Haibin Guan

Jackson Heights, NY, United States | haibin.guan@mssm.edu | www.linkedin.com/in/haibin-guan

WORK EXPERIENCE

Icahn School of Medicine at Mount Sinai — **NY, United States**
Biological Data Analyst (5/2022 - Now)

INFOTECHSoft — **Miami, Fl, United States**
Biological Data Analyst (1/2018 - 7/2018)

EDUCATION

Florida International University — **Miami, Fl, United States**
MS in Data Science (1/2017–8/2018) — **transfered to MS in Computer Science (8/2018-5/2019)**

Wenzhou Medical University — **Wenzhou, Zhejiang, China**
Bachelor in Preventive Medicine — **9/2011–7/2016**

PROJECTS

General Metabolomics Data Preprocessing Worflow — **02/2023-05/2023**
Icahn School of Medicine at Mount Sinai, United States

- **Several existing methods for correcting intra-batch and inter-batch effects in metabolomics data were explored, implemented, and compared. These included both data-driven approaches such as EigenMS, Combat, SVA, WaveICA, and WaveICA2, as well as QC-based approaches like QC-RSC, QC-RFSC, QC-SVRC, and PQN. Additionally, IS-based normalization methods like RUV, CRMN, and NOMIS were also applied.**
- **Implemented several imputation methods to overcome the missing values in untargeted metabolomics data, such as Probability PCA, Bayesian PCA, Random Forest, and K-Nearest Neighbors. The performance of different imputation methods on the same dataset was compared before choosing a specific method.**
- **Generalized/customized the data preprocessing report based on the selected optimal batch correction method, imputation algorithm, logarithmic transformation, and blank filtering based on desired SNR threshold. The processed data was then visualized through PCA plots, histogram of RSD distribution, and hierarchical clustering dendrogram, etc.**

PFAS Exposure and Thyroid Cancer Risk — **10/2022-03/2023**
Icahn School of Medicine at Mount Sinai, United States

- **A study involving 176 case-control pairs from BioMe bank at the Icahn School of Medicine at Mount Sinai was conducted to explore the potential association between PFAS exposures measured in plasma using liquid chromatography-high resolution mass spectrometry and Thyroid Cancer.**
- **Created a pre-processing workflow to improve the accuracy and reliability of downstream analyses, which includes removing compounds with a large portion of missing values, Batch Effect correction, excluding compounds with low Signal Noise Ratio, and normalization.**
- **Used Logistic Regression and Cox Proportional-Hazards model with covariates adjusted to estimate the association between Thyroid Cancer and PFAS exposures. Additionally, a statistical technique such as Inverse Probability Weights was employed to adjust for potential selection bias.**

Automated Untargeted Metabolomics Data Batch Report — **05/2022-10/2023**
Icahn School of Medicine at Mount Sinai, United States

- **An automated peak analysis workflow was modified and improved for untargeted metabolomics mass spectral data obtained from mass spectrometry (MS) instruments by using R-based tools such as XCMS and IPO (parameter optimization for XCMS). The workflow includes peak detection, retention time correction and alignment, peak grouping, and peak quality visualization tools (e.g. plot of the Total Useful Signal by injection orders to estimate the run order effect; PCA plot to visualize the potential outliers).**
- **The peak results from XCMS-IPO workflow were compared with the peak results generated from IDSL.IPA pipeline by analyzing the extracted ion chromatograms(EIC) for matched peaks with comparable m/z ranges and minimal retention time differences.**

TReNDS Neuroimaging Kaggle Competition 4/2020-6/2020
Paticipant Kaggle, United States

- Applied PCA on 2 sets of given preprocessed features: static FNC correlation features and sMRI SBM loadings to reduce the dimension.
- Trained and optimized 3 baseline regression models (SVR, KNN and ElasticNetCV) with the above top 500 PCA components to predict the corresponding age and assessment value.
- Added a simple blending model with optimized blending weights on top of those 3 models in the end of my ML pipeline boosted the performance (feature-weighted, normalized absolute errors on 10-fold CV) by 0.0127.
- The best model among my total 76 submissions got a final local CV of 0.15986 and a final LB of 0.15833. Ended up getting ranked 27/1047, a solo silver.

