Journal of the Neurological Sciences 468 (2025) 123361

Contents lists available at ScienceDirect

Journal of the Neurological Sciences

journal homepage: www.elsevier.com/locate/jns

Explainable machine learning on clinical features to predict and differentiate Alzheimer's progression by sex: Toward a clinician-tailored web interface

Fabio Massimo D'Amore [a], Marco Moscatelli [b], Antonio Malvaso [c], Fabrizia D'Antonio [d,e], Marta Rodini [d], Massimiliano Panigutti [d,e], Pierandrea Mirino [a,f], Giovanni Augusto Carlesimo [d,g], Cecilia Guariglia [d,e], Daniele Caligiore [a,f,*]

[a] *Computational and Translational Neuroscience Laboratory, Institute of Cognitive Sciences and Technologies, National Research Council (CTNLab-ISTC-CNR), Via Gian Domenico Romagnosi 18A, Rome 00196, Italy*
[b] *Research Area Milano 4, National Research Council (CNR - AdRMi4), Via Fratelli Cervi 93, Segrate (MI) 20054, Italy*
[c] *Department of Brain and Behavioral Sciences, IRCCS Mondino Foundation, National Neurological Institute, University of Pavia, Via Mondino 2, Pavia 27100, Italy*
[d] *IRCCS Santa Lucia Foundation, Via Del Fosso di Fiorano, 64, Rome 00143, Italy*
[e] *Department of Psychology, Sapienza University, Piazzale Aldo Moro, 5, Rome 00185, Italy*
[f] *AI2Life s.r.l., Innovative Start-Up, ISTC-CNR Spin-Off, Via Sebino 32, Rome 00199, Italy*
[g] *Department of System Medicine, University of Rome Tor Vergata, Via Montpellier 1, Rome 00133, Italy*

ARTICLE INFO

Keywords:
Early diagnosis
Gender differences in Alzheimer
LIME (local interpretable model-agnostic explanations)
Non-invasive predictors
Personalized medicine
SHAP (SHapley additive explanations)

ABSTRACT

Alzheimer's disease (AD), the most common neurodegenerative disorder world-wide, presents sex-specific differences in its manifestation and progression, necessitating personalized diagnostic approaches. Current procedures are often costly and invasive, lacking consideration of sex-based differences. This study introduces an explainable machine learning (ML) system to predict and differentiate the progression of AD based on sex, using non-invasive, easily collectible predictors such as neuropsychological test scores and sociodemographic data, enabling its application in every day clinical settings. The ML model uses SHapley Additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) to provide clear insights into its decision-making, making complex outcomes easier to interpret. The system includes a user-friendly graphical interface designed in collaboration with clinicians, supporting its integration into medical practice. The study extends the cohort to include healthy and Mild Cognitive Impairment subjects, aiming to support early diagnosis in AD pre-clinical stages. The ML system was trained on a large dataset of 2407 subjects from the ADNI open dataset, enhancing its robustness and applicability. By focusing on sex-specific features and utilizing longitudinal data, the system aims to improve prediction accuracy and early detection of AD, ultimately advancing personalized diagnostic and therapeutic approaches. Key findings highlight the significance of the Mini-Mental State Examination, Rey Auditory Verbal Learning Test, Logical Memory - Delayed Recall, and educational attainment in AD diagnosis and progression, with sex-based disparities. Despite performance metrics based on precision, recall, and weighted F1-score demonstrating model efficacy, future research should address the limitations of relying on a single dataset.

1. Introduction

Alzheimer's disease (AD) is the most common neurodegenerative disorder, marked by a progressive decline in memory, language, and cognitive functions [1,2]. This decline impairs daily activities, with early and middle stages often featuring depression and apathy and later stages showing neurological symptoms like dystonia and tremors [3]. The diagnosis of AD typically relies on analyzing the patient's medical history, conducting clinical tests, neurological examinations, and reviewing brain imaging data. Mild Cognitive Impairment (MCI)

[*] Corresponding author at: Computational and Translational Neuroscience Laboratory, Institute of Cognitive Sciences and Technologies, National Research Council (CTNLab-ISTC-CNR), Via Gian Domenico Romagnosi 18A, Rome 00196, Italy.
E-mail address: daniele.caligiore@istc.cnr.it (D. Caligiore).

https://doi.org/10.1016/j.jns.2024.123361
Received 4 July 2024; Received in revised form 2 October 2024; Accepted 15 December 2024
Available online 19 December 2024
0022-510X/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

currently represents the earliest identifiable stage that may indicate potential progression toward AD [4]. However, statistics show that only 20–40 % of individuals with MCI will transition to AD within three years of diagnosis [5,6]. Researchers are actively exploring numerous promising biomarkers for anticipating the onset of AD, encompassing brain imaging, proteins in cerebrospinal fluid (CSF), blood and urine tests, and genetic risk profiling [7–9]. Accuracy and timing are pivotal considerations in these diagnostic methodologies. While literature suggests biomarker changes align with AD development, no single biomarker adequately forecasts the conversion of MCI patients or healthy individuals to AD with sufficient accuracy and ample lead time before the initial appearance of overt AD symptoms. A critical aspect of current diagnostic approaches is their reliance on costly tools (such as brain imaging) and invasive clinical procedures (like amyloid-PET scans and CSF analysis), often necessitating highly specialized personnel [10,11].

Another significant limitation of current diagnostic methods is their inadequate consideration of sex differences [12,13]. Despite the growing recognition of sex-based differences in AD [14–17], there remains a significant gap in the literature regarding the underlying mechanisms by which sex influences disease manifestation. Understanding these mechanisms is critical for developing personalized, sex-specific approaches to both diagnosis and treatment. Sex-focused AD literature often prioritizes prevalence rates over the analysis of feature importance. Nevertheless, underlying factors may still render a feature significant in predicting AD, regardless of its prevalence rate. The literature often correlates variations in test scores between males and females with their respective significance in predicting the disease likelihood. However, this implicit correlation is not always accurate. Consider a scenario where a low score on a particular test implies an increased risk of AD. In males, a low score on this test does not necessarily indicate the test is a more important marker for the disease than for females. It is plausible that for females, a score slightly higher than what males typically achieve still signifies a heightened risk of developing AD. This discrepancy could arise from various factors, such as different baseline levels of the feature or complex relationships with other unmeasured features. Consequently, the feature assessed by the test may hold significant or even heightened importance for females despite their scores being comparatively higher than those of males.

Recent studies support the use of machine learning (ML) tools in AD research [18–20], noting their potential for *personalized medicine* and better decision-making [21–25]. However, challenges remain in feature selection and clinical application due to the reliance on costly, invasive methods like brain imaging and CSF biomarkers [7,26–31]. To address this, newer ML models focus on *non-invasive*, *easily collected predictors* such as neuropsychological tests, sociodemographic data, and blood biomarkers [19,32,33]. Few ML studies explicitly focus on the sex differences in AD. These works mainly proposed ML systems able to distinguish between healthy and AD patients (classification task) rather than AD progression, do not focus on clinical features, seldom use explainable ML approaches, and never, to the best of our knowledge, propose any clinical-tailored user interface usable in a real-life scenario [34–36].

This article proposes an ML-based system for predicting and differentiating the progression of AD based on sex, marking a crucial initial step toward its utilization by medical professionals. In particular, the system includes several critical added values compared to similar approaches proposed in the literature. Firstly, it pioneers the exploration of sex-specific features in AD by employing an explainable ML approach enhanced by a user interface tailored for clinicians. This approach aims to unveil patterns and relationships within the data, illuminating the distinct features characterizing AD in females and males. Elucidating the sex-specific dimensions of AD will aid in refining symptoms management and elevating the quality of life for patients. The interdisciplinary team behind this article, comprising physicians and researchers actively engaged with patients, has played a crucial role in delineating the critical characteristics of this interface and could facilitate its integration into medical practice. Transitioning from theoretical research to practical implementation remains a widely discussed yet underexplored area in the field [37]. An explainable ML algorithm provides clear, understandable insights into its decision-making process, helping non-experts grasp the rationale behind model decisions and build trust. Utilizing SHapley Additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) advanced analytical techniques, the study clarifies the complex interplay of variables and deepens our understanding of AD multifaceted nature [38].

Secondly, the ML system relies solely on non-invasive and easily detectable clinical features, making it well-suited for everyday medical settings and particularly relevant for addressing sex-specific aspects of AD [39]. Thirdly, the ML system extends the cohort of subjects by considering both healthy individuals and MCI patients drawn from the ADNI open dataset. Previous studies utilizing ML to explore sex differences primarily concentrated on the MCI population [33]. In this regard, our proposed system aims to offer support for early diagnosis in the preclinical stages, particularly in the absence of MCI, a facet overlooked in prior endeavors. Finally, the ML system was trained on a dataset with a significantly higher number of subjects, order of thousands, compared to similar ML approaches, which usually involved only hundreds. This aspect contributes to better performance, increased robustness, and adaptability across real-world applications. In addition, the ML system employs individuals whose diagnostic follow-up was available other than five years after the baseline assessment. Most ML studies focus on identifying biomarkers for early diagnosis within three years of baseline, using neuroimaging, genetic, and clinical data [40–45]. In particular, the system exploits longitudinal data to train two interacting ML models. The first one (Model 0) classifies individuals as either healthy or affected by AD, and the second one (Model 1) predicts, for healthy patients, the onset of AD based on sex within or beyond a five-year timeframe. Longitudinal analysis is essential for gaining a deeper understanding of AD, improving prediction accuracy, enabling early detection, facilitating individualized risk assessment, and monitoring treatment effectiveness over time. Overall, these aspects make the ML system proposed here a clinically translatable early diagnostic tool to predict the conversion to AD of healthy and MCI subjects based on sex and using a low number of cost-effective, fast, and easily collectible predictors.

2. Materials and methods

2.1. ADNI dataset

The paper draws upon data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.[1] Launched in 2003 through a collaborative effort between public and private sectors, led by Principal Investigator Michael W. Weiner, MD, the ADNI project aims to explore the integration of serial magnetic resonance imaging (MRI), positron emission tomography (PET), additional biological markers, and clinical and neuropsychological assessments to comprehensively measure MCI and early-stage AD progression. This study exclusively curated and utilized longitudinal neurocognitive test results and patients' demographic data. Individual patient data were extracted from the ADNI database and organized into a .csv file for efficient preprocessing, followed by processing within a neuroscientifically informed ML pipeline.

2.2. Cohort chosen for the study

11,412 observations were downloaded from September 2005 to January 2023, focusing exclusively on neuropsychological test results. The sample includes 2407 subjects, categorized as healthy individuals (475 females, 353 males), patients with MCI (324 females, 444 males), and those diagnosed with AD (345 females, 465 males). On average,

[1] adni.loni.usc.edu

each patient contributes around 4.74 observations, with a standard deviation of approximately 3.12, indicating the extent of data spread around the mean and illustrating the variability in observation numbers. The highest recorded count in the dataset is 18, signifying the maximum frequency of observations for an individual. Visits typically occur annually for each patient. Each feature was assigned the corresponding name used within the ADNI dataset. Demographic variables such as sex, age, and education were aggregated for each patient-record pairing. The features were selected to align with those commonly used at the authors' affiliated research centers. Below are the features considered to train the ML system.

- Mini-Mental State Examination (MMSE). It is a brief screening test of overall cognitive efficiency, including temporal and spatial orientation, immediate and delayed verbal memory, attention, language, and copying drawings.
- Rey Auditory Verbal Learning Test: immediate (AVTOT) and delayed (AVDEL 30MIN) recall. In this test, which evaluates episodic verbal memory for unstructured material, the examiner reads a fifteen-word list aloud five times. Immediately following each presentation and 15 min after the last one, the participant is required to recall as many words as possible without a time limit and in any order. The immediate recall score consists of the total number of words recalled in the five immediate trials (range 0–75), and the delayed score consists of the number of words recalled after the 15-min delay (range 0–15).
- Logical Memory - Immediate Recall (LIMMTOTAL) and delayed recall (LDELTOTAL). In this test, assessing episodic verbal memory for structured material, the examiner reads aloud a brief prose passage that the subject must repeat in as much detail as possible immediately and after a delay.
- Digit Span Forward (DSPANFOR) and Backward (DISPANBAC). In this test, which examines short-term and working memory for verbal material, the examiner reads aloud strings of digits of increasing length and the subject is required to repeat it in the same order. The span is established as the length of the last list recalled correctly. The same procedure is followed for the backward version (DISPANBAC) with the difference that, in this case, subjects are requested to reproduce the sequence in reversed order.
- Letter Fluency - F (FFLUENCY). In this test, evaluating lexical access and executive functions, subjects are required to generate as many words as possible that begin with the letter "F" within one minute.

2.3. Data pre-processing

The initial data analysis revealed values that fell outside the expected ranges for individual features, necessitating normalization. These anomalies likely resulted from errors during data collection or entry, a common challenge in data science [46]. Approximately 2.61 % of observations required correction. Some features, like Digital Span Backward and Forward, had a high proportion of missing data (around 66 %), while others had fewer gaps (less than 5 %). Data imputation methods, such as k-Nearest Neighbors and multivariate imputation, were applied but offered limited improvement in ML system performance during validation. XGBoost built-in capability to handle missing data effectively maintained model performance (Section 2.4.1).

The dataset provided patients' age at the first visit but lacked this information for follow-up visits. Recalculating patient ages for each subsequent visit was necessary to maintain accurate longitudinal data, which was crucial for training the ML system. Additionally, model-relevant data, such as the number of months between each visit and the onset of AD, had to be computed for patients transitioning to AD during the observation period. All features were normalized within the range [0,1], despite XGBoost not requiring it. This step facilitated the integration of missing data imputation into the pipeline. Normalization also enhanced system performance and training efficiency, especially when feature scales varied significantly.

The time-series problem was reframed as a supervised classification task, aiming to classify a patient after each visit. The model compared single-visit features with an enhanced version that incorporated data from the previous visit (if available), demonstrating improved global performance and robustness. A time-series forecasting model with a one-step time window was also implemented, incorporating the gap between visits (in days) as an input. The model learned to compare the patient's current state with their prior state to make predictions. While planned feature engineering aimed to enhance future models, the current setup required each patient to have at least two visits for classification. This transformation simplified analysis, enhanced interpretability, and likely improved prediction accuracy. By shifting from time-series analysis to classification, the model more effectively handled noisy or irregular data. However, this conversion risked information loss by ignoring long-term temporal patterns, and the uniform treatment of time gaps between visits could have introduced bias. Despite these challenges, classification algorithms offered a wider range of adaptable models and techniques, allowing for greater flexibility compared to traditional time-series methods. The transformation process unfolded as follows:

Step 1: Original Time Series (Patient Visits).

Start with a time series representing the features of patient visits recorded at successive time points.

Step 2: Lag Definition (Lag = 1).

Determine the lag, representing the number of previous time steps used as input to predict future patient visit counts. Setting the lag to 1 considers only the immediately preceding observation (patient visit count) as input.

