Ikgyu (Tom) Shin

New Haven, CT | (703) 678-7882 | tom.shin@yale.edu | linkedin.com/in/ikgyushin | github.com/ikgyu98

EDUCATION

Yale University, New Haven, CT 05/2025
M.S., Health Informatics Advised by Dr. Ted Melnick
Capstone | Zeteo Health Assistant: Retrieval-Augmented Generation (RAG)-based Hyper-Personalized Chatbot
| Automated Neurological Scoring System Using Large Language Models (LLMs) for Parkinson's Disease

University of Illinois Urbana-Champaign (UIUC), Champaign, IL 05/2023
B.S., Statistics | Computational Science and Engineering (CSE) Certificate in Statistics

RESEARCH EXPERIENCE

Graduate Researcher, Yale University 04/2024 – present
Y-IAIML Lab Drs. Andrew Taylor and Christopher L. Moore
- Technical lead for CT SB-181 compliant ED boarding data reporting initiative funded by an ACEP/CCEP grant
- Allocating $12,500 in grant funding to support the full-stack development of a public website on ED boarding
- Designing a project plan and prototype interface in Figma, focusing on database and dashboard front-end components

Undergraduate Researcher, UIUC 01/2022 – 05/2023
PaDAS Lab Dr. Pamela P. Martinez
- Spearheaded lab data management, including aggregation and validation through visualization, to ensure data integrity
- Co-authored a study on SARS-CoV-2 vaccination trends and disparities, managing data from 161 countries/territories

TEACHING EXPERIENCE

Course Manager, S&DS 365 Intermediate Machine Learning, Yale University 08/2024 – 12/2024
Instructed by Professor John Lafferty
- Managed course operations for 234 students, addressing concerns and maintaining an effective learning environment
- Oversaw grading guidelines, distributions, and aggregation for 9 TAs and handled outstanding submissions

Teaching Fellow, S&DS 100 Introductory Statistics, Yale University 01/2024 – 05/2024
Instructed by Dr. Ethan Meyers
- Developed and maintained curriculum for 143 students, including comprehensive coursework assessment
- Fostered student understanding and engagement through weekly office hours with an average attendance of 12

CAPSTONE PROJECT

Zeteo Health Assistant: RAG-based Hyper-Personalized Chatbot 10/2024 – Present
Individual Project with Zeteo Health, mentored by Eric Laundry
- Developing a RAG-based chatbot for hyper-personalized healthcare interactions, tailoring responses to user context
- Conducted exploratory data analysis on validation datasets and feedback to identify gaps and potential improvements
- Experimented with readability metrics to adapt chatbot responses dynamically to assess the feasibility of the approach
- Identified meaningful prospects, laying the groundwork for further development of advanced personalization features

Automated Neurological Scoring System Using LLMs for Parkinson's Disease 09/2024 – Present
Group Project with Drs. Hua Xu, Sule Tinaz, and Vipina K. Keloth
- Developing an LLM-based tool to analyze patient-clinician transcripts, detecting cognitive markers for progression
- Utilized a de-identification pipeline using transformer models (ClinicalBERT, RoBERTa) for HIPAA-compliance
- Conducting manual annotations and validating inter-annotator agreements via LANN to create a 'gold standard'
- Engineering prompts for LLM testing, optimizing cognitive marker extraction through zero- and few-shot learning
- Delivering findings in reports and planning manuscript preparation to standardize neurological assessments

RESEARCH PROJECT

Predicting Mortality in Heart Failure with Preserved Ejection Fraction (HFpEF) Using EHR 12/2023 – Submitted
Advised by Dr. Karthik Murugiah
- Led the development of predictive models for 30-day and 1-year mortality in HFpEF using the MIMIC-IV database
- Derived a final cohort of 3,235 hospitalizations through SQL aggregation across multiple relational tables
- Ensured data integrity via statistical validation, addressing missingness and imbalance with appropriate techniques
- Streamlined an end-to-end machine learning pipeline and exhaustive cross-validation to evaluate model performance
- Resulted in comparable-performing models: logistic regression (30-day AUC: 0.83) and HGBC (1-year AUC: 0.78)
- Presented the abstract at the AHA 2024 conference; manuscript submitted and under review at JACC Heart Failure

Racial and Ethnic Disparities in Opioid Use Disorder Treatment Among Medicaid Beneficiaries 11/2023 – 02/2024
Led by Dr. Hazar H. Khidir

