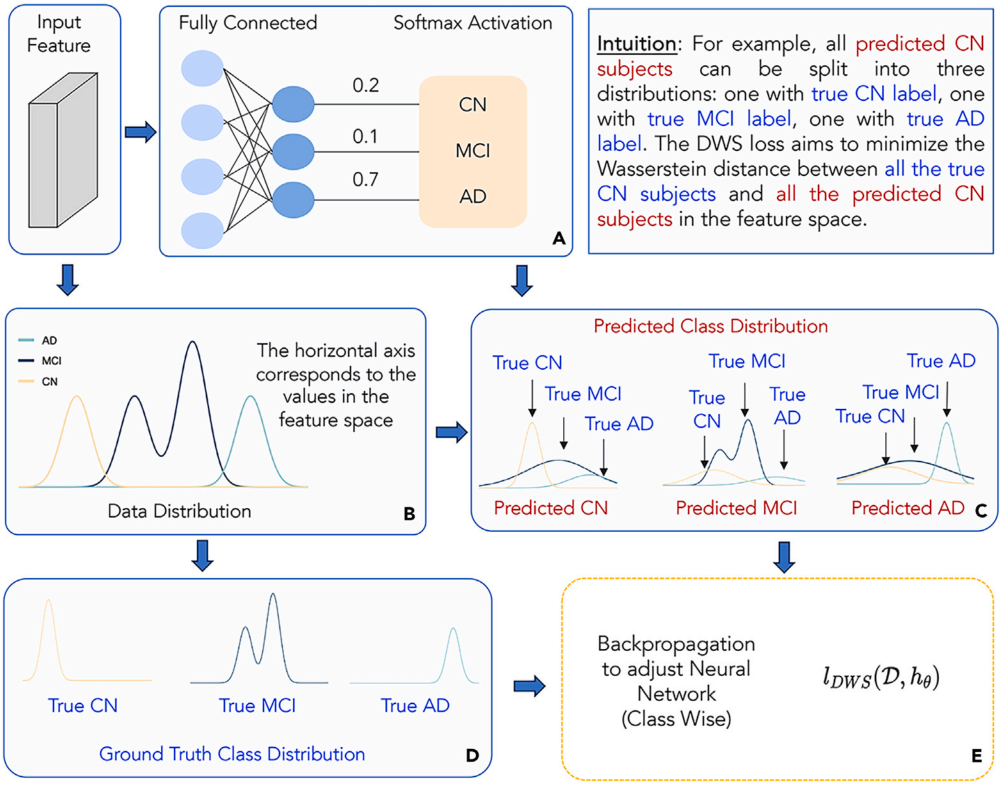
iScience

CellPress
OPEN ACCESS

Article

Distance-weighted Sinkhorn loss for Alzheimer's disease classification



Intuition: For example, all predicted CN subjects can be split into three distributions: one with true CN label, one with true MCI label, one with true AD label. The DWS loss aims to minimize the Wasserstein distance between all the true CN subjects and all the predicted CN subjects in the feature space.

The workflow of the DWS Framework.
(a) The Neural Network maps the input features to the prediction probability and classification label.
(b) The ground truth Data Distribution.
(c) Using the prediction probability as a weight to obtain predicted class distribution.
(d) Separated ground truth Data Distribution into class-wise ground truth Data Distribution.
(e) Calculate the DWS loss and backpropagation to adjust the Neural Network.

Zexuan Wang, Qipeng Zhan, Boning Tong, ..., Christos Davatzikos, Li Shen, Alzheimer's Disease Neuroimaging Initiative

li.shen@pennmedicine.upenn.edu

Highlights

Our loss function aims to learn the data-wise label distribution

Our loss function is theoretically based on Wasserstein distance

Our method better classifies Alzheimer's disease stages

The code is publicly available on GitHub

Wang et al., iScience *27*, 109212
March 15, 2024 © 2024 The Author(s).
https://doi.org/10.1016/j.isci.2024.109212

iScience

CellPress OPEN ACCESS

Article

Distance-weighted Sinkhorn loss for Alzheimer's disease classification

Zexuan Wang,[1,5] Qipeng Zhan,[1,5] Boning Tong,[1,5] Shu Yang,[1] Bojian Hou,[1] Heng Huang,[2] Andrew J. Saykin,[3] Paul M. Thompson,[4] Christos Davatzikos,[1] Li Shen,[1,6,*] and for the Alzheimer's Disease Neuroimaging Initiative

SUMMARY

Traditional loss functions such as cross-entropy loss often quantify the penalty for each mis-classified training sample without adequately considering its distance from the ground truth class distribution in the feature space. Intuitively, the larger this distance is, the higher the penalty should be. With this observation, we propose a penalty called distance-weighted Sinkhorn (DWS) loss. For each mis-classified training sample (with predicted label *A* and true label *B*), its contribution to the DWS loss positively correlates to the distance the training sample needs to travel to reach the ground truth distribution of all the *A* samples. We apply the DWS framework with a neural network to classify different stages of Alzheimer's disease. Our empirical results demonstrate that the DWS framework outperforms the traditional neural network loss functions and is comparable or better to traditional machine learning methods, highlighting its potential in biomedical informatics and data science.

INTRODUCTION

Alzheimer's disease (AD) is a progressive and degenerative condition that affects the brain's neurons and causes memory loss, cognitive decline, and behavioral issues.[1–3] The increasing prevalence of AD and its devastating impact on individuals and societies has necessitated the development of strategies for early and accurate diagnosis. Since there are currently no viable therapies for AD, it is thought that the best way to slow its progression is to start with an early diagnosis. Traditionally, AD diagnosis has relied on clinical evaluations and cognitive tests. However, these approaches often fail to detect the disease until it has significantly progressed, making effective intervention more challenging. Therefore, there is a growing interest in exploring alternative diagnostic methods.

To achieve this, researchers from a variety of fields have dedicated their work to comprehending the mechanisms underlying these diseases and identifying pathological biomarkers for the diagnosis or prognosis of AD and/or mild cognitive impairment (MCI, a prodromal stage of AD), by examining various neuroimaging modalities, such as magnetic resonance imaging (MRI),[4] positron emission tomography (PET),[5] functional MRI (fMRI),[6] etc.

Neuroimaging techniques have made significant contributions to our understanding of the brain. Over the past few decades, there has been a growing concern regarding the need for breakthroughs in efficiently analyzing and interpreting observed data. Machine learning (ML), which can handle highly dimensional and complicated data, has recently become a potent tool for disease classification and prediction.[7–9]

