Kunxiao Gao

kgao9@jh.edu | GitHub: https://github.com/kunxiaogao

EDUCATION

University of Michigan - Ann Arbor, MI **09/2025 – Present**
Ph.D. in Computer Science and Engineering, Overall GPA: 4.0/4.0

Johns Hopkins University - Baltimore, MD **08/2023 – 05/2025**
*Master of Science in Engineering in Data Science, Overall GPA: **3.87**/4.0*
Core course: Artificial Intelligence, Deep Learning, Machine Learning for Medical Applications, Machine Translation, Reinforcement Learning, Applied Stats and Data Analysis, Algorithm

University of California, Santa Barbara - Santa Barbara, CA **09/2019 - 04/2023**
*B.Sc. in Statistics and Data Science, Overall GPA: **3.94**/4.0*
Core courses: Regression Analysis, Adv Stats Models, Machine Learning, Bayes Data Analysis, Big Data Analytic
Honors: Dean's Honor List, University of California, Santa Barbara (Spring/2020 - Spring/2023)

PUBLICATIONS

Gao, K., Favaro, A., Dehak, N., & Moro-Velázquez, L. (2024). "Identifying Early Markers of Alzheimer's Disease through Longitudinal Analysis of Language and Speech Patterns," npj Dementia-Nature (to be submitted at the end of Dec 2024).

SELECTED ACADEMIC & PROFESSIONAL RESEARCH EXPERIENCE

Identifying Early Markers of Alzheimer's Disease through Longitudinal Analysis of Language and Speech Patterns **03/2024 - Present**
Supervisor: Associate Prof. Laureano Moro-Velazquez and Assistant Prof. Najim Dehak, Center for Speech and Language Processing, JHU

- Investigated early Alzheimer's Disease (AD) detection using longitudinal speech analysis, incorporating both acoustic and linguistic features for binary classification and identifying prodromal cognitive decline markers.
- Collected a balanced dataset with over 5,000 recordings, spanning up to 10 years pre-diagnosis, ensuring quality through noise reduction, diarization, and segment filtering.
- Extracted acoustic and linguistic features, utilizing advanced models (Llama3, Wav2Vec 2.0, and so on) to generate both interpretable and non-interpretable features for detecting early AD signs.
- Developed and tested classification models (SVM, XGBoost, MLPs and so on), performed longitudinal statistical analysis, identifying significant speech markers related to pauses, pitch variation, and syntactic complexity of early AD for clinical applications in detection and monitoring.
- Achieved peak classification accuracy of 0.80 in pre-diagnosis speech patterns using NIFMs, indicating their effectiveness in capturing subtle linguistic and acoustic signs of AD.

GNN Efficiency Improvements by Using Graph Sampling Strategies **02/2023 - Present**
Supervisor: Assistant Prof. Luana Ruiz, AMS department, JHU

- Developed a Graph Neural Network (GNN) model for correlation network prediction, optimizing computational efficiency while preserving predictive accuracy.
- Employed graph signal sampling techniques, testing strategies such as highest leveraging scores and adapted BFS, to construct a smaller, representative graph.Trained the GNN on the reduced graph, ensuring convergence through mathematical derivation and evaluating performance with a Graph Shift Operator (GSO) on the full dataset.
- Applied mathematical derivations using Lipschitz continuous graphons W to establish error bounds for non-uniform sampling strategies. Specifically, derived that sampling the top n nodes based on leverage scores results in better convergence compared to uniform random sampling with small exponential decay rate a.
- Achieved effective predictions, almost equal to the results from the model trained on original graph, with reduced computational costs, demonstrating that a well-sampled smaller graph can outperform traditional graph sampling methods, which achieved the lowest Mean Squared Error about 0.30, 26.8% lower than the traditional one.

Glodon - Beijing, China **06/2023 - 08/2023**

Data Scientist Intern

- Aimed to predict customer purchasing behavior and provide customer data with high purchasing power.
- Cleaned the data and developed a fuzzy matching algorithm by using jieba tokenizer and regexp to acquire the 200 most popular words which will be ruled out from the perfect matching stage to create the term match table.
- Combined the customer and base data frames on match table and wrote the aggregation back-tracking algorithm by retrieving the data in certain time intervals before different specific dates for each id and aggregated the data frame by project types, budgets, etc. for each time interval.
- Implemented XGBoost model with Randomized Search and Bayesian Optimization to train model and fine-tuned the hyper-parameters, like tree depth and maximum leaf nodes, acquiring the best model base on 65.893 AUC.
- Evaluated the model on test dataset, drew ROC and K-S curve to ensure no overfitting with approximately 66% AUC.

