Revealing the Roles of Part-of-Speech Taggers in Alzheimer’s Disease Detection: A Scientific Discovery Using One-intervention Causal Explanation

Bingyang Wen, Ning Wang, K.P. Subbalakshmi, R. Chandramouli

Submitted to: JMIR Formative Research
on: January 27, 2022

Disclaimer: **© The authors. All rights reserved.** This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Revealing the Roles of Part-of-Speech Taggers in Alzheimer’s Disease Detection: A Scientific Discovery Using One-intervention Causal Explanation

Bingyang Wen[1] MSc; Ning Wang[1] PhD; K.P. Subbalakshmi[1] PhD; R. Chandramouli[1] PhD

[1]Department of Electrical and Computer Engineering Stevens Institute of Technology Hoboken US

Corresponding Author:
Bingyang Wen MSc
Department of Electrical and Computer Engineering
Stevens Institute of Technology
Room 315
524 River Street
Hoboken
US

Abstract

Background: Machine learning-based Alzheimer's detection using natural language processing has drawn increasing attention because of its low cost compared with traditional methods. However, most of these models are black-boxes, and the decision mechanisms of the AI are obscure. In some fields like medicine, this obscurity gets in the way of widespread adoption. This has led to the development of a new class of techniques that are generally referred to as explainable AI (XAI). One approach to this problem is counter-factual explanations which answer “what if” questions like “What would have happened to Y, had I not done X?”.

Objective: This study aims to improve the transparency of a the-state-of-art language-based Alzheimer’s disease (AD) detection model and discover linguistic biomarkers that are indicative of AD and hence can be used as tools for automated diagnosis of AD.

Methods: In this paper, a new explainable artificial intelligence (XAI) method is proposed and named one-intervention counterfactual explanation (OICE). This method works on the state-of-the-art language-based, deep learning method for AD detection and provides an explanation of that method. The proposed OICE incorporates causal factors among the features used in the detection of AD, to provide more transparency of the AI’s decision. This is in contrast to conventional counterfactual explanation methods which do not incorporate causal mechanisms. An understanding of causal factors can go beyond mere statistical correlation to provide a better understanding of the underlying physical phenomenon. The proposed OICE generates counterfactual explanations from a predefined deep-based structural causal model (SCM). The proposed method generated explanations of the AI’s decision by only intervening on one feature at a time. Since OICE provides explanations for individual samples, we then analyze the counterfactual explanations statistically and define some metrics to quantify the effect of every feature.

Results: We find 11 language level biomarkers for Alzheimer’s disease detection such as adverb, pronoun, noun, preposition, etc. Previous work in psychology and NLP points out adverbs, pronouns, and nouns as potential biomarkers. Our study concurs. We also find new biomarkers that were not reported in previous studies, such as preposition, predeterminer, etc. Our results also reveal how these biomarkers are involved in the diagnostic process from a causal perspective. For example, an on-average 20.2% increase in predeterminer, causes determiner, verb (present particle), and grammatical particles change, resulting in flipping in the diagnosis from control to Alzheimer’s disease. This implies that predeterminer is potentially a strong indicator of the individual’s health and can function as a strong biomarker.

Conclusions: Our findings show consistency with previous works in psychology and natural language processing (NLP). Additionally, we offer a new explanation about how intervening a feature can affect the model's decisions using the pre-defined SCM.

(JMIR Preprints 27/01/2022:36590)

DOI: https://doi.org/10.2196/preprints.36590

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Revealing the Roles of Part-of-Speech Taggers in Alzheimer's Disease Detection: A Scientific Discovery Using One-intervention Causal Explanation

Abstract

Background:
Recently rich computational methods that use deep learning (DL) or machine learning (ML) have been developed using linguistic biomarkers for early-stage Alzheimer's Disease (AD) diagnosis. Moreover, some qualitative and quantitative studies have indicated that some part-of-speech (PoS) features/tags could be good indicators of AD. However, there has not been a systematic attempt to discover the underlying relationships between the PoS features and AD. There has also not been any attempt to quantify the relative importance of these PoS features in detecting AD.

Objective:
Our goal is to disclose the underlying relationship between PoS features and AD. Understand whether PoS features are useful in AD diagnosis and explore which of the PoS features play a vital role in the diagnosis.

Methods:
The DementiaBank, containing 1049 transcripts from 208 AD patients and 243 transcripts from 104 elderly controls is used. 27 parts-of-speech (PoS) features from are extracted from each record. Then, the relationship between AD and each of the PoS features is explored. A transformer-based deep learning model for AD prediction using the PoS features is trained. Then a global explainable artificial intelligence (XAI) method is proposed and used to discover which PoS features were most important in AD diagnosis by the transformer based predictor.
A global (model-level) feature importance measure is derived as a summarization from the local (example-level) feature importance metric, which is obtained using the proposed casually-ware counterfactual explanation method. The unique feature of this method is that it considers causal relations among PoS features and hence can preclude counterfactuals that are improbable and hence result in more reliable explanations.

Results:
The deep learning-based AD predictor achieves an accuracy of 92.2% and an F1-score of 0.955 when distinguishing AD patients from healthy controls. The proposed explanation method identified 12 PoS features as being important to the diagnosis of AD from healthy control. Of these, 3 features have been identified by other researchers in previous work in psychology and natural language processing (NLP). Nine other PoS features have not been previously identified. We believe that this is an interesting finding that can be used in creating tests that might aid in diagnosing AD. Note that although, our method is focused on PoS features, it should be possible to extend to more types of features, perhaps even derived from other biomarkers, like syntactic features.

Conclusions:
The high classification accuracy of the proposed deep-learner indicates that PoS features are

strong clues in AD diagnosis. There are 12 PoS features that are strongly tied to AD and since language is a non-invasive and potentially cheap method for detecting AD, this work shows some promising directions in this field.

Keywords: explainable machine learning; Alzheimer’s disease; natural language processing, causal inference.

Introduction

Background

Alzheimer's Disease (AD) is a serious and the most common dementia worldwide. In the US, more than 5 million individuals are living with AD, and AD Related Dementia (ADRD), costing the nation $244B in 2019. The National Academy of Sciences, the National Plan to Address Alzheimer’s Disease, and the Affordable Care Act through the Medicare Annual Wellness, identify earlier detection of ADRD as a core aim for improving brain health for millions of Americans.

Traditionally, brief cognitive screening tests and biological marker methods (usually neuroimaging [1-4] or cerebrospinal fluid examination [5]) have been used for identification. However, these approaches tend to be invasive, expensive, and/or trigger patient compliance problems. Alternatively, spoken language is a rich and inexpensive source of information in the detection of cognitive status even at the early stage.

Robinson et al. have [6] showed that AD patients are more likely to have a reduction of vocabulary size and difficulty in correctly using verbs and nouns. Croisile et al. [7] have showed that AD patients give a shorter speech, more implausible details, and syntactically simplified descriptions.

Recently, machine learning (ML) or deep learning (DL)-based automated early-stage AD detection using linguistic features have been proposed and demonstrate outstanding diagnosis accuracy. Eyigoz et al. [8] have demonstrated that a patient's language performance in naturalistic probes can expose subtle early linguistic signs of progression to AD much before a clinical diagnosis of the impairment. Khodabakhsh et al. [9] have studied the diagnosis of AD, using speech features extracted from a spontaneous conversation and obtained 90% AD detection accuracy. ML/DL-based methods allow for the use of latent features which go beyond handcraft features and represent more sophisticated concepts. For example, word (sentence) embeddings maps words (sentences) from a vocabulary to a vector of real numbers. Good embeddings will encode similar concepts to adjacent vectors. Studies that use word embeddings for AD diagnosis include [10-13]. In addition to the use of word embeddings, [10] uses PoS features; [11] uses PoS features and sentence embeddings; [12] uses targeted psycholinguistic, sentiment, and demographic features; In [13], recurrent neural networks (RNN) are used to capture the temporal dynamics in speech recordings for improving the diagnosis accuracy.

However, most previous works are performance-oriented and construct more complex models with an increasing number of features and modalities. Though better diagnosis accuracy has been achieved, they usually sacrifice transparency in the diagnosis-making process. This is because most of these complex models are deep-learning-based, which are inherently opaque and not all the features are human interpretable. This is especially true if their influence on the prediction is not well understood. This opaqueness and lack of understanding of the contributions of individual features to the prediction has resulted in a reluctance by the clinical community to use these methods in practice [14].

Explainable Artificial Intelligence (XAI) refers to methods that can reduce the opaqueness of deep learning models. XAI methods can be classified according to various criteria. One taxonomy is based on the format of explanation. Local explanation or example-based explanation explains an individual prediction while the global explanation explains the model behaviour (e.g., feature importance).

Beyond explaining the model's internal mechanism, recent works have used XAI methods for scientific discovery. XAI-based scientific discovery enables the discovery of insightful scientific concepts from model explanations obtained by XAI methods. Ginsburg et al. [15] propose FINE (feature importance in nonlinear embeddings) for the analysis of cancer patterns in breast cancer tissue slides. FINE automatically determines the important features which revealed previously unknown scientific attributes. Li et al. [16] have shown that similar concepts to Kepler's laws of planetary motion and the Newton's law of universal gravitation can be obtained by XAI methods.

Objectives

Our goal is to disclose the underlying relationship between PoS features and AD. Our work firstly explores the predictive power of PoS features for AD diagnosis by using a well performing transformer-based [17] model, which is trained to use PoS features for AD diagnosis. If a feature does not impact the decision of this predictor, then it stands to reason that this feature does not have much predictive power. Note that, though PoS features are used in previous works for AD diagnosis, and impressive accuracies have been achieved, they are usually combined with other features as inputs and hence the effect of PoS features alone is unclear. In our study, we find that using only PoS features can still yield a high AD diagnosis performance with 92.2% accuracy. Hence it is interesting to discover which PoS features play vital roles in this prediction.

In order to understand the importance of any given feature for a particular problem, it is important to study the effect this feature has *globally*, on all samples. To achieve this goal, we use example-based explanation called counterfactual explanation (CFE) [18] on our predictor. Example-based explanation gives explanations for individual data samples. Then, we analyze the statistical summary of the counterfactual explanations of a *group* of data samples to show the global effect of each input feature.

Conventionally, counterfactual explanation aims to answer "Why" questions such as "Why the model's decision is Y" or "What would have happened to Y, had I not done X?". The first step in obtaining the counterfactual explanation is to search for the counterfactual examples which are defined as the examples obtained by applying minimal changes to the features of the original example and having the predefined outputs. Then, the counterfactual explanations can be extracted by comparing the differences between the original example and its counterfactual examples. For example, if the model's prediction is changed from AD patient to healthy control as we manually increase the appearance of nouns by the minimal unit (e.g., 1) in a data sample, then the counterfactual explanation would indicate that the number of nouns used is as an important factor in classifying the sample as being from an AD patient.

However, when generating counterfactual examples, the conventional counterfactual explanations assume features are *independent* of each other. This can result in the counterfactual examples that are not feasible in the real world. For example, an infeasible counterfactual explanation can suggest that the number of nouns be decreased while the number of adjectives be increased, which is anti-causal since adjective words are usually used to decorate noun words and hence its appearance is supposed to increase or be unchanged as the number of noun words increasing.

It is clear that conclusions drawn from potentially infeasible counterfactuals cannot be reliable. Hence, it is important to develop a causally away counterfactual intervention method for our purposes. We argue that the key point to making the generated counterfactual examples feasible is to ensure the generation process of counterfactual examples obeys causal rules. That is, as we generate counterfactual examples by making changes to some features, the causal consequences of these changes (e.g., increase number of nouns cause the increase of number of adjectives) have to be considered.

To generate feasible counterfactual examples, we propose to use a causal model, which contains a directed graph that models the random variables by nodes and their causal relation by directed edges. Each edge in the causal model also encodes the causal function f: P→C, where C is any variable that is modeled in the causal model and P represents the variables that cause variable C. Then one can generate counterfactual examples of the original example by doing interventions in the causal model. Performing interventions is the process where some variables within a sample are changed to fixed values and the rest of the variables are generated according to the causal functions (e.g., f). A counterfactual sample can be regarded as a counterfactual explanation if it can yield the predefined output.

