FairAD-XAI: Evaluation Framework for Explainable AI Methods in Alzheimer's Disease Detection with Fairness-in-the-loop

Quoc-Toan Nguyen
quoctoan.nguyen@student.uts.edu.au
GrapheneX-UTS Human-Centric Artificial Intelligence Centre, Faculty of Engineering and Information Technology, University of Technology Sydney
Sydney, NSW, Australia

Linh Le
linh.le@uts.edu.au
GrapheneX-UTS Human-Centric Artificial Intelligence Centre, Faculty of Engineering and Information Technology, University of Technology Sydney
Sydney, NSW, Australia

Xuan-The Tran
xuanthe.tran@student.uts.edu.au
GrapheneX-UTS Human-Centric Artificial Intelligence Centre, Faculty of Engineering and Information Technology, University of Technology Sydney
Sydney, NSW, Australia

Thomas Do*
thomas.do@uts.edu.au
GrapheneX-UTS Human-Centric Artificial Intelligence Centre, Faculty of Engineering and Information Technology, University of Technology Sydney
Sydney, NSW, Australia

Chin-Teng Lin*
chin-teng.lin@uts.edu.au
GrapheneX-UTS Human-Centric Artificial Intelligence Centre, Faculty of Engineering and Information Technology, University of Technology Sydney
Sydney, NSW, Australia

Abstract

Despite significant progress in model developments, evaluating eXplainable Artificial Intelligence (XAI) remains elusive and challenging in Alzheimer's Disease (AD) detection using modalities from low-cost or wearable devices. This paper introduces a fine-grained validation framework named 'FairAD-XAI', which provides a comprehensive assessment through twelve properties of explanations, forming a detailed Likert questionnaire. This framework ensures a thorough evaluation of XAI methods, capturing their fairness aspects and supporting the improvement of how humans assess the reliability and transparency of these methods. Moreover, fairness in XAI evaluation is critical, as users from diverse demographic backgrounds may have different perspectives and perceptions towards the system. These variations can lead to biases in human-grounded evaluations and, subsequently, biased decisions from the AI system when deploying. To mitigate this risk, we installed two fairness metrics tailored to assess and ensure fairness in XAI evaluations, promoting more equitable outcomes. In summary, the proposed 'FairAD-XAI' framework provides a comprehensive tool for evaluating XAI methods and assessing the essential aspect of fairness. This makes it a multifactoral tool for developing unbiased XAI methods for AI-based AD detection tools, ensuring these technologies are both effective and equitable.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **General and reference** → **Evaluation**; • **Human-centered computing** → *Mobile computing*.

*Corresponding authors.
UbiComp Companion '24: , October 5–9, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-1058-2/24/10
https://doi.org/10.1145/3675094.3678998

Keywords

XAI; Evaluation; Fairness; Alzheimer's; Telemedicine; Digital Health

ACM Reference Format:
Quoc-Toan Nguyen, Linh Le, Xuan-The Tran, Thomas Do*, and Chin-Teng Lin*. 2024. FairAD-XAI: Evaluation Framework for Explainable AI Methods in Alzheimer's Disease Detection with Fairness-in-the-loop. In *Companion of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '24), October 5–9, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3675094.3678998

1 Introduction

Alzheimer's Disease (AD) is a degenerative brain condition that progressively leads to dementia, mainly impacting people over the age of 65 [1]. AD detection can be challenging due to its complex signs [2]. However, the advancements in Clinical Decision Support Systems (CDSS) using Artificial Intelligence (AI) (AI-based CDSS) pave a promising path for both healthcare professionals and individuals to early detect and slow down the progression properly, improving Quality of Life (QoL) [2, 3]. There are many modalities to detect or screen AD using AI [4, 5], such as Magnetic resonance imaging (MRI), Cerebrospinal Fluid (CSF), and Positron Emission Tomography (PET). However, the techniques above are limited to the masses due to their high cost.

Therefore, cost-effective techniques can be alternatives to deliver screening access to support healthcare for more individuals, such as Electroencephalogram (EEG) [6, 7] or handwriting/drawing [8, 9]. Moreover, in recent years, digital biomarkers [3] utilising mobile and wearable technologies [10], like smartphones, smartwatches, and smart suits, have offered a noticing option to screen AD efficiently. This is due to their widespread use, instant information access, and advanced onboard sensors leveraging their capability to monitor physical and cognitive status. However, AI-based CDSS

for AD detection in general, or the ones using mobile or wearable modalities, are not yet widely adopted clinically due to their black-box nature. The outcomes from AI-based CDSS should be explained understandably by eXplainable AI (XAI) methods for various stakeholders, such as developers, medical professionals, and lay users.

