

Saransh Gupta

[Email](#)
[LinkedIn](#)
[Github](#)
[Website](#)
[+91-9530277421](#)

ACADEMIC PROFILE

Year	Institution	Degree	CGPA
2022	Indian Institute of Technology Kharagpur	B. Tech. + M. Tech. (Engineering Product Design)	8.09 / 10.00

PUBLICATIONS

- S. Gupta *et al.*, "Integrative Network Modeling Highlights the Crucial Roles of Rho-GDI Signaling Pathway in the Progression of Non-Small Cell Lung Cancer," in *IEEE - JBHI*, 2022, [doi: 10.1109/JBHI.2022.3190038](#)
- Entity-aware Question-Answer Extraction for Shopping Guidance, **Amazon Machine Learning Conference - 2022**
- (Gupta et al.) Development of a virtual reality-based fire training simulator and machine learning-based path guidance system (working paper), **IHIET-AI, 2020**, Centre Hospitalier Universitaire Vaudois, **Lausanne, Switzerland**

INTERNSHIPS AND PROJECTS

Amazon Development Center (India) Applied Scientist - Intern		Jan'22 – June '22
Project - 1: Build a demo tool to help in the navigation and exploration of the Pre-curated Question Bank (PCQB)		
<ul style="list-style-type: none"> ▪ Created a dashboard using streamlit enabling a user to input their query and get relevant questions accordingly ▪ Integrated the frontend with the backend and a BERT based model to fetch relevant questions based on queries input ▪ Demonstrated the coverage of PCQB with respect to user queries using the query-question relevance feature 		
Project - 2: Generate Pre-curated Question Bank (PCQB) Question and Answer extraction from articles		
<ul style="list-style-type: none"> ▪ Developed a Transformers based two-step model for the Question Generation followed by the answer extraction ▪ Scrapped Texts, People Also Ask (PAA) questions and answers using certain queries related to E-Commerce domain ▪ Increased the size of training dataset by 20 times by paraphrasing the dataset using T5 Text to Text Generator model ▪ Achieved a Perplexity score of 82.3 on Question Generation by fine-tuning pre-trained T5 model on the PAA dataset ▪ Attained an F-1 score of 0.79 on the answer extraction task by fine-tuning encoders of T5-large model on PAA dataset ▪ Deployed the two step model pipeline on the streamlit based demo web-application that accept user input as text 		
Tools and Software: streamlit, Python, PyTorch, Transformers, BeautifulSoup, BERT, T5 (text to text generator)		
ZS Associates Inc. Data Science Associate - Intern		Jan'21 – June '21
Project - 1: Extract biomedical text dataset, identify entities, and classify if there exists a relation between entities		
<ul style="list-style-type: none"> ▪ Created a pipeline to extract texts from PubMed database, identifying the entities using Selenium and PubTator ▪ Implemented Binary Classification rules, devised four labeling functions using bio-verbs, co-occurrence of entities ▪ Generated a training dataset utilizing the four labeling functions in Snorkel by applying the Weak Supervision ▪ Achieved F1 score of 0.88 on the gold-standard dataset in relation-classification by training RoBERTa base model 		
Project - 2: Identify the type of relationship between two entities if it exists from the results of the Project-1		
<ul style="list-style-type: none"> ▪ Created a new set of three labelling functions for relation-type identification by using the results of the project-1 ▪ Attained F1 score of 0.83 on the gold-standard dataset using XGBoost Model followed by feature engineering 		
Tools and Software: Python, TensorFlow, Transfer Learning, Medline-Plus API, PubTator, Selenium, Snorkel		
Osaka University, Japan Remote Research Assistant		Jan '20 – Dec '20
Guide: Dr. Kenji Mizuguchi , Mizuguchi Lab, Osaka University, Osaka, Japan		
Project: Predict the Non-Small Cell Lung Cancer (NSCLC) using Machine Learning, identify its potential drug targets		
<ul style="list-style-type: none"> ▪ Extracted 412 essential genes out of 10,077 by applying Boruta Feature selection on their gene expression dataset ▪ Obtained F-1 score of 1.0 on validation and 0.98 on test dataset by using the XGBoost model to predict NSCLC ▪ Predicted drug targets for the NSCLC by simulating a Bayesian Network Model on the Rho-GDI signaling pathway ▪ Discovered methodology leads to an accurate treatment of the disease impacting 85% of the lung cancer patients 		
Tools and Software: Python, TargetMine, scikit-learn, smote, NetworkX, NumPy, pandas, Plotly, joblib		
ACHIEVEMENTS		
<ul style="list-style-type: none"> ▪ Featured as one of the Top 30 Undergraduate Achievers of IIT Kharagpur in the UG Achievers Directory 2020 ▪ Conferred merit-based scholarship of 2200 € by The A*Midex Foundation of Aix-Marseille University, France ▪ Selected among Top 5% out of all for the summer fellowship at The Institute of Science & Technology Austria ▪ Got featured in the ISE Newsletter Autumn-2020 under the Department Spotlight of ISE fights COVID-19, 2020 ▪ Awarded as Intern of The Month for my contribution as a Data Analyst at Sapio Analytics by the CEO in July 2020 		
COMPETITIONS / CONFERENCES		
Annual Amazon Machine Learning Conference (AMLC) – Bengaluru, Karnataka		[Aug 2022]
23rd World Business Dialogue, Creation Lab at Evonik - Cologne, Germany		[Jun 2022]
International Conference on Human Interaction & Emerging Technologies: Future Applications		[Aug 2020]
Young Data Scientists annual meetup at Kaggle - days, Dubai World Trade Centre		[Mar 2020]
Winner Databuzz(Data Analytics Competition) DoMS IIT Madras		[Jan 2020]
Problem Statement: Prediction of the defaulters on lending credit cards to minimize loss incurred to the banks		