Isha Arora

Boston, MA | (617) 652-6728 | arora.isha4128@gmail.com | Portfolio | LinkedIn | GitHub

Education

Master of Science, Data Science — **December 2023**
Northeastern University, Khoury College of Computer Sciences, Boston, MA — GPA: 3.83/4.00

Bachelor of Technology, Computer Science and Engineering — **July 2020**
Vellore Institute of Technology, Vellore, India — GPA: 8.67/10.00

Technical Skills

Languages:	Python, R, SQL, MATLAB, NoSQL, C, C++, Java, JavaScript, HTML/CSS
Frameworks:	AWS, Google Cloud, Git, MySQL, PostgreSQL, PL/SQL, PySpark, Apache Kafka, Docker
ML Libraries:	TensorFlow, Keras, PyTorch, OpenCV, BeautifulSoup, SkLearn, SciPy, Plotly, HuggingFace, ggplot2, NLTK, dlib
ML Tools:	Tableau, Power BI, SPSS, SAS, VBA, Qlik, OpenRefine, MS Azure, Snowflake

Experience

Data Science Researcher — **May 2024 - Present**
Data Science Research Student — **January 2023 – August 2023**
Massachusetts General Hospital, Boston, MA

- Conducted comprehensive analysis to investigate the **impact of partner deployment** during wartime on the mental health of expecting and postpartum women, and examined the **influence of delivery mode** on postpartum mothers' well-being
- Evaluated PTSD detection models, correlating self-reports with clinical metrics, for accessible PTSD detection achieving 0.94 AUC-ROC in a study of **59 postpartum patients**, leading to first authorship on a paper in AJOG
- Optimized data visualizations for NIH presentations, executing t-tests that resulted in a **25% increase in data interpretability**, enhancing decision-making and clarity in research findings
- Synthesized insights from **over 100** Machine Learning publications, enhancing data quality for a study on CB-PTSD, aligning **8000 data points** with IRB standards, contributing to the model's development for identifying risk factors in health records

Data Science Research Assistant — **February 2023 – November 2024**
Northeastern University, Boston, MA

- Refined deep learning models, such as **EfficientNet, ResNet and PROMOTEO**, to boost prostate cancer grading precision, demonstrating a commitment to enhancing diagnostic accuracy through advanced technological solutions
- Elevated grading accuracy **significantly by 8%** developing an innovative custom EfficientNet-B1 model on **AWS cloud platform**, achieving **0.66 QWK** and **0.81 WA** (weighted accuracy), demonstrating impactful model optimization in medical image analysis
- Designed, validated, and deployed **software tools** to streamline **ETL** of whole slide images and RNA sequencing data in a **high-volume data environment**, contributing to the **optimization of data-driven decision-making processes**
- Investigated and applied **advanced data visualization** to convey and explain analytical insights effectively, fostering a **greater than 50% increase** in comprehension of results among team members

Associate Engineer — **August 2020 – August 2021**
Virtusa Consulting Services Pvt. Ltd, India

- Constructed a **PostgreSQL** system to host US banking and insurance regulations, improving lookup efficiency by **30%**
- Expedited client data retrieval by **45%** in **5 months**, surpassing targets; spearheaded a **modeling** project using OpenRefine, **leading** an **8**-member **cross-functional team** to a **40% increase** in analytical capacity
- Applied Agile methodologies and JIRA for project management, focusing on **data storage monitoring** to develop **visualizations**
- Fostered **collaboration** with Wolters Kluwer USA, enhancing delivery and inter-organizational ties

Data Science Intern — **December 2019 – May 2020**
Financial Software and Systems Pvt. Ltd, India

- Developed a spam detection system for a banking app reviews, attaining **65%** accuracy, reducing manual review costs by **26%**
- Progressed product development through review sentiment analysis employing VADER (Valence Aware Dictionary for sEntiment Reasoning), leading to a **40% increase** in positive feedback with a **15% rise** in user engagement
- Designed an AI algorithm for app review authenticity, **boosting relevance by 50%**, presenting **strategic impacts** detailed in a comprehensive **stakeholder report**

