# Saurabh Kumar

**Lead ML/MLOps Engineer**

ia.kumar0121@gmail.com | (321) 213-9581 | saurabh-ai.vercel.app

## PROFILE SUMMARY

Innovative and results-driven Machine Learning Engineer with over 11 years of hands-on experience in designing, deploying, and optimizing **AI/ML** systems. Proven success across **NLP**, Computer Vision (**CV**), **LLMs**, **Voice AI**, and **Generative AI**. Skilled in building **AI agents** using **RAG** pipelines, fine-tuning open-source models, and deploying robust cloud-native solutions on **AWS** and **Azure**. Adept at integrating AI into production workflows with a strong focus on performance, scalability, and user impact.

## CORE COMPETENCIES

- **Model Development & Deployment:** Proven expertise in building and deploying ML models using Python, PyTorch, TensorFlow, and scikit-learn, with a focus on performance and scalability.
- **Generative AI & LLMs:** Hands-on experience with fine-tuning open-source LLMs, building RAG pipelines, and integrating GenAI into production systems.
- **Cloud & Infrastructure**: Experienced in deploying AI solutions on AWS and Azure using Kubernetes, Terraform, and serverless architectures.
- **MLOps & Automation**: Skilled in ML lifecycle management using MLflow, Kubeflow, and SageMaker, with CI/CD integration via GitHub Actions, Jenkins, and Docker.
- **Cross-Functional Collaboration**: Effective communicator and mentor, driving alignment between data science, engineering, and product teams to deliver impactful AI solutions.

## SKILLS HIGHLIGHTS

| | |
|---|---|
| **AI/ML** | LLMs (GPT, Claude, Mistral, Llama, Phi-3), STT/TTS (Whisper, Wav2Vec, Bark, Tortoise), CV, NLP, Sentiment Analysis, Text2SQL, RAG, Prompt Engineering |
| **DevOps & Deployment** | AWS (EC2, Sagemaker, Lambda, Bedrock), Azure (VM, AI Search), Docker, Kubernetes, GitHub Actions, CI/CD |
| **Programming & Tools** | Python, PyTorch, TensorFlow, NumPy, Pandas, JavaScript, LangChain, HuggingFace, SQL, FastAPI |
| **Frameworks & Platforms** | LangChain, Azure AI Studio, Stable Diffusion, Streamlit, Autocad |
| **Version Control & CI/CD** | Git, GitHub, Jenkins, Docker, Terraform |

## PROFESSIONAL EXPERIENCE

**Lead Machine Learning Engineer & MLOps Engineer** | Blue Health Intelligence | *Chicago, IL* **Jun 2023 – Present**

- Fine-tuned advanced ML models including BERT, Mistral 7B, and Llama 2 to deliver domain-specific NLP solutions, boosting task accuracy by 30%.
- Built AI agents using Retrieval-Augmented Generation (RAG) pipelines with fine-tuned LLMs, enabling dynamic enterprise knowledge retrieval across unstructured datasets.
- Developed a multilingual voice-based AI chatbot, integrating STT (Whisper, Wav2Vec) and TTS (Bark, Tortoise) for real-time, natural user interaction across diverse languages.
- Led the development of an academic chatbot powered by Phi-3 (SLM), combining RAG and Text2SQL to support structured and unstructured query handling for institutional data.
- Architected and deployed full-stack GenAI solutions using LangChain Router, Azure VM, Azure AI Search (JSON index), and secure REST APIs for scalable enterprise integration.
- Delivered production-grade Chatbots and Text2SQL systems leveraging GPT-3.5/4 and prompt engineering, enabling human-like interactions with structured and semi-structured data.
- Evaluated and deployed AWS Bedrock models including Claude 2, Titan, and Llama-2-70B-chat to assess performance and suitability for enterprise-grade AI features.
- Managed CI/CD pipelines using GitHub Actions, AWS CLI, and Terraform to automate model deployment, testing, and infrastructure provisioning.
- Integrated monitoring and observability tools to track model performance, latency, and usage metrics, ensuring reliability and continuous improvement.
- Collaborated cross-functionally with product, data, and engineering teams to align ML solutions with business goals and user needs.

***Key Accomplishments***

→ Improved NLP model accuracy by 30% through targeted fine-tuning and domain adaptation.
→ Delivered multilingual voice AI chatbot with real-time STT/TTS, enhancing accessibility and user engagement.
→ Reduced model deployment time by 40% via automated CI/CD and infrastructure-as-code workflows.

**Senior AI/ML Engineer** | Turing | San Francisco, CA **Apr 2020– May 2023**

- Designed and fine-tuned ML models focused on medical data processing, enabling accurate diagnostics and predictive analytics for healthcare applications.
- Implemented patient classification systems using decision tree algorithms, improving clinical decision support and streamlining triage workflows.
- Built secure RESTful APIs to integrate ML models with frontend systems, ensuring seamless data exchange and real-time insights for medical staff.
- Managed HIPAA compliance across backend services and data pipelines, enforcing strict access controls and encryption standards.
- Deployed ML services using Docker and Kubernetes, enabling scalable, fault-tolerant infrastructure for healthcare analytics platforms.
- Integrated monitoring tools to track model performance and system health, ensuring reliability and rapid response to anomalies.
- Collaborated with cross-functional teams to align ML solutions with clinical needs, driving adoption and improving patient outcomes.

