

Divya Dhaipullay

Data Scientist & Machine Learning

Indianapolis, IN • (812)-543-3634 • ddhaipul@iu.edu _
<https://divyadhaipulle09.github.io/divyadhaipullay.github.io/>
<https://www.linkedin.com/in/divya-dhaipulle/>

Available to work immediately • Open to relocation within the US.

- Enthusiastic Data Scientist with a strong aptitude for tackling advanced computer vision challenges in object detection, classification, and segmentation. Demonstrated proficiency in deploying image recognition solutions in real-time environments, backed by **two years of focused research**.
- Pioneering efforts in Large Language Models (LLMs) to address challenges in knowledge representation and reduce hallucinations for next-generation conversational AI. Skilled in prompt engineering, fine-tuning, and constructing knowledge graphs through advanced Natural Language Processing (NLP) techniques.
- Extensive background in designing and implementing comprehensive machine learning systems for production, with a strong command of both supervised and unsupervised models. Possesses a deep understanding of the mathematical principles underpinning ML algorithms, fostering innovation and research capabilities.
- Proven expertise in executing thorough data cleaning, manipulation, and advanced analytics to extract critical insights that inform business strategies and guide organizational decision-making. One such innovative model is actively deployed on the NCPA platform. **LIVE on the NCPA website.**

TECHNICAL SKILLS

Programming Languages:	Python Java C MATLAB R (Statistics)
Machine Learning	Regression Models (Linear, Logistic, Decision Trees, Random Forest) Classification Models (K-means Clustering, Logistic Regression, KNN) XGBoost Gradient Boosting Methods Ensemble Models Support Vector Machine
Deep Learning:	Neural Networks (Convolutional Neural Networks (CNN), Deep NN, Graph NN, Recurrent NN) Word/Image Embedding Techniques LSTMs, GRU Gen AI (Generative Adversarial Networks, Transformers, Self-attention models, Vision Transformers).
Computer Vision	Object detection, image classification, semantic segmentation (R-CNN Family, Mask R-CNN, YOLO, SSD) EfficientNet, ResNet, VGG Vision Transformers.
Natural language Processing	BERT Language Models GenAI, GANs, NLU, Gensim, Topic Modelling, LDA
Large Language Models	OpenAI, LLAMAv2, Falcon 180B, Claude AI, LangChain, Prompt- Engineering (Text, Vision), LangChain, OpenAI, GPT, Dream Booth, OWL-ViT, Comet_ML, LoRA, TinySolar.
ML and DL Tools:	PyTorch TensorFlow Keras HuggingFace OpenCV NLTK Spacy Numpy Pandas Matplotlib Seaborn scikit-learn/sklearn Scipy CUDA XGBoost Plotly
Data Engineering & cloud:	SQL PySpark Azure (Databricks, Data Factory, Data Lake) AWS (S3, EC2, Lambda, Sagemaker).
Data Visualization & other tools:	Power BI Statistical Analysis using R (ggplot) Altair Hypothesis testing Microsoft Office (Advanced MS Excel) JIRA CUDA Git GitHub GitLab Putty

WORK AND RESEARCH EXPERIENCE

Tata Elxsi, Senior Engineer

Sept 2018 - July 2022 | Bangalore, India

Enhanced decision-making and operational efficiency for clients like **Caterpillar Inc., Jaguar Land Rover, Defense Research Development Organization (DRDO) and National Centre for the Performing Arts (NCPA)** using data models

- Developed and deployed a facial recognition model on AWS using CNN and VGG on AWS, achieving **95%** accuracy.
- Reduced manual verification time by **60%** and enhanced security protocols lead to a **40%** decrease in fraud incidents.
- Enhanced system reliability by increasing operational reliability by **10%** and reducing failure rates by **12%** for **DRDO**.
- Increased customer engagement by **28%** and ATC (add-to-cart) rate by **12%** by developing personalized recommender systems using XGBoost regressor with **97%** accuracy on historical sales data of **NCPA**.
- Projected a **25%** increase in customer retention and reduced churn rate by **20%** using customer churn prediction model.
- Enhanced region-based customer segmentation using DBSCAN algorithm and achieved a high cluster similarity of **93%**.