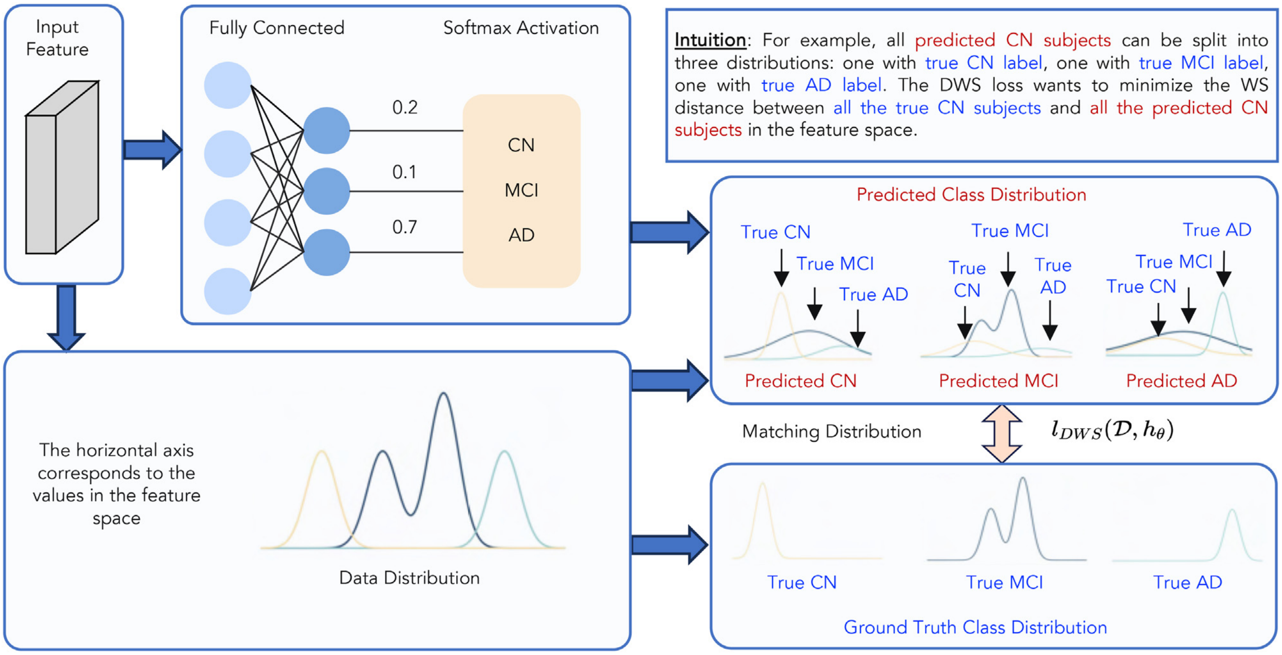
Feedforward deep neural network (DNN) has been employed in various research studies for the purpose of AD and MCI diagnosis. For example, Ning proposed a model that takes the image and genetic data to classify AD occurrence and identify the most crucial AD risk factors.[10] Magni presented a support vector machine (SVM)-based automated technique of whole-brain anatomical MRI image to distinguish between people with AD and older control participants.[11] By introducing a group lasso penalty to induce structure sparsity, Sun improved the traditional SVM-based model, comparable to or better than the state-of-the-art methods.[12]

Wasserstein distance is defined between two probability distributions on a given metric space. It has been applied to solve numerous problems, including generative network,[13] barycenter estimation,[14] and multi-class classification.[15] Besides the field of deep learning, the optimal transport (OT) theory[16] has also been applied to computational geometry,[17] surface modeling,[18] and casual inference.[19]

In multi-class classification models, traditional loss functions such as cross-entropy loss often quantify the penalty for a mis-classified training sample without adequately considering its distance from the target ground truth class distribution in the feature space. Intuitively, the larger this distance is, the higher the penalty should be. With this observation, In this paper, for each class label *A*, the proposed

[1] University of Pennsylvania, B301 Richards Building, 3700 Hamilton Walk, Philadelphia, PA 19104, USA
[2] University of Maryland, College Park, 8125 Paint Branch Drive, College Park, MD 20742, USA
[3] Indiana University, 355 West 16th Street, Indianapolis, IN 46202, USA
[4] University of Southern California, 4676 Admiralty Way, Marina Del Rey, CA 90292, USA
[5] These authors contributed equally
[6] Lead contact
*Correspondence: li.shen@pennmedicine.upenn.edu

https://doi.org/10.1016/j.isci.2024.109212



Intuition: For example, all predicted CN subjects can be split into three distributions: one with true CN label, one with true MCI label, one with true AD label. The DWS loss wants to minimize the WS distance between all the true CN subjects and all the predicted CN subjects in the feature space.

Figure 1. Schematic design of the proposed algorithm using the distance-weighted Sinkhorn (DWS) loss function

In our proposed model, we first send the data into the fully connected deep neural network. After applying the softmax activation functions, the row logits turn into probability. The probability is used as the weight for the predicted distribution. To be specific, we have three probabilities for one instance and using the whole data will give as three weighted distributions. The Sinkhorn algorithm is then used to match the distribution between the predicted distribution and the ground truth distribution for each class label (i.e., CN, MCI, or AD). Finally, backpropagation is taken to adjust the neural network. In this flowchart, we assumed that the model had correctly predicted each instance. Therefore, the dominated distribution of each class is similar to its ground truth distribution. For simplicity, we assume the features live in a one dimensional continuous space.

distance-weighted Sinkhorn (DWS) loss is explicitly designed to match the distributions between all the samples with predicted label *A* and all the samples with true label *A*; see Figure 1 for a schematic design. The Wasserstein distance and OT theory is used to capture the difference between distributions.

This strategy could be considered as the data-wise label distribution learning (LDL) problem. This loss function is fundamentally based on the OT theory as it could capture the underlying data space's geometric details. On the other hand, the LDL[20] aims to minimize the metric of the model output and the ground truth labels, trying to find the best label trending of each instance. We implement a neural network with the DWS loss and apply it to a diagnosis task on classifying AD, MCI, and cognitively normal (CN) subjects using F-fluorodeoxyglucose positron emission tomography (FDG-PET) imaging data capturing glucose metabolism. The data used in this study are obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) Database: https://adni.loni.usc.edu. The results are robust in binary and multi-class classification compared to other loss functions and comparable across methods besides DNNs. Our proposed DWS loss, to the best of our knowledge, is the first one that considers the data-wise distribution of the output model.

The rest of this paper is structured as follows: In the method details section, we introduce our problem formulation and elaborate on the proposed model in detail. The results section compares the DWS loss function with commonly used loss functions within the DNN framework and traditional ML methods (e.g., SVM). We also generate the feature importance map from the DWS loss model, ranking the importance in terms of their attribution to model decision, thus highlighting key cognitive markers. The paper concludes with a discussion of the current scope of our work, its limitations, and potential direction for future works.

RESULTS

In this section, we evaluate the effectiveness of our proposed DWS loss through conducting an empirical study on classifying CN, MCI, and AD participants using the ADNI FDG-PET dataset. This dataset contains 789 participants with 116 features, including 264 CN, 390 MCI, and 135

Table 1. ADNI participant age distribution

Age range	Less than 55	55–60	61–65	66–70	71–75	76–80	81–85	86–90	91–95
Number of participants	0	30	82	168	204	174	102	25	4

Table 2. ADNI participant years of education

Year range	0–7	8–12	13–15	16 and above
Number of participants	1	117	150	521

Table 3. ADNI participant Hispanic/Latino ethnicity

Ethnicity	Hispanic	Latino	Unknown
Number of participants	771	15	3

Table 4. ADNI participant race categories

Race	White	Asian	Am Indian/Alaskan	More than one
Number of participants	778	9	1	1

AD subjects. Participant demographic information is detailed in Table 1, 2, 3 and 4. We examine the ability of DWS on three binary classification tasks and one multi-class classification task. The three binary classification tasks are CN vs. AD, CN vs. MCI, and AD vs. MCI. The multi-class (three) task is CN vs. AD vs. MCI.

