# Zhirui Li

Personal Webpage: https://zhiruili1.github.io | GitHub: https://github.com/ZhiruiLi1 | zhirui_li@brown.edu | (805)878-6503


## EDUCATION

**Brown University** Providence, RI | Sept. 2021–Dec. 2023

M.S. Data Science, 4.0/4.0 GPA

- Relevant Courses: Computational Probability and Statistics, Mathematics for Data Science, Deep Learning, Machine Learning, Statistical Learning, Data Engineering, Computer Vision, Data Ethics, Blockchain and Cryptocurrency, Real Analysis, Measure Theory

**University of California, Santa Barbara** Santa Barbara, CA | Sept. 2017–Jul. 2021

B.A. Economics, B.A. Statistics and Data Science, 3.87/4.0 GPA

- Awards: Dean's Honor (2017–2021); graduated with High Honors (top 6%)
- Relevant Courses: Linear Algebra, Multivariable Calculus, Differential Equations, Transition to Higher Math, Object-oriented Programming, Data Structures and Algorithms, SAS Programming, Data Science in R, Probability Theory, Mathematical Statistics, Stochastic Processes, Markov Chain, Statistical Inference, Design of Experiments, Regression Analysis, Statistical Machine Learning, Microeconomic Theory, Macroeconomic Theory, Econometrics, Finance, Applied Economics, Big Data in Economics

## RESEARCH EXPERIENCE

### Interactive RNA-Seq Analysis via Shiny App & Gene Regulatory Network Model Benchmarking

Brown University, Center for Computation and Visualization, Research Assistant for Dr. Joselynn Wallace Oct. 2022–Present

- Developed a robust Shiny app in RStudio for RNA-seq data analysis, leveraging Limma-Voom and empirical Bayes for differential gene expression. Applied batch correction and voom transformation for linear modeling and visualized results with an interactive volcano plot and heatmap. Packaged the app into a Docker image for easy replication and distribution
- Benchmarked 12 gene regulatory inference models from the BEELINE repository using Docker on single-cell transcriptomic data. Determined SCRIBE as the top-performing model with the highest AUROC of 65% and AUPRC of 33%
- Participated in weekly meetings, presenting key findings, tracking project milestones, and collaborating with the team to propose next steps

### Contrastive Multimodal Learning & Advanced Single-Cell Segmentation

Brown University, Singh Lab, Research Assistant for Dr. Ritambhara Singh Jul. 2022–Present

- Designed a contrastive learning-based multimodal model in Python to integrate pathology images, annotated texts, and clinical information from the Kather colon dataset into a unified embedding space, enabling zero-shot classification and cross-modal retrieval
- Formulated an innovative contrastive loss that accounts for all modalities at once instead of averaging pairwise computations, improving classification and retrieval accuracy. Engineered a specialized pipeline to download and extract video, audio, and caption features across a million YouTube samples, tailored for evaluating the novel loss function
- Developed a convolutional neural network (CNN) in Python using Gumbel-SoftMax activation for enhanced single-cell segmentation. When benchmarked against state-of-the-art models, this CNN, paired with soft Gumbel and binary cross-entropy loss, demonstrated superior performance
- Engaged in weekly meetings, providing updates on current progress, presenting recent advancements in machine learning papers quarterly, and delegating tasks to an undergraduate researcher

### Time Series Analysis for Listeriosis Outbreak Prediction

Brown University, School of Public Health, Research Assistant for Dr. Alice Paul Jun. 2022–Aug. 2022

- Generated weekly, monthly, and SNP cluster-specific time series for listeriosis infection from 2013 to 2022, utilizing datasets sourced from the National Center for Biotechnology Information
- Constructed time series models in RStudio to predict potential listeriosis outbreaks; through nested cross-validation and exhaustive grid search, identified the optimal model as a generalized seasonal ARIMA with an autoregressive order of one, a moving average order of two, using monthly seasonal pattern, and monthly count data
- Conducted a thorough literature review on machine learning methods in population health, focusing on predicting epidemic outbreaks

## WORK EXPERIENCE

### Teaching Assistant for Graduate-Level Machine Learning Theory Course

Brown University, Data Science Institute, Teaching Assistant for Dr. Andras Zsom Jan. 2023–Present

- Fulfilled responsibilities as a TA for a graduate-level course, "Machine Learning: From Theory to Algorithms," for two semesters
- Provided comprehensive support to a class of 70 students through mentorship on final projects, managing an online question forum, hosting weekly office hours, facilitating interactive workshops, and assisting in the preparation and grading of assignments and exams