- Identified top-ranked item categories for each user, thereby triggering personalized recommendations to different clusters.
- Created a comprehensive supply chain management dashboard for *Jaguar Land Rover*, reducing overhead by **30%**.
- Processed large datasets to predict failures, improving system efficiency by **14%**, reducing maintenance costs by **10%**.
- Boosted sales by **32%** by identifying top-performing product categories and implemented targeted user recommendations.
- Designed an SQL model that matched user preferences with **94%** accuracy, improving personalized shopping experiences.
- Led PMST activities on Antenna/PIVI, enhancing component management efficiency, reducing delivery time by **20%**.
- Designed and implemented a Power BI dashboard for *Caterpillar*, consolidating **20** release operations into **one** interface.
- Conducted advanced data analytics in Python and SQL, identifying KPIs that improved operational efficiency by **15%**.

PROJECTS

Sentiment Analysis of Tweets Extracted on Covid-19 Pandemic

Jan 2023 – May 2023 | Bloomington, Indiana

- Modeled Twitter data for sentiment analysis and topic modeling to identify prevailing emotions and key discussion topics in the comment section, leveraging NLTK, Empath, and LDA

PySpark and GCP Data Pipeline for NYC Transportation Analysis

Jan 2023 – May 2023 | Bloomington, Indiana

- Developed a comprehensive Big Data analytics pipeline on Google Cloud Platform using PySpark for exploratory data analysis, preprocessing, and predictive modeling on NYC yellow taxi data

COVID-19 Data Pipeline and ETL Automation

Jan 2024 – May 2024 | Bloomington, Indiana

- Developed an automated ETL pipeline for COVID-19 data integration, transformation, and storage, ensuring accurate and up-to-date datasets for analytical purposes.

Azure Data Pipeline and Transformation for Tokyo Olympics Data

Aug 2024 – Dec 2024 | Bloomington, Indiana

- Implemented end-to-end data integration and transformation for the Tokyo Olympics dataset using Azure Data Factory and Databricks, including pipeline design, data synchronization, and PySpark-based processing.

TEACHING EXPERIENCE

Luddy School of Informatics, Computing, and Engineering

Jan 2023 – May 2024 | Bloomington, Indiana

- *Machine Learning in Bioinformatics (Fall 2023)*: Graded assignments, guided projects, and answered students' queries.
- *Embedded Systems (Spring 23,24)*: Conducted labs, debugged code, graded assignments, and addressed student queries

EDUCATION

Indiana University Bloomington, Masters in Data Science, GPA: 3.69/4

Aug 2022 - May 2024 | Bloomington, IN

Indiana University, Research Assistant

May 2023 - Present | Bloomington, IN

- Generated 4k lane image segmentation datasets using MobileSAM, SAM, Semantic-SAM, FastSAM, and Mask2Former.
- Annotated 1k images in Roboflow for lane segmentation in Vehicle Autonomous Intelligence Lab under Prof. Lantao Liu.
- Achieved an MSE of 1.2, R^2 of 0.82, and correlation coefficients ~ 1 in MDS coordinate prediction using DenseNet, Beta - VAE, ResNet, and Hugging Face Transformers (ViT, BEiT), I-JEPA in the Cognitive Science under Prof. Zoran Tiganj.
- Increased processing efficiency by 30% for whole slide image analysis with optimized GAN and Cycle GAN models on BigRed200/Slate HPC systems in the Computational Biology Lab under Prof. Xuhong Zhang.

Fraunhofer Institute for Industrial Engineering IAO, Research Assistant

June 2023 - Present | Remote

- Achieved 80% accuracy in anxiety detection using EEG and RES signals by implementing ML algorithms (SVM, KNN, Random Forest, Naive Bayes, XGBoost) with an opinion pooling technique.

University of Texas at Austin, Post-Graduation in Machine Learning, GPA: 4.00/4.33

2021 - 2023 | Remote

Visveswaraya Technological University, Bachelors in Electronics and Communication Engg., GPA: 3.76/4.00

2014 - 2018 | India