Prediction of Alzheimer Disease on the DARWIN Dataset with Dimensionality Reduction and Explainability Techniques

Alexandre Moreira[1], Artur Ferreira[1,2][a] and Nuno Leite[1][b]
[1]*ISEL, Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa, Portugal*
[2]*Instituto de Telecomunicações, Lisboa, Portugal*

Keywords: Alzheimer Disease Prediction, Classification, Dimensionality Reduction, Explainability, Feature Selection, Handwriting Tasks, Neurodegenerative Diseases.

Abstract: The progressive degeneration of nerve cells causes neurodegenerative diseases. For instance, Alzheimer and Parkinson diseases progressively decrease the cognitive abilities and the motor skills of an individual. Without the knowledge for a cure, we aim to slow down their impact by resorting to rehabilitative therapies and medicines. Thus, early diagnosis plays a key role to delay the progression of these diseases. The analysis of handwriting dynamics for specific tasks is found to be an effective tool to provide early diagnosis of these diseases. Recently, the Diagnosis AlzheimeR WIth haNdwriting (DARWIN) dataset was introduced. It contains records of handwriting samples from 174 participants (diagnosed as having Alzheimer's or not), performing 25 specific handwriting tasks, including dictation, graphics, and copies. In this paper, we explore the use of the DARWIN dataset with dimensionality reduction, explainability, and classification techniques. We identify the most relevant and decisive handwriting features for predicting Alzheimer. From the original set of 450 features with different groups, we found small subsets of features showing that the time spent to perform the in-air movements are the most decisive type of features for predicting Alzheimer.

1 INTRODUCTION

Machine learning (ML) aims to solve problems by learning a model from input data. The model is then applied on a decision process. The foundations of this decision-making are often opaque to the user, such that one is unable to establish causal links such as: *"Model X made decision Y because of Z"*. This may be a problem in certain fields (e.g. medical, military, and economic) where the consequences of an error can be harmful. Moreover, knowing the cause of decision-making helps the research itself to analyze the reasons behind each decision.

As an attempt to clarify algorithmic decision making, the concept of *explainable artificial intelligence* (XAI) arises. XAI techniques aim to provide explanations for the decisions of a given ML model (Bastani et al., 2017; Kim et al., 2018; Lakkaraju and Bastani, 2020; Lou et al., 2013a; Mothilal et al., 2020; Ribeiro et al., 2016a). These explanations can take various forms, such as visual, numerical, textual, or rule-based. They can be provided covering a *global* scope, where the overall behavior of a model is described as a whole, or a *local* scope, in which partial explanations of the model's behavior are reported, being valid only for some or even a single decision. Extracting explanations from models can be model-independent (model-agnostic) or model-dependent (the extraction method is specific to the model). Also there are transparent models which are inherently interpretable and do not require external components to provide explanations.

Neurodegenerative diseases are caused by the progressive degeneration of nerve cells, being incurable at this point. Alzheimer, Parkinson, and Huntington are probably the most well-known neurodegenerative diseases with consequences such as the cognitive abilities and the motor skills of an individual are increasingly affected over time. Without a cure, the medical staff actions aim to slow down their impact by resorting to rehabilitative therapies and medicines. Thus, the early diagnosis of these diseases is still a key factor to delay their progression. This diagnosis can be carried out by analyzing the way an individual performs some specific writing tasks and actions.

[a] https://orcid.org/0000-0002-6508-0932
[b] https://orcid.org/0000-0001-6328-3579

Moreira, A., Ferreira, A. and Leite, N.
Prediction of Alzheimer Disease on the DARWIN Dataset with Dimensionality Reduction and Explainability Techniques.
DOI: 10.5220/0013017400003886
In *Proceedings of the 1st International Conference on Explainable AI for Neural and Symbolic Methods (EXPLAINS 2024)*, pages 38-49
ISBN: 978-989-758-720-7

In this paper, we explore the use of dimensionality reduction and XAI techniques to build ML models for the prediction of the Alzheimer disease. The study involves the use of the recently introduced Diagnosis AlzheimeR WIth haNdwriting (DARWIN) dataset, by Cilia et al. (2022), with the purpose of detecting Alzheimer through the analysis of handwriting tasks.

The remainder of this paper is organized as follows. Section 2 briefly reviews related work for Alzheimer's disease detection and the techniques employed on our study. In section 3, the proposed machine learning approach to detect the Alzheimer disease through on-line writing is described as well as the dataset used for this purpose. Section 4 presents details about the undertaken experimental evaluation. In section 5, we draw the main conclusions of our study and pinpoint future directions.

2 RELATED WORK

In this section, we review *Alzheimer's disease* (AD) detection as well as to ML techniques considered in our approach. In section 2.1, we refer to existing data and initiatives for Alzheimer detection. Some related work on the DARWIN dataset is described in section 2.2. In section 2.3, we briefly review some explainability techniques considered in our experiments. Finally, section 2.4 addresses dimensionality reduction with feature selection to improve classification accuracy and leverage explainability.

2.1 Alzheimer Detection Initiatives

The detection of AD has been attracting the attention of the ML research community. We now have different sources of data for the detection of this disease. One of them is provided by Kumar and Shastri (2022). This dataset consists of pre-processed magnetic resonance images, to detect four different levels of dementia: *"Non Demented"*, *"Very Mild Demented"*, *"Mild Demented"*, and *"Moderate Demented"*.

The *Alzheimer's disease neuroimaging initiative* (ADNI) repository stores brain images and genetic biometric data[1]. It provides several sets of data to study and improve techniques for detecting Alzheimer's. Another repository based on Alzheimer's detection is *open access series of imaging studies* (OASIS)[2]. In Koenig et al. (2020), we have a volumetric biometric model to predict Alzheimer's that uses an instance of OASIS repository.

The *national institute on aging genetics of Alzheimer's disease data storage site* (NIAGADS) repository provides datasets for Alzheimer's disease in different formats[3].

2.2 The DARWIN Dataset

In Cilia et al. (2022), the authors present the DARWIN dataset, and propose a ML approach for detecting Alzheimer's by analysing writing tasks that comprise the DARWIN dataset, available on the public domain. This dataset was chosen for our study due to its novelty and to the fact that it was considered interesting and worth to explore, after analyzing the existent literature. Moreover, it was easy to explore, since image or gene expression data has a much larger dimension and could be more difficult to process. In addition to dimensionality, the size of other datasets can reach 1.5 GB such as *OASIS-1*, being computationally demanding to analyze.

The DARWIN dataset contains handwriting samples from people affected by Alzheimer's as well as a control group, in a total of 174 participants. The data was acquired during the execution of 25 specific handwriting tasks, following a protocol for the early detection of Alzheimer's. These tasks address individuals performing dictation, graphic, and copy tasks, and their performance is expressed on a total of 450 features. On the following, we briefly review recent techniques that use the DARWIN dataset to predict the AD.

In Azzali et al. (2024), the authors present a novel approach based on *vectorial genetic programming* (VE_GP) to recognize the handwriting of AD patients. VE_GP is an improved version of GP to manage time series directly. VE_GP was applied to the DARWIN dataset and the experimental results confirmed the effectiveness of the proposed approach in terms of classification performance, size, and simplicity.

