RESEARCH

Learning Cognitive-Test-Based Interpretable Rules for Prediction and Early Diagnosis of Dementia using Neural Networks

Zhuo Wang[1], Jie Wang[2], Ning Liu[1], Caiyan Liu[2], Xiuxing Li[1], Liling Dong[2], Rui Zhang[1], Chenhui Mao[2], Zhichao Duan[1], Wei Zhang[3], Jing Gao[2*], Jianyong Wang[1*] and for the Alzheimer's Disease Neuroimaging Initiative (ADNI) [†]

Abstract

Background: Accurate, cheap, and easy to promote methods for dementia prediction and early diagnosis are urgently needed in low- and middle-income countries. Integrating various cognitive tests using machine learning provides promising solutions. However, most effective machine learning models are black-box models that are hard to understand for doctors and could hide potential biases and risks.

Objective: To apply cognitive-test-based machine learning models in practical dementia prediction and diagnosis by ensuring both interpretability and accuracy.

Methods: We design a framework adopting Rule-based Representation Learner (RRL) to build interpretable diagnostic rules based on the cognitive tests selected by doctors. According to the visualization and test results, doctors can easily select the final rules after analysis and trade-off. Our framework is verified on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset (n=606) and Peking Union Medical College Hospital (PUMCH) dataset (n=375).

Results: The predictive or diagnostic rules learned by RRL offer a better trade-off between accuracy and model interpretability than other representative machine learning models. For mild cognitive impairment (MCI) conversion prediction, the cognitive-test-based rules achieve an average area under the curve (AUC) of 0.904 on ADNI. For dementia diagnosis on subjects with a normal Mini-Mental State Exam (MMSE) score, the learned rules achieve an AUC of 0.863 on PUMCH. The visualization analyses also verify the good interpretability of the learned rules.

Conclusion: With the help of doctors and RRL, we can obtain predictive and diagnostic rules for dementia with high accuracy and good interpretability even if only cognitive tests are used.

Keywords: Machine Learning; Interpretability; Dementia; Neuropsychological Tests; Deep Learning

INTRODUCTION

Nowadays, dementia is one of the major causes of disability and dependency among older people. Around 50 million people have dementia worldwide, and nearly 10 million new cases are reported every year [1]. It is estimated that nearly 60% of people with dementia are living in low- and middle-income countries (LMIC) [2,3]. The diagnostic coverage in LMIC, although few available estimates, is much lower than high-income countries (HIC) and unlikely to exceed 5-10% in most settings [4–6]. Low diagnostic coverage, especially low early diagnosis coverage, causes barely any early intervention for people with mild cognitive impairment (MCI) or dementia in LMIC. Most patients are diag-

[*]Correspondence: gj107@163.com; jianyong@tsinghua.edu.cn

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, P.R. China

[2]Department of Neurology, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical Science/Peking Union Medical College, Beijing, P.R. China

Full list of author information is available at the end of the article

[†]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

nosed after having severe symptoms, i.e., in the later stage of dementia, when the best time for the interventions has already passed [7,8]. As a result, patients suffer from more harm, risk and cost, carers of people with dementia experience high strain, and society faces a heavy economic burden [1]. Hence, it is a priority for most LMIC to increase the coverage of timely dementia diagnoses [3].

Promoting cognitive tests is an effective way to increase the diagnostic coverage in LMIC. As widely used and verified diagnostic methods for dementia, cognitive tests take advantage of being cheap, non-invasive, time-saving, and easy to promote [9–11]. Take Mini-Mental State Exam (MMSE) [12] as an example. It is a 30-point questionnaire including tests of orientation, attention, memory, language, and visual-spatial skills. During the MMSE, only simple questions are asked, and no additional equipment or specialist is required.

However, cognitive tests also have their own drawbacks. First, compared with other dementia diagnosis or prediction methods, the sensitivity of one single test could be poor for the early diagnosis [13–15]. Second, the effect of one single test is limited since different cognitive tests may focus on different functions. For example, compared with MMSE, the Functional Assessment Questionnaire (FAQ) [16] pays more attention to daily living functions. Third, the combination of different cognitive tests is difficult since: (i) The number of combinations increases exponentially along with the number of candidate cognitive tests, which will lead to a combinatorial explosion. (ii) The scores of different cognitive tests cannot be directly added. (iii) The cut-off value and threshold are hard to determine for the combined tests.

Other methods, e.g., MRI image, PET image, and genotyping data, are also inappropriate for LMIC since their data acquisition is expensive, time-consuming, and requires well-trained doctors or specialist equipment [17–19]. Even with a large amount of investment, these methods may still struggle to meet the increasing demand for dementia diagnosis.

Therefore, we urgently need to find a new effective method for dementia prediction and early diagnosis that can be easily promoted in both LMIC and HIC.

Machine learning, especially deep learning, has achieved impressive results in many medical tasks [20, 21]. It is promising to integrate various cognitive tests using machine learning models to achieve higher accuracy while keeping the advantages of cognitive tests. However, most of the effective machine learning models, e.g., deep learning and ensemble models, are black-box models [22–24]. Since we can hardly understand their decision mechanism, potential biases and risks could hide in these models, which is unacceptable for the clinical diagnosis. Additionally, these black-box models make the diagnosis separated from the doctors, ignoring the important role of doctors in the diagnosis. Another issue is the black-box models need lots of computation resources when diagnosing [25]. Therefore, the deployments of black-box models could be difficult and costly for the hospitals in LMIC.

The usage of interpretable machine learning models, e.g., decision trees and linear models, however, is also limited because of their low classification performance [23]. These models sacrifice their model capacity to obtain good interpretability. Hence, it is hard for these interpretable models to deal with complex problems like dementia prediction and diagnosis. Recent studies try to tackle the drawbacks of conventional interpretable models by interpreting black-box models using post-hoc methods, e.g., LIME [26,27] and SHAP [28]. However, the consistency between their interpretations and the original models is not guaranteed [23, 29]. Therefore, these post-hoc methods could be misleading in some cases, which is unacceptable for clinical applications.

In this study, we design a new framework that learns interpretable rules for the prediction and early diagnosis of dementia using a tailored neural network called Rule-based Representation Learner (RRL) [30]. The overall framework is shown in Figure 1a. In our framework, the whole process takes full advantage of the cooperation of the doctor and neural network, and the doctor and neural network both play an important role. Using their medical professional knowledge and clinical experience, doctors aim to select appropriate cognitive tests as the candidate features for the neural network training and do the trade-off to select the final predictive or diagnostic rule set among all the candidate rule sets generated by RRL. RRL aims to learn from the training data with selected features and generate candidate diagnostic rule sets with different classification performance and model complexities. We also propose a new visualization method for the rules learned by RRL to make this process more intuitive and easier for doctors. Experiments on two datasets verify that we can obtain highly accurate and interpretable predictive (diagnostic) rules for dementia even if only the results of several cognitive tests are used in our framework.

MATERIALS AND METHODS

Data origin and acquisition

Two datasets are used in this study. The first dataset is obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [31] database (adni.loni.usc.edu), including ADNI 1, 2/GO, and 3. The ADNI is a longi-

Table 1 Demographic characteristics of the ADNI and PUMCH datasets.

	ADNI			PUMCH		
	MCI-NC n=253	MCI-C n=353	P value	Non-dementia n=241	Dementia n=134	P value
Age (years)	70.83 (7.26)	73.89 (7.11)	<0.001	63.90 (11.70)	68.41 (10.44)	<0.001
Gender (female)	102 (40.32%)	140 (39.66%)	0.937	142 (58.92%)	72 (53.73%)	0.388
Education (years)	16.11 (2.78)	15.91 (2.75)	0.384	12.47 (3.90)	11.96 (3.92)	0.237
MMSE	28.24 (1.61)	27.07 (1.76)	<0.001	28.16 (1.25)	27.15 (1.17)	<0.001
MoCA	24.35 (2.76)	21.73 (2.79)	<0.001	25.36 (2.75)	22.54 (2.82)	<0.001

Data are shown as mean (s.d.) or n (%). Abbreviations: MCI = mild cognitive impairment; MCI-C = MCI converter; MCI-NC = MCI non-converter; MMSE = Mini-Mental State Examination; MoCA = Montreal Cognitive Assessment.

tudinal multicenter study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD). For the ADNI dataset, we mainly aim at identifying patients with mild cognitive impairment (MCI) who progress to AD, i.e., MCI converter (MCI-C), and patients with MCI who do not progress to AD, i.e., MCI non-converter (MCI-NC). Subjects are included consecutively. After data preprocessing and removing invalid records, there are 606 participants left, with 253 (41.7%) MCI-NC and 353 (58.3%) MCI-C. We ensure that all MCI-NC patients did not convert to AD after at least 48 months of follow-up. Each patient has 51 features, including the demographic information, the results of selected cognitive tests (e.g., MMSE score), and other biomarkers (e.g., APOE4, AV45, and pTau).

The second dataset was collected by the Peking Union Medical College Hospital (PUMCH) from May 2009 to April 2021 [32]. Only subjects with a normal MMSE score ($\geq$ 26) and the ability to complete all neuropsychological tests are included consecutively. The clinical history, neuropsychological tests, laboratory tests, and head CT or MRI are leveraged to make diagnoses. A total of 375 subjects are included, among which 241 (64.3%) subjects are diagnosed with cognitively normal (CN) or MCI, i.e., non-dementia, and 134 (35.7%) are diagnosed with dementia. After data preprocessing, the demographic information and the results of selected cognitive tests in each record are converted into 64 features.

The demographic characteristics of the ADNI and PUMCH datasets are shown in Table 1. See Table 3 and 4 in the Appendix for all the available features in ADNI and PUMCH.

