Sensitivity analysis for feature importance in predicting Alzheimer's Disease

Sensitivity Analysis for Feature Importance in Predicting Alzheimer's Disease

Akhila Atmakuru[1] Giuseppe Di Fatta[2] Giuseppe Nicosia[3]AliVarzandian[4] and Atta Badii[5]

[1] University of Reading, UK, akhila.atmakurupt@gmail.com
[2] Free University of Bozen-Bolzano, Italy, giuseppe.difatta@unibz.it
[3] University of Catania, Italy, giuseppe.nicosia.1@gmail.com
[4] University of Reading, UK, varzandian.ali@gmail.com
[5] University of Reading, UK, atta.badii@reading.ac.uk

Abstract. Artificial Intelligence (AI) classifier models based on Deep Neural Networks (DNN) have demonstrated superior performance in medical diagnostics. However, DNN models are regarded as "black boxes" as they are not intrinsically interpretable and, thus, are reluctantly considered for deployment in healthcare and other safety-critical domains. In such domains explainability is considered a fundamental requisite to foster trust and acceptability of automatic decision-making processes based on data-driven machine learning models. To overcome this limitation, DNN models require additional and careful post-processing analysis and evaluation to generate suitable explainability of their predictions. This paper analyses a DNN model developed for predicting Alzheimer's Disease to generate and assess explainability analysis of the predictions based on feature importance scores computed using sensitivity analysis techniques. In this study, a high dimensional dataset was obtained from Magnetic Resonance Imaging of the brain for healthy subjects and for Alzheimer's Disease patients. The dataset was annotated with two labels, Alzheimer's Disease (AD) and Cognitively Normal (CN), which were used to build and test a DNN model for binary classification. Three Global Sensitivity Analysis (G-SA) methodologies (Sobol, Morris, and FAST) as well as the SHapley Additive exPlanations (SHAP) were used to compute feature importance scores. The results from these methods were evaluated for their usefulness to explain the classification behaviour of the DNN model. The feature importance scores from sensitivity analysis methods were assessed and combined based on similarity for robustness. The results indicated that features related to specific brain regions (e.g., the hippocampal sub-regions, the temporal horn of the lateral ventricle) can be considered very important in predicting Alzheimer's Disease. The findings are consistent with earlier results from the relevant specialised literature on Alzheimer's Disease. The proposed explainability approach can facilitate the adoption of black-box classifiers, such as DNN, in medical and other application domains.

Keywords: Sensitivity Analysis, Explainability, Neural Network, predicting Alzheimer's, Feature Importance, SHAP, Sobol, Morris, Fast and High Dimensional Dataset.

1 Introduction

In recent times, there has been a surge in the usage of AI tools in the field of healthcare. These AI tools, including Machine Learning (ML) and Deep Neural Network (DNN) based models, are being used for the diagnosis and prediction of degenerative neurological disorders such as Alzheimer's Disease (AD), Parkinson's Disease (PD), and Mild Cognitive Impairment (MCI). In situations that involve critical healthcare, it is essential that the AI models output should be explainable, interpretable, accurate, reliable, and trustworthy. However, the DNN models are often perceived to lack transparency and re-adaptability, and their black-box nature and high level of complexity make for poor explainability and interpretability of the basis of their inferences. Therefore, it is imperative to evaluate and enhance their explainability to improve their reliability and acceptability. Sensitivity analysis is an effective approach for assessing and improving the explainability of the model.

Sensitivity Analysis (SA) examines the effects of variations in independent input variables and model parameters on the output of the model. This analysis is useful for researchers to gain a deeper understanding of the internal state vectors of the model and the rationale behind the predictions. Additionally, this study can facilitate the refinement or modification of the model, focusing on significant aspects or addressing any underlying defects. Consequently, SA can contribute to the valuation and enhancement of the interpretability and explainability of the model, ultimately leading to greater reliability, transparency, and reduced complexity. While various interpretations of interpretability and explainability exist in the literature, this analysis adopts the definition in [1].

Interpretability refers to the investigation of the model parameters and their impact on the output. The process requires comprehending the internal mechanics of the model and applying that understanding to make forecasts. This enhances the credibility of the model and ensures equitable predictions. Machine Learning (ML) models are highly interpretable as they are simple, use fewer parameters, and adhere to a set of operational principles. Therefore, ML models are highly interpretable. Conversely, DNNs are essentially black boxes by nature and exhibit intricate structure with numerous layers and complex operations, which makes it difficult to understand their inner workings and grasp the rationale behind their predictions.

Explainability refers to a detailed examination of the model input to evaluate its impact on the output. It refers to the ability to clearly explain predictions and to pinpoint the variables that affected them. Explainability leads to a clear and intuitive explanation for determining the most essential or important input features for the developed model and its predictions. This also enables understanding of the faults and biases in the model that impact its performance. While achieving interpretability for DNNs is challenging, explainability can be utilised to comprehend the most influential input features of the model and the rationale behind predictions.

Sensitivity Analysis [2] encompasses two complementary techniques: Global Sensitivity Analysis (GSA) and Local Sensitivity Analysis (LSA). Although both methods can be used for interpretability and explainability, this study primarily focuses on explainability. GSA examines variations in the model output behaviour with respect to the entire input space. The method involves varying the value of all input features at once and measuring the resulting output changes, with the objective of identifying the input features having the most influence on the output. LSA examines how variations in one independent input variable impact the model output. The method involves altering the value of one input feature at a time while observing the output changes. Therefore, GSA is used to understand key features across the entire dataset, while LSA is used to understand essential features in a single case. The global sensitivity offers a comprehensive explanation of the model. Two popular GSA methodologies and tools that are useful for assessing explainability are SHAP and SALib python libraries consisting of Sobol, Morris, and FAST methods.

Sobol and FAST are variance-based methods, which provide a quantifiable measure of the relative importance of each variable and its interactions by breaking down the output of the model variation into contributions from each input variable and their interactions. Sobol is computationally demanding for large datasets because of the enormous number of model assessments that must be conducted. Whereas, FAST computes sensitivity indices quickly using the Fast Fourier Transform, making it an efficient tool for high- dimensional datasets.