Step 3: Generating Input-Output Pairs.

At each time point, construct a pair consisting of input and output. The input encompasses the features of patient visits from the previous and current time: Features(t-1) and Features(t).

Step 4: Removal of Rows with Missing Values.

Remove the initial rows of the dataset lacking corresponding lagged values. This adjustment is necessary considering that, with lag 1, there will be an initial row without a lagged value.

Step 5: Dataset Structure.

Ultimately, obtain a structured dataset where each row represents an instance with an input (count of patient visits at the current time) and the output (count of patient visits at the next time point). This structured format facilitates the transformation of the time series problem into a classification task.

For each pair of patient observations, defining the expected output was essential. The ML system aimed to distinguish between patients with Alzheimer's Disease (AD) and healthy individuals (Model 0). Among the healthy subjects, it sought to predict whether they would develop AD within five years or afterward (Model 1). Three classes were established: 0 - AD (converted); 1 - Healthy individuals who will convert within five years; and 2 - Healthy individuals who may convert after five years, or potentially never. The MCI group was included in the Healthy individuals group. The features underwent normalization, which enhances model portability by ensuring consistent scaling across datasets. This reduces sensitivity to feature scale, improves convergence during training, standardizes inputs across environments, and enhances generalization to new data. Consequently, this process leads to more stable and predictable model performance across various scenarios.

The training dataset included 931 feature pairs for class 0, 458 for class 1, and 1638 for class 2. Patients were divided into three sets for training, validation, and testing, using stratification by sex and months to conversion. This same approach was applied during the k-fold validation step. The training set (65 % of subjects) was used to train the ML model, while the validation set (15 %) fine-tuned hyperparameters and monitored performance during training, helping to prevent overfitting. The remaining 20 % formed the test set, used to evaluate the model ability to generalize to unseen data. To tackle class imbalance, class weights were used to adjust the training algorithm according to the varying sizes of each class. The XGBoost model received individual example weights, calculated based on the number of examples in each class, to enhance its performance on the imbalanced dataset during training.

2.4. Explainable ML system to predict and differentiate AD progression by sex

Fig. 1 illustrates the architecture of the explainable ML system developed to predict and differentiate AD progression by sex. It comprises a pipeline consisting of two cascade classifiers: the first (Model 0) detects AD at the current observation (time t), while the second (Model 1), for non-AD observations, predicts AD onset in five years or more (potentially never). The ML system input comprises two consecutive patient observations of a specific clinical feature (CF): the current (t) and the past (t-1). Each model is an optimized XGBoost classifier utilizing a bagging boosting algorithm, which trains multiple decision trees and then combines the results (refer to Section 2.4.1 for further details). The ML system was trained, validated, and tested under two distinct scenarios: using data from females and using data from males.

To build Model 0 and Model 1, we initially explored several ML algorithms, such as RandomForest, XGBoost, and LightGBM, considering specific characteristics of the problem and available data. Primary considerations included model explainability and the algorithm capacity to get good performance. We then conducted a preliminary coarse optimization of hyperparameters for each algorithm, followed by performance comparison using grid search with stratified k-fold cross-validation. Among these algorithms, XGBoost showed the most promise and underwent further hyperparameter tuning.

2.4.1. XGBoost algorithm

XGBoost, short for"eXtreme Gradient Boosting," is a widely used ML algorithm for classification and regression problems. Our study focuses on non-invasive and easily collectible clinical features. XGBoost flexibility in handling diverse types of input data allows us to effectively utilize these non-invasive measures, enhancing the practical applicability of our model in routine medical settings. In addition, the algorithm has a built-in mechanism to handle missing data, which is common in clinical datasets. This feature ensures that it can manage incomplete records without significant loss of accuracy, maintaining the integrity and robustness of our predictions [47]. XGBoost offers several advantages to studying AD progression by sex. Firstly, the algorithm optimizes an objective function that includes both a loss function and a regularization term. For our study, we utilize mean squared error as the loss function, which helps to minimize prediction errors. The regularization component prevents overfitting, ensuring that our model generalizes well to new, unseen data. This balance between accuracy and generalizability is essential for developing a reliable clinical tool.

Secondly, it constructs a robust predictive model through tree boosting, an ensemble method that combines multiple weak learners, typically decision trees. Each tree iteratively trains to correct errors made by previous trees, resulting in a highly accurate and refined model. This iterative boosting process helps to capture complex patterns and relationships within the data, which is crucial for understanding AD progression and sex-specific differences.

Thirdly, XGBoost is renowned for its speed and efficiency, thanks to its ability to handle large datasets and perform parallel processing. This computational efficiency is particularly beneficial given the extensive dataset used in our study, which includes thousands of subjects and longitudinal data over five years. The algorithm scalability allows us to process and analyze this large volume of data quickly and effectively. Finally, the ability of XGBoost to provide insights into feature importance aligns well with our goal of creating an explainable ML model. By identifying which clinical features most influence AD progression, especially across different sexes, we can offer clear and actionable insights to clinicians. This interpretability fosters trust and facilitates the integration of our ML system into medical practice.

2.4.2. Model optimization and evaluation

All tests were developed in Python using Scikit-learn and Keras as the main libraries. Hyperparameter optimization fine-tunes the settings of a machine learning model to enhance its performance. These hyperparameters influence the behavior of the learning algorithm, affecting aspects such as model complexity, learning rate, and regularization strength. In this case, we use Grid Search, a hyperparameter optimization technique that systematically explores a predefined set of hyperparameter values to find the combination that yields the best model performance. We searched for canonical hyperparameters such as regularization, learning rate, sample parameters, and tree-related parameters like maximum depth and minimum child weight. The Grid Search algorithm was used to identify the optimal hyperparameters with the following parameter grid: *subsample*: [0.80, 0.82, …, 1.0]; *reg_lambda*: [0.0, 0.5, …, 5.0]; *reg_alpha*: [0.0, 0.5, …, 5.0]; *min_child_weight*: [0, 1, …, 10]; *max_depth*: [0, 1, …, 10]; *learning_rate*: [0.0, 0.1, …, 1.0]; *gamma*: [0.0, 0.2, …, 3.0]; *early_stopping_rounds*: [10, 12, …, 40].

The process involved cross-validation using a validation set, which also helped mitigate overfitting by limiting the number of trees. Precision, recall, and the weighted F1-score were the primary evaluation metrics. The weighted F1-score calculates the F1-score for each class and averages them based on the number of instances per class, addressing class imbalance and providing a more accurate assessment. The *learning rate* was identified as the most critical optimization parameter, with a reverse correlation, indicating its role in reducing overfitting and stabilizing performance. The parameters *gamma* and max *depth* also showed reverse correlations, reflecting efforts to simplify the model. The parameters *gamma*, *reg_lambda*, and *reg_alpha* were less significant but generally showed direct correlations.

Running the entire pipeline in both scenarios (females and males) for the tasks performed by Model 0 and Model 1, we obtained four optimized models, each with specific hyperparameters. We conducted optimization over a subset of hyperparameters within a range found iteratively through recursion. For each task in every scenario, we compared 200 models, selecting the very best. Table 1 summarizes the performance of the Model 0 and Model 1 on the test set in each scenario.

2.4.3. Features importance and statistical analysis

We employed SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) ML techniques to assess the importance of features within both Model 0 and Model 1. SHAP values offer insights into the model output by attributing the contribution of each feature to the prediction. These values, rooted in game theory, assign an importance value to each feature, quantifying its impact on the model output relative to the average prediction [48]. LIME provides local interpretability by approximating complex models with simpler ones, offering transparent explanations for individual predictions, which can be particularly useful for understanding model behavior on specific instances [49].

Table 1
Performance of the ML system by sex. This table presents the performance of two models across different scenarios (females and males), evaluated on the test dataset. Model 0: classifies individuals as either healthy or having Alzheimer's disease; Model 1: predicts whether Alzheimer's disease will onset within five years or after five years; F1-Score: a metric combining precision and recall into a single score, weighted by class distribution; Precision (P): the proportion of true positive predictions among all positive predictions (true positives divided by the sum of true positives and false positives); Recall (R): the ability to correctly identify positive cases (true positives divided by the sum of true positives and false negatives).

	Model 0			Model 1		
	F1-score	P	R	F1-score	P	R
Females	0.92	0.92	0.97	0.93	0.92	0.95
Males	0.88	0.89	0.93	0.89	0.92	0.88

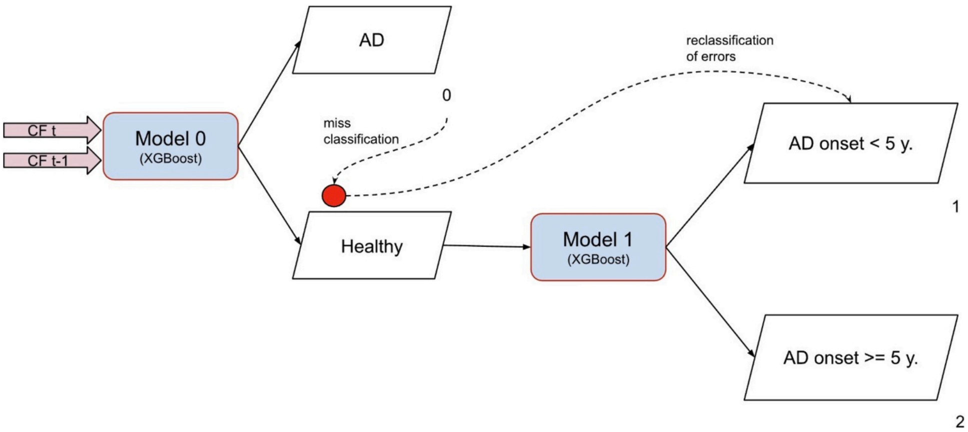


Fig. 1. ML system architecture. The outputs of the two models are the classes (0,1,2). To increase overall robustness, we trained Model 1 to classify AD patients as "AD onset < five years" to minimize the impact of any errors in Model 0 (see the dotted line). Training the second model with more data variability enhances its robustness. In this setup, Model 0 might produce an incorrect prediction. When this happens, Model 1 receives this data for further evaluation. Given that an incorrect prediction by Model 0 would likely relate to an AD case, we trained Model 1 to recognize these situations. Therefore, during its training phase, Model 1 learns to classify AD as 'AD onset less than five years'. CF: current clinical feature; y: years.

These techniques enhance the interpretation and understanding of ML models applied to clinical and biomedical data. They can help identify which patient characteristics are most influential in model predictions, providing valuable insights for clinical practice. For example, in disease prediction models, SHAP values can identify significant risk factors, potentially aiding in patient risk assessment and treatment planning. LIME complements this by offering local interpretability, enabling clinicians to understand the rationale behind individual predictions, which can be crucial for personalized patient care.

We used the Mann-Whitney U statistical test to compare SHAP values between males (M) and females (F). We set the significance level at 0.05. Each test included the computation of 95 % confidence intervals for the sampled differences between the datasets, achieved through bootstrapping. Notably, the Mann-Whitney test does not require equal sample sizes. One of its assumptions is that the instances or records in the dataset are independent. Since our dataset is anonymous, directly verifying this assumption is not feasible, and any patient relationships remain unknown. Nevertheless, we implemented precautions, such as using a single record for each patient and partitioning the data based on patient identifiers (ID).

2.5. Software used to develop the explainable ML web interface

The web interface, called EMA (ExplAIn Medical Analysis),[2] was developed to make the AI techniques used in our ML system explainable and accessible for diagnostic purposes. The interface was built using the Python programming language and the Dash framework to create interactive web applications. Dash is an open-source framework for building interactive web applications in Python. Developed by Plotly, Dash allows users to create web applications with interactive graphs and user controls without writing JavaScript code. Dash is built on Flask, Plotly.js, and React, combining the power of these libraries to create dynamic and responsive web applications. Dash offers a simple and intuitive syntax for designing layouts and components, making it easy for users to build complex user interfaces. It is useful for creating data visualization applications, enabling the creation of interactive graphs and dashboards with just a few lines of code. The Plotly library, included in Dash, supports the development of responsive and reactive web applications with interactive and animated graphs, providing an efficient and user-friendly experience. Plotly supports several chart types, including line charts, bar charts, scatter plots, and more, which are highly customizable and interactive. These features make Plotly a powerful tool for data visualization, allowing users to explore and analyze data intuitively and engagingly. To ensure portability and scalability, the application was containerized in a Docker environment. This environment includes a NGINX server that handles communication encryption for secure data transmission and a Gunicorn WSGI HTTP server that efficiently processes web requests, allowing smooth interaction with the application.

The Results section offers an in-depth look at the structure and functionality of the EMA web interface (Section 3.2). It will summarize the key findings of the ML analysis, emphasizing the differences in critical features for predicting AD onset and progression between sexes. The section will also explain how these features influence model predictions and detail the methods used to assess their contributions. Furthermore, it will demonstrate how the EMA web interface can provide clinicians with practical insights for predicting AD progression based on sex.

3. Results

3.1. Key features to predict and differentiate AD progression in females and males

Fig. 2 presents two plots that compare the most important features to predict AD onset (Model 0), as determined by SHAP values obtained from the ML algorithm, separately for females and males. We can interpret the sign of SHAP values concerning the target variable (AD or healthy controls): SHAP values less than zero lean toward AD outcome, whereas SHAP values greater than zero lean toward a healthy controls outcome.

Similarly, Fig. 3 shows two plots that compare the most important features to predict AD progression (Model 1), as determined by SHAP values obtained from the ML algorithm, separately for females and males. In this case, negative SHAP values indicate conversion to AD within five years, while positive indicate conversion to AD over five years.

