Diagnostic AI Modeling and Pseudo Time Series Profiling of AD and PD Based on Individualized Serum Proteome Data

Jianhu Zhang

Fujian Provincial Key Laboratory of Brain Aging and Neurodegenerative Diseases, School of Basic Medical Sciences, Fujian Medical University

Xiuli zhang (✉ zhxiuli@gmail.com)

CAS Key Laboratory of standardization and measurement for nanotechnology, C

Yuan Sh

Fujian Provincial Key Laboratory of Brain Aging and Neurodegenerative Disease, School of Basic Medical Sciences, Fujian Medical University

Benliang Liu

Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

Zhiyuan Hu

National Center for Nanoscience and Technology

Research

Keywords: Alzheimer's disease, Parkinson's disease, Mild cognitive impairment, artificial intelligence, predictive diagnostics

Posted Date: July 27th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-721593/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

Parkinson's disease (PD), Alzheimer's disease (AD) are common neurodegenerative disease, and mild cognitive impairment (MCI) may be happened in the early stage of AD or PD. Blood biomarkers are considered to be less invasive, less cost and more convenient, and there is tremendous potentia for the diagnosis and prediction of neurodegenerative diseases. As a recently mentioned field, artificial intelligence (AI) is often applied in biology and shows excellent results.

Method

Human blood protein microarray profiles including 156 CT, 50 MCI, 132 PD, 50 AD samples are collected from Gene Expression Omnibus (GEO). First, we used bioinformatics methods and feature engineering in machine learning to screen important features, constructed ANNclassifier models based on these features to distinguish samples, and evaluated the model's performance with classification accuracy and Area Under Curve (AUC). Secenod, we used Ingenuity Pathway Anaylsis (IPA) methods to analyse the pathways and functions in early stage and late stage samples of different diseases, and potential targets for drug intervention by predicting upstream regulators.

Result

Overall, we incorporated six indicators, including EPHA2, MRPL19, SGK2, to build a diagnostic model for AD with a test set accuracy of up to 98.07%. Incorporating 15 indicators such as ERO1LB, FAM73B, IL1RN to build a diagnostic model for PD, with a test set accuracy of 97.05%. Thirty indicators such as XG, FGFR3 and CDC37 were incorporated to establish a four-category diagnostic model for both AD and PD, with a test set accuracy of 98.71%. In addition,we found that early PD may occur earlier than early MCI. Finally, there are 24 proteins that may serve as potential therapeutic targets

Conclusion

Using deep learning methods to build classifiers based on blood protein profile can achieve better classification performance, and it helps us to diagnose the disease early. In total, it is important for us to study neurodegenerative diseases from both diagnostic and interventional aspects.

Introduction

Neurodegenerative diseases are nervous system disorders that manifest the progressive loss of function or structure of neurons, including the death of neurons. The most widely studied neurodegenerative diseases are Alzheimer's disease (AD) and Parkinson's disease (PD). AD, the leading cause of dementia [1], is a disease associated with cognitive impairment, presents with learning, language, and memory impairment [2]. PD is a complex brain system disorder that not only affects movement such as rigidity, slowness, and tremor but also impacts cognition. Mild cognitive impairment (MCI) is characterized by

ongoing memory problems, be regarded as the symptomatic predementia phase of AD and as non-motor symptom occurred in early-stage PD [3, 4]. It's important that distinguish MCI status since some studies indicate that MCI may lead to progression to AD and PD with MCI may have higher risks of developing Parkinson's disease dementia (PDD) [5, 6].

When an individual is diagnosed with AD or PD, the pathological damage in the brain has actually occurred for some time, which is irreversible [7, 8]. Therefore, early diagnosis of the two types of disease is very necessary, blood biomarkers are getting more attention because of more convenient sampling, less cost, and less risk. Eric P. Nagele found that differentially expressed proteins (DEPs) in human blood could better distinguish AD, PD, and MCI from normal samples respectively [9–11]. For example, the accuracy of distinguishing AD from normal samples was 93.4%, while the accuracy of distinguishing PD from normal samples was 97.1%. Although there have been many studies in this field, few researchers have studied MCI, PD and AD together. In this paper, on the basis of the protein expression profile of serum, we build model to distinguish which kind of a sample belong to between AD, MCI, PD, and CT and search possible biological phenomenon and drug targets.

