

Shubham Gupta

+91-8737041988 | shubham.sg53147@gmail.com | linkedin.com/in/anon-cypher

Summary

Machine Learning Engineer with 3+ years of experience delivering scalable AI/ML and GenAI solutions across computer vision, deep learning, LLMs, and RAG systems. Experienced in building agentic AI workflows, LLM-powered automation, and high-performance MLOps pipelines. Recognized for developing award-winning systems that reduced manual QA by 60% and accelerated engineering processes with autonomous, intelligent pipelines.

SKILLS

Languages & Frameworks: Python, JavaScript, SQL, Bash, LangChain, LangGraph, AutoGen, Django, AngularJS

Libraries: TensorFlow, Keras, PyTorch, Transformers, Scikit-Learn, Pandas, NumPy, Matplotlib, Seaborn, OpenCV

Technical Skills: Artificial Intelligence, Machine Learning, Deep Learning, GenAI, LLMs, Agentic AI, RAG systems, rOCR

Tools & Technologies: Docker, Git, MLflow, Jenkins, OpenAI API, Hugging Face, Pinecone, ChromaDB, FAISS, Streamlit, AWS (EC2, S3, Lambda), FastAPI

Databases: MySQL, MongoDB, PostgreSQL, Vector Databases (FAISS, ChromaDB, Pinecone)

Experience | 3+ years

Engineer, HL Mando – Gurgaon, India

May 2023 – Present

- Developed and deployed AI-driven optimization workflows using Machine Learning, Deep Learning, and metamodeling techniques to automate and enhance internal engineering processes.
- Built a deep learning-based image analysis system for automated defect detection, **achieving 89% accuracy** and reducing manual inspection workload by **60%** in visual quality assurance pipelines.
- Designed and implemented end-to-end pipelines for 2D engineering drawing comparison using YOLOv8 and TrOCR, reducing manual QA effort by **30%**.
- Built deep learning systems for steel microstructure classification and mechanical property prediction, integrating CNN-based image models with regression heads to estimate yield strength, hardness, etc.
- Applied Genetic Algorithms and Reinforcement Learning to solve complex, high-dimensional optimization problems in design parameter tuning and decision-making systems.
- Created and deployed RESTful APIs using Django and FastAPI, containerized with Docker and deployed on Linux-based servers for real-time access to ML services across cross-functional teams.
- Collaborated with R&D, QA, and design engineering teams to integrate AI solutions into existing CAD/CAE workflows, improving process throughput and traceability.
- Recognized with multiple internal awards (**Spotlight Award, Best Project, MSI Contribution Award**) for delivering impactful, production-grade AI systems.

Security Automation Engineer, BreachLock Inc. – Noida, India

Feb 2022 – Jun 2022

- Worked as part of the cybersecurity automation team to develop automated vulnerability scanning and test orchestration pipelines for cloud-based infrastructure.
- Designed comprehensive test automation frameworks to validate the functionality, security, and reliability of internal tools across platforms.
- Led the integration of automated compatibility testing for web-based services using Selenium, REST APIs, and Python scripting, improving QA efficiency.
- Identified system performance bottlenecks through monitoring tools and participated in optimizing testing infrastructure for scalability.

Machine Learning Intern, ThinkingStack India Pvt Ltd – Remote

Aug 2021 – Jan 2022

- Contributed to the development of an Automated Number Plate Recognition (ANPR) system using YOLO-based object detection models, enabling license plate localization from video frames and surveillance data.
- Designed and implemented data preprocessing and annotation workflows using tools like OpenCV and LabelImg, ensuring high-quality labeled datasets for robust model training.
- Automated training and testing processes using Python-based scripting, supporting scalable experimentation and model iteration cycles.
- Conducted error analysis, model evaluation, and performance benchmarking using metrics like mAP, precision, recall, and confusion matrix interpretation.

Projects

Smart Factory – Automated Edge-to-Cloud ML & MLOps Pipeline

- Developed a full-stack factory monitoring dashboard using React and FastAPI, integrated with edge devices capturing images at each assembly point and streaming data to a Kubernetes-based inference server.
- Implemented a self-service model training workflow where training requests trigger Kubeflow Pipelines for dataset preparation, training, evaluation, and storing model artifacts in MinIO with versioning.
- Built an automated model deployment pipeline where approved deployment requests invoke Tekton, delivering model versions to specific factory locations or edge devices through a fully containerized Docker + K8s setup.

Intelligent Document-to-JSON Parser using LLM and Langchain

- Designed a system to convert unstructured documents (PDFs, scanned reports) into structured JSON using LLM-powered parsing.
- Used schema-driven prompts with Langchain and GPT-4 to extract key fields into a structured JSON format.
- Tools & Tech: GPT-4, Langchain, TrOCR, FAISS, FastAPI, Docker

Semiconductor Visual Defect Detection

- Developed a high-resolution image analysis system for defect classification using object detection and optimized IoU-based bounding box regressors.
- Focused on false-positive suppression via data augmentation, anchor box optimization, and multi-scale feature extraction.
- Tools & Tech: YOLOv7, OpenCV, PyTorch, LabelImg

Material Microstructure Classification & Mechanical Property Prediction

- Built a dual-head CNN architecture to classify microstructure images and predict associated mechanical properties (e.g., hardness, Young's modulus).
- Applied batch normalization, dropout, early stopping, and custom loss functions for improved generalization; integrated into a Django-based internal tool.
- Tools & Tech: Tensorflow-Keras, TensorBoard, NumPy, Django, Matplotlib

Automated Defect Detection in Engineering Drawings

- Designed a deep learning pipeline for visual quality control using YOLOv8 for object detection and TrOCR for OCR-based verification in 2D CAD drawings.
- Integrated geometric validation logic for tolerance matching; achieved 89% model accuracy, reducing manual QA by 60%.
- Tools & Tech: PyTorch, OpenCV, TrOCR, FastAPI, Docker

Real-Time Pothole Detection System

- Trained and deployed a YOLOv5-based real-time object detection model on road surveillance feeds; implemented bounding box tracking, achieving 72% mAP@0.5:0.95.
- Integrated alerting system with location tagging via metadata streams.
- YOLOv5, OpenCV, NumPy, Flask, AWS Lambda

Achievements

2 nd Prize – IRTC Research Paper, HL InnoVision	2024
Spotlight Award, HL Mando	2024
Best Project Award – CAE Automation, HL Mando	2023
MSI Contribution Award, HL Mando	2023

Certifications

Machine Learning Specialization, Coursera Link	Jul 2024
Build Basic Generative Adversarial Networks (GANs), Coursera	Jul 2021
Deep Learning Masters, iNeuron.ai	Jun 2021

Education

Center for Development of Advanced Computing (CDAC) - PG Diploma in Artificial Intelligence	Sep 2022 – Mar 2023
B. N. College of Engineering and Technology - B.Tech CSE	Aug 2018 – Jul 2022