In particular, there is a research gap in the evaluation of current XAI methods applied in early AD detection using these techniques [11]. Firstly, XAI researchers often rely on self-intuition without consulting medical professionals [12]. The derived AI explanations are data-driven and lack domain experts' input. Secondly, explanations vary with domain expertise, causing confusion and doubt when explanations contradict intuition. Thereby, objective validation with ground truth data is needed [13]. Importantly, deploying inefficient and biased XAI methods in clinical settings for AD detection can lead stakeholders to make inaccurate and biased decisions. This, in turn, may unfairly penalise minority groups, as these inaccurate decisions disproportionately affect them.

Thus, evaluating to ensure fairness in XAI method evaluation for AD plays a vital role in the reliability of AI-based CSDDs because these evaluations are mainly accorded to the human-grounded view of the users. Fairness in XAI means providing explanations that are equitable and unbiased, ensuring that all user groups receive accurate and just information. Each demographic group may have different perceptions, and explainability is a non-binary characteristic [14]. Hence, assessing XAI methods based on 'formal fairness' [15] and multiple explanation quality properties [14] is crucial from the users' point of view to have a final general good mental model [16] (trusting the AI system and performing well when using it), mitigate biases and promote equitable outcomes in the detection of AD, addressing potential disparities in AI-based CSDDs. Therefore, this study proposes a 'FairAD-XAI' (fairness-in-the-loop XAI evaluation for AD) framework with the following contributions:

- Develop the 'FairAD-XAI' framework, including a Likert questionnaire to evaluate XAI methods in AD detection. This framework assesses explanation quality and fairness across demographic groups to mitigate bias and ensure reliable, equitable AI-based CSDDs for early AD detection.
- To demonstrate the practical application of the FairAD-XAI framework, an AI model for AD detection was developed.

2 Methodologies

As described in Figure 1, the proposed FairAD-XAI is placed after the XAI method being applicable to multiple explanation types [14]. In particular, it explains results from the AD detection model using low-cost/mobile/wearable modalities, acting as an evaluation phase in the process. Characteristics of each modality can be found from the research by Kourtis *et al.*[3]. The methodologies are described in the following Sections 2.1, 2.2, and 2.3.

2.1 The Likert Questionnaire

To begin with, regarding XAI methods' evaluation for AD detection, addressing the question of 'What to evaluate?' and 'Is the XAI method good?' sets the tone for the questionnaire. In this section, twenty-four Likert questionnaire questions have been developed leveraging the notion of 'co-twelve', twelve explanation quality

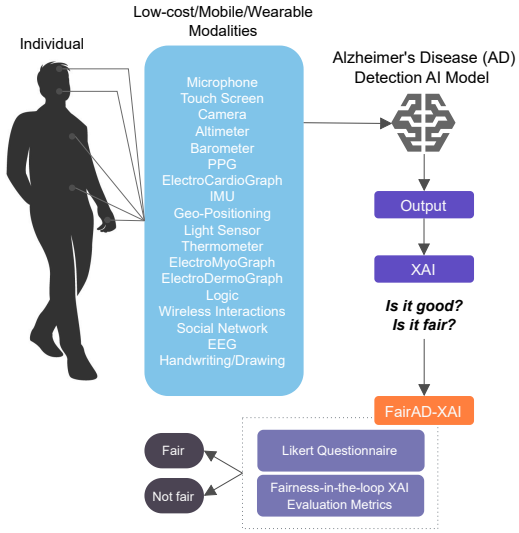

Figure 1: Illustration of Proposed FairAD-XAI AD Detection AI Model using Low-cost/Mobile/Wearable Modalities. PPG (Photoplethysmography), IMU (Inertia Measurement Unit).

properties proposed by Nauta *et al.* [14]. Co-twelve is grouped by 3 dimensions, including 'Content', 'Presentation' and 'User'. Table 1 details all proposed questions, with the Likert scale from 0 to 5 applied for all questions, obtaining the final scores to range from 0 to 120 with six overall evaluating ranges. The total score is described as LQ_{XAI}. The higher LQ_{XAI} is, the better the XAI method. These are descriptions of the overall score range:

- **Poor (0-20):** The XAI method is highly inadequate and fails to meet basic requirements.
- **Subpar (21-40):** The XAI method is below standard and lacks important elements.
- **Average (41-60):** The XAI method is fair but needs improvement.
- **Good (61-80):** The XAI method is solid and meets most expectations.
- **Very Good (81-100):** The XAI method is highly satisfactory and covers most aspects well.
- **Excellent (101-120):** The XAI method is outstanding and comprehensively addresses all criteria.

