Shaan Pakala

San Jose, CA | Riverside, CA | shaan.pakala@gmail.com | (408) 891-0158 | [website, linkedin, github]

EDUCATION

University of California, Riverside — Sept. 2021 - June 2025

Data Science & Engineering, B.S – **3.91** Upper Division GPA

Coursework: Machine Learning & Data Mining, Artificial Intelligence, Database Management Systems, Algorithms for Bioinformatics, Statistics for Data Science I & II, Regression Analysis

RELEVANT EXPERIENCE

Data Science Research Intern – UC Riverside [github] — Summer 2024

- **Submitted paper as first author** to IEEE Big Data in December 2024 (under review)
  - Paper title: "Automating Data Science Pipelines with Tensor Completion"
- Funded by National Science Foundation (NSF) Research Experience for Undergrads (REU)
  - Under the guidance of Professors Vassilis Tsotras, Vagelis Papalexakis, and Jia Chen
- Leveraged tensor completion algorithms to **accelerate hyperparameter tuning** by inferring the results of all predefined hyperparameter combinations **(5x – 25x faster than gridsearch)**
  - Includes **deep learning architecture tuning** in addition to non-deep learning models
- Also applied this to estimate SQL query cardinalities for the purpose of query optimization

Bioinformatics Research Assistant – UC Riverside — March 2024 - Present

- Bioinformatics Research Assistant under Professor Stefano Lonardi
- Recreated, verified, and extending the work of other bioinformatics research papers
  - Applying Machine Learning algorithms to protein sequence analysis
- Fine-tuned large-language model for protein cleavage site prediction for drug discovery
- Revised previous input data for machine learning model to highlight patterns more easily

Data Science Challenge – Lawrence Livermore National Laboratory [github] — July 2023

- Designed Neural Network for automated heart disease diagnosis using ECG data with PyTorch
  - **Improved minority classes' accuracy ~10%** with custom loss function
- Reconstructed spatio-temporal map of hearts with a Convolutional Neural Network
  - Mapped ECG data from 12 leads over 500ms into heart activity of 75 locations over 500ms
- Composed EDA graphs and applied normalization techniques to highlight patterns in data
- Presented results at poster session with an audience comprised of lab scientists and interns

PROJECTS

Simulated Online Store Database [github]

- Implemented database for simulated online store (Database Management Systems final project)
  - Used Java for User Interface, as well as PostgreSQL for backend database functionality
- Integrated triggers, indexes, and private functions (e.g. admin and store manager functions)

Image Segmentation Projects [github]

- Developed custom **U-Net with skip connections** for image segmentation (using PyTorch)
- Segmented images of people, cells, and for multiclass segmentation of bears & deers
- Experimented with different image preprocessing techniques for better results (using OpenCV)

Coding Languages: Python, PostgreSQL, R, C++, Java, HTML, Tableau

Libraries: PyTorch, TensorFlow, tensorly, scikit-learn, OpenCV, Pandas, Numpy, matplotlib, tidyverse