The Morris method uses a screening technique, Morris elementary effects, to estimate the influence of each input variable on the output by changing one variable at a time and estimating the change in output. The Morris method generates sample space using a different sampling strategy, and various strategies yield different sensitivity analyses. For the Morris method to provide sample data that is typical of actual data, the input variables are required to be independent in nature however, the method reveals interactions between the input features as an output. SHAP is a game theory-based method that generates feature significance as a Shapley value, which indicates the average contribution of the feature to the model output. SHAP delivers consistent results and clearly explains complicated interactions.

The present paper describes assessment of explainability of DNN models developed for detecting Alzheimer's Disease, based on feature importance scores determined using global sensitivity analysis. The classification was performed on a high-dimensional Alzheimer's dataset with two labels: Alzheimer's Disease and Cognitively Normal. Two datasets were used: one that comprised all feature metrics gathered from the FreeSurfer and the other that included a subset with fewer metrics. The GSA methodologies or tools, namely SHAP and SALib libraries containing Sobol, Morris, and FAST methods, were used to identify the feature importance scores that are significant for explaining the model. The first approach used the SALib python package and implemented Sobol, Morris, and FAST methods for determining the feature importance scores for the two datasets and evaluating the three methods and their outcomes. The second approach computed the feature importance scores using SHAP. Finally, the results obtained from the sensitivity analysis methods were analysed to determine similarity among them and the resultant methods were combined to form the final list of the most importan features.

The remainder of this article is structured as follows: Section 2 presents a brief literature survey, Section 3 describes the datasets utilised for analysis, Section 4 elaborates the implementation of diverse global sensitivity analysis methodologies for ascertaining the feature importance, while Section 5 provides the obtained results and ensuing discussions. Lastly, Section 6 presents some conclusions.

2 Literature Review

This literature review is focused on different research approaches utilised for identifying feature importance using sensitivity analysis, the use of sensitivity analysis in the context of deep neural networks and some general aspects of the pre-diagnostics for the detection of Alzheimer's Disease.

The words interpretability and explainability are frequently used interchangeably in the broad literature, although it is more appropriate to adopt different meanings in specialised work. A distinction between interpretability and explainability has been used in [1], which correspond to a standard definition for the terms that is widely accepted. The interpretable models typically offer a justification inherently, i.e. they provide a way to understand the logic behind their output in terms of semantics of the target problem and the input variables. Interpretability is typically implicitly provided by the model itself. On the other hand, the explainability of a model refers to a posthoc justification that is explicitly created after the model is built for the specific purpose to understand a model that typically is not interpretable on its own. Explainability is typically aimed at linking predictions and the input variables that influenced them by building a supplemental model for that purpose.

Some of the popular methodologies used for sensitivity analysis are Sobol, FAST and Morris. Shapely Additive exPlanations (SHAP) is a popular method used to generate posthoc explanations for black-box machine learning models. SHAP is based on the Shapley values from game theory and measures the contributions of each input feature to the predictions. In order to do that, SHAP compares the results of two scenarios, one which includes a feature and another without that feature. This comparison is performed for all possible combinations of the features. The obtained Shapley values indicate feature importance scores, which are computed as the average marginal contribution of the feature when included in the feature subset and when excluded from the subset. SHAP was found to be more robust and consistent with results [3], when compared with LIME (Local Interpretable Model-Agnostic Explanations), another explainability method. However, another study found that SHAP could be computationally very expensive for large numbers of input variables [4].

Sobol is a variance-based GSA method which quantifies the contribution of each input feature to the variance in the output of a model. Sobol analysis involves computing two values, the first order Sobol index and the total effect Sobol index. The first-order Sobol index is used to compute the contribution of each input feature to the variation of the output, whereas the total-effect Sobol index is used to compute the overall contribution of an input feature and its interactions with other features [5]. The method assumes the features are independent and uncorrelated however second order indices which is interaction between the input features is given as an output [6].

The Morris approach is a variance-based GSA method whereby each input feature is perturbed (changed) one at a time and the corresponding output variance is measured. A scaled variance of a uniform distribution determines the amount of required perturbation. Each component is perturbed numerous times to produce a series of basic effects. The average of these Morris elementary effects absolute values offers an approximation of the most important of those input features [7]. However, this approach requires a large number of samples for computing the feature importance. Also, it requires the input data to be independent in nature for it to generate a sample space which is similar to the input data. The methodology produces an infinity in the generated samples for the inputs that are not meeting the expectations of the Morris approach [8].

The Fourier Amplitude Sensitivity Test (FAST) [9] method involves variance based global sensitivity analysis. The approach employs Fourier analysis which involves producing a set of Fourier coefficients based on the input features and utilising these coefficients to estimate the variation of the output of the model. The Fourier coefficients are then used to determine the relative contribution of the variation of each input feature and rank the features depending on their contribution to the output variance. The approach is quite fast and efficient even while dealing with a high dimensional dataset as it uses Fourier transform techniques.

In [10], the SHAP methodology was applied along with other sensitivity analysis for identifying the feature importance of multiple subsets of Alzheimer's Disease dataset from ADNI. The procedures carried out by the study were feature selection, model training, and sensitivity analysis using machine learning models. The most significant features found in one of the datasets were the hippocampal and the cortical thickness.

Alzheimer's Disease (AD) is a neurodegenerative disease resulting in permanent damage to the neurological system of th brain. AD is externally characterised by behavioural changes such as language and short-term memory issues, trouble in executing tasks that require cognition, changes in personality, and loss of social and inter-personal skills, which act as early indicators of the disease (Radiology.org) [11]. The onset and the progress of the disease can be detected from observable changes in the physiological structures in the brain, which may be obtained from radiological images. An accurate diagnostic tool should be able to select the already established physiological changes with high accuracy. Research shows that AD is directly linked to structural changes in the hippocampus and in the entorhinal cortex located in the medial temporal lobe of the brain [12].