For both Figs. 2 and 3, each plot illustrates the SHAP values for the features utilized in XGBoost model training across all examples (records) sampled from the test set. These values indicate the extent to which a feature contributes to altering the XGBoost prediction in comparison to the expected value (refer to Section 2.4.1). The SHAP values associated with certain features appear to be spread across the x-axis, while those linked to other features show less dispersion. In the latter scenario, these features not only prove to be significant but also provide a robust

[2] https://ctnlab.it/index.php/explain-medical-analysis-ema/

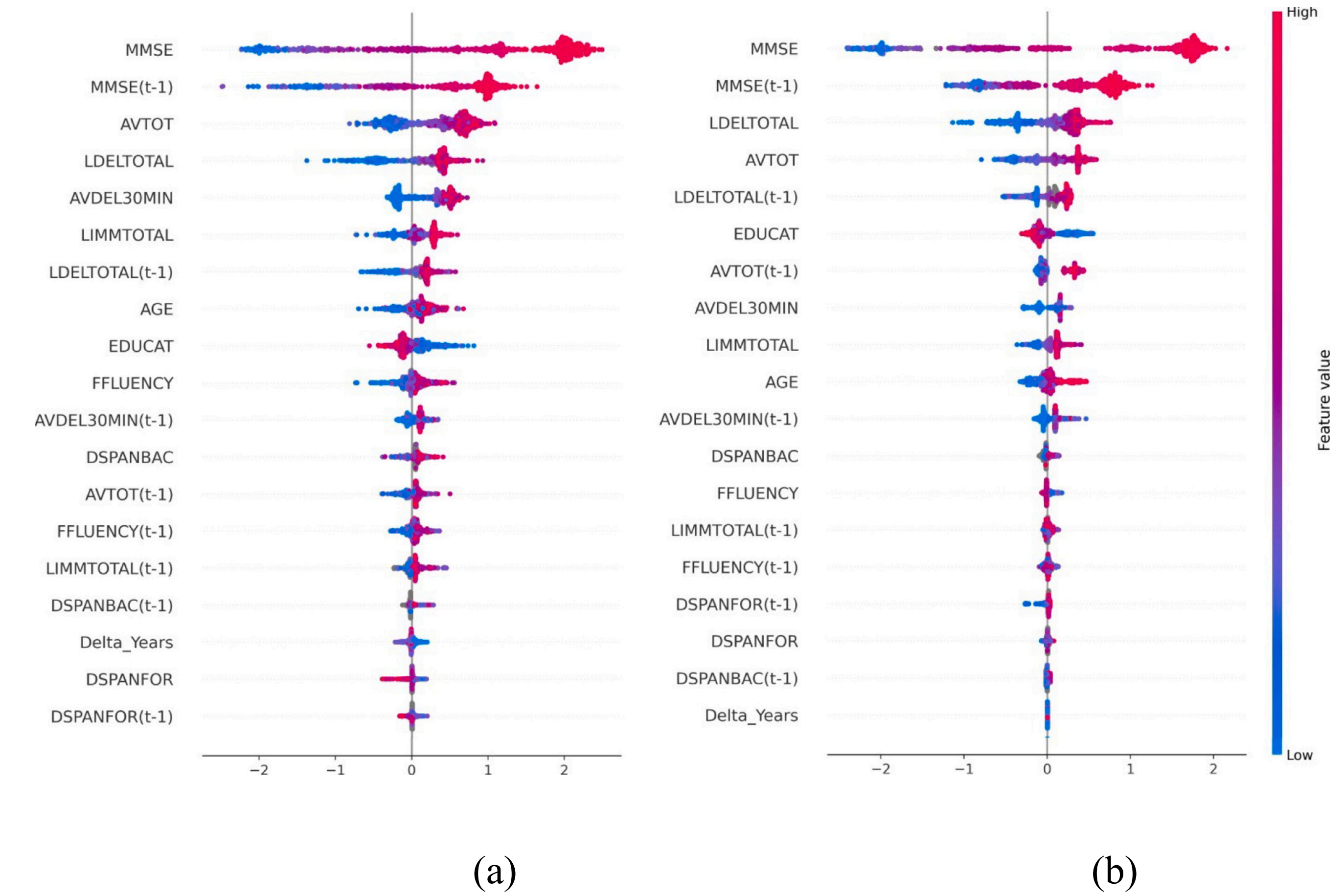


Fig. 2. SHAP beeswarm plot of Model 0 (AD/healthy) from Females (a) and Males (b) individuals. Negative values indicate AD, while positive values indicate healthy controls. Each data point corresponds to a specific record within the test set. The vertical axis represents training features, sorted by their respective importance. Feature names remain consistent with their original source from the ADNI dataset. On the horizontal axis, each point signifies an individual record and expresses the SHAP value attributed to that particular feature during prediction (AD or healthy controls). The color of each point reflects the true value of the feature in the corresponding record used during prediction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

indication of the likelihood for females or males to belong to either a healthy controls/AD group (Model 0) or within/over five years AD onset (Model 1) with greater accuracy compared to instances with wide dispersion of SHAP values.

Fig. 4 provides a concise representation of the most influential features identified through SHAP analysis for both females and males. The SHAP algorithm assigns importance scores to each feature for every individual observation, representing the significance of each feature in the models decision-making process. For each model, to combine the importance of each feature across two different time points, t-1 and t, we computed the average SHAP value for each feature using the recorded values from both times, and then we expressed the result in absolute terms. Next, we calculated the average of these values across all features. Tables 2 and 3 summarize the results of the Mann-Whitney test for statistically comparing SHAP values between females and males in both models.

In addition to the SHAP approach, which analyzes the importance of global features, we also investigated the importance of local features using LIME. This local characteristics analysis examines the responses of Models 0 and Model 1 to individual data points. For example, Fig. 5 presents the results of this analysis for data from a female subject who is defined as healthy by Model 0 and AD onset >5y by Model 1. Figs. 5 show the Model 0 confidence interval of the classification, which is 8 % as AD and 92 % as healthy, reflecting the subject's healthy status with the importance values of the features for the specific sample, with MMSE having an importance of 20 % in the”healthy” classification, followed by LDELTOTAL at 12 %. For the Model 1 the confidence interval is 10 % for AD onset <5y and 90 % for AD onset >5y; the major contributor is given by MMSE with 12 %. The last panel shows the value of the features measured in the sample subject. Features highlighted in orange contribute to class 1 (healthy), while features highlighted in blue contribute to class 0 (AD). A similar figure is generated for Model 1 for the two classifications, AD onset <5 years and AD onset >5 years. Appendix A provides additional examples of local explainers provided by LIME. Specifically, explainers are reported for male and female subjects who represent all classes of Models 0 (healthy and AD) and Model 1 (AD onset <5 years and AD onset >5 years).

Fig. 6 illustrates the distribution of LIME analysis results for females and males, comparing Model 0 and Model 1. The analysis uses absolute values to underline how each feature contributes to the performance of the models.

3.2. Explainable ML clinician-tailored user interface

Fig. 7 displays the interface we developed, named EMA (ExplAIn Medical Analysis). EMA is a web application based on two explainable machine learning models, Model 0 and Model 1, to provide insights into AD progression stratified by sex. It comprises five tabs. The first one (Description) gives an overview of the interface. The second tab (Model Description) presents an overview of the ML system (i.e., Model 0 and

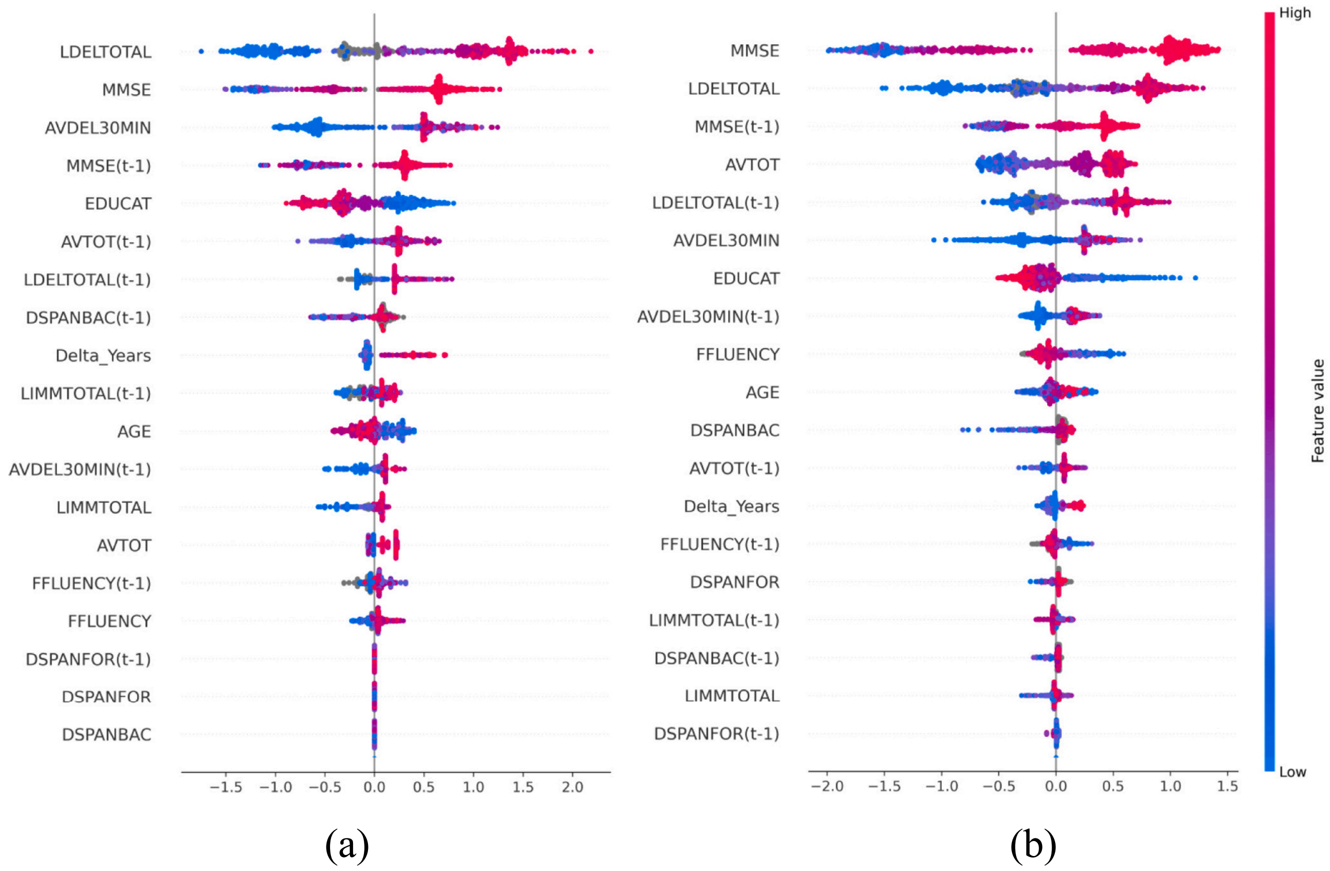
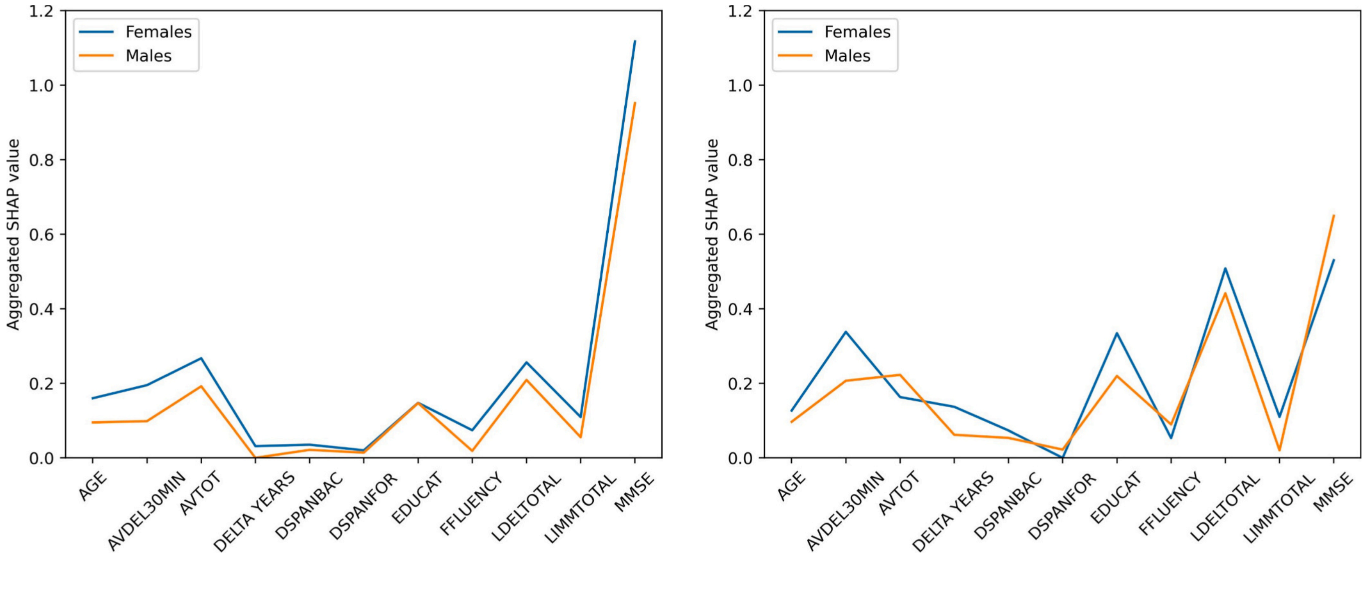


Fig. 3. SHAP beeswarm plot of Model 1 (within/over five years onset) from Females (a) and Males (b) individuals. Negative values indicate conversion to AD within five years, while positive values indicate conversion over five years. The data point description is the same as Fig. 2.



(a) Model 0 (AD/Healthy)

(b) Model 1 (within/over five years AD onset)

Fig. 4. Aggregated SHAP values magnitude for Model 0 (a) and Model 1 (b).