Under a DNN framework, we compare the proposed DWS to four loss functions: Binary Cross Entropy loss, Binary Cross Entropy loss with Logits loss, Hinge loss, and Focal loss. For a fair comparison, we choose the same neural network for these five losses. A feedforward neural network with two fully connected hidden layers is used. The neurons of the layer start from 256 and decrease to 128 at the second hidden layer. All the networks are trained to use the L^2 norm of 10^{-3}, Adaptive Moment Estimation (ADAM) optimizer with a batch size 24. The initial learning rate is 0.001 and will be decreased by a tenth in the validation loss plateau. We used the Pytorch package SampleLoss from the Geomloss to compute the Wasserstein distance.

In addition, we compare our model with several widely used classification models outside the DNN realm, including SVM, logistic regression, gradient boosting, its variant, random forest, etc. The results of these models are obtained from an automated machine learning (AutoML) pipeline, STREAMLINE,[21] which provides the models' best practice with the Bayesian optimization hyperparameter tuning. We follow the same data preprocessing and split settings for a fair comparison. The balance accuracy and accuracy are used as the evaluation metric. Twenty test runs were performed, and the average performance with standard deviation are reported. We run all of the experiments on a system with an x86 64 architecture, Intel(R) Xeon(R) CPU operating at 2.20 GHz, and 12 GB of RAM.

Comparative study with other loss functions under the same DNN framework

To evaluate the efficacy of the proposed loss function, we first compare the DWS loss with CE, BCELogit, Hinge, and Focal losses, and evaluate the performance using accuracy and balanced accuracy. In our study, we employ accuracy as a primary metric to evaluate the performance of our models. Accuracy measures the proportion of correct predictions, making it an intuitive and straightforward way to understand how well our models perform. In addition to accuracy, we also measure balanced accuracy, mainly due to the presence of imbalanced classes in our dataset. Balanced accuracy is the average of the proportion corrects of each class individually, which is especially important when the classes are imbalanced as it gives equal weight to the predictive performance of each class. This ensures that our models perform well in the

Table 5. Accuracy and balanced accuracy results compared to BCE, BCELogit, Hinge, and Focal loss

Metric	Loss function	CN vs. AD	CN vs. MCI	MCI vs. AD	CN vs. MCI vs. AD
Accuracy	BCE	0.868 ± 0.0490	0.597 ± 0.0332	0.819 ± 0.0471	0.548 ± 0.0398
Accuracy	BCELogit	0.870 ± 0.0411	0.593 ± 0.0398	0.820 ± 0.0331	0.564 ± 0.0384
Accuracy	Hinge	0.860 ± 0.0381	0.607 ± 0.0317	0.815 ± 0.0423	0.539 ± 0.0612
Accuracy	Focal	0.871 ± 0.0360	0.605 ± 0.0374	0.819 ± 0.0304	0.550 ± 0.0544
Accuracy	DWS	0.920 ± 0.0394[a]	0.608 ± 0.0369[a]	0.821 ± 0.0305[a]	0.568 ± 0.0416[a]
Balanced Accuracy	BCE	0.844 ± 0.0551	0.566 ± 0.0265	0.778 ± 0.0379	0.572 ± 0.0398
Balanced Accuracy	BCELogit	0.852 ± 0.0518	0.582 ± 0.0361	0.765 ± 0.0370	0.577 ± 0.0376
Balanced Accuracy	Hinge	0.840 ± 0.0529	0.594 ± 0.0320	0.777 ± 0.0383	0.571 ± 0.0462
Balanced Accuracy	Focal	0.846 ± 0.0453	0.588 ± 0.0364	0.770 ± 0.0349	0.578 ± 0.0414
Balanced Accuracy	DWS	0.891 ± 0.0247[a]	0.599 ± 0.0338[a]	0.782 ± 0.0381[a]	0.617 ± 0.0294[a]

[a] The best ones.

Table 6. Accuracy and balanced accuracy results compared to logistic regression, support vector machine, decision tree, random forest, gradiant boost and its variant, K-nearest neighbors, and multilayer perceptron

Metric	Method	CN vs. AD	CN vs. MCI	MCI vs. AD	CN vs. MCI vs. AD
Accuracy	LR	0.918±0.0317	0.596±0.0297	0.845±0.0277[a]	0.554±0.0384
Accuracy	SVM	0.922±0.0334[a]	0.599±0.0372	0.783±0.0394	0.590±0.0273[a]
Accuracy	DT	0.809±0.0276	0.527±0.0559	0.781±0.0421	0.519±0.0426
Accuracy	RF	0.859±0.0274	0.848±0.0399	0.789±0.0365	0.545±0.0274
Accuracy	GB	0.891±0.0271	0.607±0.0404	0.781±0.0340	0.542±0.0323
Accuracy	LGB	0.866±0.0283	0.597±0.0367	0.695±0.0443	0.582±0.303
Accuracy	KNN	0.805±0.0419	0.569±0.0336	0.819±0.0322	0.548±0.0422
Accuracy	MLP	0.875±0.0405	0.603±0.0380	0.822±0.0386	0.560±0.0405
Accuracy	DWS	0.920±0.0394	0.608±0.0369[a]	0.821±0.0305	0.568±0.0416
Balanced Accuracy	LR	0.889±0.0450	0.598±0.0335	0.767±0.0430	0.608±0.0336
Balanced Accuracy	SVM	0.896±0.0459[a]	0.594±0.0397	0.721±0.0524	0.558±0.0361
Balanced Accuracy	DT	0.786±0.0381	0.517±0.0548	0.705±0.0460	0.517±0.0534
Balanced Accuracy	RF	0.846±0.0453	0.570±0.0460	0.757±0.0490	0.577±0.0259
Balanced Accuracy	GB	0.865±0.0384	0.577±0.0411	0.702±0.0494	0.436±0.0355
Balanced Accuracy	LGB	0.835±0.0391	0.562±0.0392	0.750±0.0405	0.566±0.0372
Balanced Accuracy	KNN	0.726±0.0576	0.552±0.0343	0.648±0.0395	0.475±0.0503
Balanced Accuracy	MLP	0.842±0.0548	0.511±0.0327	0.732±0.0481	0.495±0.0495
Balanced Accuracy	DWS	0.891±0.0247	0.599±0.0338[a]	0.782±0.0381[a]	0.617±0.0294[a]

[a] The best ones.

majority class and across all classes. Table 5 reports the result across the DNN model for binary and multi-class classification tasks. The result shows that the DWS loss provides the best balance accuracy and has a smaller variance, making it robust across multiple experiments. Additionally, the standard deviation across 20 tested runs also suggested that DWS is more robust regarding randomness.

Comparative study with other classification models

In Table 6, we compare our DWS loss based DNN model with several other methods: logistic regression, SVM, decision tree, random forest, gradient boosting, light gradient boost, K-nearest neighbors, and multilayer perceptron. The DWS method has the best balanced accuracy in CN vs. MCI, MCI vs. AD, and CN vs. MCI vs. AD and ranks second in CN vs. AD, indicating that the DWS performs well on this imbalanced dataset. However, its accuracy may not be the best.

To summarize the two comparative studies, the CN vs. AD task, while easier due to the distinct nature of the two groups, saw comparable results between DWS and SVM. The real challenge lies in distinguishing subtler changes, as in CN vs. MCI, MCI vs. AD, and multi-class distinction in CN vs. MCI vs. AD. These tasks are harder due to the gradual progression of the disease but mastering them is crucial for early diagnosis in clinical settings. Here, DWS has shown to be significantly more effective.