Materials And Methods

Data sources and preprocessing

We downloaded multiple protein expression profiles from the GEO database (<https://www.ncbi.nlm.nih.gov/geo/>), including GSE29654, GSE62283, GSE74763. These datasets are all generated by the Invitrogen ProtoArray V5.0 platform. Since there are some duplicate samples in these three datasets, we preserve samples with the earlier upload time for duplicate samples. We also delete the samples from the same institution but the sample size less than 3. Finally, the number of samples in each category, as shown in Table 1.

The format of raw data is GPR, we use the R package PPA to load it, then use the Robust Linear Model (RLM) method to process data. Finally, extract the common probes and merge the expression profiles. We standardized the data using the following method, for each gene in each sample, we calculated the ratio of this gene to the total gene expression in the sample and multiplied this ratio by 1 million as the final expression value.

Table 1
Overview of the collected data

Group	Sample numbers	Age(Median(Range))	Sex(% male)
Control	156	56.50(19–86)	56.41
Mild cognitive impairment	50	73.00(55–91)	58.00
Alzheimer's disease	50	79.00(61–97)	43.47
Parkinson's disease	132	66.00(37–88)	57.69

Model of machine learning and deep learning

In the process of construct model, four types of machine learning classifiers, including naive Bayes (NB) [12], k-nearest neighbor (KNN) [13], decision tree (DT) [14], random forest (RF) [15], and one deep learning classifiers, ANN were used. Machine learning models are store in the scikit-learn 0.23.1 with Python 3.8.3 [16]. When building an ANN, we use the Keras 2.4.3 module.

Model construct

The DEPs were recognized with the limma Bioconductor package in R (Version 3.6.3) [17]. First, we extracted the union of DEPs between pairs of categories as the initial features. According to the variance of these initial features, proteins with variance less than a quarter of the population were eliminated. Then the correlation was used to remove the proteins with a correlation coefficient greater than 0.7 with other proteins. Finally, we used an SVM-based model to extract the top N features that were most important to model construction [18, 19]. For all the classifiers, the ratio of the training set, validation set, the test set is 3:1:1. The training set is used to train the model, the validation set selects the optimal model parameters, and the test set to evaluate model performance.

IPA and Protein phase separation

The IPA was used for biological analysis, which included canonical pathway analysis, disease, and function, upstream regulators. The $-\log_{10}(P\text{-value}) > 1.3$ as threshold that statistically significant, and Z-score greater than 0 is defined as active, else suppressed. For protein phase separation analysis, we uploaded the sequences of the proteins to the PLAAC (<http://plaac.wi.mit.edu/>) to get a phase separation score [20].

Results

Development of individualized diagnostic models and analysis process for AD and PD patients

Based on serum protein expression profiles, we used artificial intelligence to construct three individual disease diagnostic models. In addition, the biological pathways, functions, upstream factors, and Pseudo-timing information between diseases are mined (Fig. 1). 388 serum protein expression profiles were downloaded from the GEO database, which contained 156 Control samples (CT), 50 MCI, 50 AD, and 132 PD samples. On the one hand, the optimal feature of the model for constructing the model are firstly filtered according to the difference, variance, colinearity, and importance. Then use the optimal feature to train different classifier models, including random forest (RF), Decision Tree(DT), and Navie Bayes (NB), Artificial neural network (ANN), k-Nearest Neighbor (KNN). The trained model was applied to test set to observe the classification effectiveness (accuracy, confusion matrix, ROC) and feature effectiveness (TSNE) of the model. On the other hand, we analysed the pathways and functions between disease and

normal samples, followed by the possible order of occurrence of the disease. Finally, we analysed the upstream regulators and possible drug targets.

The AI model for the diagnosis of AD, MCI and CT based on 6 serum protein markers

AD, CT, and MCI samples were extracted from the data set, 1879 DEPs between the AD, CT, and MCI were detected (Fig. 2a). When constructing feature engineering, we follow the following principles: 1) Features with small variance do not have large impact on the classifier. 2) Highly correlated features may lead to covariance problems in the model. 3) A few important features are sufficient to represent the whole range of features. After variance, correlation and importance screening (Supplementary Fig. 1a-b), six features were finally obtained, containing *LOC728492*, *PCBD2*, *EPHA2*, *MRPL19*, *SGK2*, *LGALS1*. These six optimal features were expressed significantly differently between groups, and their importance was shown in the figure (Fig. 2b-c).