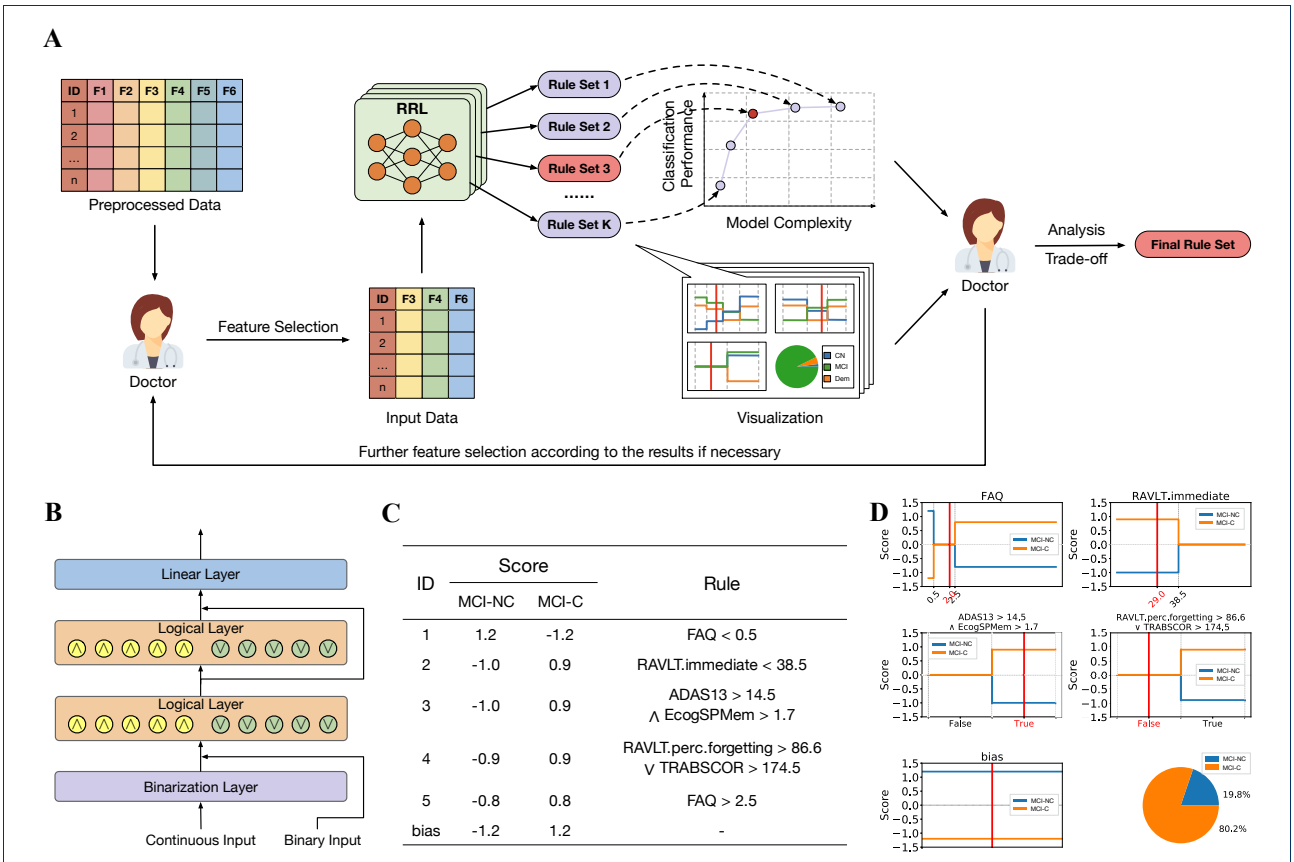
Overall framework to build interpretable rules

The overall framework to learn interpretable diagnostic rules using neural networks is shown in Figure 1a. First, after data preprocessing, doctors need to do the feature selection to ensure all the features used for the following rule construction are easily available and their corresponding cognitive tests are not time-consuming. After the feature selection, a novel neural network, called Rule-based Representation Learner (RRL) [30], is adopted to learn rules from the data. RRL can be trained like ordinary neural networks but with a different training method. After training, we can easily convert RRL into an equivalent rule set due to its tailored model structure and components. By adjusting the network structure and hyperparameters of RRL, rule sets with different model complexities and classification performances are generated. Testing these rule sets on the test set, we can obtain the relationship between model complexity and classification performance. Combined with the visualization of rule sets, doctors can select the rule set with the best trade-off as the final rule set. If all the generated rule sets do not satisfy the requirement, doctors can reselect the features according to the existing results and analyses and then retrain the RRL.

Feature selection

The feature selection step in our framework is that doctors select features according to their costs, time consumption, equipment requirement, doctor training cost to satisfy the need for different scenarios. For example, a hospital in LMIC can hardly obtain and leverage the features like AV45 or PDG due to its limited resources. Therefore, these hard-to-obtain features should be removed to match the situation of the hospital. It should be noted that feature selection by doctors is different from feature selection by machine learning models since machine learning models mainly select features according to their effects on the classification performance.

We divide all features into four types, i.e., demographics, easy-to-obtain cognitive tests, hard-to-obtain cognitive tests, and other biomarkers. ADNI-E consists of demographics and only easy-to-obtain cognitive tests, while ADNI-H consists of demographics



ID	Score		Rule
	MCI-NC	MCI-C	
1	1.2	-1.2	FAQ < 0.5
2	-1.0	0.9	RAVLT.immediate < 38.5
3	-1.0	0.9	ADAS13 > 14.5 ∧ EcogSPMem > 1.7
4	-0.9	0.9	RAVLT.perc.forgetting > 86.6 ∨ TRABSCOR > 174.5
5	-0.8	0.8	FAQ > 2.5
bias	-1.2	1.2	-

Figure 1 The overall framework and RRL examples. **a**, The overall framework to build interpretable diagnostic rules with RRL and the help of doctors. **b**, A Rule-based Representation Learner example. In RRL, one logical layer consists of one conjunction layer and one disjunction layer. One node in the conjunction layer corresponding to one conjunction operation, i.e., $\wedge$, while one node in the disjunction layer corresponds to one disjunction operation, i.e., $\vee$. **c**, One MCI-C predictive rule set learned by RRL. The second and third columns list the scores of rules for the corresponding classes. The last column lists the rules. **d**, Visualization of the rule set shown in c. The vertical red lines indicate feature values of one example subject. The pie chart shows the prediction result, i.e., the probability of each class. Abbreviations: MCI = Mild Cognitive Impairment; MCI-C = MCI converter; MCI-NC = MCI nonconverter; RRL = Rule-based Representation Learner; FAQ = Functional Assessment Questionnaire; RAVLT.immediate = RAVLT Immediate (sum of 5 trials); EcogSPMem = SP ECog - Mem; ADAS13 = Alzheimer's Disease Assessment Scale 13; RAVLT.perc.forgetting = RAVLT Percent Forgetting; TRABSCOR = Trails B.

and all the cognitive tests. ADNI-A consists of all the available features in the ADNI dataset. All the features in the PUMCH dataset are from demographics or easy-to-obtain cognitive tests. See Table 3 and 4 in the Appendix for the details of feature selection.

Rule-based representation learner

Rule-based Representation Learner (RRL) [30] is a neural network that automatically learns interpretable non-fuzzy rules for data representation and classification. RRL is designed for scenarios demanding both good classification performance and model interpretability, and we can easily adjust it to obtain a trade-off between classification performance and model complexity for different requirements. Figure 1b illustrates the structure of an example RRL.

RRL consists of three different types of layers, i.e., binarization layer, logical layer, and linear layer. Layer in RRL contains a specific number of nodes, and there are trainable edges connecting these nodes with nodes in the previous layer. The binarization layer binarizes each continuous feature into several binary features. Each node in the binarization layer corresponds to a cut-off value of a continuous feature. For the j-th continuous feature $\mathbf{c}_j$, the binarization layer randomly generates k lower bound values $(\mathcal{T}_{j,1}, \ldots, \mathcal{T}_{j,k})$ and k upper bound values $(\mathcal{H}_{j,1}, \ldots, \mathcal{H}_{j,k})$, then it will check if $\mathbf{c}_j$ satisfies the bounds and get the following binary vector as output:

$$Q_j = [\mathbf{1}_{\mathbf{c}_j > \mathcal{T}_{j,1}}, \ldots, \mathbf{1}_{\mathbf{c}_j > \mathcal{T}_{j,k}}, \mathbf{1}_{\mathbf{c}_j \leq \mathcal{H}_{j,1}}, \ldots, \mathbf{1}_{\mathbf{c}_j \leq \mathcal{H}_{j,k}}] \quad (1)$$

Where $\mathbf{1}_{(\cdot)}$ is an indicator function. The logical layers automatically learn data representations using logical rules. To build rules in more complex forms, we can stack several logical layers together. One logical layer consists of one conjunction layer and one disjunction layer. One node in the conjunction layer corresponds to one conjunction operation, while one node in the disjunction layer corresponds to one disjunction operation. The edges indicate which variables are involved in the operation. Specifically, let $\mathbf{r}_i^{(l)}$ denote the i-th conjunction node in the l-th logical layer and $\mathbf{s}_i^{(l)}$ denote the i-th disjunction node, then the two types of nodes are defined as follows:

$$\mathbf{r}_i^{(l)} = \bigwedge_{W_{i,j}^{(l,0)}=1} \mathbf{u}_j^{(l-1)}, \quad \mathbf{s}_i^{(l)} = \bigvee_{W_{i,j}^{(l,1)}=1} \mathbf{u}_j^{(l-1)}, \quad (2)$$

where $W_{i,j}^{(l,0)}$ and $W_{i,j}^{(l,1)}$ are the adjacency matrices and $\mathbf{u}^{(l-1)}$ is the output of the previous layer. The binarization layer and all the logical layers actually form a rule learner (representation learner). One logical rule, i.e., one node in the last logical layer, is formulated by the original features. The linear layer can be considered as a linear classifier based on the learned rules. In other words, the inputs of the linear classifier are the values of learned rules (0 for False and 1 for True). These logical rules are easy to analyze and understand, benefiting from their no-fuzzy form.