Feature importance

Originating from the game theory, the SHapley Additive exPlanation (SHAP)[22] method is based on Shapley's value and uses it as a unified measure of feature importance. In the SHAP method, each feature φ_i represents the effect of including that feature in model prediction, and it is computed as

$$\varphi_i = \frac{1}{|N|!}\sum_{S\subseteq N\setminus\{i\}} |S|!(|N| - |S| - 1)![f(S \cup \{i\}) - f(S)] \qquad \text{(Equation 1)}$$

where $f(S)$ is the output of the DNN model, S is the set of features to be used to explain the model, and N is the complete set of all features. In our calculation, the Shapley value of each feature is the average of its contributions across all the data, i.e., its permutation.

For the binary class classification, the SHAP values for the two classes, given a feature and observation, are just opposites of each other. Therefore, we only have a single bar value. For the three-class classification, the impact of a feature on each class is stacked to create the feature importance plot. In other words, the value tells us how much the feature is capable of helping us differentiate between classes.

Figure 2 is the feature importance map of the DWS loss associated with all four classification tasks. Taking the CN vs. MCI task in Figure 2 as an example, the left inferior parietal, left hippocampal, and different parts of temporal pole are the essential features, and previous studies[23,24] have indicated that these regions are able to detect the abnormal metabolism reduction at early stage of AD.

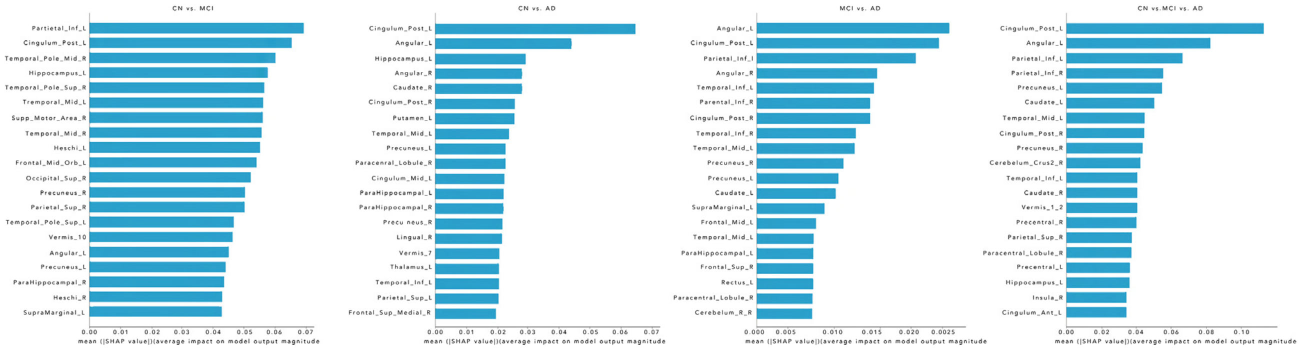


Figure 2. Feature importance of binary and multi-class classification from DWS Loss

For the MCI vs. AD task, left angular, left inferior temporal, and parietal inferior regions are shown to be the top relevant features. These regions are frequently used FDG ROIs in MCI and AD studies based on the meta-analysis, and AD patients have a significant metabolism decline among these regions compared with MCI or CN subjects.[25] In addition, posterior cingulate regions (Cingulum in the Automated Anatomical Labeling [AAL] atlas) are highly contributed to all the classifications. Studies[26,27] have demonstrated that posterior cingulate cortex showed higher hypometabolism in AD patients, and regional atrophy mainly lead to this abnormality.[28] On the clinical side, results have shown that angular gyrus has been shown to have an important role in understanding impairment in AD.[29] Left posterior cingulum has also been shown to be likely to play a remarkable role in the progressive development of cognitive impairment in AD.[30]

For the CN vs. AD task, the top three features are posterior cingulate, left angular, and left hippocampus to differentiate AD and CN patients. In clinical trials, the result shows that posterior cingulate and temporal pole are changed severely by the AD[31] pathologic.[32] The left hippocampus has also been shown to have high discriminative power in diagnosing Alzheimer's disease.[33]

For the feature importance of three class tasks, AD vs. MCI vs. CN, we plot the summation of the shapley value of each feature in order to show the global feature importance. The top three features are the left cingulate, left angular, and left inferior parietal. All numerical results are shown in Figure 2. A decrease order sorts the feature importance. The visualization of top five important features is also plotted using Mango (https://mangoviewer.com) in Figure 3.

Shown in Figure 4 is the feature importance map of the CN vs. MCI task for all the tested methods listed in Table 6. Due to the page limitation, we only show the feature importance map for the CN vs. MCI classification task, as it is a more valuable task for dementia detection at the early stage. The combined feature importance map provides an overall picture of which features are consistently important across models and how they affect predictions on average. In our result, the top five features in the combined map are left posterior cingulate, right angular, left orbital part of inferior frontal, vermis subregion, and right temporal pole mid region.

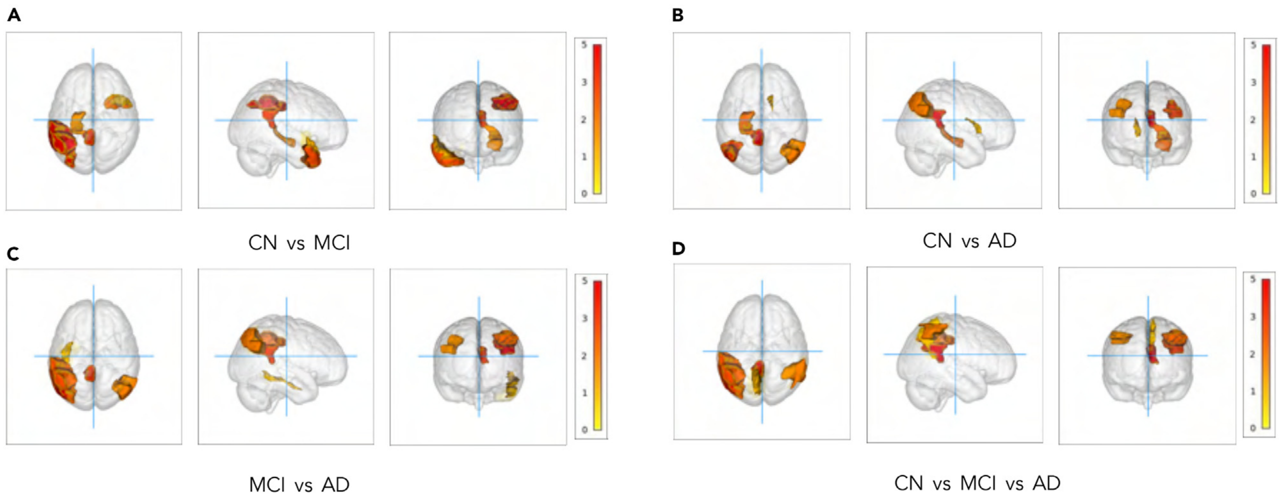


Figure 3. Brain visualization for binary and multi-class classification

Top five features are shown.

(A–D) are the four classification tasks: CN vs. MCI; CN vs. AD; MCI vs. AD; CN vs. MCI vs. AD. The color spectrum from yellow to red in the figure indicates the degrees of feature importance, with red being the most important.

