brain sciences

MDPI

Article

Dual Semi-Supervised Learning for Classification of Alzheimer's Disease and Mild Cognitive Impairment Based on Neuropsychological Data

Yan Wang [1], Xuming Gu [1], Wenju Hou [1], Meng Zhao [2], Li Sun [2] and Chunjie Guo [3,*]

[1] Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China
[2] Department of Neurology and Neuroscience Center, The First Hospital of Jilin University, Changchun 130021, China
[3] Department of Radiology, The First Hospital of Jilin University, Changchun 130021, China
* Correspondence: guocj@jlu.edu.cn; Tel.: +86-1580-430-0151

Abstract: Deep learning has shown impressive diagnostic abilities in Alzheimer's disease (AD) research in recent years. However, although neuropsychological tests play a crucial role in screening AD and mild cognitive impairment (MCI), there is still a lack of deep learning algorithms only using such basic diagnostic methods. This paper proposes a novel semi-supervised method using neuropsychological test scores and scarce labeled data, which introduces difference regularization and consistency regularization with pseudo-labeling. A total of 188 AD, 402 MCI, and 229 normal controls (NC) were enrolled in the study from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. We first chose the 15 features most associated with the diagnostic outcome by feature selection among the seven neuropsychological tests. Next, we proposed a dual semi-supervised learning (DSSL) framework that uses two encoders to learn two different feature vectors. The diagnosed 60 and 120 subjects were randomly selected as training labels for the model. The experimental results show that DSSL achieves the best accuracy and stability in classifying AD, MCI, and NC (85.47% accuracy for 60 labels and 88.40% accuracy for 120 labels) compared to other semi-supervised methods. DSSL is an excellent semi-supervised method to provide clinical insight for physicians to diagnose AD and MCI.

Keywords: Alzheimer's disease; semi-supervised Learning; neuropsychological test; difference regularization

Citation: Wang, Y.; Gu, X.; Hou, W.; Zhao, M.; Sun, L.; Guo, C. Dual Semi-Supervised Learning for Classification of Alzheimer's Disease and Mild Cognitive Impairment Based on Neuropsychological Data. *Brain Sci.* **2023**, *13*, 306. https://doi.org/10.3390/brainsci13020306

Academic Editor: Francesco Di Lorenzo and Annibale Antonioni

Received: 4 December 2022
Revised: 27 January 2023
Accepted: 7 February 2023
Published: 10 February 2023

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative brain disease, which indicates that the condition gradually worsens over time. Patients in the early stages of AD, namely mild cognitive impairment (MCI), have a greater likelihood of converting to AD years later [1]. The lesions of the disease occur mainly in the cerebral cortex and hippocampus, which causes patients to develop cognitive impairments in language, memory, and other aspects [2]. Positron emission tomography (PET), magnetic resonance imaging (MRI), and cerebrospinal fluid (CSF) biomarkers are included in A/T/N system for research [3], which highlights the importance of reliable biomarkers for AD diagnosis. However, these measures' high cost and intrusiveness limit their widespread application and potential in clinical screening patients for AD [4]. Therefore, it is vital to identify non-invasive, reliable, and widely available diagnostic biomarkers for AD.

Research has suggested that the traditional diagnosis of cognitive disorders remains limited to subjective symptoms and observable features, and that ML offers a novel paradigm that can enable automated and more objective evaluation of various psychiatric diseases [5]. In recent years, researchers have used machine learning (ML), especially

deep learning (DL), instead of traditional methods to assist in the diagnosis of AD [6–8]. In particular, the fully supervised DL-based method is the dominant approach in AD diagnosis. Specifically, convolutional neural networks (CNN) and graph convolutional networks (GCN) have demonstrated excellent performance in medical image classification tasks [9]. Amini et al. [10] compared several ML methods for AD diagnosis using functional magnetic resonance imaging (fMRI) images. They showed that CNN outperformed all other traditional ML techniques in effectively detecting AD severity. Zhou et al. [11] proposed an interpretable GCN framework using multimodal brain imaging data to classify AD, MCI, and normal controls (NC). Considering the node features and their connectivity in the network, Zhou et al. [12] further proposed a sparse interpretable GCN framework, which uses multiple modalities of brain imaging data to classify AD. However, due to the complexity of disease pathology, it is costly to obtain the ground truth labels for AD and MCI, which requires expert knowledge. The lack of labeled data remains a significant obstacle to the progress of DL in AD diagnosis [13]. Semi-supervised learning (SSL) methods in DL are particularly suitable for situations where labeled data is scarce [14].

Neuropsychological tests are commonly used in clinical practice to determine the degree of cognitive impairment including AD and MCI [15]. These tests are short-cycle, low-cost, and easy to conduct compared to medical imaging and CSF measures. Research suggests neuropsychological test results may have as much screening potential for AD patients as CSF and MRI biomarkers [16]. Grassi et al. [17] used predictors integrating sociodemographic characteristics, cognitive measures, clinical tests, etc. They used multiple supervised learning methods to identify which subjects with MCI would convert to AD in the following years. Battista et al. [18] used a combination of support vector machine (SVM) and 131 measures from 324 participants, including different neuropsychological tests to classify subjects with different clinical dementia ratings (CDR). Although ML methods such as SVM have yielded promising results, no predictors can be used as the gold standard, and some studies have found problems with some measures [19]. As advanced and prevalent ML methods, neural networks have rarely been applied to diagnosing AD using neuropsychological tests, whose widespread application will provide clinical insight for physicians to determine the degree of cognitive impairment.

To address the problem of difficulty in obtaining labeled data, this paper proposes a new method for Alzheimer's disease classification that reduces the need for labeled data based on SSL. Our proposed method applies easily available and non-invasive neuropsychological test data for the diagnosis of AD. First, we calculate the correlation of each neuropsychological test on the diagnostic results by Pearson's correlation coefficient and select features according to the magnitude of the coefficients. Then, we propose the dual semi-supervised learning (DSSL) algorithm, which uses two different encoders to learn different feature representations of the samples. In addition, we combine pseudo-labeling with consistency regularization. The two predictions obtained from the two feature representations are hard-labeled and then used as mutual pseudo-labels.

To evaluate the classification performance of DSSL, we conduct extensive experiments in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://ADNI.loni.usc.edu/ (accessed on 18 December 2021)). Experimental results show that DSSL largely outperforms existing semi-supervised methods in a variety of evaluation metrics, and the short training time of the model demonstrates its practicality in clinical diagnosis. By dividing the training data and performing training many times, we find that DSSL also has strong stability.

The contributions of this paper are:

1. We select some neuropsychological tests by feature selection, which are better predictors of automatic classification and can provide clinical diagnostic references to physicians;
2. Propose a novel semi-supervised method that introduces difference regularization in unsupervised loss computation to enhance model perturbations by learning two different feature representations;

3. Propose a tri-classification framework for cognitive impairment based on improved SSL and CNN, which identifies AD, MCI, and NC using the most straightforward method (i.e., neuropsychological tests) and fewer labels. Experimental results based on the ADNI dataset indicate that the classifier outperforms other semi-supervised methods in terms of accuracy and stability.

2. Theoretical Backgrounds

Deep neural networks contain many hidden layers, each containing a large number of hidden nodes, which gives it a powerful fitting capability to approximate almost any complex function. However, the powerful fitting ability of deep learning relies on a large amount of training data. Training with only a small amount of labeled data often leads to overfitting problems [13]. In addition, the interpretability of deep learning is still being explored by researchers [20].

2.1. Semi-Supervised Learning

SSL is a powerful method for training models on large datasets with only a small number of labels. SSL alleviates the need for labeled data by learning the connections and differences between unlabeled data. In the following sections, we discuss the background related to this work. For a three-class classification problem, let (x, p) denote a labeled example and u denote an unlabeled example, respectively. D denotes the number of labeled samples and μ_D denotes the number of unlabeled samples. Let $p_{model}(y|x)$ denote the predicted probability generated by the model with input x. Let $\mathbb{I}(condition)$ denote 1 if the condition holds and 0 if not. Let $H(p, q)$ denote the cross-entropy between two probability distributions p and q.