A review of *artificial intelligence* (AI) methods for AD diagnosis is presented in Bazarbekov et al. (2024). The review introduces the importance of diagnosing AD accurately and the potential benefits of using AI techniques and ML algorithms for this purpose. The review is based on various state-of-the-art data sources including MRI data, PET imaging, EEG and MEG signals, handwriting data, among other data sources.

Some of the authors that have proposed the DARWIN dataset have further investigated the handwriting of people affected by AD in Cilia et al. (2024). The tasks proposed in handwriting datasets such as DARWIN focused on different cognitive skills to elicitate

[1] https://adni.loni.usc.edu/data-samples

[2] https://sites.wustl.edu/oasisbrains

[3] https://www.niagads.org/home

handwriting movements. It is believed that the meaning and phonology of words to copy can compromise writing fluency. In this context, these authors investigated the impact of word semantics and phonology and how it affects the handwriting of people affected by AD. Their results confirmed that AD harms brain areas processing visual feedback.

In Erdogmus and Kabakus (2023), the authors present a novel *convolutional neural network* (CNN) as a cheap, fast, and accurate solution to detect AD. The proposed CNN was built based on the following process. The 1D features extracted from the analysis of handwriting tasks of the DARWIN dataset were transformed into 2D features, which were yielded into the proposed novel model. Then, the model was trained and evaluated on this dataset. Experimental results show that the model obtained an accuracy as high as 90.4%, which was higher than the accuracies obtained by the state-of-the-art baselines.

In Cappiello and Caruso (2024), a Quantum ML technique is applied to DARWIN dataset. Quantum ML is a recent research field that combines quantum information science and machine learning. The authors also use the DARWIN dataset to test kernel methods for the classification task and compare their performances with the ones obtained via quantum machine learning methods. They found that quantum and classical algorithms achieve similar performances and in some cases quantum methods perform better.

In a recent work by Mitra and Rehman (2024), the authors developed an ensemble ML model for analysis of handwriting kinetics (based on the DARWIN dataset), with the stacking technique to integrate multiple base-level classifiers. The experimental evaluation proved the high efficiency of the developed technique, where the proposed model surpasses all state-of-the-art models based on the DARWIN dataset for AD prediction.

2.3 Explainability Techniques

Recently, we have witnessed an increasing interest on the use of explainability and interpretability techniques (Bastani et al., 2017; Kim et al., 2018; Lakkaraju and Bastani, 2020; Lou et al., 2013a; Mothilal et al., 2020; Ribeiro et al., 2016b). In this section, we briefly review explainability techniques.

The *local interpretable model-agnostic explanations)* (LIME) (Ribeiro et al., 2016b) technique explains the predictions of any classifier with an interpretable model, locally around the prediction. LIME models the local neighborhood of a prediction, by perturbing an individual instance and by generating synthetic data. Using LIME, one can interpret an explanation in a similar way as a linear model. However, some explanations are occasionally unstable and highly dependent on the perturbation process.

The *SHapley Additive exPlanations* (SHAP) approach is a game-theoretic method that explains the output of any ML model. It resorts to optimal credit allocation with local explanations using Shapley values (a concept developed in cooperative game theory). The SHAP values provide insights into the importance of the features (Scheda and Diciotti, 2022).

The *explainable boosting machine* (EBM) (Lou et al., 2013a) is a generalized additive model with automatic interaction detection. An EBM model is often as accurate as state-of-the-art black-box models, while remaining interpretable. EBM models are often slower to train than other methods; however they are compact and provide fast prediction.

The *knowledge distillation* (KD) approach, transfers knowledge from a large model to a smaller one (Bastani et al., 2017). The large model is the black-box or teacher while the smaller one is the explainer or student. The student model is learned to imitate the behavior of the teacher, while remaining interpretable.

In Szepannek and Lübke (2022), we have the *partial dependence plots* (PDP), which follows a model-agnostic assessment strategy for each feature, evaluating its effect on the model response. The degree of model explainability extending the concept of explainability to the multiclass case is explored.

2.4 Dimensionality Reduction

In our approach, we explore the use of dimensionality reduction. We resort to the *k-fold feature selection* (KFFS) filter, proposed by Ferreira and Figueiredo (2023). It is a filter *feature selection* (FS) method, based on the idea that the discriminative power of a feature is proportional to the number of times it is chosen, on the k-folds over the training data, by some unsupervised or supervised FS filter. KFFS is controlled by two parameters: the number of folds k to sample the training data; the threshold T_h to assess the percentage of choice of a feature by the filter on the k-folds. Figure 1 depicts the input and output parameters of the KFFS algorithm, using a generic FS filter denoted as $@filter$, which is applied on k-folds of the input data.

As filters to be used in KFFS, we consider the Fisher ratio or Fisher score (Fisher, 1936). For the i-th feature, the Fisher score is defined as

$$\mathrm{FiR}_i = \frac{\left|\overline{X}_i^{(-1)} - \overline{X}_i^{(1)}\right|}{\sqrt{\mathrm{var}(X_i)^{(-1)} + \mathrm{var}(X_i)^{(1)}}}, \qquad (1)$$

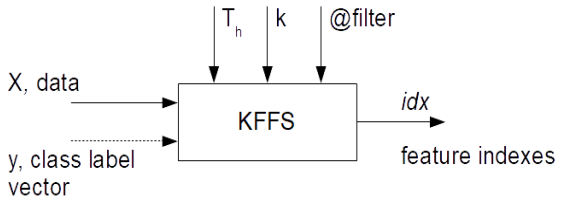

Figure 1: The k-fold feature selection (KFFS) algorithm proposed by Ferreira and Figueiredo (2023).

where $\overline{X}_i^{(-1)}$, $\overline{X}_i^{(1)}$, $\text{var}(X_i)^{(-1)}$, and $\text{var}(X_i)^{(1)}$, are the means and variances of feature X_i, for the patterns of each class. The ratio measures how well each feature alone separates the two classes (Fisher, 1936).

In this paper, we also consider the *fast correlation-based filter* (FCBF), proposed by Yu and Liu (2003, 2004). FCBF starts by selecting a set of features highly correlated with the class, with a correlation value above a threshold. This correlation is assessed by *symmetrical uncertainty* (SU) (Yu and Liu, 2003), defined as

$$SU(X_i, X_j) = \frac{2I(X_i; X_j)}{H(X_i) + H(X_j)}, \qquad (2)$$

where $H(.)$ is the entropy defined by Shannon and $I(.)$ denotes the *mutual information* (MI) (Cover and Thomas, 2006). SU is zero for independent random variables and equal to one for deterministically dependent ones. On its first step, FCBF identifies the predominant features, the ones with higher correlation with the class. In the second step, a redundancy detection analysis finds redundant features among the predominant ones. The set of redundant features is processed, removing the redundant features keeping the most relevant ones.

3 PROPOSED APPROACH

In this section, we describe our approach. Section 3.1 describes our overall ML pipeline. The DARWIN dataset preprocessing tasks are described in section 3.2. The classification and explainability techniques are decribed in section 3.3.