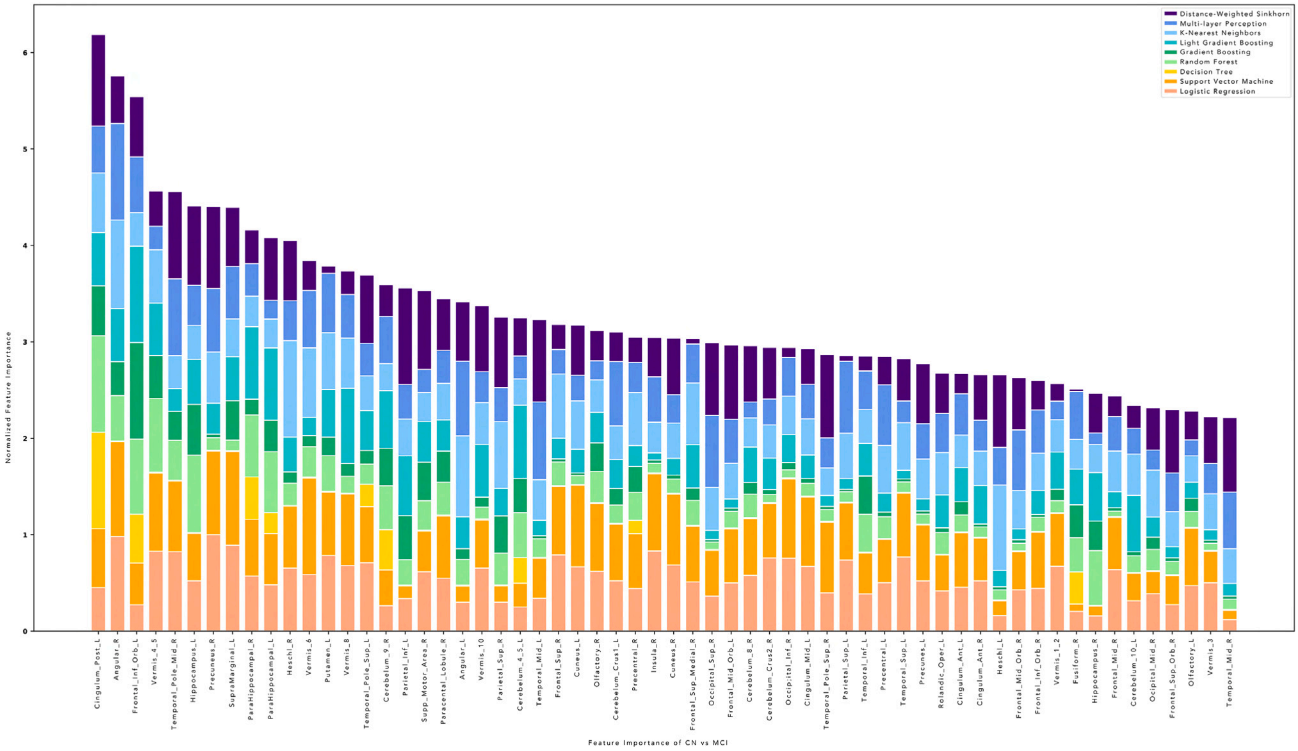


Figure 4. Feature importance of CN vs. MCI for all the methods tested in our empirical study

DISCUSSION

Based on the Wasserstein distances and Sinkhorn algorithm, we have proposed the DWS loss as an alternative to the current loss function for AD classification. The DWS incorporates ground truth distribution into the loss function, providing more information when we calculate the loss function.

In our empirical study, we have implemented a DNN with our DWS loss and applied it to a diagnosis task on classifying CN, MCI, and AD subjects using FDG-PET imaging data from the landmark ADNI database. Since the dataset is imbalanced, the balanced accuracy is more important than the accuracy as the evaluation metric. Our empirical results have demonstrated that the proposed DWS framework outperforms the traditional neural networks and yields comparable or better performances under the balanced accuracy. These experiments suggest the potential usage of our proposed method.

Limitations of the study

The proposed DWS loss function has several limitations. Similar to many deep learning methods, it requires hyper-parameters tuning. In the scheme of OT, the ground metric of the cost function plays an important role. We plan to focus on ground metric learning later to automatically determine the best ground metric to make the loss function more effective. Currently, the loss function is tailored

Algorithm 1. Updating the Weight θ of Deep Neural Network

```
1: Input: Mapping function $z = h_\theta(x)$, $y \in R^{1 \times k}$.
    2: Calculate: $S_T$, $p_j$, $S_{p,j}$, $C$.
    3: Initialize: $u = 1$, $K = e^{-\lambda C}$.
    4: for $t = 1, 2, \ldots$ do.
    5: $v \leftarrow y / K^T u$
    6: $u \leftarrow z / Kv$.
    7: end for.
    8: $\nabla_{h\theta}(x) \ell_{SD} = \epsilon \ln(u)$
    9: Output: Gradient of the objective function with respect to the learned mapping $h_\theta$.
```

specifically for DNNs, utilizing their architecture to calculate gradients. Future work will aim to derive an explicit gradient descent formula applicable to the DWS method. This formula will enable the integration of the DWS loss function into a broader range of ML models beyond DNNs.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

ACKNOWLEDGMENTS

This work was supported in part by the NIH grants U01 AG068057, RF1 AG063481, U01 AG066833, R01 LM013463, P30 AG073105, and R01 AG071470, and the NSF grant IIS 1837964.

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; Janssen Alzheimer Immunotherapy Research And 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.

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.

AUTHOR CONTRIBUTIONS

Conceptualization, Z.W., Q.Z., and L.S.; methodology, Z.W., Q.Z., S.Y., B.H., and L.S.; resources, H.H., A.J.S., P.M.T., C.D.; formal analysis, Z.W., Q.Z., B.T., S.Y., and L.S.; writing – original draft, Z.W., Q.Z., and B.T.; funding acquisition, H.H., A.J.S., P.M.T., C.D., L.S.; writing – review and editing, Z.W., Q.Z., B.T., S.Y., B.H., H.H., A.J.S., P.M.T., C.D., and L.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 16, 2024
Revised: January 27, 2024
Accepted: February 7, 2024
Published: February 12, 2024