We use different classifiers (KNN, RF, ANN, NB, DT) and different features (optimal features, random features, all features) to build models. Finally, the optimal features can achieve similar even better classification performance than all features, and this is not due to randomness (Fig. 2d). The accuracy and loss curves for these six features during ANN model training (Fig. 2e) show that we stopped training when the model was stable. The Micro-AUC for the optimal features was 0.9994, higher than 0.9191 for all features and 0.6385 for random features (Fig. 2f). The accuracy of this model in all three test sets is greater than 0.95 (Fig. 2g-i), and their AUC in the test set are shown in the supplemental Fig. 4a-c. The model accuracy in the all test set was 98.07%, where MCI and AD classification being completely correct, outperforming all features and random features (Fig. 2j-l). Compared to all features and random features, optimal features can distinguish samples well (Fig. 2m-o). The above results show that the optimal features selected after feature engineering help to improve performance and simplify the model. We defined 0 for the CT, 0.5 for the MCI and 1 for the AD sample to analyse the correlation between optimal features and disease progression. Most features were positively correlated with the severity of cognitive loss, except for MRPL19. EPHA2 is a neuroinflammatory factor (Supplementary Fig. 1c), which may indicate that the neuroinflammatory pathway in which EPHA2 resides is closely related to the progression of AD.

The AI model for the diagnosis of PD, MCI and CT based on 15 serum protein markers

We extracted PD, CT and MCI samples from the dataset and used 3092 DEPs as initial features (Fig. 3a). Finally, after feature selection, 15 features were retained (Supplementary Fig. 1d-e), containing *ERO1LB*, *IGLa*, *LOC400763*, *PHKG2*, *PPM1L*, *RAD51L3*, *IL23A*, *DYNLRB2*, *BCAT1*, *CDC37*, *IL1RN*, *MAB21L2*, *S100A13*, *FAM73B*, *IP6K2*. Heatmaps of the 15 features also showed significant differences between groups (Fig. 3b-c). Similarly, the ANN model with feature engineering performs best (Fig. 3d). When the model tends to be stable, the classification accuracy is the highest and the loss is the lowest (Fig. 3e). The test set accuracy of the optimal features was 97.05%, where the MCI classification was completely accurate with micro-AUC of 0.9984, while the all features were 0.83343 and the random features were 0.7897 (Fig. 3f,j-l). The accuracy of this model in all three test sets is greater than 0.94 (Fig. 3g-i), and

their AUC in the test set are shown in the supplemental Fig. 4d-f. The optimal features distinguished the MCI samples well compared to all features and random features (Fig. 3m-o). Finally, we also analysed the correlation between optimal features and disease progression (Supplementary Fig. 3f). Among these features, IL23a and IL1RN are pro-inflammatory cytokines and anti-inflammatory factors, respectively. MAB21L2 may be related to neurodevelopment [21], and BACT1 knockout may cause neuronal oxidative damage [22].

The AI model for the diagnosis of AD, PD, MCI and CT based on 30 serum protein markers

Similarly, we took out the DEPs of all samples for feature filtering to obtain the optimal model with 30 features (Fig. 4a-b), among which *PCBD2*, *LGALS1* belong to the features in model1, while *IGLa*, *ERO1LB*, *MAB21L2*, *CDC37*, *DYNLRB2*, *FAM73B*, *IP6K2*, *S100A13* belong to the features in model2, which indicates that the features extracted by feature engineering have good robustness, and the importance of these 30 features is shown in the figure (Fig. 4c). The filtered features were also optimal in the Artificial neural networks (ANN) model compared to other methods and other classifiers (Fig. 4d). The accuracy of this model in all three test sets is greater than 0.95 (Fig. 4g-i), and their AUC in the test set are shown in the supplemental Fig. 4g-i. Compared to all features and random features, the classification accuracy of all test sets is 98.71%, and all samples are correctly classified except a PD samples which are misclassified as AD, and the micro-AUC reaches 0.9999 (Fig. 4f,j-l), which was greater than 0.8541 for all features and 0.6660 for random features, and could completely identify MCI samples in the TSNE (Fig. 5m-o).