To be deployed in a real-life setting for AD detection, the suggested average overall range for the XAI method should be 'Good' or higher, with an overall score of at least 61 from all demographic groups.

The following are the descriptions of each dimension with according properties:

Dimension 1: Content

This dimension addresses the question of 'How accurately does the explanation reflect the model's behaviour, and is it comprehensive, consistent, smooth, differentiable, and simple enough to be easily understood?'

Correctness: Refers to how accurately an explanation reflects the behaviour of the predictive model (f). It measures the fidelity of the explanation. This property is about the explanation's descriptive accuracy, not the model's predictive accuracy.

Completeness: Measures how comprehensively the explanation covers the behaviour of f. The explanation should encompass 'the whole truth' in an ideal scenario.

Consistency: Ensures that identical inputs yield identical explanations. This property evaluates the determinism of the explanation method. Consistency involves 'implementation invariance' for methods that only consider input and output, meaning models producing the same outputs should generate the same explanations.

Continuity: Assesses the smoothness of the explanation function produced by the explanation method. A continuous function ensures that minor changes in the input, which result in almost identical model responses, do not cause significant differences in the explanation.

Contrastivity: Evaluates how effectively an explanation can differentiate between various events or outcomes. It aims to explain why a particular event occurred by contrasting it with an alternative event that did not happen.

Covariate Complexity: Refers to how complicated the features in an explanation are and how they relate to the outcome. The features should be simple and easy to understand. This means they might differ from the original input features to make the explanation clearer and more understandable.

Dimension 2: Presentation

This dimension addresses the question, 'Is the explanation concise, clear, and well-organised, and does it effectively convey the certainty or probability of the model's predictions?'

Compactness: Addresses the brevity and conciseness of the explanation, considering human cognitive limitations.

Composition: Involves the explanation's format, organisation, and structure to enhance its clarity. It focuses on how the explanation is presented rather than the content itself.

Confidence: Relates to whether the explanation measures certainty or probability. This can involve two aspects: i) the confidence level of the black box model's prediction or ii) the reliability or likelihood of the explanation itself.

Dimension 3: User

This dimension addresses 'How well the explanation is tailored to the user's needs and expertise. Does it align with their existing knowledge and allow for interactive engagement and control?'

Context: Considers how well the explanation is tailored to the user's needs and level of expertise.

Coherence: Evaluates how well the explanation aligns with existing background knowledge and rational evidence. It addresses the reasonableness, plausibility, and agreement with human rationales.

Controllability: Measures how much a user can manage, adjust, or engage with an explanation. It is based on the idea that 'explanations are interactive' and should allow users to influence the model itself [17] and refine them.

2.2 Fairness-in-the-loop XAI Evaluation Metrics

The installed metrics for assessing fairness in evaluating XAI methods for AD detection are employed by tailoring two of the most important metrics in algorithm fairness evaluation [18], including Disparate Impact (*DI*) and Demographic Parity (*DP*). DI measures the ratio of favourable outcomes between different demographic groups. DP assesses whether different demographic groups receive favourable outcomes at similar rates. To evaluate the fairness of XAI methods' evaluation in AD detection, in our proposed framework, the 2 metrics are named DI XAI evaluation (DI_{XAI}) and DP XAI evaluation (DP_{XAI}). Their equations are described below:

$$DI_{XAI} = \frac{\mathbb{E}[LQ_{XAI} \mid S \neq 1]}{\mathbb{E}[LQ_{XAI} \mid S = 1]} \geq \tau_{DI}(0.8, 1.2), \tag{1}$$

$$DP_{XAI} = |\mathbb{E}[LQ_{XAI} \mid S = 1] - \mathbb{E}[LQ_{XAI} \mid S \neq 1]| \leq \tau_{DP}(0, 24). \tag{2}$$

Notations:

- DI_{XAI}: Disparate Impact of the XAI evaluation metric.
- DP_{XAI}: Disparate Parity of the XAI evaluation metric.
- $\mathbb{E}[LQ_{XAI} \mid S = 1]$: Expected (average) scores (LQ_{XAI}) from users in the reference group ($S = 1$).
- $\mathbb{E}[LQ_{XAI} \mid S \neq 1]$: Expected (average) scores (LQ_{XAI}) from users in groups other than the reference group ($S \neq 1$).
- S: Sensitive attribute (e.g., sex, ethnicity, race, marital status, occupation, age).
- $S = 1$: Reference group for the sensitive attribute.
- $S \neq 1$: Other group for the sensitive attribute.
- τ_{DI}: Threshold for DI_{XAI}, typically in the acceptable range of 0.8 to 1.2.
- τ_{DP}: Threshold for DP_{XAI}, typically in the acceptable range of 0 to 24.

To illustrate, DI_{XAI} is calculated by taking the ratio of the average LQ_{XAI} between the reference group ($S \neq 1$) and the other group(s) ($S = 1$), ensuring that the expected LQ_{XAI} is proportionate across different demographic groups. Whereas, DP_{XAI} is calculated by taking the absolute difference between the average LQ_{XAI} of the reference and other group(s), ensuring that different demographic groups have an equal chance of receiving a LQ_{XAI}. Acceptable thresholds, τ_{DI} and τ_{DP} follow the '80% rule' [18], requiring any group's acceptance rate to be at least 80% of the highest rate. These metrics ensure that the evaluations of XAI methods are both fair and unbiased, enhancing the reliability and transparency of these methods across various user demographics. The development team can decide which of these metrics to use in their projects. Importantly, the evaluation process can be conducted with one or multiple sensitive attributes, depending on the project's resources and objectives.

2.3 Implementing Algorithm

Algorithm 1 is used to detail the workflow of implementing the proposed 'FairAD-XAI' framework. It begins with defining stakeholder(s) Z, described in [19] and sensitive attributes such as sex, ethnicity, race, marital status, occupation, and age to have different groups G of Z. The total number of evaluating samples (N) is set, along with the number of evaluation loops (k). Next, n is calculated

Table 1: FairAD-XAI Framework, Likert Questionnaire for Evaluating XAI Methods in AD Detection. The rating assesses each property on a 0 to 5 scale, while the overall score is a comprehensive average, ranging from Poor to Excellent.

Dimension	Property	Question	Scale
Content	Correctness	How well does the explanation reflect the model's behaviour?	**Rating** 0: Not at all 1: Very Little 2: Little 3: Moderate 4: Much 5: Very much **Overall Score** 0 - 20: Poor 21 - 40: Subpar 41 - 60: Average 61 - 80: Good 81 - 100: Very Good 101 - 120: Excellent
		How accurately does the explanation represent the model's decisions?	
	Completeness	How well does the explanation cover the model's behavior?	
		How thorough is the explanation in providing necessary information?	
	Consistency	How consistently do similar inputs yield identical explanations?	
		How reliable is the explanation method across similar cases?	
	Continuity	How smoothly does the explanation respond to minor input changes?	
		How proportional are minor input changes in the explanation?	
	Contrastivity	How well does the explanation differentiate between outcomes?	
		How clearly does the explanation contrast different diagnoses?	
	Covariate Complexity	How understandable are the features in the explanation?	
		How well do the features relate to the diagnosis?	
Presentation	Compactness	How concise is the explanation?	
		How well does the explanation avoid unnecessary complexity?	
	Composition	How well-organized is the explanation?	
		How effectively does the format aid understanding?	
	Confidence	How well does the explanation measure certainty?	
		How clearly does the explanation convey confidence?	
User	Context	How well is the explanation tailored to the user's expertise?	
		How sufficient is the context provided by the explanation?	
	Coherence	How well does the explanation align with AD domain knowledge?	
		How plausible is the explanation?	
	Controllability	How much control does the user have over the explanation?	
		How interactive is the explanation?	

as $\frac{N}{s}k$. The development team then decides which metric to use. Thresholds are established for the two installed evaluation metrics in acceptable ranges: τ_{DI} (0.8,1.2) or τ_{DP} (0,24). Additionally, data of LQ_{XAI} is collected from every user g in each group G). During each of the k evaluation loops, the LQ_{XAI} is calculated based on all samples for the current loop is calculated. Depending on the chosen metric, either DI_{XAI} or DP_{XAI} is calculated for the current loop, and the result is saved. The $final\_average$ of all folds is calculated after all k loops. If the final average meets its respective threshold, the XAI method is labelled by Z as fair or unfair if not. This structured approach ensures a thorough and fair assessment of XAI methods in the context of AD detection, addressing the stated problems.