To understand the significance of a single feature, we propose only to intervene on one feature at a time for counterfactual generation. We hence name our proposed method: one-intervention-causal-explanation (OICE). We then use the one-intervention counterfactual examples to explain the importance of each feature by asking, “What would have happened to the output, had I intervened on feature A?”. Moreover, using one-intervention can allow us to systematically study the impact of the different features. Each feature (and its descendants) that is impacted by the parent feature in this one-intervention approach, can be further analyzed by the structural causal model (SCM). Finally, we define three metrics to quantify the importance of features in the decisions.

Related Work

Counterfactual Explanation

Counterfactual explanations are a widely used method for generating explanations of a model's decision and aim to answer “How the world would have to be different for a desirable outcome to occur” [18]. By studying these counterfactual instances, one can explain why the model arrives at the outcome, by comparing the difference between the hypothesis and the original scenarios or a possible suggestion about how the desired outcome can be obtained by changing some of the features. Generally, counterfactual explanations are generated by finding the minimal changes that are needed to change the classification of this instance to the desired class. Wachter et al. [18] formulates a general form for finding the counterfactual explanations x^{CF}:

$$x^{CF} = argmax_{x'} \lambda \left(f_w \left(x' \right) - y' \right) + d \left(x, x' \right) (1)$$

where x is the query instance, f_w is the classifier, y' is the desired output, and $d(\cdot,\cdot)$ is a distance function. In practice, maximization over λ is done by iteratively solving for x' and increasing λ until a sufficiently close solution is found.

The quality of counterfactual explanations is measured in terms of actionability, feasibility, diversity, and sparsity. The meaning of each metric is stated as follow:

- Actionability: A CFE that changes any immutable features (e.g., gender: male → female) is un-actionable and vice versa.
- Feasibility: Features that are changed by a CFE should be in a reasonable range/population. An infeasible CFE could be changing the number of credit card from 5 to -1.
- Diversity: The ability to generate diverse CFEs.
- Sparsity: The number of features that are changed in CFES. Fewer changes/high sparsity is favorable since humans can only extract limited information.

Most existing approaches in the literature of counterfactual explanations are dedicated to improving the metrics mentioned above. Recent studies [19, 20] consider the distribution of data and generate counterfactual instance from the relatively high-density region of the input space. These methods improve the feasibility by avoiding unlikely or unrealistic counterfactual instances under the data distribution. Ustun et al. [21] improves the actionability and feasibility by allowing the counterfactual instances that optimize a user-specified cost function and prevent counterfactuals from changing immutable variables like age, sex, gender. Russell [22] proposes a Mixed Integer Programming (MIP) formulation to handle mixed data types and offers counterfactual explanations for linear classifiers that respect the original data structure. This formulation is guaranteed to find coherent solutions by only searching within the “mixed-polytope” structure defined by a suitable choice of linear constraints.

The work most similar to ours is [23], which shifts the paradigm from nearest counterfactual explanations to minimal interventions. Specifically, in [23], counterfactual examples are generated by the predefined SCM and a set of possible interventions to reach the desired outcomes. The optimal intervention set is obtained by choosing the one that induces the minimum cost, where the cost is measured by a predefined cost function on the intervention sets. Additionally, they prove the necessity of considering all inter-variable causal dependencies and demonstrate efficiency on some toy datasets. We use a more complex SCM, known as Causal Generative Neural Network (CGNN) [24], to capture the inter-variable causal dependencies and generate counterfactual explanations by the intervention. We additionally statistically analyze the derived explanations to inspect the global behaviour of the model.

Methods

For scientific discovery purposes, our method incorporates three phases: knowledge learning, knowledge extraction, and knowledge verification. As shown in Fig. 1, in the knowledge learning phase, we use a transformer-based classifier to learn the underlying mechanism between PoS features and AD; In the knowledge extraction phase, we use our proposed XAI method, OICE to extract the learned mechanism. Specially, OICE would quantitatively indicate the importance of PoS features used by the model in AD classification; The extracted knowledge (i.e., feature importance) would be verified with findings of previous works in phase 3. A model that is verified to have high consistency with previous findings is more plausible and hence is more likely to bring reliable insights about the underlying mechanism among PoS features and AD.

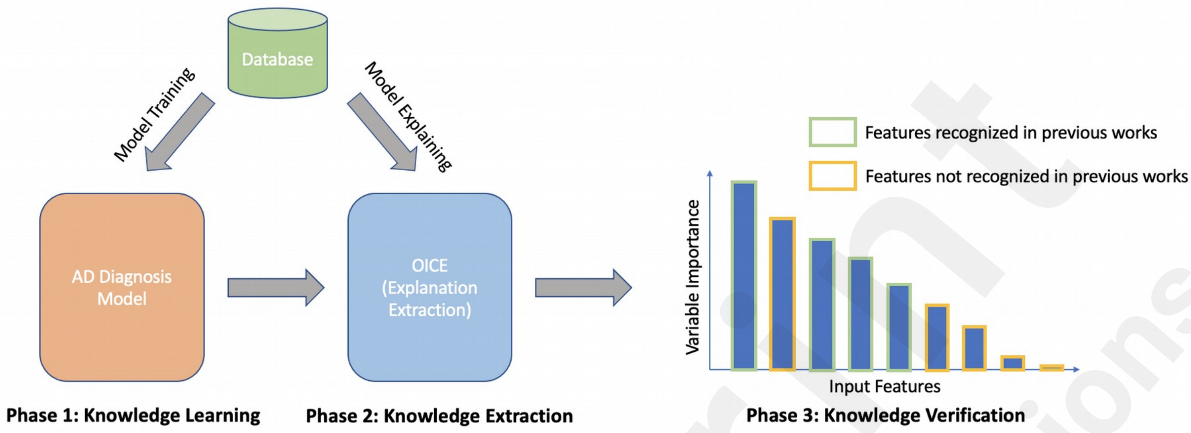

Database
Model Training
Model Explaining
AD Diagnosis Model
OICE (Explanation Extraction)
Features recognized in previous works
Features not recognized in previous works
Variable Importance
Input Features
Phase 1: Knowledge Learning
Phase 2: Knowledge Extraction
Phase 3: Knowledge Verification

Figure 1 Method Overview. Procedures of using XAI for scientific discovery.

In the following sections, we will describe in detail the methods we used in the first two phases. We verify the extracted knowledge (phase 3) in the Result Section. We first introduce the dataset followed by the structure of the transformer-based classifier. Then we introduce the proposed model explanation method, OICE. Finally, we describe details in implementing the introduced methods.

Dataset Description

DementiaBank [25] is a database of multimedia interactions for the study of communications in dementia patients. This dataset comprises of the transcripts of individuals (dementia and control) who are given four tasks: (1) Cookie theft description: participants in both the control group and dementia group were given a picture of a child attempting to steal a cookie and asked to describe what they saw. (2) Word fluency: which measured their fluency (dementia group only); (3) Recall: participants tested on their memory recall (dementia group only) and (4) Sentence construction: where they were tested on sentence construction (dementia group only). In total, the corpus contains 1049 transcripts from 208 AD patients and 243 transcripts from 104 elderly control individuals for a total of 1292 t`ranscripts. Two examples of DementiaBank dataset are illustrated in Table 1. In this study, we use all the transcripts described above.

The transcripts were tokenized into single word tokens, and each token is computed with PoS tags using NLTK toolkit [27]. Upon each transcript, we generate a PoS feature vector with the counts of 27 PoS tags. The names and the meanings of the 27 PoS features are introduced in Table 2.

Table 1 Two examples of DementiaBank data sample. In our experiment, we analyze the PoS features that are extracted from the speech records.

Label	Speech Record
Healthy Control	okay, well the mother is drying the dishes, the sink is overflowing, um the little girl’s reaching for a cookie, and her brother’s taking cookies out of the cookie jar, and the stool is going to f knock him on the floor laughs, he’s going to fall on the floor because the stool’s not uh what, with gravity, whatever, uh the uh curtains are blowing I think, that’s all I can see
AD Patient	I would like to have a lead pencil, the tree is blossoming, I hope my child doesn't hafta go to the hospital , I hope my child doesn't hafta go to the hospital, I shouldn't say that because we have a daughter who's pregnant, and I do want her to go to the hospital, okay then, this winter has been a very cold one, the doctor said I, I sat in the chair by a the doctor, brief, I'm not, I forgot to try make them brief, the bureau drawer stands open

Ethical Consideration

We use the Dementiabank dataset which is archived by TalkBank. TalkBank is subject to its own Code of Ethics (detailed in [26]) which supplements but does not replace the generally accepted professional codes of American Psychological Associatgion Code of Ethics and American Anthropological Association Code of Ethics.

Transformer-based AD Classification Model

Recently we proposed a transformer-based [11] classifier to exploit PoS features, as shown in Fig. 2. In our architecture we use the multi-head attention (MHA) module and the encoder structure of the transformer to process these features. Our motivation for this stems from the success of this architecture in creating state-of-the-art language embeddings as demonstrated in [11]. This architecture comprises of a self-attention module that captures the intra-feature relationships; an attention layer together with a following 1-D CNN layer. The MHA module is the same as that proposed in [11] for the popular transformer architecture. Let $R=\{r_1,r_2,I,r_n\}$ be the set of records, then r_i is the i^{th} record in the dataset. We compute PoS features for each record. Let $P=\{p_1,p_2,I,p_n\}$ be the set of PoS feature vectors and p_i be the i^{th} vector in the PoS matrix. We use h Multi-Head-Attention (MHA) layers on $P=\{p_1,p_2,I,p_n\}$ to capture the relationship between the PoS features. The MHA transforms $P$to another matrix of n-dimensional vectors $A=\{a_1,a_2,I,a_n\}$. The MHA module is followed by a 1-layer CNN and a SoftMax layer to get the final classification.

Table 2 PoS features & meanings

Tag	Meaning
NN	common nouns
PRP	personal pronoun
VBG	verb, gerund or present participle
UH	interjection
NNS	noun, plural
MD	modal
JJR	adjective, comparative
VB	verb, base form
IN	preposition or subordinating conjunction
JJ	adjective
RP	particles
PRP$	possessive pronoun
CC	coordinating conjunction
CD	cardinal number
PDT	predeterminer
NNP	proper noun, singular
TO	to
DT	determiner
RB	adverb
VBZ	verb, 3rd person singular present
VBN	verb, past participle
WP	wh-pronoun
VBP	verb, non-3rd person singular present
JJS	adjective, superlative
VBD	verb, past tense
EX	existential there
WP$	possessive wh-pronoun