The serum proteins of patients in the MCI, AD and PD groups all showed different differences from the healthy sample

In this paper, we first analysed the DEPs of disease. The number of DEPs in MCI, AD and PD compared to CT was 1010, 839 and 2122 respectively. The number of DEPs in AD and PD compared to MCI was 1221 and 1467 respectively. Finally, the number of DEPs between AD and PD was 2082 (Supplementary Fig. 2a). Firstly, PD was very different from CT, MCI, AD in terms of the number of DEPs, but was closest to MCI (Supplementary Fig. 2b). Next, we found that the phase change proteins in MCI, AD, and PD had relatively large differences in location and cell type, with PD is mostly distributed in the nucleus and enzymes, AD is mostly distributed in the extracellular space and transcriptional regulators, and MCI is mostly distributed in the cytoplasm (Supplementary Fig. 2c-d). In addition, we found differences in phase separation scores in the cytoplasm between PD and normal samples, which may indicate that phase separation in PD is associated with the cytoplasm (Supplementary Fig. 2e). Further analysis of the cell type scores in the cytoplasm revealed that the differences may lie in other cell types. Finally, we show the 10 proteins with the largest differences in disease relative to normal (Supplementary Fig. 2g-i), where *EMG1*, *IFI6* are the most up-regulated and down-regulated DEPs for MCI relative to normal, *ZCD2*, *IFI6* are the most up-regulated and down-regulated DEPs for AD, and *CCT7*, *RANBP6* are the most up-regulated and down-regulated DEPs for PD.

Early PD may occur before early MCI

Serum molecules flow with the blood and can affect the body's cells, tissues and organs in a comprehensive way. With regard to the biological events influenced by serum molecules, we further analysed the activation levels of individual biological events based on conventional significance analysis. We classified the disease in more detail based on the underlying information, dividing MCI into early MCI (EMCI) and late MCI (LMCI), AD into early mild-moderate AD (EMMAD) and late mild-moderate AD (LMMAD), and PD into early PD (ESPD) and mild-moderate PD (MMPD). By observing the canonical pathways and disease and bio functions, we could find that the number of up-regulated pathways increased and the number of down-regulated pathways decreased in the process from EMCI/CT to LMMAD/CT (Fig. 5a-b). Z-scores, the mean change in pathway relative to control samples, showed the same trend. ESPD followed the same trend as EMCI but with greater variation. The results show that there is a continuum of inertia between multiple biological events in the organism of MCI and AD patients, while PD is more distinct from both. The occurrence of biological events in the organism of patients with early PD was intermediate between healthy and early MCI. This suggests that early PD may precede early MCI.

Similarly, IPA was used for the analysis of 7 groups of samples (Fig. 5c). Among the canonical pathways, we choose the 10 pathways with the largest relative differences between MMPD and LMMAD. Long-term activation of EIF2 leads to continuous decline in protein synthesis, which leads to memory impairment and neuronal damage [23]. The up-regulation ratio of EIF2 in MCI is small, while in AD and PD is larger, which may indicate that EIF2 is more related to neuronal damage.

Among the classical pathways, we identified two pathways associated with the Coronavirus, namely the "Coronavirus Replication Pathway" and the "Coronavirus Pathogenesis Pathway". Coronavirus Replication in disease was enhanced but pathogenicity was reduced compared to normal samples. The Coronavirus replication ability of AD was stronger than that of PD. It is known from the literature that patients with COVID-19 appear to be more susceptible to AD and that patients with AD may be more susceptible to severe infection with COVID-19 [24]. In contrast, the current literature does not clearly indicate whether PD patients are more susceptible to COVID-19. This may reveal a greater susceptibility to COVID in AD.

The level of cell maturation is relatively low in the early stages of disease compared to normal samples, while in the middle and late stages of disease progression, cell maturation begins to increase abnormally to near even greater than normal levels. In terms of molecular function, excessive increases in activating nuclear factor NF- κ B have been shown to play an important role in driving Abeta deposition, neuroinflammation and neurodegenerative disease in AD, but NF- κ B levels are not increased in PD, which may suggest that NF- κ B does not promote α -SYN deposition [25].

