Devika Gumaste

516-841-6033 dg3370@columbia.edu linkedin.com/in/devika-gumaste devika3370 devika3370.github.io

Education

Columbia University Expected May 2025

Master of Science - Electrical Engineering (Machine Learning Research | MS Honors Scholar) (4.1/4.0) *New York, NY*

- **Key Coursework:** Advanced DL, DL on the Edge, Applied ML, Mathematics of DL, Algorithms, NLP, Statistical Learning

BITS Pilani Aug 2021

Bachelor of Engineering - Electronics and Communication Engineering (8.3/10) *Goa, India*

- **Key Coursework:** Digital Signal Processing, Embedded System Design, ML for Communication Systems, Computer Architecture

Technical Skills

Languages: Python, R, SQL, C, C++, MATLAB, HTML/CSS, Bash
Technologies/Frameworks: AWS(S3, SageMaker, ECS, EC2, Lambda), Google Cloud Platform, PostgreSQL, MySQL, ROS (Robot Operating System), Flask, RESTful API design, Docker, ONNX, Nvidia Jetson SDK, LangChain, HuggingFace, Linux
Libraries: Tensorflow, PyTorch, Keras, Scikit-Learn, NumPy, Pandas, TensorRT, CUDA, OpenCL, Dask, Matplotlib, openCV
Tools: Git, Bitbucket, JIRA, Excel, MATLAB, Jenkins, Nexus, TeamCity, Jupyter Lab

Experience

COSMOS Lab, Columbia University May 2024 – Present

Graduate Research Assistant *New York, NY*

- Conducting cutting-edge research under the guidance of Prof. Zoran Kostic at Columbia University. Currently a member of the Cosmos Computer Vision team. Research is funded by Center for Smart Streetscapes (CS3) NSF grant.
- Developed and published a novel distributed architecture for Vision Language Models (VLMs) at PerCom 2025, which partitions model components between heterogeneous devices. This research optimizes real-time processing and computational efficiency, demonstrating a 33% throughput improvement and scalable deployment of VLMs on NVIDIA edge computing platforms without traditional model compression techniques.
- Advancing monocular 3D object detection through innovative modeling and algorithmic strategies, aiming to enhance accuracy and robustness using single-camera inputs in resource-constrained environments. Developed a novel 3D object detection approach by leveraging YOLO pose estimation models and re-purposing key points to define 3D bounding box corners from RGB images.

ZS Associates Jun 2021 – Jul 2023

Senior Software Engineer (Business Technology Solutions) *Pune, India*

- Led a 5-member cross-functional team in the design and development of a tool to optimize the placement of pharmaceutical sales representatives to territories based on geographical and business factors, using Angular (frontend), Python Flask (backend), and PostgreSQL (database). Implemented data ingestion workflows, dynamic configuration features, and custom SQL query generation. Optimized memory usage and API performance, and developed boundary-based geographical visualizations using Geopandas and Google Maps APIs. Fostered an Agile environment with Scrum, setting clear goals and empowering team members, delivering the project 11.6% ahead of schedule.
- Developed and deployed machine learning-based applications to simplify complex ML workflows for end users with an intuitive UI in Jupyter Notebooks. Implemented regression, clustering, decision trees, random forests, and dimensionality reduction algorithms, incorporating hyperparameter tuning, SHAP-based interpretability, and result visualizations to improve model performance and decision-making.
- Designed and implemented an automated diagnostic tool for EDA and advanced analytics, creating modules for data ingestion, transformations, and over 50 visualizations. Leveraged Dask to scale Python code for large distributed clusters, ensuring efficient data handling, with FastAPI as the backend and a Jupyter-based frontend. Implemented parallel processing for multiple diagnoses, reducing processing time by 35%.

Projects

Cautious Speculative Decoding, *Columbia University* Dec 2024

- Developed a novel pipeline combining safety-first content filtering with speculative decoding, using a small draft model fine-tuned with LoRA for rapid toxicity filtering and token generation.
- Achieved significant improvements in toxicity detection (F1 score from 0.1682 to 0.8205) and computational efficiency (35% speedup) while maintaining high-quality text generation (BERTScore of 0.71), demonstrating an effective balance between safety and performance.

Arrhythmia Detection using Vision Transformers, *Columbia University* May 2024

- Implemented vision transformers for arrhythmia detection from ECG images. Adopted knowledge distillation methods to condense the model from 12-lead ECG classification into one suitable for single-lead data while preserving accuracy.
- Conducted comprehensive experiments on Google-ViT, BEiT (BERT for image classification), Swin Transformer and Meta's DEiT to evaluate their performance on arrhythmia detection, achieving accuracy up to 98%.

Real-time Pose Estimation and Correction on Edge Devices, *Columbia University* May 2024

- Designed a system for identifying and analyzing yoga poses, achieving 94% accuracy. Leveraged deep learning models to locate and track key body landmarks, providing real-time feedback on pose alignment and correctness.
- Deployed on NVIDIA Jetson Nano™ with a processing speed of 18 FPS, the application provides real-time feedback and corrective guidance, with a user-friendly interface for seamless interaction.