Yujie Dai

RIGHT TO WORK IN THE UK | Yujie.Dai@bristol.ac.uk | +447419987968
linkedin.com/in/yujiedai/ | github.com/DaisyDDD | https://daisyddd.github.io/YujieDai.github.io/

PhD researcher in eXplainable AI (XAI) and Machine Learning (ML) with expertise in large-scale electronic health records (EHR), model interpretability, and applied data analysis. Experienced in developing interpretable ML models using Python-based frameworks (scikit-learn, PyTorch) and explainability techniques such as SHAP and LIME. Passionate about bridging research and application by using data-driven analysis to address complex real-world challenges.

EDUCATION

Ph.D. in Digital Health and Care (Population Health Data Science), University of Bristol, UK Sept 2022 – Present

- **EPSRC-funded** CDT student focusing on interpretable machine learning for healthcare applications.
- **Supervisors:** Prof Andrew Dowsey, Prof Raul Santos-Rodriguez, Dr Brian Sullivan
- **Thesis: ML and XAI in UTI Risk Stratification and Antibiotic Resistance Prediction on a Linked EHR Dataset.**
  - Processed and integrated a large linked EHR dataset (N = 962,237) from both primary and secondary care; implemented parallel data preprocessing using **ProcessPoolExecutor** to handle large-scale feature engineering and data cleaning tasks.
  - Performed hyperparameter tuning with **Hyperopt** package for model optimisations and robustness improvements.
  - Achieved **AUC 0.62–0.98** for UTI risk classification and **AUC 0.60–0.80** for antibiotic resistance prediction across 6 risk groups and 7 antibiotics with **XGBoost** model.
  - Explored the transparency and robustness of XAI models using **SHAP, LIME, and bLIMEy** to identify key clinical and demographic predictors for both ordinal and categorical outcomes.
  - **Other Relevant Packages**: orf, PyTorch for LSTM, statsmodels, geopandas, pickle, pandas, NumPy, matplotlib, seaborn

MSc (Distinction) in Artificial Intelligence, University of St Andrews, UK Sept 2021 – Sept 2022

- **Dissertation** (17.5/20): **Investigating the relationship between network properties and disease spread using SIR models.**
  - **Supervisor**: Prof. Simon Dobson
  - Implemented computational simulations in Python to analyze disease transmission dynamics across modular and core-periphery network structures.
  - Constructed network topologies using the **NetworkX** library and simulated disease spread using the **epydemic** and **epyc** packages based on the SIR (Susceptible–Infected–Recovered) model framework.
  - Investigated how network size, connectivity, and modularity influence infection rate, speed, and epidemic thresholds.
- **Key modules**: AI Practice (17.3/20), Object-Oriented Modelling (17.6/20), Constraint Programming (17.1/20)

BSc in Software Engineering, Beijing Institute of Technology, China Sept 2016 – Jul 2020

- **GPA:** 84/100 including key modules: Artificial Intelligence (95/100), Design and Analysis of Algorithms (88/100), Data Mining (88/100), Software Architecture (93/100), Graduation Project (95/100).
- Awarded 3 university **scholarships** for academic excellence: 2016–2017, 2017–2018, and 2018–2019.

WORKING & RESEARCH EXPERIENCE

Research Data Scientist (part-time), The Jean Golding Institute, UK Sept 2024 – Sept 2025

- Developed and optimized machine learning models for biological classification using Python libraries such as **scikit-learn, XGBoost, and TensorFlow**, implementing algorithms including Logistic Regression, SVM, Random Forest, and Neural Networks.
- Performed feature engineering and dimensionality reduction using **PCA, UMAP, and t-SNE** to analyze and visualize patterns in high-dimensional datasets.
- Achieved the **best F1-scores of 0.98, 0.99, and 0.97** across three target classes using an XGBoost model.
- Collaborated with multidisciplinary teams to communicate results and support research-driven insights.

Data Scientist Collaborator, Turing Data Study Groups (DSG), The Alan Turing Institute, UK Jan 2025 – Feb 2025

- Participated in the AI for Decarbonisation (ADViCE) challenge on heat pump efficiency using the Electrification of Heat dataset (740 UK installations, 2020–2023). *(PROJECT DETAILS)*
- Conducted data preprocessing and time-series data quality assessment using the **tsfresh** package.
- Extracted operational features (e.g., peak patterns and seasonal trends) using Empirical Mode Decomposition and z-score-based peak detection, implemented with the **emd** package.