Possible therapeutic targets and drugs

Finally, we predicted the upstream regulators that may cause differential protein profiles in patients. The upstream regulator of DEPs and the corresponding drug treatment information were obtained through IPA annotation, of which 85 upstream regulators correspond to 837 drugs. In addition, 170 DEPs correspond to 911 drugs. These 231 kinds of DEPs and upstream regulators are potential therapeutic targets, and

1445 kinds of drugs can be considered as treatment options (Fig. 6a-b). The expression of 231 potential targets in the 7 groups of samples is shown in Fig. 6c. Among them, we can observe that ESPD is the closest to normal, which may also reflect the earlier onset of ESPD. Then, in order to further narrow the scope, we extracted 24 proteins that belong to both upstream regulator and DEPs. The predicted expressions of these 24 upstream regulators are shown in Fig. 6D and the drugs corresponding to all proteins are listed in Supplementary table 1. In addition, machine learning models of LGALS1 also appeared in 24 upstream factors. In the early stages of the disease, LGALS1 expression levels were reduced, along with reduced protein activity. In both AD and PD patients, LGLAS1 expression and protein activity are activated, which we speculate may be related to the overreaction of the organism. This may suggest the use of activators in the early stages and inhibitors in the late stages, and OTX008 is a target drug for LGALS1. We predict potentially intervenable drugs based on the activation levels of upstream factors, laying the foundation for diagnosis and intervention in neurodegenerative diseases such as AD and PD.

Discussion

In neurodegenerative diseases, patients with MCI are a vague intermediate state, which may not only appear in the early symptoms of AD and PD, but may also turn into normal. There is currently no good treatment for neurodegenerative diseases, and by the time they are diagnosed it is too late. Therefore, the early diagnosis of neurodegenerative diseases is particularly important. It can help us have more time to think and cope with clinical symptoms before they appear.

Although there are lots of researches, many researchers' models still have relatively large limitations. Zehra Karapinar Senturk uses voice data to identify PD samples and normal samples based on feature engineering and SVM classifiers. As a result, the classification accuracy is only 93.84% [26], which may be caused by feature engineering steps that features are only screened for importance. In Jörn Löttsch's study, a classifier was constructed using both olfactory and culinary information, and the machine learning model was able to distinguish non-PD samples with 94.1% accuracy, but only 58.9% for PD samples, which may be due to the extreme sample imbalance in the model training process [27]. In Sanghee Moon's study, which collected data from wearable devices and also used multiple data models, the highest accuracy was only 0.92, with the maximum f1 score was 0.61. In this study, despite the authors' simple feature engineering and oversampling methods, the sample imbalance was still exposed [28]. In Marek Wodzinski's study, audio information was used, but the classification accuracy of the test set was only 0.90 [29]. More importantly, the above-mentioned models are all binary models, which may lead to limited applications. Finally, in this paper we have constructed a reasonable feature engineering and three multi-classifiers with high classification accuracy, all of which can achieve over 97% accuracy.

We use IPA to analyze the activation level of each biological event based on between MCI, AD, PD and CT, and three of them deserve our attention, namely "Neuroinflammatory signaling pathway", "JAK/Stat signaling pathway", "Acute phase response signaling" (Supplementary Fig. 3). Neuroinflammatory signals are immune responses activated by microglia and astrocytes in the central nervous system (CNS),

and are generally considered to be related to neurodegenerative diseases. The JAK/STAT pathway is the primary signaling mechanism for a variety of cytokines and growth factors [30]. Inhibition of the JAK/STAT pathway can prevent neuroinflammation and neurodegeneration by inhibiting the activation of α -SYN by innate and adaptive immune responses [31]. In patients with AD and PD, the JAK/STAT signal pathway was activated, whereas it was reversed in MCI, which is consistent with the neuroinflammatory pathway. Compared with the traditional view that the inflammation that occurs in neurodegenerative diseases is chronic, IPA analysis believes that acute inflammation also occurs in neurodegenerative diseases and plays an important role. Previous studies have shown that the formation of senile plaques in patients with AD may involve acute inflammation [32], and acute inflammation is also related to the severity of PD [33], which suggests that we need to re-examine the role of acute inflammation in neurodegenerative diseases.

