Rupapriya Naskar

Integrated Dual Degree (B.Tech + M.Tech)
Department of Aerospace Engineering
Indian Institute of Technology, Kharagpur

+91-7047382727
rupapriyanaskar2002@gmail.com
linkedin.com
Kharagpur, West Bengal, 721302

Education

Degree/Certificate	Institute/Board	CGPA/Percentage	Year
Dual Degree, AE	Indian Institute of Technology, Kharagpur	7.71	2024
Higher Secondary	WBCHSE	92%	2020
Secondary	WBBSE	91.71%	2018

Internships

- **Research Intern: National University of Singapore** *May. 2023 - July. 2023*
  Objective: To extract key voter issues and predict preferences from large-scale multilingual data
  - Researched statistically analyzing the salience biases of voters and contextualizing biases to devise effective political strategies
  - Built a probabilistic model using multilingual dataset of **6,00,000**+ human responses to effectively predict citizens' preferences
  - Optimized and implemented the **BERTopic** to generate the topics and effectively subdivide the whole corpus based on them
- **Data Science Intern: Chi SquareX, Bengaluru, India** *Nov. 2023 - Dec. 2023*
  Objective: Created a credit risk classifier on the German Credit dataset to predict loan defaults
  - Conducted EDA and FE, using Chi-square and ANOVA for univariate screening and applied **PCA** to address multicollinearity
  - Compared 8 models (LR, RF, SVM, XGBoost, LightGBM); LightGBM achieved best **F1** of **0.86 (test)** and **0.83 (5-fold CV)**
  - Tuned LightGBM with GridSearchCV (lr, depth, estimators), improving F1 score under class imbalance via cross-validation
- **Research Intern: Indian Institute of Technology, Kharagpur** *May. 2025 - July. 2025*
  Objective: To develop an end-to-end climate diffusion pipeline for long-horizon SST generation
  - Built a conditional **DDPM** with an attention-augmented **UNet** to generate seasonally conditioned global SST anomaly maps
  - Designed month-conditioned diffusion pipeline with cross-attention embeddings for seasonally coherent global SST generation
  - Developed an autoregressive SST generation and evaluation pipeline validated via spatial and grid-point temporal correlations
  - Achieved high model skill with an average spatial correlation of **0.982** and temporal correlation of **0.098** across training epochs

Publications

- **Title: A Comparative Study Of Generative Models as Surrogates of Earth System Models to Simulate Global SST (Under Review)** *July. 2025*
  Author(s): Deepayan Chakraborty,Raj Kishore,Rupapriya Naskar,Adway Mitra
  - Studied data-driven climate emulation to generate the long-term global SST as a low-cost alternative to Earth System Models
  - Evaluated GAN, VAE and DDPM with similar architectures for long-horizon global SST generation using CMIP6 GISS data
  - Found GAN-based emulations most consistent with the observed SST patterns, while other models matched in specific aspects

Projects

- **Project A: Medical Chatbot, Self Project** *Jan. 2025 - Feb. 2025*
  Objective: To offer accessible, accurate medical assistance through automated question answering [GitHub]
  - Built an end-to-end AI-driven medical question–answering system capable of reasoning over large unstructured clinical content
  - Designed a scalable pipeline that ingests, structures, and preserves the contextual relationships within long-form medical texts
  - Incorporated a **vector database** (Pinecone) to enable precise, context-aware retrieval of medical information across documents
  - Implemented a retrieval-augmented generation**(RAG)** framework to produce grounded, reliable, and context-specific responses
  - Built a deployment pipeline using Docker, GitHub Actions, and AWS to automate containerization and cloud-based delivery
- **Project B: Twitter Hate Speech Detection, Self Project** *Mar. 2025 - Feb. 2025*
  Objective: To enhance social media safety through an accurate, production-ready hate speech detection model [GitHub]
  - Built hate speech detection system with RNN, achieving **93.8% validation accuracy** and **91% recall** on dataset of 56745 tweets
  - Designed custom NLP tokenizer handling 50K words, converting texts to sequences and padding to 300 tokens for the training
  - Trained a **5.1M-param** LSTM with 100-dim embeddings using PyTorch Lightning, optimized with **Adam** at **1e-3** learning rate
  - Developed CI/CD pipeline with Docker, GitHub Actions and AWS (ECR, EC2) to automate containerization and deployment
- **Project C: Identifying Question Pairs with Similar Intent, Self Project** *Oct. 2024*
  Objective: To eliminate duplicate questions by identifying semantically similar question pairs using machine learning techniques [GitHub]
  - Used Random Forest with Gaussian Naive Bayes, achieving **75% accuracy** and **60% recall** in detecting similar question pairs
  - Performed text preprocessing with tokenization, stemming and removal of HTML tags, stopwords and patterns from the data
  - Utilized **TF-IDF** vectorization to transform textual data into meaningful numerical features for the machine learning models
  - Deployed the machine learning model using Streamlit, facilitating real-time data handling and interactive results visualization

Skills

- **Programming & ML Stack:** C, C++, SQL, Python, NumPy, Pandas, PyTorch, Scikit-learn, Matplotlib, Seaborn, NLTK, LangChain, RAG
- **Developer Tools:** Git, GitHub, Docker, Jupyter Notebook, VS Code

Extra Curricular Activities

- **Cultural:** Participated in Cartooning Competition at the General Championship 2023 and won gold medal for JCB Hall, IIT KGP
- **Sports:** Silver Medal in General Championship Football (2022–23); represented JCB Hall in Cricket, Athletics, and Badminton

Coursework Information

- Programming & Data Structures | Probability & Statistics | Deep Learning | Foundations of Artificial Intelligence & Machine Learning | Introduction to LLMs