YUHENG YAO

YUHENG004@e.ntu.edu.sg | +65 81440450
150 Nanyang Cres, Crescent Hall of Residence,Singapore, 637122

EDUCATION

Nanyang Technological University (NTU), Singapore 08/2024-now

- *Master of Science in Data Science*

Beihang University (BUAA), Beijing, China 09/2020-06/2024

- *Bachelor of Science in Statistics* | **GPA**: 3.68/4.0 **WES GPA**: 3.71/4.0
- 2nd Prize, the 33rd Fengru Cup Competition (05/2023)
- Scholarship for Academic Excellence (05/2021)

PUBLICATIONS

Yuheng Yao. Exploratory Data Analysis & Data Mining on NBA Match Prediction. Accepted by the 2023 6th International Conference on Computing and Big Data (ICCBD 2023)

RESEARCH EXPERIENCES

LLM-based Query rewrite(on going) 2/2025 - now
Author | Instructed Professor Cong Gao at NTU Singapore

- Optimized baseline for LLM-based query rewrite baseline.
- Automated pipeline for data preparation, management, generative model and analysis of query rewrite.
- Induce novel AI techniques to construct the state of art model.

Machine Learning-based Intelligent Medical Diagnostic Assistance System 2/2024-05/2024
Author | Instructed by Associate Professor Jian Ma at BUAA Beijing, China

- Independently designed and established workflows including data processing, model training and hyperparameter selection, and decision strategies
- Conducted data analysis and data mining on the datasets corresponding to the five diseases and identified significant features from these datasets.
- Utilized machine learning models such as Decision Tree, Random Forest, XGBoost, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) to train, test, and predict disease outcomes
- Incorporated the Crow Search Algorithm as the hyper-parameter search algorithm, significantly improving the model's accuracy, with the highest improvement for KNN reaching 18%.
- Introduced an ensemble learning algorithm named stacked generalization, which takes the predictive results from seven predictive models as input to obtain the final prediction results，raising the system's prediction accuracy for liver disease from 79% to 92%.

Machine Learning-based Sports Data Analysis 12/2022-06/2023
Research Assistant | Instructed by Dr. Jiaxi Yang at Columbia University Remote

- Independently completed data cleaning, data mining, and analysis of 2003-2022 NBA games data (25,786 matches)
- Built a multivariate linear regression model to identify the factors affecting the prediction of NBA games outcomes
- Utilized machine learning models such as Naive Bayes, Random Forest, Logistic Regression (LR), KNN and LDA to train, test and predict outcomes based on previous matches
- Refined the models to obtain the optimal performance and found that the most critical factor was the recent win rate, with the LR model performing the best, achieving an average accuracy of 67% over all seasons

A Novel Meta Learning-based Model for Parametric Partial Differential Equations (PDEs) 08/2022-05/2023
Major Participant | Instructed by Associate Professor Xiao Zhang at BUAA Beijing, China

- Developed the overall design of the model from extracting data features via Style Transfer, to introducing a Neural ODE to solve the model and PDEs, and to applying a MAML algorithm to train the model
- Assessed the feasibility of the model and reproduced the DyAd model for improved generalization of the prediction
- Defined the experiment design and implementation including data sampling, analysis, and future improvement
- Proved the model to be effective in reducing relative errors in less time (0.74%±0.08%, 12s) compared with a Res-Net model (1.02%±0.07%, 0.8h)

ACADEMIC HIGHLIGHTS

A Network Science's View on Academic Collaboration at NTU CCDS*(Instructed by Professor Sourav S Bhowmick)* 04/2025

- Led data preprocessing using Python to clean and structure co-authorship data from DBLP, resolving issues like missing values, duplication, and outdated links.
- Constructed and visualized a dynamic, undirected, and weighted collaboration network, with nodes representing faculty and edges indicating co-authorships over time.
- Analyzed network properties (e.g., degree distribution, average degree growth, preferential attachment) to uncover structural patterns such as the emergence of hubs and densification trends.

Chicago Food Inspection Data Quality Profiling and Preparation*(Instructed by Professor Sourav S Bhowmick)* 04/2025

- Cleaned and analyzed a large-scale public dataset (~287K rows, 17 columns) on food inspections from the City of Chicago Open Data Portal.
- Applied association rule mining (Apriori) to uncover patterns in violations across facility types, inspection results, and risk levels.
- Discovered and verified (approximate) functional dependencies using TANE to enhance schema understanding and imputation logic.

Virality and Sentiment Analysis of Online Fitness Videos on Bilibili*(Instructed by Professor Sourav S Bhowmick)* 11/2024

- Collected and processed over 320,000+ user comments and metadata from Pamela Reif's fitness videos on Bilibili, using Open API integration and web scraping tools.
- Conducted exploratory data analysis (EDA) to identify trends in video popularity, user demographics (gender, user level, region), and content tags.
- Developed a feature selection and regression model to identify key predictors of high-performing videos (e.g., tags, video length, time of release).

Deep Learning Fundamentals and Practices*(Instructed by Lecturer Xiaoyu Chen)* 03-06/2024

- Hands-on implementation of MLP, CNN, RNN, and Transformer models and related experimental processes to solve age prediction, image recognition, fast information extraction and Math Word problems.

Regression Analysis *(Instructed by Professor Chao Liu)* 03-04/2023

- Explored the logarithmic relationship between composition ratios and strength of concrete through multivariate regression modeling, increasing the R^2 from 0.2 to 0.4

Optimization Theory and Algorithm *(Instructed by Professor Deren Han)* 12/2022-01/2023

- Performed theoretical reasoning and C programming of common optimization algorithms such as Steepest Descent, Newton's Method, Backtracking Line Search, DFP, and Gauss-Newton (with Linear Search)

INTERNSHIP

Chinese Institute of Electronics, Beijing, China (40hrs/week) 06-08/2023

- Contributed to the planning, organization, promotion and holding of the 18th "GigaDevice Cup" China Graduate Electronics Design Contest
- Completed data cleaning, filtering, and merging in R to provide data management support for the Contest
- Analyzed the Contest data to identify trends and patterns; visualized results for better presentation