Table 2
Mann-Whitney *U* test comparing SHAP value distributions between females (F) and males (M) in Model 0 for the most important features. A $p < 0.05$ (bold) indicates a statistically significant difference. h_1 represents the alternative hypothesis.

	mean		std.		median			p
Feature	F	M	F	M	F	h1	M	
MMSE	1.118	0.952	0.543	0.431	1.343	>	1.113	**0.000**
AVTOT	0.267	0.192	0.128	0.124	0.273	>	0.188	**0.000**
AVDEL30MIN	0.195	0.098	0.105	0.044	0.186	>	0.104	**0.000**
LDELTOTAL	0.256	0.209	0.129	0.107	0.275	>	0.219	**0.000**

Table 3
Mann-Whitney U test comparing SHAP value distributions between females (F) and males (M) in Model 1 for the most important features. A p < 0.05 (bold) indicates a statistically significant difference. h_1 represents the alternative hypothesis.

	mean		std.		median			p
Feature	F	M	F	M	F	h1	M	
MMSE	0.530	0.649	0.242	0.298	0.486	<	0.693	**0.000**
AVTOT	0.163	0.223	0.080	0.092	0.165	<	0.239	**0.000**
AVDEL30MIN	0.338	0.207	0.114	0.098	0.332	>	0.203	**0.000**
LDELTOTAL	0.508	0.441	0.265	0.246	0.579	>	0.440	**0.000**
EDUCAT	0.334	0.220	0.186	0.180	0.309	>	0.181	**0.000**

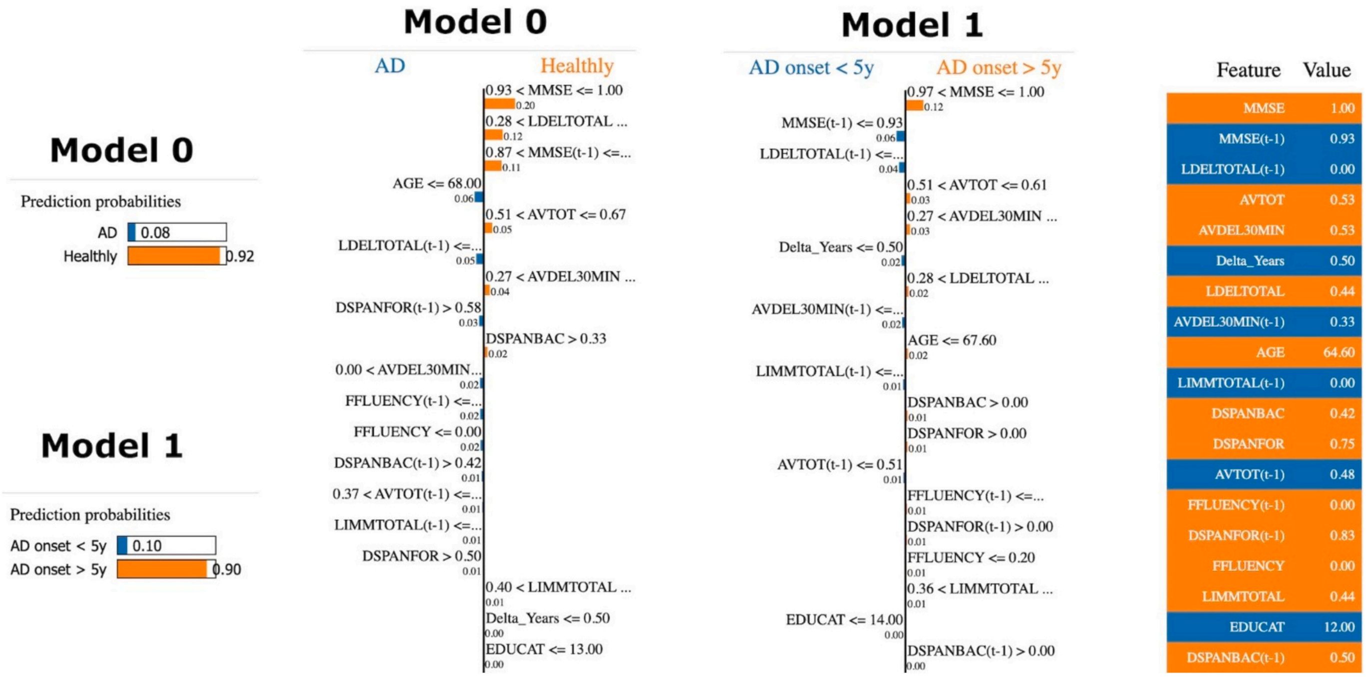


Fig. 5. LIME explainer for a female subject in Model 0 and Model 1.

Model 1) and explains its functionality. The subsequent two tabs (SHAP Evaluation and LIME Evaluation) illustrate feature importance through the SHAP and LIME methodologies, enabling users to grasp the most relevant features. Lastly, the section about the dataset under analysis (Value Prediction tab) involves employing trained models to predict new observations. The depicted screen in the figure first displays a form comprising two empty tables corresponding to observations at time t and (t-1), along with a section showing the model graphical outputs.

The users can manually enter the data. In this case, they need to select the patient sex and each cell of the table to input the values. If the users want to compare multiple observations with the same previous observation (t-1), they can use the "+" function below the second table, which allows them to add new rows. This function enables testing different values for the critical features, helping to find those with greater weight in predicting the onset of the disease and its temporal definition.

The "Load/Generate Data" button supports data insertion and loading processes. When clicked, a modal window opens, allowing the users to load existing data or generate new data. Data loading involves selecting the desired diagnosis category: patient, onset within five years, or onset after five years. The users could insert values from a randomly selected record in the ADNI dataset corresponding to the chosen diagnosis category.

The second section of the modal window provides the option to generate random values. The user can specify the number of records to add and which attributes to modify. It is important to note that the ability to change specific features depends on existing observations in the table and requires activating the function to retain previously inserted values before adding new ones.

After entering the values, the "Predict" button applies the ML system

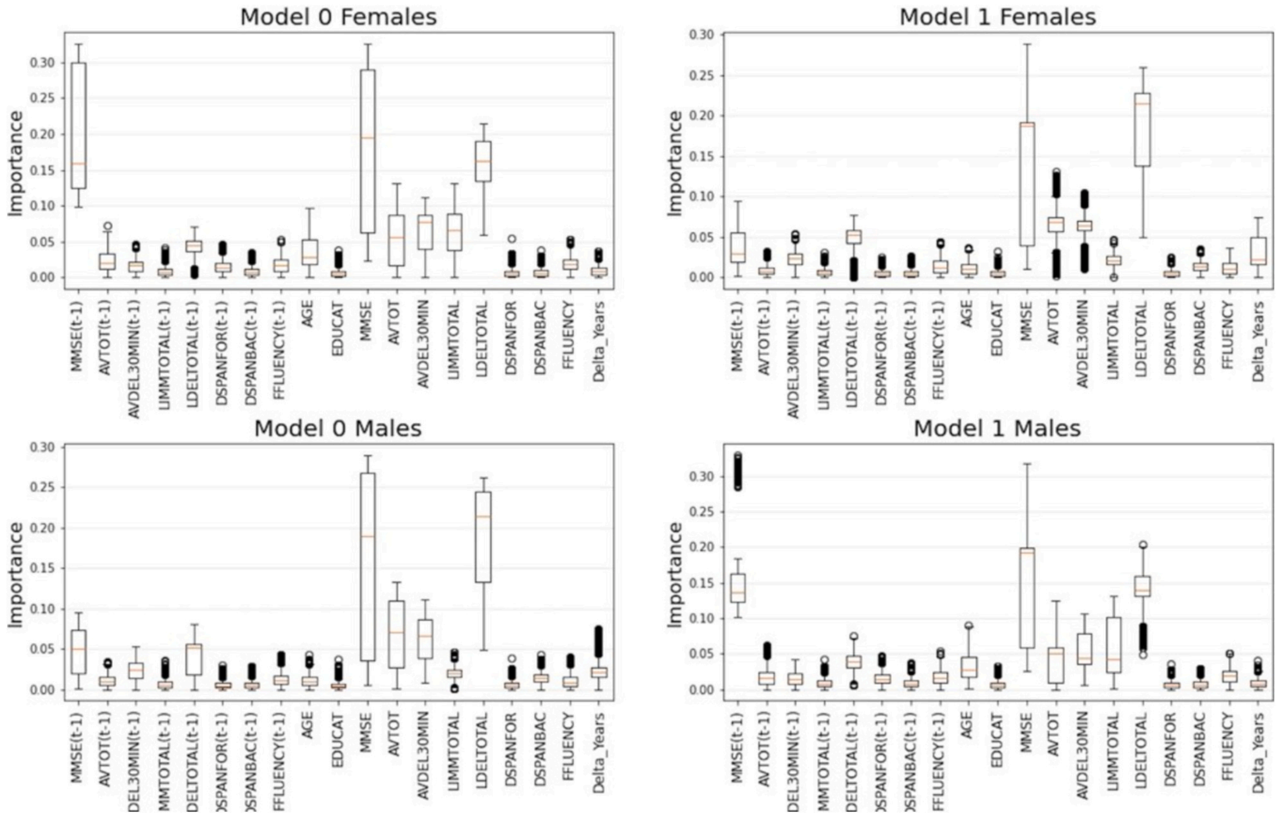
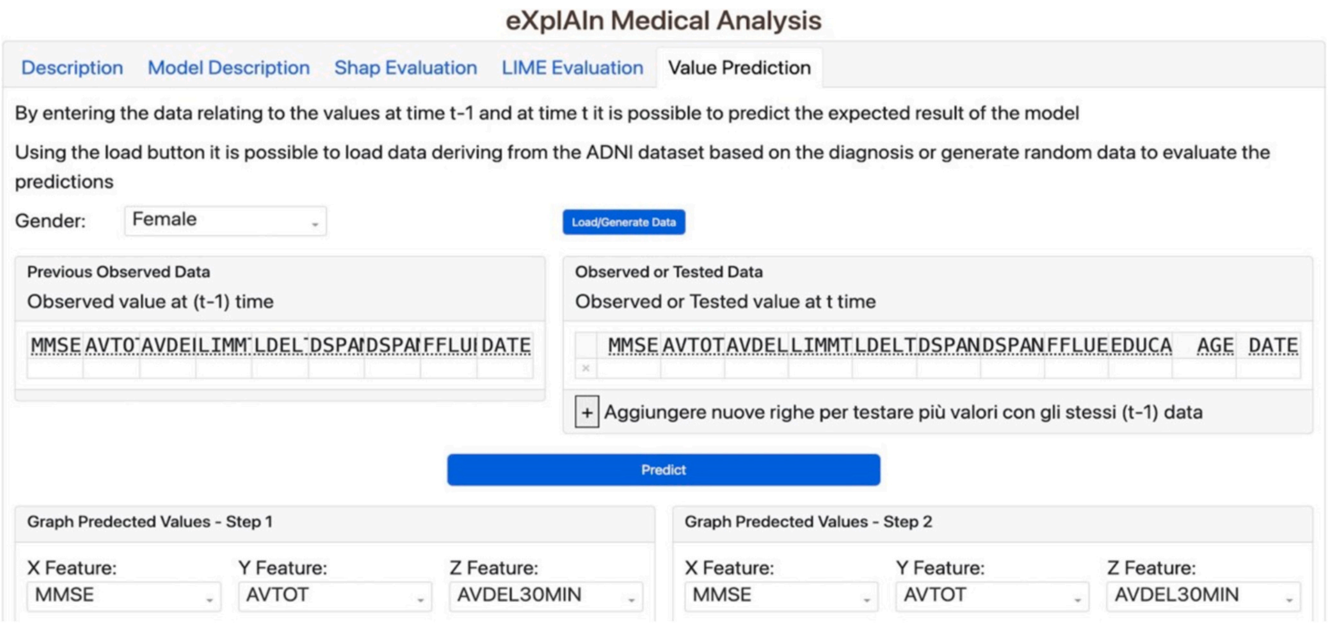


Fig. 6. Distribution of LIME analysis results.



eXplAIn Medical Analysis

By entering the data relating to the values at time t-1 and at time t it is possible to predict the expected result of the model

Using the load button it is possible to load data deriving from the ADNI dataset based on the diagnosis or generate random data to evaluate the predictions

Gender: Female

Previous Observed Data

Observed value at (t-1) time

MMSE	AVTO	AVDEI	LIMM	LDEL	DSPAI	DSPAI	FFLUI	DATE

Observed or Tested Data

Observed or Tested value at t time

MMSE	AVTOT	AVDEL	LIMMT	LDELT	DSPAN	DSPAN	FFLUE	EDUCA	AGE	DATE

Aggiungere nuove righe per testare più valori con gli stessi (t-1) data

Graph Predected Values - Step 1

X Feature: MMSE
Y Feature: AVTOT
Z Feature: AVDEL30MIN

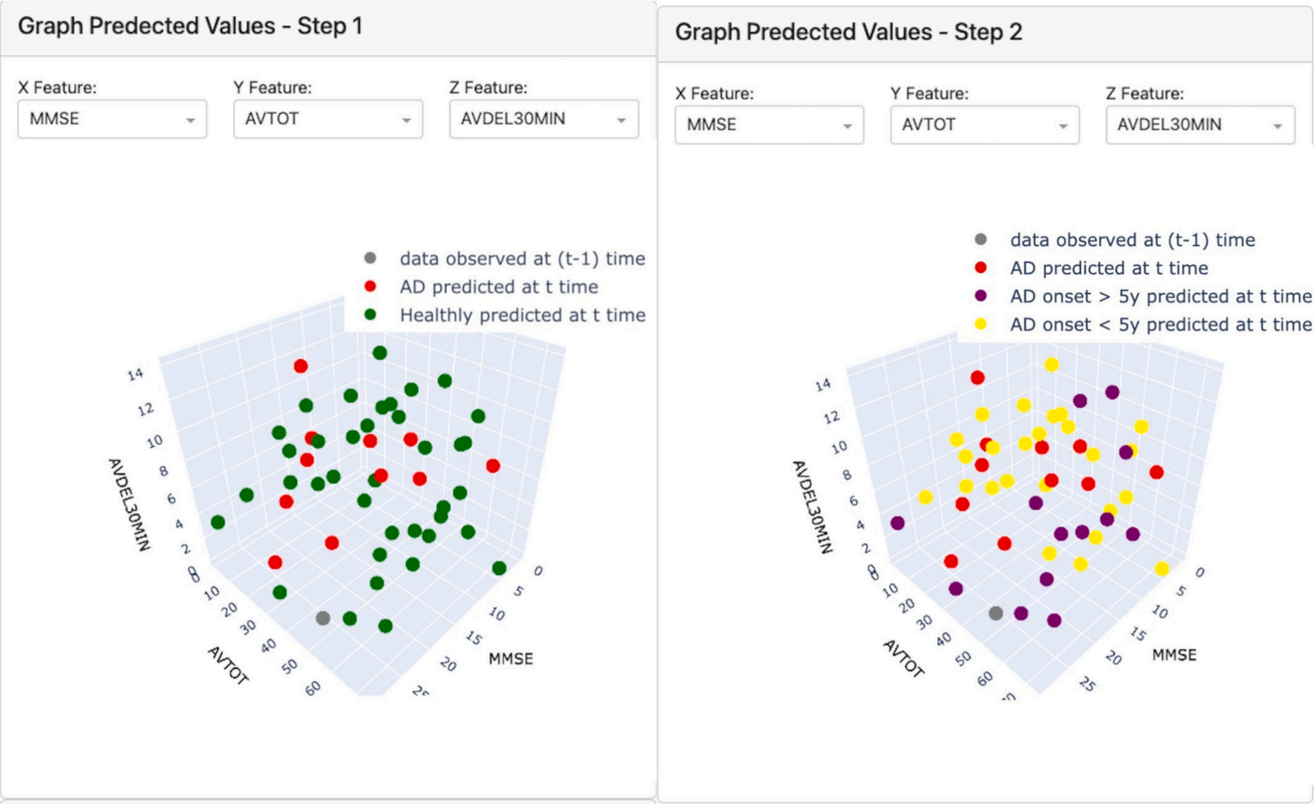
Graph Predected Values - Step 2

X Feature: MMSE
Y Feature: AVTOT
Z Feature: AVDEL30MIN

Fig. 7. The EMA web user interface. The figure illustrates the section relating to the prediction and observation of the model results (Value Prediction tab).

to all rows of the observation table (t) relative to the preceding observation (t-1). Following this, graphs are generated for both stages of the pipeline if the initial prediction classifies as "healthy"; otherwise, the interface shows only the first graph.

Fig. 8 illustrates the predictions for various records, displaying the two analysis steps of the model. Specifically, the image on the right showcases all the output classifications of the model pipeline. By selecting a marker using the mouse, the users can identify the record that generated the prediction and experiment with modifying different features. In this way, the users could assess the significance of individual features relative to the observation at the previous time (t-1). The users could interact with the graphs, rotating and zooming to better visualize and analyze different records. Additionally, the users could adjust the characteristics on the three axes (x, y, z) to observe the marker movement compared to the previous reference (t-1).



