# Fanyi Kong

Kong.fa@northeastern.edu ♦ kfy2018@outlook.com

## EDUCATION

**Northeastern University, Boston, MA** Sep 2018 - May 2020
Masters of Science: Industrial Engineering (Healthcare Focus)
Core Courses: Data Mining(A), Deep Learning(A-), Computation and Visualization(A-)

**China University of Labor Relations, Beijing, China** Sep 2012 - June 2016
Bachelor of Engineering: Safety Engineering (Occupational Health Focus)
**★ Awarded Scholarship(5%)| Merit Student(3%)| Led Robot Training Group won the third award of "the Fifth Tsinghua University Undergraduate Engineering Training Competition"| Led team to complete project funded by Beijing University Students' Scientific Research and Entrepreneurial Project**

## TECHNICAL STRENGTHS

**Programming Languages**: Python, Shell Scripting, Batch Script, SQL, R, Java, SAS, Julia, Lingo
**Operating System& Tools**: Linux, Windows, PyCharm, Git, Docker, Tableau, Jupyter
**Framework**: PyTorch, TensorFlow, ESPnet, Mozilla DeepSpeech, Wav2Letter++, Kaldi, Numpy, Pandas
**Cloud Service**: Amazon Web Services(AWS), Alibaba Cloud, Tencent Cloud
**Technical Skills**: AI, Neural Network, Deep Learning, Machine Learning, Algorithm Design, Data Mining

## JOURNAL PUBLICATIONS

Active Learning: Encoder-Decoder-Outlayer & Vector Space Diversification Sampling **Published 2023**
**Co-first Author**
**Mathematics, MDPI4353,Mathematics 2023, 11(13), 2819; https://doi.org/10.3390/math11132819**

Predicting Drug-Drug Interactions using Heterogeneous Graph Neural Networks: HGNN-DDI **Accepted**
**Co-first Author**
**The 4th International Conference on Signal Processing and Machine Learning**

## RESEARCH & PROJECT EXPERIENCE

**Predicting Drug-Drug Interactions using Heterogeneous Graph Neural Networks: HGNN-DDI** *July 2023 - August 2023*

***Collaborative project with Shanghai Jiao Tong University*** ***Shanghai, China***
**Developed a deep learning approach to predicting drug-drug interactions (DDIs) from the chemical structure of drugs, as represented by the Simplified Molecular Input Line Entry System (SMILES).**

- Crawled datasets from the website and prepared datasets through Python.
- Applied ChemBertA for data preprocessing and Heterogeneous Graph for feature extraction; created a robust Graph Neural Network model for high-accuracy DDI prediction.
- Enhanced model performance through rigorous training, applying cross-validation and hyperparameter tuning, including dropout and batch normalization, to ensure accuracy and robustness. Integrated t-SNE visualization to further analyze and interpret the model's feature space, providing insightful perspectives on data clustering and model behavior. The accuracy of the best is 96.86%.

**Active Learning: Encoder-Decoder-Outlayer & Vector Space Diversification Sampling** *January 2023 - March 2023*

***Got A- level grade in the course "Algorithms for Big Data Program" Supervised by Pro. David Woodruff, CMU.*** ***Online***
**Developed a machine learning training pipeline integrating Encoder-Decoder-Outlayer framework with Vector Space Diversification Sampling, aimed at enhancing data diversity and reducing human labeling efforts.**

- Compared VSD to traditional clustering techniques like DBSCAN and dimension reduction methods like t-SNE and PCA. Incorporated active learning to iteratively select and label data points.
- Utilizing a pre-trained "all-mpnet-base-v2" Sentence-BERT model as the encoder and training a 3-layer resnet framework with Nadam optimizer, this approach optimizes GPU memory and data buffering for efficiency. The methodology demonstrates significant performance improvements in neural network models with less data, reducing computational and labor costs.

**Data Analysis of Career Village Website Users by SQL and Tableau** *October 2019 - December 2019*

***Focused on website users' operating habits to provide advice to the website to make progress.*** ***Boston, MA***

- Redesigned and organized Excel tables of columns about users' feedback, students, and professors from the original dataset. Drew an Entity Relationship Diagram to show the relations between new data entities directly.
- Built a database including three tables with SQL, then visualized data with Tableau. Analyzed data (such as students in which major used this website the most) and effectively presented findings via Tableau and PowerPoint.

**Predicting In-Hospital Mortality of ICU Patients by Machine Learning** *February 2019 - April 2019*
***Team Leader*** ***Boston, MA***
***Focused on prediction of the effect of ICU on patients and helped people make better decisions.***

- Searched related materials and cleaned 299,264*42 datasets using Python. Used PFE to reduce the data dimension. Selected better features to train the model.