Amgen Knowledge Graphs Based on PubMedBERT model **01/2023 - 06/2023**

Supervisor: Dr. Maxim Ivanov, Amgen, University of California, Santa Barbara

- Applied NLP and text-mining algorithms to automate the creation and updating of Knowledge Graphs (KGs), for drug repurposing by extracting relationships among biomedical entities.
- Replicated the BERT model from the BioRED paper and benchmarked performance. Fine-tuned PubMedBERT on the BioRED dataset for Named Entity Recognition (NER) and achieved high precision (62%) and recall (75%) on unseen datasets, confirming its robustness across varied biomedical text sources.
- Completed data preprocessing tasks and trained fine-tuned models using Amgen's database, leveraging the pretrained PubMedBERT model for Named Entity Recognition (NER).

OTHER PROJECT EXPERIENCE

Machine Learning in Medical Applications **02/2024 - 05/2024**

Instructor: Associate Prof. Laureano Moro-Velazquez, JHU

- Created a classification model for harmful brain activity using EEG and spectrogram data from the Harvard Challenge, achieving an accuracy of over 86.04%.
- Converted EEG signals to spectrograms, organized electrode data into four panels, and applied Fourier Transforms to improve data visualization and interpretability.
- Implemented ResNet and EfficientNet with Multiple instance learning, dividing each EEG embedding into four horizontal windows, creating distinct sub-segments for each instance in the dataset, which allowed the model to treat each window as a separate "instance" within a larger EEG 'bag,' thereby enabling it to identify meaningful patterns across different portions of the signal, to address challenges in EEG signal variability, finding EfficientNet effective for multi-scale information.
- Compared and tuned MIL tune hyperparameter on dev set, identifying optimal kernel and stride settings to enhance model adaptability and accuracy, with applications for seizure and harmful brain activity detection, advancing automated medical diagnostics.

Machine Translation **09/2023 - 12/2023**

Instructor: Prof. Philipp Koehn, JHU

- Designed and implemented a translation model combining statistical and neural network techniques to improve accuracy and fluency in French-to-English translation.
- Applied the Metropolis-Hastings algorithm to optimize beam search decoding, achieving higher translation quality with reduced computational cost.
- Integrated phrase-based models, language models, and reordering strategies, benchmarking performance against standard metrics and baselines.
- Worked extensively with decoding processes in neural translation models, back-propagation, and calculating language model probabilities to improve overall translation effectiveness.

Data Science Capstone **09/2022 - 06/2023**

Supervised by Prof. Trevor Ruiz, University of California, Santa Barbara

- Gained expertise in data analysis methods such as generalized linear regression, KNN regression, Random Forest, NLP, and neural networks, while mastering professional data processing tools.
- Directed group-based learning on real-world problems, including identifying blood biomarkers for ASD detection, developing fraud risk prediction models, and applying clustering methods to classify countries by socio-economic and health factors.
- Conducted literature reviews, analyzed emerging research trends, and presented group findings, contributing to the academic discourse on data science applications in diverse fields.

OTHER WORKING EXPERIENCE

John Hopkins University - Baltimore, MD **08/2024 - Present**

Teaching Assistant for Applied Statistics and Data Analytics

- Conducted problem sessions to reinforce course concepts and address student questions.
- Graded assignments, ensuring timely and accurate feedback for students.
- Supervised all exams, maintaining academic integrity and facilitating the exam process.
- Updated and revised solutions for assignments and exams to reflect course material.

China Unicom - Beijing, China **06/2021 - 08/2021**

Software engineer Intern, Technical Department

- Worked in the technical department. Assisted the team in doing a Web crawler with Python to obtain, clean and analyze data of 30 banks (including longitude, latitude, address, name, telephone, and so on)
- Helped create and improve the user database of 9 provinces' and cities' data like populations, regional GDP, average disposable income per resident, and so on in China.
- Developed a program in python that can automatically make the Power Point of ICBC Urban Market Analysis Report of 4 cities by passing in different parameter values.

AWARDS AND HONORS

- Second Prize, CYPT (China Young Physicists' Tournament), Nankai University, China 08/2017

SKILLS

Programming Language: Python (Pytorch, sklearn, Tensorflow, spacy, requests etc.), R, SAS, Latex, SQL, Spark, Shell

Language: Chinese (native), English (GRE 325, V 158, Q 167, AW 4.0)