BenchCouncil Transactions on Benchmarks, Standards and Evaluations 4 (2024) 100198

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/

Full Length Article

COADBench: A benchmark for revealing the relationship between AI models and clinical outcomes

Jiyue Xie [a], Wenjing Liu [b,*], Li Ma [b], Caiqin Yao [c], Qi Liang [a], Suqin Tang [a], Yunyou Huang [a]

[a] Key Lab of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, No. 15 Yucai Road, Qixing District, Guilin, 541004, Guangxi, China
[b] Guilin Medical University, 20 Lequn Road, Xiufeng District, Guilin City, 541001, Guangxi, China
[c] The Sencond Nanning People's Hospital, No. 13, Dan Village Road, Jiangnan District, 530000, Guangxi, China

ARTICLE INFO

Keywords:
Alzheimer's disease
Benchmark
Clinical outcome
Artificial intelligence

ABSTRACT

Alzheimer's disease (AD), due to its irreversible nature and the severe social burden it causes, has garnered significant attention from AI researchers. Numerous auxiliary diagnostic models have been developed with the aim of improving AD diagnostic services and thereby reducing the social burden. However, due to a lack of validation regarding the clinical value of these models, no AD diagnostic model has been widely accepted by clinicians or officially approved for use in enhancing AD diagnostic services. The clinical value of traditional medical devices is validated through rigorous randomized controlled trials to prove their impact on clinical outcomes. In contrast, current AD diagnostic models are only validated based on their accuracy, and the relationship between these models and patient outcomes remains unknown. This gap has hindered the acceptance and clinical use of AD diagnostic models by healthcare professionals. To address this issue, we introduce the COADBench, a benchmark centered on clinical outcomes for evaluating the clinical value of AD diagnostic models. COADBench curated subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database who have at least two cognitive score records (the most commonly used clinical endpoint in AD clinical trials) from different follow-up visits. To the best of our knowledge, for the first time, it links the cognitive scores of subjects with model performance, using patient cognitive scores as clinical outcomes after intervention to evaluate the models. Through the benchmarking of current mainstream AD diagnostic algorithms using COADBench, we find that there was no significant correlation between the subjects' cognitive improvement and the model's performance, which means that the current performance evaluation criteria of mainstream AD diagnostic algorithms are not combined with clinical value.

1. Introduction

Alzheimer's disease is the most common type of dementia, accounting for the largest proportion of dementia, because of its irreversible, high cost of diagnosis, no cure and other characteristics, to society has brought a very serious burden. In order to reduce the diagnostic cost and improve the diagnostic effect, artificial intelligence (AI) researchers have developed various deep learning models to assist the diagnosis of Alzheimer's disease. For example, *Qiu et al.* [1] use a multi-modal input model based on 3D CNN to make three classifications of subjects, and the best model achieves an AUC of 0.971; *Xing et al.* [2] use a binary classification of subjects based on dynamic images and a pre-trained CNN model, and the best model achieves an AUC of 0.95.

Alzheimer's disease currently lacks a cure, so the main purpose of diagnosis is to identify patients with reversible or delayed symptoms for treatment, improving clinical outcomes and thus benefiting patients. The clinical assessment of the effectiveness of Alzheimer's disease diagnosis is mainly based on the calculation of benefits (such as cognitive improvement) based on changes in clinical endpoints or alternative endpoints. However, the evaluation indicators (Accuracy, AUC, etc.) of the current AI models used to diagnose Alzheimer's disease are not directly or indirectly related to clinical value. This means that although an AI model achieves a high value of AUC in the diagnostic task of categorizing or multicategorizing subjects (normal, mild cognitive impairment, Alzheimer's), the clinical value based on patient benefit does not necessarily improve. For example, *Zhang et al.* [3] use a fusion input model based on 3D CNN and Transformer to binary classify subjects. The accuracy of the best model reaches 0.929, but the index of cognitive improvement of patients in clinical practice is only 0.806.

Currently, in other areas where AI models have been introduced, the correlation between model evaluation and clinical outcomes is low. *Tyler et al.* [4] propose an algorithm based on KNN-DSS to provide

* Corresponding author.
E-mail address: liuwenjing@stu.gxnu.edu.cn (W. Liu).

https://doi.org/10.1016/j.tbench.2025.100198
Received 24 October 2024; Accepted 1 March 2025
Available online 13 March 2025

weekly insulin injection recommendations for patients with type 1 diabetes (T1D), using the duration of time that the patient's HbA1c level remains within the safe range as the clinical outcome in conjunction with the algorithm; *Komorowski et al.* [5] propose an AI clinician who gives reinforcement learning to provide the best medical strategy to the patient, and use mortality rates to evaluate the AI doctor's medical strategy. *Adams et al.* [6] develop a sepsis alert system based on machine learning, deploy it in hospitals to monitor the situation of sepsis patients, and evaluated the performance of the system using in-hospital mortality as the clinical outcome of patients. But there is no comparable example of a model for diagnosing Alzheimer's disease. This can lead to high classification evaluation metrics such as AUC or Accuracy, but poor clinical outcomes. For example, when the model tends to accurately identify patients whose cognition cannot be improved, a high model accuracy does not result in improved clinical outcomes.

