Harichandana Neralla

Harichandana.Neralla@utdallas.edu| harichandana10899@gmail.com | +1 (332)-254-4284 | https://harineralla.github.io/

EDUCATION

Master of Science, Computer Science (Specializing in Data science) **August 2022 – April 2024**

The University of Texas at Dallas GPA – 3.77

***Course Work*:** *Natural Language Processing, Machine Learning, Deep Learning, Data Mining, Big Data Analytics*

Bachelor of Science, Computer Science and Engineering **June 2016 – May 2020**

KL University, Guntur, India GPA - 9.22/ 10

Publication - Expression Invariant Features for Face Recognition – IJEAT https://www.ijeat.org/portfolio-item/B3099129219/.

CERTIFICATION

AWS Certified– Solutions Architect Associate - https://www.credly.com/badges/1ef68e5e-4946-4e57-ba9c-5b51a3499da/public_url

THESIS RESEARCH

The University of Texas – Dallas, *Machine learning, Deep learning, Natural Language Processing* **August 2023 – Present**

- As a part of the research work, worked on ensembling deep learning models from multiple SOTA research papers to solve/study the code summarization, code description generation, on combining cross-modal transfer learning with large language models like LLaMa, CodeBERT, PaLM, GPT-3.5 and GPT-4.
- Implemented a Git-based version control workflow to track changes to the codebase, model configurations, and experiment logs across different model ensembling approaches.
- Designed an automated process for Story Intention Graph (SIG) generation from code summaries, using prompting on LLMs for intermediate outputs. Implemented GraphCNN to generate encoded representation of the SIGs for autoregressive transformer decoding.
- Conducting research work on SOTA open-source tools, frameworks to develop on them using innovative research methodologies to detect vulnerable code and generate software patches.
- Study includes Bugsplainer a Vue.js, Flask Application which leverages code structures to explain software bugs with Neural Machine Translation offering 3 solutions 60M, 220M and fine-tuned Code T5.
- Implementing a robust data ingestion and preprocessing pipeline to handle the large dataset of 150K git commits and PySecDB, ensuring data quality and consistency.

TECHNICAL SKILLS

Languages: Python, R, Matlab, Scala, C, C++, Java.

Deep Learning Libraries: PyTorch**,** NumPy, Pandas, Matplotlib, TensorFlow, NLTK, OpenCV, CoreFlow, Sklearn, PySpark, MLlib, Keras.

ML and DL Skills: Regression, Clustering, SVM, Decision Trees, Naïve Bayes, Random Forest, NLP, CNN, RNN, LSTM, GAN, YOLO.

Other Frameworks for Analytics: Spark, Hadoop, Hive, DataBricks, ETL, AWS Services, Flask, Tableau, Excel, PowerBi.

PROFESSIONAL EXPERIENCE

University of Texas at Dallas – *PMClub Data/Cloud Engineer Intern* **May 2023 – August 2023**

- Designed and implemented a comprehensive data analytics and visualization tool utilizing React Js, Python and MongoDB.
- Improved Node.js backend response time 10x by optimizing data handling with MongoDB, ElasticDB and DynamoDB.
- Digitalized on-prem ETL spark jobs to AWS cloud reduced response time from 4hrs to 2hrs using AWS Glue, and data catalog.
- Optimized data architecture by migrating all S3 data to a unified data lake house with Databricks integration.

Innova Solutions *(formerly ACS Solutions),* **Hyderabad, India -** *Software Engineer* **December 2019 – July 2022**

- Architected, devised, and produced multiple web applications and microservices, utilizing Python, Flask, and REST APIs for backend, as well as MongoDB for data storage. Wrote comprehensive unit tests achieving 97% coverage.
- Optimized data workflows utilizing Kafka messaging, improving cut latency and increased transaction volume by 2x.
- Spearheaded the design and development of 60 high-impact UI applications and trading dashboards using React Js, MERN Stack, and Redux, ensuring robustness through rigorous testing with Cypress.
- Produced orchestration using Kubernetes and Docker, and automated scaling the deployment of projects utilizing AWS EC2.
- Created script for memory management resolving 60% of server issues based on AWS CloudWatch monitoring analysis.
- Managed high-traffic ELK stack for log aggregation using kibana, reducing downtime by 50%, and enhancing monitoring.
- Created script for memory management resolving 60% of server issues based on AWS CloudWatch monitoring analysis.