- Built models with two measures, unsupervised model Markov Chain and supervised models Random Forest, SVM, ANN.
- Applied three boosting algorithms (GBR, LightGBM, XGBoost) to improve output after bagging models.

**Statistical Analysis of Occupational Hazard Factors on Professional Drivers** *March 2016 - June 2016*

***Analyzed questionnaires to get occupational hazard factors and used the Perceived Stress Scale to quantify subjects' stresses.*** ***Beijing, China***

- Designed and collected over 240 questionnaires from random samples of professional drivers. The survey included information regarding drivers' physical, chemical, mental, and dietary elements.
- Developed SAS statistical analysis with data; determined corresponding 12 occupational hazard factors by analyzing data. Proposed improvement measures based on existed occupational hazard factors.

**Optimization Project of Enterprise Occupational Disease** *March 2015 - March 2016*

***Team Leader*** ***Beijing, China***

***Funded by Beijing University Students' Scientific Research and Entrepreneurial Project***

- Conducted field investigation in 3 companies, and suggested optimization measures on aspects such as reporting process, government functions, and employees. Analyzed the last 5 years of healthcare management investigation data by SAS.
- Compared other countries' health systems with the Chinese healthcare system, analyzed and built an optimal system for 6 controlling occupational diseases in China. Sought advice from occupational safety experts and health managers.

## PROFESSIONAL EXPERIENCE

**Algorithm Engineer**

**At the Intelligent Research Institute, Nanjing FiberHome Technology Co., Ltd., Nanjing, China**

**English Speech Recognition** May 2022 - April 2023

***Led the development of the company's first English Speech Recognition product, achieving accuracy comparable to our top-tier ASR systems developed over several years.***

- Crawled over 60,000 hours of labeled open-source English audio from websites. Collected and labeled 600 recordings featuring Chinese accents, thereby diversifying the dataset and enhancing recognition accuracy.
- Utilized SpecAugment for data augmentation, enhancing the model's ability to handle diverse speech variations. Employed Python for dataset and label preparation.
- Built and trained a transformer-based model. Focused on hyperparameter tuning and utilized early stopping strategies during training, based on gradient descent insights. Finalized the best model with a WER of 0.163.
- Thoroughly tested the optimal model, monitoring gradient descent behavior. Prepared and packaged the model for efficient deployment.

**Mandarin Speech Recognition** February 2022 - May 2022

***Improved Mandarin speech recognition accuracy by 0.124 in comparison with older versions of the product.***

- Prepared datasets and labels for training and testing. Developed a phoneme dictionary.
- Completed data augmentation using Clipping, adding noises, Time Shift Augmentation, Pitch Shift Augmentation, and SpecAugment. Compared these methods, ultimately selecting SpecAugment for its superior performance.
- Built and optimized a Conformer-based model, adjusting learning rates, batch sizes, and other relevant hyperparameters. Applied layer normalization to prevent overfitting, ensuring that the model generalizes well to new data.

**Cantonese Speech Recognition** October 2021 - December 2021

***Improved Cantonese speech recognition accuracy by 0.163 in comparison with older versions of the product.***

- Investigated Cantonese pronunciation rules, such as the division of pitch; used Python to generate a Cantonese pinyin dictionary. Employed the Ali synthesis interface to synthesize 70,000 audios of different timbres in batches.
- Built and trained a DeepSpeech-based model, observed the convergence of the model and the change of Wer; adjusted the parameters based on the gradient descent's observation; and cut the model, achieving a WER of 0.167 in the best model iteration.

**Multi-channel speech recognition** June 2021- September 2021

***Assisted the team in winning the second-place prize in the "2021 China Hualu Cup Data Lake Algorithm Competition".***

- Crawled data from websites to compile a training high-quality dataset; collected a diverse range of voice data to cover various speech patterns and accents, laying a solid foundation for efficient model training.
- Innovated in merging and weighting voice data to create multi-channel audio, simulating real-world conversations. Developed an automated labeling system, enhancing supervised model training.
- Applied beamforming techniques to enhance signals from specific directions, enabling effective separation and identification of multiple sources in mixed audio, significantly boosting signal recognition accuracy.

**Uighur Speech Recognition** March 2021- May 2021

***Increased the accuracy of Uyghur speech recognition by 0.137 in comparison with older versions of the product.***

- Researched features of Uyghur speech data, including the number of characters contained in the longest and shortest sentences, etc.; visualized the results by Python; then cut audio.
- Screened the missing characters in Uyghur labels against existing labels, and generated new labels; used Python to convert audio data from ".flac" format to ".wav" format in batches, and generated new labels in the corresponding format.
- Built and trained the model, then observed the convergence of the model and change of Wer.