

2022 – 2026

B.Tech, Mechanical Engineering | Minor in Machine Learning & Applications

Relevant Courses: Introduction to Machine Learning, Statistical Natural Language Processing, Data Structures & Algorithms

Scholastic Achievements

- Secured **AIR 2889** in JEE Advanced and **AIR 2306** in JEE Mains 2022 among **1.027 million** candidates
- Secured **AIR 122** in NEST 2022 and **AIR 387** in IISER Aptitude Test 2022 among **22,000+** candidates

Experience

Junior Machine Learning Engineer | Arivihan Technologies Pvt. Ltd., Indore

Mar 2026 – Sep 2026

- Built a production **multi-agent LLM tutor** with **6 agents**, **30+ tools**, **RAG**, using **GPT-5.4/5.4-mini** and **Gemini 3.1 Flash**
- Engineered **hybrid agent memory** using summarization and **embedding retrieval** for persistent context within user sessions
- Optimized **LLM inference costs** by migrating from **GPT-5.4** to **GPT-5.6-Luna**, reducing **token costs by 75%**
- Built a fault-tolerant **AI serving pipeline** with **FastAPI**, **Redis**, **DynamoDB**, **SQS/DLQ**, supporting **100K concurrent users**
- Optimized live-chat analytics from **1,500** to **1 DynamoDB query/class**, increasing AI-category coverage from **4%** to **100%**
- Built a **complaint and feedback agent** with **FastAPI live retrieval** and **RAG over OpenAI Vector Stores** for query resolution
- Built an automated **call-auditing pipeline** using **Soniox STT** for transcription and **GPT-5.4-mini** for abuse detection, and audit reports, automating **S3-to-DynamoDB** processing via a scheduler with frontend report rendering

Artificial Intelligence Systems Engineer Intern | InfraDock AI, Delhi

Sep 2025 – Feb 2026

- Built parsing and ingestion pipelines for construction **DPRs**, **Work Orders** and **BOQs** supporting **92%** field accuracy
- Modelled entities into a **work-graph** for **Graph RAG**, improving multi-hop query accuracy **35%** over vector-only
- Added layout-aware chunking, dedup and incremental re-indexing, cutting ingestion per project pack to **under 10 min**
- Revamped the analyst frontend with conversational querying and citation drill-down to the source BOQ row

Product Engineering Intern | UnifyApps, Gurugram

May 2025 – Jul 2025

- Engineered scalable **RAG retrieval** services and ingestion pipelines for enterprise PDF, image and Excel sources
- Designed **RBAC-enabled agentic indexers** with semantic filters, cutting query latency **25%** to **sub-200ms**
- Shipped **150+ API integrations** (Java, Groovy) and evaluated OCR agents, achieving **15%** higher extraction accuracy

Projects

Agentic AI Tutoring Question Generator | Self Project

Jan 2026 – Feb 2026

- Built a **LangGraph** planner-first multi-agent system emitting structured JSON execution plans per request
- Added vision, solver and evaluator agents with **Pydantic** typed state for deterministic, testable workflows

Vision-Driven Adaptive Traversability for Legged Robots | Mobile Robotics Lab, IIT Kanpur

Nov 2024 – Aug 2025

- Developed a **vision-proprioception fusion** framework for terrain-aware autonomous navigation of **Unitree Go2**
- Built a **ResNet-50** image reliability estimator using brightness and blur metrics, improving robustness **25%**
- Trained **GaNAV** on **8,000+** RUGD images and deployed a **ROS2** terrain-aware DWA planner for real-time navigation

Industrial Coal Pellet Inspection (Computer Vision) | IndustANTRA, IIT Kanpur

May 2024 – Aug 2024

- Annotated **3,000+** images (CVAT, Fiji) and trained **YOLOv8** and **U-Net** models in PyTorch to **92%** accuracy
- Deployed the real-time inspection pipeline on-site at a steel plant, cutting manual reprocessing **25%**

Fault Detection & Classification using Deep Learning | Vibrations Lab, IIT Kanpur

Nov 2023 – Feb 2024

- Redesigned a **VGG-16** CNN in TensorFlow on LabVIEW rotor-bearing signals, reaching **95%** multi-class accuracy
- Applied **t-SNE** for fault separability analysis and shipped a Tkinter GUI for operator-side inference

Technical Skills

Languages: Python, C++, Java, SQL, MATLAB, TypeScript

ML / DL: PyTorch, TensorFlow, Hugging Face, scikit-learn, OpenCV, XGBoost, LightGBM

GenAI & RAG: LangChain, LangGraph, FAISS, Pinecone, GraphRAG, LLM Evals & Guardrails, Agentic Workflows

Cloud & MLOps: AWS (Lambda, API Gateway, DynamoDB, S3, EC2), Vertex AI, Docker, Kubernetes, Jenkins, GitHub Actions

Backend & Data: FastAPI, Flask, Streamlit, REST APIs, PostgreSQL, MySQL, MongoDB, Git, Linux