Pradyumna Kumar Sahoo

pradyumna.sahoo@outlook.in | +91-7540885886 | prady029.github.io | Hyderabad, India

Summary

Impact-driven Data Scientist with close to 5 years of experience building production-grade AI/ML systems across Medical and Finance domain for Computer Vision, Audio and Generative AI use-cases. Proven track record in architecting real-time audio chatbot systems, Knowledge-graph based GraphRAG pipelines, fine-tuning large language models, developing multimodal agents and leading cross-functional teams. Seeking a Senior Data Scientist role to drive AI innovation at scale.

Experience

Data Scientist August 2025 – Present

Mondee Pvt. Ltd. *Hyderabad, India*

- Architected a **medical-grade GraphRAG audio chatbot** for our flagship clinical decision support system **CDSS** deployed across **Surekha Hosptial Chain** and **BhaktiVedant Hospital**, by constructing structured knowledge graphs from medical textbooks using *NER* and *Neo4J*, enabling doctors to query alternative possible diagnosis and treatment protocols with traceable, hallucination-resistant responses via a *LangChain*-powered retrieval layer.
- Engineered a **drug–drug interaction checker and dosage scheduler agent** served through a *Model Context Protocol(MCP)* interface for consumption by in-house agents, integrating real-time data from **RxNorm** and **PubMed** APIs to surface conflict alerts and patient-specific dosage recommendations within the chat session.
- Led **end-to-end fine-tuning and serving of Medgemma-27b-text-it** for domain-specific clinical NLP tasks using *Unsloth* and *vLLM*, coordinating data curation, training, and evaluation pipelines with research scholars from **IIT Madras and IIT Hyderabad**.
- Directed **large-scale Speech-to-Text (STT/ASR) data preparation and multimodal fine-tuning of gemma-3n-e2b-it** via *LoRA* and *TGI*-based serving, overseeing live clinical audio collection, Subject Matter Expert annotation, and quality control to build a medical voice agent capable of real-time clinical transcription and Doctor's Note generation.

Senior Member Technical (AI/ML) December 2023 – July 2025

ADP India Pvt. Ltd. *Hyderabad, India*

- Designed and deployed a **Knowledge-Graph based RAG pipeline** on *AWS Neptune* & *AWS Bedrock* orchestrated via *LangGraph*, enhancing financial data processing accuracy and minimizing AI hallucinations. Recognized as **runner-up in the ADP Global Hackathon(2024).**
- Developed an intelligent **Process Mining Chatbot solution for Global Payroll Services** using *Microsoft Power Automate* and *Databricks* to analyze transaction patterns across millions of client records, optimizing payment workflows and eliminating operational bottlenecks across **73 payroll cycles per client on average**.
- Engineered an **agentic assistant** built on *Google Agent Development Kit* and *AWS Opensearch* that drafts context-aware emails and schedules meetings in real-time by checking live calendars within user chat sessions, saving equivalent to **24 hours per user per month**; currently rolled out across all ADP employees.
- Built a **scalable QR code detection, decoding, and masking pipeline** with a fine-tuned *YOLOv8* for multi-orientation detection and *OpenCV* for automated region masking, sanitising financial documents prior to downstream processing at scale across millions of payroll documents.
- Developed an **Indic PII detection and redaction system** for financial regulatory compliance using fine-tuned NER models via *HuggingFace Transformers* and *SpaCy* with *IndicNLP*, identifying sensitive entities — Aadhaar, PAN, account numbers, names across 10+ Indic scripts — from payroll and HR documents.

Junior Data Scientist June 2021 – December 2023

Claim Genius Pvt. Ltd. *Remote, India*

- Built a high-performance **Instance Segmentation pipeline** using *Detectron2* served via *FastAPI* for vehicle parts identification, improving segmentation accuracy to **95% mAP** and accelerating assessment throughput by **26%**; integrated *GradCAM++* visualisations for regulatory model interpretability, enabling stakeholders to audit spatial reasoning behind predictions.
- Engineered an **automated ML pipeline failure tracing system** using *MLflow* to monitor inference-time distributions and surface root-cause analysis on mispredictions, **reducing manual diagnosis time by 30%** and enabling faster corrective action.

- Deployed an **image super-resolution and denoising ensemble** combining *SWIN2SR Transformer* and *NAFNet* pre-trained models to upscale and denoise compressed input images, **reducing model failures by 16%** and improving downstream prediction accuracy by **12%**.
- Improved data pipeline quality by building an **automatic labelling error-detection service** using *Scikit-learn* confidence scoring that flagged curation errors — **saving 6 man-hours per head per sprint** — and trained *PyTorch* **GAN models for synthetic image generation** to resolve class imbalance in rare damage categories.
- Extended damage assessment capabilities by designing a **geometric flat-tyre detection approach** via *OpenCV* polygon analysis of tyre and rim regions — enabling reliable detection with zero curated data — and boosted **vehicle damage severity classification by 3% per class** through a fusion ensemble combining *PyTorch* CNN features with *XGBoost* structured metadata.

Technical Skills

LLM Training & Inference: Full fine-tuning, PEFT / LoRA / QLoRA, Instruction Tuning, RLHF / DPO, Mixed-precision (bf16/fp16), Gradient Checkpointing, vLLM, Quantisation (GPTQ / AWQ / bitsandbytes), HuggingFace Transformers, DSPy

Agentic & RAG Systems: LangChain, LangGraph, LiveKit, Tool-use / Function Calling, Multi-agent Orchestration, Multimodal Retrieval-Augmented Generation, Model Concept Protocols (MCP)

Knowledge Graphs & Vector Search: Neo4j, AWS Neptune, Qdrant, Knowledge Graph Construction

ML & Deep Learning: PyTorch, TensorFlow, Scikit-learn, XGBoost, LightGBM, Detectron2, OpenCV, SpaCy

MLOps & Productionisation: Weights & Biases, Docker, FastAPI, Apache Airflow, Apache Spark(Databricks), ETL Pipeline Design, Process Mining, CI/CD for ML, Git

Cloud & Infrastructure: AWS (Bedrock, Neptune, S3, Lambda, EC2, Opensearch), GCP (Vertex AI – familiar), Azure ML (familiar), Kibana

Databases: LanceDB, PostgreSQL, MongoDB, SQLite

Languages : Python, SQL, MATLAB, Bash

Certifications

Agentic Knowledge Graph Construction – DeepLearning.AI × Neo4j *August, 2025*

Building AI Voice Agents for Production – DeepLearning.AI × LiveKit *July, 2025*

Neo4j Fundamentals – Neo4j GraphAcademy *July, 2025*

Pretraining LLMs – DeepLearning.AI × Upstage *February, 2025*

TensorFlow Developer Certificate – Coursera *August, 2020*

Deep Learning Specialization – DeepLearning.AI, Coursera *April, 2020*

Research Projects(TCS Big Data Lab)

Multi-label Classification Enhancement – Leveraged MLSMOTE + LLSF-DL deep learning hybrid to generate synthetic data for rare tail-labels using *Python*, *MATLAB*, and *TensorFlow*, significantly improving multi-label classification performance on imbalanced datasets.

Session-based Recommendation with Graph Neural Networks – Designed and implemented a *PyTorch*-based Graph Neural Network engine that captures graph-structure dependencies to deliver personalized item recommendations during active user sessions, enhancing recommendation accuracy and relevance.

Education

M.Sc. Computer Science (Big Data Analytics)
Central University of Rajasthan *Kishangarh, India*

Integrated B.Sc. B.Ed. (Physical Sciences and Education)
Regional Institute of Education (NCERT), Bhubaneswar *Bhubaneswar, India*