3 Datasets

The present study utilized a data set that contained brain T1-weighted structural MRI scans with a slice thickness of 1.5 mm obtained from 1901 research participants. The data for this study was sourced from three publicly available data repositories, namely the Australian Imaging Biomarker & Lifestyle Flagship Study of Ageing (AIBL), Alzheimer's Disease Neuroimaging Initiative (ADNI), and Information eXtraction from Images (IXI).

The main purpose of this study was to develop a tool to support improvement in the diagnosis of AD. It was noted that the ADNI library is composed of multiple images of the same subject taken from various studies. To ensure consistency, screening and baseline scans were selected as the adopted photos since they were the earliest accessible scans of a subject. When there were multiple images of the same subject the image with the greatest contrast-to-noise ratio (CNR) was chosen.

To facilitate operations such as skull stripping, image registration, cortical and subcortical segmentation, hippocampal subfields segmentation, and calculation of cortical thickness, surface, and volume, all images were pre-processed using FreeSurfer version 6.0 [13].

A significant number of files containing numerical measurements connected to specific regions of interest were produced during the pre-processing stage (ROI). KNIME [14] and its extension KSurfer [15] were used to extract, filter, and clean the data created by the pre-processing.

A total of 446 traits were obtained from the data generated by FreeSurfer. The utilisation of ICV normalisation and the estimated total intracranial volume (ICV) were not included. Inaccurate or duplicate features were discarded during the data cleansing phase, resulting in the elimination of 42 traits. The numerical measurements of the brain were referred to as the feature set F = {fi}, with |F| = 404. The selection of traits was not based on specific domain expertise.

The raw dataset included information from both the right and left-brain hemispheres. Within each hemisphere, five distinct types of metrics were calculated from MRI scan images, comprising volume, thickness, thickness standard deviation, mean curvature, and area. Our concentration for feature selection was specifically on AD and CN as targets to perform experiments. Dataset 1 consisted of 401 features for classes AD and CN that encompassed all of the metrics excluding Gender, Age and Label. Dataset 2 consisted of 265 features for classes AD and CN that encompassed all of the metrics excluding Gender, Age and Label.

4 Methodologies

This section outlines the two methodologies implemented for conducting sensitivity analysis on two distinct Deep Neural Network (DNN) models coded in Python. The sensitivity analysis leveraged two Python method libraries, namely SHAP and the SALib for Sobol, Morris, and FAST methods. The goal of the analysis methodologies was to execute sensitivity analysis on a specified DNN model for a given input dataset and produce output predictions utilising different methods. The analysis outcome comprises of a set of features and their corresponding feature importance scores, specific to a designated DNN model, input dataset, and analysis methodology. Subsequently, the resulting feature importance scores across all methods are evaluated to determine the most significant features.

The sensitivity analysis was conducted on two distinct DNN models, which differ by the input dataset and the internal architecture of the model. As previously discussed in Section 3 Datasets, there exist two distinct datasets of features. The dataset 1 consists of 401 features, excluding the Label, whereas the 2 dataset 2 consists of a subset of 265 features, excluding the Label.

The DNN Model 1 meant for analysing 401 features and is composed of an initial layer that utilises the 'ReLU' activation function, 3 sets of hidden layers along with dropout layers, and a last output layer that uses a Sigmoid activation function. The input layer contained neurons tailored for 401 features with the activation of 'ReLU'. Each hidden layer, which contained the 'ReLU' activation function, was followed by a dropout layer with a dropout rate of 30%. The initial hidden layer set comprises 200 neurons, the second hidden layer set comprises 100 neurons, and the last hidden layer set comprises 50 neurons. The ultimate output layer included one neuron that employed a sigmoid activation function for binary classification. Prediction accuracy of the DNN Model 01 obtained with Dataset 1 using Stratified 10 Cross validation was approx. 91% with 3% standard deviation.

The DNN Model 2 meant for analysing 265 features. It featured an input layer that utilised the 'ReLU' activation function, 3 sets of hidden layers along with dropout layers, and a final output layer that employed the Sigmoid activation function. The input layer contained neurons specifically designed for 265 features and used the 'ReLU' activation function. Each hidden layer started with the 'ReLU' activation function and ended with a dropout layer that has a 30% dropout rate. The first hidden layer had 150 neurons, the second had 75, and the last hidden layer had 30 neurons. The final output layer consisted of a single neuron that was equipped with a sigmoid activation function for binary classification. Prediction accuracy of the DNN Model 2 obtained with Dataset 2 using Stratified 10 Cross validation was approx. 91.4% with 4.4 as standard deviation.

4.1 Methodology 1

This section describes the Python implementation of the SALib library for the Sobol, Morris, and FAST methods. The implementation made use of two distinct datasets, namely Dataset 1 and Dataset 2. Individual DNN models were created for each dataset. The objective of this methodology is to perform various sensitivity analyses on the two DNN models and to identify the most important features in predicting Alzheimer's Disease.

The implementation procedure begins with the preparation of the dataset for model training. Datasets 1 and 2 are used to define training features and the target label, the training dataset is then scaled in preparation for normalisation. The DNN Models 1 and 2 were trained using Dataset 1 and Dataset 2, respectively. Using the trained DNN models and corresponding datasets, the Sobol, Morris, and FAST methods were utilised for prediction.

For each of the methods, a definition of the original data was provided which includes the data mean, standard deviation, and distribution as parameters for generating the sample dataset. The sample size parameter was also provided which determined the resultant sample size of the dataset. The resultant sample dataset was created by following the perturbation techniques of the respective specified methods which involved adding small amounts of noise to the original dataset and increasing the dataset size.

The generated resultant sample dataset was used for prediction. The predicted target and the definition of the dataset was provided to the ‘analyse’ function of specified methods. The output from the ‘analyse’ function provided the results of sensitivity analysis containing feature importance scores.

To obtain stable, and reliable results, the above procedure was run a number of times by training the DNN model with a specified dataset, then prediction using the resultant sample dataset generated by the specified method and subsequent analysis for feature importance.

The DNN models are usually randomly initialised with weights and biases which can lead to a slightly different output for each run. While training, stochastic optimisation techniques such as ADAM were used. This leads to a selection of different subsets of training data resulting in different outputs each time. Carrying out the training and prediction in the DNN models for multiple iterations will produce robust, reliable estimates of the performance of the model as well as the feature importance score.

