Divyesh Pratap Singh

+1 (716) 573 1947 | dsingh27@buffalo.edu | github.com/Divyeshpratap | linkedin.com/in/divyesh-pratap-singh | My Personal Website

May 2025 Masters in Artificial Intelligence graduate seeking Full Time roles in Machine Learning and Data Science.

Profile

- **Data Scientist** with **5 years of full-time experience** in automotive and credit card industry.
- Working as **Senior Research Assistant** in UB CUBS and HBML Lab on Active Speaker Detection, & Child Speech modelling.
- Awarded **Best Graduate Engineering Trainee** in Suzuki Motors, and **Best Performing Decision Scientist** in EXL Banking vertical.

Experience

CUBS (Center for Unified Biometrics and Sensors), UB, *Senior Research Assistant | NY, USA)* GitHub ***July 2024 - Present***

- **AVATAR**: Audio Visual Active Tracking and Annotation Rendering (Advised by Dr. Venu Govindaraju)

- Modelled AVATAR, a near **Real-Time Multi Modal Active Speaker Localization** and Transcription Framework using a **Producer Consumer Multi-Threading** Architecture, achieving a **latency** of **< 0.5 seconds**.
- Producer thread performs **Face tracking** on live camera feed using **S3FD, InsightFace** and **SORT**. Trained a **25 M parameter** speaker separation model on **Oxford Voxceleb** & **Google AVSpeech** dataset consisting of **2M videos** to give **Lip Sync** scores with **0.89 F1** score.

The Research Foundation SUNY, University at Buffalo, *Research Aide | NY, USA)* GitHub ***April 2024 - Present***

- **ChildSpeak**: LLM Powered Speech Pattern Analysis for Child Language Development (Advised by Dr. Ifeoma Nwogu)

- Designed a robust causal modeling pipeline to identify a toddler as Late Talker or Typically Developing using parent child audio conversations.
- Implemented three key components: Audio **transcription** and **diarization** using OpenAI **Whisper** and Pyannote, identification of **novel part-of-speech** tags, and **causal inferencing** to analyze word usage representative of late talkers.
- Annotated a large dataset comprising **2.5 million words** and a **40K vocabulary** across new categories: Shape/Non-Shape nouns and Result/Manner verbs. Leveraged **GPT 4o-mini** and **LLama 3.1** 405B models to **annotate** the dataset with a proxy tree-of-thought **prompt**.
- **Finetuned Roberta** base, achieving **94% accuracy** in **classifying Shape/ Non-shape nouns** and **91%** on **Result/ Manner Verbs**.

EXL Analytics, *Data Science Lead Assistant Manager | Hybrid (Gurugram, India)* ***November 2021 - June 2023***

- **Continuous Data Integrity Tool** (Cross-Functional Collaboration, Data Streamlining, A/B testing).

- Implemented robust **data quality controls** to bolster integrity of all Credit & Fraud risk decision science models in American Express. This automates extract & transform (**ETL**) pipeline to safeguard against **data discrepancies** at an early stage.
- Engineered via a scalable **anomaly** detection framework (ADF) that **ensembles four time series** algorithms, optimized for high-volume data at the scale of load append **3 million transactions** per day.
- Leveraged **Hadoop** for optimal data management and **Hive/ PySpark** for querying, enabling the big data pipeline to process millions of rows across thousands of features within minutes, ensuring real-time data fidelity.
- Attained an impressive **81% mean outlier alert accuracy** on 240 plus datasets, **reducing manual verification** efforts by **30%**.

Suzuki Motors India Limited, *Analytics Deputy Manager | Onsite (Gurugram, India)* ***July 2017 - November 2021***

- **Inventory Management & SKU Demand Forecasting** (Process Optimization, Traceability, Data Storytelling, & Dashboarding).

- Crafted a forecasting model by integrating ABC and XYZ analysis to provide **live insights on SKU availability, safety stock, & reorder points (**using **Power-BI)** for over 4,000 parts. Integrated **SARIMA** with **LSTM** framework to predict delivery times of critical spare parts, enhancing the Order to Delivery Process Analysis and Inventory Optimization. Awarded with Best GET.
- Increased **inventory turnover ratio** from **3.2 to 4.7** leading to annual cost savings of **$120K** by reducing wastage.

Projects

Automobile Inspector: AI-Powered Car Damage Detection and Chatbot Platform GitHub ***June 2024 - August 2024***

- Deployed an AI driven web application that integrates computer vision and NLP to provide comprehensive car repair solutions, from visual damage assessment to a **Retrieval-Augmented Generation (RAG)** based conversational chatbot.
- Engineered segmentation pipeline uses **Mask R-CNN** with **ResNet-101** backbone, and **Deformable Convolution Networks (DCN)** to accurately localize vehicle damages into dents, scratches, broken lamps, glass shatter, flat tire, and cracks.
- Designed **carBot**, a context aware **chatbot** leveraging **chat history** to enhance prompt accuracy using **Ollama**. Documents are stored in Facebook's **FAISS vector store** for efficient retrieval and processed via Llama for context-specific responses.
- Architected a user-centric application using **Flask**, enabling damage detection, cost estimation & repair requests in just **5 clicks**.

Education

3.8/4.0 **MS in Artificial Intelligence**, *University at Buffalo, The State University of New York* | NY, USA ***August 2023- May 2025***

8.2/10 **B.Tech in Electrical Engineering**, *Thapar University* | India ***July 2013-June 2017***

Courses: Machine Learning, Pattern Recognition, Computer Vision & Image Processing, Fundamental of A.I., Reinforcement Learning, Numerical Math, Advanced Algorithm Analysis & Data Structures, Data Intensive Computing, Computational Linguistics, Information Retrieval, Linear Algebra

Skills

Programming:	**Python**, **HIVE**, **PySpark**, **SQL**, Shell scripting , HTML, LaTeX, Apache Solr, Java
Machine Learning:	**Transformers**, Parameter Efficient Training, **LORA**, **Predictive Analytic**, Feature Engineering, Clustering, Advanced Statistics.
Framework/ Library:	**PyTorch**, TensorFlow, **Hugging Face**, **Hadoop**, **LangChain**, **sklearn**, **pandas**, DataParallel, OpenMMLab, CUDA, SpaCy.
Information Tech:	**Linux**, Cloud Computing, Distributed Computing, **Big Data**, OpenShift, GIT, Postman, GCP, **Agile**, **JIRA**, Kanban.