To effectively learn the discrete rules in an end-to-end way, RRL adopts logical activation functions and a novel gradient-based discrete model training method called Gradient Grafting. The logical activation functions use multiplications of real-value variables to simulate the logical operations. With logical activation functions, we can obtain a continuous version of RRL, which is differentiable. Using Gradient Grafting, we can build a complete backward path from the loss function to the parameters of the discrete RRL by combining the gradients from both discrete and continuous RRL. Hence, the whole RRL is differentiable, and we can optimize discrete RRL with gradient descent. For more details, please refer to [30].

Model interpretation

As we mentioned before, RRL can be considered as a rule learner and a linear classifier. Therefore, we can interpret and understand RRL in two steps. First, the weights of the linear classifier tell us how each rule contributes to the final decision on each class. For each rule, we call its corresponding weights its scores. In addition, by sorting the scores, we can get important rules that we should pay more attention to. Second, in RRL, each rule is formulated by several original features. We can easily understand one rule by analyzing the original features and logical operators in it.

For example, one example predictive rule set learned by RRL is shown in Figure 1c. The second and third columns list the scores of the rules. The last column lists the rules. All rules are sorted by the maximum absolute value of their corresponding scores. When diagnosing with these rules, we find all the rules with True value (satisfied) and sum up their scores in each score column to get the score for each class. Let denote the score of the i-th class by $\mathbf{z}_i$. The predicted probability of the i-th class is:

$$softmax(\mathbf{z})_i = \frac{e^{\mathbf{z}_i}}{\sum_{j=1}^{\mathcal{C}} e^{\mathbf{z}_j}}, \quad (3)$$

where $\mathcal{C}$ is the number of classes. For one class, a high score indicates a high probability. Take the rule "ADAS13 $>$ 14.5 $\wedge$ EcogSPMem $>$ 1.7" as an example. Since the scores of class MCI-NC and MCI-C are -1.0 and 0.9, respectively, when the value of this rule is True, the score of being MCI-C will increase while the score of being MCI-NC will decrease. Furthermore, this rule contains the conjunction of ADAS13 and EcogSPMem, which tells us that the combination of these two features could be useful for the diagnosis.

The model complexity of RRL has a negative effect on the model interpretability. To limit the model complexity and obtain a trade-off between classification performance and model interpretability, we can reduce the number of logical layers in RRL and reduce the number of nodes in each logical layer. Moreover, to search for an RRL with shorter rules during the training, we can increase the coefficient of the L1/L2 regularization term [33] in the loss function. The dead nodes detection, redundant rules elimination [34], and skip connections are also helpful for a simpler RRL.

Rule visualization

Although the rule set learned by RRL is much more interpretable than ordinary neural networks, understanding rules and their real value scores may take too much time for doctors. Additionally, the cut-off values of one feature appearing in different rules also make it harder to understand. Inspired by the nomogram [35], we visualize the rule set learned by RRL in a novel form to make it more intuitive and convenient to use. Figure 1d is the visualization of the rule set shown in Figure 1c.

It takes three steps to generate the visualization results. First, we merge rules of length one that have the same feature (sum up the scores of all satisfied rules according to the feature value). Then, for each feature,

we draw the scores of all the classes (e.g., MCI-NC and MCI-C) in the feature value range. Scores between different cut-off values are different. For example, in Figure 1c, the rule "FAQ < 0.5" and "FAQ > 2.5" can be merged and visualized together, i.e., the first subfigure in Figure 1d. In this subfigure, we can directly see how the value of the cognitive test FAQ affects the score. If FAQ < 0.5, i.e., Rule #1 in Figure 1c is satisfied, then the total score of being MCI-NC increased by 1.2, and the total score of being MCI-C decreased by 1.2. If $0.5 \leq$ FAQ ≤ 2.5, i.e., no rule is satisfied, then the total score will not change. Similarly, if FAQ > 2.5, we add the scores of Rule #5.

Second, for all the rules whose length is greater than one, we consider each of them a new feature and draw the scores corresponding to the True and False states. For instance, the third subfigure in Figure 1d is the visualization of Rule #3. When the value of Rule #3 is True, we will add the total score of being MCI-NC and MCI-C by -1.0 and 0.9, respectively. Otherwise, add the total scores by zero.

Finally, we draw the biases of the linear layer, e.g., the penultimate subfigure in Figure 1d. When diagnosing one subject, we also draw red vertical lines to represent the feature values of this subject. The ordinates of the intersection points of the red line and other horizontal lines are the scores. The pie chart shows the final result. For the example subject in Figure 1d, the total score of being MCI-C is 0+0.9+0.9+0-1.2=0.6, and the total score of being MCI-NC is 0-1.0-1.0+0+1.2=-0.8. According to Softmax, the probabilities of MCI-C and MCI-NC are 80.2% and 19.8%, respectively.

With visualization, doctors can intuitively understand how the scores change with the value of one feature and how the combination of features affects the diagnosis.

Evaluation

We adopt the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) to evaluate the classification performance. We use the total number of edges as the metric of model complexity for rule-based models. Compared to the total length of all rules, the total number of edges takes the reused structures in rule-based models into consideration. Hence, it can evaluate the model complexity more accurately. 10-fold cross-validation is adopted to split the datasets and reduce the biases of the evaluations. We split the dataset according to the roster ID. Therefore, no patient is included in both the training and test sets, and the risk of data leakage is avoided [36]. We implement RRL with Python and PyTorch [37]. All experiments are conducted on a Linux server with an Intel Xeon E5 v4 CPU at 2.10GHz and one GeForce RTX 2080 Ti GPU.

Comparison with other models

The performance of RRL is compared with six representative machine learning models, including interpretable models and complex models (black-box models) that are hard to interpret and understand. CART [38] is a rule-based model that builds a decision tree. Logistic Regression (LR) [39] is a linear model. These two models and linear Support Vector Machines (SVM) [40] are considered interpretable models. Piecewise Linear Neural Network (PLNN) [41], nonlinear SVM (using RBF or Poly kernels), Random Forest (RF) [42] and eXtreme Gradient Boosting (XGBoost) [43] are considered complex models. For SVM, the regularization parameter C is in $\{2^{-4}, 2^{-2}, 1, 2^2, 2^4, 2^6\}$, and the tolerance for stopping criteria is set to 0.001. PLNN is a Multilayer Perceptron (MLP) that adopts piecewise linear activation functions, e.g., ReLU [44]. RF and XGBoost are ensemble models consisting of hundreds of decisions trees. Nonlinear SVM, PLNN, RF, and XGBoost are hard to interpret due to their complex inner structures.

RESULTS

Performance of MCI-C prediction

To show the effectiveness of RRL on the MCI-C prediction task, we train RRL on the ADNI dataset and compare it with three interpretable models (i.e., CART, LR, and linear SVM) and four black-box models (i.e., nonlinear SVM, PLNN, RF, and XGBoost). The 10-fold cross validated AUCs of the MCI-C prediction of these models trained on the ADNI dataset are shown in Table 2. ADNI-E, ADNI-H, and ADNI-A represent different feature selections, as we described before.

First, we can observe that RRL outperforms all the baseline models, including the black-box models, regardless of the feature selection we chose. Second, the average AUC of RRL trained on ADNI-E is 0.904, which means with only demographics and easily available cognitive tests, we can obtain an accurate rule set for the MCI-C prediction task. We can also see that, for the AUCs of highly accurate models, e.g., RRL and RF, the differences between ADNI-E and ADNI-H are quite small. It indicates that, for the MCI-C prediction task, the easily available cognitive tests already have the same useful information as those cognitive tests that are hard to obtain. However, compared with ADNI-H, the hidden useful information in ADNI-E could be more difficult to find and may cause the trained model to be much more complex. Third, the average AUC of RRL trained on ADNI-A is 0.933, which is better than the results on ADNI-E and ADNI-H. This observation indicates there is still some useful information for MCI-C prediction that cognitive tests cannot capture. Only depending on other biomarkers,

Table 2 10-fold cross validated AUC of RRL and other baseline models on the ADNI and PUMCH data sets.

	RRL(Ours)	CART	LR	SVM(Linear)	SVM(RBF)	SVM(Poly)	PLNN	RF	XGBoost
ADNI-E	**0.904±0.05**	0.836±0.07	0.884±0.06	0.887±0.05	0.881±0.06	0.873±0.05	0.888±0.06	0.898±0.05	0.886±0.06
ADNI-H	**0.904±0.04**	0.835±0.06	0.894±0.04	0.894±0.03	0.896±0.04	0.890±0.04	0.902±0.04	0.902±0.04	0.903±0.05
ADNI-A	**0.933±0.03**	0.855±0.05	0.911±0.03	0.912±0.03	0.921±0.04	0.912±0.03	0.929±0.03	0.927±0.03	0.920±0.03
PUMCH	**0.863±0.08**	0.752±0.09	0.834±0.04	0.841±0.03	0.850±0.05	0.850±0.06	0.841±0.05	0.861±0.08	0.836±0.07

Data are shown as mean ± std. RRL = Rule-based Representation Learner; CART = Classification and Regression Trees; LR = Logistic Regression; SVM = Support Vector Machine; RBF = Radial Basis Function; PLNN = Piecewise Linear Neural Network; RF = Random Forest; XGBoost = eXtreme Gradient Boosting (gradient boosted decision tree).