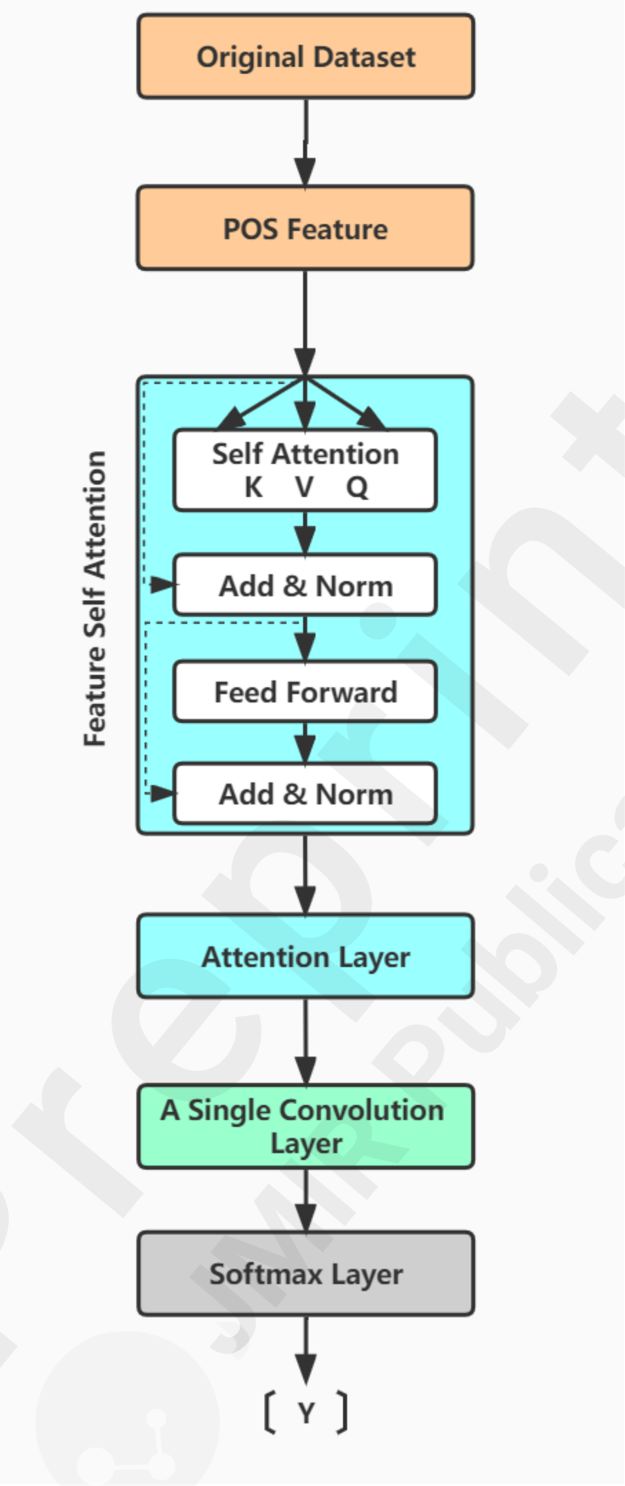


Figure 2 The proposed transformer-based classifier that uses the PoS features of the patient/control's description.

One Interventional Counterfactual Explanation (OICE)

To derive an explanation, OICE first calculates the counterfactual explanations for each single sample. Each single counterfactual explanation can be simply seen as a vote for features' importance by each sample. Then, OICE groups these counterfactual explanations to summarize the global explanation about feature importance. In this subsection, we first outline the preliminary information on structural causal models (SCM), which is an essential element for obtaining counterfactual explanations. We then describe how we learn an SCM from the data. Next, we discuss how we formulate the OICE and how OICE generates individual counterfactual explanations by using the pretrained SCM. Then, we introduce the metrics that we propose to measure the feature importance (global explanation) according to a group of counterfactual explanations.

Structural Causal Model

In this section, we review the concepts of structural causal models (SCM) and interventions. An SCM, M, can be represented by a triplet, $M=\langle X,F,U\rangle$, that contains a set of endogenous variables, $X=\{X_1,X_2,I,X_d\}$, a set of causal mechanisms, $F=\{F_1,F_2,I,F_d\}$, and a set of exogenous variables, $U=\{U_1,U_2,I,U_d\}$, where each U_i is independently drawn from distribution, U. Any endogenous variable X_i can be obtained by its causal mechanism F_i as $X_i=F_i(PA_i,U_i)$, where U_i U and PA_i denotes the parent nodes of X_i and $PA_i \in X\{X\text{¿}_i$.

In our case, the endogenous variables are the random variables of the PoS features. The causal effect between two PoS features in hence encoded in the causal mechanisms between them (can be null if no causal relation between them). The exogenous variables are seen as the set of unknown factors that can cause PoS features.

We denote an intervention in SCM by a do-operator $do(\cdot)$. Intervening the set of X to the value a can be then described as $do(\{X_i=a\}_{i\in I})$ where I is a set of indices of the subset of endogenous variables to be intervened upon. By intervention, causal relations and causal mechanisms defined in the original SCM can be changed. Endogenous variables from I can be obtained by $do(X_i=a)$ rather than $X_i=F_i(PA_i,U_i)$. Therefore, by performing the intervention, the original SCM M can be changed to a post-intervention SCM M_I.

Structural Causal Model via Generative Network

We use the CGNN proposed in [24] to represent the SCM since it does not limit the types of causal mechanisms (e.g., linear or non-linear). Given a causal graph, a CGNN can be trained to learn the causal mechanisms underlying the causal graph by reducing the Max Mean Discrepancy (MMD) [28] between the ground-truth data and the generated data. CGNN generates each endogenous variable by $X_i=F_i^{\theta_i}(PA_i,U_i)$, where $F_i^{\theta_i}$ is a generative neural network parameterized by θ_i. For simplicity, we use F_i to represent $F_i^{\theta_i}$ in the rest of paper. U_i are random samples drawn from Gaussian distribution. Fig. 3 illustrates an example of constructing SCM by CGNN.

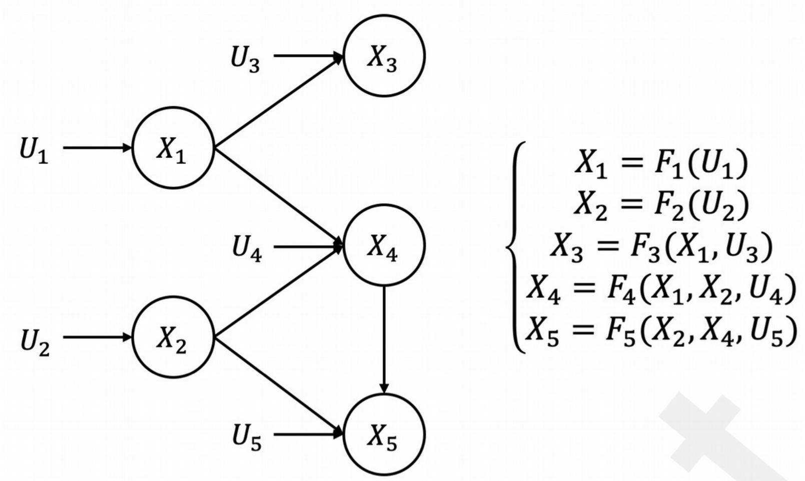


Figure 3 Example of a Structural Causal Model (SCM). Left: causal graph, right: causal mechanisms. As for CGNN, each causal mechanism is implemented with a generative neural network.

The weights of causal mechanisms (i.e., θ_i) are updated to minimize the MMD between the ground-truth samples and the samples generated by the CGNN. In our experiment, we discover the causal relations from the dementiabank dataset by using the PC algorithm [29]. PC algorithm is a constraint-based causal discovery method, under the assumption of causal sufficiency (i.e., no latent confounders). We discover causal relations among PoS features from the dementiabank dataset rather than use generic PoS causal rules as the former would better capture the causal relations among PoS features in the dementia group.

Explanation by Minimal Intervention

We now introduce some notations and discuss the formulation of OICE. Let $x^F \in R^d$ denote the original factual sample and $x^{CF} \in R^d$ denote the counterfactual sample that is obtained by a set of interventions I. Here, we re-define $I = \{I_1, I_2, \ldots, I_d\}$ to be an intervention set that has the same length as the sample x^F. For each element I_i, if $I_i = 0$, it denotes no intervention on x_i^F (the i^{th} element of x^F), otherwise do intervention $x_i^F = I_i$. Generally, any sample x (both factual and counterfactual) can be generated by the SCM $\langle X, F, U \rangle$ using the equation: $x = G(U^F, I; F)$, where G represents a sequence of processes to generate x. G contains a causal graph and the corresponding causal mechanisms between variables. The variables of a sample, x, are generated in sequence from root to leaf of the causal graph. Generating factual sample x^F, can be done by setting all the elements in I to zero. For a given x^F, its corresponding exogenous variables U^F can be obtained by inverting the generating process: $U^F = G^{-1}(x^F; F^{-1})$.

We formulate the problem of one-intervention counterfactual explanation as searching for the optimal $I^{¿}$ that results in a counterfactual example x^{CF}, which would flip the outcome from y to y'. One-intervention is implemented by fixing the $\|I\|_0$ to be 1. It is formulated as:

$$I^{¿} = \| h(G(U^F, I; F)) - y' \|^2 \text{, subject to } \|I\|_0 = 1 (2)$$

where h is the predictive model. In most cases, the model h is a probabilistic model, we then select the optimal solutions $I^{¿}$ as the one that results in counterfactual examples that can achieve a certain degree of certainty to be y' (e.g., $h(G(U^F, I; F))$ is 80% certain to be y'). By doing so, multiple optimal solutions are obtained, which contain different intervened features. Note that the same kind of intervened features may have different intervention values.

Consequently, we further distill our optimal solutions set by only keeping one solution for each subset with the same intervention that causes the minimum distance weighted by Median Absolute Deviation (MAD) [18].

Note that OICE implicitly assumes the causal relation from variables x^F to outcomes y by the predictive model h. However, OICE does not rely on this relation to generate counterfactual examples x^{CF}. Model h in OICE only helps to solve the optimization problem stated in Eq. 1.

Metrics for Measuring Importance

So far, we have introduced how to obtain explanations for individual instances by OICE. We then make the inference of model's global behaviour (i.e., importance of features) by statistically analyzing the explanations derived from a batch of samples. In the section, we introduce some metrics to measure the impact of intervening a feature to cause a flip in the outcome. The impact of features can be further associated with its importance for a machine learning model in making a decision.

Let $S=\{S^{(1)}, S^{(2)}, \ldots, S^{(n)}\}$ represents a set of n samples that belong to class y (i.e., $h(S^i)=y, for\, i=1,2,\ldots,n$). In our case, the problem is a binary classification problem, and the classes are: "control" or "Alzheimer's". Let $C_k^{(i)}$ denote the counterfactual explanation of the i^{th} sample obtained by intervening on the feature k and hence $h(C_k^{(i)}) \neq y$. To measure the impact on flipping the outcome that is caused by intervening feature k, we introduce our first metric, *Impact Score (IS).* IS_k can be interpreted as the proportion of counterfactual samples for which feature k must be intervened to flip the outcome and is defined as:

$$IS_k = \frac{|I_k|}{n} (3)$$

where $I_k = \{i : h(C_k^{(i)}) \neq y, i=1,2,\ldots,n\}$ is a set that contains the indices of samples in S that have a counterfactual explanation obtained by intervening on feature k. The *IS* score describes the overall impact and does not consider the cost of the intervention (i.e., how much a feature has been increased or decreased). Accordingly, we introduce another metric, *weighted impact score (wIS),* to measure the impact made by changing the unit value of a feature. This measure trades off the impact with the cost of impact. *wI*S can be used to compare among the features. Features with higher *wIS* value have more importance in flipping the outcome. To define *wIS*, we first introduce the parameter, *cost of impact (CI)*, to measure the average absolute change that must be made to achieve the impact (i.e., *impact score*). Using subscript j to index the j^{th} feature of a sample $S^{(i)}$ or $C_k^{(i)}$, the *cost of impact (CI)* for feature k can be defined as follow:

$$CI_k = \frac{1}{|I_k|} \sum_{i \in I_k}^{\square} \square \frac{|C_{k,j}^{(i)} - S_j^{(i)}|}{R_k}, j=k \wedge CI_k \in [0,1] (4)$$

where R_k is the range of feature k. Next, we define the *weighted Impact Score* as follows:

$$wIS_k = \frac{IS_k}{CI_k} (5)$$

Note that the *wIS* defined in Eq. 5 does not consider the trends of change in a feature (i.e., increasing or decreasing). To take care of this, we separate wIS_k into $wIS_k^{+¿¿}$ and $wIS_k^{-¿¿}$ to represent the *weighted impact score* for increasing and decreasing the value of feature k respectively. They are calculated using the following rules: (i) if all the trends of change (i.e., sign($C_{k,j}^{(i)} - S_j^{(i)}$)) are same, then wIS_k^{δ} is calculated using Eq. 5 where δ is + if the changes are positive and - for negative. (ii) if both positive changes and negatives change exist, $wIS_k^{+\text{¿}\text{¿}}$ and $wIS_k^{-\text{¿}\text{¿}}$ are calculated using a modified version of Eq.4 so that the summation is done only on the positive and negative changes and do normalization respectively. Additionally, the impact score introduced above measures the overall importance of changing both the intervened feature and its descendent features (caused by intervention on this feature).