In order to solve the above problems, COADBench first considers the use of clinical outcomes to evaluate the diagnostic model of Alzheimer's disease.

In most current clinical trials, the endpoint of Alzheimer's disease is cognitive improvement, so cognitive improvement is a quantitative model and a clinically significant endpoint acceptable to experts [7–11]. Thus, we propose clinical benefit measures based on changes in patients' ADAS scores (which reflect patients' cognitive ability) during follow-up after diagnosis and treatment, which could be used to evaluate model performance.

Second, we select samples from Alzheimer's Disease Neuroimaging Initiative (ADNI). The sample inclusion criteria: patients have at least two follow-up visits, in the form of 3D imaging data and demographic non-imaging data, with three categories of subjects: normal, mild cognitive impairment, and Alzheimer's disease.

Third, we build COADBench based on clinical benefit indicators and benchmark datasets, and conduct benchmark testing on mainstream Alzheimer's diagnosis models using the constructed COADBench. Our contributions are as follows:

- To the best of our knowledge, for the first time, we introduce ADAS scores as surrogate outcomes in the evaluation of an Alzheimer's disease model, correlating the model's performance with clinical value.
- To the best of our knowledge, with ADAS scores as the center, we construct the first clinically valuable benchmark for evaluating Alzheimer's disease models.
- The evaluation of current mainstream Alzheimer's disease models based on COADBench reveal that: (1) When classification evaluation indicators such as Accuracy and AUC are used to evaluate the model, the model with the best performance may not be the model with the highest clinical value; (2) There was no significant positive correlation between the classified evaluation indicators and clinical benefit indicators based on ADAS scores.

The paper is structured as follows. Section 2 describes the definition of the problem. Section 3 reviews recent research on diagnostic models for Alzheimer's disease. Section 4 covers COADBench in detail. Section 5 introduces the experimental results and analysis based on COADBench. Section 6 summarizes the findings.

2. Problem definition

2.1. Definition of the AD diagnosis problem

The AD diagnosis task in the current mainstream research is defined as a classification problem as follows:

$$min\left\{\mathbb{E}_{(x,y)\sim\mathcal{D}_{Tr}}\alpha L\big(m(x),y\big)+\beta R(m)\right\} \tag{1}$$

Where $\mathcal{D}_{Tr}$ is the training set, The $L\big(m(x),y\big)$ indicates the loss at data point (x,y) with AD diagnosis model m, $R(m)$ indicates the regularization term of the model m. The coefficients α and β trade off these terms.

2.2. Clinical assessment of patient cognition

Clinically, the main way to enhance patient benefit is by improving the patient's cognitive function, which is quantified through the ADAS scores obtained from multiple follow-ups after treatment. Additionally, it is important to reduce the various losses caused by inaccurate diagnoses.

$$\begin{cases} max\left\{\sum\limits_{\mathcal{D}_{Te}} f\big(m(x)\big) * p * max\left\{0, A'-A\right\}\right\} \\ f\big(m(x)\big)=\begin{cases}1, m(x)=1\\ 0, other\end{cases} \\ min\left\{L_{FPR}\right\} \\ L_{FPR}=\dfrac{FP}{FP+TN}\end{cases} \tag{2}$$

Where $\mathcal{D}_{Te}$ is the test set, $m(x)$ represents the prediction result of the model, and A and A' represent the ADAS score values of the patient at the current and next follow-up visits, respectively. p is equal to 1 when the model prediction is correct; otherwise, p is equal to 0.

L represents the psychological impact on non-AD subjects when they are misdiagnosed as AD patients, as well as the losses incurred from further medical consultations. Since this part is difficult to quantify, we use the model's False Positive Rate on the test set as a substitute. FP represents false positive rate and TN represents true negative rate.