The resultant feature importance scores from all the methodologies namely Sobol, Morris and FAST require post-processing of the data to identify the most important features for each method.

Figure 01 is a schematic flow chart representation of the Methodology 1. The same procedure wasfollowed for Sobol, Morris, and FAST methods.

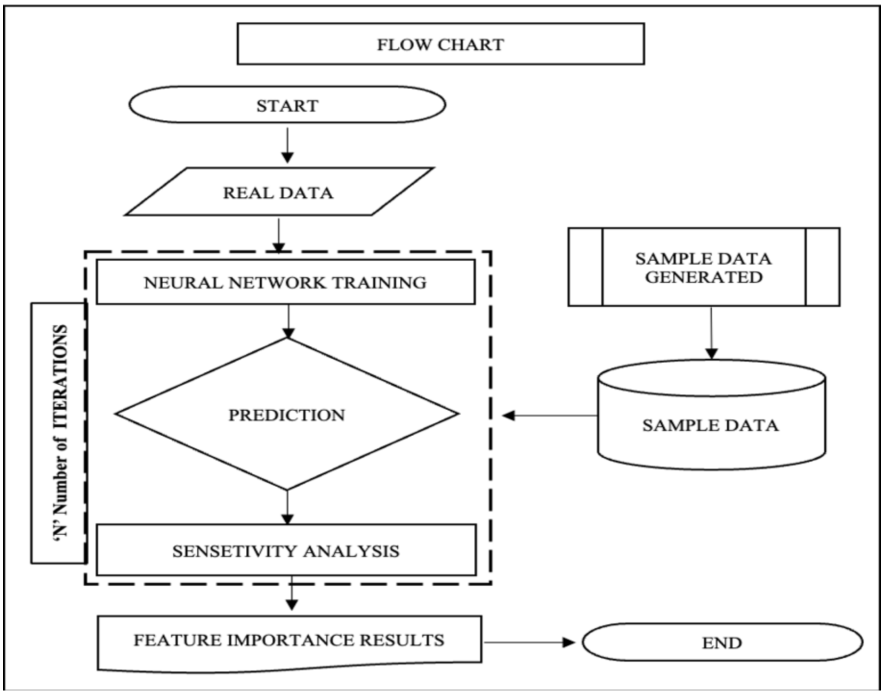

Fig. 1. Schematic Flowchart of Methodology 1

4.2 Methodology 2

This section describes the implementation of the SHAP method using python library. The implementation uses DNN model 1 and DNN model 2 which were created for dataset 1 and 2 respectively as described in the previous section. The objective of this methodology was to perform SHAP sensitivity analyses on the two DNN models and to identify the most important features in predicting Alzheimer's Disease.

The implementation procedure begins with the preparation of the dataset for model training. The dataset 1 and 2 was used to define training features and the target label. The training dataset was then scaled in preparation for normalisation. The corresponding datasets were used to train both DNN Model 1 and Model 2.

In order to carry out Shapley analysis, the SHAP explainer function was initialised using the input data and the trained DNN model. for generating Shapley values. The analysis focus was to explicate the effect of the input data on the output for the respective model.

The procedure followed for Methodology 2 comprised of multiple runs of the DNN models to produce robust and reliable estimates of feature importance. The resultant feature importance scores from SHAP require post processing to identify the most important features for the methodology.

Figure 02 is a schematic flow chart representation of the Methodology 2 using the SHAP method.

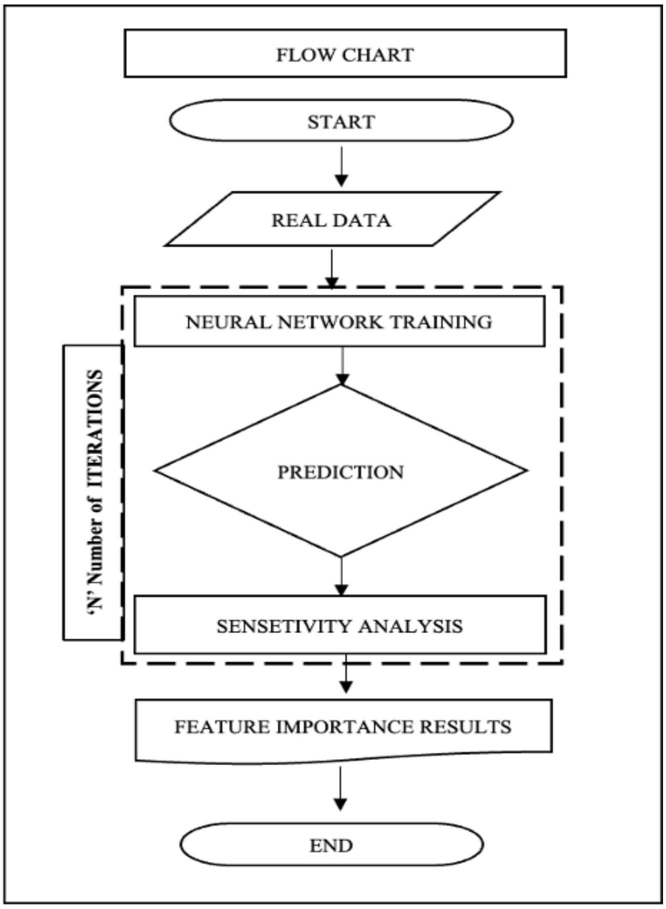

Fig. 2. Schematic Flowchart of Methodology 2

5 Results and Discussions

Explainability of DNN classifier models developed for detection of Alzheimer's Disease was assessed using compatible Alzheimer's datasets and sensitivity analysis methods, namely SHAP and SALib containing Sobol, Morris and FAST.

The two DNN models 1 and 2 were developed and trained using datasets 1 and 2 respectively. The developed models were analysed using SHAP, Sobol, Morris, and FAST methods. Under Methodology 1, Sobol, Morris and FAST methods were used to analyse both DNN models. Under Methodology 2, SHAP method was used to analyse both DNN models. In each analysis, the chosen method was run 500 times for Dataset 1 with 401 features and 300 times for Dataset 2 having 265 features so as to average out any fluctuations in the obtained outputs. In each analysis, feature importance scores were obtained as outputs.