3.1 Machine Learning Pipeline

Our approach relies on the use of classification and explainability techniques over the DARWIN dataset. We aim to accurately detect Alzheimer's disease and to identify the most decisive features, following two approaches:

- a model-agnostic explainer, depicted in Figure 2;
- a transparent explainer, shown in Figure 3.

Moreover, we apply these approaches on the original dataset with $d = 450$ features as well as on the dimensionality reduced dataset with $m < d$ features, attained by the KFFS algorithm, described in section 2.4.

3.2 Dataset Preprocessing

We started by analyzing the domain of values of each column to identify whether it would be categorical or numerical. Often, in the case of categorical features, it is necessary to transform them into a numerical domain. On the DARWIN dataset, there was only a single column with categorical values, the "ID" column to identify the individual being tested; it was decided to discard this column, making a total of 450 features. The domain value of the classes from categorical to numeric is as follows: class H (Healthy) labelled as 0, and class P (Patient) labelled as 1. After the pre-processing stage, we proceeded with data partition into a training set (80%) and a test set (20%). We have also chosen the instance that will be used to extract local explanations. The selected instance was number 3 of the test set (belonging to class "0").

We have carried out experiments with the original dataset and other experiments with the dimensionality reduced datasets, by using KFFS as described in section 2.4. Our aim is to assess the impact of dimensionality reduction on the DARWIN dataset and to analyze the results of explainability techniques on the original and on the reduced space.

With the extraction of explanations we aim to find the most relevant features and groups of features for this dataset. These explanations will facilitate the work of the clinical domain expert since each group of features, or at least some groups, may exhibit certain values for different causes. For example, perhaps a high "total_time" may have different causes than a high "pressure_var", which may pin point the region of the brain that is affected, but ultimately this interpretation should be carried by the domain expert (medical doctor).

3.3 Machine Learning Techniques

As classification models, based on the existing literature, we have chosen well-known algorithms, such as *logistic regression* (LR) (David W. Hosmer Jr., 2013), *random forest* (RF) (Breiman, 2001), *support vector machines* (SVM) (Vapnik, 1999), and *explainable boosting machine* (EBM) (Lou et al., 2013b). LR learns a statistical model to estimate a single binary dependent variable, coded by an indicator variable.

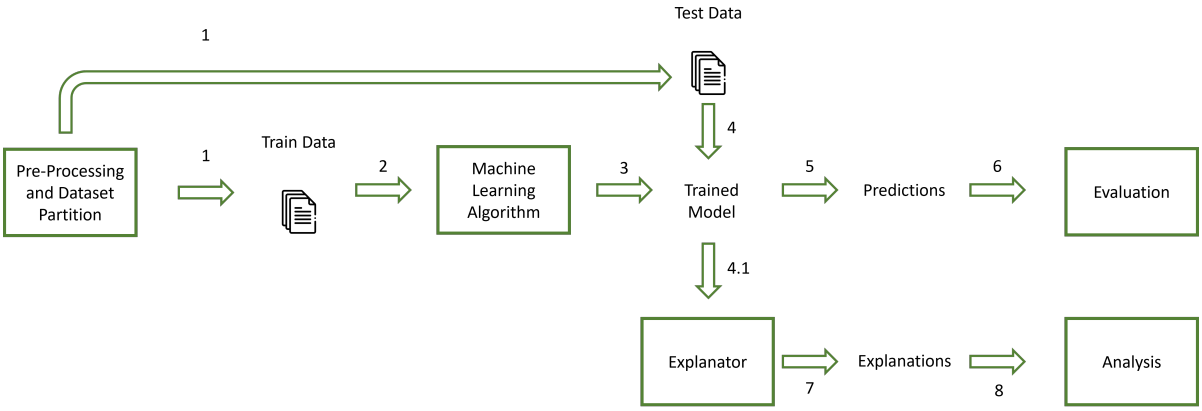

Figure 2: System block diagram for model-agnostic explainers.

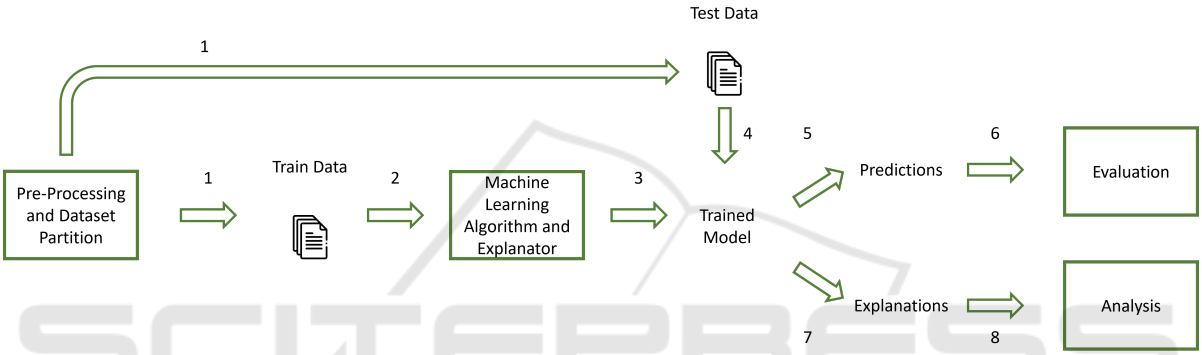

Figure 3: System block diagram for transparent explainers.

RF is an ensemble method that combines several decision trees to build a classifier. SVM map the input data to a high-dimensional feature space to classify the data points.

As explainability techniques, we consider the use of EBM, LIME, and SHAP, mentioned in section 2.3. We also use EBM for classification purposes.

4 EXPERIMENTAL EVALUATION

In this section, we present details about the undertaken experiments. Section 4.1 describes the experimental settings followed by our approach and its evaluation. In section 4.2, we report the baseline (no feature selection) classification results without and with hyperparameter tunning. Section 4.3 addresses local explanation results while section 4.4 presents global explanation results. The dimensionality reduction evaluation results are reported in section 4.5. Finally, the discussion on the experimental results is the topic of section 4.6.

4.1 Experimental Settings

The experiments were carried out on an AMD Ryzen 7 1700 Eight-Core Processor 3.00 GHz PC, 16 GB RAM, with Radeon RX 580 Graphics card. The Operating System is Windows 10 Home. The source code was written in the Python language and run using the Jupyter Notebook. The ML models under evaluation have the following hyperparameters:

- LR - solver="lbfgs", max_iter=500, C=1;
- RF - n_estimators=100, boostrap=True, max_features = "sqrt";
- SVM - kernel="linear", C=1, decision_function_shape = "ovr", gamma="scale";
- EBM - smoothing_round=200, max_bins = 1024, cyclic_progress=1;

As compared to the default parameter values, we modified LR's *max_iter* from 100 to 500 (due to a warning about insufficient number of iterations in order for the algorithm to converge), and in the case of SVM, we have changed the *kernel* from "rbf" to "linear", which allowed for improved results.