PROJECTS

Online Shoppers Purchase Intension Prediction *(Python, Scala, SVM, Random Forest, XGBoost)* **January 2023 – April 2023**

- Developed customized implementations of Decision Trees and Support Vector Machines (SVM) with linear and Gaussian kernels to predict online shoppers' purchase intent from session data. Compared multiple ratios of positive to negative samples and achieved best accuracy of 88.3% using decision trees.
- Employed ensemble and multi-class classification models in Scikit-Learn including Gradient Boosting (87.9% accuracy), Random Forest (88.4%), SVM with RBF kernel (88.3%), and Extreme Gradient Boosting (XGBoost - 88.5%) to optimize purchase prediction from user sessions. XGBoost provided the strongest performance.
- Architected complete predictive pipeline integrating data preprocessing, feature engineering, grid search hyperparameter tuning and cross validation for rigorous evaluation. Analyzed performance using metrics like accuracy, recall, precision, F1 score, ROC curves for 12K user sessions.
- Utilized MLflow for experiment tracking, logging relevant metrics, parameters, and artifacts across the different model combinations to analyze performance and facilitate model selection.

RNN for Stock Market Prediction *(RNN, LSTM, Hadoop, Kafka, PySpark, MLlib, MatplotLib)* **January 2023 – April 2023**

- Developed Recurrent Neural Network (RNN) architecture with Long Short-Term Memory (LSTM) cells utilizing Apache Kafka and PySpark for distributed data processing within a Hadoop ecosystem resulting in 88% accuracy. Achieved lowest MSE loss of 49.4 on S&P500 dataset across 10 epochs leveraging PySpark for scalable data processing across 300K samples of historical stock market data.
- Engineered features from raw historical stock data and optimized data pipelines, applying techniques like MinMaxScaler for effective normalization. Reduced memory requirements by 70% compared to baseline.
- Analyzed model performance using evaluation metrics including Mean Squared Error (MSE) loss, plots of predicted vs actual values, and monitoring losses across training epochs. Identified optimal batch size and number of epochs to enhance model accuracy.

Expression Invariant Features for Face Recognition, Undergraduate Major Project **August 2019 – January 2020**

- Created face recognition application using Convolutional Neural Networks (CNN), C++ and OpenCV for face detection and segmentation. Compared use of PCA and LDA for dimensionality reduction before classification.
- Implemented 3 variations of Support Vector Machine (SVM) classifiers – linear, polynomial and RBF kernels for multi-class facial expression recognition across 5 categories. RBF kernel provided highest accuracy of 76.8% on dataset of 70 individuals.
- Engineered pipeline for data collection, preprocessing including resizing, normalization and shuffling of 7200 facial images. Published conference paper demonstrating model effectiveness in recognizing faces despite changes in facial expression.

LEADERSHIP EXPERIENCE

Graduate Teaching Assistant, *University of Texas at Dallas* **June 2023 – present**

- Instructing Database Systems and DSA for undergraduate students, grading their exams and assisting professor.

HackDFW by SYTD, *Sponsor Challenge Winner* (https://devpost.com/software/smoke-app) **September 2022**

- Won the Sponsor Challenge for developing an mobile application aimed at helping teenagers quit smoking/vaping, using React Native and Node Js.

KL University, *GDSC Member and Mentor at LINUX Club and Elite Club* **September 2018- December 2019**

- Spearheaded and taught 120 students on 60 Flutter open-source apps, developed ML applications on LINUX, for 4 years.