It is important to understand how much each changed feature contributes to flipping the outcome. Consequently, we introduce another metric, called *pure impact score (PIS)*, to quantify the importance of every changed feature within the counterfactual explanations obtained by the same intervention.

Hence, the *PIS* for a feature is calculated by subtracting the impact (on flipping the outcome) caused by its child nodes from the *IS* score of this feature. As the weighted impact score representing the change of impact score per unit change of the value of the feature, the impact of each child node is m can be hence quantified as the average of the changes of the m's values multiply by the weighted impact score of m. The impact caused by feature m when m is causally affected by feature k is defined as follows:

$$PIS_k^m = wIS_m^{+\text{¿}\times\frac{1}{|I_k|}\sum_{i\in I_k^{+\text{¿}}\text{¿}}\square\frac{\left|C_{k,j}^{(i)}-S_j^{(i)}\right|}{R_k} - wIS_m^{-\text{¿}\times\frac{1}{|I_k|}\sum_{i\in I_k^{-\text{¿}}\text{¿}}\square\frac{\left|C_{k,j}^{(i)}-S_j^{(i)}\right|}{R_k}(6)\text{¿}}\text{¿}}$$

While the *pure impact score* for the intervened feature k, PIS_k^k, is defined as:

$$PIS_k^k = IS_k - \sum_{m\in CH_k}^{\square} PIS_k^m (7)$$

where CH_k is the set of indices of the child nodes of feature k. The value of PIS_k^m is then normalized over IS_k to represent the percentage of effort for flipping the outcome.

Implementation Details

Model Settings

In our experiments, we have 6 layers for the multi-head attention (MHA) module. We used stochastic gradient descent + momentum (SGD + Momentum) as the optimizer for training. Since the DementiaBank is an unbalanced dataset, we added a class weight correction by increasing the penalty for misclassifying the less frequent class

During model training to reduce the effect of data bias. The class weight correction ratio used in this paper is 7:3. We randomly split the original data into 81% training, 9% validation and 10% testing over multiple seeds. Our proposed model achieves a high accuracy of 92.2%, F1 score of 0.952, precision of 0.935, recall of 0.971, and AUC of 0.971 on the DementiaBank dataset.

PoS Features Causal Relation Discovery

As mentioned earlier, we use the PC Algorithm [29] to discover the intra-feature dependencies. The causal graphs returned by the PC Algorithm contain undirected edges. We hence further revise the returned graph by orienting the undirected edges. The edges are oriented according to our knowledge of the linguistic features. For example, we make the causal direction NN->JJ since NN (nouns) causes the use of JJ (adjectives). The full causal graph for the 27 linguistic features used in our experiment is illustrated in Fig. 4.

Problem Solver

Solving the l_0 norm constraints in Eq.2 is a non-trivial task. However, the parts-of-speech (PoS) features used by the proposed classifier are all integers and within narrow ranges. It makes it possible to solve our problem by exhausting all the solutions and then select the optimal ones. Gradient-based methods can be used for solving continuous values. In the work, we focus on classifying the text into Alzheimer's disease or control and is hence discrete. Additionally, we set the certainty parameter to 80%, this implies all solutions, I, that satisfy $\| h\left(G\left(U^{|F|}, I; F\right)\right) - y' \|^2 < \alpha$, where $\alpha = 0.04$, are considered optimal. The value of α is chosen to reflect 80% certainty.

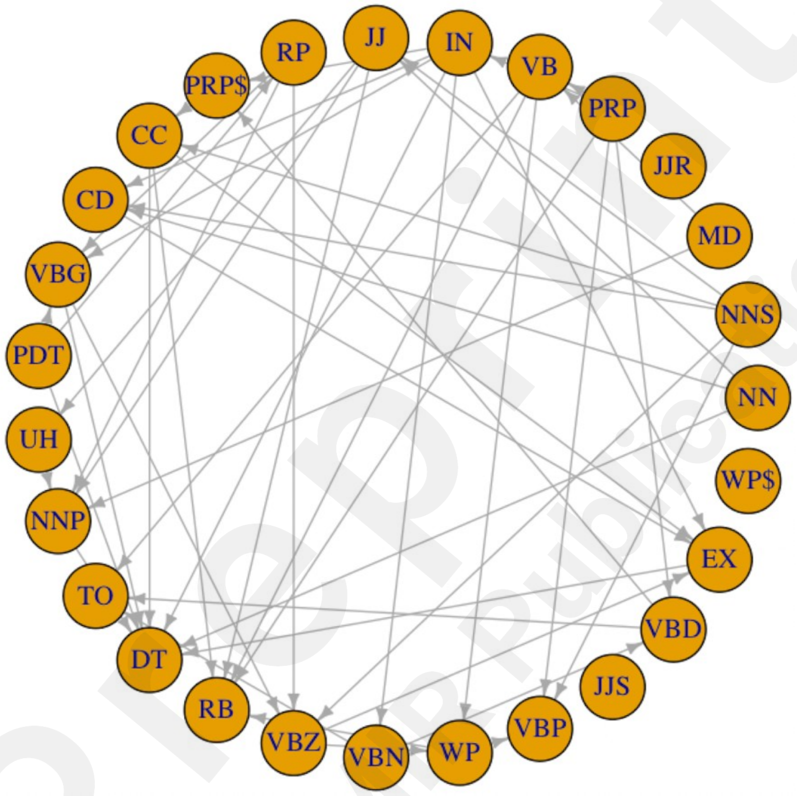


Figure 4 Causal Graph for 27 Linguistic features. The starting variable of each directed edge represents the cause, and the ending variable represents the effect.

Results

Predictive Power of PoS Features

Table 3 Evaluation of the trained AD diagnosis model.

Accuracy	Precision	Recall	F1-Score ↑	AUC
92.2%	0.935	0.971	0.955	0.971

All PoS features described in Table 2 are used for model training. The models' performance has been evaluated using the accuracy, precision, recall, F1-score and area under the receiver operating characteristic curve (AUC) metrics. All these scores are reported in Table 3. The high performance illustrates that PoS features extracted from speech can help to distinguish AD patients from the health controls. This finding encourages us to move forward to explore which of the PoS features is playing vital role.

Knowledge Extracted from Model Explanation

In this section, we continue to reveal the significant PoS features that direct the model’s decision. We analyze the counterfactual examples from a statistical perspective and analyze the important features derived from this analysis. We study the counterfactual explanations for a control sample (i.e., an individual without Alzheimer). The important features are derived by analyzing which feature plays a vital role in misclassifying a control sample as an Alzheimer's patient. In this experiment, we report the results of 210 of 243 controls. These 210 control samples were classified correctly by the classifier. The optimal counterfactual explanation for all the 210 results can be achieved by only intervening one feature. Other samples are excluded because of misclassification.

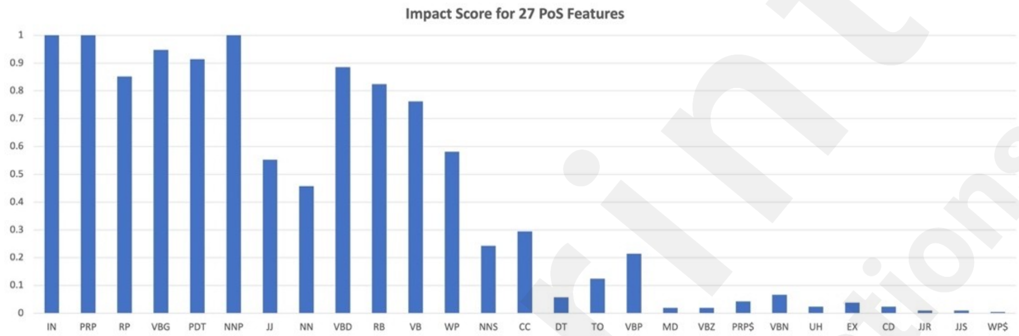


Figure 5 The impact score (IS) for 27 PoS features. Feature with higher IS value denotes more samples successfully flipping the model’s outcome by intervening on it.

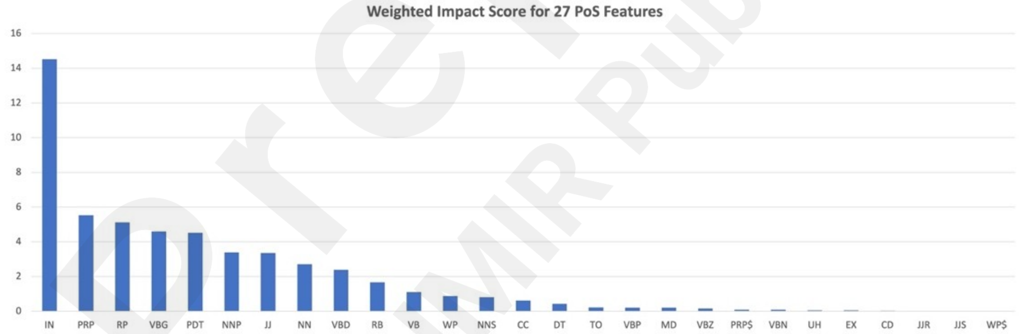


Figure 6 The weighted impact score (wIS) for 27 PoS features. Features with higher values denote more importance for machine learning in making decisions.

We plot the both the *impact score* and *weighted impact score* for all PoS features in Fig. 5 and Fig. 6. We regard the top twelve features (IN, PRP, RP, VBG, PDT, NNP, JJ, NN, VBD, RB, VB and WP) as our primary findings about important PoS features in AD diagnosis. The selection considers PoS features that have both high *IS* scores and *wIS* scores. Features with low *IS* scores indicates that few samples adopt them for flipping the model’s output which is less reliable as the lack of agreement by the majority. In Fig.7, we also illustrate the examples of AD and healthy control from original dataset and the counterfactual examples (explanation) in a spider plot. It shows that, the generated counterfactual examples capture the difference of PoS features between AD patients and healthy controls. The PoS features, we used in this work are shown in Table 2. Further information of those features can be found in [30], [31].

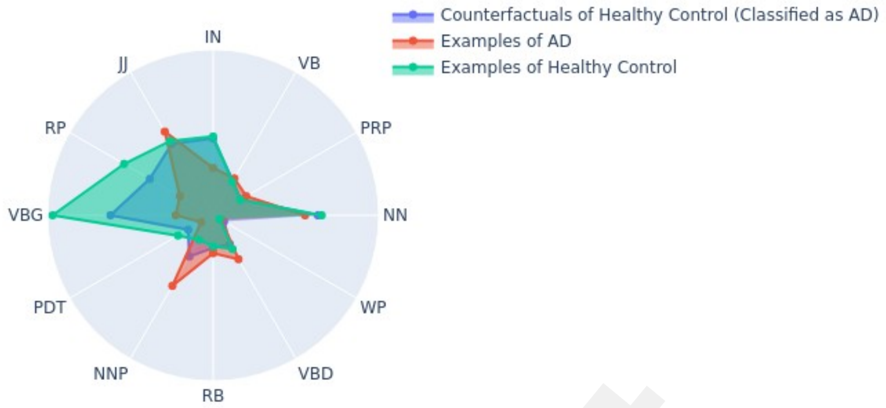


Figure 7 Spider plot of samples for AD patients, healthy control and counterfactual samples (classified as AD patients).

Table 4 Impact Cost in percentage and the direction of change for all 27 PoS features. A smaller CI value denotes smaller changes are needed.

PoS feature name	Cost of Impact (CI) value
NN ↓[b]	16.9%
NNS ↑[c]	30.0%
MD ↑	9.0%
JJR ↑	83.3%
PRP ↑	18.1%
VB ↑	69.3%
IN ↑↓	5.5%
JJ ↑	16.5%
RP ↑	16.7%
PRP$ ↑	30.3%
CC ↑	48.7%
CD ↑	75.6%
VBG ↓	20.7%
PDT ↑	20.2%
UH ↑	33.3%
NNP ↑	29.5%
TO ↑	57.4%
DT ↓	13.6%
RB ↑	49.6%
VBZ ↓	12.5%
VBN ↑	86.7%
WP ↑	67.1%
VBP ↑	73.8%
JJS ↑	100%
VBD ↑	37.1%
EX ↑	67.2%
WP$ ↑	100%

[a] Please refer Table 2 for all abbreviations for the feature names.

