Kexin (Summer) Shang

4044097577 | ks4254@drexel.edu | linkedin.com/in/kexin-shang5301

EDUCATION

Drexel University, PA Sep 2023 - Present
Doctor of Philosophy in Information Science **GPA: 4.00/4.00**
(Focus on LLM in Healthcare)

Washington University in St. Louis, MO Sep 2021 - Dec 2022
Master of Science in Biostatistics and Data Science **GPA: 3.94/4.00**

Georgia State University, GA Jun 2019 - May 2023
Bachelor of Science in Mathematics (Statistics) **GPA: 3.85/4.30**
Bachelor of Science in Biology (Double Major, Co-diploma)

Southwest Jiaotong University, Chengdu, China Sep 2017 - Jun 2019
Bachelor of Engineering in Bioengineering (Co-diploma) **GPA: 3.63/4.00**

Main courses: *Recommender System, Nature Language Processing (Pytorch), Network Analysis, Applied Machine Learning in Data Science (Sklearn), Biostatistics (SAS), Analysis, Optimization, Survival Analysis, Bioinformatics (Linux), etc*.

RESEARCH EXPERIENCE

Healthcare Informatics Research Lab, CCI, Drexel Research Assistant Sep 2023 - Present
Topic: Multi-LLM Collaboration on Medical QA
1st author manuscript ready for International Conference on Healthcare Informatics (ICHI) 2025 submission

- Developed an agent-based LLM collaboration framework that improves both consistency and accuracy across all three LLM participants on medical QA task
- Measured the "confidence" of a LLM and investigate how it affects LLM's teamworking strategy
- Serve local LLMs through Text Generation Inference (TGI) method to significantly reduce inference time cost
- Elicit LLMs' generation via a combined prompting of Zero-shot Chain-of-Thought and Self-Consistency

Center for Healthy Weight and Wellness, Psychiatry, WUSTL Intern May - Dec 2022
Topic: Harnessing Mobile Technology to Reduce Mental Health Disorders in College Population
Poster presented on International Conference on Eating Disorders (ICED) 2023

- Constructed composite variables from over 200 features via PCA regression, which determines nearly 40% of the variation of response rate to follow-up surveys
- Designed a factorial design on 4 treatment components and identified moderator variables with each component using logistic regression models and simple slope analysis
- Conducted a cross-sectional survey the prevalence of 11 types of clinical and subclinical eating disorders in rural areas, suburban areas, and urban areas in U.S. applying pairwise T-tests with Holm's corrections

Department of Developmental Biology, WUSTL Research Assistant Sep 2021- Jun 2022
Topic: Role of Transposable Element in Transcript-level Expression Regulation

- Developed a Shell-based pipeline to obtain TE-derived transcripts' expression contribution from GTEx database
- Located age-sensitive TE-derived transcripts in skin tissue by plotting time-series Z-scores across age intervals
- Removed unwanted variation using residuals (RUVr with k=4) from RNAseq data and plot a 3D PCA which successfully showed clear separations between sun-exposed skin genes and sun-unexposed skin genes

IN-CLASS PROJECTS

DSCI 691 Nature Language Processing Drexel 2024 Summer

Topic: Text-based Emotion Recognition across BERT Family and LSTM

Source of data: a dataset of English Twitter messages with six basic emotions: anger, fear, joy, love, sadness, and surprise. https://huggingface.co/datasets/dair-ai/emotion

- Benchmarked four BERT family models (BERT, RoBERTa, DistilBERT, and XLNet) on 'emotion' dataset with bidirectional LSTM, the previous generation state-of-the-art model as the baseline
- Analyzed the potential reasons of the significant difference between our result and a similar research

GitHub page: https://github.com/summer5301/Text-based-Emotion-Recognition-across-BERT-Family-and-LSTM

DSCI 511 Data Acquisition and Preprocessing Drexel 2023 Fall

Topic: Analysis of the Canadian Wildfires Effect Posed on the Air Quality in US Cities.
Source of data: Scraped Wikipedia for Top 20 most populous US Cities and the "Open-Meteo" API for weather data and air quality data.
Individual contribution:

- Used Pandas Python package to restructured time-series weather data of each city scarped from Wikipedia and API into 1-year span by date and month.
- Represented continuous variables such as pm 2.5 index by mean and categorical variables such as "air quality level" by major vote and store cleaned data in Json files.

GitHub page: https://github.com/summer5301/4_smokwatchers_project/tree/main

MSB 660 01 Biomedical Data Mining WUSTL 2022 Spring

- Leveraged the Medical Expenditure Panel Survey (MEPS) database to predict medical cost across 3376 patients by fitting models of multiple linear regression, bagged random forest regression, and logistic regression w/t lasso penalty
- Adopted LDA and Naïve Bayes classifier to classify patients with high medical cost, achieving 96.9% and 94.1% specificity respectively
- Used Inverse normal transformation (INT) to normalize highly skewed data (change skewness from 4.9 to 0.074)

CONFERENCE

2022 ASA Women in Statistics and Data Science Conference, St. Louis, MO — Audience
International Conference on Eating Disorders (ICED) 2023, Washington, DC — Poster Presenter

HONOURS & AWARDS

Valedictorian of the Recognition Ceremony, WUSTL — 2022
Merit Scholarship ($11,886), WUSTL — 2021
Wiley M. Suttles Math Award ($750), GSU — 2023
In-state Scholarship; Presidential List; Member of the Honors College, GSU — 2020 - 2021
Second-class Scholarship (¥3000); National Scholarship Nominated, SWJTU — 2018 - 2019

EXTRACURRICULAR ACTIVITY

Publicity Department of the Chinese Student Union, GSU — Minister

- 2020 Atlanta Chinese Students and Scholars Spring Festival Gala
- Social media account management and operation

SKILLS

Analytics: Machine Learning Models, Deep Learning Architectures (implementation on CNN, RNN, transformers) on various data types (text, image, and tabular), OpenAI API, LangChain, vLLM and Text Generation Inference (TGI) inference
Programming: Python (Pandas, Numpy, Tensorflow, Pytorch), Shell, R, SAS, MySQL, Latex
Language: English (fluent); Chinese (native)