In the semi-supervised task, we aim to predict the classification using several image labels accurately. Especially for a reasonably large dataset, labeling these images manually could be a tedious and challenging task. Therefore, it is now understandable why we chose the semi-supervised algorithm for our study.

2.1.1. Consistency Regularization

Consistency regularization is an essential component of the deep neural network model in the SSL algorithm. Consistency regularization employs a perturbation strategy in which the same sample is altered to yield various outputs. The perturbation approach assumes that the model should output similar predictions when the same input sample is perturbed. This idea was first proposed in [21] and promoted by [22,23]. The perturbation methods can be divided into sample perturbation methods and model perturbation methods according to the different perturbation stages. Sample perturbation refers to the data augmentation of the input sample to obtain a new sample that is different from the previous sample but mostly similar; model perturbation is a change in the model, where the same sample undergoes a different model to produce a difference in the output results. Consistency regularization in the model is mainly trained on unlabeled data by the loss function:

$$\| p_{model}(y|\mathcal{A}(u)) - p_{model}(y|\mathcal{A}(u)) \|_2^2, \tag{1}$$

where $\|\cdot\|_2$ denotes the L2 norm and $\mathcal{A}(u)$ denotes data augmentation. Note that both $\mathcal{A}(u)$ and p_{model} are random functions, so the two terms in Equation (1) are not the same. The consistency regularization with different $\mathcal{A}(u)$ belongs to the sample perturbation method. Virtual adversarial training [24] (VAT) uses adversarial perturbation to generate an adversarial sample that forms a difference from the original sample, and MixMatch [25] uses the mixup [26] method to perform data augmentation on the input samples. FixMatch [27] uses both strong and weak augmentations, and experiments with strong augmentations based on RandAugment [28] and CTAugment [29]. Most of the existing sample perturbation methods, however, are data augmentation methods used for image data. It is not widely applicable to other types of data. The consistency regularization with different p_{model} belongs

to the sample perturbation method. Π-model [23] uses the randomness of dropout [30] to perturb the model so that the outputs of the same input sample are different. Temporal ensembling [23] uses the average of previous model checkpoints when generating artificial labels for comparison with the current prediction. Mean teacher [31] divides the model into two types: the student model, which is a general training model, and the teacher model, which is obtained by an exponential moving average of the parameters of the student model. For the same input, the different outputs obtained by the student and teacher models constitute consistency regularization.

2.1.2. Pseudo-Labeling

The low-density assumption is a common fundamental assumption in SSL, referring to the classification boundary not passing through high-density regions in the input space. One way to achieve this assumption requires SSL models to output low-entropy predictions for unlabeled data. Pseudo-labeling [32] implicitly minimizes entropy by generating a hard (one-hot) label on the high-confidence prediction results of unlabeled data and using this hard label along with the model prediction result as parameters for the standard cross-entropy loss. Letting $q = p_{model}(y|u)$ and $\hat{q} = arg\ \max(q)$, the loss function used for the pseudo-labeling can be expressed as:

$$\mathbb{I}(\max(q) \geq \tau) H(\hat{q}, q), \tag{2}$$

where τ denotes the threshold. Pseudo-labeling treats the predictions of SSL classifiers on unlabeled data as artificial labels.

2.1.3. Label Propagation

Label propagation is a graph-based SSL method that associates all labeled and unlabeled samples by constructing a graph. The nodes in the graph include labeled and unlabeled samples, and the weights of the edges represent the similarity between two nodes. The labels of the samples are propagated through the edges between the nodes. Recently, it has been combined with pseudo-labels as a novel way of giving pseudo-labels or calculating losses based on pseudo-labels. Iscen et al. [33] used a label propagation method based on the manifold assumption to predict the current node based on the k nodes with high similarity, and used the predicted results to generate pseudo-labels for unlabeled samples. SimPLE [34] introduces pair loss in addition to supervised loss and consistency loss, which decrease the noise of pseudo-labels by setting a confidence threshold and similarity threshold.

2.2. Contrastive Learning

Self-supervised learning, unlike supervised learning which requires expensive labeling, is able to use unlabeled data to learn the underlying representation. Contrast learning, one of the important methods of self-supervised learning, aims to learn an encoder that encodes data of the same kind similarly and makes the encoding results of different classes of data as different as possible. The Pretext task is a self-supervised task using pseudo-labels to learn data representation. How to design the pretext task to better fit the SSL downstream tasks is the key to incorporating self-supervised learning into the SSL model. The CCSSL [35] framework introduces class-aware contrast loss on top of the SSL model, seamlessly integrating clustering and comparison in the feature space. LaSSL [36] learns differentiated feature representations that enable aggregation of same-class samples and dispersion of different class samples by minimizing class-aware contrast loss and performs label propagation based on the feature representations.

3. Materials and Methods

3.1. ADNI Database

Data used in this study is obtained from the ADNI database. ADNI was launched in 2003 as a longitudinal multicenter study led by Principal Investigator Michael W. Weiner. The initial objective of ADNI was to develop MRI, PET, and other biomarkers for early detection and tracking. For up-to-date information, see www.adni-info.org (accessed on 18 December 2021). In this study, we chose baseline neuropsychological data from the preliminary phase of the project (ADNI-1). The data we used are from 819 subjects including 188 AD subjects, 402 MCI subjects, and 229 NC subjects. The characteristics of the subjects selected for this study are shown in Table 1.

Table 1. Subject characteristics.

Characteristic	AD (n = 188)	MCI (n = 402)	NC (n = 229)	p-Value
Gender (M/F)	99/80	257/143	119/110	-
Age	75.3 ± 7.5	74.8 ± 7.4	75.9 ± 5.0	-
MMSE	23.3 ± 2.0	27.0 ± 1.8	29.1 ± 1.0	<0.001
CDR	0.7 ± 0.3	0.5	0	<0.001
FAQ	13.1 ± 6.8	3.9 ± 4.5	0.1 ± 0.6	<0.001
ADAS1	6.1 ± 1.5	4.6 ± 1.4	2.9 ± 1.1	<0.001
RAVLT	23.2 ± 7.7	30.6 ± 9.0	43.3 ± 9.0	<0.001
NPIQ	3.5 ± 3.4	1.9 ± 2.7	0.3 ± 0.9	<0.001
GDS	1.7 ± 1.4	1.6 ± 1.4	0.8 ± 1.1	0.14

Data are expressed as mean ± standard deviation. MMSE = mini-mental state examination, CDR = clinical dementia rating, FAQ = functional activity questionnaire, ADAS1 = word list non-learning (mean) RAVLT = Anterograde episodic memory-verbal, NPIQ = neuropsychiatric inventory Q, GDS = geriatric depression scale. The p-values for the differences between AD, MCI and NC are based on two-way t-tests with Bonferroni correction.

3.2. Neuropsychological Data

The itemized scores of seven neuropsychological tests are used, including the Alzheimer's disease assessment scale-cognitive (ADAS-Cog) [37], the mini-mental state exam (MMSE) [38], the clinical dementia rating (CDR) [39], the Rey auditory verbal learning test (RAVLT) [40], the functional activity questionnaire (FAQ) [41], the neuropsychiatric inventory Q (NPIQ) [42], and the geriatric depression scale (GDS) [43]. These neuropsychological tests are widely used to determine the degree of cognitive impairment in clinical settings. Appendix A.1 details the cognitive functions associated with each test. A total of 64 itemized scores are derived from these seven tests. For each test, we use a different number of sub-scores, including 15 rubric scores from ADAS-cog, 31 rubric scores from MMSE, 1 rubric score from CDR, 4 rubric scores from RAVLT, 11 rubric scores from FAQ, 1 rubric score from NPIQ, and 1 rubric score from GDS. In the semi-supervised learning task of this paper, each itemized score is considered a feature of the sample. We provide a brief introduction of the neuropsychological tests selected as features in Appendix .1.