[b] The down-arrow indicates the decreasing the values.

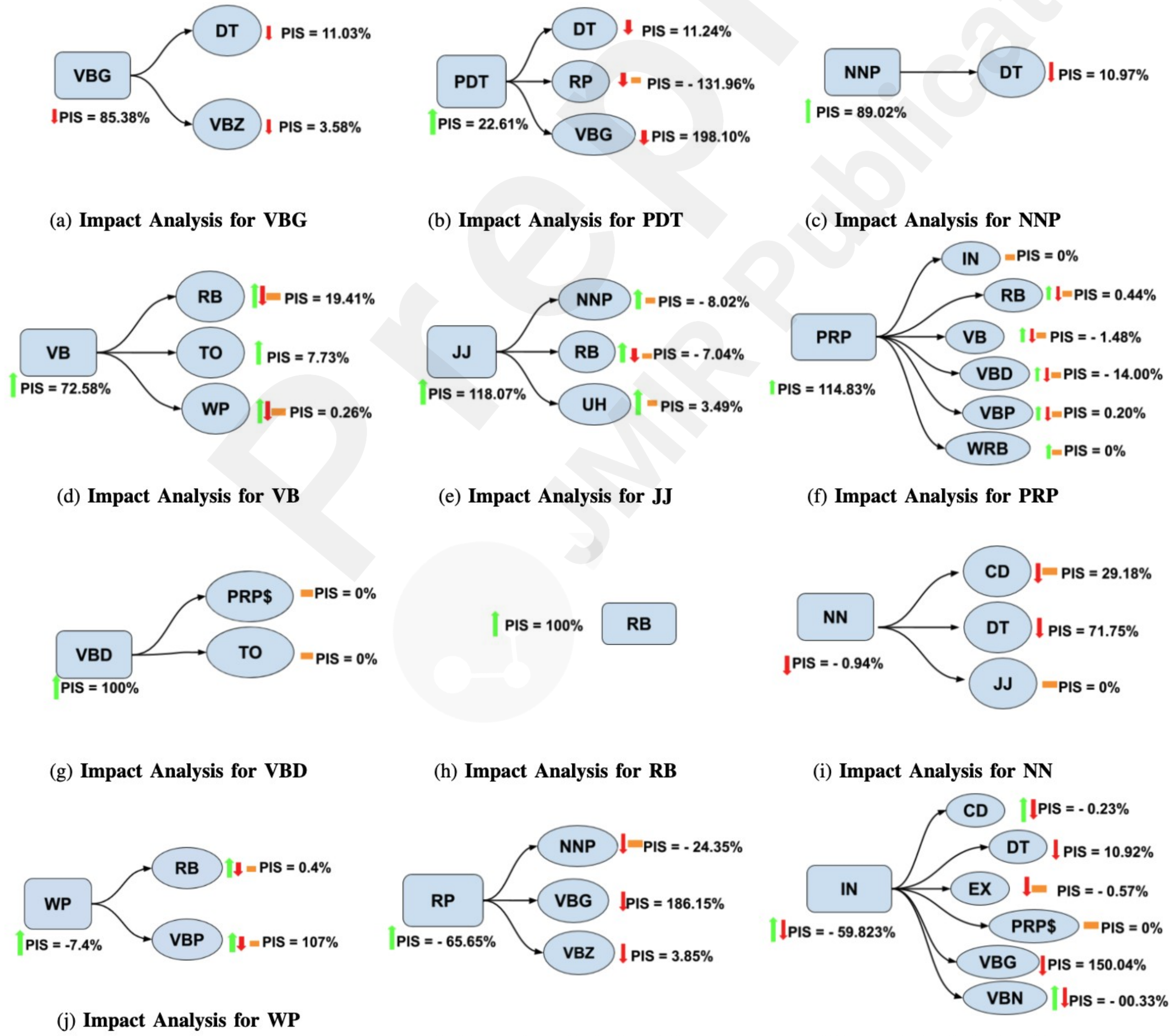
[c] The up-arrow indicates increasing the values.

We then analyze the important features, to answer the question:

How exactly does intervening a feature cause the outcome to flip?

To answer the above question, we need to consider the children features of the intervened feature given by the SCM. More specifically, knowing that the counterfactual examples have moved across the decision boundary (i.e., the outcome has flipped), we examine how each changed feature (i.e., intervened features and its children) affects this movement of the original examples towards or away from the decision boundary. We use the normalized *PIS* (in terms of percentage) to quantify this effect. Positive *PIS* denotes moving the original examples towards the decision boundary and vice versa. In Fig. 8, we show four representative features as examples and illustrate how changes in each feature contribute to flipping the outcome.

To complete the explanation that we promised at the beginning of this section, we use *cost of impact (CI)* to quantitatively describe the average minimal changes that must be done to flip the outcome. In Table 4, we report CI and the changing direction (an up-arrow means an increase in the value is required while a down-arrow means a decrease is required). Take NN (nouns) for example, reducing the use of it by 16.88% of the total range of NN (nouns) feature, will make the classifier flip the final decision.



(a) **Impact Analysis for VBG** (b) **Impact Analysis for PDT** (c) **Impact Analysis for NNP**

(d) **Impact Analysis for VB** (e) **Impact Analysis for JJ** (f) **Impact Analysis for PRP**

(g) **Impact Analysis for VBD** (h) **Impact Analysis for RB** (i) **Impact Analysis for NN**

(j) **Impact Analysis for WP** (k) **Impact Analysis for RP** (l) **Impact Analysis for IN**

Figure 8 Explanations for the representative features: For an intervened feature: the red down-arrow indicates a decrease of value is

required for flipping the outcome, for a child node (feature): the red down-arrow indicates the changing direction caused by the intervention. The same rule applies to the green up-arrow (an increase of value) and the orange horizontal line (no change of value) (a)-(d) Cooperative: We consider the features to be "cooperative" if both the intervened feature and its descendent features contribute to flip the outcome. (e)-(h) Dominant: we define the feature as dominant if the intervened feature significantly contributes to flip the outcome while its descendent features make no or opposite contribution. (i)-(j) Idling: We define the intervened feature as "idling" if it itself only contributes to flip the outcome slightly while the child features make a significant contribution. (k)-(l) Inverse: We term the feature "inverse" if the change of the intervened feature moves the original instances away from the decision boundary, but it causes other features to significantly push the original instances forward to the decision boundary.

Now, we combine the results from both Table 4 and Fig. 6 to offer explanations for all important features. For clarity, in the following explanation, we do not imply the words "increase", "decrease" or "change" as the actions that can modify the values of features. These three words are used to represent the pattern of how much the divergence of a feature from its real value can affect the decision of the model. We use "contribution" or "contribute" to denote the positive effort (measured by *PIS*) or process to flip the outcome. As an opposite to "flip the outcome", we use the terminology "consolidate the outcome" to denote that changing a feature causes the outcome to move further away from the decision boundary.

- *VBG:* Decreasing the value of VBG by 20.66% causes both values of DT and VBZ to decrease. The decrements of VBG, DT, and VBZ contribute to flipping the outcome.
- *PDT:* Increasing PDT by 20.21% causes VBG, DT and RP to decrease or remain unchanged. VBG and DT contribute significantly to flip the outcome, while PDT makes partial contributions.
- *NNP:* Increasing NNP by 29.46% will cause DT to decrease. Increasing NNP contributes significantly to flipping the outcome, while the resulting decrements of DT make a partial contribution.
- *VB:* Increasing VB by at least 69.25% will cause RB and WP to change or remain unchanged and cause TO to increase. The changes of VB, RB, and TO contribute significantly to flip the outcome. The changes in WP makes small contributions.
- *JJ:* Increasing JJ by at least 16.51% will cause NNP and UH to increase or remain unchanged, and cause RB to change or remain unchanged. Even though the change of NNP and RB consolidate the outcome, increasing JJ can significantly contribute to flipping the outcome. Additionally, the change of UH makes a negligible contribution compared with the increment of JJ.
- *PRP:* Increasing PRP by at least 18.09% will cause WRB to increase or remain unchanged, and cause VB, IN, RB, VBP, and VBD to change or remain the same. However, by analyzing the *PIS* for the changes in these features, we conclude that PRP contributes significantly to flipping the outcome.
- *VBD:* Increasing VBD by at least 37.11% will not cause PRP and TO change. We conclude that VBD solely contributes to flipping the outcome.
- *RB:* RB does not have any descendants. We conclude that increasing RB by 49.62% will cause a flip in the outcome.
- *NN:* Decreasing NN by 16.88% can cause CD, DT and JJ to decrease or stay unchanged. Though the change of NN does not contribute to flipping the result, the resultant changes of CD and DT are enough to flip the outcome.
- *WP:* Increasing WP by 67.1% can cause RB and VBP to increase, decrease or stay unchanged. Though the changes of WP and RB do not contribute to flipping the result, the resultant change of VBP are enough to flip the outcome.
- *RP:* Increasing RP by 16.67% causes VBG and VBZ to decrease and NNP to either

decrease or remain unchanged. The changes of RP consolidate the outcome. However, increasing RP can still flip the outcome since intervening on RP will cause the descendent features to change. These changes significantly contribute to flipping the outcome.

- *IN:* On an average, either increasing or decreasing IN by 5.53% can cause CD, DT, EEX, PRP, VBG, and VBN to change or remain unchanged. Among all the descendants of IN, the change of CD, EX, PRP, and VBN make negligible contributions to flipping the result. The change of IN consolidates the outcome, and the major contributions to flipping the outcome are influenced significantly by decreasing VBG and slightly, by decreasing DT.

Discussion

Principal Findings

Firstly, the high performance of the AD diagnosis model on PoS features indicates that PoS features have a rich of clues of speech/language impairments that happens in AD patients. Later by explaining the model using our proposed OICE XAI method, we reveal several important linguistic biomarkers in early-stage Alzheimer’s disease detection. Some of the findings are consistent with the prior findings in psychology and NLP:

- Adverb (RB) is highly relevant to semantic impairment: [32] claims that adverb shows a deictic purpose, which is more common in aphasics with a semantic impairment, and further in [33], adverb was proved to have higher correlations with a diagnosis of AD. Our one intervention method shows that increasing the usage of RB, causes the same speech to be classified as a patient (from a control). Hence, our experiments align with previous findings that increased use of adverbs is an indicator of AD.
- Increased pronoun (PRP) usage is an important sign of semantic dementia: [34] shows that dementia patients with semantic dementia produced an increased number of pronouns than controls. The result is in line with our conclusion that increasing the number of PRP in a control's speech classifies it as a speech sample of a dementia patient.
- Noun (NN) naming deficits indicate cognitive deficits: AD patients show graceful degradation of using living and non-living nouns [35]. We see the same decline in noun usage when shifting from a control sample to a dementia sample.

The consistency between findings of this study and previous studies implies the model possibly learns useful clues about PoS feature. It somewhat supports the point that the rest of the not studied features can offer new insights. To sum up, three of twelve important features (i.e., RB, PRP, and NN) found by our method are consistent with prior findings. We additionally find another eight important features that have not been reported yet, which are IN, RP, VBG, PDT, NNP, JJ, VBD, VB and WP. Our work also seems to suggest that the most important feature may be IN or the use of prepositions. Further clinical studies may be necessary to verify this insight.

Limitations and Further Study

For the scope of work that we consider here, we do not see any limitations; however, we do believe that there is good scope for further study in this area. More modalities can be used in designing an Alzheimer’s disease predictor. These modalities could include brain imagery and other traditional bio-markers. The OICE method can then be applied to all the features used to detect AD leading to a much more nuanced understanding of the causal relations of these biomarkers. This could then lead to clinical trials that test these findings. A subset of non-invasive biomarkers may then emerge as important in predicting AD and this might in turn lead to easier to implement screens for the disease.

