# Developing an Interpretable Machine Learning Model for Brain Health Research: Phase 3

*Natasha Rinta Dunstan*

Master of Science
Data Science
School of Informatics
University of Edinburgh
2022

# Abstract

Globally, over 55 million people live with dementia, of which Alzheimer's Disease (AD) accounts for the majority of dementia diagnoses. The onset of AD could be delayed through early detection and greater understanding of the risk factors associated with the disease. While existing machine learning-based AD risk prediction models have achieved promising results, their clinical utility is restricted by an over-reliance on diagnostic imaging and biomarker sampling data, low model interpretability, and poor generalisability. By employing a transfer learning framework, our work aims to circumvent these limitations, facilitating AD risk prediction in younger, asymptomatic populations. Our best-performing source model demonstrated a sensitivity of 71.72%, specificity of 89.26%, AUROC of 86.90% and GA of 80.01 %, with our target model achieving sensitivity 47.06%, specificity 57.14%, GA 51.86%, and AUROC 50.84%. Our findings from feature importance analysis concur with previous reports of important modifiable AD risk factors, and suggest the presence of thyroid disease and kidney disease as putative AD risk factors. We propose that our model could potentially be utilised as a population-wide screening tool for prediction of future AD risk, and thus contribute to lowering the disease burden of AD.

# Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Natasha Rinta Dunstan*)

# Acknowledgements

I would like to express my heartfelt gratitude to my supervisor, Dr Samuel Danso, who provided me with much mentorship and technical insight throughout my time working on this project; to the faculty members of the MSc Data Science course, who equipped me with many of the skills necessary to complete my project; to Yu Cheng Foo, medical student at Edinburgh Medical School, who provided clinical perspective and assistance with proofreading; and lastly, to my family, for their unconditional love and support.

# Table of Contents

# Chapter 1

# Introduction

## 1.1 Dementia and Alzheimer's Disease

Dementia is a clinical syndrome associated with progressive, irreversible cognitive decline – affecting memory, comprehension, ability to learn, social functioning, and language skills, leading to increasing loss of independence and function [1]. Worldwide, dementia is thought to affect 10 million new patients annually, with global prevalence estimated to reach 78 million by the year 2030 [18]. In addition to being the seventh leading cause of death globally, its socioeconomic burden has been estimated at US$ 1.3 trillion in 2019, with figures set to double by 2030 [1]. While there are a number of causes of dementia, Alzheimer's disease (AD) is by far the commonest, accounting for 60-70% of all cases [1].

The pathophysiology of AD is characterised by the formation of abnormal protein aggregates, known as "plaques" or "tangles", in the nerve cells (or "neurons") of the brain. These abnormal protein aggregates disrupt cellular processes, leading to neuronal damage, and inhibiting normal neurotransmission [3]. This cycle of abnormal protein aggregate formation and neuronal damage continues over a period of years to decades, resulting in a progressive loss of brain tissue. While there is currently no cure for AD, early diagnosis and intervention could delay AD onset and slow disease progression. As the brain changes seen in AD often precede dementia onset by more than 20 years [2], identifying potential risk factors could potentially inform strategies to reduce the risk of developing AD in the population.

## 1.2 Adopting Machine Learning for Alzheimer's Disease Risk Prediction

Recent advancements in computational approaches like machine learning (ML) have made it possible to personalise patient management plans, and have been adapted for use in detecting heart diseases, breast cancer and even real-time prediction of septic shock [4, 15, 49]. Given the success of these medical applications, the development of ML-based AD risk prediction models have become increasingly popular in the recent years [20, 29], as they are able to capture complex interactions of risk factors, a clear advantage over traditional statistical methods [19].

Current approaches to developing ML-based AD risk prediction models suffer from a number of limitations that restrict their clinical utility [19, 34], including lack of model interpretability (particularly in "black-box" models) [30, 38], the absence of external validation, overfitting, poor model generalisability, and an over-reliance on data acquired through diagnostic neuroimaging (e.g. Magnetic Resonance Imaging MRI brain scans) and biomarker sampling techniques (e.g. cerebrospinal fluid sampling) [19, 34, 43].

In an attempt to circumvent these challenges, our work aimed to build on the transfer learning framework for AD risk prediction, as proposed by Danso et al. [14]. We hypothesise that the adoption of a transfer learning framework will efficiently improve model generalisability, as it avoids training multiple models from scratch on different datasets [32, 47]. Since no ML-based AD risk prediction models have been developed using the EPAD dataset [39], we adopted this dataset as our source domain, predicting that novel relationships between AD risk factors could potentially be uncovered from this data. Upon developing the source model on the EPAD dataset, the knowledge learned by this model was subsequently "transferred" for the development of a target model, which relied on the PREVENT dataset [37] as our target domain. This enabled us to both assess and enhance model generalisability to a younger and undiagnosed population drawn from the PREVENT dataset (i.e. the target domain).

For our study, we relied on two tree-based classifiers, namely the Random Forest (RF) and the XGBoost algorithm. While both classifiers are considered ensemble-based algorithms (i.e. algorithms that combine multiple weak learners to form a strong learner), they differ in terms of the learning strategy employed. For instance, RF relies on the concept of bagging, that is by training many decision trees on different bootstrapped samples (i.e. sampling with replacement) of the data [8]. Meanwhile, the

XGBoost algorithm is based on the gradient boosting technique, whereby decision trees are sequentially fitted and added to the ensemble, with each newly added decision tree being trained to correct the prediction errors made by the prior tree [10].