REFERENCES

1. Srivastava, S., Ahmad, R., and Khare, S.K. (2021). Alzheimer's disease and its treatment by different approaches: A review. Eur. J. Med. Chem. *216*, 113320.
2. Breijyeh, Z., and Karaman, R. (2020). Comprehensive review on Alzheimer's disease: causes and treatment. Molecules *25*, 5789.
3. Alzheimer's Association (2015). 2015 Alzheimer's disease facts and figures. Alzheimers Dement. *11*, 332–384.
4. Jack, C.R., Jr., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imag. *27*, 685–691.
5. Nordberg, A. (2004). PET imaging of amyloid in Alzheimer's disease. Lancet Neurol. *3*, 519–527.
6. Dennis, E.L., and Thompson, P.M. (2014). Functional brain connectivity using fMRI in aging and Alzheimer's disease. Neuropsychol. Rev. *24*, 49–62.
7. Shen, L., Kim, S., Qi, Y., Inlow, M., Swaminathan, S., Nho, K., Wan, J., Risacher, S.L., Shaw, L.M., Trojanowski, J.Q., et al. (2011). Identifying Neuroimaging and Proteomic Biomarkers for MCI and AD via the Elastic Net. Multimodal Brain Image Anal. *7012*, 27–34.
8. Wan, J., Zhang, Z., Rao, B.D., Fang, S., Yan, J., Saykin, A.J., and Shen, L. (2014). Identifying the neuroanatomical basis of cognitive impairment in Alzheimer's disease by correlation- and nonlinearity-aware sparse Bayesian learning. IEEE Trans. Med. Imag. *33*, 1475–1487.
9. Shen, L., and Thompson, P.M. (2020). Brain Imaging Genomics: Integrated Analysis and Machine Learning. Proc. IEEE *108*, 125–162. https://doi.org/10.1109/JPROC.2019.2947272.
10. Ning, K., Chen, B., Sun, F., Hobel, Z., Zhao, L., Matloff, W.; Alzheimer's Disease Neuroimaging Initiative, and Toga, A.W., et al. (2018). Classifying Alzheimer's disease with brain imaging and genetic data using a neural network framework. Neurobiol. Aging *68*, 151–158.
11. Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., Dubois, B., Lehéricy, S., and Benali, H. (2009). Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology *51*, 73–83.
12. Sun, Z., Qiao, Y., Lelieveldt, B.P.F., Staring, M., et al.; Alzheimer's Disease NeuroImaging Initiative (2018). Integrating spatial-anatomical regularization and structure sparsity into SVM: Improving interpretation of Alzheimer's disease classification. Neuroimage *178*, 445–460.
13. Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In International conference on machine learning (PMLR), pp. 214–223.
14. Cuturi, M., and Doucet, A. (2014). Fast computation of Wasserstein barycenters. In International conference on machine learning (PMLR), pp. 685–693.
15. Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T.A. (2015). Learning with a Wasserstein loss. Adv. Neural Inf. Process. Syst. *28*.
16. Bogachev, V.I., and Kolesnikov, A.V. (2012). The Monge-Kantorovich problem: achievements, connections, and perspectives. Russ. Math. Surv. *67*, 785–890.
17. Cui, L., Qi, X., Wen, C., Lei, N., Li, X., Zhang, M., and Gu, X. (2019). Spherical optimal transportation. Comput. Aided Des. *115*, 181–193.
18. Wang, Z., Yang, W., Ryan, K., Garai, S., Auerbach, B.M., and Shen, L. (2022). Using Optimal Transport to Improve Spherical Harmonic Quantification of Complex Biological Shapes. In 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), *2022*, pp. 1255–1261.
19. Torous, W., Gunsilius, F., and Rigollet, P. (2021). An optimal transport approach to causal inference. Preprint at arXiv *8*. https://doi.org/10.48550/arXiv.2108.05858.
20. Geng, X. (2016). Label distribution learning. IEEE Trans. Knowl. Data Eng. *28*, 1734–1748.
21. Urbanowicz, R.J., Zhang, R., et al. (2022). STREAMLINE: A Simple, Transparent, End-To-End Automated Machine Learning Pipeline Facilitating Data Analysis and Algorithm Comparison. Preprint at arXiv *2*. https://doi.org/10.48550/arXiv.2206.12002.
22. Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. *30*.
23. Dukart, J., Kherif, F., Mueller, K., Adaszewski, S., Schroeter, M.L., Frackowiak, R.S.J., and Draganski, B.; Alzheimer's Disease Neuroimaging Initiative (2013). Generative FDG-PET and MRI model of aging and disease progression in Alzheimer's disease. PLoS Comput. Biol. *9*, e1002987.
24. Jiang, J., Sun, Y., Zhou, H., Li, S., Huang, Z., Wu, P., Shi, K., Zuo, C., Initiative, N., et al. (2018). Study of the influence of age in 18F-FDG PET images using a data-driven approach and its evaluation in Alzheimer's disease. Contrast Media Mol. Imaging *2018*.
25. Landau, S.M., Harvey, D., Madison, C.M., Koeppe, R.A., Reiman, E.M., Foster, N.L., Weiner, M.W., Jagust, W.J., et al.; Alzheimer's Disease Neuroimaging Initiative (2011). Associations between cognitive, functional, and FDG-PET measures of decline in AD and MCI. Neurobiol. Aging *32*, 1207–1218.
26. Salmon, E., Collette, F., Degueldre, C., Lemaire, C., and Franck, G. (2000). Voxel-based analysis of confounding effects of age and dementia severity on cerebral metabolism in Alzheimer's disease. Hum. Brain Mapp. *10*, 39–48.
27. Sakamoto, S., Ishii, K., Sasaki, M., Hosaka, K., Mori, T., Matsui, M., Hirono, N., and Mori, E. (2002). Differences in cerebral metabolic impairment between early and late onset types of Alzheimer's disease. J. Neurol. Sci. *200*, 27–32.
28. Chételat, G., Villain, N., Desgranges, B., Eustache, F., and Baron, J.-C. (2009). Posterior cingulate hypometabolism in early Alzheimer's disease: what is the contribution of local atrophy versus disconnection? Brain *132*, e133–e134.
29. Penniello, M.-J., Lambert, J., Eustache, F., Petit-Taboué, M.C., Barré, L., Viader, F., Morin, P., Lechevalier, B., and Baron, J.-C. (1995). A PET study of the functional neuroanatomy of writing impairment in Alzheimer's disease The role of the left supramarginal and left angular gyri. Brain *118* (Pt 3), 697–706.
30. Bozzali, M., Giulietti, G., Basile, B., Serra, L., Spanò, B., Perri, R., Giubilei, F., Marra, C., Caltagirone, C., and Cercignani, M. (2012). Damage to the cingulum contributes to Alzheimer's disease pathophysiology by deafferentation mechanism. Hum. Brain Mapp. *33*, 1295–1308.
31. Minoshima, S., Foster, N.L., and Kuhl, D.E. (1994). Posterior Cingulate Cortex in Alzheimer's Disease.
32. Arnold, S.E., Hyman, B.T., and Van Hoesen, G.W. (1994). Neuropathologic changes of the temporal pole in Alzheimer's disease and Pick's disease. Arch. Neurol. *51*, 145–150.
33. Juottonen, K., Laakso, M.P., Partanen, K., and Soininen, H. (1999). Comparative MR analysis of the entorhinal cortex and hippocampus in diagnosing Alzheimer disease. AJNR. Am. J. Neuroradiol. *20*, 139–144.
34. Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jr., Jagust, W., Morris, J.C., et al. (2017). Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. Alzheimers Dement. *13*, e1–e85.
35. Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., et al. (2013). The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimers Dement. *9*, e111–e194.
36. Jagust, W.J., Landau, S.M., Koeppe, R.A., Reiman, E.M., Chen, K., Mathis, C.A., Price, J.C., Foster, N.L., and Wang, A.Y. (2015). The Alzheimer's disease neuroimaging initiative 2 PET core: 2015. Alzheimers Dement. *11*, 757–771.
37. Penny, W.D., Friston, K.J., Ashburner, J.T., Kiebel, S.J., and Nichols, T.E. (2011). Statistical Parametric Mapping: The Analysis of Functional Brain Images (Elsevier).
38. Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage *15*, 273–289.
39. Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S., Trouvé, A., and Peyré, G. (2018). Interpolating between Optimal Transport and MMD Using Sinkhorn Divergences. https://doi.org/10.48550/ARXIV.1810.08278.
40. Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. (2020). Faster Wasserstein distance estimation with the Sinkhorn divergence. Adv. Neural Inf. Process. Syst. *33*, 2257–2269.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
ADNI	Weiner[34]	https://adni.loni.usc.edu/
Software and algorithms		
DWS Loss	This Paper	https://github.com/PennShenLab/DWS-Loss

RESOURCE AVALIABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Li Shen (Li.Shen@pennmedicine.upenn.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession URLs for the datasets are listed in the key resources table.
- The Source code and tutorials for implementing the DWS Loss has been deposited at GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Data for this study were sourced from the ADNI database. The experiment involved 789 participants, comprising 390 women and 399 men. Participant demographics such as age, total years of education, and ethnicity are detailed in the Tables 1, 2, 3, and 4. All participants were diagnosed with one of the following conditions: Alzheimer's disease (AD), MCI, or were deemed CN.