Metabolomics Data Integration System 1/2018-7/2018
Biomedical Data Analyst Intern INFOTECHSoft, Miami, Fl, United States

- Assisted in an NIH-funded research project aiming to develop a Metabolomics Data Integration System.
- Integrated multiple metabolite data across various laboratory platforms.
- Performed a complete QIIME 2 workflow including demultiplex, denoise/cluster, taxonomy classification, alignment, and diversity analysis with the FASTQ data.
- Worked collaboratively with the bioinformatics researcher and software engineers.

Multiple Myeloma DREAM Challenge 7/2017-12/2017
Participant Biorg Lab, Florida International University

- Used an algorithm that creates datasets with balanced distributions by combining oversampling by SMOTE and undersampling by Tomek due to the given four training data sets in the competition being known to have distinct subject profiles. The R Bioconductor limma package was used for assessing differential gene expression and returned a ranking of the genes. Utilized SVM-RFECV and MRMR approaches to perform feature selection further.
- The selected genes were then used to train six baseline classifiers: Support Vector Machine (SVM), Neural Network (NN), Random Forest (RF), Gradient Boosting Machine (GBM), Learning Vector Quantization (LVQ), and Generalized Linear Model (GLM). Applied stacking strategy in the end by training ensemble classifiers to combine the results of the mentioned baseline classifiers (In the 2nd round of competition, this model achieved the first place out of 40 teams).

Modeling of influenza-like illness prediction based on Elman neural network 1/2016-6/2016
CDC Disease Surveillance Intern Zhejiang Provincial CDC, Hangzhou, Zhejiang, China

- 10 highly correlated weather factors were selected by examining the correlation between time series variables contemporaneously and at 7 lagged values in R.
- Generated a predictive model by implementing Elman Neural Network (a 3-layer RNN) in MATLAB to provide early detection of the influenza pandemic in Zhejiang Province (the optimal model with NN structure 10-15-1 obtained 10.58 % mean error rate and 0.8767 nonlinear correlation coefficient).

TECHNICAL SKILLS

Programming Languages:	R, Python, MATLAB, Linux, XML/HTML, D3.js, CSS, LaTeX, SPSS
Frameworks:	TensorFlow, PyTorch
Development Tools:	Git, Docker, AWS
Database:	MySQL, PostgreSQL
Bioinformatical analysis Tools:	QIIME2, CAMERA, PICRUSt, Kraken, XCMS, GNPS

PUBLICATIONS

- van Gerwen, Maaike and Colicino, Elena and Guan, Haibin and Dolios, Georgia and Nadkarni, Girish N. and Vermeulen, Roel C.H. and Wolff, Mary and Arora, Manish and Genden, Eric Michael and Petrick, Lauren M., Per- and Polyfluoroalkyl Substances (PFAS) Exposure and Thyroid Cancer Risk. Available at SSRN: https://ssrn.com/abstract=4397033
- Pedro Soto, Ilia Ilmer, Haibin Guan, Jun Li, “Lightweight Projective Derivative Codes for Compressed Asynchronous Gradient Descent”. Proceedings of the 39th International Conference on Machine Learning, PMLR 162:20444-20458, 2022. Available: https://proceedings.mlr.press/v162/soto22a.html

- Pedro Soto, Haibin Guan, Jun Li. Locally Random P-adic Alloy Codes with Channel Coding Theorems for Distributed Coded Tensors. Available: https://arxiv.org/abs/2202.03469v2
- Ruiz-Perez, D., Guan, H., Madhivanan, P. et al. So you think you can PLS-DA?. BMC Bioinformatics 21, 2 (2020). https://doi.org/10.1186/s12859-019-3310-7
- T. Zhang, H. Guan, F.Li, and F.He, "Modeling of influenzalike illness prediction based on Elman neural network," Preventive Medicine, vol. 31, no. 2, p. 113, 2019. [Online]. Available: http://www.zjyfyxzz.com/CN/Y2019/V31/I2/113

EXTRA-CURRICULAR

Oxford Machine Learning Summer School Participant	Oxford, United Kingdom 7/2023
YOUTH ASSEMBLY OF THE UNITED NATIONS Youth Assembly Delegate	New York, the United States 2/2016
Summer Work Travel Program Exchange Visitor	Estes Park,Colorado, the United States 6/2015-9/2015
International Study in Gachon University Exchange Visitor	Seoul, Korea 1/2015-2/2015