3. Related work

To evaluate the effectiveness of a model in diagnosing a particular disease, it is necessary to ensure that its correct diagnostic predictions have a positive impact on patients. For example, *Komorowski et al.* [5] use a model to provide medication strategies for sepsis patients. They demonstrate the model's effectiveness by showing that the lowest mortality rates occurred in patients whose actual dosages matched the AI's recommendations. *Tyler et al.* [4] demonstrate the effectiveness of their model by showing that patients' blood sugar levels improved after adjusting the medication dosage according to the model's recommendations. *Arbabshirani et al.* [12] not only demonstrate the accuracy and specificity of their model in diagnosing intracranial hemorrhage but also highlight its clinical impact. The model successfully identify patients initially deemed to require only routine examinations, upgrading them to needing immediate examinations. Radiologists confirm that 64% of these upgraded patients indeed have intracranial hemorrhages, thereby proving the model's effectiveness. Because deep learning and similar technologies must ensure improved patient outcomes before being applied clinically, it is not sufficient to merely focus on increasing the accuracy of disease diagnosis models.

Since there are no treatments that can stop or reverse AD, existing medications may alleviate symptoms but are typically only effective in the early stages of the disease [13]. As a result, much research focuses on accurately identifying early-stage AD patients. The effectiveness of these models is often evaluated based on accuracy, a computer-based metric. However, when these models are applied clinically, it is essential to consider not only their accuracy but also whether early intervention, following a model's identification of an early AD patient, can improve actual patient outcomes. Currently, there are no prospective studies to validate this aspect. Most models are trained and tested using publicly available Alzheimer's disease datasets and evaluated based on metrics such as accuracy, sensitivity, precision, specificity, and F-measure. For example, studies by *Suk et al.* [14], *liu et al.* [15], *Martinez-Murcia et al.* [16], *Feng et al.* [17], *Raza et al.* [18] are all based on these publicly available datasets and performance metrics. In prospective studies on the effectiveness of drugs in improving patient symptoms [19], the impact on patients is typically assessed using the

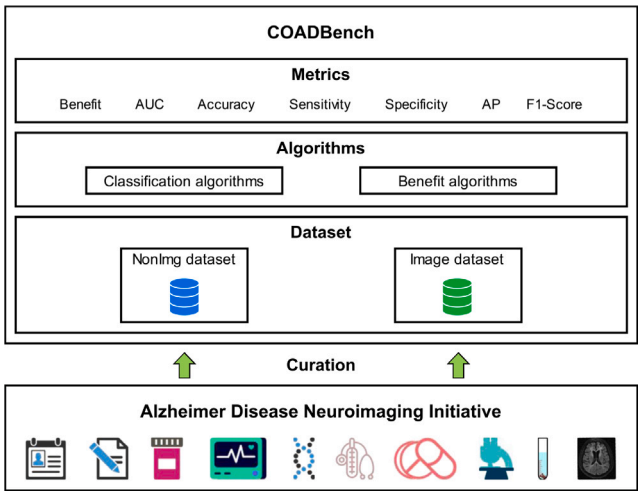

Fig. 1. The summary of COADBench benchmark framework.

Mini-Mental State Examination (MMSE) and the Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS-cog).

In numerous medical domains, the assessment of model performance is frequently closely tied to actual patient outcomes. For example, in sepsis, which can result in rapid patient deterioration and death, model efficacy is often evaluated based on mortality rates. In contrast, AD remains incurable and progresses slowly [20], rendering mortality an impractical outcome measure. Consequently, current deep learning research on Alzheimer's disease focuses on early diagnosis, with model performance evaluation primarily relying on computational metrics such as accuracy, sensitivity, etc. However, reliance on these metrics alone is insufficient to demonstrate the model's positive impact on individual patients. Moreover, the absence of prospective studies further complicates the validation of the model's effectiveness in clinical practice.

4. COADBench

The structural block diagram of COADBench is shown in Fig. 1. The structural block diagram is viewed from bottom to top. Data from 13 types of medical examinations commonly used in AD diagnosis are selected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (http://adni.loni.usc.edu) and divided into non-imaging and imaging datasets to match different model inputs. AD diagnosis models generally use classification algorithms to diagnose subjects. In COADBench, we also use the benefit calculation algorithm to compute benefit metrics. For model evaluation, classification metrics (such as AUC, Accuracy, etc.) are used to assess the model's performance in classification, while benefit metrics are used to evaluate the model's clinical benefits for patients.

For model evaluation, classification algorithms and benefit calculation algorithms are used to obtain classification evaluation metrics (AUC, Accuracy, etc.) and benefit indicators, respectively.

4.1. Data sources

COADBench involves 10 tables and 3 categories of images which represent 13 categories of medical examinations data commonly used in AD diagnosis. The data are collected from 67 sites in the United States and Canada, contains 1543 subjects with 6225 visits, and all visits are labeled by one of three labels: AD (Alzheimer's disease), CN (Cognitively normal) and MCI (Mild cognitive impairment).

ADNI 13 kinds of medical tests shown in the list below:

Table 1
Characteristics of subjects.