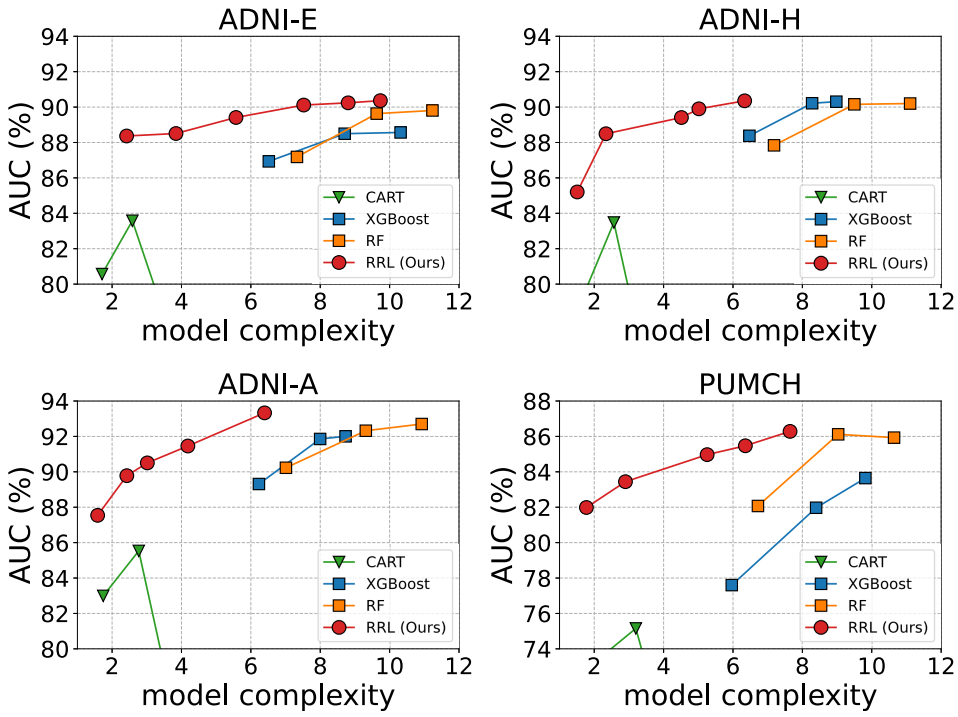


Figure 2 Scatter plots of AUC against model complexity for RRL and baseline models on ADNI-E, ADNI-H, ADNI-A, and PUMCH. AUC is used to evaluate the classification performance while log(#edges) is used to evaluate the model complexity. Higher AUC and lower model complexity are better. Each point is the average result of 10-fold cross-validation of one model with a specific hyperparameter setting. Abbreviations: ADNI = Alzheimer's Disease Neuroimaging Initiative; AUC = Area Under the Curve; PUMCH = Peking Union Medical College Hospital; RF = Random Forest; RRL = Rule-based Representation Learner; XGBoost = eXtreme Gradient Boosting.

e.g., data from MRI and PET, can we leverage more useful information and further improve the results.

Model complexity of MCI-C predictive rules

One crucial factor affecting interpretability is model complexity. Even models that are commonly considered as interpretable, e.g., decision trees, could become hard to understand if their model complexities are high. For example, we can hardly understand a decision tree with more than one hundred layers. Moreover, low model complexity without acceptable accuracy is meaningless. Therefore, interpretable models seek to keep low model complexity while ensuring high accuracy, and what we really care about is the relationship between accuracy (prediction performance) and model complexity of the models. In Figure 2, we draw scatter plots of average AUC (on the test set) against log(#edges) for rule-based models or ensemble methods trained on ADNI-E, ADNI-H, and ADNI-A. The logarithm of the number of edges, i.e., log(#edges), is used to evaluate the model complexity.

We can observe that, compared with other baseline models, RRL can obtain a better trade-off between prediction performance and model complexity regardless of the feature selection we chose. In other words, if the AUC of RRL is close to one baseline model, then

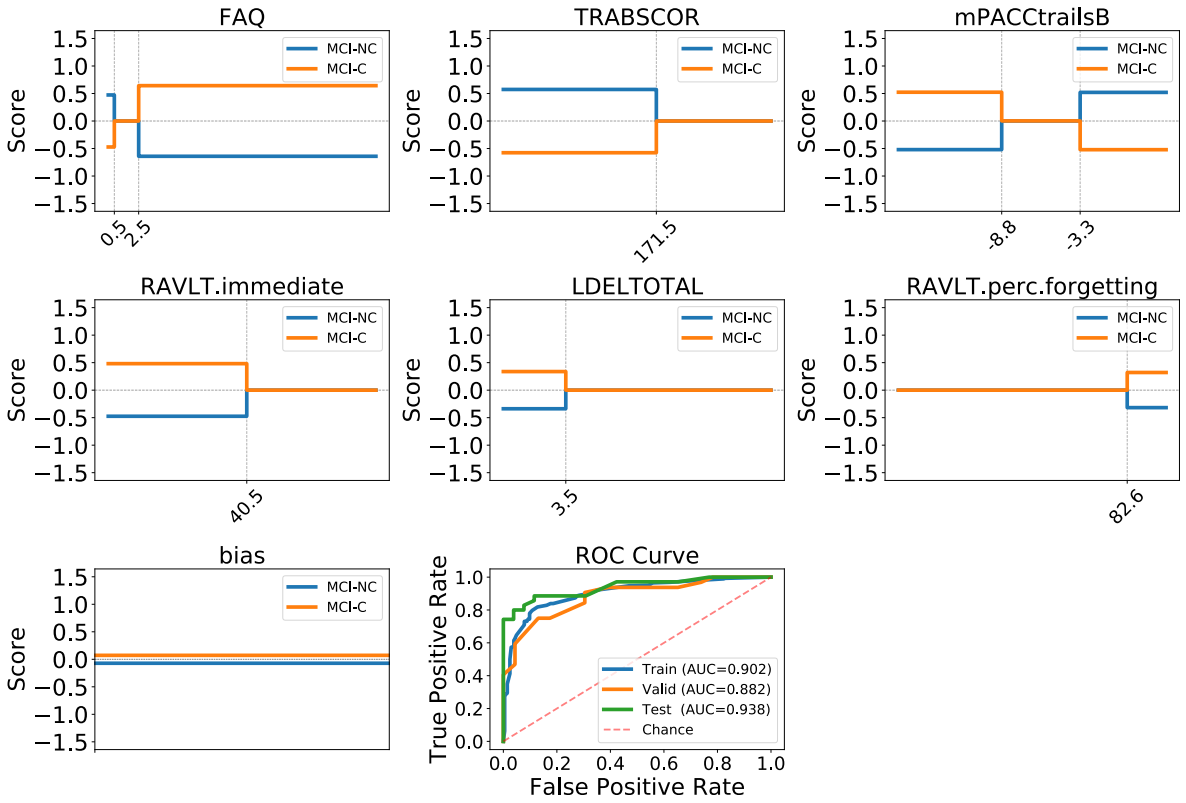


Figure 3 Visualization of the entire predictive rule set for ADNI-E, i.e., RS-E. Only easy-to-obtain cognitive tests are used to build the rule set. The last and the penultimate subfigures show the ROC curve and the bias, respectively. Abbreviations: FAQ = Functional Assessment Questionnaire; TRABSCOR = Trail Making Test-B. mPACCtrailsB = ADNI modified Preclinical Alzheimer's Cognitive Composite (PACC) with Trails B; RAVLT.immediate = RAVLT Immediate recall (sum of 5 trials); LDELTOTAL = Logical Memory (Delayed Recall); RAVLT.perc.forgetting = RAVLT Percent Forgetting.

the model complexity of RRL will be lower. If RRL has a close model complexity with one baseline model, then the AUC of RRL will be higher. For example, to get an AUC close to 0.895 in ADNI-E, the log(#edges) of RRL is 5.6 while the log(#edges) of RF is 9.6, which indicates that the model complexity of RF is about fifty times that of RRL (i.e., $e^{9.6-5.6} \approx 55$).

Another observation is that RRL trained on ADNI-E needs more model complexity than RRL trained on ADNI-H to get a close AUC. The cognitive tests in ADNI-H that are hard to obtain can simplify the learned rules for MCI-C prediction. The main reason is that one hard cognitive test will collect more information than one easily available cognitive test. Therefore, RRL trained on ADNI-H could use fewer or shorter rules than RRL trained on ADNI-E to build the model with comparable performance. We can also see that with the help of other biomarkers, RRL trained on ADNI-A has better prediction performance than RRL trained on ADNI-E or ADNI-H when their model complexities are close.

Visualization of MCI-C predictive rules

After the procedure shown in Figure 1a, we obtain the final predictive rule sets for ADNI-E, ADNI-H, and ADNI-A. Figure 3, Figure 4, and Figure 5 show the visualization of the entire predictive rule set for ADNI-E, ADNI-H, and ADNI-A, respectively. Let us call these predictive rule sets RS-E, RS-H, and RS-A, respectively. All these predictive rule sets are selected by the doctors according to their prediction performance, rule complexity, costs, time consumption, equipment requirement, doctor training cost, and consistency with clinical experience. The prediction performance, i.e., the ROC curve and AUC, of each rule set on its corresponding fold is shown in the last subfigure of the corresponding figure, and the bias (of the linear layer) in each rule set is also shown in the penultimate subfigure. The remaining subfigures are sorted by their span of scores.