The evaluation metrics are: *Accuracy*, *False-positive rate*, *False-negative rate*, *Precision*, *Recall*, and *F1-score* (Rainio et al., 2024).

4.2 Baseline Classification Results

Table 1 reports the confusion matrices, on the test set for all models while Table 2 presents the classification metrics.

In Table 1, all classifiers achieved the same confusion matrices except for SVM. This behavior can also be seen on the metrics in Table 2. According to these results the accuracy for LR, RF, SVM and EBM was, respectively, 89%, 89%, 86% and 89%. However, after calculating the confusion matrices for one train/test partition we proceeded to calculating the performance for 10 different partitions. The results can be seen in Table 2. In terms of overall performance EBM was the leading classifier and RF achieved very similar results with, respectively 89% and 88% of hit rate, in addition, they achieved the same FN rate and recall. Then LR in third place with 79% of hit rate and in the last place was SVM with 74%. These results show that the partition used to generate the confusion matrices in Table 1, did not represent the overall results, in the case of LR and SVM, which have a considerable lower mean accuracy. In spite of this difference the explanations were extracted from the first step and according to some tests made, they do not differ very much from those extracted from a partition that resulted in metrics closer to the mean.

We have also performed *GridSearchCV* from *scikit-learn* to find the best values for the hyperparameters of the classifiers, that would result in better performance. We have found the following: for LR, C=0.01, max_iter=500, solver= "lbfgs"; for SVM kernel="rbf", C=500, decision_function_shape="ovr", gamma ="scale"; for EBM max_bins=1024, smoothing_rounds=500, cyclic_progress=0. The results obtained with these hyperparameters, are also reported in Table 2.

After hyperparameter tunning, all models achieve improvements in metrics specially SVM, which now surpassed LR in terms of hit rate. EBM and RF also improved performance having both a mean accuracy of 91% being, once again, the classifiers with the superior metrics.

Table 1: Confusion matrices for LR, RF, SVM, and EBM.

LR

14	1
3	17

RF

14	1
3	17

SVM

13	2
3	17

EBM

14	1
3	17

4.3 Local Explainability Results

We now report the experimental results of the explainability techniques, namely, LIME, SHAP, and EBM local explainers. For the LIME explainer, one instance from the test set was chosen to provide explanations. We have chosen one instance from class "0". Figure 4 depicts the extraction of local explanations for the SVM classifier using LIME. The graphical explanation is organized into three components, from left to right: probabilities of the classifier itself; importance of the most relevant features; selected features and their value. The feature considered as the most important with a strong influence on class "1" is "airtime_19", indicating the time of the pen off the paper for task 19. In general, it seems that an "air_time" high, even for other tasks, contributes to the positive class, which seems to make sense, given that the individual would spend more time without performing the task itself, which indicates difficulty. The second most relevant feature is "total_time19" yielding that a shorter time to perform the task contributes to the negative class. Another group of characteristics considered important was "pressure_var", with 2 appearances among the 10 most important characteristics. The less time an individual takes to perform a task, the greater the dexterity and, therefore, the person will be healthier.

We have also considered the SHAP explanations for our models, on the same training data. Since EBM is a transparent explainer, it is able of simultaneously carrying out the classification task and to extract explanations. Also to make EBM's comparison smoother, changes were made to the visualization of this method to resemble that of SHAP and despite their slight difference, they present the same type of information, as depicted in Figure 5.

From Figure 5, we notice the difference in fidelity as compared to LIME, given that the value of the decision function, for this case, is always the same in relation to the original model, as can be found directly below the titles. We observe a predominance of the "pressure_var" group of features, with 5 features similar to the 9 most important ones. Another interesting aspect is that the 3 most important features in the case of SVM lean towards the positive class, which can contribute to an increase in uncertainty, since the true class is the negative one. EBM found that the most important feature is the extension

Table 2: Evaluation metrics on the baseline dataset (all features), for LR, RF, SVM, and EBM. The best result is in boldface.

	No tunning				Parameter tunning			
Metric	**LR**	**RF**	**SVM**	**EBM**	**LR**	**RF**	**SVM**	**EBM**
Accuracy	0.79 ± 0.06	0.88 ± 0.07	0.74 ± 0.07	**0.89 ± 0.06**	0.81 ± 0.08	**0.91 ± 0.04**	0.82 ± 0.09	**0.91 ± 0.04**
False-positive rate	0.19 ± 0.10	0.12 ± 0.09	0.24 ± 0.10	**0.09 ± 0.07**	0.17 ± 0.09	0.08 ± 0.09	0.13 ± 0.11	**0.07 ± 0.07**
False-negative rate	0.23 ± 0.11	**0.13 ± 0.10**	0.27 ± 0.14	**0.13 ± 0.08**	0.22 ± 0.10	**0.10 ± 0.06**	0.23 ± 0.09	0.11 ± 0.07
Precision	0.82 ± 0.08	0.89 ± 0.08	0.77 ± 0.08	**0.91 ± 0.06**	0.83 ± 0.09	**0.93 ± 0.07**	0.87 ± 0.11	**0.93 ± 0.06**
Recall	0.77 ± 0.11	**0.87 ± 0.10**	0.73 ± 0.14	**0.87 ± 0.08**	0.78 ± 0.10	**0.90 ± 0.06**	0.77 ± 0.09	0.89 ± 0.07
F1-score	0.78 ± 0.07	0.88 ± 0.07	0.74 ± 0.09	**0.89 ± 0.06**	0.80 ± 0.09	**0.91 ± 0.04**	0.81 ± 0.09	**0.91 ± 0.04**