The features importance score obtained using 4 methods was thoroughly analysed to determine their similarities. Each feature importance score list was converted into a corresponding ranking pattern to show difference between the corresponding rankings over the number of specified features.

The absolute difference between two rankings was computed using equation (1) for a specified number of selected features. This formula quantifies the relative discrepancy by dividing the absolute difference between rankings by the specified number of features. The averaging of rank differences provides a measure of central tendency that represents the overall similarity between the compared lists. This averaging of the results yields a single aggregate value to represent the collective similarity. This analysis as shown in Figure 03 was performed over various sets of selected features values

$$\frac{abs(A-B)}{SNSF}$$

$$\text{If } A \leq SNSF \text{ or } B \leq SNSF \qquad (1)$$

where A is a rank from Rank list A, B is rank from rank list B and SNSF is the Specified Number of Selected Features.

Upon comparison, it was found that the results obtained from SHAP and Sobol methods demonstrated the highest degree of similarity. To ensure the reliability of the methodology, only the results from SHAP and Sobol methods were selected for further analysis.

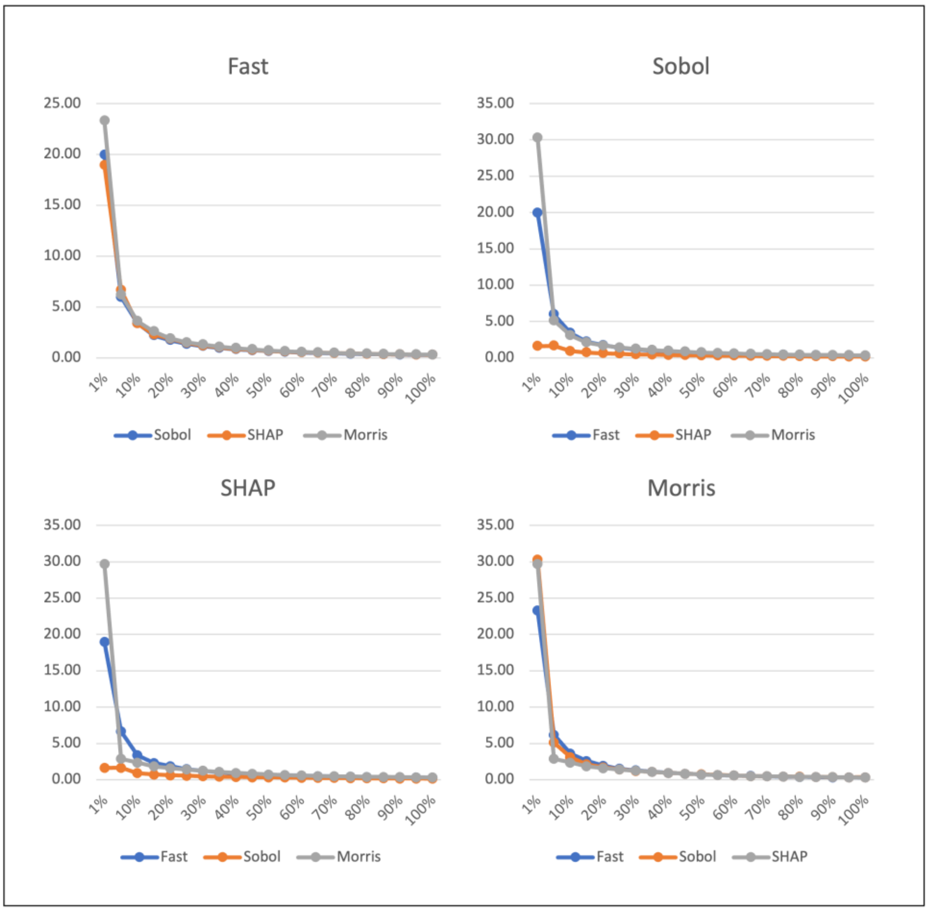

Fig. 3. Similarity analysis for 4 different approaches and 401 features dataset

To create the final feature importance ranking, the results obtained from SHAP and Sobol methods were combinedusing the Rank position (reciprocal rank) method [16] which is illustrated in equation (2). The rank score for document 'i' is computed, utilising its position information across all retrieval systems (j = 1 . . . n).

$$r(d_i) = \frac{1}{\sum_j \frac{1}{position(d_{ij})}} \qquad (2)$$

The Rank Position score was calculated for each document to be combined. These scores were then used to sort the documents into a non-decreasing order. The scores derived from this method were subsequently ranked to form the ultimate feature importance ranking.

Table 01 describes the 20 most important features recognised by the rank reciprocal method out of the 401 features. The results were obtained from DNN model 01 using SHAP and Sobol methods. The table contains a list of features along with their corresponding medical terminologies and references to medical literature.

Table 1. 20 most important features recognised by the rank reciprocal method out of the 401 features.

Feature Name	Medical Names	Medical Reference
Left-Inf-Lat-Vent	Temporal horn of left lateral ventricle	(Vernooij et al, 2020) [17]
Right-Inf-Lat-Vent	Temporal horn of right lateralventricle	(Vernooij et al, 2020) [17]
left_Hippocampal_tail	Hippocampal tail	(Zhao et al,2019) [18]
left_presubiculum	Pre subiculum	(Carlesimo et al,2015) [19]
Le ft_Whole_hippocampus	Hippocampus	(Rao et al,2022) [20]
left_molecular_layer_HP	Molecular Layer Hippocampus	(Stephen et al,1996) [21]
left_subiculum	Subiculum	(Carlesimo et al,2015) [19]
right_Hippocampal_tail	Hippocampal tail	(Zhao et al,2019) [18]
lh_bankssts_volume	Banks of Superior Temporal Sulcus	(Sacchi et al,2023) [22]
lh_bankssts_thicknessstd	Banks of Superior Temporal Sulcus	(Sacchi et al,2023) [22]
lh_parahippocampal_thickness	Para Hippocampal	(Van et al,2000) [23]
rh_paracentral_thicknessstd	Paracentral	(Yang et al,2019) [24]
right_subiculum	Subiculum	(Carlesimo et al,2015) [19]
rh_inferiorparietal_thickness	Inferior Parietal	(Jacobs et al,2012) [25]
lh_transversetemporal_meancurv	Transverse Temporal	(Peters et al,2009) [26]
Left-Amygdala	Amygdala	(Poulin et al,2011) [27]
left_hippocampal fissure	Hippocampal Sulcus	(Bastos et al,2006) [28]
left_GC-ML-DG	Granule Cell (GC) and Molecular Layer (ML) of the Dentate Gyrus (DG)	(Ohm et al, 2007) [29]
Right-Amygdala	Amygdala	(Poulin et al,2011) [27]
rh_inferiortemporal_volume	Inferior Temporal	(Scheff et al,2011) [30]