Graph Predected Values - Step 1

X Feature: MMSE

Y Feature: AVTOT

Z Feature: AVDEL30MIN

Graph Predected Values - Step 2

X Feature: MMSE

Y Feature: AVTOT

Z Feature: AVDEL30MIN

● data observed at (t-1) time
● AD predicted at t time
● Healthy predicted at t time
AVDEL30MIN
14
12
10
8
6
4
2
0
AVTOT
10
20
30
40
50
60
MMSE
0
5
10
15
20
● data observed at (t-1) time
● AD predicted at t time
● AD onset > 5y predicted at t time
● AD onset < 5y predicted at t time
AVDEL30MIN
14
12
10
8
6
4
2
0
AVTOT
10
20
30
40
50
60
MMSE
0
5
10
15
20
25

Fig. 8. The figure shows the result of the random generation of 50 records in which it is possible to observe the prediction of the models used.

4. Discussion

4.1. Model results framed within the literature on sex difference in AD

Recent evidence suggests crucial sex-based differences in AD manifestation and progression, highlighting the necessity for personalized, sex-specific approaches to both diagnosis and treatment [12–14]. Biological explanations for sex differences in AD manifestation include hormonal influences, genetic susceptibility, and differences in brain structure and function [15,16]. For example, sex hormones and sex chromosomes interact with various disease mechanisms during aging, encompassing inflammation, metabolism, and autophagy, leading to unique characteristics in disease progression between males and females [16]. Multimodal brain imaging indicates sex differences in the development of the AD endophenotype, suggesting that the preclinical AD phase is early in the female aging process and coincides with the endocrine transition of perimenopause. These data indicate that the optimal window of opportunity for therapeutic intervention in females is early in the endocrine aging process [17].

Several works support the use of ML tools in AD research and clinical practice to provide predictions with a certain degree of confidence, pivoting on information about the specific person [18–20]. These *personalized medicine* approaches support improved and more effective decision making by researchers and clinicians [21–24]. Although some of these models could reach high levels of accuracy [25], consistency regarding what combination of features is more informative to predict AD as well as the translation into clinical practice are still lacking. One possible reason for this is that current ML algorithms still generally rely on expensive and invasive predictors, such as brain imaging or CSF biomarkers [7,23,24,26–31]. As such, these studies only serve the purpose of proof of concept but do not represent a viable substitute of standard approaches with which they share application complexities and economic costs. To overcome these limitations, recent works proposed ML algorithms elaborating only *non-invasive* and *easy-to-collect* predictors (e.g., neuropsychological test scores, sociodemographic and clinical features, blood biomarkers) [19,32,33].

Few ML studies focus on the sex differences in AD [34–36]. Klingenberg and colleagues compare the performance of a not-explainable ML system in classifying healthy versus AD individuals stratified by sex. They do not conduct a feature importance analysis [35]. Sarica and colleagues confirm that females had a higher risk of progressing to dementia. Influential variables across the sexes included brain glucose metabolism and CSF biomarkers. Hippocampus volume is critical only to predict male conversion, while verbal memory and executive function are key contributors only in predicting female conversion [36]. The ML analysis proposed by Cieri and colleagues shows that sex differences in memory and cortical thickness can play a key role in the different vulnerability and progression of AD in females compared to males [34]. Despite some efforts utilizing ML to investigate sex differences in AD, there remains a significant gap in research addressing the most important clinical features for characterizing AD in females and males.

The primary focus of these studies lies in ML systems capable of discerning between healthy individuals and those with AD, primarily for classification purposes. There is a notable absence of emphasis on tracking AD progression or incorporating clinical features into the models. Furthermore, these works have utilized explainable ML approaches sparingly, and there have been no attempts to craft a clinical-focused user interface that enhances the models interpretability for real-world implementation.

The explainable ML system proposed here addresses these points. It only relies on non-invasive and readily detectable clinical features to support its use in everyday medical settings. Non-invasive tests are more accessible, comfortable, and safer than invasive procedures, reducing patient anxiety and increasing participation in screenings. They are cost-effective, quicker, and require less preparation, enhancing efficiency in medical settings. In this way, they are also well-suited for application in countries with limited infrastructure and developing economies. Moreover, non-invasive features enable ongoing monitoring of health status and response to treatment, aiding in the management of chronic conditions and early detection of complications, particularly significant in addressing sex-specific aspects of AD [39].

In addition, the ML system proposed here incorporates a user-friendly graphical interface co-designed with end-users, particularly clinicians, streamlining its assimilation into medical routines. In this way, the ML system forecasts and distinguishes AD progression based on sex, representing a pivotal initial stride toward its adoption by healthcare practitioners. In the literature discussing sex differences in AD, the focus often lies on variations in test scores, neglecting to explicitly delve into an analysis of feature importance as discussed in this article. Here, we highlight the importance of differentiating between scoring disparities and the significance of features in predicting AD. For example, if research indicates that a particular feature consistently yields higher scores in females compared to males, and higher scores are associated with a greater likelihood of developing AD. This finding does not diminish the feature importance in predicting AD in males; it merely underscores its lower prevalence within this group. Various factors may hide the importance of a feature for a particular sex. For example, the necessity for different scoring scales may arise for males and females undertaking the same test. Our explainable ML analysis through Model 0 and Model 1 allows us to clarify these underlying factors. This insight holds significant promise for personalized medicine, indicating the necessity to tailor data collection methods for males and females to optimize diagnostic and therapeutic outcomes.

Fig. 2 shows that the most important feature to detect non-AD vs AD subjects (Model 0) is MMSE for both sexes. The ML analysis also identifies the tests to evaluate verbal memory and learning capacity (AVTOT, AVDEL30MIN) and logical memory (LDELTOTAL), as highly significant to detect not-AD vs AD subjects in both females and males, albeit with some differences (see Fig. 2). For all these features the SHAP value for females is significantly higher than that for males (Fig. 4a and Table 2), suggesting that they may be more predictive factors in determining the risk of AD in females than in males. MMSE is still an important feature for sex-based AD onset prediction within or over five years (Model 1) for both sexes (Fig. 3), even though in this case the SHAP value for males is significantly higher than that for females (Fig. 4b and Table 3), suggesting that the MMSE may be a more predictive factor in males than in females. By contrast, for this latter group, the most important feature to predict AD onset within or over five years is LDELTOTAL (Fig. 3). AVTOT is more important for males compared to females. The LIME analysis largely confirms these results for MMSE and LDELTOTAL (Fig. 6).

These results agree with literature showing that females have lower MMSE scores (indicating worse global cognitive status) than males at initial diagnosis of AD (after adjustment for age and education level) [13,50]. Previous studies have also observed that females had significantly worse cognitive function scores than males in the areas of episodic memory, semantic memory, working memory, perceptual speed, and visuospatial ability [51–54]. However, these findings do not diminish the importance of MMSE in predicting AD in males; they merely underscore its higher prevalence in females. The existing literature does not explicitly investigate whether MMSE significantly contributes to detect not-AD vs AD subjects when stratified by sex. The Model 0 addresses this gap. It suggests that MMSE is a significant predictor also for males (Fig. 2), despite the lower SHAP values compared to females (Fig. 4a and Table 2). In addition, Model 1 suggests that the MMSE could be an important predictor for monitoring AD progression in males.

Finally, the ML analysis suggests that for both females and males, the level of education (EDUCAT) is important to detect not-AD vs AD subjects (Model 0) or AD onset within or over five years (Model 1), but with quite differences. It is more important for females in the Model 1 (Fig. 4). These findings align with existing literature indicating historical constraints on females access to cognitive reserve, thereby heightening their risk of AD, albeit mitigated by educational attainment. Indeed, the elevated incidence of AD in females could be due to disparities in reserve levels between genders [55–57]. More specifically, previous studies have shown that females with lower and moderate levels of education, compared to those with a higher level of education, had 4.3 (95 % confidence interval: 1.5, 11.9) and 2.6 (95 % confidence interval: 1.0, 7.1) times higher risks for AD, respectively [58]. Nevertheless, alternative studies have suggested that females continue to experience a heightened risk of AD even after adjusting for educational level, calling into question the notion that lower educational attainment alone explains the increased risk of AD in females [59]. Consequently, further research is warranted to elucidate this matter.

4.2. Integrating sex-specific diagnostic approaches into AD clinical practice

The model results reveal that certain features, such as MMSE and memory tests, have differing predictive power for AD in males and females. Specifically, the higher SHAP values for females suggest that MMSE may be a stronger indicator of AD risk for women, while it remains significant for long-term monitoring in men. Similarly, tests like AVTOT and LDELTOTAL show varying levels of importance, underscoring the need for sex-specific memory assessments to accurately evaluate AD risk. Additionally, the differential impact of education on AD risk between sexes highlights the necessity for tailored risk assessments and treatment plans based on educational background.

These results underline notable distinctions in AD detection and progression between males and females, emphasizing the requirement for customized diagnostic strategies. Raising awareness among healthcare professionals about these differences is the first step to incorporating sex-specific diagnostic approaches for AD into current clinical practice, enhancing patient care, and advancing personalized medicine in neurology.

Updating diagnostic guidelines is another crucial step. Collaborating with medical associations to incorporate sex-specific neuropsychological tests into these guidelines ensures that healthcare providers have clear directions on which assessments to prioritize and how to interpret their results in the context of sex differences. Tailoring treatment plans based on sex-specific neuropsychological test results is essential for optimizing patient care. For instance, if a test indicates a higher risk or different progression pattern in males, treatment plans could be adjusted accordingly.

Enhancing patient education is also crucial. Educating patients about the significance of sex-specific diagnostic tools empowers them to advocate for comprehensive assessments and personalized care based on their characteristics, ultimately improving their treatment outcomes. Furthermore, fostering ongoing research and development in this area is vital. Exploring and refining sex-specific neuropsychological tests and biomarkers for AD will contribute to improving diagnostic accuracy and informing tailored treatment approaches in the future.

4.3. The importance of a “clinician-in-the-loop” approach

Although numerous explainable ML algorithms exist in the medical field, few efforts extend beyond foundational research to create software and graphical interfaces that help healthcare professionals utilize the algorithm results [60]. ML researchers often rely on their intuition to determine what makes a good explanation of the ML analysis results, neglecting validation from medical professionals [61]. To maximize stakeholder benefit, the concurrent involvement of medical and AI experts is imperative in refining interpretability within the explainable ML framework. Effectively, research in this domain inadequately adheres to the “human-in-the-loop” principle, which needs explicit consideration of end-users, such as healthcare professionals in our scenario, for evaluating and developing explainable systems. This approach represents a significant limitation in current research endeavors [62].

This article tackles this issue, leveraging the expertise of healthcare professionals, particularly neurologists, working in medical and clinical domains with AD patients. Their insights were pivotal in interpreting model results within the AD context and designing the EMA web interface to enhance model explainability, catering to experts in AD. The EMA application enhances explainability by providing graphical tools to evaluate feature importance using results from SHAP and LIME algorithms. It also displays predictions based on patients' data for AD detection and progression through an interactive 3D graph. The 3D visualization allows end users to observe deviations from initial values and assess the patients' status regarding AD detection and progression through chromatic identification. EMA web application marks a first step toward the support of AI systems in precision medicine, enabling clinicians to investigate the "black box" and assist in data interpretation. Subsequent efforts will involve disseminating this software among medical staff and soliciting feedback to refine user experience and explainability.

5. Conclusions

This article proposes an explainable ML system designed to predict and differentiate the progression of AD by sex. As discussed in Sections 1 and 4, the ML system demonstrates several critical advantages over existing methods. It represents a significant advancement toward personalized diagnostic approaches, with a user-friendly graphical interface developed in collaboration with clinicians to ensure its practical application in medical settings. By focusing on sex-specific features and utilizing non-invasive clinical data, the ML system offers a novel and accessible approach to understanding AD heterogeneity.

Despite the promising results, there are several limitations that warrant further investigation and refinement to enhance the robustness and generalizability of the proposed ML system. Below are these limitations, along with future actions to address each one. Firstly, the dataset used to train the ML models predominantly comprises individuals from North America. Future research should include cohorts from diverse geographical regions, such as Europe and Asia. This broader representation is essential for enhancing generalizability and ensuring that ML algorithms are reliable and applicable in various real-world contexts. Additionally, the dataset potential over-representation of some demographics and limited diversity in lifestyle factors could impact the generalizability of our results. Future work should address this by incorporating comorbidities and lifestyle factors from the ADNI dataset as covariates to control for their effects. Balancing confounders across groups will also be crucial for reducing bias and improving the robustness of the analysis.

Secondly, further steps will be necessary to make the ML system and the related EMA web application usable in clinical practice. It will be required to ensure compliance with stringent regulatory standards for AI in medicine, such as those set by the Food and Drug Administration in the United States and the European Medicines Agency in Europe. These regulations require thorough validation of the AI system safety, efficacy, and reliability. Additionally, comprehensive clinical trials will be essential to evaluate the system performance in real-world settings. These trials should include diverse patient populations and assess the system impact on diagnostic accuracy, treatment outcomes, and overall patient care. Collaboration with medical professionals and regulatory bodies throughout this process will be crucial to address ethical and practical considerations, ultimately ensuring the system integration into standard medical practice. Additionally, although the current EMA implementation incorporated informal clinician feedback, a more structured evaluation of the system's explainability is needed. Conducting formal user studies and feedback sessions with clinicians from various centers would help assess the clarity and usefulness of the explanations provided, highlighting areas for improvement.

Thirdly, while SHAP and LIME effectively assess feature importance, they may introduce biases due to model-specific assumptions and local approximations. For example, LIME can misinterpret complex interactions if the local linearity assumption fails. Future model improvements should include cross-verifying findings with alternative feature importance procedures, such as permutation or tree-based measures. These approaches can provide complementary insights and help mitigate any biases introduced by SHAP and LIME.

Finally, the current system does not integrate with software capable of autonomously collecting and organizing data from neuropsychological tests directly from patients. Leveraging digital devices like tablets to collect and manage data can streamline the pipeline of data collection plus data analysis [63]. This integration will facilitate preventive fast screening during routine healthcare visits and improve the overall efficiency of the diagnostic process.

CRediT authorship contribution statement

Fabio Massimo D'Amore: Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Marco Moscatelli:** Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Formal analysis. **Antonio Malvaso:** Writing – review & editing, Visualization, Validation, Methodology, Data curation. **Fabrizia D'Antonio:** Visualization, Resources, Methodology, Data curation. **Marta Rodini:** Visualization, Resources, Methodology, Data curation. **Massimiliano Panigutti:** Resources, Data curation. **Pierandrea Mirino:** Resources, Data curation. **Giovanni Augusto Carlesimo:** Writing – review & editing, Visualization, Validation, Methodology. **Cecilia Guariglia:** Writing – review & editing, Visualization, Validation, Methodology. **Daniele Caligiore:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Acknowledgements

GAC research activities are supported by #NEXTGENERATIONEU (NGEU) and funded by the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), project MNESYS (PE0000006) a multiscale integrated approach to the study of the nervous system in health and disease (DN. 1553 11.10.2022).