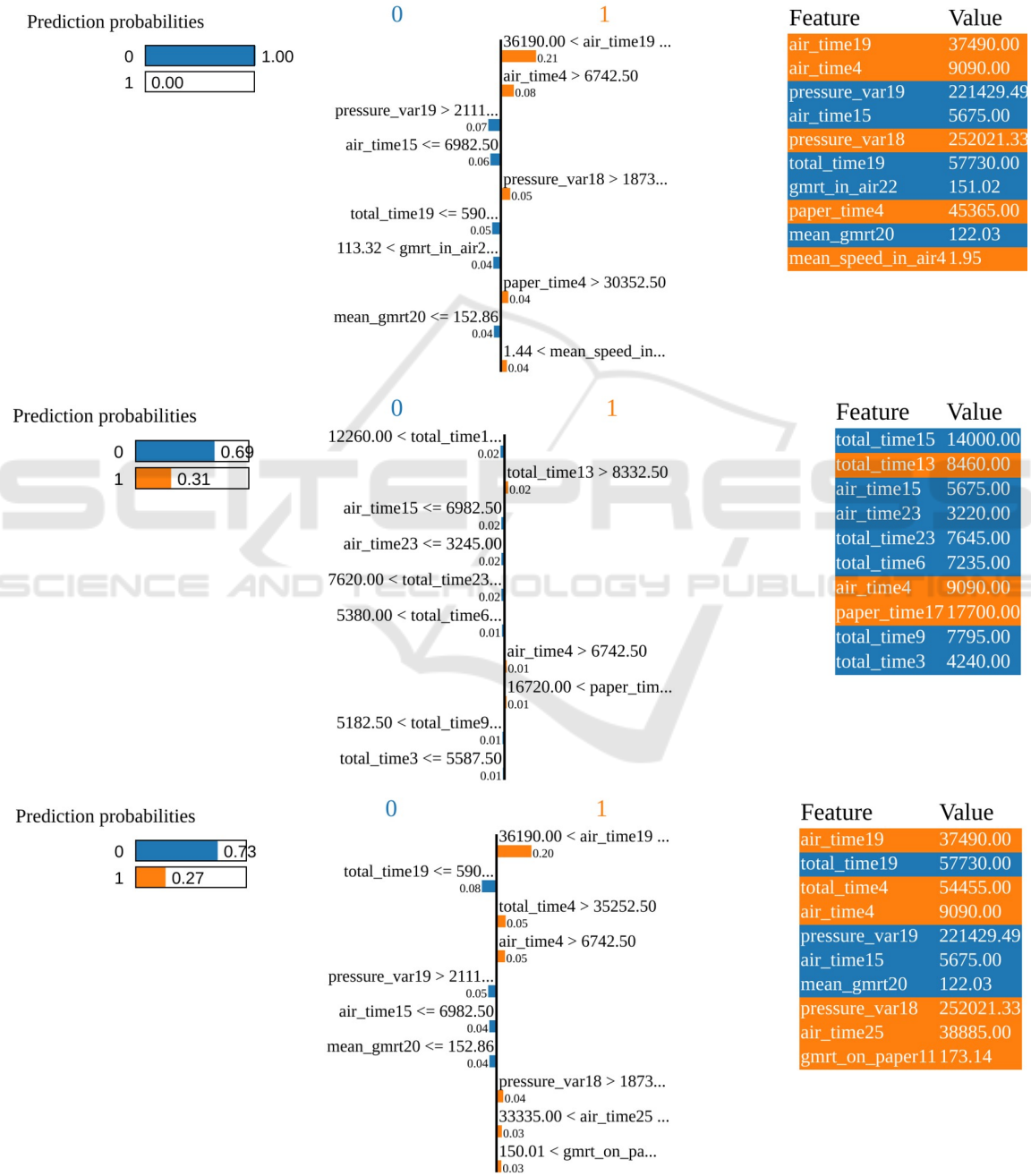

Figure 4: Local LIME explanations for the LR (top), RF (middle), and SVM (bottom) classifiers.

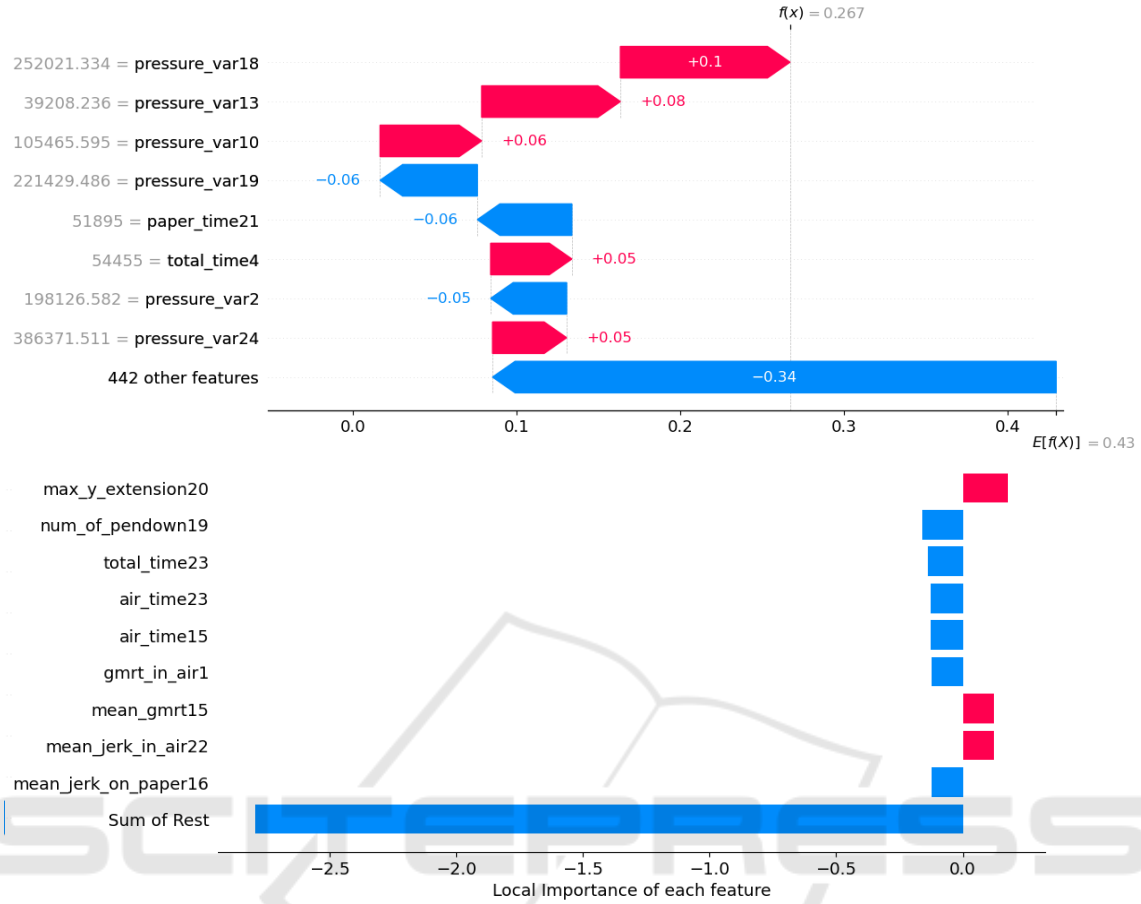

Figure 5: Local explanations provided by SHAP for the SVM (top) classifier and EBM (bottom).

in y of task 20, contributing to the positive class. The "num_of_pendown" group of features also seem to be relevant. This also identifies importance in the total time to perform a task and the time the pen is off the paper. In general, the sum of the importance of characteristics that are not considered most important has a greater impact than the most important characteristics individually, and it is not just the most relevant characteristics that influence explanations, but rather their complete set.

4.4 Global Explainability Results

We also present the application of SHAP and EBM as global explainers, as illustrated in Figure 6.

The models that used SHAP to extract explanations present two-color bars symbolizing each class. In binary classification, the bars of both classes often have the same dimension, that is, the features contribute with equal intensity to both classes. The most relevant features at a global level are similar to the most relevant at a local level, with the group of characteristics "pressure_var" predominating as being the most relevant with 13 appearances in both. In second place in terms of number of appearances were the groups "air_time" and "total_time" both with 3. In SVM, the group "pressure_var" also predominates with 13 appearances, with the group "total_time" appearing 5 times, ranking second in terms of frequency.