3.3. Method

3.3.1. Features Selection

Feature selection has a highly important role in DL. Pearson's correlation coefficient (PCC) [44], one of the most common feature selection methods, is applied to neuropsychological tests in this study. Although PCC cannot assess how similar a combination of multiple variables is to a single variable, it is still the most popular method for calculating the similarity between two variables. PCC evaluates the degree of correlation between two variables by calculating the standard deviation of the two variables and the covariance between them. PCC between the two variables X and Y is defined as:

$$\rho_{X,Y} = \frac{COV(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}, \quad (3)$$

where COV denotes the covariance, μ_X denotes the mean of X, μ_Y denotes the mean of Y, σ_X denotes the standard deviation of X, σ_Y denotes the standard deviation of Y, and E denotes the expectation. The value calculated by Equation (3) varies from -1 to 1. A value between 0 and 1 denotes that the two variables are positively correlated, while a value between -1 and 0 denotes that they are negatively correlated. The closer the absolute value is to 1, the stronger the correlation between the two variables.

3.3.2. Dual Semi-Supervised Learning

In this subsection, we introduce DSSL, a novel semi-supervised method, as a convenient and accurate classifier for the clinical diagnosis of AD. Inspired by fixMatch [27], DSSL combines consistency regularization and pseudo-labeling, two SSL methods discussed in the previous section. Figure 1 shows the overall view of the model for the supervised and unsupervised parts. DSSL applies model perturbation through two different encoders. To make the two encoders learn as different features as possible, DSSL introduces difference regularization, which stretches the distance between the features extracted from the input by the two encoders. The network architecture of the encoders is shown in Figure 2. For a sample, two different feature vectors are obtained through Encoder1 and Encoder2, respectively. These two vectors are then fed into the multilayer perceptron (MLP) network to obtain two prediction results. They serve each other as pseudo-labels for the different prediction results, which constitutes consistency regularization. Algorithm 1 provides the complete DSSL algorithm.

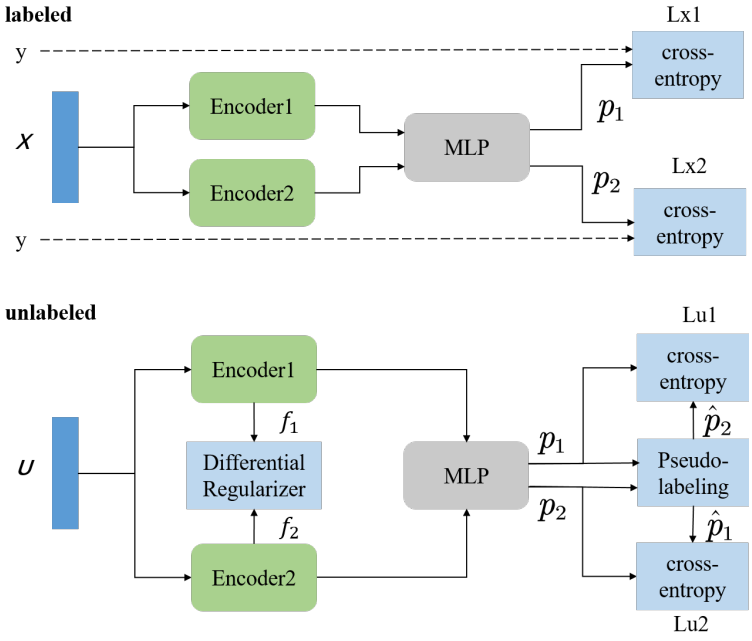


Figure 1. Overview of the proposed model.

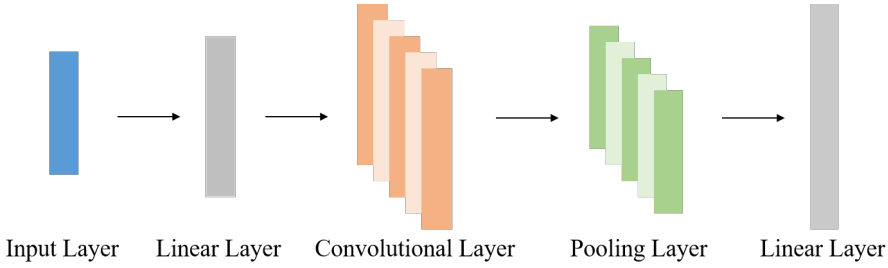


Figure 2. An illustration of the composition of the encoder. The encoder mainly consists of a connection layer, a convolutional layer, a pooling layer, and a connection layer. The main alteration part of different encoders is in the pooling layer.

Algorithm 1 DSSL algorithm.

Input: Labeled examples and their labels $\mathcal{X} = (x_d, p_d); d \in (1, \ldots, D)$, unlabeled examples $\mathcal{U} = u_d; d \in (1, \ldots, \mu_D)$, confidence threshold τ, differential regularizer loss weight β, unlabeled loss weight λ.

// Cross-entropy loss for labeled data through Encoder1
$l_{x1} = \frac{1}{D} \sum_{d=1}^{D} H(p_d, p_{model1}(y|x_d); \theta)$
// Cross-entropy loss for labeled data through Encoder2
$l_{x2} = \frac{1}{D} \sum_{d=1}^{D} H(p_d, p_{model2}(y|x_d); \theta)$
for *d=1* ***to*** μ_D **do**
  $f_1 = Encoder1(u_d)$ // Features of u_d extracted by Encoder1
  $f_2 = Encoder2(u_d)$ // Features of u_d extracted by Encoder2
  $q_1 = MLP(f_1)$ // Predictions of u_d through Encoeder1-MLP module
  $q_2 = MLP(f_2)$ // Predictions of u_d through Encoeder2-MLP module
end
// Difference regularization between two features
$R_D = \frac{1}{\|(Norm(f_1) - Norm(f_2))\|_F}$
// Cross-entropy loss with q_2 pseudo-label and q_1
$l_{u1} = \frac{1}{\mu_D} \sum_{\mu_D} \mathbb{I}(\max(q_2) \geq \tau) H(argmax(q_2), q_1)$
// Cross-entropy loss with q_1 pseudo-label and q_2
$l_{u2} = \frac{1}{\mu_D} \sum_{\mu_D} \mathbb{I}(\max(q_1) \geq \tau) H(argmax(q_1), q_2)$
return $l_{x1} + l_{x2} + \lambda(l_{u1} + l_{u2}) + \beta R_D$

3.3.3. Regularization of DSSL

Two regularizations are introduced in our approach, a difference regularization so that the two encoders learn different features, and a consistency regularization combined with pseudo-labeling.

Differential Regularizer (R_D) — We expect to learn two different aspects of the feature representation from Encoder1 and Encoder2. Therefore, we apply a difference regularization between the two features output by the two encoders. The distance between the two feature vectors is appropriately widened to increase the perturbation and to prepare for the consistency regularization later. The concrete implementation is shown below:

$$R_D = \frac{1}{\|(Norm(f_1) - Norm(f_2))\|_F}, \tag{4}$$

where *Norm* is the normalization operation, which aims to put two feature representations into an order of magnitude to compare, f_1 and f_2 are the feature vectors learned by the two encoders, and $\|\cdot\|_F$ denotes the Frobenius norm.