### Data Processing Pipeline Development for Medicare Advantage Advertising Analysis

Brown University, School of Public Health, Data Scientist for Dr. Andrew Ryan Jun. 2023–Aug. 2023

- Managed and processed a Nielsen Ad Intel relational database with over 300 million entries using RStudio to study US Medicare Advantage (MA) advertising trends
- Designed and implemented a sophisticated filtering strategy to pinpoint MA TV advertisements; subsequently analyzed, visualized, and presented insights derived from the data

**Automating Data Preprocessing & Visualizations for Oceanographic Analysis**

Brown University, Department of Environmental Sciences, Data Scientist for Dr. Baylor Fox-Kemper Jun. 2022–Aug. 2022

- Devised automated preprocessing procedures in Python to transform raw data from buoys into structured, analyzable datasets
- Compiled, visualized, and transformed raw oceanographic data spanning 2017 to 2022 into the standardized NetCDF format for more efficient analysis and storage

## PROJECTS

**Customer Churn Analysis** Sept. 2023–Present

- Developed six classification models in Python, including Logistic Regression (L1, L2, ElasticNet regularization), Random Forest, SVM, and XGBoost, to predict customer churn using demographic and transaction data for a graduate machine learning course
- Achieved a F2-beta score of 0.9, nearly doubling the performance of the baseline model. Employed permutation feature importance and SHAP values for comprehensive model interpretation
- Delivered a report and presentation to communicate complex data insights to non-technical audiences effectively and recommended proper strategic actions

**Blockchain System Architecture** Feb. 2023–May 2023

- Engineered a robust blockchain infrastructure in Go for a graduate blockchain and cryptocurrencies course, encompassing core components like 'Blockchain' for transaction recordings and 'Miner' for transaction validation
- Optimized network efficiency and consistency through 'Node' synchronization, 'Server' data exchange, and off-chain transactions via 'Lightning Node' and 'Channel'
- Strengthened system security with the 'Wallet' for cryptographic key storage and the 'Watchtower' for vigilant monitoring against malicious activities

**Diamond Clarity Classifier** Sept. 2022–Dec. 2022

- Designed Convolutional Neural Networks in Python for a graduate computer vision course to classify diamond images into six clarity labels, achieving 60% accuracy with the optimal model
- Applied image processing techniques to remove the background from diamond images, using the Harris Corner Detector for intersection identification, enhancing classification accuracy

**Russia-Ukraine War Sentiment Analysis** Feb. 2022–May 2022

- Fine-tuned Recurrent Neural Networks (Bidirectional LSTM, LSTM, GRU) in Python to categorize tweets into 11 sentiment types with 83% accuracy for the optimal model and leveraged Latent Dirichlet Allocation for underlying topic discovery across the dataset for a graduate deep learning course
- Preprocessed tweets on the Russia-Ukraine War from February 24 to May 1, 2022, through tokenization, lemmatization, and the removal of semantically vacuous and rare words

**French to English Translator** Feb. 2022–May 2022

- Developed Transformer-based sequence-to-sequence Neural Networks in Python to translate French to English; attained per-symbol accuracies of 73%, with a perplexity score of 5.52
- Preprocessed the French and English Hansard, converting sentences to ID forms via word-to-ID dictionaries for optimized processing

**Fatal Police Shootings Analysis** Sept. 2021–Dec. 2021

- Implemented mixed-effect and forward stepwise logistic regression models in RStudio to probe racial disparities in fatal police shootings during a graduate statistical learning course, achieving ~80% accuracy on the test set
- Analyzed, identified, and presented significant predictors contributing to the performance of the fatal police shooting models

**Salary for College Graduate Analysis** Mar. 2020–Jun. 2020

- Trained a linear regression model using RStudio to predict the median salary for a student ten years post-enrollment with an adjusted R-squared of 73% for a linear regression class
- Assessed multicollinearity using Variance Inflation Factor (VIF) and pinpointed influential points with Cook's distance; validated model assumptions through diagnostic plots, predictor interactions, and non-linear transformations

## SKILLS

- Proficient with TensorFlow, Keras, and PyTorch in Python for the development and application of deep learning models
- Skilled in utilizing SKLearn, NumPy, Pandas, and Matplotlib in Python to build comprehensive data science pipelines encompassing exploratory data analysis, data processing, model development, tuning, and interpretation
- Proficient in RStudio for statistical learning, including generalized linear, longitudinal, multilevel, and time series models. Experienced with Bioconductor packages such as sva, edgeR, and limma, and adept at creating interactive visualizations with Shiny
- Experienced in utilizing Git for version control, employing SQL for database management, using SAS for statistical programming, and creating and deploying Docker images