The predominant group for EBM is also "total_time" with 7 appearances, although the most important feature is "air_time23" which belongs to the second most frequent group with 6 appearances. Something that all models seem to agree on is that the "total_time" group has significant importance at a global level, which is intuitive given that a person with impaired motor capacity, which is one of the symptoms of Alzheimer's, will take more time to complete tasks. Furthermore, tasks 17 and 19 appear in all global explanations, which may indicate that they have a greater ability to detect the presence of this disease.

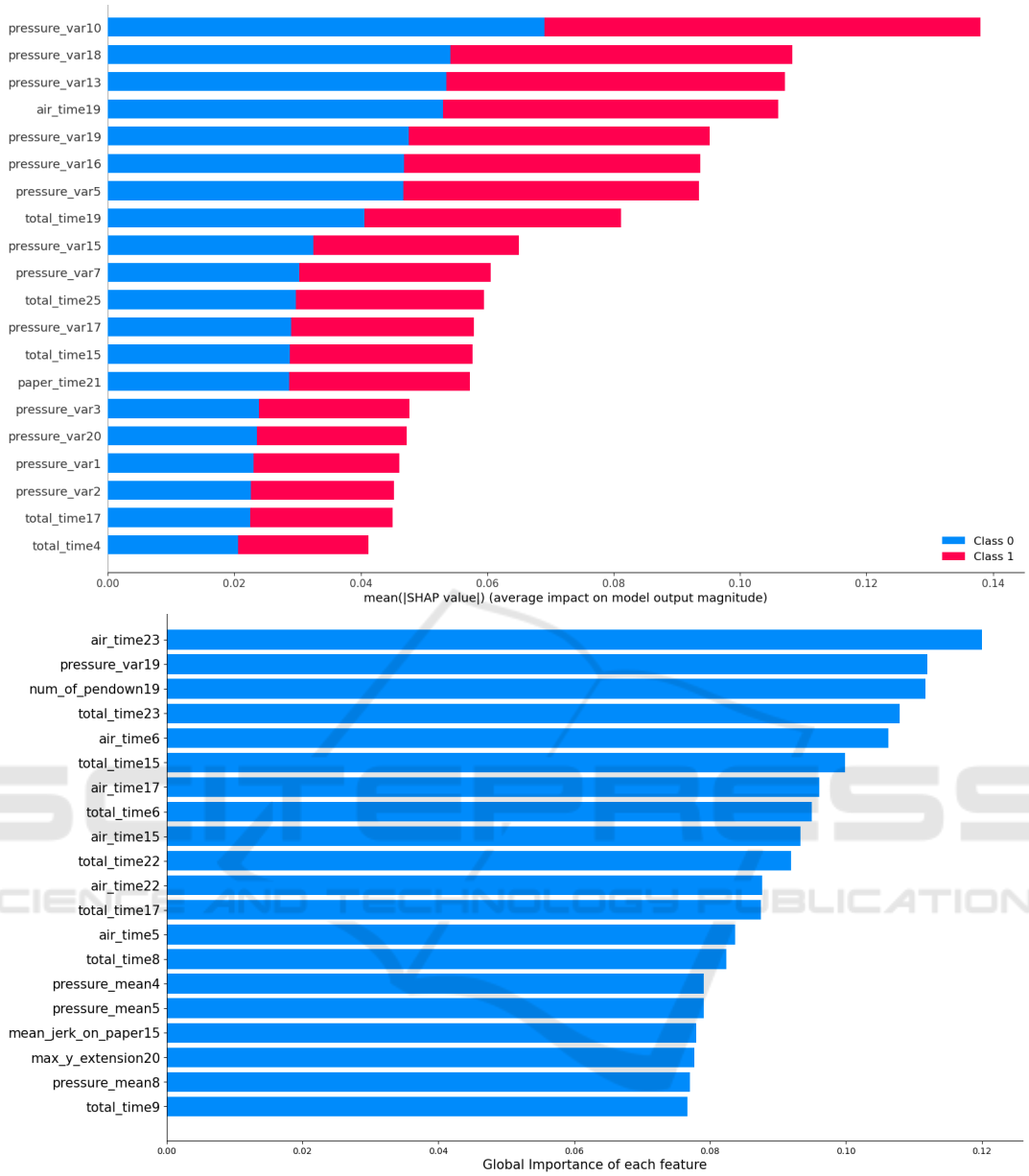

Figure 6: Global explanations provided by SHAP, for the SVM (top) classifier and EBM (bottom).

4.5 Dimensionality Reduction

We now report on the experimental results of the dimensionality reduction evaluation with KFFS. We have applied FS as a pre-processing stage to the dataset; the reduced dimensionality data is applied to the pipelines depicted in Figure 2 and Figure 3.

We have considered KFFS with $k = 10$ folds and $T_h \in \{70, 80, 90, 95\}$. As filters for KFFS, we have FiR and FCBF techniques described in section 2.4; we also considered the *union* and the *intersection* of the feature subspaces provided by these filters to assess on how differentiated these subspaces are (e.g., a small number of features in the intersection implies large differentiation). Table 3 reports the reduced dimensionality of the dataset, denoted as m, with KFFS, for the mentioned filters and thresholds. We observe that the increase of the T_h parameter yields a decrease on the dimensionality of the dataset. These threshold values achieve a significant degree of reduction on the dimensionality of the dataset. Another important aspect of these results is that the intersection of the feature subspaces provided by both filters yields a very reduced version of the dataset.

Table 3: Dimensionality reduction with KFFS. The reduced dimensionality m, from the original dimensionality $d = 450$.

	KFFS Filter, k = 10			
T_h	FiR	FCBF	Union	Inter.
70	25	34	53	6
80	23	28	45	6
90	22	23	40	5
95	22	21	39	4

Table 4: Confusion matrices for EBM with: $T_h = 70$, $m = 34$ features (left); $T_h = 80$, $m = 28$ features (right).

EBM	
14	1
2	18

EBM	
14	1
2	18

Table 4 shows the confusion matrices, for the EBM model with the KFFS(FCBF) filter with $k = 10$ and $T_h \in \{70, 80\}$. As compared to Table 1, we have a false negative reduction from 3 to 2; as a consequence, the accuracy of the classifier is improved.

Table 5 reports the values for the same metrics as in Table 2, considering the reduced dimensionality dataset. We observe that on the reduced dimensionality dataset, the EBM technique achieves the best results, with a large reduction of the number of features.

Table 6 reports a similar experimental evaluation as in Table 5, using the KFFS(FiR) filter. The RF and EBM classifiers perform very similarly, with 1% difference on both cases.

Figure 7 depicts the local and global SHAP explanations for the EBM classifier on KFFS(FCBF) reduced space (the best performing in Table 5). Figure 8 shows the local and global SHAP explanations for the RF classifier on KFFS(FiR) reduced space (the best performing in Table 6). For both cases, the "air_time" group of features is usually ranked at the top.

4.6 Discussion

The experimental evaluation in Cilia et al. (2022) addresses all available features for the classification task. It was found that, in almost all cases, the tasks analyzed separately obtain a lower success rate than considering all tasks. This may prove the relevance of carrying out different tasks, as they test different aspects relevant to the detection of Alzheimer's. The reported accuracy rates are 81.86 % (±7.20), 88.29 % (±4.90), and 79.00 % (±7.55), for LR, RF, and SVM, respectively. We got an improvement with the RF and SVM classifiers, after parameter tuning. Regarding sensitivity, the following values were presented: 84.17 %, 90.28 %, and 77.50 %. In this case, worse results were achieved for LR, and similar results for RF and SVM.