Conclusions

In this work, we propose a novel counterfactual explanation method, called one-intervention counterfactual explanation (OICE), to analyze the dominant linguistic features, specifically PoS features, that can be used for AD disease detection. We propose three metrics to evaluate the contributions of these features to the final decision of the model. We collect the explanations from the AD detection model of high accuracy and analyze these explanations by the metrics we define. The features declared as important in the detection of AD by our methods are consistent with previous works in psychology and the NLP, such as adverb, pronoun, and noun. We also find a few other features that are important, and which have not yet been reported. Finally, by leveraging the structural causal model, we further explain how these important features affect the decision-making process.

Data Availability

The DementiaBank Dataset [25] used in this work is password protected and restricted to members of the DementiaBank consortium group. Accessibility to this dataset can be granted after joining DementiaBank consortium group as a member. For details about accessing the dataset, please refer to [25].

Conflicts of Interest

none declared

Abbreviations

XAI: explainable artificial intelligence
DL: deep learning
ML: machine learning
SCM: structural causal model
NLP: natural language processing
AD: Alzheimer’s disease
ADRD: Alzheimer’s disease Related Dementia
OICE: one-intervention counterfactual explanation
CGNN: causal generative neural network
MMD: Max Mean Discrepancy
MAD: Median Absolute Deviation
IS: impact score
CI: cost of impact
wIS: weighted impact score
PIS: pure impact score

References

1. Killiany RJ, Gomez-Isla T, Moss M, Kikinis R, Sandor T, Jolesz F, Tanzi R, Jones K, Hyman BT, Albert MS. Use of structural magnetic resonance imaging to predict who will get

Alzheimer's disease. Ann Neurol 2000 Apr;47(4):430-9. [Medline]

2. Arco J E, Ramirez J, Gorriz, J M, Ruz M. Data fusion based on Searchlight analysis for the prediction of Alzheimer's disease. Expert Systems with Applications 2021 Dec;185. [CrossRef]
3. Raza M, Awais M, Ellahi W, Aslam N, Nguyen H X, Le-Minh H. Diagnosis and monitoring of Alzheimer's patients using classical and deep learning techniques. Expert Systems with Applications 2019 Dec;136:353-364. [CrossRef]
4. De A, Chowdhury A S. DTI based Alzheimer's disease classification with rank modulated fusion of CNNs and random forest. Expert Systems with Applications 2021 May;161. [CrossRef]
5. Pitschke M, Prior R, Haupt M, Riesner D. Detection of single amyloid beta-protein aggregates in the cerebrospinal fluid of Alzheimer's patients by fluorescence correlation spectroscopy. Nat Med 1998 Jul;4(7):832-4. [Medline]
6. Robinson KM, Grossman M, White-Devine T, D'Esposito M. Category-specific difficulty naming with verbs in Alzheimer's disease. Neurology 1996 Jul;47(1):178-82. [Medline]
7. Croisile B, Ska B, Brabant M J, Duchene A, Lepage Y, Aimard G, Trillet M. Comparative study of oral and written picture description in patients with Alzheimer's disease. Brain Lang. 1996 Apr;53(1):1-19. [Medline]
8. Eyigoz E, Mathur S, Santamaria M, Cecchi G, Naylor M. Linguistic markers predict onset of Alzheimer's disease. EClinicalMedicine 2020 Oct 22;28:100583. [Medline]
9. Khodabakhsh A, Kus¸xuo˘glu S, Demiro˘glu C. Natural language features for detection of alzheimer's disease in conversational speech. IEEE-EMBS International Conference on Biomedical and Health Informatics; July 2014; Valencia, Spain p. 581–584. [CrossRef]
10. Karlekar S, Niu T, Bansal M. Detecting Linguistic Characteristics of Alzheimer's Dementia by Interpreting Neural Models. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 2. Association for Computational Linguistics, USA, p.701-707. [CrossRef]
11. Wang N, Chen M, Subbalakshmi K P. arXiv. 2021. Explainable CNN-attention networks (c-attention network) for automated detection of alzheimer's disease URL: https://arxiv.org/abs/2006.14135 [accessed 2021-2-18]
12. Palo F D and Parde N. 2019. Enriching Neural Models with Targeted Features for Dementia Detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 302–308, Florence, Italy. Association for Computational Linguistics. [CrossRef]
13. Mahajan P, Baths V. Acoustic and language based deep learning approaches for alzheimer's dementia detection from spontaneous speech. Frontiers. Retrieved May 9, 2022. [CrossRef]
14. Das D, Ito J, Kadowaki T, Tsuda K. 2019. An interpretable machine learning model for diagnosis of Alzheimer's disease. PeerJ 7:e6543. [CrossRef]
15. S. B. Ginsburg, G. Lee, S. Ali and A. Madabhushi, "Feature Importance in Nonlinear Embeddings (FINE): Applications in Digital Pathology," in *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 76-88, Jan. 2016. [CrossRef]
16. Li, Zelong et al. "From Kepler to Newton: Explainable AI for Science Discovery." https://arxiv.org/abs/2111.12210 [accessed 2022-5-6]
17. Ashish V, Noam S, Niki P, Jakob U, Llion J, Aidan N. Gomez, Łukasz K, Illia P. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA,

6000–6010. [CrossRef]

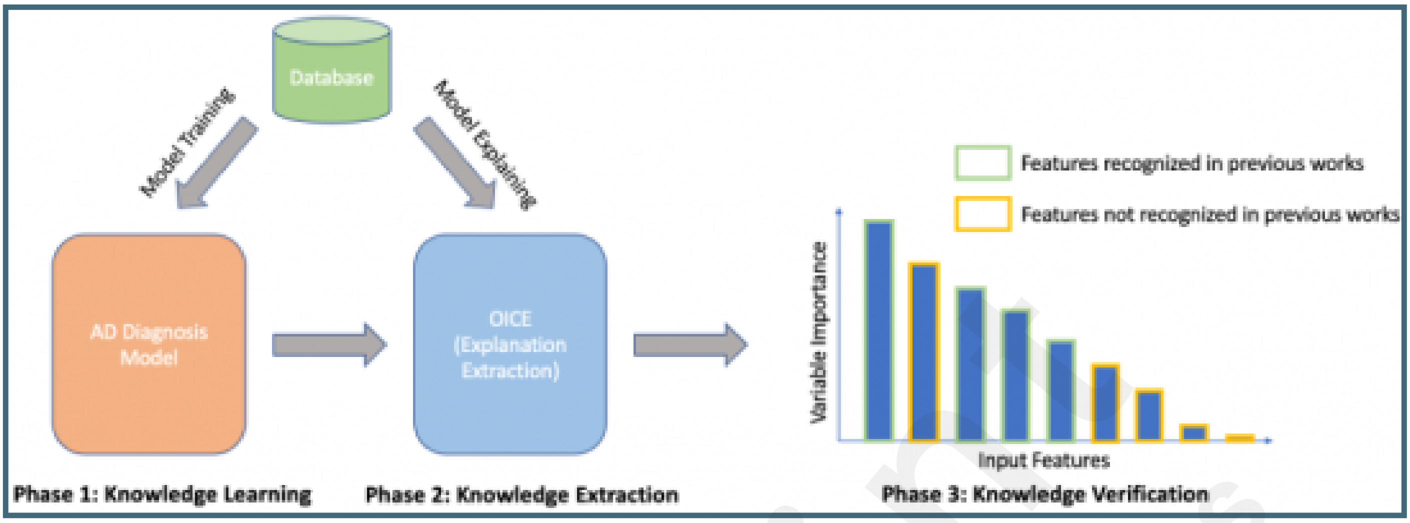
18. Wachter S, Mittelstadt B, Russell C. arXiv. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR URL: https://arxiv.org/abs/1711.00399 [accessed 2021-2-18]
19. Poyiadzi R, Sokol K, Santos-Rodriguez R, De Bie T, Flach P. FACE: Feasible and Actionable Counterfactual Explanations. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, New York, NY, USA, p.344–350. [CrossRef]
20. Joshi S, Koyejo O, Vijitbenjaronk W, Kim B, Ghosh J. arXiv. 2019. Towards realistic individual recourse and actionable explanations in black-box decision making systems URL: https://arxiv.org/abs/1907.09615 [accessed 2021-02-18]
21. Ustun B, Spangher A, Liu Y. Actionable Recourse in Linear Classification. In Proceedings of the Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, p.10–19. [CrossRef]
22. Russell C. Efficient Search for Diverse Coherent Explanations. In Proceedings of the Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, p.20–28. [CrossRef]
23. Karimi A H, Schölkopf B, Valera I. Algorithmic Recourse: from Counterfactual Explanations to Interventions. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, p.353–362. [CrossRef]
24. Goudet O, Kalainathan D, Caillou P, Guyon I, Lopez-Paz D, Sebag M. arXiv. 2017. Causal generative neural networks URL: https://arxiv.org/abs/1711.08936 [accessed 2021-02-18]
25. Boller F and Becker J. 2005. Dementiabank database guide. University of Pittsburgh. URL: https://dementia.talkbank.org/
26. Code of Ethics, TalkBank. https://talkbank.org/share/ethics.html
27. Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc. URL: https://www.nltk.org/book/
28. Gretton A, Borgwardt K M, Rasch M J, Schölkopf B, Smola A. A kernel two-sample test 2012 Mar;13:723-773. [CrossRef]
29. Spirtes P, Glymour C, Scheines R. Causation, prediction, and search. 2nd edition. 2001. ISBN: 9780262194402
30. Part-of-speech tutorial URL: https://sites.google.com/site/partofspeechhelp/
31. Toutanova K, Klein D, Manning C D, Singer Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. Association for Computational Linguistics, USA, p.173–180. [CrossRef]
32. Varley R. Deictic terms, lexical retrieval and utterance length in aphasia: an investigation of inter-relations. Eur J Disord Commun 1993;28(1):23-41. [Medline]
33. Fraser K C, Meltzer J A, Rudzicz F. Linguistic Features Identify Alzheimer's Disease in Narrative Speech. J Alzheimers Dis 2016;49(2):407-22. [Medline]
34. Almor A, Kempler D, MacDonald M C, Andersen E S, Tyler L K. Why do Alzheimer patients have difficulty with pronouns? Working memory, semantics, and reference in comprehension and production in Alzheimer's disease. Brain Lang. 1999 May;67(3):202-27. [Medline]
35. Almor A, Aronoff J M, MacDonald M C, Gonnerman L M, Kempler D, Hintiryan H, Hayes U L, Arunachalam S, Andersen E S. A common mechanism in verb and noun naming

deficits in Alzheimer's patients. Brain Lang 2009 Oct;111(1):8-19. [Medline]