Appendix A. Additional Examples of Local Explainers

This section presents supplementary examples of local explainers generated using LIME. These examples are provided for male and female subjects across all classes of Models 0 (comprising healthy individuals and those with AD) and Model 1 (distinguishing between individuals with AD onset <5 years and those with AD onset >5 years). Fig. A.9 illustrates the results of the analysis for data from a female subject classified as AD by Model 0. The figure displays the Model 0 confidence intervals for classification, showing 89 % confidence as AD and 11 % as healthy, indicating the subject's healthy status. The figure also depicts the importance values of features specific to the sample, where MMSE is most important with 30 % importance in the "AD" classification, followed by LDELTOTAL at 9 %. Fig. A.10 presents the results of the LIME analysis for data from a female subject who is defined as healthy by Model 0. The figure shows the Model 0 confidence interval of the classification, which is 8 % as AD and 92 % as healthy, reflecting the subject's healthy status with the importance values of the feature for the specific sample, with MMSE having an importance of 20 % in the “Healthy” classification, followed by LDELTOTAL at 12 %.

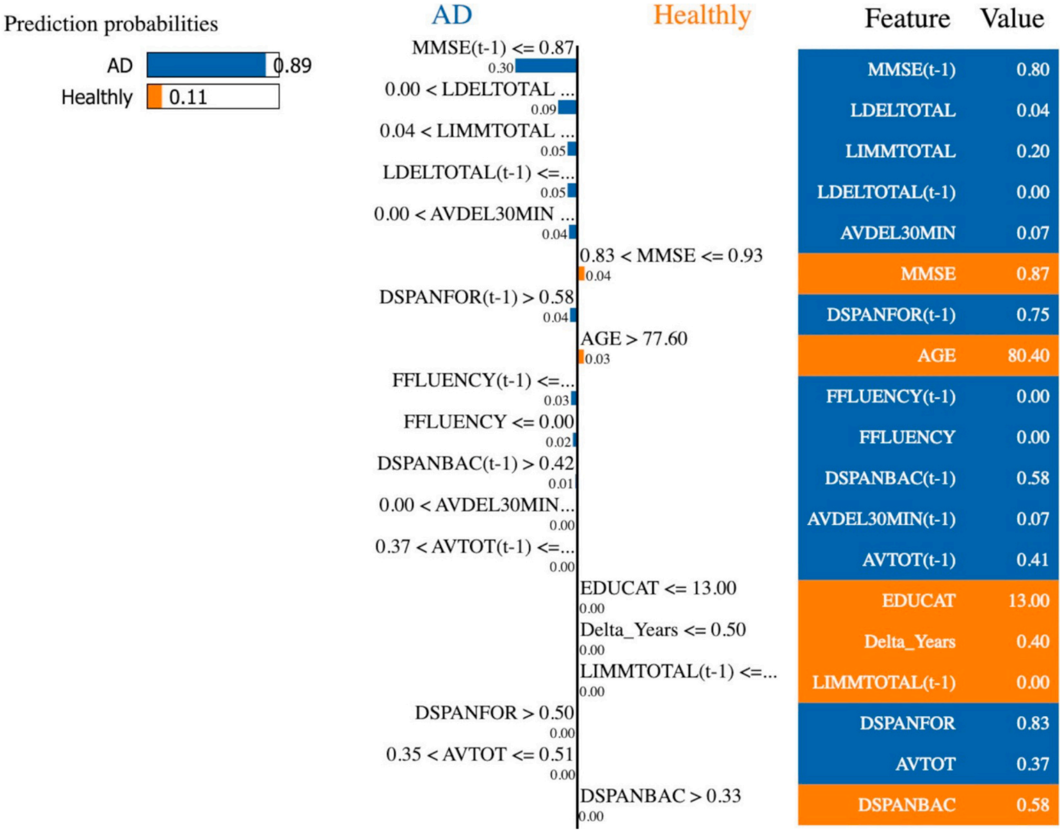


Fig. A.9. LIME explainer for female subject for Model 0 and AD classification.

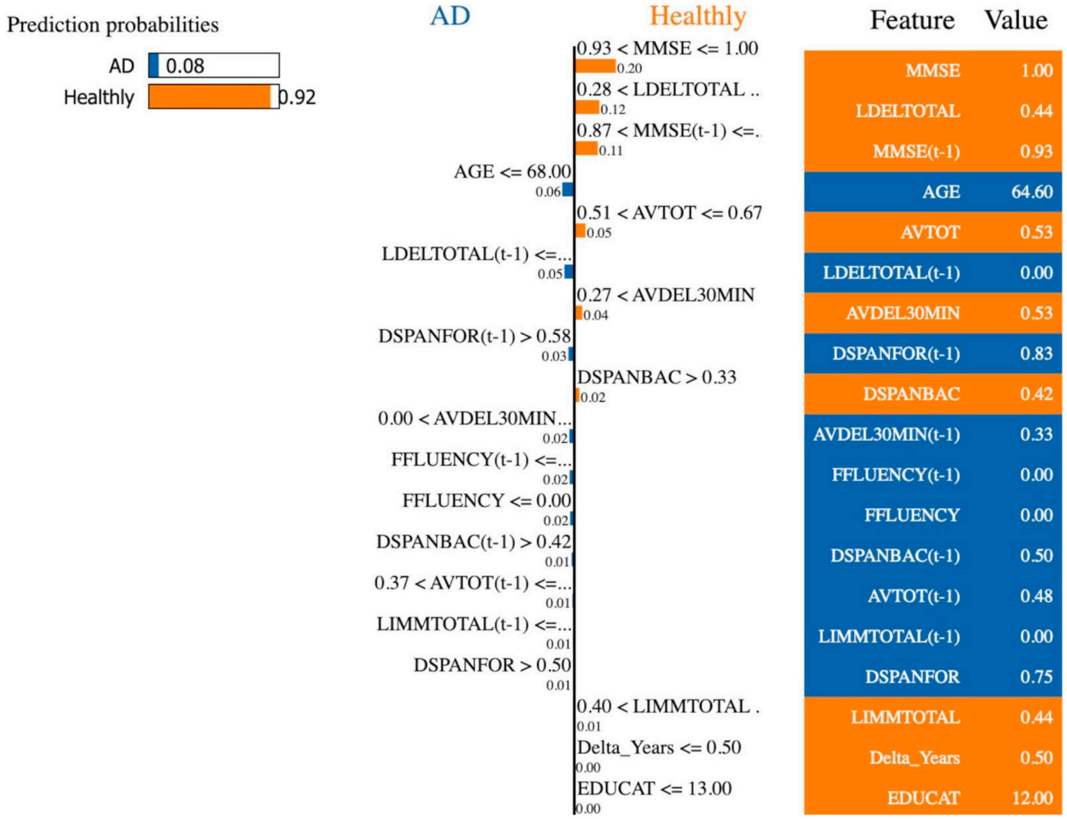


Fig. A.10. LIME explainer for female subject for Model 0 and Healthy classification.

Fig. A.11 shows the Model 1 confidence interval of the classification, which is 78 % as AD onset < 5 years and 22 % as AD onset >5 years, reflecting the subject's healthy status with the importance values of the features for the specific sample, with LDELTOTAL having an importance of 24 % in the"AD onset < 5 years" classification, followed by AVTOT at 11 %. Similarly, according to Fig. A.12 the Model 1 assigns a 90 % probability to the onset of AD occurring more than 5 years in the future and a 10 % probability to the onset occurring in less than 5 years. The importance values of the features for this specific sample are also shown, with the MMSE score contributing 12 % to the “AD onset >5 years” classification, followed by the AVTOT score at 3 %.

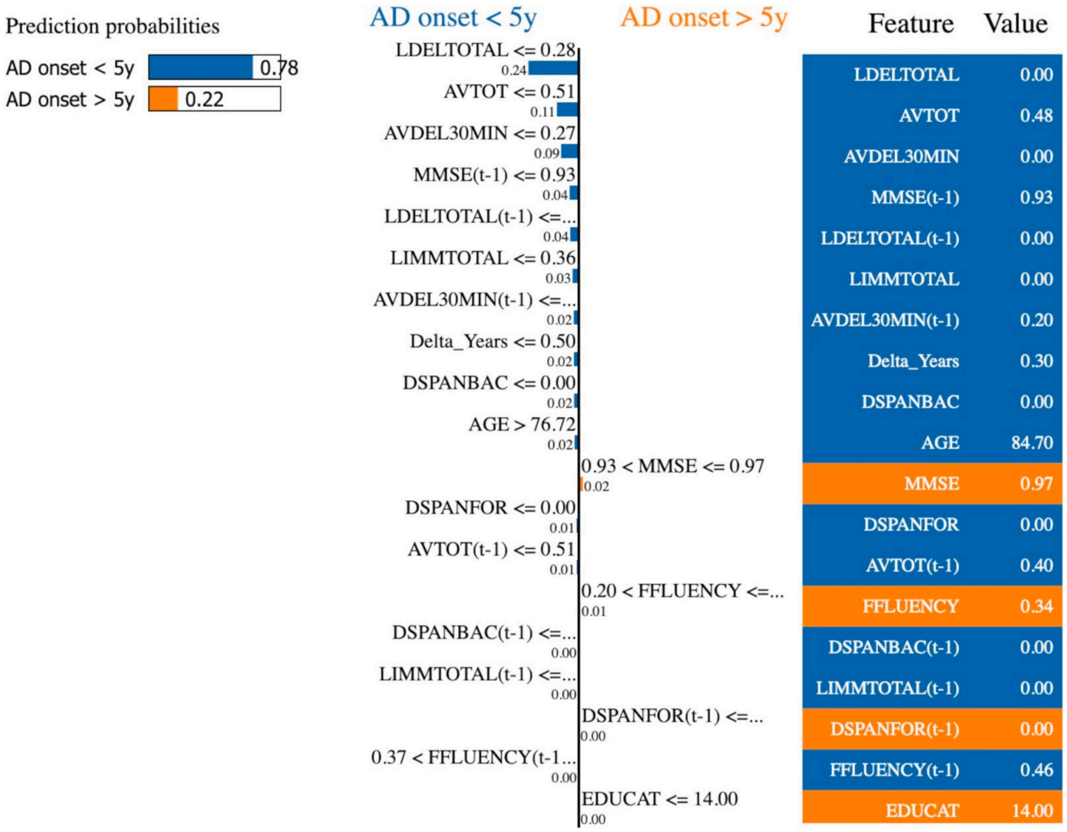


Fig. A.11. LIME explainer for female subject for Model 1 and AD onset <5y classification.

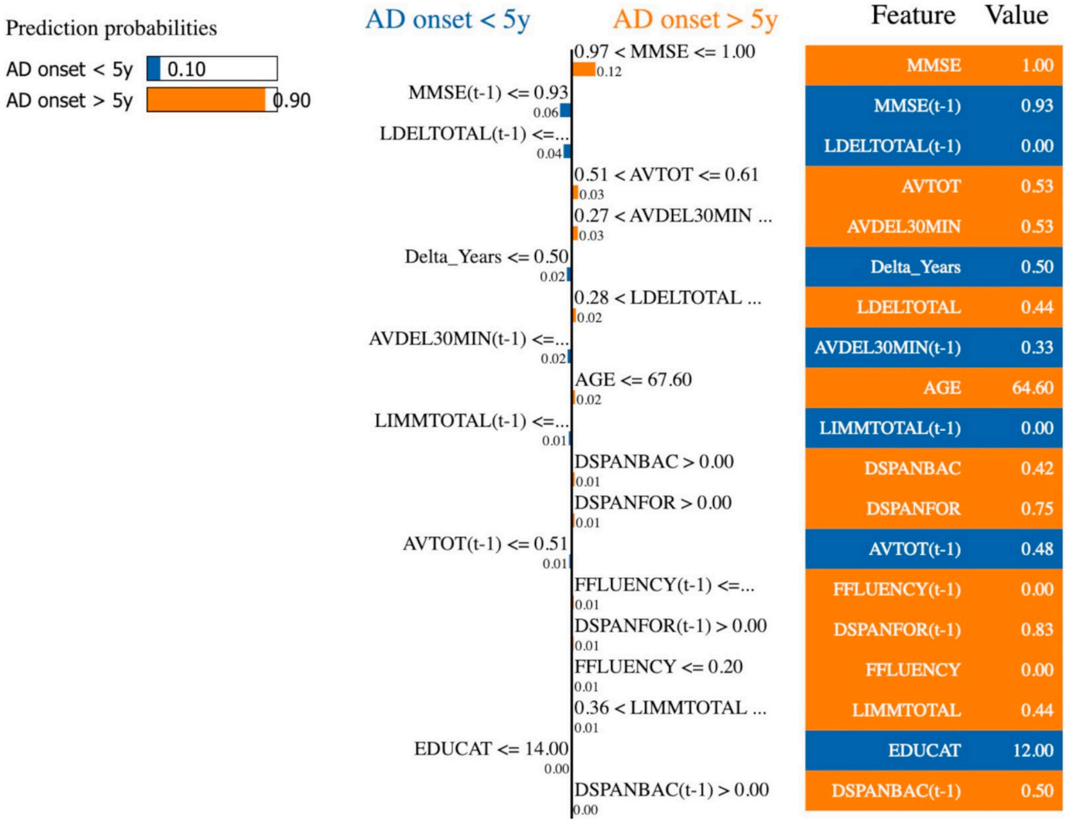


Fig. A.12. LIME explainer for female subject for Model 1 and AD onset >5y classification.

Fig. A.13 shows the analysis results for data from a male subject classified as having AD by Model 0. The figure displays Model 0 confidence intervals for classification, indicating a 96 % confidence level for AD and 4 % for healthy, thereby identifying the subject's condition as AD. Additionally, the figure highlights the importance values of features specific to the sample, with MMSE being the most significant at 24 % in the “AD” classification, followed by LDELTOTAL at 20 %. Fig. A.14 presents the results of the LIME analysis for data from a male subject classified as healthy by Model 0. The figure shows Model 0 confidence intervals for the classification, indicating 14 % confidence as AD and 86 % as healthy, thereby reflecting the subject's healthy status. Additionally, the figure highlights the importance values of features for the specific sample, with LDELTOTAL having an importance of 22 % in the “Healthy” classification, followed by AVTOT at 10 %.