2.4 Material

Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [20] was used for the model's development. The selected features are five cognitive and functional tests: Clinical Dementia Rating-Sum of Boxes (CDRSB), Alzheimer's Disease Assessment Scale 11 (ADAS11), Alzheimer's Disease Assessment Scale 13 (ADAS13), Alzheimer's Disease Assessment Scale Q4 (ADASQ4), and Mini-Mental State Examination (MMSE). These tests can be done by mobile devices via touch screen, camera, and microphone with medical professionals via telemedicine, in which individuals are supported and interact with medical professionals remotely [21], assessing cognitive and functional functions, effectively aiding in AD detection [22], including memory, attention, and language skills. There are a total of 1094 subjects included in this study, 699 males and 395 females. There are 699 and 395 subjects, with 3577 and 1098 records (4657 in total) used for $model_{AD}$ development for NC and AD, respectively. The training sets contain 80% of the total records; the rest is a test set.

2.5 ML Methods

To support in illustrating the proposed FairAD-XAI framework, a binary AI model for AD detection ($model_{AD}$) with two classes consisting of normal control (NC) and AD was developed. Because the model only utilises five features, there is no need for complex models. Five methods were included in this research due to their high performance and widely applied with effectiveness in the domain of AD detection [23]. They are Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), K-nearest Neighbours (KNN), and eXtreme Gradient Boosting (XGBoost). Evaluation metrics consist of accuracy (Acc), true positive rate (TPR), false positive rate (FPR), and F1score. More information on these metrics can be referred to in this paper [24]

3 Results

3.1 ML Model

Table 2 details the results of these four methods for $model_{AD}$. As we can see, RF outperformed other methods in three out of four metrics, being the best-performing method for $model_{AD}$. Therefore, the results from this model will be explained using the XAI method

Algorithm 1 Implementation of FairAD-XAI Framework for XAI Methods' Evaluations in AD Detection

Develop the AI model for AD detection ($model_{AD}$).
Evaluate $model_{AD}$ and obtain the output results.
Decide stakeholder(s), Z, being accountable for the decision.
Define sensitive attributes: sex, ethnicity, race, marital status, occupation, and age to have groups of users, G.
Set N as the total number of evaluating samples (explained results from outputs of $model_{AD}$) (e.g., $N = 50$).
Set k as the number of evaluation loops for separation (e.g., $k = 5$)
Calculate $n = \frac{N}{k}$ (e.g., $n = 10$)
Choose DI_{XAI} or DP_{XAI} to use
Set thresholds τ_{DI} in range (0.8,1.2) or τ_{DP} in range (0,24) depending on the chosen metric
Collect LQ_{XAI} data from $g \in G$
Initialise an empty list fold_averages
for each of the k evaluation loops **do**
    Collect n samples for the current loop
    Calculate the LQ_{XAI} for the current loop based on n
    if using DI_{XAI} **then**
        Calculate DI_{XAI} for the current loop
        Add the DI_{XAI} result to fold_averages
    else if using DP_{XAI} **then**
        Calculate DP_{XAI} for the current loop
        Add the DP_{XAI} result to fold_averages
    end if
end for
Calculate the final average of all folds final_average = $\frac{\sum \text{fold\_averages}}{k}$
if using DI_{XAI} **then**
    if final_average $\geq 1 - \tau_{DI}$ **then**
        Z label the XAI method as fair
    else
        Z label the XAI method as unfair
    end if
else if using DP_{XAI} **then**
    if final_average $\leq \tau_{DP}$ **then**
        Z label the XAI method as fair
    else
        Z label the XAI method as unfair
    end if
end if

Table 2: Results of Methods Developed for $model_{AD}$. Bold Value Represents the Best Value in The Metric.

Method	Acc	TPR	FPR	F1score
DT	98.71	98.21	1.78	98.21
RF	98.39	**98.67**	**1.32**	**98.51**
SVM	**98.82**	98.44	1.57	98.36
KNN	98.50	97.29	2.70	97.89
XGBoost	98.60	98.30	1.69	98.07

and then evaluated by the FairAD-XAI framework. Figure 2 is the confusion matrix of $model_{AD}$ (RF) performed on the test set.