		Number of subjects
Age	[55, 60)	39
	[60, 70)	311
	[70, 80)	790
	[80, 90)	391
	[90, 92]	12
Educate	[0, 3]	1525
	[4, 20]	9
Ethnic category	Hisp/Latino	46
	Not Hisp/Latino	1488
	Unknown	9
Racial category	White	1431
	More than one	14
	Black	64
	Asian	27
	Hawaiian/Other PI	1
	Unknown	6
Marriage	Married	1176
	Never married	53
	Widowed	178
	Divorced	130
	Unknown	6
Category	AD	330
	CN	408
	MCI	805

(1) Base information (Base), usually obtained through consultation, includes demographics, family history, medical history, and symptoms.
(2) Cognition information (Cog), usually obtained through consultation and testing, includes Alzheimer's Disease Assessment Scale, Mini-Mental State Exam, Montreal Cognitive Assessment, Clinical Dementia Rating, and Cognitive Change Index.
(3) Cognition testing (CE), usually obtained through testing, includes ANART, Boston Naming Test, Category Fluency-Animals, Clock Drawing Test, Logical Memory-Immediate Recall, Logical Memory-Delayed Recall, Rey Auditory Verbal Learning Test, Trail Making Test.
(4) Neuropsychiatric information (Neur), usually obtained through consultation, includes Geriatric Depression Scale, Neuropsychiatric Inventory, and Neuropsychiatric Inventory Questionnaire.
(5) Function and behavior information (FB), usually obtained through consultation, includes Function Assessment Question, Everyday Cognitive Participant Self Report, Everyday Cognition Study Partner Report.
(6) Physical neurological examination (PE), usually obtained through testing, includes Physical Characteristics, Vitals, and neurological examination.

The rest of the examinations include blood testing (Blood), urine testing (Urine), nuclear magnetic resonance scan (MRI), positron emission computed tomography scan with 18-FDG (FDG), positron emission computed tomography scan with AV45 (AV45), gene analysis (Gene), and cerebrospinal fluid analysis (CSF).

4.2. Benchmark datasets

To assess different AD diagnosis model, COADBench data source into two parts: image data and non-image data. The image data includes nuclear magnetic resonance scan imaging (MRI), positron emission computed tomography (PET) image, while the non-image data includes the remaining 10 types of tabular data from ADNI.

The demographic information of benchmark datasets subjects is shown in Table 1.

4.3. Classification algorithms

AD diagnosis models typically use classification algorithms, usually binary classification (normal individuals, Alzheimer's disease patients) or three-way classification (normal individuals, mild cognitive impairment, Alzheimer's disease patients). Based on the model's output format, the following classification methods are used:

- The model's output consists of a number of values ranging from [0, 1], corresponding to the number of classes, which represent the probabilities of the subjects belonging to each category. The category associated with the highest probability is then selected as the model's judgment result for the subject.
- The model's output is a score representing the subject's level of cognitive impairment. A threshold (in the case of binary classification) or two thresholds (in the case of three-way classification) are needed to map the score to a specific category. For three-way classification, for example, when the model outputs a COG_Score, thresholds a and b can be used to determine the specific category according to the following formula:

$$Category = \begin{cases} CN, COG\_Score \leq a \\ MCI, a < COG\_Score < b \\ AD, b \leq COG\_Score \end{cases} \tag{3}$$

4.4. Metrics

In COADBench, in addition to the common classification evaluation indicators such as AUC, Accuracy, Sensitivity, Specificity and AP, we also introduce the clinical indicator benefit to evaluate the benefit of a model to the subject. Benfit computation formula is as follows:

$$M = \sum_{i}^{n} l * b_i \tag{4}$$

$$B = \frac{1}{m} \sum_{i}^{n} l * p * b_i \tag{5}$$

Where l indicates the label of the subject (AD is 1 and others are 0), p indicates the prediction of the subject, If the cognition of the subject has not improved, then b=0, otherwise b is the difference between the subject's current ADAS-Cog and the follow-up ADAS-Cog.

Please note that all operations involving the subtraction of metrics in the paper assume that the difference between the two confidence intervals of the respective metrics is both independent and normally distributed.

5. Experimental results and analysis

COADBench is constructed based on the mainstream four Alzheimer's diagnosis models for the benchmark test. The benchmarking process for each diagnosis model is roughly the same, requiring data preprocessing, model training, and evaluation using both classification indices and image evaluation metrics.

5.1. Experimental setup

Experiments were conducted on a machine equipped with an NVIDIA A100 80 GB PCIe GPU, Intel Xeon Silver 4208 CPU, 256 GB RAM, and a 16TB HDD running CentOS 7.9. The hyperparameters of the experimental models are shown in Table 3.

Our process for evaluating AD diagnostic models is as follows: First, we save multiple intermediate models at different stages of training. For each intermediate model, we calculate classification evaluation metrics such as AUC and Accuracy, as well as the benefit metric on the test set.