Since the rule set shown in Figure 3, i.e., RS-E, is obtained from RRL trained on ADNI-E, it only uses demographics and easily available cognitive tests for MCI-C prediction. We can see that the

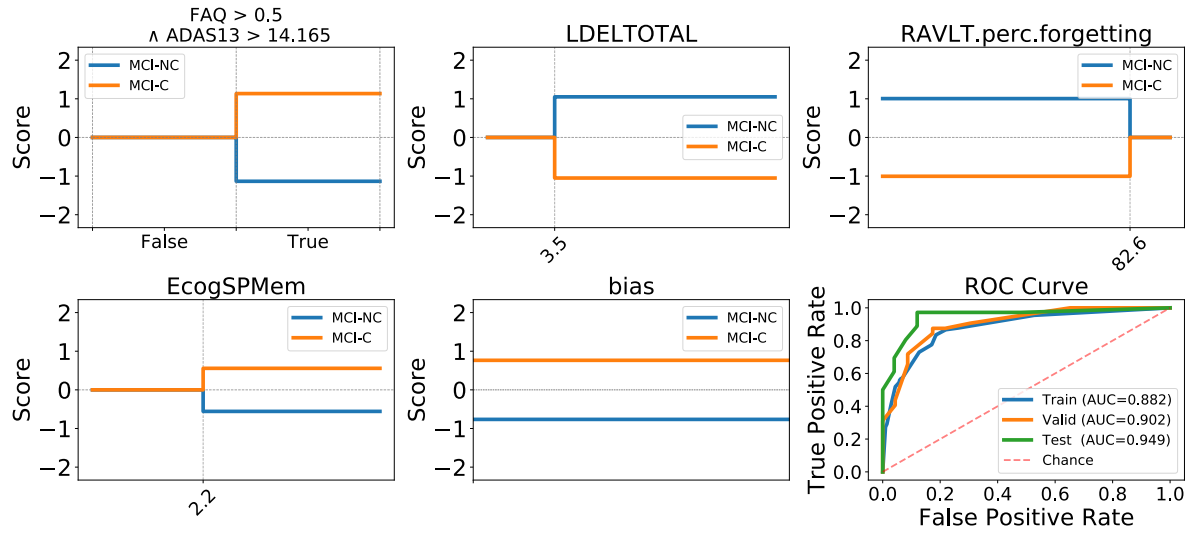


Figure 4 Visualization of the entire predictive rule set for ADNI-H, i.e., RS-H. Only cognitive tests are used to build the rule set. The last and the penultimate subfigures show the ROC curve and the bias, respectively. Abbreviations: FAQ = Functional Assessment Questionnaire; ADAS13 = Alzheimer's Disease Assessment Scale 13 (including Delayed Word Recall and Number Cancellation); LDELTOTAL = Logical Memory (Delayed Recall); RAVLT.perc.forgetting = RAVLT Percent Forgetting; EcogSPMem = Everyday Cognition filled by the patient's study partner (Memory domain with eight items).

whole rule set is very simple, and only six features are needed, including FAQ, TRABSCOR, mPACCtrailsB, RAVLT.immediate, RAVLT.perc.forgetting and LDELTOTAL. These six features are generated from five cognitive tests (both TRABSCOR and mPACCtrailsB are from the same cognitive tests) and evaluate MCI patients from different perspectives. FAQ evaluates the daily living functions. TRABSCOR and mPACCtrailsB evaluate executive functions and attention. RAVLT.immediate, RAVLT.perc.forgetting, and LDELTOTAL evaluate immediate, delayed, and episodic memory, respectively. Although RS-E is simple, it gets a good prediction performance, i.e., the AUCs on the training set, validation set, and test set are 0.902, 0.882, and 0.938, respectively. Through the visualization, doctors can easily understand how the values of cognitive tests affect the prediction result. For example, FAQ has two cut-off values, i.e., 0.5 and 2.5, affecting the outcome. When FAQ < 0.5, the score of being MCI-NC is higher, while when FAQ > 2.5, the score of being MCI-C is higher. Therefore, doctors only need to get the values of these six features and compare them with their corresponding cut-off values to obtain the scores for MCI-C prediction.

Figure 4 shows the rule set for ADNI-H, i.e., RS-H. Compared with the rule set for ADNI-E, i.e., RS-E, shown in Figure 3, RS-H uses additional features, i.e., features from hard-to-obtain cognitive tests, to build the rules. We can see that RS-H is even simpler than RS-E, but its prediction performance is good. The AUCs of RS-H on the training set, validation set, and test set are 0.882, 0.902, and 0.949, respectively. Similar to RS-E, RS-H still keeps FAQ, LDELTOTAL, and RAVLT.perc.forgetting. The difference is that RS-H replaces TRABSCOR, mPACCtrailsB, and RAVLT.immediate with ADAS13 and EcogSPMem. ADAS13 evaluates the global cognition of the patient while EcogSPMem evaluates their memory. Another observation is that RS-H uses a combination of two original features to build a new feature, i.e., "FAQ $>$ 0.5 $\wedge$ ADAS13 $>$ 14.165". For this new feature, only when "FAQ > 0.5" and "ADAS13 > 14.165" are both satisfied, the score of being MCI-C will be higher. Otherwise, RS-H uses the remaining features to predict.

Figure 5 shows the rule set for ADNI-A, i.e., RS-A. Since RS-A uses all the available features in ADNI, the procedure of generating RS-A not only shows how to select important features automatically but also shows how to combine cognitive tests with other effective biomarkers. We can observe that the prediction performance of RS-A is the best among RS-E, RS-H, and RS-A, and the rule complexity of RS-A is still very low. The AUCs of RS-A on the training set, validation set, and test set are 0.923, 0.915, and 0.955, respectively. Similar to RS-E and RS-H, RS-A also uses FAQ, mPACCtrailsB, RAVLT.immediate, and EcogSPMem. These four features are generated from cognitive tests. The difference is that AV45, FDG, and Hippocampus, these three non-cognitive test features play an important role in the MCI-C prediction. AV45 detects the

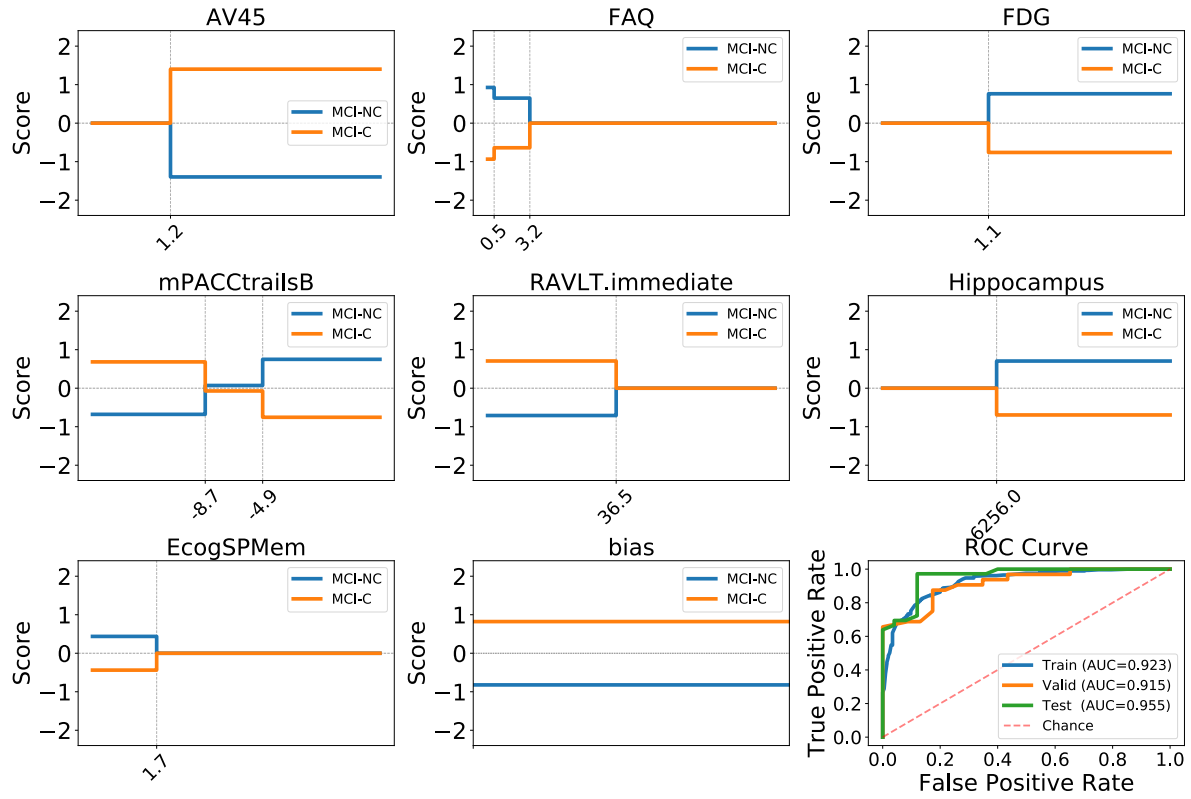


Figure 5 Visualization of the entire predictive rule set for ADNI-A, i.e., RS-A. All the available features in the ADNI dataset are used to build the rule set. The last and the penultimate subfigures show the ROC curve and the bias, respectively. Abbreviations: AV45 = AV45 ratio (cortical grey matter/whole cerebellum); FAQ = Functional Assessment Questionnaire; FDG = Average FDG-PET of angular, temporal, and posterior cingulate; mPACCtrailsB = ADNI modified Preclinical Alzheimer's Cognitive Composite (PACC) with Trails B; RAVLT.immediate = RAVLT Immediate recall (sum of 5 trials); Hippocampus = UCSF Hippocampus volume (right and left); EcogSPMem = Everyday Cognition filled by the patient's study partner (Memory domain with eight items).

brain amyloid-β (Aβ) protein deposition, whose effectiveness of early diagnosis of Alzheimer's disease has been verified by other studies [45]. FDG is the average FDG-PET of angular, temporal, and posterior cingulate. Hippocampus is essential for spatial learning. FDG and Hippocampus are related to neurodegeneration or neuronal injury. Additionally, the cut-off values of cognitive test features, e.g., mPACCtrailsB, are slightly adjusted according to other non-cognitive test features.

Performance of early dementia diagnosis

To verify the effectiveness of RRL on the early dementia diagnosis task, we train RRL on the PUMCH dataset and compare it with the same baseline models as we used in the MCI-C prediction task. The 10-fold cross validated AUCs of the dementia diagnosis of all models are shown in the last row of Table 2.

Similar to the observation we get from the results of the MCI-C prediction task on the ADNI dataset, RRL outperforms all the baseline models on the early dementia diagnosis task. Since all the subjects in PUMCH have a normal MMSE score, it is difficult to diagnose them precisely. Only RRL and RF can achieve an average AUC of 0.86, significantly outperforming all the interpretable baseline models.

Model complexity of diagnostic rules

To show the relationship between model complexity and diagnosis performance, we draw scatter plots of average AUC (on the test set) against log(#edges) for rule-based models or ensemble methods trained on PUMCH in Figure 2 (the last subfigure). The logarithm of the number of edges, i.e., log(#edges), is used to evaluate the model complexity.