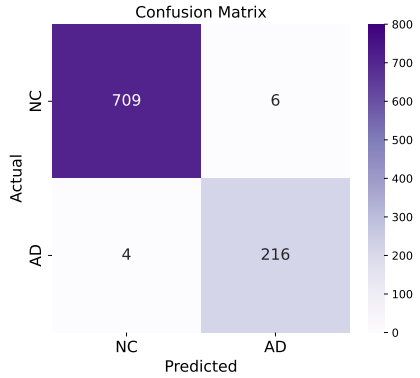

Figure 2: Confusion Matrix of $model_{AD}$ Performed on The Test Set.

3.2 XAI Evaluation using FairAD-XAI

Using the developed $model_{AD}$ (RF) with classes NC and AD, we evaluate the XAI method using the OmniXAI library [25]. This utilizes Local Interpretable Model-Agnostic Explanations (LIME) as detailed by Ribeiro *et al.* [26]. Figures 3 and 4 show LIME results for NC and AD predictions, respectively. Features are the model's input features, with importance scores calculated by perturbing feature values and observing prediction changes. Higher scores indicate greater influence, with positive scores supporting the predicted class and negative scores opposing it, pushing the prediction away from the predicted class [26]; 'Positive' on the right means supporting the prediction of the label class. For more details, refer to the OmniXAI framework [25].

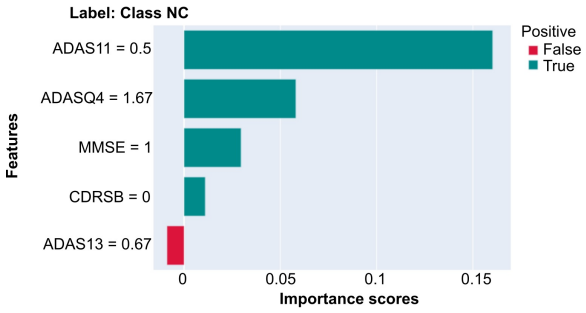

Figure 3: Example of XAI Result of Output from $model_{AD}$ with NC.

The following contents are for the demonstration of how to use FairAD-XAI to assess the XAI results, illustrated in Algorithm 1, assuming that the results like Figure 3 and 4 are samples in N. It is important to emphasise that all stakeholders, users, and data in the example below are hypothetical and are provided solely for illustrative purposes to demonstrate the practical implementation of FairAD-XAI. Stakeholders Z include two team leaders: one from

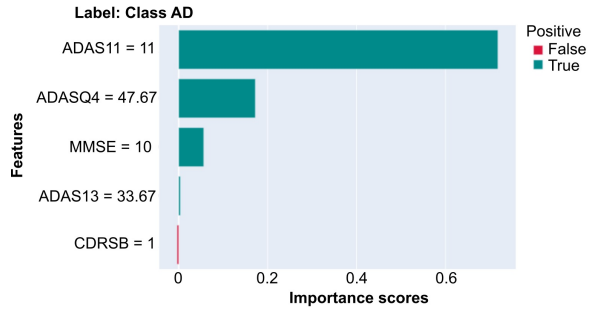

Figure 4: Example of XAI Result of Output from $model_{AD}$ with AD.

the AI development team and one from the medical professional team. Groups G contain individuals (g) grouped by sex (Male, Female), race (White, Asian, Black, Others), and marital status (Never married, Widowed, Divorced, Separated, Married). These example subsets of G are selected as prior research in fairness literature has identified bias related to these characteristics [27, 28].

(1) Develop and evaluate $model_{AD}$.
(2) Identify stakeholders (Z): AI team leader and medical team leader.
(3) Define sensitive attributes: sex, race, marital status.
(4) Set total samples $N = 50$ and evaluation loops $k = 5$.
(5) Calculate samples per loop $n = 10$.
(6) Choose metric: Disparate Impact (DI_{XAI}).
(7) Set threshold τ_{DI} (0.8, 1.2).
(8) Set fold_averages storing the values of each G each fold
(9) Collect LQ_{XAI} data from groups G.