METHOD DETAILS

Data

Data used in the preparation of this article were obtained from the ADNI database,[34,35] which was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and AD. All participants provided written informed consent and study protocols were approved by each participating site's Institutional Review Board. Up-to-date information about the ADNI is available at www.adni-info.org.

In this study, we downloaded and analyzed FDG-PET imaging data (measuring glucose metabolism) from the ADNI database.[36] The FDG-PET imaging data play a crucial role in the diagnosis and assessment of AD. The Statistical Parametric Mapping software tool[37] was used to register FDG-PET scans into the standard brain space defined by the Montreal Neurological Institute (i.e., MNI-space). After that, the FDGPET scans were segmented based on the AAL atlas.[38] We calculated the average voxel signal intensity for each of 116 AAL regions and used these regional average measures as our features in the subsequent classification studies.

Optimal Transport

Below, we explain in brief the OT theory and its special instrument, the Wasserstein metric and the sinkhorn divergence.[39]

The OT problem is the optimal cost of changing one probability vector to match the shape of another probability vector. This gives us a measure of how similar the two probability vectors are. Most of the time, the least expensive method of moving mass from one distribution to another is also called Wasserstein distances.

Mathematically, consider two multi-variate distributions, with position $\{\mathbf{x}_i\}_{i=1}^{n}$ and $\{\mathbf{y}_i\}_{i=1}^{n}$, and then we have the discretized distributions:

$$\mathrm{P} = \sum_{i=1}^{n} p_i \delta_{\mathbf{x}_i} \text{ and } \mathrm{Q} = \sum_{j=1}^{m} q_j \delta_{\mathbf{y}_j} \tag{Equation 2}$$

where $\delta_{\mathbf{x}}$ denotes a Dirac delta function placed at a location $\mathbf{x} \in \mathbb{R}^n$. In this way, the p_i is vector of weights and $\{\mathbf{x}_i\}_{i=1}^{n}$ is the mass locations. The ground cost matrix represents the transportation costs between each pair of mass locations:

$$\mathbf{C}_{ij} = \| \mathbf{x}_i - \mathbf{x}_j \|^2 \qquad \text{(Equation 3)}$$

The transportation plan T, which tells us how much mass needs to be moved from $\mathbf{x}_i \rightarrow \mathbf{y}_j$ is a matrix $\mathbf{T} \in \mathbb{R}^{n \times m}$.

The total cost of a transport plan is then:

$$\begin{aligned} \underset{T}{\text{minimize}} \quad & \ell_{OT}(P,Q) = \langle T, C \rangle \\ \text{subject to} \quad & \mathbf{T} \in \mathbb{R}_{+}^{n \times m}, \\ & \mathbf{T}^T 1_n = \mathbf{q}, \\ & \mathbf{T} 1_m = \mathbf{p} \end{aligned} \qquad \text{(Equation 4)}$$

where $\langle \mathbf{T}, \mathbf{C} \rangle$ is the Frobenius inner product between the transport plan $\mathbf{T}$ and the cost matrix $\mathbf{C}$.

Solving the above optimization problem can be computationally challenging and unstable, requiring $\mathcal{O}(n^3 \log n)$ calculations. Because of this, it is challenging to apply Wasserstein distances in two-sample tests consistently. The entropic regularized Wasserstein distance is created by adding a regularization term $\gamma \mathbf{H}(\mathbf{T})$ to address these issues. This is called Sinkhorn algorithm, and its mathematical formulation is:

$$\begin{aligned} \underset{T}{\text{minimize}} \quad & \ell_{ROT}(P,Q) = \langle T, C \rangle - \gamma \mathbf{H}(T) \\ \text{subject to} \quad & \mathbf{T} \in \mathbb{R}_{+}^{n \times m}, \\ & \mathbf{T}^T 1_n = \mathbf{q}, \\ & \mathbf{T} 1_m = \mathbf{p} \end{aligned} \qquad \text{(Equation 5)}$$

where $\mathbf{H}(T)$ *is the entropy of the transport plan matrix* $\mathbf{T}$ *and is given by* $\mathbf{H}(T) = \sum_{i=1}^{n} \sum_{j=1}^{m} T_{i,j}(\log T_{i,j} - 1)$. *Note that the regularized Wasserstein distance is biased as* $\ell_{ROT}(P,P) \neq 0$. Therefore, combining two regularized Wasserstein distances can build an unbiased divergence, and it is called the Sinkhorn divergence:

$$\ell_{SD}(\mathbf{P},\mathbf{Q}) = \ell_{ROT}(\mathbf{P},\mathbf{Q}) - \ell_{ROT}(\mathbf{P},\mathbf{P}) - \ell_{ROT}(\mathbf{Q},\mathbf{Q}) \qquad \text{(Equation 6)}$$

Problem formulation and proposed method

Figure 1 shows the schematic design of the proposed algorithm using our DWS loss function. In our proposed model, we first send the data into a fully connected DNN. After applying the softmax activation functions, the row logits turn into probability. The probability is used as the weight for the predicted distribution. To be specific, we have three probabilites for one instance and using the whole data will give as three weighted distributions. The Sinkhorn algorithm is then used to match the distribution between the predicted distribution and the ground truth distribution for each class label (i.e., CN, MCI, or AD). Finally, backpropagation is taken to adjust the neural network. In this flowchart, we assumed that the model had correctly predicted each instance. Therefore, the dominated distribution of each class is similar to its ground truth distribution. Below, we discuss our problem formulation.

Let $\mathcal{X}$ represent the feature space and $\mathcal{Y}$ denote the label space $\mathcal{L} = \{l_1, l_2, \ldots, l_k\}$, where l_k is the k th label. In this paper, the aim is to learn an optimal mapping function h_θ: $\mathcal{X} \rightarrow \mathcal{Y}$, parameterized by θ, over space of hypotheses $\mathcal{H}$. Given an input $\mathbf{x}$, h_θ maps it into a vector $= [y_{l_1}, y_{l_2}, \ldots, y_{l_k}]$. Therefore, the vector $\mathbf{y}$ represents the probability that the instances belong to each label. The ground truth probability is also defined by $\mathbf{z} = [z_{l_1}, z_{l_2}, \ldots, z_{l_K}]$, where the component is equal to 1 for true label and 0 otherwise.