Data Science Projects

Speech Emotion Recognition — **September 2024 – Present**

- Initiating a project to analyze and classify emotions from speech patterns, aiming to enhance human-computer interaction
- Developing a preprocessing pipeline to extract relevant features from audio data, laying groundwork for emotion prediction
- Implementing and fine-tuning machine learning models to focus on improving the accuracy of emotion detection

Exploring User Accessibility and Human-Machine Interaction Using EMG — **October 2023 – December 2023**

- Crafted a user identification and gesture recognition model using Electromyogram data, aiding those with mobility issues

- Built XGBoost, neural network models for gesture recognition, securing **91% accuracy** alongside an **F-1 score of 0.9**
- Attained **94% accuracy** in LSTM user classification, with a notable **cross-day rank-5 accuracy of 80.3%**

The Song Search **October 2022 – December 2022**

- Implemented a music retrieval system, harnessing the GTZAN collection for a curated audio dataset, advancing music search
- Secured top 5 **candidate set accuracy of 74%** with **0.68 MAP**, utilizing TensorFlow MAGENTA's MT3 model (based on T5 architecture), highlighting enhanced precision in music information retrieval

Deep Clustering for Unsupervised Learning of Visual Features - A Reproduction **March 2022 – May 2022**

- Replicated Facebook AI Research's DeepCluster network, integrating Power Iteration Clustering (PIC) and AlexNet clustering with ImageNet dataset, encompassing 64 classes with 600 images, extending analysis to an external dataset of 28000 images
- Assessed Normalized Mutual Information (**NMI**) between clusters, reaching an **approximate value of 0.8**, indicating robust clustering

A Literature Review on BERT, RoBERTa, and T5 **March 2022 – May 2022**

- Analyzed distinctions between attention models BERT, RoBERTa, T5, delineating their characteristics and performance metrics
- Reviewed foundational papers of attention models, detailing the evolutionary improvements from BERT to subsequent models

Cryptocurrency Price Prediction **November 2021 – December 2021**

- Led a comprehensive project aimed at predicting daily price of cryptocurrency, leveraging 5 years of historical data
- Performed Data Analysis and Visualization on pricing using Python and Tableau, emphasizing daily price fluctuations
- Optimized a Random Forest model, achieving an **RMSE score of 0.222**, evidencing model accuracy

Facial Emotion Recognition **October 2021 – December 2021**

- Orchestrated development of a model classifying 28273 facial images into 6 emotions: anger, sadness, happiness, fear, surprise, and neutral
- Executed a range of classification algorithms including Decision Tree, Random Forest, Gaussian Naïve Bayes, k-NN, VGG-16, and incorporated the dlib library for image processing, demonstrating versatile modelling skills
- Realized **63% accuracy** benchmark with top-performing VGG-16 model, highlighting model optimization and application

Publications

Arora, Isha Hemant, et al. "A diagnostic questionnaire for childbirth related posttraumatic stress disorder: a validation study." American Journal of Obstetrics & Gynecology, July 2024.

From work at Massachusetts General Hospital as a Data Science Researcher

- Highlighted effectiveness of the self-reporting PTSD Checklist (PCL-5) as a preliminary diagnostic tool for CB-PTSD to significantly streamline and enhance efficiency of the diagnostic approach
- Established cutoff value of **28**, optimizing sensitivity (**0.8**) and specificity (**0.93**), which led to accurate diagnosis in **86%** of women

Mudavadkar, G.R., et al. "Gastric Cancer Detection with Ensemble Learning on Digital Pathology: Use Case of Gastric Cancer on GasHisSDB Dataset." Diagnostics MDPI, August 2024.

From work at Northeastern University as a Data Science Research Assistant

- Histopathological images from the GasHisSDB dataset with three image resolutions (160 x 160, 120 x 120, 80 x 80) used to classify them as normal or abnormal.
- Ensemble model (VGG-16 and ResNet-34) obtained an **average accuracy of more than 99%**
- ResNet50, VGGNet, and ResNet34 performed better than EfficientNet and VitNet, with the ensemble model continuously delivering higher accuracy