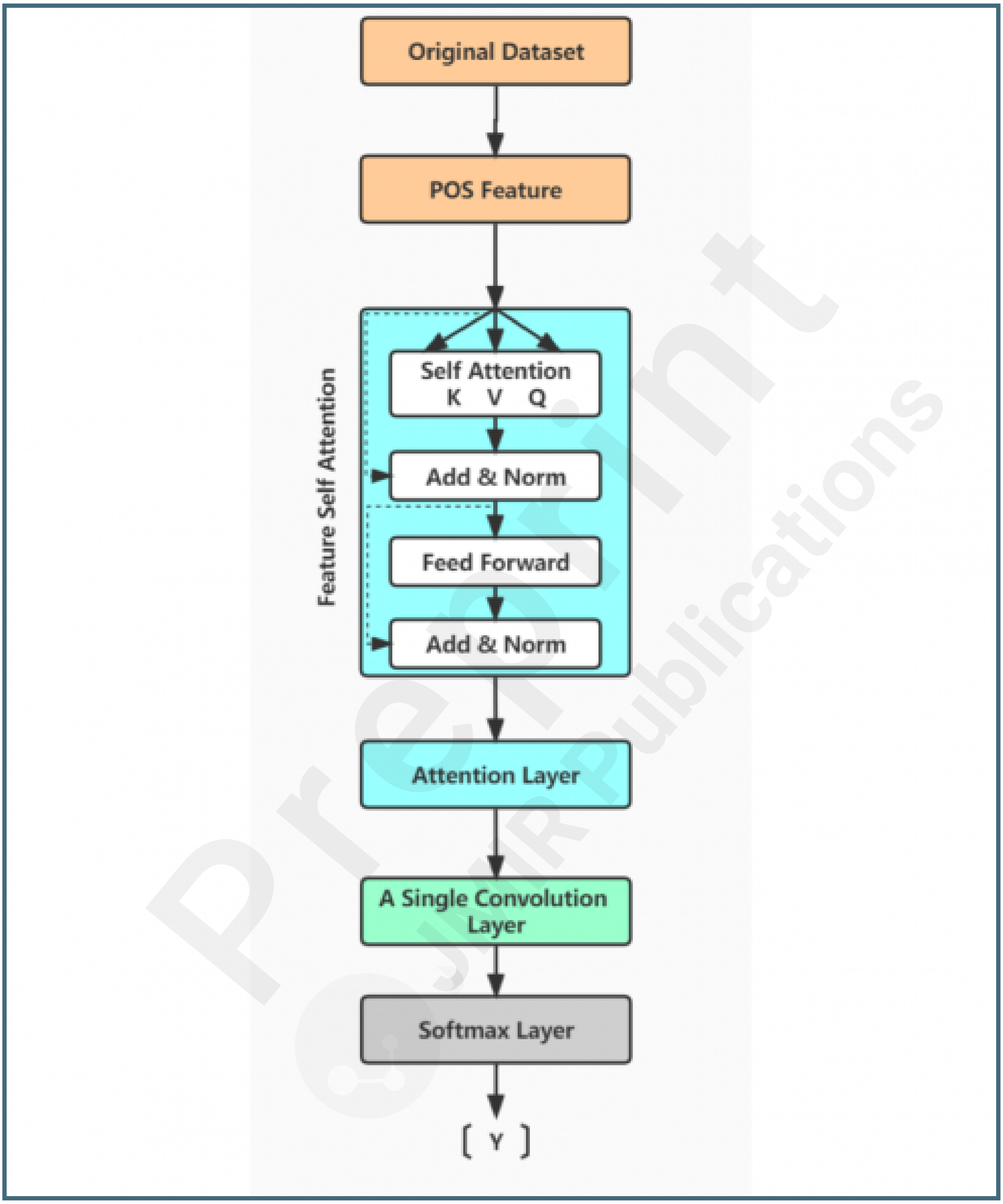
Supplementary Files

Figures

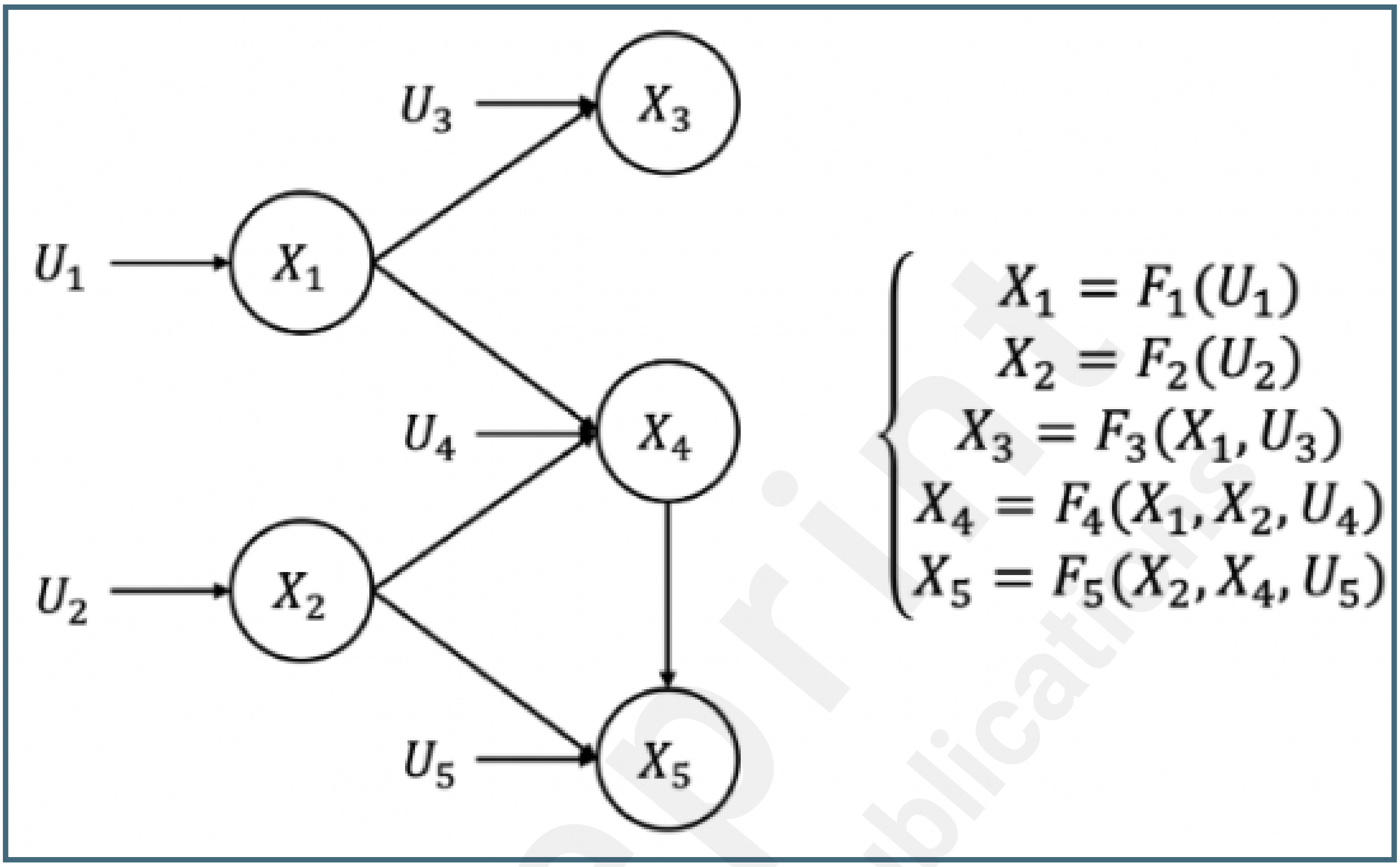
Method Overview. Procedures of using XAI for scientific discovery.


Database
Model Training
Model Explaining
AD Diagnosis Model
OICE (Explanation Extraction)
Features recognized in previous works
Features not recognized in previous works
Variable Importance
Input Features
Phase 1: Knowledge Learning
Phase 2: Knowledge Extraction
Phase 3: Knowledge Verification

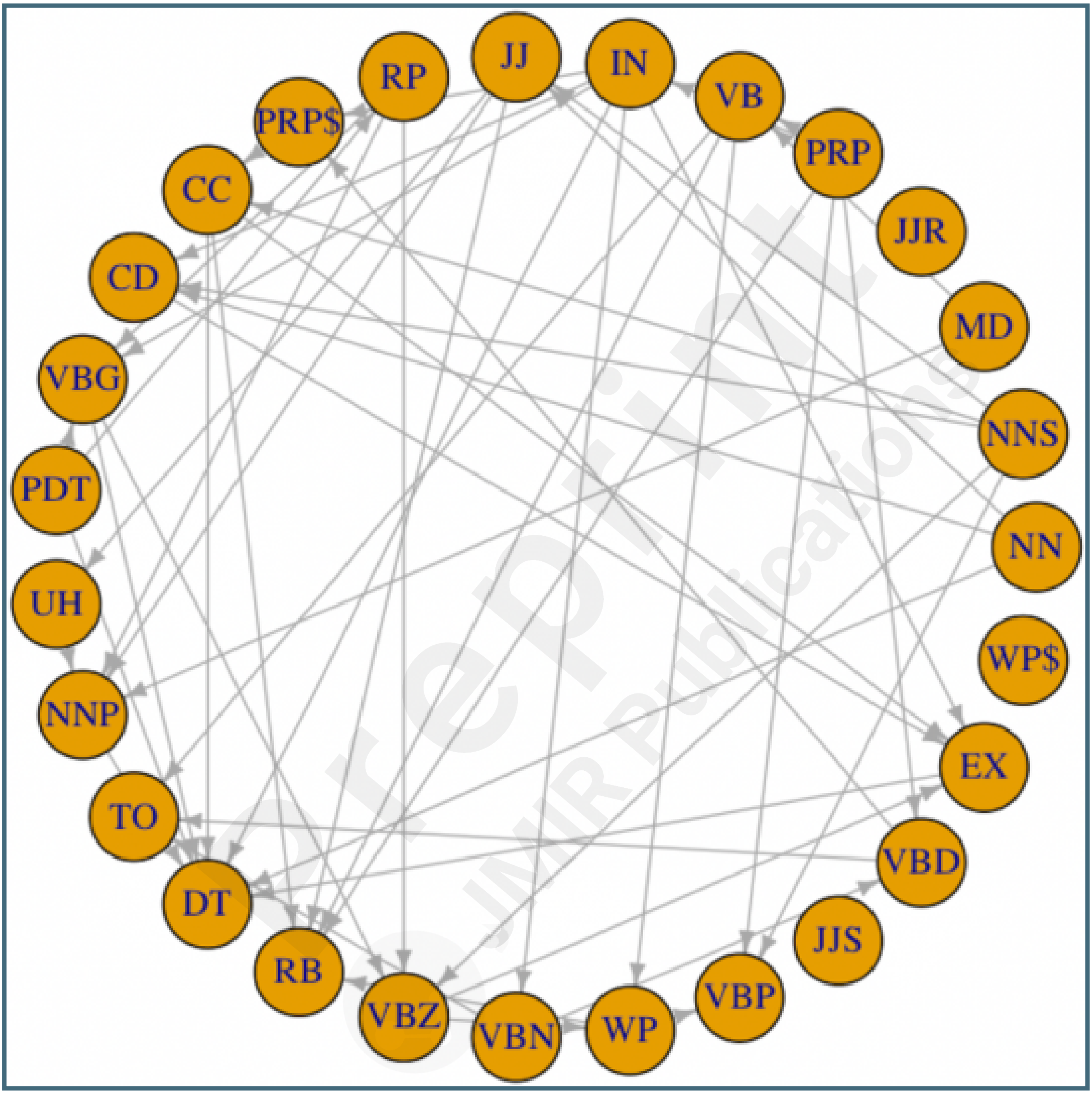
The proposed transformer-based classifier that uses the PoS features of the patient/control’s description.



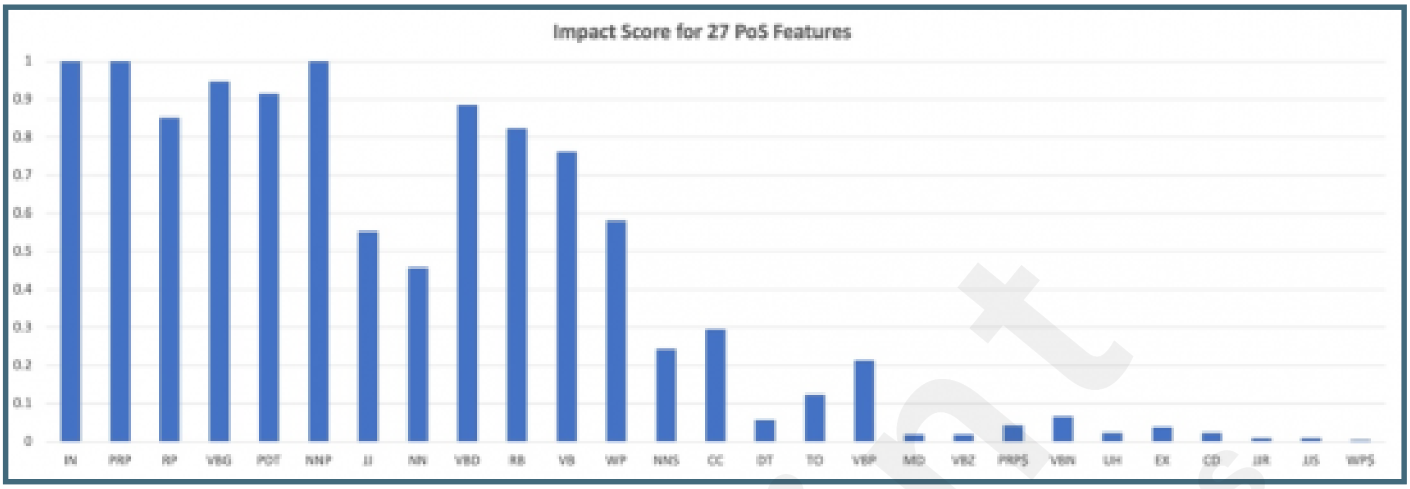
Example of a Structural Causal Model (SCM). Left: causal graph, right: causal mechanisms. As for CGNN, each causal mechanism is implemented with a generative neural network.