Consistency Regularization— DSSL combines consistency regularization with the pseudo-labeling approach by turning the model's predictions into hard labels. Not all hard labels of the samples are involved in the operation as parameters of the model's loss function. The model keeps only the pseudo-labels whose maximum prediction probability is higher than a predefined threshold. Assuming $q_2 = p_{model2}(y|u)$, where p_{model2} is the prediction of the Encoder2-MLP module, and q_2 is the prediction probability. Similarly, p_{model1} is the prediction of the Encoder1-MLP module, and q_1 is the prediction probability. We use $\hat{q}_2 = arg\ \max(q_2)$ as a pseudo-label. In other words, the category with the highest prediction probability is obtained as the pseudo-label of the sample. More specifically, consistency regularization is defined as:

$$l_{u1} = \frac{1}{\mu_D} \sum_{\mu_D} \mathbb{I}(\max(q_2) \geq \tau) H(\hat{q}_2, p_{model1}(y|u)), \tag{5}$$

where l_{u1} is the consistency loss of q_1 with $\hat{q}_2$ as the pseudo-label, τ is a scalar hyper-parameter representing the threshold value used to determine which samples participate in calculating the loss function.

3.3.4. Loss Function of DSSL

The training objective of DSSL is to minimize the following total objective function:

$$l_T = l_{x1} + l_{x2} + \lambda(l_{u1} + l_{u2}) + \beta R_D, \tag{6}$$

where λ and β are regularization coefficients. l_{u2} is similar to l_{u1}, which computes the cross-entropy loss of the hard label of q_1 with q_2. l_{x1} and l_{x2} are the standard cross-entropy loss between the true labels and the output of the Encoder1-MLP module, the Encoder2-MLP module, respectively. l_{x1} is formulated as the following expression:

$$l_{x1} = -\frac{1}{D}\sum_{D} y \log\ p_{model1}(y|x), \tag{7}$$

where x is the labeled data and y is the accurate label of the data. Since l_{x2} is similar to l_{x1}, it will not be discussed further here.

4. Results

4.1. Features Selection

We select a total of 64 itemized scores from 7 neuropsychological tests. To find characteristics that significantly discriminate Alzheimer's disease, we do PCC calculations between their scores and labels. Then, the correlation coefficients are ranked in descending order of absolute value, and the top 15 features are selected as input for the subsequent semi-supervised experiments. Their corresponding PCCs are shown in Table 2. The table shows that their total scores correlate more strongly with the degree of cognitive impairment compared to the sub-scores of each test.

Table 2. The 15 items with the highest absolute PCC.

Feature Name	Absolute PCC	Feature Name	Absolute PCC
CDR-SB	0.827928	FAQFORM	0.647591
MMSETOTAL	0.766936	RAVLT_immediate	0.629216
ADASMOD	0.743978	FAQFINAN	0.620305
ADAS_Q4	0.721948	FAQTRAVL	0.595336
FAQTOTAL	0.691619	FAQSHOP	0.574347
ADAS11	0.691199	RAVLT_perc_forgetting	0.561252
FAQREM	0.657881	FAQMEAL	0.550885
ADAS_Q1	0.656050		

CDR-SB = CDR sum of boxes, MMSETOTAL = total score of MMSE, FAQTOTAL = total score of FAQ, ADASMOD = total score of ADAS-Cog including Q4 and Q14, ADAS_Q4 = ADAS delayed word recall, FAQTOTAL = total score of FAQ, ADAS11 = total score of ADAS-Cog excluding Q4 and Q14, FAQREM = FAQ remember appointments, ADAS_Q1 = ADAS word recall, FAQFORM = complete forms, RAVLT_immediate = RAVLT immediate recall, FAQFINAN = FAQ manage finance, FAQTRAVL = FAQ travel out of the neighborhood, FAQSHOP = FAQ shop, RAVLT_perc_forgetting = RAVLT Percent Forgetting, FAQMEAL = prepare a balanced meal.

4.2. Implementation

To determine the optimal parameters of the DSSL framework, we use 5-fold cross-validation, i.e., the dataset is randomly divided into 5 folds. Each time, one fold is selected for testing and the remaining 4 folds are used for training. DSSL uses the adam optimizer to optimize the model parameters. As with FixMatch [27], we use an exponential moving average of the parameters with a decay of 0.999 to update the model instead of the decay learning rate. This allows the model to converge more smoothly at a higher number

of iterations and improves the accuracy of the final prediction results [31]. Since we consider supervised loss and consistency loss to be equally important, we set the consistency regularization coefficient λ to 1.

In our implementation, the confidence threshold τ in the DSSL loss function plays a key role in the classification accuracy. To determine the optimal value of τ, we conduct experiments in which τ is varied from 0 to 0.99. To better understand the role of confidence threshold in DSSL, we refer to two measures proposed in the FixMatch approach: impurity rate (the prediction error rate of samples exceeding the threshold) and passing rate (the number of instances above the threshold as a percentage of the total), calculated as follows:

$$\text{impurity rate} = \frac{\sum_{d=1}^{D} \mathbb{I}(\max(q_1) \geq \tau)\mathbb{I}(y_d \neq \hat{q}_1)}{\sum_{d=1}^{D} \mathbb{I}(\max(q_1) \geq \tau)}, \tag{8}$$

$$\text{mask rate} = \frac{1}{\mu_D}\mathbb{I}(\max(q_1) \geq \tau). \tag{9}$$

Table 3 shows the quantity and quality of pseudo-labels and the DSSL classification accuracy at different τ in the 60-label case. From the results, we can see that there is a positive correlation between these two indicators, i.e., when the sample pass rate increases, the impurity rate also increases, which is in line with our expectation. Next, to determine the optimal value of the difference regularization coefficient β, we report the accuracy scores for multiple selected values of this parameter at 60 labels in Figure 3a. It can be seen that the proposed method achieves high prediction accuracy (over 82%) for different values of β, where the highest accuracy is obtained for β = 2. We also experiment with the performance variation of DSSL when trained using different training set sizes. In this experiment, we keep the number of samples with labels below 40% of the number of samples in the training set. As can be seen in Figure 3b, the performance of DSSL gradually improves as the training data increases and plateaus after the size of the training data exceeds 500.

Table 3. Passing rate, impurity rate, and accuracy of test set for DSSL with different thresholds in the 60-label case.

τ	Passing Rate	Impurity Rate	Accuracy
0.25	100	18.55	82.05
0.5	100	16.48	84.12
0.75	99.67	16.45	85.1
0.85	98.95	16.58	84.24
0.9	97.79	15.33	85.47
0.95	95.31	14.32	85.22
0.97	93.36	13.16	85.47
0.99	87.64	10.88	85.22

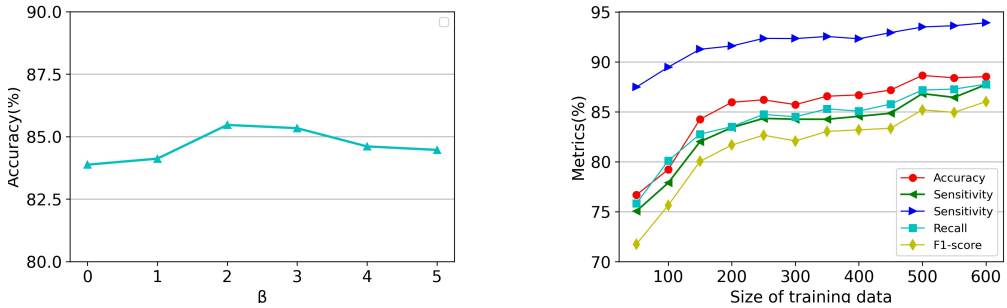


Figure 3. Illustrating (**a**) classification accuracy of the proposed method on different values of β in the 60-label case and (**b**) classification performance of the proposed method on different sizes of training data.

