SHARVARI KALGUTKAR

Data Scientist

sharvari.kalgutkar82000@gmail.com | Los Angeles, CA | Tableau | Sharvari289 | sharvarikalgutkar | Portfolio

EDUCATION

Masters in Applied Data Science, *University of Southern California* **Aug 2022-May 2024**
Machine Learning, Data Mining, Data Management, Database Systems, Deep Learning, Data Visualization, Research Methods, Experimental Design and Analysis for User Studies, Fairness in AI and Responsible AI (**CGPA 3.85/4**)

Bachelor of Technology Electronics and Telecommunication, *Sardar Patel Institute of Technology, India* **Aug 2018-Jun 2022**
Data Structures & Algorithms, Statistical Computational Lab, Object Oriented Programming, Applied Mathematics (**CGPA 9.52/10**)
Minor: **Marketing** and Management, *SP Jain Institute of Management*

TECHNICAL SKILLS

Machine Learning and Deep Learning: Python, R, TensorFlow, PyTorch, NumPy, Scikit-learn, Pandas, Matplotlib, Plotly, Seaborn, OpenCV
Tools and Technologies: PySpark, Databricks, Hadoop, Tableau, Alteryx, Power BI, D3.js, Amazon Web Services, Docker, CVAT
Databases and Infrastructure: SQL, Firebase, MongoDB, XML, Excel, AWS S3, AWS RDS, AWS DynamoDB, PostgreSQL, Linux, Git
Professional skills: Statistics, Data Mining, Unsupervised/Supervised ML, Data Visualization a nd Analysis, Computer Vision, Big data, NLP

PROFESSIONAL EXPERIENCE (1 Year)

Data Science Researcher, *CKIDS University of Southern California* **Feb 2024-Present**

- Research **neural network** forgetting, its impact on learning from **non-IID** data distributions and energy efficiency.
- Train and evaluate neural network models using **TensorFlow** in **distributed computing** environment such as **federated learning** versus **round-robin** training, assessing performance across both IID and non-IID datasets**.**

AI Engineer, *Scientist Technologies* **Nov 2021-May 2022**

- Cross-collaborated to develop 5 **Python**-based algorithms for road safety analysis in an **agile** environment, achieving a **91% R2 score.**
- Implemented **OpenCV video processing** for enhanced safety **visualization**, delivering a **3x efficiency** boost in quality checks.
- Devised **Computer Vision** Road quality tracking system using models like **Faster R-CNN**, yielding highest **precision** of **84%.**
- Orchestrated an end-to-end **ML workflow**, leveraging **AWS EC2** for efficient model training, managing **data quality, annotation** and **cleaning** through **CVAT**, and storing data using **AWS S3**.
- Automated **data migration** of **720+ hours** from **Google Drive to AWS S3**, using **Google Cloud REST API**, drastically reducing time.

Deep Learning Research Engineer, *Skinzy Software Solutions* **Oct 2020-Jan 2021**

- Constructed a **Mask-RCNN instance segmentation** model in **TensorFlow** to detect skin diseases, achieving an **IOU of 0.6**.
- Deployed a **ResNet-50 Transfer Learning** model for skin disease classification, yielding an **accuracy of 85%.**

PROJECTS

Starbucks Stores Analysis | *Data Visualization, Statistical Analysis, Dashboard, D3.js, Map box, HTML, CSS*

- Designed a dynamic **D3.js Dashboard** to analyze Starbucks' global store location strategy, KPI's optimizing decisions.
- Built a custom **Mapbox** Starbucks **store locator map** for LA, improving user navigation and accessibility to nearby stores.
- Executed global, country, and state-level analysis using **diverse data visualizations**, including Bar Charts, Scatterplots, Proportional Symbol maps, and Choropleth maps.

E-commerce Global Market Analysis | *Data Analysis, Python, Matplotlib, Seaborn, Plotly, Communication*

- Qualified as the **National Finalist** with a **rank of 7 out of 600+ teams** at the Business Data Analytics at IIT Delhi.
- Engineered impactful **data visualizations** using **Python**, **Matplotlib**, **Seaborn**, and **Plotly,** featuring **Barplots, Line Charts, Box Plots, Squarify plots, and World maps** to analyze pivotal sales trends across 6 e-commerce markets in a team of 3.
- **Communicated to stakeholders'** seasonality trends, customer retention, RFM analysis to identify top performing markets.

HappinessQ | *Data Analysis, Data Management, Firebase, NoSQL databases, MySQL, Hadoop MapReduce, Flask, JavaScript*

- Built **Firebase and SQL**-based **distributed file storage** for analyzing the World Happiness Index, GDP and unemployment.
- Engineered a web-based command-line interface in **Python** and **JavaScript** for manipulating user-uploaded files, enabling commands like directory creation, reading file partitions etc.
- Deployed a **Flask** website with **Hadoop**-like partition-based **MapReduce** for faster parallel data search and analysis.

Yelp Business Recommendation System | *Data Mining, Big Data, PySpark, Collaborative filtering*

- Built an **Apache Spark** Recommendation System for user-business rating prediction for **1.5M** users and **200k** businesses.
- Executed **Item-based** and **model-based Collaborative filtering** using **XGBoost regression**, yielding **RMSE** of **1**.
- Created an enhanced **hybrid recommendation system** with **feature mining**, reducing **RMSE** to **0.97**.

PUBLICATIONS

Pneumonia Detection from Chest X-ray using Transfer Learning (Team Lead*)* **|** *Deep Learning, Image processing, Data Augmentation*

- **Led a team** of five to engineer three transfer learning models, namely **ResNet50, VGG-16, and Inception V3** to aid pneumonia detection with a maximum **recall of 98%** and **accuracy of 94%.**
- Employed **Image Processing** techniques, including **Data Augmentation**, to increase the dataset size by **5x.**

EEG Brainwave Emotion Detection Using Stacked Ensembling | *Feature Selection, Machine Learning, Python, SVM, Decision Trees*

- Optimized and trained 8 **Machine Learning** models, including Random Forest, Decision Trees, K-Nearest Neighbors, Logistic Regression, Support Vector Classifier, XGBoost, Lightgbm, and neural networks, resulting in **96% accuracy**.
- Developed a **Stacked Ensemble ML model** for emotion classification from EEG signals achieving improved **97% accuracy**.
- Conducted **Principal Component Analysis** to mitigate the dataset's high dimensionality, **reducing it by 94%.**