Given an i.i.d. set of N training samples $\mathcal{D} = ((\mathbf{x}_1,\mathbf{y}_1)\ldots(\mathbf{x}_N,\mathbf{y}_N))$, the overarching goal of the algorithm is to find the mapping function h_θ that minimizes the empirical risk

$$\min_{h_\theta \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \ell(\mathbf{z}_i = h_\theta(\mathbf{x}_i), \mathbf{y}_i) \qquad \text{(Equation 7)}$$

where $\ell(\bullet,\bullet)$ is a loss function. Instead of minimizing the empirical risk among each instance, we would like to minimize the empirical risk w.r.t each class

$$\min_{h_\theta \in \mathcal{H}} \frac{1}{K} \sum_{j=1}^{k} l_{SD}(\mathbf{S}_j, \mathbf{S}_{p,j}) \qquad \text{(Equation 8)}$$

where K is the number of class in our sample.

Distance-weighted Sinkhorn loss

In this section, the DWS Loss will be introduced. In the following sections, we use the shorthand DWS for brevity. Instead of considering the difference between two probability vectors of the same instance, we consider the difference between the total distribution of data and its weight-predicted distribution. The formulation of DWS Loss is as follows:

$$l_{DWS}(\mathcal{D}, h_\theta) = \sum_{j=1}^{k} l_{SD}(\mathbf{S}_j, \mathbf{S}_{p,j}) \tag{Equation 9}$$

The $\mathbf{S}_j$ is the empirical distribution of the data in class j in the dataset and could be represented as

$$\mathbf{S}_j = \frac{1}{N_j} \sum_{i=1}^{N} \mathbf{z}_{i,j} \delta_{\mathbf{x}_i} \tag{Equation 10}$$

where N_j is the number of instances in class j, and the sum is over all instances in the dataset. Notice that $\mathbf{z}_i$ is the one-hot encoded ground truth label of instance i, where $z_{i,j} = 1$ if instance i belongs to class j and $z_{i,j} = 0$ otherwise. $\delta_{\mathbf{x}_i}$ is a Dirac delta function centered at the location of instance i. Similarly, let $\mathbf{y}_i = h_\theta(\mathbf{x}_i)$ be the predicted label distribution for instance i. We can then define p_j as a vector that collects the predicted probabilities for class j across all instances. It estimates the probability of each instance being assigned to class j according to the model h_θ.

The predicted distribution of the data for class j is then represented as:

$$\mathbf{S}_{p,j} = \frac{1}{C_j} \sum_{i=1}^{N} h_\theta(\mathbf{x}_i)[j] \delta_{\mathbf{x}_i} \tag{Equation 11}$$

Here C_j is a normalizing factor to ensure $S_{p,j}$ is indeed a distribution.

We then illustrate the exact computation method for calculating the DWS loss; see also Algorithm 1. Notice that optimizing and differentiating the DWS loss consists of 2 regularized Wasserstein distances. Therefore, we first focus on the computation of a single regularized Wasserstein. The dual form could be written as

$$Dual(\alpha, \beta) = \alpha^\top \mathbf{P} + \beta^\top \mathbf{Q} - \epsilon \sum_{i,j=1}^{n,m} e^{\frac{\left(C_{ij} - \alpha_i - \beta_j\right)}{\epsilon}} \tag{Equation 12}$$

Then, by the Sinkhorn's scaling theorem, one could show that the optimal solution for the primal problem is related to its dual form solution:

$$T_* = \mathrm{diag}(e^{\epsilon\alpha_*}) e^{-\frac{C}{\epsilon}} \mathrm{diag}(e^{\epsilon\beta_*}) \tag{Equation 13}$$

where α^* and β^* are the minimizers of the dual Lagrange problem. The above optimal solution for dual problem $\alpha^\bigstar, \beta^\bigstar$ can be computed using Sinkhorn's algorithm. From the Sinkhorn algorithm, one can show that $\alpha^\bigstar = \varepsilon \log u^\bigstar, \beta^\bigstar = \varepsilon \log v^\bigstar$, where the $u^\bigstar, v^\bigstar$ are the outputs of Sinkhorn algorithm. Referring to the dual formulation, one could notice that $\nabla_{h_\theta(\mathbf{x})} \ell_{SD} = \alpha^*$.

Also, for each instance i in the dataset, the DWS loss also maintains the point wise convergence, that is:

$$\mathbf{y}_i \rightarrow \mathbf{z}_i \Leftrightarrow l_{SD}(\mathbf{S}_i, \mathbf{S}_{p,i}) \rightarrow 0 \Leftrightarrow l_{DWS}(\mathcal{D}, h_\theta) \rightarrow 0 \tag{Equation 14}$$

Theoretical result

Here, we provide a theoretical bound using Rademacher Complexity.

Definition 1. Let G be a family of functions mapping a set Z into $\mathbb{R}$. A list $S = (z_1, \ldots, z_m)$ of elements. Let $\sigma = (\sigma_1, \ldots, \sigma_m)$ be a list of independent random variables, where, for each $i \in \{1, \ldots, m\}$, σ_i takes value $+1$ with probability $1/2$ and takes value -1 with probability $1/2$. Then the empirical Rademacher complexity of G with respect to S is defined to be

$$R_S(G) = E_\sigma\left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i)\right] \tag{Equation 15}$$

The Rademacher complexity of G with respect to samples of size m drawn according to D is

$$R_m(G) = \underset{S \sim D^m}{E}[R_S(G)] \tag{Equation 16}$$

We also have the definition of Emprical Risk and Risk as follows:

$$E(h) = \underset{(x,y) \sim \mathcal{D}}{\mathbb{E}}[\ell(h(x), y)]; \widehat{E}_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i) \tag{Equation 17}$$

Then, based on the general theorem of Rademacher complexity, we have that.

Theorem 1. Let G be a family of functions mapping a set Z to the unit interval $[0,1]$. Suppose that a sample S of size m is drawn according to distribution D on Z. Then for any $\delta > 0$, with probability at least $1 - \delta$ the following holds for all functions $g \in G$:

$$E(g) \leq \widehat{E}_S(g) + 2R_m(G) + O\left(\sqrt{\frac{\log\frac{1}{\delta}}{m}}\right) \qquad \text{(Equation 18)}$$

To connect the Sinkhorn divergence with the Rademacher complexity, we have the following approximation between the original OT formulation and the Sinkhorn divergence. We adopted Proposition 11 from Chizat.[40]

Theorem 2. Assume that $\mu_n = \sum_{i=1}^{n} p_i \delta_{x_i}$ and $\nu_n = \sum_{j=1}^{n} q_j \delta_{y_j}$ are discrete measures with n atoms such that $p_i, q_j \geq \alpha / n$ for some $\alpha > 0$. Then, we have that

$$0 \leq l_{SD}(\mu, \nu) - l_{OT}(\mu, \nu) \leq 2\lambda H(\gamma^*, \mu \otimes \nu) \leq 4\lambda(\log n + \log(1/\alpha)) \qquad \text{(Equation 19)}$$

Therefore, we could have that

$$R_m(l_{OT}) \leq R_m(l_{SD}) \leq R_m(l_{OT}) + 2\lambda(\log n + \log(1/\alpha)) \qquad \text{(Equation 20)}$$

Finally, we can conclude that with at least $1 - \delta$ probability, $E(l_{DWS}) \leq \inf_{h \in \mathcal{H}} E(h) + 2\lambda(\log n + \log(1/\alpha)) + O\left(\sqrt{\frac{\log\frac{1}{\delta}}{m}}\right)$.