We can also observe that, on PUMCH, RRL obtain a better trade-off between diagnosis performance and model complexity than other baseline models. The results in Figure 2 show that we can easily adjust the model complexity of RRL trained on PUMCH by setting the hyperparameters. Therefore, there are more

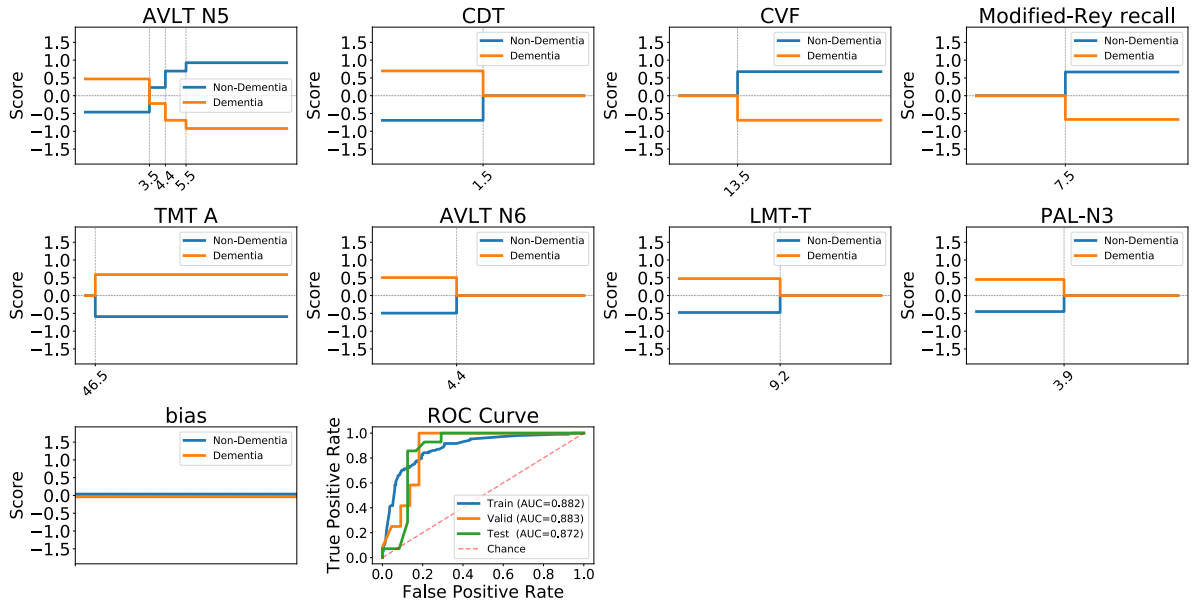


Figure 6 Visualization of the rule set for early dementia diagnosis. The entire rule set is learned from PUMCH, i.e., RS-P. All the features in the PUMCH dataset are from demographics or easy-to-obtain cognitive tests. The last and the penultimate subfigures show the ROC curve and the bias, respectively. Abbreviations: AVLT-H = the first learning trial of Auditory Verbal Learning Test–Huashan version; AVLT N5 = the fifth long delayed free recall trial of AVLT-H; CDT = Clock drawing test; CVF = Category verbal fluency; Modified-Rey recall = Modified Rey-Osterreith figure with a 10-minute free recall; TMT A = Trail making test A; AVLT N6 = the sixth delayed category cue recall trial of AVLT-H; LMT-T = the total score of the first story of logical memory test of modified Wechsler Memory Scale; PAL-N3 = the third learning trial of Paired-associate learning of the Clinical Memory Test.

choices for the doctors, and it is more likely to find suitable diagnosis rules.

Visualization of dementia diagnostic rules

In Figure 6, we show the visualization of the final dementia diagnostic rule set for PUMCH, and we call this rule set RS-P. RS-P is generated from RRL trained on the PUMCH dataset and selected by doctors according to its prediction performance, rule complexity, time consumption, equipment requirement, doctor training cost, and consistency with clinical experience.

As we mentioned before, since all the subjects in PUMCH have a normal MMSE score, it is difficult to diagnose them precisely. Therefore, to achieve an acceptable diagnosis performance, RS-P has to use eight features to build the rule set. These eight features evaluate patients from different perspectives. AVLT N5 and AVLT N6 evaluate the delayed memory. CDT evaluates visuospatial and executive functions. CVF evaluates the executive function and motor speed. Modified-Rey recall evaluates the visuospatial function and nonverbal memory. TMT A evaluates attention and executive function. LMT-T evaluates episodic memory. PAL-N3 evaluates verbal memory and executive function. As the last subfigure of Figure 6 shows, the AUCs of RS-P (on its corresponding fold) on the training set, validation set, and test set are 0.882, 0.883, and 0.872, respectively.

DISCUSSION

Our study demonstrates that, by combining doctors and tailored neural networks, i.e., RRL, we can obtain accurate and interpretable rules for the prediction and early diagnosis of dementia based on cognitive tests. For the MCI-C prediction task, our study suggests that even if only easily available cognitive tests are used, the predictive rules learned by RRL can get a good prediction performance. With the help of hard-to-obtain cognitive tests and other biomarkers, rules learned by RRL can further improve their interpretability and prediction performance. For the dementia diagnosis task, our study suggests that the integration of cognitive tests is able to diagnose subjects with normal MMSE scores (≥ 26) with acceptable performance. This also suggests that one single cognitive test is insufficient for the early dementia diagnosis, and integrating several cognitive tests could be a promising direction. In addition, our study demonstrates that RRL can achieve a better trade-off between accuracy and model interpretability than other rule-based models in the prediction or diagnosis of dementia.

Cognitive tests are widely used for the screening and diagnosis of dementia [10, 46]. They take advantage of

being cheap, non-invasive, and easy to promote. However, since dementia is a heterogeneous disease [47,48] and different cognitive tests focus on different functions [10, 49], the effect of one single test is limited. Studies have found that the sensitivity of one single cognitive test could be poor for the early diagnosis of dementia [13–15].

To overcome the shortcomings of single cognitive tests, one possible solution is finding other effective biomarkers, e.g., AV45 and FDG [50, 51]. However, these biomarkers are commonly expensive, time-consuming, and require well-trained doctors or specialist equipment. Furthermore, recent studies find that cognitive tests can capture useful information that other biomarkers can not capture, and combining cognitive tests and other biomarkers could significantly improve the results [52]. In our study, we have similar findings on the predictive rule sets for ADNI-A. Therefore, the role of the cognitive tests remains irreplaceable.

Another solution is integrating existing cognitive tests and biomarkers using machine learning models to improve performance [52–54]. However, conventional interpretable models, e.g., decision trees and linear models, can hardly deal with complex tasks like the prediction or diagnosis of dementia due to their limited model capacity. Hence, deep learning models and ensemble models are widely used in dementia related tasks. However, these models are commonly considered black-box models, and we can hardly understand their decision mechanism. Therefore, potential biases or errors could hide in these black-box models, and the cooperation of models and doctors is also hard to carry out. Even if we can try to interpret the black-box models using post-hoc methods, e.g., LIME and SHAP [26–28], the consistency between the interpretation and the original model is not guaranteed. The complexities and performance of baseline models shown in our study also verify the drawbacks of conventional interpretable models and black-box models.

Compared with other machine learning models, RRL not only has good classification performance but also obtains a better trade-off between performance and interpretability in our study. This suggests that RRL could be the best choice for doctors in most cases. Furthermore, after the visualization of trained RRL, the simulatability of RRL is good (similar to decision trees). To get the same result with RRL, doctors only need to compare the features with their corresponding cut-off values and calculate the score of each class. Therefore, no operation training is needed for doctors using RRL, and it is easy to deploy RRL.

Different from other studies [52,55,56], we also consider promoting the predictive and diagnostic rules learned by our model, especially in LMIC. In the feature selection step, doctors are asked to select features according to their costs, time consumption, equipment requirement, and doctor training cost to suit different scenarios. For example, the predictive rule set for ADNI-E, i.e., RS-E, only uses demographics and features from easily available cognitive tests. These easily available cognitive tests cost less time and need less training for the doctors. Therefore, hospitals lacking doctors and equipment, e.g., hospitals in LMIC, can still use RS-E for MCI-C prediction. In contrast, the predictive rule set for ADNI-A, i.e., RS-A, uses features from hard-to-obtain cognitive tests and biomarkers. These hard-to-obtain cognitive tests could cost a long time and need well-trained doctors. The hard-to-obtain biomarkers also need well-trained doctors and expensive equipment. Hence, RS-A is more suitable for hospitals with sufficient resources, e.g., a tertiary hospital, to further improve the performance of MCI-C prediction.

The framework we proposed is also promising for other medical tasks that can be handled with supervised learning, e.g., the diagnosis of depression [57,58]. Using our framework in these tasks is not only able to obtain diagnostic rule sets with high performance and good interpretability but also helpful for doctors to discover new knowledge hidden in the data.

Our study has the following limitations. First, unlike Long Short-Term Memory (LSTM) [59], RRL can not directly deal with longitudinal time sequence data, especially variable-length sequence data. Therefore, in our study, our model only uses the state of subjects at a certain stage, which may lose some useful information hidden in the longitudinal time sequence data. Second, some of the features used in our study are coarse-grained. Their corresponding fine-grained features could be used to further improve the learned rules, e.g., saving time in testing. For example, we can consider the score of each sub-item of MMSE as one feature, rather than only the total score of MMSE as one feature. Third, to keep the interpretability of our rules, we can not directly use unstructured data, e.g., MRI and PET images, for training. Instead, we need to extract interpretable features from these unstructured data, which could result in information loss compared with end-to-end methods. Fourth, our study can not accurately identify cognitively normal (CN) subjects who progress to dementia since the sample size of CN-to-Dementia subjects is small. Further data collections are therefore warranted.