Fig. A.15 presents the results of the LIME analysis for data from a male subject classified as AD onset <5 years by Model 1. The figure shows the Model 1 confidence interval of the classification, which is 85 % as AD onset <5 years and 15 % as AD onset >5 years, reflecting the subject's healthy status with the importance values of the features for the specific sample, with MMSE having an importance of 8 % in the “AD onset <5 years” classification, followed by LDELTOTAL at 5 %. Fig. A.16 presents the results of the LIME analysis for data from a male subject classified as having an AD onset >5 years by Model 1. The figure shows Model 1 confidence intervals for the classification, indicating 99 % confidence for AD onset less than 5 years and 1 % for AD onset more than 5 years. This reflects the subject's status with the importance values of the features for the specific sample, with AVTOT having an importance of 3 % in the “AD onset <5 years” classification, followed by LIMMTOTAL at 3 %.

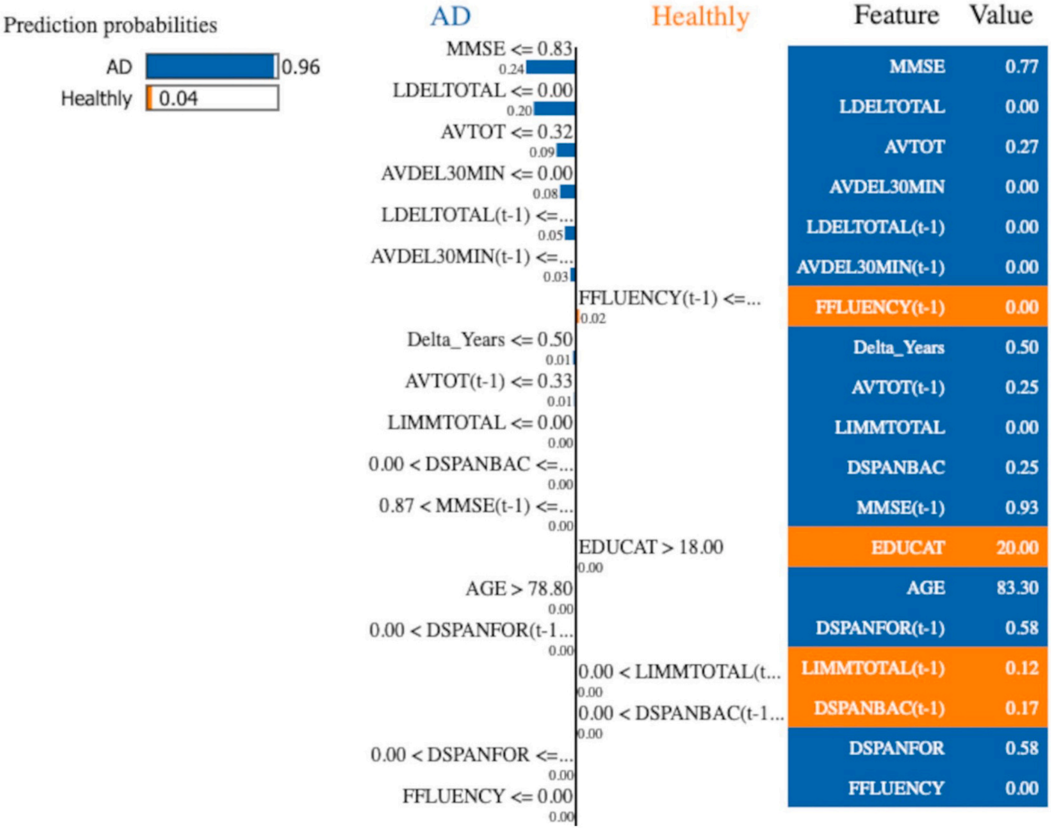


Fig. A.13. LIME explainer for male subject for Model 0 and AD classi.

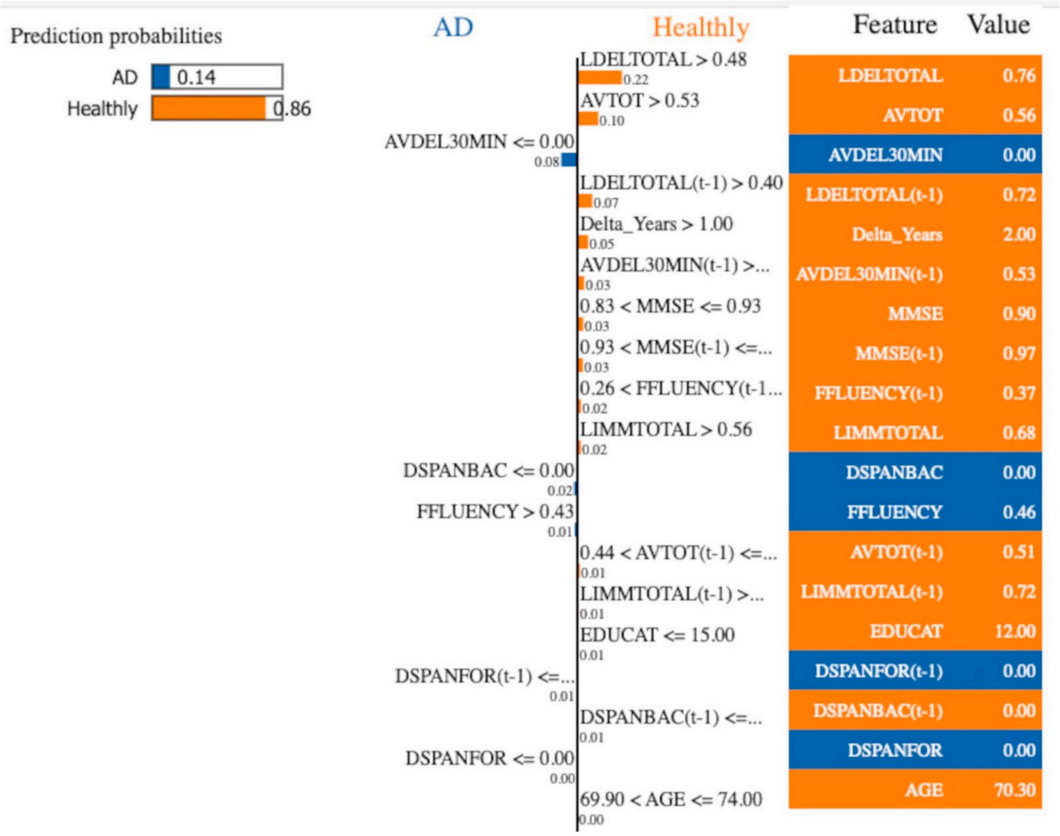


Fig. A.14. LIME explainer for male subject for Model 0 and Healthy classification.

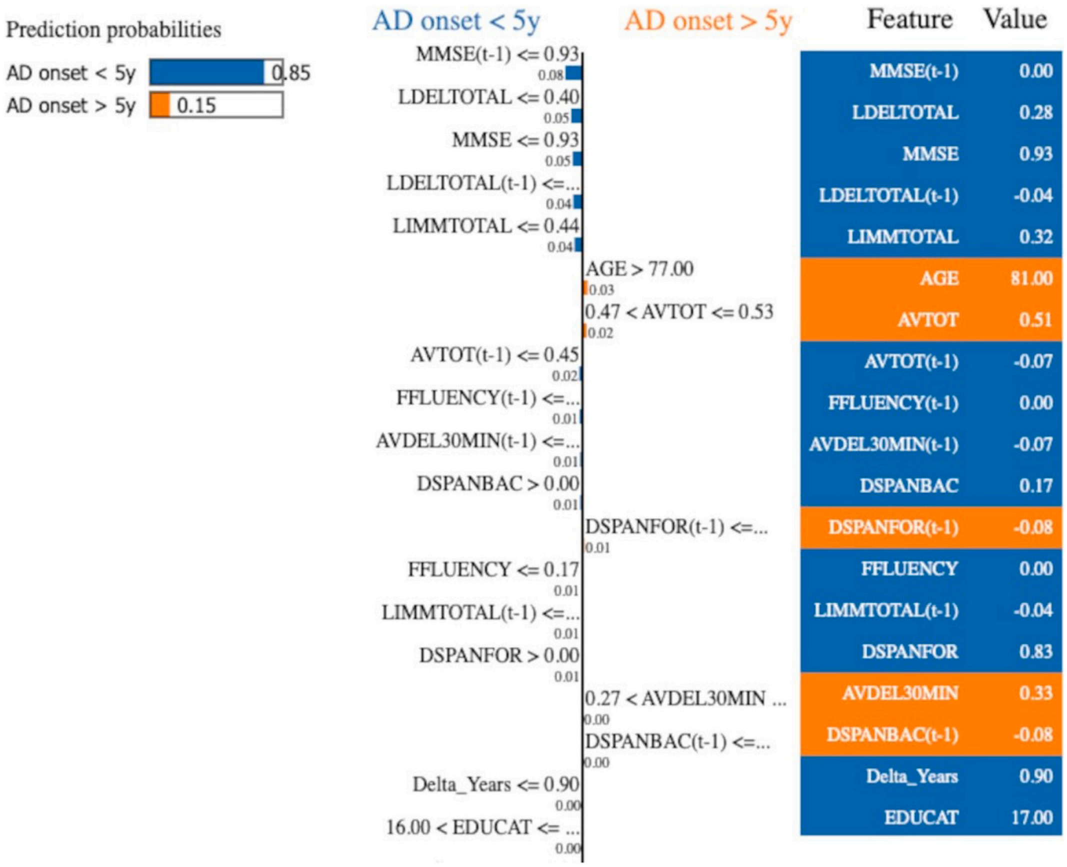


Fig. A.15. LIME explainer for male subject for Model 1 and AD onset <5y classification.

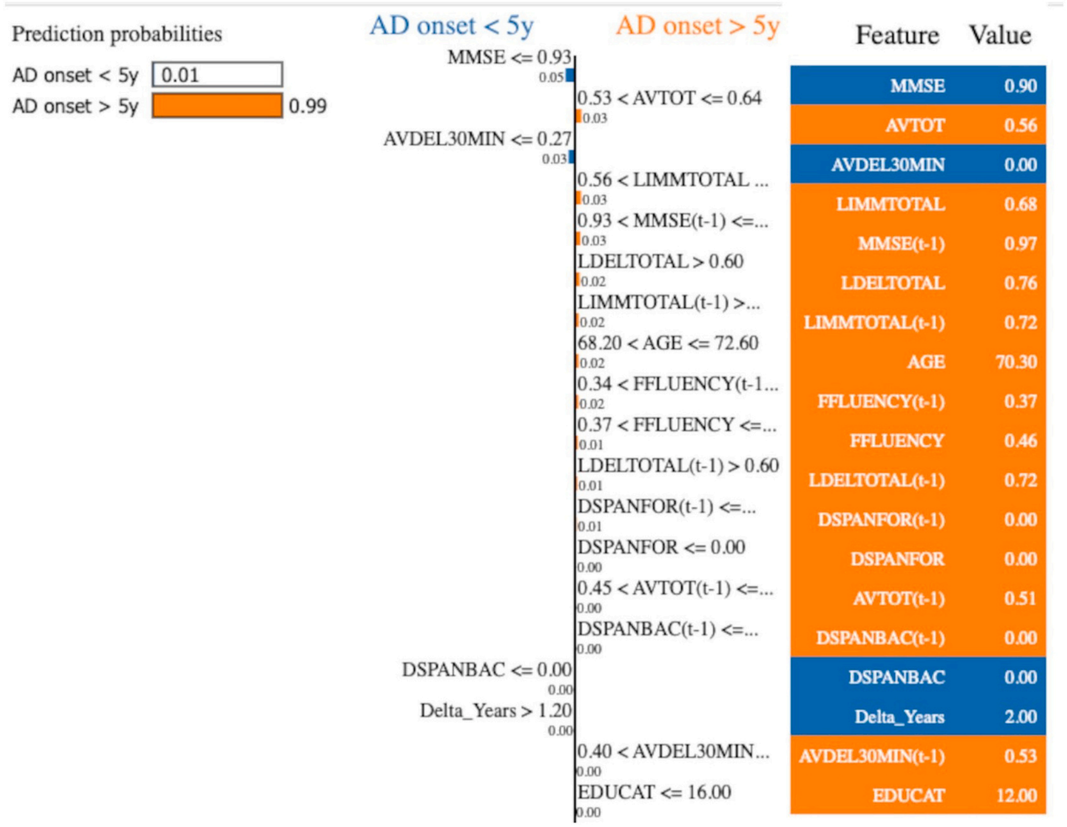


Fig. A.16. LIME explainer for male subject for Model 1 and AD onset >5y classification.

References

[1] F.J. Wolters, L.B. Chibnik, R. Waziry, R. Anderson, C. Berr, A. Beiser, J.C. Bis, D. Blacker, D. Bos, C. Brayne, J.-F. Dartigues, S.K. Darweesh, K.L. Davis-Plourde, F. de Wolf, S. Debette, C. Dufouil, M. Fornage, J. Goudsmit, L. Grasset, V. Gudnason, C. Hadjichrysanthou, C. Helmer, M.A. Ikram, M.K. Ikram, E. Joas, S. Kern, L.H. Kuller, L. Launer, O.L. Lopez, F.E. Matthews, K. McRae-McKee, O. Meirelles, T.H. Mosley, M.P. Pase, B.M. Psaty, C.L. Satizabal, S. Seshadri, I. Skoog, B.C. Stephan, H. Wetterberg, M.M. Wong, A. Zettergren, A. Hofman, Twenty-seven-year time trends in dementia incidence in Europe and the United States, Neurology 95 (2020) e519–e531, https://doi.org/10.1212/WNL.0000000000010022.

[2] X.-X. Zhang, Y. Tian, Z.-T. Wang, Y.-H. Ma, L. Tan, J.-T. Yu, The epidemiology of Alzheimer's disease modifiable risk factors and prevention, J. Prev. Alzheimers Dis. 8 (2021) 313–321.

[3] P. Scheltens, B.D. Strooper, M. Kivipelto, H. Holstege, G. Ch'etelat, C.E. Teunissen, J. Cummings, W.M. van der Flier, Alzheimer's disease, Lancet 397 (2021) 1577–1590.

[4] R.C. Petersen, B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, L. Fratiglioni, Mild cognitive impairment: a concept in evolution, J. Intern. Med. 275 (3) (2014) 214–228.

[5] R. Petersen, J. Parisi, D. Dickson, K. Johnson, D. Knopman, B. Boeve, G. Jicha, R. Ivnik, G. Smith, E. Tangalos, H. Braak, E. Kokmen, Neuropathologic features of amnestic mild cognitive impairment, Arch. Neurol. 63 (2006) 665–672.

[6] R. Roberts, D. Knopman, M. Mielke, R. Cha, V. Pankratz, T. Christianson, Y. Geda, B. Boeve, E.R.J. Ivnik, W.A. Tangalos, R. Petersen Rocca, Higher risk of progression to dementia in mild cognitive impairment cases who revert to normal, Neurology 82 (2014) 317–325.

[7] J. Dukart, F. Sambataro, A. Bertolino, Accurate prediction of conversion to Alzheimer's disease using imaging, genetic, and neuropsychological biomarkers, J. Alzheimers Dis. 49 (2015) 1143–1159.