- Contributed to a study on Medicaid claims examining racial/ethnic disparities in MOUD prescriptions after ED visits
- Ensured data quality by selecting states with low to medium missingness in race/ethnicity data using CMS DQ Atlas
- Cross-referenced Medicaid TAF data to validate and expand HCPCS, ICD, and NDC code lists for OUD treatment

Biomedical Text Summarization with Domain-Specific Transformer-based NLP Integration 02/2024 – 05/2024
CPSC 577 Natural Language Processing

- Conducted in-depth analysis and refinement of NLP models (LED, BART, T5) for summarizing biomedical literature
- Designed in and cross-domain experiments, optimizing models for specialized biomedical and mixed-domain datasets
- Improved summary accuracy, relevance, and readability, achieving up to 47.3 ROUGE-1 and 86.5 BERTScore

Predicting Systemic Lupus Erythematosus (SLE)-Associated Epitopes using Deep Learning 03/2024 – 05/2024
CPSC 452 Deep Learning Theory and Applications

- Leveraged state-of-the-art DL models (ProtBERT, ReLSO, BEETLE, MITNet) to predict SLE-associated epitopes
- Integrated contextual biological information into the learning process for model interpretability and prediction
- Boosted ProtBERT accuracy from 76.5% to 91.8% and AUC from 0.5 to 0.93 with the AdamW optimizer

Health Supplement Analysis: National Health and Nutrition Examination Survey (NHANES) 01/2023 – 11/2023
BIS 634 Computational Methods for Informatics

- Analyzed NHANES data to assess the impact of health supplements on liver health by identifying key risk factors
- Processed 5569 entries with 11 continuous variables, introducing ordinal and one-hot encoded nourishment features
- Attained comparable classification results, expecting improvements upon solving class imbalance on NHANES data

LEADERSHIP EXPERIENCE

Yale Graduate Student Assembly, Public Health Representative, Yale University 05/2024 – present

- Establishing effective faculty-student communication channels, strengthening department-wide engagement
- Leading student collaboration initiatives through organized meetings and events, managing departmental budget

Data Science & Artificial Intelligence Society, President/Co-founder, UIUC 06/2022 – 05/2024

- Assembled a group of 30 scholars from diverse fields united by an interest in data science and artificial intelligence
- Supervised each project group, including introductory courses and data competitions, alongside weekly seminar

PUBLICATION

Shin, I., Bhatt, N., Alashi, A., Kandala, K., & Murugiah, K. (2024). Predicting 30-day and 1-year mortality in heart failure with preserved ejection fraction (HFpEF). medRxiv. *Manuscript currently under review at JACC: Heart Failure*. https://doi.org/10.1101/2024.10.15.24315524

Larsen, S. L., Shin, I., Joseph, J., West, H., Anorga, R., Mena, G. E., Mahmud, A. S., & Martinez, P. P. (2023). Quantifying the impact of SARS-CoV-2 temporal vaccination trends and disparities on disease control. Science Advances, 9(31), eadh9920. https://doi.org/10.1126/sciadv.adh9920

ORAL PRESENTATION

Khidir, H., Nedelec, L., Sebok-Syer, S., Sabbatini, A., Shin, I., Melnick, E., & Lin, M. (2025). Racial and Ethnic Disparities in Opioid Use Disorder Treatment Among Medicaid Beneficiaries. *Abstract submitted to the 2025 SAEM Annual Meeting, May 13–16, Philadelphia Marriott Downtown, Philadelphia, PA*.

Shin, I., Bhatt, N., Alashi, A., Kandala, K., & Murugiah, K. (2024). Predicting 30-Day and 1-Year Mortality in HFpEF Using EHR Data. *Presented at AHA2024 Scientific Sessions, November 16, McCormick Place, Chicago, IL*. Circulation, 150(Suppl_1), Abstract 4140208. https://doi.org/10.1161/circ.150.suppl_1.414020

TECHNICAL SKILLS

Programming Languages: Python (PyTorch, TensorFlow, Scikit-Learn, Gradio, LangChain), R, SQL, Excel, SAS
Methods: Data Preprocessing & Transformation, Statistical Analyses & Modeling, Database, Dashboard, Prompt Engineering
Tools and Platforms: Hugging Face, EHR (EPIC, ICD-9/10 code, MIMIC-IV), GitHub, Microsoft Azure, LANN, Figma