Chuqi Wang

Irvine, CA 92617 | (949) 849-3189 | chuqiw4@uci.edu
GitHub | LinkedIn | Website

RESEARCH INTERESTS

Statistical methods and applications, machine learning and data mining, causal inference, biostatistics, with a focus on health-related problems, including mental health, environmental issues, food safety, public health, and other related areas. I am particularly motivated to explore statistical methods in interdisciplinary research, such as personalized medicine, and to develop AI systems for health monitoring and prevention through multimodal models.

EDUCATION

University of California, Irvine ***Irvine, CA, USA***
Master of Data Science *Sept. 2023 – Expected Dec. 2024*

- GPA: 3.97/4.0
- Relevant Courses: Databases & Data Management, Big Data Management, Artificial Intelligence, Probability & Statistical Theory, Statistical Methods, Bayesian Data Analysis, Machine Learning & Data Mining

McGill University ***Montreal, QC, Canada***
Bachelor of Science in Statistics, minor in Computer Science *Sept. 2018 – May 2022*

- GPA: 3.56/4.0, Major GPA: 3.76/4.0
- Relevant Courses: Algorithm & Data Structures, Advanced Calculus, Algebra & Analysis, Statistical Learning, Mathematical Statistics, Generalized Linear Models

PUBLICATION

- Wang, C. (2023, January). A REVIEW on 3D convolutional neural network. In 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA) (pp. 1204-1208). IEEE. [Link]
- Zhang, L., Cai, W., Liu, Z., Yang, Z., Dai, W., Liao, Y., ... & Chen, Y. (2023). Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. arXiv preprint arXiv:2308.09975. (under review via ACL Rolling Review, submitted October 2024) [Link]

EXPERIENCE

Survey Research on AI Applications in Sports *Oct. 2024 – Present*
Graduate Student Researcher (Advisor: Prof. Weining Shen)

- Reviewed and summarized critical findings of research papers in AI Applications in Sports, focusing on the application of computer vision, NLP and multimodal LLMs in sports to support a comprehensive survey. Presenting findings in team discussion to facilitate project alignment and knowledge sharing

Olivares Lab, UCI Civil & Environmental Engineering *Jun. 2024 – Present*
Research Team Member (Advisor: Prof. Christopher Olivares)

- Collected, cleaned and processed PFAS contamination data for drinking water from 20 public sources across 50 U.S. states. Utilized Pandas and Camelot to extract and clean raw data from various formats (PDFs, GIS data, lab reports, etc.), resulting in a structured dataset of 2.5 million samples for accurate analysis of PFAS in drinking water.
- Conducted in-depth exploratory data analysis (EDA) and spatial analysis using Matplotlib, Plotly, and Seaborn visualize PFAS contaminant concentrations across U.S. states, including interactive maps and summary statistics tables, to investigate PFAS concentration distributions, trends over time, and regional comparisons across states.
- Review at least five research papers on PFAS in drinking water each week and present findings at weekly group meetings, currently applying machine learning models to predict per-and polyfluoroalkyl substances (PFAS) concentration across different states and co-authoring a research paper on findings.

Financial Large Language Model Research *Jun. 2024 – Oct. 2024*
Graduate Student Researcher (Advisor: Prof. Weining Shen)

- Conducted weekly literature reviews of 5 papers on financial multimodal LLMs and collected financial datasets to support financial LLM survey research. Presented one selected paper in team meetings to highlight key findings and advancements.
- Evaluated large language models on financial tasks using the FinEval benchmark via API calls, testing four scenarios: zero-shot, zero-shot CoT, five-shot, and five-shot CoT. Conducted comparative analysis to identify 19 different models' strengths and weaknesses, guiding model selection for financial applications.

- Co-authored a research paper on the FinEval benchmark with Xin Guo et al., contributing data presentation through extensive tables and figures formatted in LaTeX. Enhanced clarity and interpretability of benchmark questions and model results, supporting robust comparative analysis within the study.

Pelvic Floor Disorders Research Lab, UCI Health *Jun. 2024 – Aug. 2024*

Individual Study Researcher (Advisor: Dr. Olivia Chang)

- Utilized Pandas to clean, filter, and merge datasets containing over 4 million patient records from the 2019-2022 NSQIP database and conducted comparative statistical analysis of patients who underwent "Vaginoplasty with peritoneal pull-through" versus "Vaginoplasty alone".
- Developed interactive dashboards using Tableau to present data visualizations for researchers. Performed logistic regression analysis with stepwise selection to predict composite outcomes of patients who underwent transgender surgeries, achieving 94.89% accuracy and an AUC of 0.865, with a cross-validation error of 0.056.

PROJECTS

Climate AI-Capstone Project [GitHub] *(Industry Partner: CCEX Company)* *Sept. 2024 – Present*

- Developing Python scripts to scrape and preprocess approximately 2,000 verified carbon standard reports, requirements and templates from the Verra registry, including content extraction, including segmenting PDF reports and creating embeddings to enable fast retrieval and efficient search capabilities.
- Leading the integration of LLMs via API within a Retrieval-Augmented Generation (RAG) framework to develop Climate AI-a conversational tool designed to provide users with project report recommendations and generate project description documents meeting Verified Carbon Standard (VCS) requirements.
- Utilized Docker containers to develop a multimodal chatbot with a knowledge base, retrieval engine, and follow-up question suggestions. Focused on generating expert guidance, drafting project descriptions for certification, and achieving customer satisfaction, cost efficiency, and risk mitigation.

Stroke Prediction Using Bayesian Logistics Regression [GitHub] *Feb. 2024 – Mar. 2024*

- Developed a Bayesian logistic regression model using the rstan package in R, fitted with 2000 iterations and 4 Markov chains via Markov Chain Monte Carlo (MCMC), to predict stroke occurrence based on 5110 patients' demographic, medical, and lifestyle data. Achieved a 95.2% test accuracy and improved model sensitivity through decision threshold adjustment.
- Performed data preprocessing and exploratory data analysis (EDA) using dplyr and ggplot, and applied diagnostic tools like Bulk ESS, Tail ESS, and trace plots to ensure model convergence and reliability.

Midwifery Services Database Application [GitHub] *Jan. 2022 – Apr. 2022*

- Designed an Entity-Relationship (ER) diagram and a relational database model for the Quebec Ministry of Health to efficiently manage midwifery services. Established and executed a database schema using DB2.
- Authored comprehensive SQL queries for data population, maintenance, and updates. Engineered a user-friendly database application tailored for midwives via Java Database Connectivity (JDBC).

SKILLS

Programming Languages:	Python, R, Java, SQL, C/C++, MATLAB
Frameworks & Tools:	Jupyter, SciKit-Learn, TensorFlow, PyTorch, Tableau, AWS, LaTeX
Databases:	MySQL, PostgreSQL, Apache Cassandra, MongoDB, Neo4J, Apache Spark