After calculating the various metrics, the model with the highest AUC or Accuracy is selected as the one with the best classification performance, while the model with the highest benefit is considered the most beneficial for patients.

5.2. Data preprocessing

Each Alzheimer's diagnosis model the required data form is not the same, some model using only the image data, some model only using the image data, and some models use a mixed input of image and image data.

Image data preprocessing typically involves only standardizing the image size, while non-image data requires data cleaning. This includes removing features with too many missing values, removing features with excessive single-value entries, and filling in the missing values. The meanings of some of the main columns in the data are shown in Table 4.

Our benchmark data set of each record according to follow-up time and ADAS — cog difference to calculate the practice guideline values, so that the follow-up evaluation model of calculating the practice guideline values, the calculation formula of the practice of index in 4.4. The dataset was divided into training, validation, and test sets in a 6:2:2 ratio.

5.3. Model

We selected the following AD diagnostic models for evaluation:

- *Qiu et al.* [1] proposed three models to classify subjects into three categories: an MRI model based on 3D CNN and multilayer perceptron, using only MRI images as input; CatBoost based nonImg model uses only non-image data as input; the Fusion model based on CatBoost uses a mixed input of non-image data and image data. In Qiu et al.'s paper, the Fusion model finally achieved the best performance. We benchmarked for all three models.
- *Xing et al.* [2] used dynamic image-based and pre-trained CNN models to dichotomize subjects (AD vs CN). In Xing et al.'s paper, they used approximate rank pooling to convert 3D MRI into 2D dynamic image. The pre-trained CNN model was then input.
- *Zhang et al.* [3] use CNN and the Transformer based on 3 d model of the subjects for binary classification (AD vs CN), use only the image data as input.
- *Hosseini et al.* [21] proposed a deep 3D convolutional neural network for three-classification of subjects with Alzheimer's disease, using MRI images as input.

5.4. Results

The results of the evaluation of the mainstream AD diagnostic models are shown in Table 2. As can be seen from the table, when the AUC and other indicators of the AD diagnostic model reach their highest, the benefit value is not the highest in most of cases, which means that it is problematic to use AUC and other classification evaluation indicators to select the most effective model, because the model selected according to this method is not necessarily the most beneficial model for patients.

If we only look at the situation with the highest index, there is not much difference between the index value of the model with the best classification effect and the model that is most beneficial to patients, but not all AD diagnostic models can achieve good classification effects. For example, the Multimodal Nonimg model in Table 1 has the highest accuracy of 0.7619. The corresponding benefit is 0.8310, but when taking the model with the highest benefit, the benefit reaches 0.8886 when the accuracy is only 0.6577. This indicates that when the classification effect of the model is not very good, the index values of the model with the best classification effect and the model that is most beneficial to the patient may differ greatly.

In order to better analyze the experimental results and illustrate our point, we plot the scatter plot 2 with categorical metrics on the X-axis and benefit on the Y-axis. Each point in the scatter plot represents a