In conclusion, we propose a new framework that adopts a Rule-based Representation Learner to learn interpretable rules for the prediction and early diagnosis of dementia mainly based on cognitive tests. To

make the rule sets learned by RRL more intuitive and convenient to use for doctors, we propose a novel visualization form of RRL. To ensure the learned rules are easy to promote and deploy, especially in low- and middle-income countries, feature selection and rule set selection are also carried out by doctors considering the situation of different scenarios. The results on ADNI and PUMCH verify that, even if only cognitive tests are used, we can still obtain rule sets with high performance and good interpretability for the prediction and early diagnosis of dementia with the help of doctors and RRL.

APPENDIX

Tables 3 and 4 list all the features used in the ADNI and PUMCH data sets, respectively. The type of each feature is decided by doctors in the feature selection step.

Acknowledgements

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Funding

This work was supported in part by National Key Research and Development Program of China under Grant No. 2020YFA0804503, 2020YFA0804501, National Natural Science Foundation of China under Grant No. 61521002, and Beijing Academy of Artificial Intelligence (BAAI).

Abbreviations

MCI: Mild Cognitive Impairment; MCI-C: MCI converter; MCI-NC: MCI nonconverter; AD: Alzheimer's disease; CN: Cognitively Normal; AUC: Area Under the Curve; ROC: Receiver Operating Characteristic; LMIC: Low- and Middle-Income Countries; HIC: High Income Countries; MRI: Magnetic Resonance Imaging; PET: Positron Emission Tomography; MMSE: Mini-Mental State Exam; FAQ: Functional Assessment Questionnaire; ADNI: Alzheimers Disease Neuroimaging Initiative; PUMCH: Peking Union Medical College Hospital; RRL: Rule-based Representation Learner; LR: Logistic Regression; PLNN: Piecewise Linear Neural Network; SVM: Support Vector Machines; RF: Random Forest; XGBoost: eXtreme Gradient Boosting; MLP: Multilayer Perceptron; LSTM: Long Short-Term Memory;

Availability of data and materials

The ADNI data that support the findings of this study are publicly available in the LONI database (https://ida.loni.usc.edu). The PUMCH data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethics approval and consent to participate

All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of PUMCH (No. JS1836).

Competing interests

The authors report no competing interests.

Consent for publication

Not Applicable

Authors' contributions

Z.W. and Jianyong Wang conceived the work. Z.W., Jie Wang, C.L., L.D., R.Z., C.M. and J.G. contributed to the data acquisition and resource allocation. Z.W., N.L., X.L., R.Z., Z.D. and W.Z. contributed to the design and development of the models, software and the experiments. Z.W., Jie Wang, C.L., J.G. and Jianyong Wang interpreted, analysed and presented the experimental results. Z.W., Jie Wang, N.L., X.L., R.Z., Z.D., W.Z., J.G. and Jianyong Wang contributed to drafting and revising the manuscript. All authors read and approved the final manuscript and are personally accountable for its content.

Author details

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, P.R. China. [2]Department of Neurology, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical Science/Peking Union Medical College, Beijing, P.R. China. [3]School of Computer Science and Technology, East China Normal University, Shanghai, P.R. China.

References

1. World Health Organization: Dementia (2020). https://www.who.int/en/news-room/fact-sheets/detail/dementia Accessed 21 Sep 2020
2. Organization, W.H., et al.: Risk reduction of cognitive decline and dementia: Who guidelines (2019)
3. Patterson, C., et al.: World alzheimer report 2018: The state of the art of dementia research: New frontiers (2018)
4. Nakamura, A.E., Opaleye, D., Tani, G., Ferri, C.P.: Dementia underdiagnosis in brazil. The Lancet **385**(9966), 418–419 (2015)
5. Jitapunkul, S., Chansirikanjana, S., Thamarpirat, J.: Undiagnosed dementia and value of serial cognitive impairment screening in developing countries: A population-based study. Geriatrics & gerontology international **9**(1), 47–53 (2009)
6. Dias, A., Patel, V.: Closing the treatment gap for dementia in india. Indian journal of psychiatry **51**(Suppl1), 93 (2009)
7. Olazarán, J., Reisberg, B., Clare, L., Cruz, I., Peña-Casanova, J., Del Ser, T., Woods, B., Beck, C., Auer, S., Lai, C., *et al.*: Nonpharmacological therapies in alzheimer's disease: a systematic review of efficacy. Dementia and geriatric cognitive disorders **30**(2), 161–178 (2010)
8. Prince, M., Bryce, R., Ferri, C.: World alzheimer report 2011: The benefits of early diagnosis and intervention (2018)
9. Nasreddine, Z.S., Phillips, N.A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J.L., Chertkow, H.: The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. Journal of the American Geriatrics Society **53**(4), 695–699 (2005)
10. Tsoi, K.K., Chan, J.Y., Hirai, H.W., Wong, S.Y., Kwok, T.C.: Cognitive tests to detect dementia: a systematic review and meta-analysis. JAMA internal medicine **175**(9), 1450–1458 (2015)
11. Bellio, M., Oxtoby, N.P., Walker, Z., Henley, S., Ribbens, A., Blandford, A., Alexander, D.C., Yong, K.X.: Analyzing large alzheimer's disease cognitive datasets: Considerations and challenges. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring **12**(1), 12135 (2020)

Table 3 All the features used in the ADNI dataset.

Type	Feature	Description
Demographics	AGE	Age
	PTGENDER	Sex
	PTEDUCAT	Education
	PTETHCAT	Ethnicity
	PTRACCAT	Race
	PTMARRY	Marital
Easy-to-obtain Cognitive Tests	MMSE	Mini-Mental State Examination
	RAVLT.immediate	RAVLT Immediate (sum of 5 trials)
	RAVLT.learning	RAVLT Learning (trial 5 - trial 1)
	RAVLT.forgetting	RAVLT Forgetting (trial 5 - delayed)
	RAVLT.perc.forgetting	RAVLT Percent Forgetting
	LDELTOTAL	Logical Memory - Delayed Recall
	DIGITSCOR	Digit Symbol Substitution
	TRABSCOR	Trails B
	FAQ	Functional Assessment Questionnaire
	MOCA	Montreal Cognitive Assessment
	mPACCdigit	ADNI modified Preclinical Alzheimer's Cognitive Composite (PACC) with Digit Symbol Substitution
	mPACCtrailsB	ADNI modified Preclinical Alzheimer's Cognitive Composite (PACC) with Trails B
Hard-to-obtain Cognitive Tests	CDRSB	Clinical Dementia Rating scale Sum of Boxes
	ADAS11	Alzheimer's Disease Assessment Scale (ADAS) 11
	ADAS13	ADAS 13 (including Delayed Word Recall and Number Cancellation)
	ADASQ4	ADAS Delayed Word Recall
	EcogPtMem	Pt ECog - Mem
	EcogPtLang	Pt ECog - Lang
	EcogPtVisspat	Pt ECog - Vis/Spat
	EcogPtPlan	Pt ECog - Plan
	EcogPtOrgan	Pt ECog - Organ
	EcogPtDivatt	Pt ECog - Div atten
	EcogPtTotal	Pt ECog - Total
	EcogSPMem	SP ECog - Mem
	EcogSPLang	SP ECog - Lang
	EcogSPVisspat	SP ECog - Vis/Spat
	EcogSPPlan	SP ECog - Plan
	EcogSPOrgan	SP ECog - Organ
	EcogSPDivatt	SP ECog - Div atten
	EcogSPTotal	SP ECog - Total
Other Biomarkers	APOE4	Number of APOEe4 alleles
	FDG	Average FDG-PET of angular, temporal, and posterior cingulate
	PIB	Average PIB SUVR of frontal cortex, anterior cingulate, precuneus cortex, and parietal cortex
	AV45	AV45 ratio (cortical grey matter/whole cerebellum) Summary florbetapir cortical SUVR normalized by whole cerebellum.
	FBB	FBB ratio (cortical grey matter/whole cerebellum) Summary florbetaben cortical SUVR normalized by whole cerebellum.
	ABETA	Cerebrospinal Fluid (CSF) ABETA
	TAU	CSF TAU
	PTAU	CSF PTAU
	Ventricles	UCSF Ventricles
	Hippocampus	UCSF Hippocampus
	WholeBrain	UCSF WholeBrain
	Entorhinal	UCSF Entorhinal
	Fusiform	UCSF Fusiform
	MidTemp	UCSF Med Temp
	ICV	UCSF ICV

12. Tombaugh, T.N., McIntyre, N.J.: The mini-mental state examination: a comprehensive review. Journal of the American Geriatrics Society **40**(9), 922–935 (1992)
13. Cornelis, E., Gorus, E., Beyer, I., Bautmans, I., De Vriendt, P.: Early diagnosis of mild cognitive impairment and mild dementia through basic and instrumental activities of daily living: Development of a new evaluation tool. PLoS medicine **14**(3), 1002250 (2017)
14. Zamrini, E., De Santi, S., Tolar, M.: Imaging is superior to cognitive testing for early diagnosis of alzheimer's disease. Neurobiology of aging **25**(5), 685–691 (2004)
15. Querbes, O., Aubry, F., Pariente, J., Lotterie, J.-A., Démonet, J.-F., Duret, V., Puel, M., Berry, I., Fort, J.-C., Celsis, P., *et al.*: Early

Table 4 All the features used in the PUMCH dataset.