Using local and global explanation techniques on the baseline version of the dataset (the original number of features), we were able to find consistent results and to identify the groups of features that are the most meaningful. Out of the set of 450 features, we have conclude that the "air_time", "total_time", and "pressure_var" groups of features seem to be carry the decisive information for the detection and prediction of Alzheimer disease.

The use of feature selection with the KFFS algorithm yielded significant reduction on the dimensionality of the dataset, keeping or improving the classification performance. This implies that only a small subset of features conveys information about the detection of Alzheimer disease. The use of explainability techniques on the reduced dimensionality dataset has revealed that the "air_time" group of features, that is, the time that the patient takes to perform some writing tasks is the most relevant to detect the disease.

5 CONCLUSIONS

Neurodegenerative diseases impose a progressive decrease of the patient cognitive and motor skills. Their cause is yet to be determined, thus the early diagnosis is a key factor to delay their progression and symptoms. In this context, ML and XAI approaches have been proposed to devise diagnostic and prediction systems.

In this paper, we described an *explainable ML* system that is able to detect Alzheimer's disease through the analysis of specific handwriting tasks, with the DARWIN dataset. We have confirmed and improved previous results, showing the relevance of carrying out different writing tasks, as they test several aspects relevant to the detection of Alzheimer's. We have assessed the effect of applying hyperparameter fine-tuning and we noticed an improvement on the models performance. Running fine-tuning optimization, at the cost of running time, we have improved the classification results, namely the false negative rate. Moreover, we have also performed experiments using different local and global explainability techniques on the original dataset and on the reduced dimensionality dataset. From these experiments, we identify the groups of features that are the most decisive to Alzheimer detection from handwriting. In detail, from the original set of 450 features, we found that the time spent to perform the in-air movements, plays a decisive role at predicting Alzheimer.

As future work, we aim to fine tune the parameters of feature selection techniques, to improve the classi-

Table 5: Evaluation metrics on the reduced dimensionality dataset, for LR, RF, SVM, and EBM. We use KFFS(FCBF) filter with $k = 10$ and $T_h \in \{70, 80\}$. The best result is in boldface.

KFFS(FCBF), k = 10	**$T_h = 70, m = 34$**				**$T_h = 80, m = 28$**			
Metric	**LR**	**RF**	**SVM**	**EBM**	**LR**	**RF**	**SVM**	**EBM**
Accuracy	0.85 ± 0.07	0.85 ± 0.07	0.76 ± 0.07	$\mathbf{0.89 \pm 0.05}$	0.80 ± 0.07	0.86 ± 0.04	0.77 ± 0.08	$\mathbf{0.89 \pm 0.04}$
False-positive rate	0.18 ± 0.09	0.20 ± 0.11	0.19 ± 0.06	$\mathbf{0.11 \pm 0.07}$	0.21 ± 0.10	0.20 ± 0.08	0.20 ± 0.07	$\mathbf{0.13 \pm 0.06}$
False-negative rate	0.13 ± 0.11	$\mathbf{0.11 \pm 0.09}$	0.29 ± 0.11	$\mathbf{0.11 \pm 0.09}$	0.20 ± 0.09	0.09 ± 0.07	0.26 ± 0.11	$\mathbf{0.08 \pm 0.07}$
Precision	0.84 ± 0.07	0.83 ± 0.08	0.80 ± 0.07	$\mathbf{0.90 \pm 0.06}$	0.81 ± 0.08	0.84 ± 0.05	0.80 ± 0.07	$\mathbf{0.88 \pm 0.05}$
Recall	0.87 ± 0.11	$\mathbf{0.89 \pm 0.09}$	0.71 ± 0.11	$\mathbf{0.89 \pm 0.09}$	0.80 ± 0.09	0.91 ± 0.07	0.74 ± 0.11	$\mathbf{0.92 \pm 0.07}$
F1-score	0.85 ± 0.07	0.86 ± 0.06	0.75 ± 0.08	$\mathbf{0.89 \pm 0.06}$	0.80 ± 0.07	0.87 ± 0.04	0.77 ± 0.09	$\mathbf{0.90 \pm 0.05}$

Table 6: Evaluation metrics on the reduced dimensionality dataset, for LR, RF, SVM, and EBM. We use KFFS(FiR) filter with $k = 10$ and $T_h \in \{70, 80\}$. The best result is in boldface.

KFFS(FiR), k = 10	**$T_h = 70, m = 25$**				**$T_h = 80, m = 23$**			
Metric	**LR**	**RF**	**SVM**	**EBM**	**LR**	**RF**	**SVM**	**EBM**
Accuracy	0.77 ± 0.08	0.83 ± 0.06	0.76 ± 0.05	$\mathbf{0.84 \pm 0.06}$	0.77 ± 0.08	0.82 ± 0.06	0.76 ± 0.05	$\mathbf{0.83 \pm 0.06}$
False-positive rate	0.22 ± 0.12	0.21 ± 0.11	$\mathbf{0.15 \pm 0.10}$	0.17 ± 0.10	0.22 ± 0.13	0.21 ± 0.11	$\mathbf{0.16 \pm 0.08}$	0.18 ± 0.10
False-negative rate	0.24 ± 0.08	$\mathbf{0.13 \pm 0.06}$	0.33 ± 0.11	0.15 ± 0.08	0.25 ± 0.08	$\mathbf{0.15 \pm 0.05}$	0.31 ± 0.09	0.17 ± 0.07
Precision	0.80 ± 0.11	0.83 ± 0.08	0.84 ± 0.09	$\mathbf{0.85 \pm 0.08}$	0.80 ± 0.11	0.82 ± 0.08	0.83 ± 0.07	$\mathbf{0.84 \pm 0.08}$
Recall	0.76 ± 0.08	$\mathbf{0.87 \pm 0.06}$	0.67 ± 0.11	0.85 ± 0.08	0.75 ± 0.08	$\mathbf{0.85 \pm 0.05}$	0.69 ± 0.09	0.83 ± 0.07
F1-score	0.77 ± 0.07	$\mathbf{0.84 \pm 0.05}$	0.74 ± 0.07	$\mathbf{0.84 \pm 0.06}$	0.77 ± 0.07	$\mathbf{0.83 \pm 0.05}$	0.75 ± 0.06	$\mathbf{0.83 \pm 0.05}$

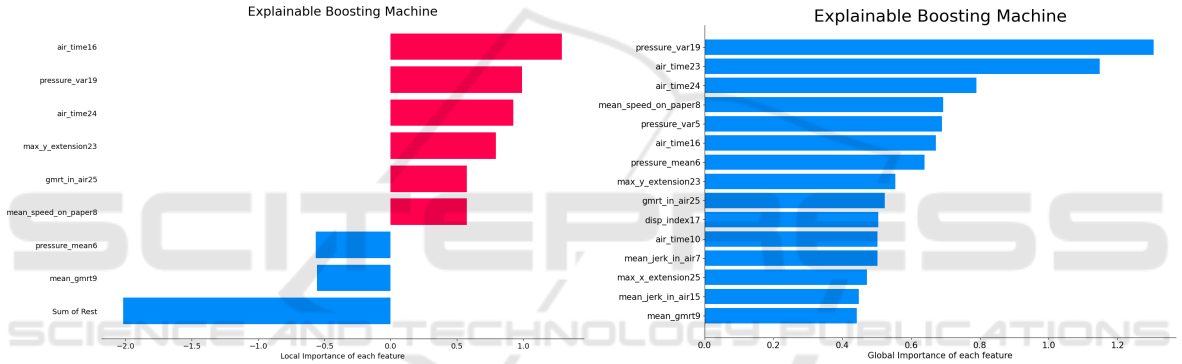

Figure 7: Local and global SHAP explanations for the EBM classifier on KFFS(FCBF) reduced space.