***Key Accomplishments***

→ Improved patient classification accuracy by 25% through decision tree optimization and targeted feature engineering.

→ Reduced API response time by 40% via backend refactoring and efficient data serialization, enhancing real-time usability for clinicians.

→ Achieved HIPAA-compliant deployment across cloud infrastructure, enabling secure integration with third-party healthcare platforms.

**Machine Learning Engineer** | Visionet Systems Inc | Cranbury, NJ **Dec 2017 – Mar 2020**

- Advanced image processing and generative vision systems using GANs, diffusion models, and transformer-based vision architectures (ViT, CLIP, DALL·E-style) to support large-scale labeling.
- Designed NLP pipelines for entity recognition, summarization, and semantic similarity using transformers (BERT, RoBERTa, GPT-based models) for enterprise annotation tasks.
- Scaled data processing pipelines for petabyte-scale datasets with Spark, Ray, Dask, and Kafka, integrated into AWS and GCP cloud infrastructure.
- Optimized model training and inference workflows with ONNX Runtime, TensorRT, and distributed GPU clusters, achieving higher throughput and lower latency.
- Contributed to cross-functional teams, mentoring engineers on ML best practices, data engineering, and production software development, ensuring high-quality and scalable AI solutions post-acquisition.

***Key Accomplishments***

→ Boosted image labeling throughput by 40% through deployment of generative vision systems using GANs, diffusion models and transformer-based architectures (ViT, CLIP, DALL·E-style).

→ Improved annotation accuracy and speed by 30% with NLP pipelines for entity recognition, summarization, and semantic similarity using BERT, RoBERTa, and GPT-based models.

**AI Engineer Intern** | Visionet Systems Inc | Cranbury, NJ **Dec 2014 – Aug 2018**

- Extracted and processed large datasets using Python, Pandas, and NumPy, improving model prediction accuracy by 15%.
- Delivered multiple end-to-end AI/ML solutions for global clients, focusing on predictive analytics, NLP, and recommendation systems.
- Developed a loan defaulter prediction model using ensemble techniques, achieving high precision in financial risk profiling.
- Built a vehicle recommendation engine using user behavior data, clustering techniques, and ML classification, increasing user engagement and conversion rates.
- Consulted on cloud-based ML deployments using Dockerized environments and RESTful API integration for web applications.

***Key Accomplishments***

→ Improved model prediction accuracy by 15% through advanced data preprocessing and feature engineering using Python, Pandas, and NumPy.

→ Built a vehicle recommendation engine leveraging user behavior data and clustering algorithms, increasing user engagement and conversion by 25%.

## EDUCATION

| | | |
|---|---|---|
| **Master of Artificial Intelligence** | \| University of Houston–Downtown – Downtown Houston, TX | **2020 – 2023** |
| **Bachelor of Computer Science** | \| Indian Institute of Technology Delhi – New Delhi, India | **2010 – 2014** |

## CERTIFICATIONS

| | |
|---|---|
| **AWS Certified Solutions Architect — Associate** | **Aug 2017** |
| **AWS Certified Cloud Practitioner** | **Aug 2017** |

## PROJECTS

**Voice AI Chatbot**

- Building a real-time voice assistant with STT/TTS pipelines and LLM-based conversation engine; designed for multilingual users.

**Chatbot with Phi-3 and RAG**

- Created a chatbot powered by Phi-3 (SLM) with RAG over unstructured academic rules and regulation documents; deployed using Azure VM, LangChain Router, and Azure AI Search (JSON index).
- Enabled Text2SQL interactions with relational databases using LLM routing and prompt chaining.

**AI Chatbot with LLMs**

- Developed a GPT-4 powered chatbot with memory persistence and dynamic context switching using LangChain.

**Renewable Energy Predictor**

- Neural network model predicting solar/wind output, achieving 92% accuracy across multiple conditions.

**NLP Resume Screener**

- Resume ranking tool with LLM-based NLP filtering, reducing manual screening time by 50%.

**YOLOv5 Real-Time Object Detector**

- Edge-deployed detection system for surveillance with 85%+ real-time detection accuracy.

**Text-to-Image Generator (Stable Diffusion)**

- Developed a photorealistic image generation system using diffusion models and natural language prompts.

**Breast Cancer Classifier**

- Designed a recommendation engine leveraging collaborative filtering and content-based techniques; integrated user profiling and clustering for enhanced personalization.

**Vehicle Recommendation System**

- Developed and deployed a predictive model using XGBoost and Random Forest to identify potential loan defaulters with high recall.

**Loan Defaulter Prediction Model**

- Building a real-time voice assistant with STT/TTS pipelines and LLM-based conversation engine; designed for multilingual users.

**Autonomous Weed Remover**

- Engineered a real-time weed detection system using Raspberry Pi and computer vision algorithms for agricultural automation.