4.3. Results of Disease Classification

To evaluate the performance of the SSL method, five evaluation metrics are chosen: Accuracy, Sensitivity, Specificity, Recall, and F1-score. The true positive (TP), true negative (TN), false positive (FP), and false negative (FN) rates are each related to these factors. The definitions of these evaluation measures are provided below:

$$\text{Accuracy (ACC)} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}, \tag{10}$$

$$\text{Sensitivity (SEN)} = \frac{\text{TP}}{\text{TP+FP}}, \tag{11}$$

$$\text{Specificity (SPE)} = \frac{\text{TN}}{\text{TN+FP}}, \tag{12}$$

$$\text{Recall (REC)} = \frac{\text{TP}}{\text{TP+FN}}, \tag{13}$$

$$\text{F1-score (F1)} = \frac{\text{2TP}}{\text{2TP+FP+FN}}. \tag{14}$$

To compare the effect of different labeled sample sizes in the training set on the classification performance, our experiments are designed with two labeled sample sizes: 60 labeled and 120 labeled. It should be noted that the rest of the training data are unlabeled samples. In the proposed model, the test data achieved an accuracy of 85.47% with 60-label training and 88.40% with 120-label training. Figure 4 shows the prediction results for the test set samples, where the boxes indicate the actual labels of the samples and the dots indicate the prediction results of the samples by DSSL. T-distributed stochastic neighbor embedding (t-SNE) can reduce high-dimensional data to two or three dimensions for data visualization. As shown in the figure, most of the sample points predicted by DSSL fall correctly in the boxes of the authentic samples.

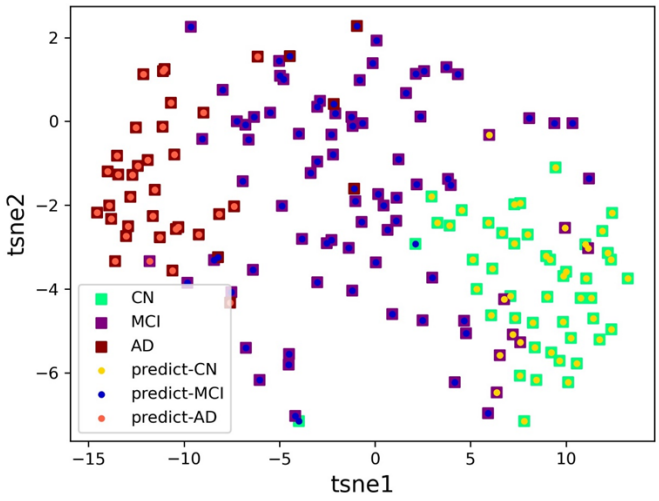


Figure 4. The two-dimensional t-SNE plot of the prediction results of DSSL for the test set samples.

The architecture of the two encoders in the DSSL model significantly impacts the results of the semi-supervised experiments. Figure 2 depicts the internal structure of the encoders. We experimentally test the effect of changing the encoder structure on the classification performance, especially when Encoder1 and Encoder2 have the same structure. The changes to the encoder are mainly focused on the pooling layer, applying max pooling and average pooling. Table 4 compares the classification results of the DSSL framework applying different combinations of encoders. It can be seen that the DSSL with different structures of Encoder1 and Encoder2 has better classification results.

Table 4. Test set evaluation results for DSSL with different encoders in the 60-label case.

	Pooling Layer	ACC (%)	SEN (%)	SPE (%)	REC (%)	F1 (%)
60-label	Max + Max	84.49	82.42	83.05	91.29	80.46
	Avg + Avg	84.00	82.05	83.00	91.19	80.42
	Max + Avg	85.47	83.77	84.14	91.82	81.92
120-label	Max + Max	88.15	85.89	86.06	93.03	84.60
	Avg + Avg	88.27	86.72	87.37	93.41	85.50
	Max + Avg	88.40	86.99	87.07	93.20	85.53

Max + Max means the pooling layer of Encoder1 is max pooling, and the pooling layer of Encoder2 is also max pooling. The other pooling layers are similar.

4.4. Comparison with Other Methods

We compare our proposed method with other existing semi-supervised methods. The five methods described in Section 2: MixMatch [25], FixMatch [27], SimPLE [34], CCSSL [35], and LaSSL [36] are considered as baseline methods. To fairly compare these methods, we reimplement them using the same deep learning framework (i.e., PyTorch) and model. Considering that the strong augmentation part of the baseline methods is only applicable to image data, we choose mixup [26] as an alternative to RandAugment [28] or CTAugment [29] for data augmentation. Table 5 compares the performance of all baselines and DSSL. We compute the evaluation results for both cases with labeled samples of 60 and 120. All results are averaged for the 5-fold cross-validation. It can be seen that DSSL outperforms all baselines to a large extent, both in the 60-label and 120-label cases. Figure 5 illustrates box plots of the accuracy of the 5-fold cross-validation experiments for the cases of 60 and 120 labels, respectively.

Table 5. The semi-supervised evaluation results of each model for the ADNI database data in the 60-label and 120-label cases.

	Method	ACC (%)	SEN (%)	SPE (%)	REC (%)	F1 (%)	Training Time (Minute)
60-label	MixMatch [25] (2019)	77.29	75.40	88.21	76.73	72.40	1.27
	FixMatch [27] (2020)	81.44	79.01	90.27	80.29	76.90	1.15
	SimPLE [34] (2021)	80.34	77.39	89.66	78.80	75.26	2.53
	CCSSL [35] (2022)	81.07	79.74	89.99	80.05	77.11	2.48
	LaSSL [36] (2022)	79.47	76.06	89.15	78.49	74.26	2.82
	DSSL	85.47	83.77	84.14	91.82	81.92	2.42
120-label	MixMatch [25] (2019)	82.42	80.55	91.13	81.57	78.16	1.24
	FixMatch [27] (2020)	84.49	82.10	91.80	83.71	80.46	1.11
	SimPLE [34] (2021)	84.98	82.94	91.92	83.60	85.15	2.51
	CCSSL [35] (2022)	78.64	76.03	88.83	76.92	73.23	2.42
	LaSSL [36] (2022)	85.10	82.63	91.60	83.66	81.07	2.85
	DSSL	88.40	86.99	87.07	93.20	85.53	2.27

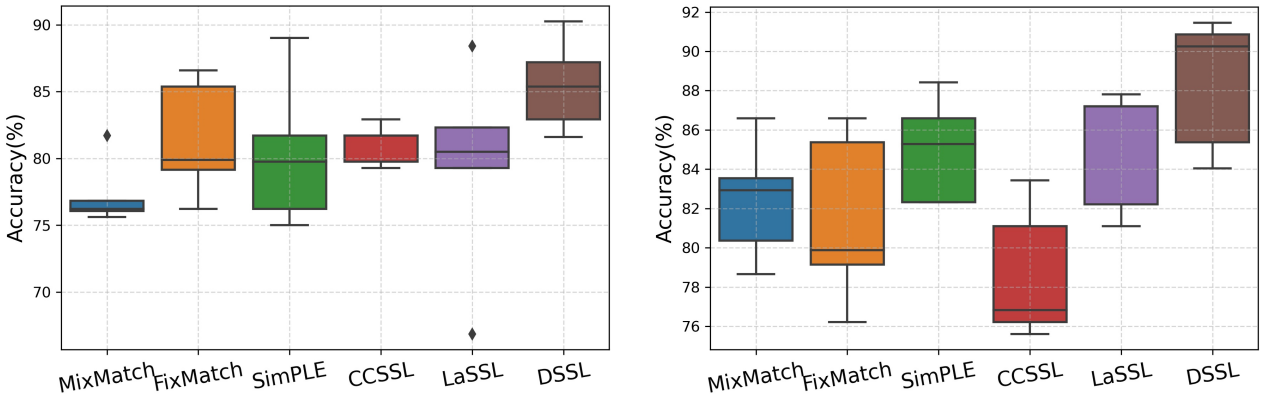


Figure 5. Illustrating (**a**) each model's accuracy in a 5-fold cross-validation experiment in the 60-label case and (**b**) each model's accuracy in a 5-fold cross-validation experiment in the 120-label case.

