# Prasuk Jain

Morena, India | P: +91 8819855540 | p.jain161202@gmail.com | LinkedIn | Github

## EXPERIENCE

**Xcelyst Partners** *August 2023 – Present*
*Data Scientist*

***Project :*** *AI Interview Bot* *January 2025 - Present*

- Led a research team to assess the candidate's expression and body language using computer vision techniques.
- Architected the backend of the project and seamless integration of the backend with AI model as a separate service.
- Partnered with product stakeholders to translate GTM needs into data science deliverables.
- Integrated Kyutai STT 1b for testing purposes, resulting in a reduction of the TTS and STT latency by 40%.
- Designed & deployed a **Databricks-based ingestion + compute platform** for large scale structured + unstructured processing, enabling data readiness for downstream analytics & models.
- Hardened infra for production workloads using **Docker + AWS EC2 + GCP**, improved model reliability, scalability and **reduced infra cost footprint.**

***Project :*** *IndiGrid Maximo* *April 2025 – July 2025*

- Designed and implemented the **backend architecture** for a web application, replacing IBM Maximo for Indigrid.
- Identified and developed **Generative AI** use-cases within the power system ecosystem to enhance operational efficiency.
- Automated **data scraping processes using Python**, significantly reducing manual technical tasks and improving productivity.
- Designed a **multi-agent RAG pipeline** combining retrieval, tool-calling and iterative refinement loops to improve reasoning quality and reduce hallucinations in domain-specific LLM use-cases.

***Project :*** *Candidates Screening Model* *Aug 2023 – Dec 2024*

- Cleaned and pre-processed large-scale job descriptions and resumes for text similarity analysis.
- Built **real-time scoring models** using BM25 + Embeddings + Flask to rank candidate-job similarity, improving recommendation precision by **25%** with continuous monitoring.
- Dockerize applications and deploy them on **AWS EC2** instances, optimising the infrastructure for scalability and efficient resource management.
- Worked on a candidate-job role matching model using Jaccard Similarity, enhancing role-based recommendations and making it production-ready by using **FLASK.**
- Optimised **full-text search** queries in **PostgreSQL** to boost resume and job search efficiency.
- Scaled data pipelines to process **8M+ profiles** using PostgreSQL + NLP, improved match quality by **30%** and reduced query latency by **85%** enabling more accurate & faster decisioning.
- Fine-tuned **Meta LLaMA 3.1** with SFT + QLoRA (4-bit) using **Unsloth**, delivering **domain-adapted language understanding** that reduced bad matches / noisy suggestions.
- Partnered with product stakeholders to translate GTM needs into data science deliverables.

## SKILLS .

**Programming Languages:** Python, PySpark, SQL (PostgreSQL).
**Tools & Platforms:** AWS, GCP, Vertex AI, Docker, Git, Databricks, MS Office Suite (Excel, PowerPoint).
**ML & AI Frameworks:** TensorFlow 2, PyTorch, OpenCV, Scikit-learn, Scikit-image, NumPy, Pandas, Matplotlib, Seaborn, Flask, Hugging Face, OpenAI APIs.
**Soft Skills:** Leadership, Team Collaboration, Communication, Strategic Thinking, Analytical Problem Solving.

**Certifications:**

- Databricks Certified Data Engineer Associate by Databricks
- Image Processing with Python by DataCamp
- Generative AI by Udemy
- TensorFlow 2: A Complete Guide by Udemy

## PROJECTS

**Implementation of Denoising Diffusion Probabilistic Models using Tensorflow (Demo)** *Oct 2025*

- Implemented DDPM from scratch (CelebA) with FP16 mixed precision, achieved **FID < 10** and final loss **0.014** within 6 epochs, stable sampling with reproducibility.
- Optimised diffusion training pipeline delivering **1.95× GPU speedup** on P100 and **20× faster DDIM sampling** vs vanilla DDPM with minimal quality degradation.
- Authored a detailed technical explainer on mathematical formulation + implementation nuances (LinkedIn), gaining strong engagement from the deep learning community.

**MagicFace Webapp (Code)** *May 2023 – Ongoing*

- Independently developed "Magicface" project using **Flask** framework and **machine learning** algorithms.
- Implemented facial analysis features, including gender detection, facial expression recognition, and age estimation.
- Integrated real-time filter application functionality for image enhancement and modification.

**Kyutai STT-1B: Real-Time Streaming Speech Recognition System (Code)** *Aug 2025 – Sept 2025*

- Built a real-time speech recognition demo using Kyutai STT-1B with streaming inference via the Moshi framework.
- Achieved **0.5s latency** and **40% lower memory usage** vs Whisper through Delayed Streams Modelling (DSM).
- Integrated **semantic Voice Activity Detection (VAD)** for speech boundary detection without external models.

## EDUCATION .

**Madhav Institute of Technology and Science, Gwalior** *2020-2024*

Bachelor of Technology, Artificial Intelligence and Robotics
Cumulative CGPA: 8.23/10
Core Team Member of Robotics Club and Coding Club

## ADDITIONAL .

**Languages:** Fluent in English.

**Extra:**

- Participated in Wittyhacks 3.0, organised by Datacode and Major League Hacking (MLH)
- Participated in HackCBS 6.0, organised by Major League Hacking (MLH)