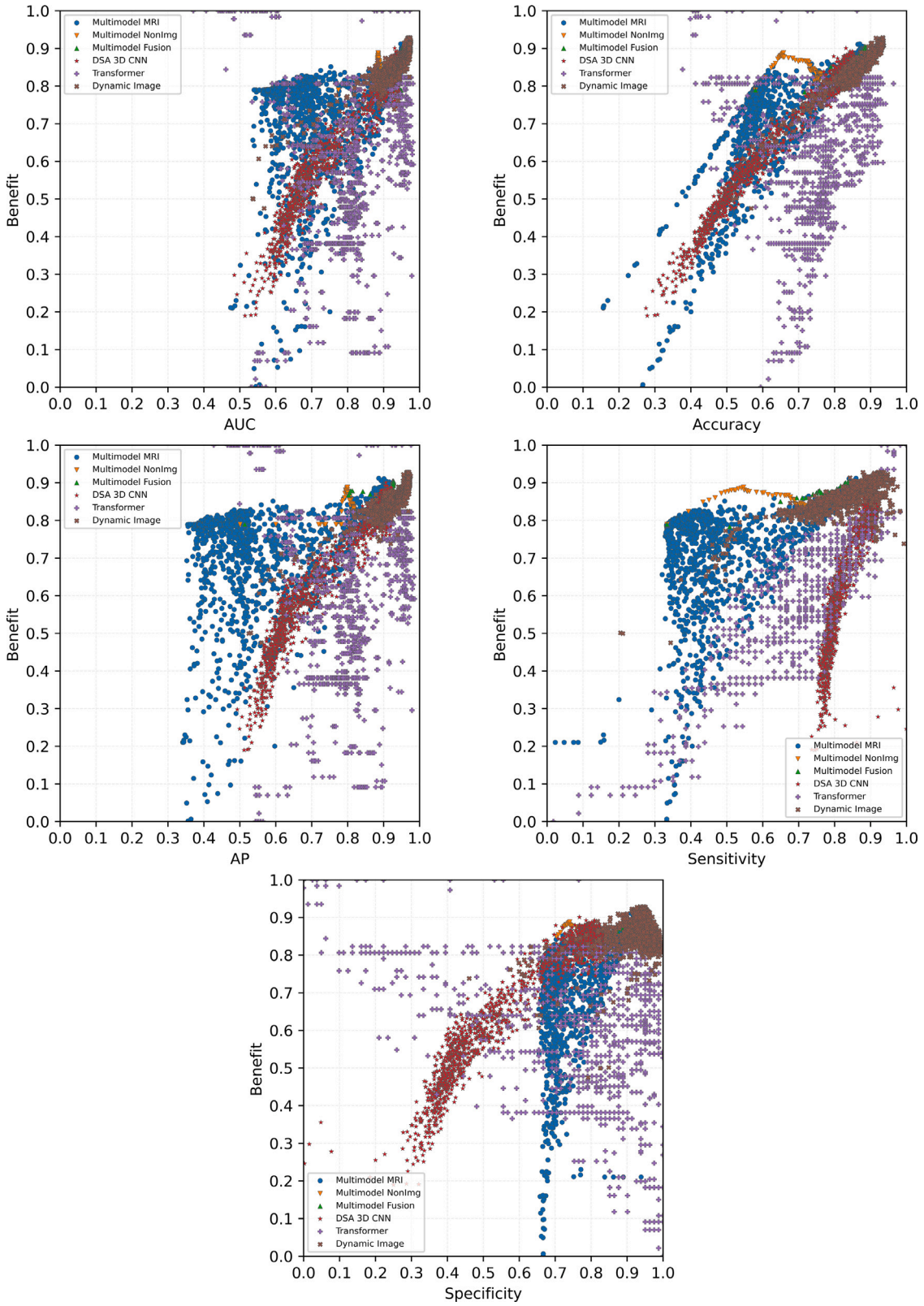

Fig. 2. Scatter plot of metrics versus benefit.

Table 2
Classification Metrics vs Benefit.

	best AUC		best Benefit		best Accuracy		best Benefit	
	AUC	Benefit	AUC	Benefit	Accuracy	Benefit	Accuracy	Benefit
Multimodal MRI [1]	0.9591	0.8926	0.9477	0.9119	0.8806	0.9020	0.8698	0.9119
Multimodal NonImg [1]	0.8978	0.8139	0.8846	0.8886	0.7619	0.8310	0.6577	0.8886
Multimodal Fusion [1]	0.9548	0.8770	0.9528	0.9051	0.8857	0.9045	0.8828	0.9051
DSA 3D CNN [21]	0.9477	0.8624	0.9307	0.9007	0.8596	0.8712	0.8314	0.9007
Transformer [3]	0.9830	0.5913	0.6008	0.9839	0.9638	0.8064	0.4275	0.9839
Dynamic Image [2]	0.9753	0.9180	0.9737	0.9286	0.9367	0.9079	0.9340	0.9286

	best AP		best Benefit		best Sensitivity		best Benefit	
	AP	Benefit	AP	Benefit	Sensitivity	Benefit	Sensitivity	Benefit
Multimodal MRI [1]	0.9186	0.8965	0.8863	0.9119	0.8595	0.9119	0.8595	0.9119
Multimodal NonImg [1]	0.8241	0.8128	0.7993	0.8886	0.7503	0.8173	0.5466	0.8886
Multimodal Fusion [1]	0.9280	0.8679	0.9255	0.9051	0.8518	0.9025	0.8386	0.9051
DSA 3D CNN [21]	0.9295	0.8624	0.9060	0.9007	0.9230	0.8518	0.8929	0.9007
Transformer [3]	0.9859	0.5913	0.6135	0.9839	0.9825	0.8225	0.9474	0.9839
Dynamic Image [2]	0.9748	0.9180	0.9702	0.9286	0.9926	0.7379	0.9213	0.9286

	best Specificity		best Benefit	
	Specificity	Benefit	Specificity	Benefit
Multimodal MRI [1]	0.9223	0.8905	0.9191	0.9119
Multimodal NonImg [1]	0.8488	0.8251	0.7404	0.8886
Multimodal Fusion [1]	0.9227	0.9045	0.9190	0.9051
DSA 3D CNN [21]	0.8459	0.8377	0.7673	0.9007
Transformer [3]	0.9877	0.7204	0.0617	0.9839
Dynamic Image [2]	0.9968	0.8224	0.9468	0.9286

Table 3
Model hyperparameters. Since the Multimodal NonImg and Multimodal Fusion models are based on the CatBoost regressor, there is no need to set batch size, optimizer, or loss function.