The results in the Table 01 and 02 highlight the significance of the brain regions indicated by the selected features in early detection of Alzheimer's Disease.

Table 02 outlines the 20 most important features as identified by the reciprocal ranking method out of the 265 features 60% of which are identical to the Table 01. The table findings have adopted the DNN model 02 for SHAP and Sobol along with their corresponding medical terminologies. Also, references to medical literature are provided,emphasising the importance of these features or brain regions in the early detection of Alzheimer's Disease.

Table 2. 20 most important features as identified by the reciprocal ranking method out of the 265 features.

Feature Name	Medical Names	Medical Reference
Left-Inf-Lat-Vent	Temporal horn of left lateralventricle	(Vernooij et al, 2020) [17]
Right-Inf-Lat-Vent	Temporal horn of right lateralventricle	(Vernooij et al, 2020) [17]

Feature Name	Medical Names	Medical Reference
right_Hippocampal_tail	Hippocampal tail	(Zhao et al,2019) [18]
left_presubiculum	Presubiculum	(Carlesimo et al,2015) [19]
left_subiculum	Subiculum	(Carlesimo et al,2015) [19]
left_Hippocampal_tail	Hippocampal tail	(Zhao et al,2019) [18]
left_hippocampal-fissure	Hippocampal Sulcus	(Bastos et al,2006) [28]
lh_parahippocampal_thickness	Para Hippocampal	(Van Hoesen et al,2000) [23]
left_molecular_layer_HP	Molecular Layer Hippocampus	(Stephen et al,1996) [21]
rh_entorhinal_thickness	Entorhinal	(Van Hoesen et al,1991) [31]
rh_rostralmiddlefrontal_thickness	Rostral Middle Frontal	(Vasconcelos et al,2014) [32]
rh_inferiorparietal_thickness	Inferior Parietal	(Greene SJ et al,2010) [33]
left_Whole_hippocampus	Hippocampus	(Rao et al,2022) [20]
lh_precuneus_thickness	Precuneus	(Giacomo et al, 2022) [34]
Left-Amygdala	Amygdala	(Poulin et al,2011) [27]
Optic-Chiasm	Optic-Chiasm	(Sadun AA et al,1990) [35]
Right-Pallidum	Pallidum	(Miklossy J et al,2011) [36]
rh_entorhinal_volume	Entorhinal	(Van Hoesen et al,1991) [31]
right_presubiculum	Pre-Subiculum	(Carlesimo et al,2015) [19]
Left-Pallidum	Pallidum	(Miklossy J et al,2011) [36]

The comparison between Table 01 and Table 02 highlights the abiding significance of specific brain regions in the early detection of Alzheimer's Disease. The fact that both tables demonstrate a significant 60% overlap in characteristics, while using distinct models for validation, serves to highlight the resolute nature of these common traits. This prevailing trend as also confirmed by findings reported in relevant medical literature, serves to underscore the paramount importance of the specific brain regions in the field of Alzheimer's research.

The trustworthiness of medical diagnostic systems relies heavily on their repeatability. Utilising a vast number of input features, the DNN-based classifier accurately predicts outcomes, particularly with a select subset of features. To ensure the explainability of the DNN classifier, sensitivity analysis methods are employed to identify the most significant features. The repeatability of these crucial features signifies the consistency of the diagnosis. For evaluating the developed model, repeatability is assessed by analysing the important features obtained through SHAP and Sobol analyses. In SHAP analysis, the output of each iteration is scrutinised to determine the input feature that yields the highest score. By analysing the output of 500 iterations, it was possible to determine how frequently an input feature appeared with the maximum score. The results are presented as histograms in Figure 4 below which shows the similar outputs obtained through Sobol analysis. The recurrence of these features instils confidence in the accuracy of the diagnosis.

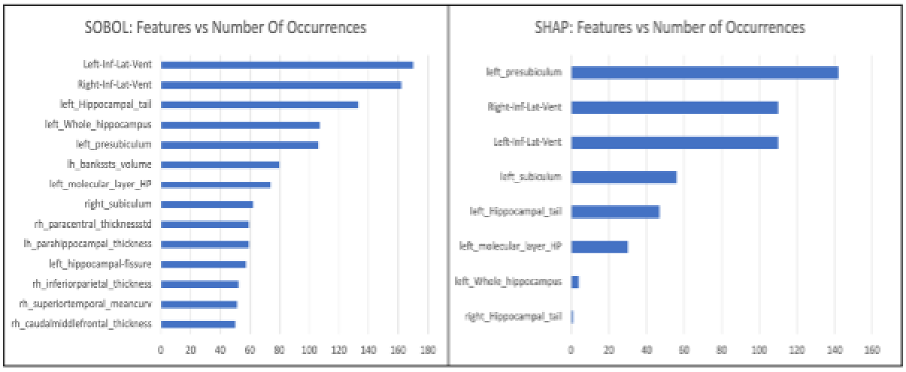

Fig. 4. Repeatability analysis for 2 different approaches with 401 features dataset

6 Conclusions

Classification models based on Deep Neural Networks (DNN) were developed for predicting Alzheimer’s Disease and were studied using sensitivity analysis techniques to assess model explainability based on feature importance scores. An Alzheimer’s dataset with two labels, Alzheimer’s Disease and Cognitively Normal, was utilised for the classification task. Two DNN models were developed to analyse two datasets of different sizes. The analysis aimed at assessing the extent of the explainability and used two approaches based, respectively, on SHAP and on Global Sensitivity Analysis (G-SA) techniques (Sobol, Morris and FAST) from the python library SALib.