Although we achieve the best classification results in the 5-fold cross-validation experiments, the selection of different labeled data can seriously affect the classification performance for the SSL algorithm. We randomly select labeled samples from the training set and repeat this process 100 times to obtain 100 division results. We train these 100 divisions sequentially to observe the stability of the algorithm. The variance of the 100 times predictions for the DSSL and each baseline are shown in Table 6. For visualization purposes, we select the three models with the slightest variance in each of the two cases and plot their 100 times results as line graphs, as shown in Figure 6. It can be seen that the variance of DSSL is the lowest in both the 60-label and 120-label cases, which indicates that DSSL is more stable than the other baseline methods. In addition, the variance of the model with 120 labels is generally smaller than that of the case with 60 labels, suggesting that the increase in the number of labeled samples improves the stability of the SSL algorithm.

Table 6. The variance of the results of each model after 100 experiments in the 60 and 120 label cases.

Method	60 Labels	120 Labels
MixMatch [25] (2019)	3.84	2.87
FixMatch [27] (2020)	3.62	2.48
SimPLE [34] (2021)	3.41	2.62
CCSSL [35] (2022)	3.38	3.38
LaSSL [36] (2022)	4.34	3.07
DSSL	2.91	2.30

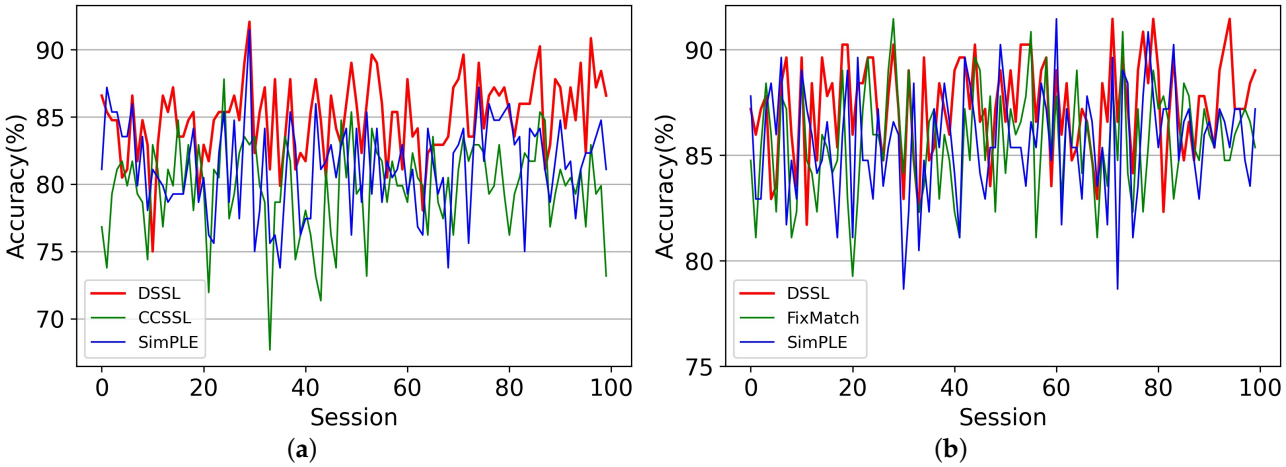


Figure 6. Illustrating (**a**) line graphs of the three methods with a minor variance in the results of 100 experiments in the 60-label case and (**b**) line graphs of the three methods with a minor variance in the results of 100 experiments in the 120-label case.

5. Discussion

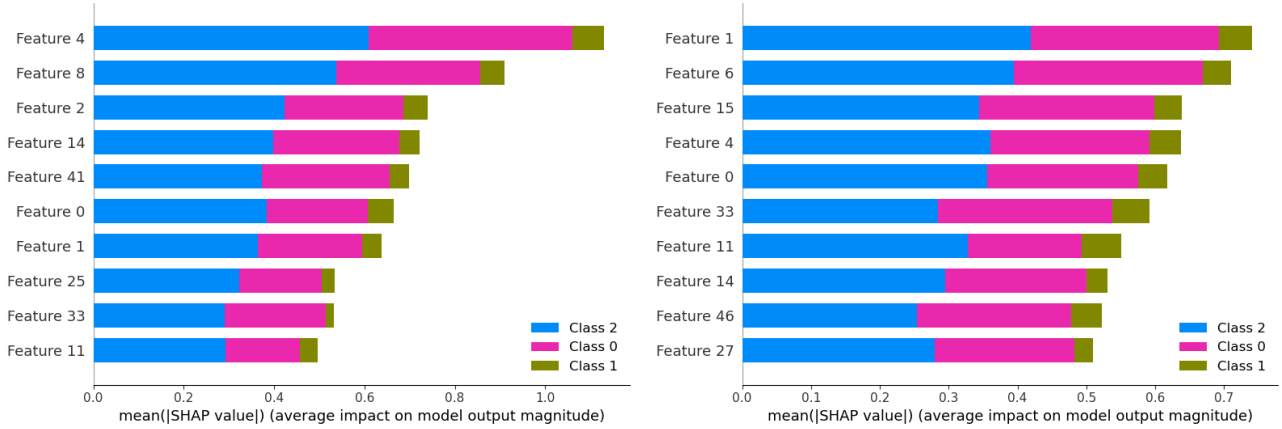
In this study, two encoders are used to learn different features of the sample for predicting different degrees of cognitive impairment: AD, MCI, and NC. With the ADNI neuropsychological dataset and a small number of labels, DSSL achieved an accuracy of 85.47% in the 60-label case and 88.40% in the 120-label case. The comparison results in Table 5 show that our proposed semi-supervised method outperforms the existing semi-supervised methods in terms of accuracy, sensitivity, specificity, recall, and F1-score. The comparison results in Table 6 show that our proposed algorithm is more stable than the existing semi-supervised methods.

Feature selection has an essential role as a precursor to the classification task. PCC is one of the most typical and popular similarity measures. The reason we chose PCC for feature selection is that PCC has the property that shifts in the position and scale of the variable do not cause a change in this coefficient. This property allows the correlation between the neuropsychological test scores after normalization and the diagnosis to be the same as the original values. It helps to improve classification performance while providing physicians with biomarker references for clinical diagnosis. As seen in Table 2, CDR, MMSE, ADAS, and FAQ have strong correlations with the degree of cognitive impairment and their total scores correlate more strongly with the diagnostic outcome compared to the sub-scores.

For computational complexity, Table 5 shows the training time for DSSL and other comparative methods. It can be seen that MixMatch and FixMatch take the shortest time, and our proposed method takes a little longer because it requires updating the parameters of both encoders. All the experiments are performed on a PC with 2.0 GHz, 8-core CPU, and 8 GB RAM on a Windows 10 operating system. Overall, all experiments applying

neuropsychological test data for training require less than 3 min, which demonstrates the usability of the proposed method for clinical applications.

The confidence threshold seriously affects the quality of the generated pseudo-labels. Although we find the optimal value of τ in Table 3 through extensive experiments, this is time-consuming, and there is no guarantee that the set threshold will work for each data division. The question to be considered is how to weigh the number of unlabeled samples exceeding the threshold and the consistency rate of pseudo-labels with valid labels. Perhaps automatic learning of this parameter using neural networks would be a better approach. This is also how the model will be improved in the future. DSSL diagnoses AD by using two encoders to learn different features of the sample. To facilitate the visualization of the learned feature representations, we use Shapley values [45] to quantify the importance of features in the algorithm predictions. We sort each feature in the feature representation by its contribution to the model output. Figure 7 shows the top 10 features with the highest contribution in each of the two feature representations, where class 2, 1, and 0 denote AD, MCI, and NC, respectively. As seen in the figure, all features have higher impact scores for AD and NC, while MCI as an intermediate stage is weakly influenced by these features. Moreover, the same features in the two feature representations do not contribute consistently to the algorithm output, which indicates that the two encoders in the proposed method do learn different feature representations. However, there are still limitations in the medical interpretation of these features in correlation with disease pathology. Using expert knowledge to correct the learned feature representation may yield better classification results.



(**a**) (**b**)

Figure 7. Illustrating (**a**) 10 features with the highest contribution in the feature representation learned by Encoder1 and (**b**) 10 features with the highest contribution in the feature representation learned by Encoder2.