Loop Calculations

Collect data from subsets of G (sensitive attributes) with g

Loop 1: - Collect LQ_{XAI} for each group using 10 samples:

Sex:

- Male: $LQ_{XAI} = \{100, 98, 102, 96, 101\}$ (5 users)
- Female: $LQ_{XAI} = \{104, 106, 102, 110, 105\}$ (5 users)

Race:

- White: $LQ_{XAI} = \{100, 101, 99, 98, 104\}$ (5 users)
- Asian: $LQ_{XAI} = \{104, 105, 103, 100, 99\}$ (5 users)
- Black: $LQ_{XAI} = \{100, 102, 98, 104, 104\}$ (5 users)
- Others: $LQ_{XAI} = \{102, 103, 101, 99, 108\}$ (5 users)

Marital Status:

- Never married: $LQ_{XAI} = \{102, 103, 101, 95, 99\}$ (5 users)
- Widowed: $LQ_{XAI} = \{100, 99, 101, 112, 97\}$ (5 users)
- Divorced: $LQ_{XAI} = \{104, 105, 103, 110, 98\}$ (5 users)
- Separated: $LQ_{XAI} = \{102, 101, 103, 115, 100\}$ (5 users)
- Married: $LQ_{XAI} = \{100, 99, 101, 98, 112\}$ (5 users)

LQ_{XAI} is averaged from users g. For sex, there are only two G, so we calculate the DI_{XAI} by taking the ratio of the LQ_{XAI}. For the sensitive attribute of race, we have multiple groups: White, Asian, Black, and Others. We calculate the pairwise DI_{XAI} for each combination of these groups and compute the DI_{XAI} for race as the average of these pairwise values. Similarly, for the sensitive attribute of marital status, we have multiple groups: Never married, Widowed, Divorced, Separated, and Married. We calculate the pairwise DI_{XAI} for each combination of these groups and compute the DI_{XAI} for marital status as the average of these pairwise values.

These DI_{XAI} values are added to fold_averages for each subset of G with sensitive attributes (sex, race, marital status) and move to the next loop with other samples in n.

Repeat for Loops 2-5. Finally, the overall average DI_{XAI} for each subset of G (sensitive attributes) is calculated by averaging the fold_averages across all loops.

Final Calculation

Calculate the final average of each G all folds (loops). Assuming that the final average DI_{XAI} is 1.02 for sex, 0.9 for race, and 1.12 for marital status.

Decision

Since all the final averages are within the threshold (0.8, 1.2), the XAI method is labelled as fair by Z.

Overall Score: Additionally, this can be calculated by storing the average LQ_{XAI} from subsets of G in fold_averages instead of DI_{XAI}, then quality is concluded based on the range in Section 2.1.

4 Conclusion and Discussion

In conclusion, the proposed 'FairAD-XAI' framework is a vital element in tackling the current problems of XAI methods for early AD detection [11]. It can be a reference for multiple ML tasks, not limited to classification in the example. Moreover, it can be also a versatile, comprehensive and quantitative method for evaluating the effectiveness of XAI techniques using various explanation types [14]. Furthermore, this adaptable framework makes it possibly applicable beyond AD detection. Projects in other domains can adopt and tailor the 'FairAD-XAI' framework to fit their specific needs, leveraging its emphasis on explanation quality to assess the reliability and fairness of their XAI implementations.

ACKNOWLEDGMENTS

This work was partially funded by the Australian Research Council (ARC) through discovery grants DP210101093 and DP220100803, and the Australian National Health and Medical Research Council (NHMRC) Ideas Grant APP2021183. Additional support was provided by the UTS Human-Centric AI Centre, sponsored by GrapheneX (2023-2031). Part of the research received funding from the Australia Defence Innovation Hub under Contract No. P18-650825, the Australian Cooperative Research Centres Projects (CRC-P) Round 11 CRCPXI000007. We would like to send our appreciation to all of the funders.

References

[1] Zhaomin Yao, Hongyu Wang, Wencheng Yan, Zheling Wang, Wenwen Zhang, Zhiguo Wang, and Guoxu Zhang. Artificial intelligence-based diagnosis of alzheimer's disease with brain mri images. *European Journal of Radiology*, page 110934, 2023.

[2] Shaker El-Sappagh, Jose M Alonso-Moral, Tamer Abuhmed, Farman Ali, and Alberto Bugarín-Diz. Trustworthy artificial intelligence in alzheimer's disease: state of the art, opportunities, and challenges. *Artificial Intelligence Review*, 56(10):11149–11296, 2023.