In the study of Alzheimer's Disease diagnosis, numerous significant characteristics have been recognised as noteworthy in terms of their correlation with the evolution and diagnosis of the disease. These characteristics incorporate the temporal horn of the left and right lateral ventricles, the hippocampal tail, the pre-subiculum, the whole hippocampus, the molecular layer of the hippocampus, the subiculum, the banks of the superior temporal sulcus, the para hippocampal region, the paracentral area, the inferior parietal region, the transverse temporal area, the amygdala, the hippocampal sulcus, and the inferior temporal area. These features have been extensively scrutinized and linked with various aspects of Alzheimer's Disease, such as neuroimaging, structural modifications, atrophy, and clinical correlations. The assessment of the significance of these features has contributed to enhancing our understanding of the use of sensitivity analysis techniques for the explainability of machine learning models and for feature selection. A subset of important features was effectively utilised by the DNN-based classifier to accurately forecast outcomes.

By using sensitivity analysis techniques significant features were identified. The repeatability of the results of a medical diagnostic system is a crucial factor in determining its reliability. The consistent presence of some features in multiple analyses serves to enhance the credibility of the diagnostic model.

This research has made notable contributions to the field of computer science by advancing the use of machine learning models for Alzheimer's Disease prediction and addressing the critical aspect of explainability. By developing and analysing the results from Deep Neural Network (DNN) models, the study has contributed to insights into their performance and the extent of explainability. The research has focused on assessing the explainability of these models through sensitivity analysis techniques, shedding light on the significant variables. Furthermore, the study has conducted a comprehensive feature analysis, exploring key features such as the temporal horn of the lateral ventricles, hippocampal regions, and other relevant areas, providing a deep understanding of the complex dynamics of the disease. Additionally, by presenting a comparative analysis of SHAP and sensitivity analysis techniques, the research has provided valuable insights into their performance and suitability for feature importance assessment, guiding researchers in selecting the most effective approach for predictive models.

References

1. Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5), pp.206-215.
2. Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Piano, S.L., Iwanaga, T., Becker, W. and Tarantola, S., 2021. The future of sensitivity analysis: An essential discipline for systems modelling and policy support. Environmental Modelling & Software, 137, p.104954.
3. Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
4. Jia, R., Dao, D., Wang, B., Hubis, F.A., Hynes, N., Gürel, N.M., Li, B., Zhang, C., Song, D. and Spanos, C.J., 2019, April. Towards efficient data valuation based on the Shapley value. In The 22nd International Conference on Artificial Intelligence and Statistics (pp. 1167-1176). PMLR.
5. Sobol, I.M., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and computers in simulation, 55(1-3), pp.271-280.
6. Helton, J.C. and Davis, F.J., 2003. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliability Engineering & System Safety, 81(1), pp.23-69.
7. Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics, 33(2), pp.161-174.
8. Borgonovo, E. and Plischke, E., 2016. Sensitivity analysis: A review of recent advances. European Journal of Operational Research, 248(3), pp.869-887.
9. Cukier, R.I., Levine, H.B. and Shuler, K.E., 1978. Nonlinear sensitivity analysis of multiparameter model systems. Journal of computational physics, 26(1), pp.1-42.
10. Bloch, L., Friedrich, C.M. and Alzheimer's Disease Neuroimaging Initiative, 2022. Machine learning workflow to explain black-box models for early Alzheimer's Disease classification evaluated for multiple datasets. SN Computer Science, 3(6), p.509.
11. Radiology for patients, 'https://www.radiologyinfo.org/en/info/alzheimers'
12. Raji, C.A., Lopez, O.L., Kuller, L.H., Carmichael, O.T. and Becker, J.T., 2009. Age, Alzheimer Disease, and brain structure. Neurology, 73(22), pp.1899-1905.
13. Fischl, B., 2012. FreeSurfer. Neuroimage, 62(2), pp.774-781.
14. M. Berthold, N. Cebron, F. Dill, G. Di Fatta, T. Gabriel, F. Georg, T. Meinl, P. Ohl, C. Sieb, B. Wiswedel, "KNIME: the Konstanz Information Miner", Workshop on Multi-Agent Systems and Simulation (MAS&S), 4th Annual Industrial Simulation Conference (ISC), Palermo, Italy, June 5-7, 2006, pp.58-61.
15. Sarica, A., Di Fatta, G. and Cannataro, M., 2014. K-Surfer: a KNIME extension for the management and analysis of human brain MRI FreeSurfer/FSL data. In Brain Informatics and Health: International Conference, BIH 2014, Warsaw, Poland, August 11-14, 2014, Proceedings (pp. 481-492). Springer International Publishing.
16. Nuray-Turan, Rabia & Can, Fazli. (2006). Automatic ranking of retrieval systems using fusion data. Information Processing & Management. 42. 595-614. 10.1016/j.ipm.2005.03.023.
17. Vernooij, M.W., van Buchem, M.A. (2020). Neuroimaging in Dementia. In: Hodler, J., Kubik-Huch, R., von Schulthess, G. (eds) Diseases of the Brain, Head and Neck, Spine 2020–2023. IDKD Springer Series. Springer,Cham. https://doi.org/10.1007/978-3-030-38490-6_11
18. Zhao W, Wang X, Yin C, He M, Li S, Han Y. Trajectories of the Hippocampal Subfields Atrophy in the Alzheimer's Disease: A Structural Imaging Study. Front Neuroinform. 2019 Mar 22;13:13. doi: 10.3389/fninf.2019.00013. PMID: 30983985; PMCID: PMC6450438.
19. Carlesimo, G.A., Piras, F., Orfei, M.D., Iorio, M., Caltagirone, C. and Spalletta, G. (2015), Atrophy of pre-subiculum and subiculum is the earliest hippocampal anatom-ical marker of Alzheimer's Disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 1: 24-32. https://doi.org/10.1016/j.dadm.2014.12.001
20. Rao YL, Ganaraja B, Murlimanju BV, Joy T, Krishnamurthy A, Agrawal A. Hippocampus and its involvement in Alzheimer's Disease: a review. 3 Biotech. 2022 Feb;12(2):55. doi: 10.1007/s13205-022-03123-4. Epub 2022 Feb 1. PMID: 35116217; PMCID: PMC8807768.
21. Stephen Scheff, D. Larry Sparks, Douglas Price; Quantitative Assessment of Synaptic Density in the Outer Molecular Layer of the Hippocampal Dentate Gyrus in Alzheimer's Disease. Dementia 1 April 1996; 7 (4): 226– 232. https://doi.org/10.1159/000106884
22. Sacchi L, Contarino VE, Siggillino S, Carandini T, Fumagalli GG, Pietroboni AM, Arcaro M, Fenoglio C, Orunesu E, Castellani M, Casale S, Conte G, Liu C, Triulzi F, Galimberti D, Scarpini E, Arighi A. Banks of the Superior Temporal Sulcus in Alzheimer's Disease: A Pilot Quantitative Susceptibility Mapping Study. J Alzheimer's Dis. 2023;93(3):1125-1134. doi: 10.3233/JAD-230095. PMID: 37182885.
23. Van Hoesen, G.W., Augustinack, J.C., Dierking, J., Redman, S.J. And Thangavel, R. (2000), The Parahippocampal Gyrus in Alzheimer's Disease: Clinical and Preclinical Neuroanatomical Correlates. Annals of the New York Academy of Sciences, 911: 254-274. https://doi.org/10.1111/j.1749-6632.2000.tb06731.x