Conclusion

In this article, we use blood data to construct three classifiers. These three classifiers can diagnose individual states well. Their classification accuracy is greater than 95%, and the model's classifier AUC is greater than 0.98. The classifier may be effective for early diagnosis of patients, and we have revealed that early PD may be earlier than early MCI samples through IPA analysis.

Abbreviations

PD Parkinson's disease

AD Alzheimer's disease

MCI mild cognitive impairment

GEO Gene Expression Omnibus

AUC Area Under Curve

IPA Ingenuity Pathway Analysis

PDD Parkinson's disease dementia

DEPs differentially expressed proteins

RLM Robust Linear Model

NB naive Bayes

KNN k-nearest neighbor

DT decision tree

RF random forest

ANN artificial neural network

CT control samples

EMCI early MCI

LMCI late MCI

EMMAD early mild-moderate AD

LMMAD late mild-moderate AD

ESPD early PD

MMPD mild-moderate PD

Declarations

Funding

This work was supported by the National Natural Science Foundation of China (32027801, 31870992, 21775031), the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDB36000000, XDB38010400), CAS-JSPS (Grant No.GJHZ2094), Science and Technology Service Network Initiative of the Chinese Academy of Sciences (Grant No.KFJ-STS-ZDTP-079), Research Foundation for Advanced Talents of Fujian Medical University (XRCZX2017020, XRCZX2019005), Beijing Natural Science Foundation Haidian original innovation joint fund (L202023). The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval and Consent to participate

Not applicable

Consent for publication

Not applicable

Acknowledgements

Not applicable

Author information

Affiliations

Fujian Provincial Key Laboratory of Brain Aging and Neurodegenerative Diseases, School of Basic Medical Sciences, Fujian Medical University, Fuzhou 350108, Fujian Province, China

Jianhu Zhang, Yuan Sh, Zhiyuan Hu

CAS Key Laboratory of Standardization and Measurement for Nanotechnology, CAS Key Laboratory for Biomedical Effects of Nanomaterials and Nanosafety, CAS Center for Excellence in Nanoscience, National Center for Nanoscience and Technology of China, Beijing 100190, China

Xiuli Zhang

Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

Benliang Liu

Contributions

Z.H. and X.Z. designed the study. J.Z., X.Z., Y.S. and B.L. performed the analyses and interpreted the results. J.Z. and X.Z. wrote the manuscript. X.Z. and Z.H. conducted this study. All authors read and approved the final manuscript.

Corresponding authors

Correspondence to Zhiyuan Hu.

Data availability

Raw data can be download from the NCBI Gene Expression Omnibus under accession code GSE29654 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29654>), GSE62283 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62283>), and GSE74763 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74763>).

Code availability

This study did not produce new data. The project data analysis codes are deposited in GitHub (<https://github.com/zhxiuli/Al.git>).

Supplementary information

References

1. Long, J. M. & Holtzman, D. M. Alzheimer Disease: An Update on Pathobiology and Treatment Strategies. *Cell* 179, 312-339, doi:10.1016/j.cell.2019.09.001 (2019).
2. McKhann, G. M. et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 7, 263-269, doi:10.1016/j.jalz.2011.03.005 (2011).
3. Albert, M. S. et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 7, 270-279, doi:10.1016/j.jalz.2011.03.008 (2011).
4. Poewe, W. et al. Parkinson disease. *Nature reviews. Disease primers* 3, 17013, doi:10.1038/nrdp.2017.13 (2017).
5. Pagani, M. et al. Early identification of MCI converting to AD: a FDG PET study. *European journal of nuclear medicine and molecular imaging* 44, 2042-2052, doi:10.1007/s00259-017-3761-x (2017).
6. Saredakis, D., Collins-Praino, L. E., Gutteridge, D. S., Stephan, B. C. M. & Keage, H. A. D. Conversion to MCI and dementia in Parkinson's disease: a systematic review and meta-analysis. *Parkinsonism & related disorders* 65, 20-31, doi:10.1016/j.parkreldis.2019.04.020 (2019).
7. Petersen, R. C. Early diagnosis of Alzheimer's disease: is MCI too late? *Current Alzheimer research* 6, 324-330, doi:10.2174/156720509788929237 (2009).
8. Gaig, C. & Tolosa, E. When does Parkinson's disease begin? *Movement disorders : official journal of the Movement Disorder Society* 24 Suppl 2, S656-664, doi:10.1002/mds.22672 (2009).
9. DeMarshall, C. A. et al. Detection of Alzheimer's disease at mild cognitive impairment and disease progression using autoantibodies as blood-based biomarkers. *Alzheimer's & dementia (Amsterdam, Netherlands)* 3, 51-62, doi:10.1016/j.dadm.2016.03.002 (2016).
10. Nagele, E., Han, M., Demarshall, C., Belinka, B. & Nagele, R. Diagnosis of Alzheimer's disease based on disease-specific autoantibody profiles in human sera. *PloS one* 6, e23112, doi:10.1371/journal.pone.0023112 (2011).
11. Han, M., Nagele, E., DeMarshall, C., Acharya, N. & Nagele, R. Diagnosis of Parkinson's disease based on disease-specific autoantibody profiles in human sera. *PloS one* 7, e32383, doi:10.1371/journal.pone.0032383 (2012).
12. Zhang, H. *The Optimality of Naive Bayes*. Vol. 2 (2004).
13. Troyanskaya, O. et al. Missing value estimation methods for DNA microarrays. *Bioinformatics (Oxford, England)* 17, 520-525, doi:10.1093/bioinformatics/17.6.520 (2001).
14. Breiman, L., Friedman, J., Olshen, R. & Stone, C. J.