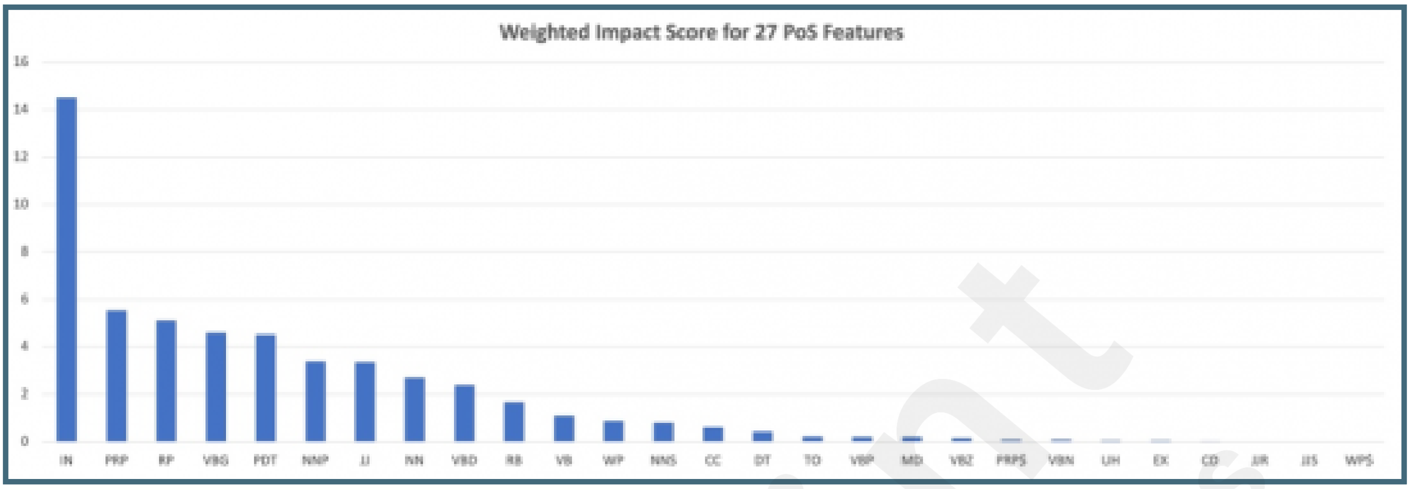
Causal Graph for 27 Linguistic features. The starting variable of each directed edge represents the cause, and the ending variable represents the effect.



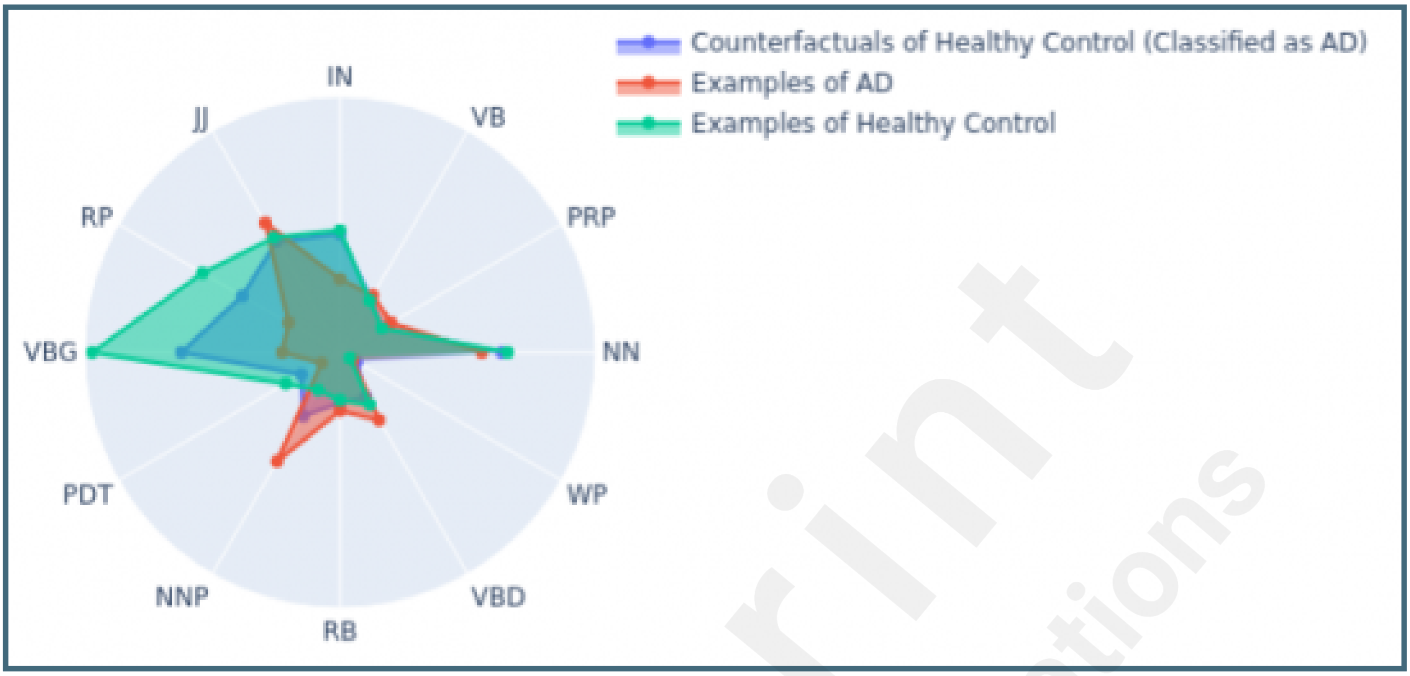
The impact score (IS) for 27 PoS features. Feature with higher IS value denotes more samples successfully flipping the model's outcome by intervening on it.



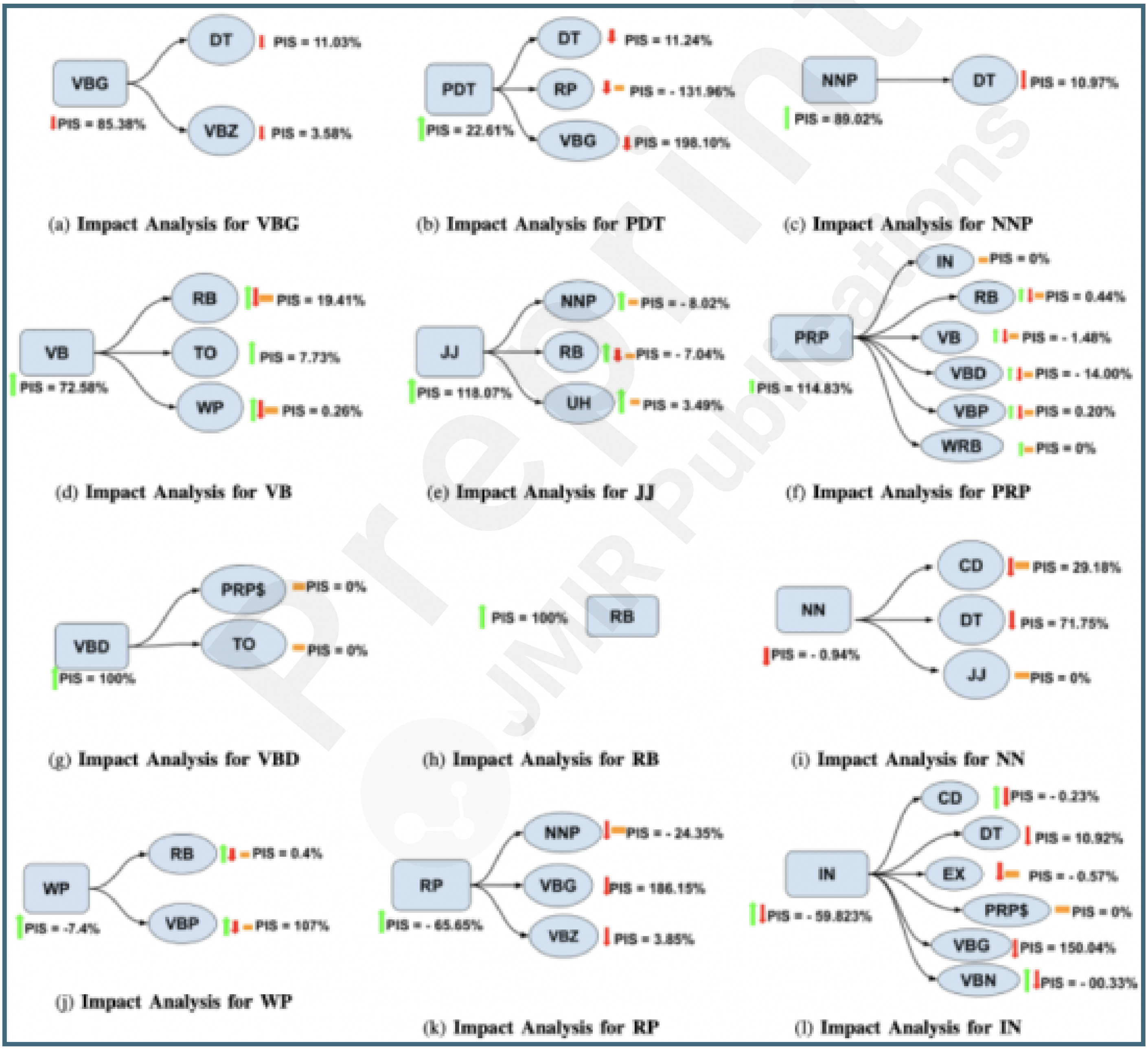
The weighted impact score (wIS) for 27 PoS features. Features with higher values denote more importance for machine learning in making decisions.



Spider plot of samples for AD patients, healthy control, and counterfactual samples (classified as AD patients).



Explanations for the representative features: For an intervened feature: the red down-arrow indicates a decrease of value is required for flipping the outcome, for a child node (feature): the red down-arrow indicates the changing direction caused by the intervention. The same rule applies to the green up-arrow (an increase of value) and the orange horizontal line (no change of value) (a)-(d) Cooperative: We consider the features to be "cooperative" if both the intervened feature and its descendent features contribute to flip the outcome. (e)-(h) Dominant: we define the feature as dominant if the intervened feature significantly contributes to flip the outcome while its descendent features make no or opposite contribution. (i)-(j) Idling: We define the intervened feature as "idling" if it itself only contributes to flip the outcome slightly while the child features make a significant contribution. (k)-(l) Inverse: We term the feature "inverse" if the change of the intervened feature moves the original instances away from the decision boundary, but it causes other features to significantly push the original instances forward to the decision boundary.



(a) Impact Analysis for VBG

(b) Impact Analysis for PDT

(c) Impact Analysis for NNP

(d) Impact Analysis for VB

(e) Impact Analysis for JJ

(f) Impact Analysis for PRP

(g) Impact Analysis for VBD

(h) Impact Analysis for RB

(i) Impact Analysis for NN

(j) Impact Analysis for WP

(k) Impact Analysis for RP

(l) Impact Analysis for IN