6. Conclusions

To accurately determine AD severity with easily available features and a limited number of labels, we propose a novel semi-supervised framework, namely DSSL. We first collect 64 itemized scores from seven neuropsychological tests and use PCC for feature selection. A total of 15 features most relevant to the diagnostic results are selected to serve as input for subsequent semi-supervised experiments. Then, the DSSL model is proposed to better screen for AD and MCI using only neuropsychological tests and a small amount of labeling, without the need for costly PET and MRI, etc. The model uses two encoders and difference regularization to learn two different features from the same sample. Finally, we empirically demonstrate the validity and stability of our method through extensive comparisons with a large number of existing semi-supervised algorithms in terms of accuracy, sensitivity, specificity, recall, F1-score, and variance.

In the future, the proposed algorithm will be applied to other AD biomarkers of multimodal data such as MRI, PET, etc. It would be a promising research direction to use other deep neural network models as encoders to extract potential feature representations of the data and to explore medical interpretations of the relationship between feature representations and disease pathology.

Author Contributions: Conceptualization and methodology, Y.W.; formal analysis and investigation, X.G.; data curation, M.Z. and C.G.; writing—original draft preparation, X.G. and C.G.; visualization, W.H.; project administration, Y.W.; funding acquisition, C.G. and L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (No. 62072212), the General Program of the National Natural Science Foundation of China (No. 82071442), the Development Project of Jilin Province of China (No. 20220508125RC), the Jilin Provincial Department of Finance (No. JLSWSRCZX2021-004), the Natural Science Foundation of Jilin Province (No. 20210101273JC), Bethune Project of Jilin University (No. 2020B47), and the Science and Technology Achievement Transformation Fund of the First Hospital of Jilin University (No. JDYY2021-A0010).

Institutional Review Board Statement: Ethical review and approval were waived for this study, due to all research data are from open-source datasets.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: We using open datasets to tested our method. The ADNI dataset can be found in https://adni.loni.usc.edu/ (accessed on 18 December 2021).

Acknowledgments: 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, and the National Institute of Biomedical Imaging and Bioengineering.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1

We chose 64 itemized scores from the 7 neuropsychological tests whose corresponding functions are shown in Table A1.

Table A1. Neuropsychological tests used in this study.

Neuropsychological Tests	
AQAS-cog	Total score of MMSE (1)
Q1: Word recall	CDR
Q2: Commands	CDR Sum of boxes
Q3: Constructional praxis	RAVLT
Q4: Delayed word recall	RAVLT Immediate recall
Q5: Object naming	RAVLT Learning
Q6: Ideational rraxis	RAVLT Forgetting delay
Q7: Orientation	RAVLT Percent forgetting
Q8: Word recognition	FAQ
Q9: Clarity of language	Q1. Manage finance
Q10: Comprehension	Q2. Complete forms
Q11: Word finding	Q3. Shop
Q12: Remembering test instructions	Q4. Perform games of skill or hobbies

Table A1. *Cont.*

Neuropsychological Tests	
Q14: Number cancellations	Q5. Prepare hot beverages
Total score of ADAS11	Q6. Prepare a balanced meal
Total score of ADASMOD	Q7. Follow current events
MMSE	Q8. Attend to TV, books, or magazines
Orientation to place (5)	Q9. Remember appointments
Orientation to time (5)	Q10. Travel out of the neighborhood
Registration (3)	Total score of FAQ
Attention and concentration (5)	NPIQ
Recall (3)	Total score of NPIQ
Language (8)	GDS
Visual construction (1)	Total score of GDS

Appendix .1

ADAS-Cog— ADAS-Cog is a screening instrument that provides a specific assessment of the severity of cognitive and non-cognitive behavioral impairments. Thirteen tests were used to assess memory (word recall and word recognition), language (naming and comprehension), reasoning (commands), orientation, constructional praxis (copying geometric designs), and ideational praxis (putting the letter in the envelope). The advantage of the ADAS-Cog over other scales is that its scores quantify the clinical and impressionistic aspects of the patient and objectively define cognitive characteristics.

MMSE—MMSE is a comprehensive screening tool commonly used in the clinical diagnosis of cognitive impairment. It consists of 30 items assessing 7 main areas: orientation to place, orientation to time, registration (repetition of words), attention and concentration (serial subtraction), recall (recall of the previous words), language (naming, writing, and comprehension), and visual construction (design copy). The total score between 0 and 30 indicates different degrees of cognitive impairment.

CDR—Washington University in St. Louis developed CDR to determine longitudinal changes in aging and dementia. It measures global cognitive impairment and evaluates domains including memory, orientation, decision-making and problem-solving, family life and personal preferences, and independent living abilities. The CDR combines the ratings of the six functions into a total score, with a more accurate measure of change by the sum of the boxes.

RAVLT—RAVLT is an anterograde verbal episodic memory test widely used in clinical practice. Fifteen irrelevant words are given verbally at a rate of one per second, and subjects are asked to recall these words immediately. The process has been performed a total of five times. After a 20-min delay filled with irrelevant tests, subjects were asked to review the initial list of 15 words. Finally, a yes/no recognition test was performed, which consisted of 30 words, including the original 15 words and 15 randomly inserted words.

FAQ—FAQ rates the subject's ability to perform daily activities based on interviews with partners, which assesses the patient's physical, mental, and social role function completion and factors that affect daily performance. FAQ uses 10 questions to evaluate the above indicators, with a total score of 30. Subjects are considered to have social activity dysfunction when the total score is greater than nine.

NPIQ—NPI is a validated, multi-item, reliable tool for assessing the psychopathology of patients with AD. The assessment of NPI is based on interviews with caregivers or eligible partners, which are relatively brief (15 min). The NPIQ is a short version of the NPI, which only screens questions and severity ratings for each domain. The highest score is 36.

GDS—GDS is a self-report assessment that is used to diagnose the degree of depression in older adults. This scale, comprising thirty entries, assesses the following areas: depressed mood, irritability, and reduced mobility. In addition, subjects are asked to answer yes or no for each entry of the GDS.

