

EDUCATION

Indiana University

Master of Science in Data Science | **GPA: 4.0/4.0**

Bloomington, IN

Aug 2022 – May 2024

Relevant Coursework: Statistics, Machine Learning, Financial Econometrics, Microeconomics, Applied Algorithms, Exploratory Data Analysis, Deep Learning Systems, Applied Database Technologies

Nirma University

Bachelor of Technology | **GPA: 8.59/10.0**

India

Aug 2016 – May 2020

Relevant Coursework: Calculus, Linear Algebra, Discrete Mathematics, Probability, Computer Vision, Robotics

SKILLS

Programming Languages and tools: Python, MySQL, PostgreSQL, R, Matlab, Tableau

Python Libraries: NumPy, Pandas, TensorFlow, Keras, Pytorch, Scikit Learn, SciPy, NLTK, Seaborn, Matplotlib, Plotly, Pillow, Numba, Dask, Transformers, PyMongo, Modin-Dask, Concurrent, Boto3

Areas of expertise: Statistics, Machine Learning, Deep Learning, Signal Processing, Image Processing, Bioinformatics, Data Visualization

Cloud technologies and version control: Docker, Singularity, Slurm, Snowflake, Spark, AWS (EC2, S3, Redshift), GitHub, MLflow

Certifications: [Machine Learning](#), [Deep Learning](#), [SQL](#), [Financial Markets](#), [PGP in AI & ML](#)

PROFESSIONAL EXPERIENCE

Dartmouth University

Machine Learning Intern

Lebanon, NH

Jun 2023 – Dec 2023

- **Enhanced cell detection** in a pathology image dataset by developing a **YOLOv8 based segmentation model** with **Dask** and **PyTorch** in collaboration with cross-functional team of medical professionals reaching **99% precision** and **84% recall**.
- **Implemented a Graph Neural Network (GNN)** for connected components analysis through **ResNet-50** based detected cell embeddings for **cancer cell classification** improving the existing **accuracy** by **22%**.
- Led a project to enhance gene mutation reporting using **entity recognition** on PubMed journals, **deploying an end-to-end ML pipeline with Singularity containers** to improve genetic data features for multimodal analysis alongside image processing.

Kelley School of Business

Research Fellow

Bloomington, IN

May 2023 – Aug 2023

- Developed a robust system using **PySpark** to perform topic mapping and count aggregation for more than **1 million web domains**, providing a comprehensive understanding of web domain interests related to financial modeling, **reducing the compute time by 70% for** near real time insights.
- Leveraged advanced techniques, including **anaphora resolution**, **Named Entity Recognition (NER)** with **Levenshtein distance** similarity, and **FinBERT**, a state-of-the-art finance language model, to extract structured information from over **1 million named entities** and perform sentiment analysis on **120,000 financial reports**, to better map reading interests to high return strategies.
- Optimized data processing pipelines through multithreading using the **Concurrent** library, resulting in faster and more efficient analysis of entity linked sentiments in the context of visualizing temporal trends in financial markets.

Neuberg Supratech

Data Scientist

India

Aug 2020 – Aug 2022

- Led the discovery of crucial biomarker genes feature sets for IVF implantation windows through sophisticated **differential gene expression** analysis using **DeSEQ2** and **Wald-Binomial** statistical test in R programming language.
- **Achieved up to 100% accuracy** in critical classes for pinpointing implantation windows in IVF patients by pioneering a high-performing classification model utilizing **XGBoost**, **Support Vector Machines**, **Random Forest**, and **scikit-learn** with identified genes, ensuring state-of-the-art testing results.
- Significantly accelerated the diagnostic process with an **80% reduction in total turnaround time**, by deploying an **ELT pipeline paired with data science pipeline**.
- Delivered extraordinary revenue growth of **\$100k** within the first-year post-algorithm deployment, showcasing the algorithm's clinical value and commercial success, in a lab handling **\$139 million** in annual revenue and serving **1.8 million patients** annually.

ACADEMIC PROJECTS

Algorithmic trading using Deep learning (*Springer Nature*, DOI: [10.1007/s11042-022-12328-x](https://doi.org/10.1007/s11042-022-12328-x)): Conceptualized and developed a predictive algorithm for the NIFTY 50 index using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models via TensorFlow, incorporating 60 diverse input features. Achieved a low Mean Absolute Percentage Error (MAPE) of 1.1% over ten years of trading data.

Synthetic cancer tissue slide generation: Used **Style-GAN2** to generate synthetic whole-slide images for cancerous tissue to tackle the lack of data for computer vision pipeline due to HIPAA regulations. Used YOLO-v8 cell detection trained on 2.2 million instances to provide a comprehensive foundation for future computer vision in pathology training.

Sentiment Analysis of COVID-19 Tweets: Scraped Twitter data with Tweepy and Twitter API, preprocessed using **snowball** and **NLTK**. Employed **collapsed Gibbs sampling** for topic clustering and cosine similarities to identify topics from predefined lexicons. Developed an analytics pipeline to identify needy regions during COVID-19 second wave using tweet sentiments in India.

Brain fiber reconstruction using Diffusion MRI (www.dipy.org): Simulated diffusion MRI signals using DIPY multi-tensor simulation module and trained a Deep Learning model to accurately resolve axonal fiber crossings under 30 degrees. Implemented a novel loss function to tackle sparse classification targets. Article to be submitted to NeurIPS 2024 under the tutelage of [Dr Eleftherios Garyfallidis](#).