15. Breiman, L. Random Forests. *Machine Learning* 45, 5-32, doi:10.1023/A:1010933404324 (2001).
16. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830 (2011).
17. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* 43, e47, doi:10.1093/nar/gkv007 (2015).
18. Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* 20, 273-297, doi:10.1007/BF00994018 (1995).
19. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* 46, 389-422, doi:10.1023/A:1012487302797 (2002).
20. Lancaster, A. K., Nutter-Upham, A., Lindquist, S. & King, O. D. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics (Oxford, England)* 30, 2501-2502, doi:10.1093/bioinformatics/btu310 (2014).
21. Wang, D. et al. Daphnetin Ameliorates Experimental Autoimmune Encephalomyelitis Through Regulating Heme Oxygenase-1. *Neurochemical research* 45, 872-881, doi:10.1007/s11064-020-02960-0 (2020).
22. Mor, D. E. & Sohrabi, S. Metformin rescues Parkinson's disease phenotypes caused by hyperactive mitochondria. 117, 26438-26447, doi:10.1073/pnas.2009838117 (2020).
23. Halliday, M. et al. Repurposed drugs targeting eIF2 α -P-mediated translational repression prevent neurodegeneration in mice. *Brain : a journal of neurology* 140, 1768-1783, doi:10.1093/brain/awx074 (2017).
24. Ciaccio, M. et al. COVID-19 and Alzheimer's Disease. 11, doi:10.3390/brainsci11030305 (2021).
25. Lindsay, A., Hickman, D. & Srinivasan, M. A nuclear factor-kappa B inhibiting peptide suppresses innate immune receptors and gliosis in a transgenic mouse model of Alzheimer's disease. *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie* 138, 111405, doi:10.1016/j.biopha.2021.111405 (2021).
26. Karapinar Senturk, Z. Early diagnosis of Parkinson's disease using machine learning algorithms. *Medical hypotheses* 138, 109603, doi:10.1016/j.mehy.2020.109603 (2020).
27. Lötsch, J., Haehner, A. & Hummel, T. Machine-learning-derived rules set excludes risk of Parkinson's disease in patients with olfactory or gustatory symptoms with high accuracy. *Journal of neurology* 267, 469-478, doi:10.1007/s00415-019-09604-6 (2020).
28. Moon, S. et al. Classification of Parkinson's disease and essential tremor based on balance and gait characteristics from wearable motion sensors via machine learning techniques: a data-driven approach. *Journal of neuroengineering and rehabilitation* 17, 125, doi:10.1186/s12984-020-00756-5 (2020).
29. Wodzinski, M., Skalski, A., Hemmerling, D., Orozco-Arroyave, J. R. & Noth, E. Deep Learning Approach to Parkinson's Disease Detection Using Voice Recordings and Convolutional Neural Network Dedicated to Image Classification. *Annual International Conference of the IEEE Engineering in*

- Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference 2019, 717-720, doi:10.1109/embc.2019.8856972 (2019).
30. Murray, P. J. The JAK-STAT signaling pathway: input and output integration. *Journal of immunology* (Baltimore, Md. : 1950) 178, 2623-2629, doi:10.4049/jimmunol.178.5.2623 (2007).
31. Qin, H. et al. Inhibition of the JAK/STAT Pathway Protects Against α -Synuclein-Induced Neuroinflammation and Dopaminergic Neurodegeneration. 36, 5144-5159, doi:10.1523/jneurosci.4658-15.2016 (2016).
32. Sawada, H. et al. Baseline C-Reactive Protein Levels and Life Prognosis in Parkinson Disease. *PloS one* 10, e0134118, doi:10.1371/journal.pone.0134118 (2015).
33. Chen, Y., Fu, A. K. & Ip, N. Y. Eph receptors at synapses: implications in neurodegenerative diseases. *Cellular signalling* 24, 606-611, doi:10.1016/j.cellsig.2011.11.016 (2012).

Figures

Figure 1

Schematic diagram of the workflow of this study. The original dataset was downloaded from the GEO database, which included 156 CT samples, 50 MCI samples, 50 AD samples, and 132 PD samples. The blue arrow represents the process of exploring the biological function of the data, and the green arrow represents the modeling process performed on the data. Biological function studies how serum molecules affect the body's cells, tissues and organs across the board as they flow with the blood, including differences in proteins between groups, as well as their distribution locations and cell types, pathways, functions, phase separation capabilities, pseudo-timing analysis, and upstream regulators and potential drugs. Data modeling constructs three models with the same process. The construction process

is as follows. We first use feature engineering to extract important features, then use different machine learning model to training data, finally evaluate the classification efficiency of the model.

Figure 2

Model 1(CT, MCI, AD) feature engineering and model evaluation. a) Heatmap of all features in the model. b) Expression profile of the optimal features in the dataset. c) Importance of the optimal features filtered after feature engineering. d) Accuracy of different classifiers on different features. e) Accuracy and loss

curves during training of the ANN model. f) AUC of the ANN model when using different features. g-i) Confusion matrix corresponding to all features, optimal features, and random features in the ANN model. j-l) TSNE consisting of all features, optimal features, and random features

Figure 3

Model 2 (CT, MCI, PD) feature engineering and model evaluation. a) Heatmap of all features in the model. b) Expression profile of the optimal features in the dataset. c) Importance of the optimal features filtered

after feature engineering. d) Accuracy of different classifiers on different features. e) Accuracy and loss curves during training of the ANN model. f) AUC of the ANN model when using different features. g-i) Confusion matrix corresponding to all features, optimal features, and random features in the ANN model. j-l) TSNE consisting of all features, optimal features, and random features

Figure 4

Model 3 (CT, MCI, AD, PD) feature engineering and model evaluation. a) Heatmap of all features in the model. b) Expression profile of the optimal features in the dataset. c) Importance of the optimal features filtered after feature engineering. d) Accuracy of different classifiers on different features. e) Accuracy and loss curves during training of the ANN model. f) AUC of the ANN model when using different features. g-i) Confusion matrix corresponding to all features, optimal features, and random features in the ANN model. j-l) TSNE consisting of all features, optimal features, and random features

Figure 5

IPA analysis under detailed classification. a) Expression, mean Z-score and the number of up-regulated and down-regulated of canonical pathways. b) Expression, mean Z-score and the number of up-regulated and down-regulated of diseases and bio function c) IPA analysis of Specific canonical pathways, diseases and biofunctions of disease compared to CT group, respectively

Figure 6

Analysis of upstream regulator and potential drugs of differentially expressed proteins. a) Venn diagram of upstream regulators with drug and DEPs with drug. b) Venn diagram of drugs corresponding to DEPs and drugs corresponding to upstream regulators. c) Expression profiles of potential therapeutic targets in each group and mean Z-score. d) Expression profiles of 24 differential upstream regulators

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFig1.pdf](#)
- [SupplementaryFig2.pdf](#)
- [SupplementaryFig3.pdf](#)
- [SupplementaryFig4.pdf](#)
- [SupplementaryTable1.csv](#)