Model	Learning rate	Batch size	Epochs	Optimizer	Loss function
Multimodal MRI [1]	0.001	3	100	Adam	MSE
Multimodal NonImg [1]	0.05	–	100	–	–
Multimodal Fusion [1]	0.05	–	100	–	–
DSA 3D CNN [21]	0.000015	4	100	Adam	Cross entropy
Transformer [3]	0.0001	4	40	Adam	Cross entropy
Dynamic Image [2]	0.00001	16	100	Adam	Cross entropy

Table 4
Non-imaging data column meaning.

Column	Meaning
RID	Unique identifier of subject
VISCODE	Follow-up time
filename	The corresponding MRI file name
COG	Sample classification
Other	Feature

model with a different level of training, and the different shapes of the points distinguish between different model architectures.

From the trend of the scatter plot, it can be seen that when the classification evaluation metrics reach higher values, the benefit metrics are also high. This indicates that when the model performs well in classification, the benefits for AD patients are significant. However, when the classification evaluation metrics are not very high, there is not always a linear relationship between the classification metrics and the benefit metrics. All classification metrics of the DSA 3D CNN and Dynamic Image models show a clear positive correlation with the benefit metrics, particularly evident with the Accuracy metric. The Accuracy metric of the Multimodal MRI model also shows a certain positive correlation with the benefit metrics, while other models did not show this relationship. This implies that when the classification performance of a model did not reach a high level, one could not simply select the best classification model as the one that provides the highest benefit to patients. Focusing solely on classification performance during model training might overlook models that were truly beneficial to patients.

It is noteworthy that the Specificity metric of the Transformer model exhibited a tendency for a negative correlation with the benefit metric, which contrasts with other models. Furthermore, when the Specificity values of multiple intermediate models are similar, the benefit values can differ significantly. This may be due to the fact that Specificity reflects the model's classification accuracy for non-AD subjects, while the increase in benefit is related to AD subjects. When the model prioritizes identifying non-AD subjects and neglects the recognition of AD subjects, the benefit value tends to be lower. Conversely, high classification accuracy across all categories is necessary to achieve high values for both Specificity and benefit metrics. This further underscores the importance of not relying solely on classification evaluation metrics when selecting the most beneficial model for patients.

6. Conclusion

To the best of our knowledge, in this work, we are the first to associate AD (Alzheimer's Disease) diagnostic algorithms with clinical outcomes for evaluation, revealing the limitations of current mainstream AD algorithms and providing guidance for future development. However, our work have limitations. Due to challenges in clinical trials, we did not evaluate the algorithms in a real clinical environment but used cognitive improvement from clinical follow-ups as a proxy outcome, which may introduce bias. Additionally, our evaluation used data solely from the ADNI database, limiting patient diversity. To address these issues, we plan to create a hybrid evaluation system combining

real-world and simulated data, expanding the scope to broader regions to reduce bias.

CRediT authorship contribution statement

Jiyue Xie: Data curation, Methodology, Resources, Validation, Visualization, Writing – original draft. **Wenjing Liu:** Supervision, Writing – original draft, Writing – review & editing. **Li Ma:** Writing – review & editing. **Caiqin Yao:** Writing – review & editing. **Qi Liang:** Writing – review & editing. **Suqin Tang:** Writing – review & editing. **Yunyou Huang:** Conceptualization, Formal analysis, Writing – review & editing.

Declaration of competing interest

The author Yunyou Huang is founding editor for BenchCouncil Transactions on Benchmarks, Standards and Evaluations and was not involved in the editorial review or the decision to publish this article. The other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

We thank MGCS for providing the reference analysis of this work [22]. This work was supported by the National Natural Science Foundation of China (No. U21A20474 to S.T.).

References

[1] S. Qiu, M.I. Miller, P.S. Joshi, J.C. Lee, C. Xue, Y. Ni, Y. Wang, I. De Anda-Duran, P.H. Hwang, J.A. Cramer, et al., Multimodal deep learning for Alzheimer's disease dementia assessment, Nat. Commun. 13 (1) (2022) 3404.

[2] X. Xing, G. Liang, H. Blanton, M.U. Rafique, C. Wang, A.-L. Lin, N. Jacobs, Dynamic image for 3D MRI image Alzheimer's disease classification, in: European Conference on Computer Vision, Springer, 2020, pp. 355–364.

[3] Y. Zhang, K. Sun, Y. Liu, D. Shen, Transformer-based multimodal fusion for early diagnosis of Alzheimer's disease using structural MRI and PET, in: 2023 IEEE 20th International Symposium on Biomedical Imaging, ISBI, IEEE, 2023, pp. 1–5.