[3] Lampros C Kourtis, Oliver B Regele, Justin M Wright, and Graham B Jones. Digital biomarkers for alzheimer's disease: the mobile/wearable devices opportunity. *NPJ digital medicine*, 2(1):9, 2019.

[4] Laura M Winchester, Eric L Harshfield, Liu Shi, AmanPreet Badhwar, Ahmad Al Khleifat, Natasha Clarke, Amir Dehsarvi, Imre Lengyel, Ilianna Lourida, Christopher R Madan, et al. Artificial intelligence for biomarker discovery in alzheimer's disease and dementia. *Alzheimer's & Dementia*, 19(12):5860–5871, 2023.

[5] Kaj Blennow and Henrik Zetterberg. Biomarkers for alzheimer's disease: current status and prospects for the future. *Journal of internal medicine*, 284(6):643–663, 2018.

[6] Siuly Siuly, Ömer Faruk Alçin, Hua Wang, Yan Li, and Peng Wen. Exploring rhythms and channels-based eeg biomarkers for early detection of alzheimer's disease. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[7] Xuan-The Tran, Linh Le, Quoc Toan Nguyen, Thomas Do, and Chin-Teng Lin. EEG-SSM: Leveraging State-Space Model for Dementia Detection, 2024.

[8] Nicole D Cilia, Claudio De Stefano, Francesco Fontanella, and Sabato Marco Siniscalchi. How word semantics and phonology affect handwriting of alzheimer's patients: a machine learning based analysis. *Computers in Biology and Medicine*, 169:107891, 2024.

[9] Irene Azzali, Nicole D Cilia, Claudio De Stefano, Francesco Fontanella, Mario Giacobini, and Leonardo Vanneschi. Automatic feature extraction with vectorial genetic programming for alzheimer's disease prediction through handwriting analysis. *Swarm and Evolutionary Computation*, 87:101571, 2024.

[10] Chin-Teng Lin and Tien-Thong Nguyen Do. Direct-sense brain–computer interfaces and wearable computers. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(1):298–312, 2020.

[11] Vimbi Viswan, Noushath Shaffi, Mufti Mahmud, Karthikeyan Subramanian, and Faizal Hajamohideen. Explainable artificial intelligence in alzheimer's disease classification: A systematic review. *Cognitive Computation*, 16(1):1–44, 2024.

[12] Yubo Kou and Xinning Gui. Mediating community-ai interaction through situated explanation: the case of ai-led moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–27, 2020.

[13] Moritz Böhle, Fabian Eitel, Martin Weygandt, and Kerstin Ritter. Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification. *Frontiers in aging neuroscience*, 11:456892, 2019.

[14] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

[15] Luca Deck, Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. A critical survey on fairness benefits of explainable ai. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1579–1595, 2024.

[16] Roberto Confalonieri and Jose M Alonso-Moral. An operational framework for guiding human evaluation in explainable and trustworthy ai. *IEEE Intelligent Systems*, 2023.

[17] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

[18] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[19] Robert R Hoffman, Shane T Mueller, Gary Klein, Mohammadreza Jalaeian, and Connor Tate. Explainable ai: roles and stakeholders, desirements and challenges. *Frontiers in Computer Science*, 5:1117848, 2023.

[20] Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

[21] Wei Bo, Suzanne S Sullivan, Xiaoyu Zhang, Mingchen Gao, and Wenyao Xu. A telemedicine analytic framework for fully and semi-automatic alzheimer's disease screening using clock drawing test. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[22] Mohammad Al Olaimat, Jared Martinez, Fahad Saeed, Serdar Bozdag, and Alzheimer's Disease Neuroimaging Initiative. Ppad: A deep learning architecture to predict progression of alzheimer's disease. *Bioinformatics*, 39(Supplement_1):i149–i157, 2023.

[23] Rahul Kumar and Chandrashekhar Azad. Comprehensive overview of alzheimer's disease utilizing machine learning approaches. *Multimedia Tools and Applications*, pages 1–53, 2024.

[24] Steven A Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A Riegler, Pål Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, 12(1):5979, 2022.

[25] Wenzhuo Yang, Hung Le, Tanmay Laud, Silvio Savarese, and Steven CH Hoi. Omnixai: A library for explainable ai. *arXiv preprint arXiv:2206.01612*, 2022.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[27] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[28] Darshali A Vyas, Leo G Eisenstein, and David S Jones. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms, 2020.