

Arpit Kumar

Data Science & ML Engineer focused on multimodal modeling, experimentation, and production-scale ML systems.

<https://arpitkumar.dev> |  arpitkumar2004 |  Arpit Kumar |  arpitkumar@kgpian.iitkgp.ac.in |  +91-9506035861

EDUCATION

Indian Institute of Technology (IIT) Kharagpur, India

SGPA: 7.86/10

2022 - 2027

Major: Integrated B.Tech - M.Tech (Dual Degree) in Chemical Engineering

Micro: AI & ML foundation and Application

TECHNICAL SKILLS

Languages: Python, C, C++, SQL, JavaScript

ML & AI: PyTorch, TensorFlow, Scikit-learn, Transformers, Gradient Boosting (LightGBM/XGBoost/CatBoost)

Data Engineering: NumPy, Pandas, SciPy, Matplotlib, Seaborn, Joblib, OpenCV, BeautifulSoup, Selenium

MLOps & Infrastructure: Docker, GCP, MLflow, DVC, Apache Airflow, Apache Kafka, YAML, Git

Software & Backend: FastAPI, Flask, PostgreSQL, MongoDB, Redis, Linux, CI/CD Workflows, Cron Jobs

ACHIEVEMENTS

- Competitive Programming: Achieved **Expert** status on Codeforces (Peak Rating: **1612**, Handle: **kumararpit**)
- Ranked **Top 20 Nationally** in the DTL Quant Challenge 2024 and **Top 0.5% Global** rank in the Amazon ML Challenge
- Secured All India Rank **1478** (Top 0.8%) in JEE Advanced and **99.2 %ile** in JEE Mains 2022 among 1 Million+ candidates
- Institute Championships: Awarded **Gold Medal** (GC Data Analytics 2024), **2x Silver Medals** (OpenIIT Data Analytics and Case Study 2023), and **Bronze Medal** (GC ChemQuest) in the Inter-hall General Championships

PROJECTS

Multimodal Price Prediction System | **PyTorch, BERT, CLIP, XGBoost, Docker**

Kharagpur, India

Developed an end-to-end multimodal ML pipeline for product price estimation from images & Description

Jun 2025 - Aug 2025

- Built multimodal regression with BERT text embeddings, CLIP image features, & tabular metadata for 75K products
- Improved SMAPE from 26.4% (tabular LightGBM baseline) to 18.2% via feature fusion and stacking ensemble
- Conducted ablation study showing text embeddings contributed +4.7% and image features +3.1% performance gain
- Reduced p95 inference latency (250ms to 50ms) on GCP n2-standard-4 via parallel feature extraction & sparse caching
- Prevented data leakage by time-aware train/validation split and feature sanitization

Conversational AI Platform for Employee Welfare | **FastAPI, Isolation Forest, Transformers, GCP**

IIT Kharagpur, India

Developed a multi-stage AI pipeline integrating ensemble anomaly detection and LangChain-powered LLMs

Sep 2025 - Dec 2025

- Built a multi-platform dashboard (**Next.js, Expo, FastAPI**) with automated tracking via CRON-based backend tasks
- Flagged at-risk users by an anomaly detection engine using EMA-weighted metrics, Isolation Forest, and LOF models
- Automated HR reporting using **LangChain**-driven LLM pipelines for context-aware chat and well-being summaries
- Containerized full-stack deployment on Google Cloud (GCE, GCR, GCS) for high-availability enterprise performance

Transport Analysis of Electrochemical CO₂ to Methanol Conversion | **Prof. Sourav Modal | BTP**

IIT Kharagpur, India

Zero-Gap Membrane Electrode Assembly (MEA) using Gas Diffusion Electrodes (GDE)

Aug 2025 - Dec 2025

- Diagnosed a 3.1x critical scalability gap (40 vs. 130 mA/cm²) by conducting a rigorous Techno-Economic Analysis, identifying transport phenomena than catalyst kinetics as the primary barrier to commercial Levelized Cost of Product
- Validated a theoretical specific energy of 15.2 kWh/kg through detailed mass-energy balance calculations, optimizing the system for >40% Energy Efficiency by minimizing ohmic resistance and maximizing single-pass conversion.
- Architected a Zero-Gap GDE-MEA reactor framework designed to sustain >200 mA/cm² current density by integrating a liquid anolyte loop, effectively mitigating thermal runaway and parasitic carbonate crossover.
- Re-engineered the reaction control strategy by modeling the mobile CO* intermediate as a 3D reaction-diffusion problem, establishing fluid dynamics as the decisive lever for Methanol Selectivity (FE >90%)

COMPETITIONS

General Championship Data Analytics by Evva Health | **Gold**

IIT Kharagpur, India

Technology Students' Gymkhana, IIT Kharagpur | Developed a Social and Healthcare Risk Scorecard

Mar 2024 - Apr 2024

- Achieved **82.89%** accuracy by Voting ensemble **BERT**, Naive Bayes on patient-reported health data for risk classification
- Reduce manual data collection time by **60%** by deploying a **Streamlit** web app with real-time scoring & dynamic feedback
- Scraped **1000+** entries by **BeautifulSoup/Selenium** and **Bifactor & MIRT** models to optimise healthcare resource in India

Open IIT Data Analytics Competition | **Silver**

IIT Kharagpur, India

Technology Students' Gymkhana, IIT Kharagpur | Prediction of Footfall on time-series data

Nov 2023 - Dec 2023

- Secured **2nd** rank among institute, predicting **city footfall** from prior years time series data using **time-series models**
- Scraped data by **BeautifulSoup** and applied **EDA**, Time series analysis, and clustering techniques by **K-Mean Cluster**.
- Utilized forecasting models **FBProphet, Random Forest, LSTM**, and ensemble methods & achieved **accuracy** of 86.63%.

RELEVANT COURSEWORK

IIT Kharagpur : Mathematics I, Mathematics II, Probability and Statistics, Optimization Techniques, Programming & Data Structures, Advanced Mathematical Techniques, Transform Calculus, Introduction to Stochastic Processes*, Partial Differential Equation, Computer-Aided Process Engineering, High Performance Computing*, Process Dynamics & Control

Standford Online : Machine Learning (CS229), Deep Learning (CS230), Natural Language Understanding (CS224U), NLP with Deep Learning (CS224N), Reinforcement Learning (CS234)