[4] N.S. Tyler, C.M. Mosquera-Lopez, L.M. Wilson, R.H. Dodier, D.L. Branigan, V.B. Gabo, F.H. Guillot, W.W. Hilts, J. El Youssef, J.R. Castle, et al., An artificial intelligence decision support system for the management of type 1 diabetes, Nat. Metab. 2 (7) (2020) 612–619.

[5] M. Komorowski, L.A. Celi, O. Badawi, A.C. Gordon, A.A. Faisal, The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care, Nature Med. 24 (11) (2018) 1716–1720.

[6] R. Adams, K.E. Henry, A. Sridharan, H. Soleimani, A. Zhan, N. Rawat, L. Johnson, D.N. Hager, S.E. Cosgrove, A. Markowski, et al., Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis, Nature Med. 28 (7) (2022) 1455–1460.

[7] S. Gavrilova, I. Kolykhalov, N. Selezneva, Y. Kalyn, G. Jarikov, N. Mikhailova, A. Bratsoun, Clinical efficacy of exelon in patients with Alzheimer's disease, Eur. Neuropsychopharmacol. (9) (1999) 330–331.

[8] D.S. Geldmacher, Donepezil (aricept®) for treatment of Alzheimer's disease and other dementing conditions, Expert. Rev. Neurother. 4 (1) (2004) 5–16.

[9] G. Razay, G.K. Wilcock, Galantamine in Alzheimer's disease, Expert. Rev. Neurother. 8 (1) (2008) 9–17.

[10] F. Smith, Mixed-model analysis of incomplete longitudinal data from a high-dose trial of tacrine (cognex®) in Alzheimer's patients, J. Biopharm. Statist. 6 (1) (1996) 59–67.

[11] R. Wolz, A.J. Schwarz, K.R. Gray, P. Yu, D.L. Hill, Alzheimer's Disease Neuroimaging Initiative, et al., Enrichment of clinical trials in MCI due to AD using markers of amyloid and neurodegeneration, Neurology 87 (12) (2016) 1235–1241.

[12] M.R. Arbabshirani, B.K. Fornwalt, G.J. Mongelluzzo, J.D. Suever, B.D. Geise, A.A. Patel, G.J. Moore, Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration, NPJ Digit. Med. 1 (1) (2018) 9.

[13] S. Fathi, M. Ahmadi, A. Dehnad, Early diagnosis of Alzheimer's disease based on deep learning: A systematic review, Comput. Biol. Med. 146 (2022) 105634.

[14] H.-I. Suk, S.-W. Lee, D. Shen, Alzheimer's Disease Neuroimaging Initiative, et al., Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis, NeuroImage 101 (2014) 569–582.

[15] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M.J. Fulham, et al., Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease, IEEE Trans. Biomed. Eng. 62 (4) (2014) 1132–1140.

[16] F.J. Martinez-Murcia, A. Ortiz, J.-M. Gorriz, J. Ramirez, D. Castillo-Barnes, Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders, IEEE J. Biomed. Heal. Inform. 24 (1) (2019) 17–26.

[17] W. Feng, N.V. Halm-Lutterodt, H. Tang, A. Mecum, M.K. Mesregah, Y. Ma, H. Li, F. Zhang, Z. Wu, E. Yao, et al., Automated MRI-based deep learning model for detection of Alzheimer's disease process, Int. J. Neural Syst. 30 (06) (2020) 2050032.

[18] M. Raza, M. Awais, W. Ellahi, N. Aslam, H.X. Nguyen, H. Le-Minh, Diagnosis and monitoring of Alzheimer's patients using classical and deep learning techniques, Expert Syst. Appl. 136 (2019) 353–364.

[19] C. Wattmo, Å.K. Wallin, E. Londos, L. Minthon, Predictors of long-term cognitive outcome in Alzheimer's disease, Alzheimer's Res. Ther. 3 (2011) 1–13.

[20] R.A. Sperling, P.S. Aisen, L.A. Beckett, D.A. Bennett, S. Craft, A.M. Fagan, T. Iwatsubo, C.R. Jack Jr., J. Kaye, T.J. Montine, et al., Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimer's Dement. 7 (3) (2011) 280–292.

[21] E. Hosseini-Asl, G. Gimel'farb, A. El-Baz, Alzheimer's disease diagnostics by a deeply supervised adaptable 3D convolutional network, 2016, arXiv preprint arXiv:1607.00556.

[22] Y. Huang, J. Zhao, D. Cui, Z. Yang, B. Xia, Q. Liang, W. Liu, L. Ma, S. Tang, T. Hao, et al., Quantifying the dynamics of harm caused by retracted research, 2024, arXiv preprint arXiv:2501.00473.