TIMESPAN

Management of chronic cardiometabolic disease and treatment discontinuity in adult ADHD patients

H2020 – 965381

D6.1. – DLNN algorithms (freely available via GitHub) (Task 1)

Dissemination level	Public
Contractual date of delivery	30. September 2021
Actual date of delivery	24. September 2021
Type	Report
Version	1
Filename	TIMESPAN_Deliverable Report_D6.1.
Workpackage	6
Workpackage leader	Stephen Faraone (SUNY)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 965381.

Author list

Organisation	Name	Contact information
SUNY Upstate Medical University	**Stephen V Faraone**	**sfaraone@childpsychresearch.org**
SUNY Upstate Medical University	**Yanli Zhang-James**	**ZhangY@upstate.edu**

Abbreviations

ML	Machine learning
DL	Deep Learning
MLP	Multilayer perceptron
GRU	Gated recurrent *unit*

Table of Contents

1. Executive Summary

The main objective of the D6.1 is to create innovative data structures DLNNs to predict cardiometabolic outcomes and treatment discontinuity using registry and clinical data. Our machine learning and deep learning framework for this objective is now complete and the codes are freely available via Github repository (https://github.com/ylzhang29/ML-DL_Framework).

2. Deliverable report

The registry and clinical data are typically tabular data. There is a number of machine learning (ML) and deep learning (DL) models that are especially suitable for tabular data and have been used successfully in our previous work with various different type of tabular data including registry and clinical data, genetic and transcriptomic data, tabular outputs of the magnetic resonance imaging data, as well as COVID-19 epidemiological data (Chen et al., 2020; Faraone, James, Chen, & Larsson, 2019; Tylee et al., 2017; Yanli Zhang-James, 2020; Y Zhang-James, Buitelaar, The ENIGMA- ASD Working Group, van Rooij, & Faraone, In Press (2021); Y. Zhang-James et al., 2020; Yanli Zhang-James, Glatt, & Faraone, 2019; Y. Zhang-James, Helminen, et al., 2021; Y. Zhang-James, Hess, et al., 2021).

To accommodate the wide varieties and sources of the data that we will be analysing within TIMESPAN, we have designed a ML/DL framework that incorporates these models and their supplementary methods aiding data input/preprocessing, feature engineering/dimension reduction, model hyperparameter search, model stacking and ensemble, as well as inferring model interpretability. The codes for these tools and models are freely accessible via our github repository (see the link above). Sufficient documentations are included within the code files to facilitate adaptation to user-specific dataset.

All codes are written in Python, using Scikit-learn (Pedregosa et al., 2012), Keras (Charles, 2013) and Tensorflow libraries (Abadi et al., 2016; GoogleResearch, 2015).

Briefly, this repository contains the following files:

1. Read input tabular data (including generate training, validation and testing subsets; scaling features and binarize targets, i.e our outcomes of interests such as cardiometabolic diagnosis or events)
2. PCA feature reduction: a commonly used feature reduction and engineering method
3. Commonly used Scikit-learn models (including ensemble models) for tabular data.
4. Scikit-learn model hyperparameter search (covering a wide range of models and hyperparameters, and all of the commonly used search algorithms)
5. Multilayer perceptron (MLP) model: A neural network model suitable for tabular data.
6. Hyperopt search for MLP: Hyperparameter search algorithm for the MLP models using Hyperopt (http://hyperopt.github.io/hyperopt/)
7. Ensemble-MLP model: generate ensemble MLP model and stabilized predictions
8. Seq2Seq model with GRUs (Dey & Salem, 2017; Wu et al., 2016): a longitudinal neural network model that will use time-series data input and predict the future events or event serials (Y. Zhang-James, Hess, et al., 2021).
9. Feature importance analysis: a collection of various methods to examine and extract feature importance scores for various of models.

3. Conclusion

We have now completed D6.1. These models (and their supporting methods) are freely available for the scientific community and will be readily implemented in our next phase analysis as soon as the data access and/or transfer are complete.

References:

Abadi, M., Barham P, Chen J, Chen, Z., Davis, A., Dean, J., . . . Brain, G. (2016, November 2–4). *TensorFlow: A system for large-scale machine learning.* Paper presented at the Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, Savannah, GA, USA.

Charles, P. W. D. (2013). Keras repository. *GitHub.*(https://github.com/charlespwd/project-title).

Chen, Q., Zhang-James, Y., Barnett, E. J., Lichtenstein, P., Jokinen, J., D'Onofrio, B. M., . . . Fazel, S. (2020). Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: A machine learning study using Swedish national registry data. *PLoS Med, 17*(11), e1003416. doi:10.1371/journal.pmed.1003416

Dey, R., & Salem, F. M. (2017, 6-9 Aug. 2017). *Gate-variants of Gated Recurrent Unit (GRU) neural networks.* Paper presented at the 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS).

Faraone, S., James, Y. Z., Chen, Q., & Larsson, H. (2019). 15. Predicting Comorbid Disorders in ADHD Using Machine Learning. *Biological Psychiatry, 85*(10), S6. doi:10.1016/j.biopsych.2019.03.029

GoogleResearch. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. doi:10.1109/TIP.2003.819861.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830. doi:10.1007/s13398-014-0173-7.2

Tylee, D. S., Hess, J. L., Quinn, T. P., Barve, R., Huang, H., Zhang-James, Y., . . . Glatt, S. J. (2017). Blood transcriptomic comparison of individuals with and without autism spectrum disorder: A combined-samples mega-analysis. *Am J Med Genet B Neuropsychiatr Genet, 174*(3), 181-201. doi:10.1002/ajmg.b.32511

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., . . . Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv, abs/1609.08144*.

Zhang-James, Y. (2020). Improved Classification Performance With Autoencoder-Based Feature Extraction Using Cross-Disorder Datasets. *Biological Psychiatry, 87*(9), S87-S88. doi:10.1016/j.biopsych.2020.02.246

Zhang-James, Y., Buitelaar, J. K., The ENIGMA- ASD Working Group, van Rooij, D., & Faraone, S. V. (In Press (2021)). Ensemble Classification of Autism Spectrum Disorder Using Structural MRI Features. *JCPP Advances*.

Zhang-James, Y., Chen, Q., Kuja-Halkola, R., Lichtenstein, P., Larsson, H., & Faraone, S. V. (2020). Machine-Learning prediction of comorbid substance use disorders in ADHD youth using Swedish registry data. *J Child Psychol Psychiatry*. doi:10.1111/jcpp.13226

Zhang-James, Y., Glatt, S. J., & Faraone, S. V. (2019). Nu Support Vector Machine in Prediction of Fluid Intelligence Using MRI Data. In K. M. Pohl, Wesley K. Thompson, Ehsan Adeli, & M. G. Linguraru

(Eds.), *Adolescent Brain Cognitive Development Neurocognitive Prediction* (pp. 92-98). Lecture Notes in Computer Science: Springer International Publishing.

Zhang-James, Y., Helminen, E. C., Liu, J., Group, E.-A. W., Franke, B., Hoogman, M., & Faraone, S. V. (2021). Evidence for similar structural brain anomalies in youth and adult attention-deficit/hyperactivity disorder: a machine learning analysis. *Transl Psychiatry, 11*(1), 82. doi:10.1038/s41398-021-01201-4

Zhang-James, Y., Hess, J., Salekin, A., Wang, D., Chen, S., Winkelstein, P., . . . Faraone, S. V. (2021). A seq2seq model to forecast the COVID-19 cases, deaths and reproductive R numbers in US counties. *medRxiv*. doi:10.1101/2021.04.14.21255507