Type	Feature	Description
Demographics	Age	years
	Age group	< 65 years old; 65-85 years old; > 85 years old
	Gender	Male or female
	Education	years
	Education group	≤6 years of education; >6 and ≤12 years of education; >12 years of education
	Handedness	Right-handedness; left-handedness; Mixed- handedness
Easy-to-obtain Cognitive Tests	HAD-anxiety	Anxiety score of Hospital Anxiety and Depression scale
	HAD-depression	Depression score of Hospital Anxiety and Depression scale
	CVF	Category verbal fluency
	DST	Digit symbol test
	TMT A	Trail making test A
	TMT B	Trail making test B
	CDT	Clock drawing test
	PAL N1	The first learning trial of Paired-associate learning of The Clinical Memory Test (PAL)
	PAL N1- Simple part	Six simple word pairs of PAL N1
	PAL N1- Difficult part	Six difficult word pairs of PAL N1
	PAL-N2	The second learning trial of PAL
	PAL N2- Simple part	Six simple word pairs of PAL N2
	PAL N2- Difficult part	Six difficult word pairs of PAL N2
	PAL-N3	The third learning trial of PAL
	PAL N3- Simple part	Six simple word pairs of PAL N3
	PAL N3- Difficult part	Six difficult word pairs of PAL N3
	PAL-T	The total score of the three learning trials of PAL
	BDT N1	The first figure of Block design test of the Aphasia Battery of Chinese (BDT)
	BDT N2	The second figure of BDT
	BDT N3	The third figure of BDT
	BDT-T	The total score of BDT
	Luria TST	Luria three-step task
	FC N1	The first figure of Figure copying of the Aphasia Battery of Chinese (FC)
	FC N2	The second figure of FC
	FC N3	The third figure of FC
	FC N4	The fourth figure of FC
	FC-T	The total score of four figures of FC
	Gesture imitation	Imitation of seven hand gestures
	Modified-Rey copy	Copy of a modified Rey-Osterrieth figure
	Speech length	Sentence length of spontaneous speech
	Speech time	Time of spontaneous speech
	Semantic paraphasia	"Yes" or "No"
	Phonemic paraphasia	"Yes" or "No"
	Repetitive language	"Yes" or "No"
	Word retrieval	Hesitation and delay in spoken production; "Yes" or "No"
	Language output	"Fluent" or "nonfluent"
	Language comprehension	Executing five commands
	Repetition	Repeating three sentences
	Object naming	The number of correctly named objects
	Color naming	The number of correctly named colors
	AVLT N1	The first learning trial of Auditory Verbal Learning Test–Huashan version (AVLT-H)
	AVLT N2	The second learning trial of AVLT-H
	AVLT N3	The third learning trial AVLT-H
	AVLT-L	Total score of three learning trials of AVLT-H
	AVLT N4	The fourth short delayed free recall trial of AVLT-H
	AVLT N5	The fifth long delayed free recall trial of AVLT-H
	AVLT-T	Total score of AVLT-L, AVLT N4 and AVLT N5
	AVLT N6	The sixth delayed category cue recall trial of AVLT-H
	AVLT-RH	Recognitions hits of AVLT-H
	AVLT-RF	Recognitions false of AVLT-H
	LMT N1	The first story of logical memory test of modified Wechsler Memory Scale (LMT)
	LMT N2	The second story of LMT
	LMT N3	The third story of LMT
	LMT-T	The total score of LMT
	Modified-Rey recall	Modified Rey-Osterreith figure with a 10-minute free recall
	Modified-Rey Recognition	Recognition of Modified Rey-Osterreith figure; "True" or "false"
	Similarities	Similarities of the Wechsler Adult Intelligence Scale
	Calculations	Calculations of the Wechsler Adult Intelligence Scale

diagnosis of alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain **132**(8), 2036–2047 (2009)
16. Pfeffer, R.I., Kurosaki, T.T., Harrah Jr, C., Chance, J.M., Filos, S.: Measurement of functional activities in older adults in the community. Journal of gerontology **37**(3), 323–329 (1982)
17. Frisoni, G.B., Fox, N.C., Jack, C.R., Scheltens, P., Thompson, P.M.: The clinical use of structural mri in alzheimer disease. Nature Reviews Neurology **6**(2), 67–77 (2010)
18. Iaccarino, L., Sala, A., Caminiti, S.P., Perani, D.: The emerging role of pet imaging in dementia. F1000Research **6** (2017)
19. Huang, W., Qiu, C., von Strauss, E., Winblad, B., Fratiglioni, L.: Apoe genotype, family history of dementia, and alzheimer disease risk: a 6-year follow-up study. Archives of neurology **61**(12), 1930–1934 (2004)
20. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J.: A guide to deep learning in healthcare. Nature medicine **25**(1), 24–29 (2019)
21. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics **19**(6), 1236–1246 (2018)
22. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
23. Molnar, C.: Interpretable Machine Learning. Lulu. com, Lulu. com (2020)
24. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 1097–1105 (2012)
26. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
27. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
28. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777 (2017)
29. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019)
30. Wang, Z., Zhang, W., Liu, N., Wang, J.: Scalable rule-based representation learning for interpretable classification. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021)
31. Petersen, R.C., Aisen, P., Beckett, L.A., Donohue, M., Gamst, A., Harvey, D.J., Jack, C., Jagust, W., Shaw, L., Toga, A., *et al.*: Alzheimer's disease neuroimaging initiative (adni): clinical characterization. Neurology **74**(3), 201–209 (2010)
32. Wang, J., Wang, Z., Liu, N., Liu, C., Mao, C., Dong, L., Li, J., Huang, X., Lei, D., Chu, S., *et al.*: Random forest model in the diagnosis of dementia patients with normal mini-mental state examination scores. Journal of personalized medicine **12**(1), 37 (2022)
33. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT press, Cambridge, MA (2016)
34. Wang, Z., Zhang, W., Ning, L., Wang, J.: Transparent classification with multilayer logical perceptrons and random binarization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 6331–6339 (2020)
35. Evesham, H.A.: The History and Development of Nomography. Docent Press, Docent Press (2010)
36. Saravanan, N., Sathish, G., Balajee, J.: Data wrangling and data leakage in machine learning for healthcare. International Journal of Emerging Technologies and Innovative Research **5**(8), 553–557 (2018)
37. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., *et al.*: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp. 8026–8037 (2019)
38. Breiman, L.: Classification and Regression Trees. Routledge, Routledge (2017)
39. Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M., Klein, M.: Logistic Regression. Springer, Springer (2002)
40. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT press, MIT press (2001)
41. Chu, L., Hu, X., Hu, J., Wang, L., Pei, J.: Exact and consistent interpretation for piecewise linear neural networks: A closed form solution. In: SIGKDD, pp. 1244–1253 (2018). ACM
42. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
43. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
44. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML, pp. 807–814 (2010)
45. Wong, D.F., Rosenberg, P.B., Zhou, Y., Kumar, A., Raymont, V., Ravert, H.T., Dannals, R.F., Nandi, A., Brašić, J.R., Ye, W., *et al.*: In vivo imaging of amyloid deposition in alzheimer disease using the radioligand 18f-av-45 (flobetapir f 18). Journal of nuclear medicine **51**(6), 913–920 (2010)
46. Belleville, S., Fouquet, C., Hudon, C., Zomahoun, H.T.V., Croteau, J.: Neuropsychological measures that predict progression from mild cognitive impairment to alzheimer's type dementia in older adults: a systematic review and meta-analysis. Neuropsychology review **27**(4), 328–353 (2017)
47. Khoury, R., Ghossoub, E.: Diagnostic biomarkers of alzheimer's disease: a state-of-the-art review. Biomarkers in Neuropsychiatry **1**, 100005 (2019)
48. Devi, G., Scheltens, P.: Heterogeneity of alzheimer's disease: consequence for drug trials? Alzheimer's research & therapy **10**(1), 1–3 (2018)
49. Velayudhan, L., Ryu, S.-H., Raczek, M., Philpot, M., Lindesay, J., Critchfield, M., Livingston, G.: Review of brief cognitive tests for patients with suspected dementia. International psychogeriatrics **26**(8), 1247–1262 (2014)
50. Johnson, K.A., Sperling, R.A., Gidicsin, C.M., Carmasin, J.S., Maye, J.E., Coleman, R.E., Reiman, E.M., Sabbagh, M.N., Sadowsky, C.H., Fleisher, A.S., *et al.*: Florbetapir (f18-av-45) pet to assess amyloid burden in alzheimer's disease dementia, mild cognitive impairment, and normal aging. Alzheimer's & Dementia **9**(5), 72–83 (2013)
51. Cohen, A.D., Klunk, W.E.: Early detection of alzheimer's disease using pib and fdg pet. Neurobiology of disease **72**, 117–122 (2014)
52. Palmqvist, S., Tideman, P., Cullen, N., Zetterberg, H., Blennow, K., Dage, J.L., Stomrud, E., Janelidze, S., Mattsson-Carlgren, N., Hansson, O.: Prediction of future alzheimer's disease dementia using plasma phospho-tau combined with other accessible measures. Nature Medicine **27**(6), 1034–1042 (2021)
53. So, A., Hooshyar, D., Park, K.W., Lim, H.S.: Early diagnosis of dementia from clinical data by machine learning techniques. Applied Sciences **7**(7), 651 (2017)
54. Dai, P., Gwadry-Sridhar, F., Bauer, M., Borrie, M., Teng, X.: Healthy cognitive aging: a hybrid random vector functional-link model for the analysis of alzheimer's disease. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
55. Westman, E., Muehlboeck, J.-S., Simmons, A.: Combining mri and csf measures for classification of alzheimer's disease and prediction of mild cognitive impairment conversion. Neuroimage **62**(1), 229–238 (2012)
56. Karikari, T.K., Benedet, A.L., Ashton, N.J., Rodriguez, J.L., Snellman, A., Suárez-Calvet, M., Saha-Chaudhuri, P., Lussier, F., Kvartsberg, H., Rial, A.M., *et al.*: Diagnostic performance and prediction of clinical progression of plasma phospho-tau181 in the alzheimer's disease neuroimaging initiative. Molecular Psychiatry **26**(2), 429–442 (2021)
57. Chekroud, A.M., Zotti, R.J., Shehzad, Z., Gueorguieva, R., Johnson, M.K., Trivedi, M.H., Cannon, T.D., Krystal, J.H., Corlett, P.R.: Cross-trial prediction of treatment outcome in depression: a machine learning approach. The Lancet Psychiatry **3**(3), 243–250 (2016)
58. Priya, A., Garg, S., Tigga, N.P.: Predicting anxiety, depression and stress in modern life using machine learning algorithms. Procedia Computer Science **167**, 1258–1267 (2020)
59. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)