SARAVANAN HARI BASKARAN

(332) 258-6908 | sh4621@columbia.edu | linkedin.com/in/saravanan1999/

EDUCATION

Columbia University *(GPA 3.7/4.0)* — New York, NY, USA
Master of Science in Data Science — Expected Dec 2025
Vellore Institute of Technology *(CGPA: 4.0/4.0)* — Vellore, Tamil Nadu, India
Bachelor of Technology in Computer Science and Engineering — May 2022

SKILLS

Programming Languages: Python, Java, Go, PHP, C/C++, R, MATLAB, SAS
Front-End Technologies: HTML, CSS3, JavaScript, ReactJS, AngularJS.
Frameworks / Tools: OpenCV, Sklearn, PyTorch, TensorFlow, Tableau, XGBoost, LightGBM, SciPy, NumPy, Pandas, Keras, Matplotlib, Seaborn, ggplot2, Git, MySQL, Jenkins, Kafka, Cassandra, Redis, AWS, Firebase, MongoDB, BeautifulSoup, Microsoft Excel, Apache Spark, REST, gRPC, Airflow, Snowflake.
Area of Expertise: *Machine Learning, Deep Learning, LLMs (Large Language Models), Data Analysis, Predictive Modeling, NLP, Computer Vision, Time Series Analysis, Feature Engineering, A/B Testing, Statistics, Data Structures and Algorithms, ETL Pipelines*

WORK EXPERIENCE

Arklex.AI — New York, New York, US
Machine Learning Engineer Intern — May 2025 - Present

- Took the lead in architecting a modular **LLM-powered agentic framework** with **Retrieval-Augmented Generation (RAG)** support, enabling context-grounded response generation using enterprise-specific data sources.
- Led development of real-time configurable agent personas, including dynamic tone modulation, sentence complexity, and behavioral rules, allowing deployment of adaptive, brand-aligned AI agents tailored to user context and intent.

MakeMyTrip India Private Limited — Bengaluru, Karnataka, India
Data Scientist — Jan 2022 - May 2024

- Benchmarked sentiment models including BERT, TF-IDF with Logistic Regression, and RoBERTa, and worked collaboratively to finalize DistilBERT for its performance–efficiency balance while helping develop an internal NLP engine revealing complaint themes and sentiment trends in hotel reviews.
- Took initiative to redesign tax document OCR pipeline combining advanced image pre-processing (adaptive thresholding, skew correction) with a fine-tuned **CRNN** model and **LSTM**-enhanced Tesseract, increasing invoice data extraction accuracy from 57% to 83.33% across highly unstructured document formats.
- Drove development of end-to-end deep learning–powered photo moderation pipeline to flag low-quality, policy-violating, or misleading hotel listing images by integrating **ResNet50** for content quality classification, OCR for text overlay detection, perceptual hashing for duplicate identification, and rule-based filters for category mismatches lessening manual review effort by ~40% and improving listing integrity.

Verzeo EduTech — Remote
Data Science Intern — Apr 2020 - May 2020

- Built a high-accuracy student performance prediction pipeline by honing **Random Forest, XGBoost, and LightGBM** models with k-fold cross-validation and hyperparameter tuning, achieving $R^2 = 0.93$ for forecasting outcomes such as course completion and engagement levels.
- Streamlined adaptive data preprocessing with IQR-based outlier removal, dynamic encoding, and missing value imputation to handle noisy LMS and enrollment data; led data quality initiatives using **NumPy**, **SciPy**, and **Pandas** to enforce statistical validation, bolstering model reliability across diverse student segments.

PROJECTS

Optimized Deep Learning Pipeline for Pneumothorax Detection: Accelerated Preprocessing & U-Net Segmentation

- Engineered an automated image preprocessing pipeline for 12,000+ chest X-ray images, leveraging **PyTorch**, **PIL**, and **torchvision** to standardize dimensions, normalize pixel intensities, and apply CLAHE-based contrast enhancement, ensuring high-quality inputs for deep learning models.
- Identified and addressed performance bottlenecks in dataset preprocessing, achieving a 66% reduction in per-image processing time and boosting throughput to 11.26 images/sec through batched transformations, multiprocessing, and memory-efficient disk I/O.
- Trained deep learning models for medical image segmentation, evaluating **U-Net**, **ResNet**-based encoders, and **EfficientNet** to balance computational efficiency and accuracy, selecting U-Net with a ResNet-34 backbone for its spatial detail retention.

Cyberbullying Detection via Multi-Stage NLP Pipeline

- Designed a multi-model NLP pipeline for robust cyberbullying detection, combining a hate speech classifier (HateBERT) with a BERT model fine-tuned on cyberbullying data, achieving 95% recall through weighted ensemble voting and enriched contextual understanding of abusive language.
- Identified gaps in traditional text classification and introduced a hybrid approach combining NER (spaCy), sentiment, and sarcasm detection (DistilBERT) to improve precision in challenging cases of indirect and ironic bullying.
- Implemented **explainable AI (XAI)** techniques using **SHAP** and **LIME**, providing model transparency and interpretability for flagged messages, enabling responsible AI and trust in AI-powered moderation systems.

NYC Taxi Data Pipeline: Scalable ML & Real-Time Insights

- Spearheaded development of a distributed data processing pipeline using **Apache Spark** on **AWS EMR**, to accelerate processing of 10M+ NYC Taxi Trip Records by 70%, facilitating real-time analytics for data-driven decision-making.
- Developed and deployed scalable ML-integrated APIs, utilizing **FastAPI**, **Docker**, and **AWS EKS** to serve models with 93.8% accuracy, decreasing response times by 60%, and enhancing user engagement through a ReactJS-powered frontend, increasing satisfaction scores by 35%.
- Automated ML deployment workflows, implementing CI/CD pipelines with GitHub Actions and optimizing data pipeline triggers using **AWS Lambda** and **CloudWatch**, reducing deployment cycles by ~50%.
- Augmented system observability and reliability through Grafana dashboards and ensured secure API access via OAuth2-based authentication protocols.