24. Yang H, Xu H, Li Q, Jin Y, Jiang W, Wang J, Wu Y, Li W, Yang C, Li X, Xiao S, Shi F, Wang T. Study of brain morphology change in Alzheimer's Disease and amnestic mild cognitive impairment compared with normal controls. Gen Psychiatr. 2019 Apr 16;32(2):e100005. doi: 10.1136/gpsych-2018-100005. PMID: 31179429; PMCID: PMC6551438.
25. Jacobs HI, Van Boxtel MP, Jolles J, Verhey FR, Uylings HB. Parietal cortex matters in Alzheimer's Disease: an overview of structural, functional and metabolic findings. Neurosci Biobehav Rev. 2012 Jan;36(1):297-309. doi: 10.1016/j.neubiorev.2011.06.009. Epub 2011 Jun 30. PMID: 21741401.
26. Peters F, Collette F, Degueldre C, Sterpenich V, Majerus S, Salmon E. The neural correlates of verbal short-term memory in Alzheimer's Disease: an fMRI study. Brain. 2009 Jul;132(Pt 7):1833-46. doi: 10.1093/brain/awp075. Epub 2009 May 11. PMID: 19433442.
27. Poulin SP, Dautoff R, Morris JC, Barrett LF, Dickerson BC; Alzheimer's Disease Neuroimaging Initiative. Amygdala atrophy is prominent in early Alzheimer's Disease and relates to symptom severity. Psychiatry Res. 2011 Oct 31;194(1):7-13. doi: 10.1016/j.pscychresns.2011.06.014. Epub 2011 Sep 14. PMID: 21920712; PMCID: PMC3185127.
28. Bastos-Leite AJ, van Waesberghe JH, Oen AL, van der Flier WM, Scheltens P, Barkhof F. Hippocampal sulcus width and cavities: comparison between patients with Alzheimer Disease and nondemented elderly subjects. AJNR Am J Neuroradiol. 2006 Nov-Dec;27(10):2141-5. PMID: 17110684; PMCID: PMC7977199.
29. Ohm TG. The dentate gyrus in Alzheimer's Disease. Prog Brain Res. 2007;163:723-40. doi: 10.1016/S0079- 6123(07)63039-8. PMID: 17765747.
30. Scheff SW, Price DA, Schmitt FA, Scheff MA, Mufson EJ. Synaptic loss in the inferior temporal gyrus in mild cognitive impairment and Alzheimer's Disease. J Alzheimer’s Dis. 2011;24(3):547-57. doi: 10.3233/JAD-2011- 101782. PMID: 21297265; PMCID: PMC3098316.
31. Van Hoesen GW, Hyman BT, Damasio AR. Entorhinal cortex pathology in Alzheimer's Disease. Hippocampus. 1991 Jan;1(1):1-8. doi: 10.1002/hipo.450010102. PMID: 1669339.
32. Vasconcelos Lde G, Jackowski AP, Oliveira MO, Flor YM, Souza AA, Bueno OF, Brucki SM. The thickness of posterior cortical areas is related to executive dysfunction in Alzheimer's Disease. Clinics (Sao Paulo). 2014 Jan;69(1):28-37. doi: 10.6061/clinics/2014(01)05. PMID: 24473557; PMCID: PMC3870310.
33. Greene SJ, Killiany RJ; Alzheimer's Disease Neuroimaging Initiative. Subregions of the inferior parietal lobule are affected in the progression to Alzheimer's Disease. Neurobiol Aging. 2010 Aug;31(8):1304-11. doi: 10.1016/j.neurobiolaging.2010.04.026. Epub 2010 Jun 8. PMID: 20570398; PMCID: PMC2907057.
34. Giacomo Koch and others, Precuneus magnetic stimulation for Alzheimer’s Disease: a randomized, sham-controlled trial, Brain, Volume 145, Issue 11, November 2022, Pages 3776–3786, https://doi.org/10.1093/brain/awac285
35. Sadun AA, Bassi CJ. Optic nerve damage in Alzheimer's Disease. Ophthalmology. 1990 Jan;97(1):9-17. doi: 10.1016/s0161-6420(90)32621-0. PMID: 2314849.
36. Miklossy J. Alzheimer's Disease - a neurospirochetosis. Analysis of the evidence following Koch's and Hill's criteria. J Neuroinflammation. 2011 Aug 4;8:90. doi: 10.1186/1742-2094-8-90. PMID: 21816039; PMCID: PMC3171359.
37.

****