Ashutosh Roy

Data Scientist

+91-9810599837 | ashu2003roy@gmail.com | linkedin.com/in/ashutosh-roy | github.com/ashutosh-roy | Gurugram, Haryana

Summary

Data Scientist with applied experience in statistical modeling, NLP, and ML pipeline development. Published researcher (RECCAP 2026, IIT Palakkad) with a track record of translating complex datasets into measurable outcomes — from benchmarking OCR systems at IIT Delhi to building predictive models for healthcare data at MitoVoid AI. Skilled in Python, Scikit-learn, Pandas, and Power BI, with strong foundations in feature engineering, evaluation, and visualization.

Experience

Research Intern **Jan 2025 – Jun 2025**
Indian Institute of Technology (IIT) Delhi *New Delhi, India*

- Conducted **benchmarking and evaluation** of 3 OCR systems (Tesseract, Marker, Gemini OCR) across document types — analyzed failure modes, produced structured comparison metrics, and recommended data-driven extraction strategies.
- Designed a hybrid retrieval pipeline combining semantic vector search (FAISS/Pinecone) and keyword retrieval (Apache Solr); evaluated retrieval precision across thousands of archaeological records.
- Co-authored **SARCH: Multimodal Search for Archaeological Archives** with IIT Delhi. *arxiv.org/abs/2511.05667*

AI/ML Intern **Oct 2024 – Nov 2024**
MitoVoid AI *Remote*

- Built predictive NLP models for healthcare data analysis; applied **feature engineering** techniques including domain-specific tokenization and clinical entity extraction to improve model performance.
- Developed end-to-end data pipelines: ingestion → preprocessing → feature extraction → model inference, reducing manual processing time significantly.
- Improved patient query resolution time by ˜**40%** through optimized intent classification and response generation.

Software Engineer Intern **Nov 2024 – Dec 2024**
LeanTactics Solutions Pvt. Ltd. *Remote*

- Integrated ML model outputs into production backend systems using Python and FastAPI; optimized data flow between inference layer and application, achieving **25–35% latency reduction**.
- Profiled API performance bottlenecks using data-driven approaches; delivered measurable improvements in request handling at scale.

Projects

Speech Emotion & Stress Detection | *Python, PyTorch, Wav2Vec2, BiLSTM, Hugging Face* | ***RECCAP 2026*** Aug 2024 – Present

- Built multi-task ML model across 3 corpora (RAVDESS, TESS, SAVEE); implemented unified preprocessing (VAD, silence removal, log-Mel spectrograms) and cross-corpus label harmonization for consistent evaluation.
- Achieved **85.9% accuracy**, **0.856 F1**, RMSE **0.1268** — evaluated using macro-F1, confusion matrix analysis, and regression RMSE; outperformed single-task baselines on all metrics. Paper accepted at RECCAP 2026, IIT Palakkad.

Exam-Helper RAG System | *Python, LLaMA 3, FAISS, Pinecone, LangChain* Dec 2024 – Apr 2025

- Designed a retrieval-augmented QA pipeline over large academic corpora; implemented vector indexing and similarity search for context-precise LLM responses with measurably reduced hallucination.

Medical Chatbot | *Mistral 7B, QLoRA, Tesseract OCR, Transformers* Mar 2024 – Apr 2024

- Fine-tuned Mistral 7B on domain-specific medical datasets; built OCR preprocessing pipeline to extract structured clinical data from unstructured reports for downstream model inference.

Music Identification App | *Python, ACRCloud API, Streamlit, NumPy* Sep 2024 – Dec 2024

- Built real-time audio recognition pipeline: audio capture → feature extraction → ACRCloud API fingerprinting → metadata retrieval; served via Streamlit UI for instant song identification.

BulkyMail | *Python, Streamlit, SMTP, Jinja2, Pandas* Dec 2024

- Built a bulk email automation tool with dynamic Jinja2 templating over recipient data (name, role, company); processed structured CSV lists to drive personalized large-scale outreach campaigns.

Technical Skills

Data Science & ML: Statistical Modeling, Feature Engineering, NLP, Regression & Classification, Evaluation & Benchmarking, Deep Learning, RAG, LLMs
Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, PyTorch, Hugging Face, TensorFlow, LangChain
Visualization & BI: Power BI, Tableau, Streamlit, Matplotlib
Databases: MySQL, FAISS, Pinecone
Languages: Python, SQL, C++
Tools: Jupyter Notebook, Google Colab, Git, Docker

Education

Chhattisgarh Swami Vivekananda Technical University (CSVTU) **2022 – 2026**
B.Tech (Hons) in Computer Science and Engineering (Artificial Intelligence) *Bhilai, Chhattisgarh*

St. Columbus School **2020 – 2021**
Higher Secondary (Class XII - Science PCMB) - 87% *Faridabad, Haryana*