- Achieved **silhouette scores of 0.477 (k=2) and 0.421 (k=5)** with the Z-Score peak finding + SDTW K-means clustering to identify high- and low-performing heat pumps.
- Co-authored the final project report: **https://doi.org/10.5281/zenodo.15877726**.

Developer Intern, Graph Data and Blockchain Lab, Beijing Institute of Technology, China Aug 2020 – Aug 2021

- Designed and implemented website front-end (JavaScript, HTML, CSS) for a static testing platform.
- Collaborated with backend developers to visualize graph-based data analytics results.

Project Management Intern, Bentley Systems (Beijing) Co., Ltd, China Aug 2019 – Jul 2020

- Facilitated partner program analytics by tracking service usage metrics and performance reports.
- Automated data reporting and quality checks, improving visibility for management and compliance teams.

Programming Tutor (part-time), Beijing Quchuangyi Technology Development Co., Ltd, China Sep 2020 – Jun 2021

- Taught programming to students aged 6–16, covering Scratch, Python, and C++.
- Delivered interactive lessons to build computational thinking and problem-solving skills, through coding exercises and algorithmic challenges in an age-appropriate way.

PUBLICATIONS & PRESENTATIONS

Publications

Dai, Y. et al. (2024). *Explainable AI for Classifying UTI Risk Groups Using a Real-World Linked EHR and Pathology Lab Dataset.* https://arxiv.org/abs/2411.17645. The 2025 AAAI Health Intelligence Workshop. *In proceedings of the Studies in Computational Intelligence (Springer/Nature).*

Zhang, L; Xong, S; **Dai, Y**. (2023). *A Deep Learning Based Intraoperative Bleeding Point Detection System*. Patent No. CN202310660999.1. Public Announcement Number: CN116385977A. Announcement Date: August 15, 2023.

Selected Presentations & Conferences

Presenter, AAAI Conference on AI Health Intelligence Workshop, Philadelphia, USA 5 Mar 2025

- On paper *'Explainable AI for Classifying UTI Risk Groups Using a Real-World Linked EHR and Pathology Lab Dataset'.*

Poster, Combatting CDI Conference 2024, Cardiff, UK 27 Feb 2024

- '*Characterising CDI in the Southwest of England with the BNSSG Systemwide Dataset*' (*VIEW POSTER*).

Poster, UK Health Security Agency 2023, Leeds, UK 15 Nov 2023

- '*UTI & CDI Detection and Analysis in the Local BNSSG Area*' (*VIEW POSTER*).

TEACHING EXPERIENCE

Instructor, "Decision Trees and Random Forests with scikit-learn" Workshop, The Jean Golding Institute 6 May 2025

- Delivered a training on supervised classification using Python scikit-learn.
- Covered key concepts, practical implementations, and model evaluations for academic and professional audiences.
- Video recording of the session: Classification with scikit-learn: Decision Trees and Random Forests.

Lab Demonstrator, Engineering Mathematics & Technology, University of Bristol Sept 2023 – March 2025

- Supported postgraduate teaching in *Programming and Analytics for Digital Health, Advanced Financial Technology and Statistical Computing and Empirical Methods.*
- Guided students through R exercises for data processing and model construction such as Linear Regression and Bayesian Logistic Regression.
- **Relevant Packages:** tidyverse, dplyr, caret, glmnet, brms, ggplot2

PROFESSIONAL ACTIVITIES & LEADERSHIP

- **Member**, The Clinical AI Interest Group, The Alan Turing Institute
- **Team Member (Gold Medalist)**, International Genetically Engineered Machine (iGEM) Competition 2019
- **Organizer**, The 52nd Society for Academic Primary Care Annual Scientific Meeting, Bristol, UK
- **Executive President**, The Student Union of School of Computer Science, Beijing Institute of Technology

OTHER SKILLS

- **Languages:** English (Fluent, IELTS 7.5), Mandarin (Native)