[8] E.M. Reiman, Y.T. Quiroz, A.S. Fleisher, K. Chen, C. Velez-Pardo, M. Jimenez-Del-Rio, A.M. Fagan, A.R. Shah, S. Alvarez, A. Arbelaez, M. Giraldo, N. Acosta-Baena, R.A. Sperling, B. Dickerson, C.E. Stern, V. Tirado, C. Munoz, R.A. Reiman, M. J. Huentelman, G.E. Alexander, J.B.S. Langbaum, K.S. Kosik, P.N. Tariot, F. Lopera, Brain imaging and fluid biomarker analysis in young adults at genetic risk for autosomal dominant alzheimer's disease in the presenilin 1 E280A kindred: a case-control study, Lancet Neurol. 11 (2012) 1048–1056.

[9] L. Younes, M. Albert, A. Moghekar, A. Soldan, C. Pettigrew, M.I. Miller, Identifying Changepoints in biomarkers during the preclinical phase of Alzheimer's disease, Front. Aging Neurosci. 11 (2019) 74, https://doi.org/10.3389/FNAGI.2019.00074.

[10] M.E. De Vugt, F.R. Verhey, The impact of early dementia diagnosis and intervention on informal caregivers, Prog. Neurobiol. 110 (2013) 54–62.

[11] J. Rasmussen, H. Langerman, Alzheimer's disease - why we need early diagnosis, Degen. Neurol. Neuromusc. Dis. 9 (2019) 123–130.

[12] L. Castro-Aldrete, M.V. Moser, G. Putignano, M.T. Ferretti, A. Schumacher Dimech, A. Santuccione Chadha, Sex and gender considerations in alzheimer's disease: the women's brain project contribution, Front. Aging Neurosci. 15 (2023) 1105620.

[13] M.T. Ferretti, M.F. Iulita, E. Cavedo, P.A. Chiesa, A. Schumacher Dimech, A. Santuccione Chadha, F. Baracchi, H. Girouard, S. Misoch, E. Giacobini, et al., Sex differences in Alzheimer disease—the gateway to precision medicine, nature reviews, Neurology 14 (2018) 457–469.

[14] C.M. Mazure, J. Swendsen, Sex differences in alzheimer's disease and other dementias, Lancet Neurol. 15 (2016) 451–452.

[15] R. Jim'enez-Herrera, A. Contreras, J.D. Navarro-L'opez, L. Jim'enez-D'ıaz, Sex differences in alzheimer's disease: an urgent research venue to follow, Neural Regen. Res. 19 (2024) 2569–2570.

[16] C. Lopez-Lee, E.R.S. Torres, G. Carling, L. Gan, Mechanisms of sex differences in alzheimer's disease, Neuron 112 (2024) 1208–1221.

[17] L. Mosconi, V. Berti, C. Quinn, P. McHugh, G. Petrongolo, I. Varsavsky, R.S. Osorio, A. Pupi, S. Vallabhajosula, R.S. Isaacson, et al., Sex differences in alzheimer risk: brain imaging of endocrine vs chronologic aging, Neurology 89 (2017) 1382–1390.

[18] D. Caligiore, M. Silvetti, M. D'Amelio, S. Puglisi-Allegra, G. Baldassarre, Computational modeling of Catecholamines dysfunction in Alzheimer's disease at pre-plaque stage, J. Alzheimers Dis. 77 (2020) 275–290.

[19] M. Grassi, N. Rouleaux, D. Caldirola, D. Loewenstein, K. Schruers, G. Perna, M. Dumontier, A novel ensemble-based machine learning algorithm to predict the conversion from mild cognitive impairment to Alzheimer's disease using socio-demographic characteristics, clinical information, and neuropsychological measures, Front. Neurol. 10 (2019) 756, https://doi.org/10.3389/fneur.2019.00756.

[20] A.A. Moustafa, Alzheimer's Disease: Understanding Biomarkers, Big Data, and Therapy, Academic Press, 2021, ISBN 9780128213346.

[21] H. Hampel, A. Vergallo, G. Perry, S. Lista, The Alzheimer precision medicine Initiative, J. Alzheimers Dis. 68 (2019) 1–24.

[22] G. Perna, M. Grassi, D. Caldirola, C. Nemeroff, The revolution of personalized psychiatry: will technology make it happen sooner? Psychol. Med. 48 (2018) 705–713.

[23] S. Grueso, R. Viejo-Sobera, Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review, Alzheimers Res. Ther. 13 (2021) 1–29.

[24] N. Pradhan, A.S. Singh, A. Singh, Alzheimer disease early diagnosis and prediction using deep learning techniques: a survey, in: Recent Trends in Communication and Electronics, 2021, pp. 590–593.

[25] M. Odusami, R. Maskeliunas, R. Damasevicius, T. Krilavicius, Analysis of features of Alzheimer's disease: detection of early stage from functional brain changes in magnetic resonance images using a Finetuned ResNet18 network, Diagnostics 11 (2021) 1071, https://doi.org/10.3390/diagnostics11061071.

[26] M. Grassi, G. Perna, D. Caldirola, K. Schruers, R. Duara, D. Loewenstein, A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion in individuals with mild and Premild cognitive impairment, J. Alzheimers Dis. 61 (2018) 1555–1573.

[27] S. Hojjati, A. Ebrahimzadeh, A. Khazaee, A. Babajani-Feremi, Predict- ing conversion from MCI to AD using resting-state fMRI, graph theo- retical approach and SVM, J. Neurosci. Methods 282 (2017) 69–80.

[28] M. Liu, D. Cheng, K. Wang, Y. Wang, Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis, Neuroinformatics 16 (2018) 295–308.

[29] X. Long, L. Chen, C. Jiang, L. Zhang, Prediction and classification of Alzheimer disease based on quantification of MRI deformation, PLoS ONE 12 (2017) e0173372, https://doi.org/10.1371/JOURNAL.PONE.0173372.

[30] D. Pan, A. Zeng, L. Jia, Y. Huang, T. Frizzell, X. Song, Early detection of Alzheimer's disease using magnetic resonance imaging: a novel approach combining convolutional neural networks and ensemble learning, Front. Neurosci. 14 (2020) 259, https://doi.org/10.3389/fnins.2020.00259.

[31] C. Platero, L. Lin, M.C. Tobar, Longitudinal neuroimaging Hippocampal markers for diagnosing Alzheimer's disease, Neuroinformatics 17 (2019) 43–61.

[32] J. Beltran, B. Wahba, N. Hose, D. Shasha, R. Kline, Inexpensive, non-invasive biomarkers predict Alzheimer transition using machine learning analysis of the Alzheimer's disease neuroimaging (ADNI) database, PLoS ONE 15 (2020) e0235663.

[33] M. Merone, S.L. D'Addario, P. Mirino, F. Bertino, C. Guariglia, R. Ventura, A. Capirchio, G. Baldassarre, M. Silvetti, D. Caligiore, A multi-expert ensemble system for predicting alzheimer transition using clinical features, Brain Inf. 9 (2022) 20.

[34] F. Cieri, X. Zhuang, D. Cordes, N. Kaplan, J. Cummings, J. Caldwell, A. D. N. I. (ADNI), Relationship of sex differences in cortical thickness and memory among cognitively healthy subjects and individuals with mild cognitive impairment and alzheimer disease, Alzheimers Res. Ther. 14 (2022) 36.

[35] M. Klingenberg, D. Stark, F. Eitel, C. Budding, M. Habes, K. Ritter, A.D.N. Initiative, Higher performance for women than men in mri-based alzheimer's disease detection, Alzheimers Res. Ther. 15 (2023) 84.

[36] A. Sarica, A. Pelagi, F. Aracri, F. Arcuri, A. Quattrone, A. Quattrone, A.D. N. Initiative, Sex differences in conversion risk from mild cognitive impairment to alzheimer's disease: an explainable machine learning study with random survival forests and shap, Brain Sci. 14 (2024) 201.

[37] B. Allen, The promise of explainable ai in digital health for precision medicine: a systematic review, J. Pers. Med. 14 (2024) 277.

[38] G. Angelini, A. Malvaso, A. Schirripa, F. Campione, S.L. D'Addario, N. Toschi, D. Caligiore, Unraveling sex differences in Parkinson's disease through explainable machine learning, J. Neurol. Sci. 123091 (2024).

[39] M.T. Ferretti, H. Ding, R. Au, C. Liu, S. Devine, S. Auerbach, J. Mez, A. Gurnani, Y. Liu, A. Santuccione, et al., Maximizing utility of neuropsychological measures in sex-specific predictive models of incident Alzheimer's disease in the Framingham heart study, Alzheimers Dement. 20 (2) (2024) 1112–1122.

[40] A.S. Alatrany, W. Khan, A. Hussain, H. Kolivand, D. Al-Jumeily, An explainable machine learning approach for Alzheimer's disease classification, Sci. Rep. 14 (2024) 2637.

[41] D.M. Cammisuli, G. Cipriani, G. Castelnuovo, Technological Solutions for diagnosis, management and treatment of Alzheimer's diseaserelated symptoms: a structured review of the recent scientific literature, Int. J. Environ. Res. Public Health 19 (2022), https://doi.org/10.3390/IJERPH19053122.

[42] M. Odusami, R. MaskeliAnas, R. Damasevicius, An intelligent system for early recognition of alzheimer's disease using neuroimaging, Sensors (2022), https://doi.org/10.3390/S22030740.

[43] A. Silva-Spınola, I. Baldeiras, J.P. Arrais, I. Santana, The road to personalized medicine in Alzheimer's disease: the use of artificial Intelligence, Biomedicines 10 (2022), https://doi.org/10.3390/BIOMEDICINES10020315.

[44] S. Khanna, D. Domingo-Fern'andez, A. Iyappan, M.A. Emon, M. Hofmann-Apitius, H. Frohlich, Using multi-scale genetic, neuroimaging and clinical data for predicting Alzheimer's disease and reconstruction of relevant biological mechanisms, Sci. Rep. 8 (2018), https://doi.org/10.1038/S41598-018-29433-3.

[45] A. Moscoso, J. Silva-Rodr'ıguez, J. M. Aldrey, J. Cort'es, A. Fern'andez-Ferreiro, N. G'omez-Lado, A ́. Ruibal, P. Aguiar, Prediction of Alzheimer's disease dementia with MRI beyond the short-term: Implications for the design of predictive models, NeuroImage 23 (2019). doi:https://doi.org/10.1016/j.nicl.2019.101837.

[46] H. Chen, P. Yu, D. Hailey, T. Cui, Identification of the essential components of quality in the data collection process for public health information systems, Health Inf. J. 26 (1) (2020) 664–682, pMID: 31140353, https://doi.org/10.1177/1460458219848622.

[47] Y. Tabei, H. Saigo, Y. Yamanishi, S.J. Puglisi, Scalable Partial Least Squares Regression on Grammar-Compressed Data Matrices, 2016, pp. 1875–1884, https://doi.org/10.1145/2939672.2939864.

[48] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From Local Explanations to Global Understanding With Explainable AI for Trees, 2020, https://doi.org/10.1038/s42256-019-0138-9.

[49] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[50] C. Pradier, C. Sakarovitch, F. Le Duff, R. Layese, A. Metelkina, S. Anthony, K. Tifratene, P. Robert, The mini mental state examination at the time of

alzheimer's disease and related disorders diagnosis, according to age, education, gender and place of residence: a cross-sectional study among the french national alzheimer database, PLoS ONE 9 (2014) e103630.

[51] L.L. Barnes, R.S. Wilson, J.L. Bienias, J.A. Schneider, D.A. Evans, D.A. Bennett, Sex differences in the clinical manifestations of alzheimer disease pathology, Arch. Gen. Psychiatry 62 (2005) 685–691.

[52] K. Irvine, K.R. Laws, T.M. Gale, T.K. Kondel, Greater cognitive deterioration in women than men with alzheimer's disease: a meta analysis, J. Clin. Exp. Neuropsychol. 34 (9) (2012) 989–998.

[53] M.L. Bleecker, K. Bolla-Wilson, J. Agnew, D.A. Meyers, Age-related sex differences in verbal memory, J. Clin. Psychol. 44 (3) (1988) 403–411.

[54] S.D. Gale, L. Baxter, D.J. Connor, A. Herring, J. Comer, Sex differences on the rey auditory verbal learning test and the brief visuospatial memory test–revised in the elderly: normative data in 172 participants, J. Clin. Exp. Neuropsychol. 29 (5) (2007) 561–567.

[55] J.J. Ryan, L. Glass Umfleet, D.S. Kreiner, A.M. Fuller, A.M. Paolo, Neuropsychological differences between men and women with alzheimer's disease, Int. J. Neurosci. 128 (4) (2018) 342–348.

[56] E.E. Sundermann, P.M. Maki, L.H. Rubin, R.B. Lipton, S. Landau, A. Biegon, A.D. N. Initiative, Female advantage in verbal memory: evidence of sex-specific cognitive reserve, Neurology 87 (18) (2016) 1916–1924.

[57] S. Subramaniapillai, A. Almey, M.N. Rajah, G. Einstein, Sex and gender differences in cognitive and brain reserve: implications for alzheimer's disease in women, Front. Neuroendocrinol. 60 (2021) 100879.

[58] L. Letenneur, J. Launer, K. Andersen, M. Dewey, A. Ott, J. Copeland, J. Dartigues, P. Kragh-Sorensen, M. Baldereschi, C. Brayne, et al., Education and risk for alzheimer's disease: sex makes a difference eurodem pooled analyses, Am. J. Epidemiol. 151 (11) (2000) 1064–1071.

[59] L. Letenneur, V. Gilleron, D. Commenges, C. Helmer, J.-M. Orgogozo, J.-F. Dartigues, Are sex and educational level independent predictors of dementia and alzheimer's disease? Incidence data from the paquid project, J. Neurol. Neurosurg. Psychiatry 66 (2) (1999) 177–183.

[60] N. Prentzas, A. Kakas, C.S. Pattichis, Explainable ai applications in the medical domain: a systematic review, arXiv (2023), 2308.05411.

[61] Y. Kou, X. Gui, Mediating community-ai interaction through situated explanation: the case of ai-led moderation, Proc. ACM Human-Comp. Interact. 4 (CSCW2) (2020) 1–27.

[62] V. Viswan, N. Shaffi, M. Mahmud, K. Subramanian, F. Hajamohideen, Explainable artificial intelligence in alzheimer's disease classification: a systematic review, Cogn. Comput. 16 (2024) 1–44.

[63] A. Di Vita, F. Vecchione, M. Boccia, A. Bocchi, M.C. Cinelli, P. Mirino, A. Teghil, F. D'Antonio, C. de Lena, L. Piccardi, et al., Diane: a new first-level computerized tool assessing memory, attention, and visuospatial processing to detect early pathological cognitive decline, J. Alzheimers Dis. 86 (2) (2022) 891–904.