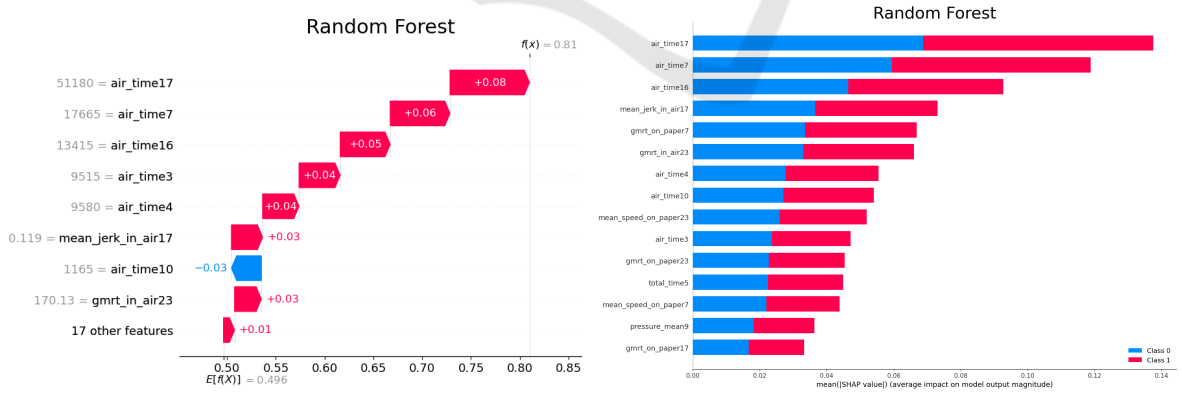

Figure 8: Local and global SHAP explanations for the RF classifier on KFFS(FiR) reduced space.

fication metrics minimizing the false negatives as well as the number of features.

ACKNOWLEDGEMENTS

This research was supported by Instituto Politécnico de Lisboa (IPL) under Grant IPL/IDI&CA2024/ML4EP_ISEL.

REFERENCES

Azzali, I., Cilia, N. D., De Stefano, C., Fontanella, F., Giacobini, M., and Vanneschi, L. (2024). Automatic feature extraction with vectorial genetic programming for Alzheimer's disease prediction through handwriting analysis. *Swarm and Evolutionary Computation*, 87:101571.

Bastani, O., Kim, C., and Bastani, H. (2017). Interpreting blackbox models via model extraction. *ArXiv*, abs/1705.08504.

Bazarbekov, I., Razaque, A., Ipalakova, M., Yoo, J., Assipova, Z., and Almisreb, A. (2024). A review of artificial intelligence methods for alzheimer's disease diagnosis: Insights from neuroimaging to sensor data analysis. *Biomedical Signal Processing and Control*, 92:106023.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Cappiello, G. and Caruso, F. (2024). Quantum ai for alzheimer's disease early screening.

Cilia, N. D., De Gregorio, G., De Stefano, C., Fontanella, F., Marcelli, A., and Parziale, A. (2022). Diagnosing Alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking. *Engineering Applications of Artificial Intelligence*, 111:104822.

Cilia, N. D., De Stefano, C., Fontanella, F., and Siniscalchi, S. M. (2024). How word semantics and phonology affect handwriting of alzheimer's patients: A machine learning based analysis. *Computers in Biology and Medicine*, 169:107891.

Cover, T. and Thomas, J. (2006). *Elements of information theory*. John Wiley & Sons, second edition.

David W. Hosmer Jr., Rodney X. Sturdivant, S. L. (2013). *Applied Logistic Regression*. 3rd edition.

Erdogmus, P. and Kabakus, A. T. (2023). The promise of convolutional neural networks for the early diagnosis of the alzheimer's disease. *Engineering Applications of Artificial Intelligence*, 123:106254.

Ferreira, A. and Figueiredo, M. (2023). Leveraging explainability with k-fold feature selection. In *12th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pages 458–465.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viégas, F., and Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In Dy, J. G. and Krause, A., editors, *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2673–2682. PMLR.

Koenig, L. N., Day, G. S., Salter, A., Keefe, S., Marple, L. M., Long, J., LaMontagne, P., Massoumzadeh, P., Snider, B. J., Kanthamneni, M., Raji, C. A., Ghoshal, N., Gordon, B. A., Miller-Thomas, M., Morris, J. C., Shimony, J. S., and Benzinger, T. L. (2020). Select atrophied regions in Alzheimer disease (sara): An improved volumetric model for identifying Alzheimer disease dementia. *NeuroImage: Clinical*, 26:102248.

Kumar, S. and Shastri, S. (2022). Alzheimer MRI preprocessed dataset.

Lakkaraju, H. and Bastani, O. (2020). How do I fool you? manipulating user trust via misleading black box explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013a). Accurate intelligible models with pairwise interactions. In Dhillon, I. S., Koren, Y., Ghani, R., Senator, T. E., Bradley, P., Parekh, R., He, J., Grossman, R. L., and Uthurusamy, R., editors, *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD, Chicago, IL, USA*, pages 623–631. ACM.

Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013b). Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 623–631, New York, NY, USA. Association for Computing Machinery.

Mitra, U. and Rehman, S. U. (2024). Ml-powered handwriting analysis for early detection of alzheimer's disease. *IEEE Access*, 12:69031–69050.

Mothilal, R., Sharma, A., and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617. ACM.

Rainio, O., Teuho, J., and Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016a). "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*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016b). Why should I trust you? explaining the predictions of any classifier. In *HLT-NAACL Demos*, pages 97–101. The Association for Computational Linguistics.

Scheda, R. and Diciotti, S. (2022). Explanations of machine learning models in repeated nested cross-validation: An application in age prediction using brain complexity features. *Applied Sciences*, 12(13).

Szepannek, G. and Lübke, K. (2022). Explaining artificial intelligence with care. *KI - Künstliche Intelligenz*.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer-Verlag.

Yu, L. and Liu, H. (2003). Feature selection for high-dimensional data: a fast correlation-based filter solution. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 856–863.

Yu, L. and Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research (JMLR)*, 5:1205–1224.