References

1. Chehrehnegar, N.; Nejati, V.; Shati, M.; Rashedi, V.; Lotfi, M.; Adelirad, F.; Foroughan, M. Early detection of cognitive disturbances in mild cognitive impairment: A systematic review of observational studies. *Psychogeriatrics* **2019**, *20*, 212–228. [CrossRef] [PubMed]
2. Selkoe, D.J. Alzheimer's disease. *Cold Spring Harb. Perspect. Biol.* **2011**, *3* , a004457. [CrossRef] [PubMed]
3. Jack, C.R.; Bennett, D.A.; Blennow, K.; Carrillo, M.C.; Feldman, H.H.; Frisoni, G.B.; Hampel, H.; Jagust, W.J.; Johnson, K.A.; Knopman, D.S.; et al. A/T/N: An unbiased descriptive classification scheme for Alzheimer disease biomarkers. *Neurology* **2016**, *87*, 539–547. [CrossRef] [PubMed]
4. Shi, Y.; Wang, Z.; Chen, P.; Cheng, P.; Zhao, K.; Zhang, H.; Shu, H.; Gu, L.; Gao, L.; Wang, Q.; et al. Episodic Memory–Related Imaging Features as Valuable Biomarkers for the Diagnosis of Alzheimer's Disease: A Multicenter Study Based on Machine Learning. *Biol. Psychiatry: Cogn. Neurosci. Neuroimaging* **2020**, *8*, 171–180. [CrossRef] [PubMed]
5. Chen, Z.S.; Kulkarni, P.; Galatzer-Levy, I.R.; Bigio, B.; Nasca, C.; Zhang, Y. Modern views of machine learning for precision psychiatry. *Patterns* **2022**, *3*, 100602. [CrossRef]
6. Kruthika, K.; Maheshappa, H. CBIR system using Capsule Networks and 3D CNN for Alzheimer's disease diagnosis. *Inform. Med. Unlocked* **2019**, *14*, 59–68. [CrossRef]
7. Zhang, Y.; Wang, S.; Xia, K.; Jiang, Y.; Qian, P. Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion. *Inf. Fusion* **2021**, *66*, 170–183. [CrossRef]
8. Fan, Z.; Xu, F.; Qi, X.; Li, C.; Yao, L. Classification of Alzheimer's disease based on brain MRI and machine learning. *Neural Comput. Appl.* **2020**, *32*, 1927–1936. [CrossRef]
9. Turkson, R.E.; Qu, H.; Mawuli, C.B.; Eghan, M.J. Classification of Alzheimer's disease using deep convolutional spiking neural network. *Neural Process. Lett.* **2021**, *53*, 2649–2663. [CrossRef]
10. Amini, M.; Pedram, M.M.; Moradi, A.; Ouchani, M. Diagnosis of Alzheimer's disease severity with fMRI images using robust multitask feature extraction method and convolutional neural network (CNN). *Comput. Math. Methods Med.* **2021**, *2021*, 5514839. [CrossRef]
11. Zhou, H.; He, L.; Zhang, Y.; Shen, L.; Chen, B. Interpretable Graph Convolutional Network Of Multi-Modality Brain Imaging For Alzheimer's Disease Diagnosis. In Proceedings of the 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), Kolkata, India, 28–31 March 2022; pp. 1–5.
12. Zhou, H.; Zhang, Y.; Chen, B.Y.; Shen, L.; He, L. Sparse Interpretation of Graph Convolutional Networks for Multi-modal Diagnosis of Alzheimer's Disease. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Singapore, 18–22 September 2022.
13. Saleem, T.J.; Zahra, S.R.; Wu, F.; Alwakeel, A.; Alwakeel, M.; Jeribi, F.; Hijji, M. Deep Learning-Based Diagnosis of Alzheimer's Disease. *J. Pers. Med.* **2022**, *12*, 815. [CrossRef]
14. Van Engelen, J.E.; Hoos, H.H. A survey on semi-supervised learning. *Mach. Learn.* **2020**, *109*, 373–440. [CrossRef]
15. Seo, E.H. Neuropsychological assessment of dementia and cognitive disorders. *Korean Neuropsychiatr. Assoc.* **2018**, *57*, 2–11. [CrossRef]
16. Ewers, M.; Walsh, C.; Trojanowski, J.Q.; Shaw, L.M.; Petersen, R.C.; Jack, C.R., Jr.; Feldman, H.H.; Bokde, A.L.; Alexander, G.E.; Scheltens, P.; et al. Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. *Neurobiol. Aging* **2012**, *33*, 1203–1214. [CrossRef]
17. Grassi, M.; Perna, G.; Caldirola, D.; Schruers, K.; Duara, R.; Loewenstein, D.A. A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion in individuals with mild and premild cognitive impairment. *J. Alzheimer's Dis.* **2018**, *61*, 1555–1573. [CrossRef]
18. Battista, P.; Salvatore, C.; Castiglioni, I. Optimizing neuropsychological assessments for cognitive, behavioral, and functional impairment classification: A machine learning study. *Behav. Neurol.* **2017**, *2017*, 1850909. [CrossRef]
19. Battista, P.; Salvatore, C.; Berlingeri, M.; Cerasa, A.; Castiglioni, I. Artificial intelligence and neuropsychological measures: The case of Alzheimer's disease. *Neurosci. Biobehav. Rev.* **2020**, *114*, 211–228. [CrossRef]
20. Huang, Z.F.; Li, F.; Wang, Z.; Wang, Z. Interpretability of Deep Learning. *Int. J. Future Comput. Commun.* **2022**, *11*. [CrossRef]
21. Bachman, P.; Alsharif, O.; Precup, D. Learning with pseudo-ensembles. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)* ; Neural Information Processing Systems Foundation, Inc. (NeurIPS): La Jolla, CA, USA, 2014; Volume 27.
22. Sajjadi, M.; Javanmardi, M.; Tasdizen, T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*; Neural Information Processing Systems Foundation, Inc. (NeurIPS): La Jolla, CA, USA, 2016; Volume 29.
23. Laine, S.; Aila, T. Temporal ensembling for semi-supervised learning. *arXiv* **2016**, arXiv:1610.02242.
24. Park, S.; Park, J.; Shin, S.J.; Moon, I.C. Adversarial dropout for supervised and semi-supervised learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
25. Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; Raffel, C.A. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*; Neural Information Processing Systems Foundation, Inc. (NeurIPS): La Jolla, CA, USA, 2019; Volume 32.
26. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.

27. Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.A.; Cubuk, E.D.; Kurakin, A.; Li, C.L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 596–608.
28. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2019; pp. 3008–3017.
29. Berthelot, D.; Carlini, N.; Cubuk, E.D.; Kurakin, A.; Sohn, K.; Zhang, H.; Raffel, C. ReMixMatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia 26–30 April 2020.
30. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
31. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Neural Information Processing Systems Foundation, Inc. (NeurIPS): La Jolla, CA, USA, 2017; Volume 30.
32. Lee, D.H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Proceedings of the Workshop on Challenges in Representation Learning, ICML, Atlanta, USA, 16–21 June 2013; Volume 3, p. 896.
33. Iscen, A.; Tolias, G.; Avrithis, Y.; Chum, O. Label Propagation for Deep Semi-Supervised Learning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5065–5074.
34. Hu, Z.; Yang, Z.; Hu, X.; Nevatia, R. SimPLE: Similar Pseudo Label Exploitation for Semi-Supervised Classification. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 15094–15103.
35. Yang, F.; Wu, K.; Zhang, S.; Jiang, G.; Liu, Y.; Zheng, F.; Zhang, W.; Wang, C.; Zeng, L. Class-Aware Contrastive Semi-Supervised Learning. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14401–14410.
36. Zhao, Z.; Zhou, L.; Wang, L.; Shi, Y.; Gao, Y. LaSSL: Label-Guided Self-Training for Semi-supervised Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022.
37. Kueper, J.K.; Speechley, M.; Montero-Odasso, M. The Alzheimer's disease assessment scale–cognitive subscale (ADAS-Cog): Modifications and responsiveness in pre-dementia populations. a narrative review. *J. Alzheimer's Dis.* **2018**, *63*, 423–444. [CrossRef]
38. Folstein, M.F.; Folstein, S.E.; McHugh, P.R. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **1975**, *12*, 189–198. [CrossRef] [PubMed]
39. Morris, J.C. The clinical dementia rating (cdr): Current version and. *Young* **1991**, *41*, 1588–1592.
40. Rey, A. *L'examen Clinique en Psychologie*; University Press of France: Paris, France, 1958.
41. Pfeffer, R.I.; Kurosaki, T.T.; Harrah, C., Jr.; Chance, J.M.; Filos, S. Measurement of functional activities in older adults in the community. *J. Gerontol.* **1982**, *37*, 323–329. [CrossRef]
42. Kaufer, D.I.; Cummings, J.L.; Ketchel, P.; Smith, V.; MacMillan, A.; Shelley, T.; Lopez, O.L.; DeKosky, S.T. Validation of the NPI-Q, a brief clinical form of the Neuropsychiatric Inventory. *J. Neuropsychiatry Clin. Neurosci.* **2000**, *12*, 233–239. [CrossRef]
43. Sheikh, J.I.; Yesavage, J.A. Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. *Clin. Gerontol. J. Aging Ment. Health* **1986**, *5*, 165–173.
44. Schober, P.; Boer, C.; Schwarte, L.A. Correlation coefficients: Appropriate use and interpretation. *Anesth. Analg.* **2018**, *126*, 1763–1768. [CrossRef]
45. Ghorbani, A.; Zou, J. Data shapley: Equitable valuation of data for machine learning. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 2242–2251.