Furthermore, to address existing concerns over model explainability, we explored effective approaches for visualising risk factors that drive the predictions made by our models.

## 1.3 Rationale and Significance

Considering the growing socioeconomic and health burden associated with AD, it is crucial that timely efforts to mitigate its effects are initiated. Through developing an intuitive ML model for predicting AD risk within asymptomatic individuals, we aim to facilitate the identification of patients at high risk of developing AD, as well as the identification of modifiable risk factors that could potentially reduce the risk of developing AD. Equally significant is our efforts to generate a highly-interpretable model, which we envision will provide clinical insight and engender trust in algorithmic decision-making among clinicians.

The development of a ML model trained solely on patient socio-demographic data, without relying on data from advanced diagnostic techniques such as neuroimaging or biomarker sampling, lends itself well for use as a population-wide screening tool for high AD risk. While common AD diagnostic techniques such as MRI scanning or cerebrospinal fluid sampling are effective at picking up patients with early-stage AD that do not display overt clinical symptoms [6, 7, 45], the relatively high cost of performing these interventions and their side effect profile make it impractical to apply these techniques for population-wide screening of asymptomatic patients, resulting in missed opportunities for early intervention in AD. Thus, since the majority of asymptomatic patients will not routinely receive MRI scans or undergo CSF sampling for AD risk stratification [36, 14], an alternative AD risk prediction model that does not rely on biomarkers or medical imaging would enable a sufficiently large number of potential AD cases to be picked up, enabling clinicians to initiate disease-modifying interventions earlier in the disease course.

# Chapter 2

# Materials and Methods

## 2.1 Machine learning framework for AD risk prediction

As mentioned previously, the development of AD is known to precede the onset of clinical symptoms and behavioural signs by decades, resulting in considerable difficulty in the prevention and treatment of AD. Nevertheless, this suggests that there is a substantial window of opportunity for which early intervention could be taken to delay AD onset and slow disease progression. The aim of this project was to thus develop a machine learning-based AD risk prediction model, capable of identifying individuals at increased risk of developing AD given a range of characteristics or risk factors (i.e. features), decades before the onset of clinically-apparent AD. The problem was formulated as a binary classification task which involved the prediction two discrete labels ("High-Risk" vs. "Low-Risk").

We employed a transfer learning framework in developing our AD risk prediction model. As traditional ML methods rely on the assumption that the training and test data share the same feature space and data distribution, such models may suffer from a decline in predictive performance when presented with new data with different distributions to the training set, resulting in an inability to adapt to novel scenarios. A potential solution to this would be to retrain the model from scratch using newly collected training data that reflects the new distribution. In most real-world applications, as the collection of sufficient data for training is often expensive, a more efficient approach would be to leverage knowledge learned by the original model and readapt it for the new data distribution. By adopting a transfer learning framework, we were able to further relax the identical distribution assumption, thus enabling the generalisation of our model across two different patient populations. This is illustrated in Figure

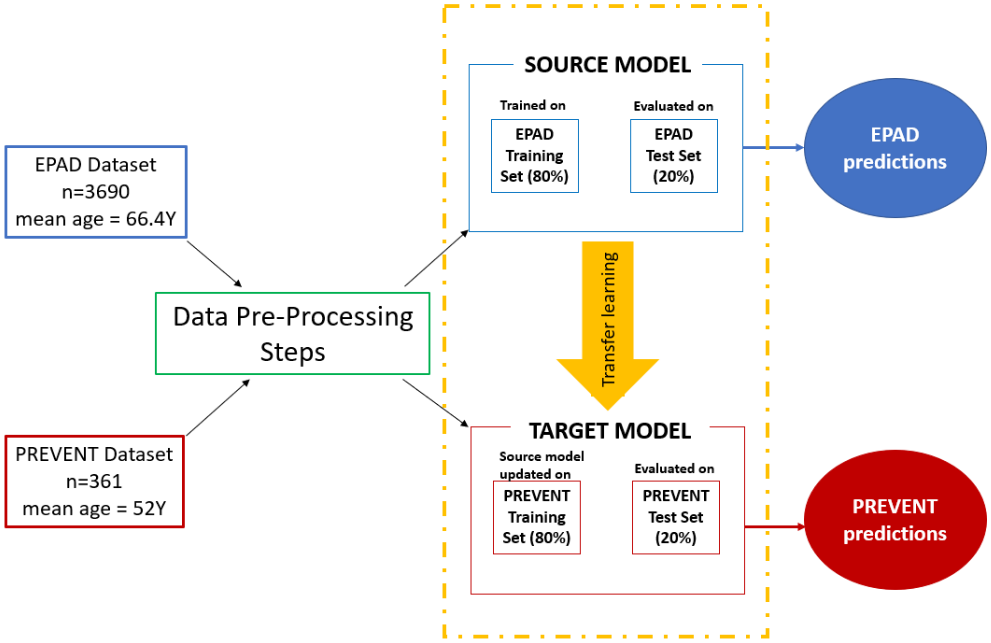


Figure 2.1: Figure depicting transfer learning framework (adapted from Danso et al. [14])

2.1, whereby a "source model" was first trained on the EPAD cohort data (i.e. the "source domain") [39], and later adapted for building the "target model" which is able to predict AD risk in the PREVENT cohort data (i.e. the "target domain"). Details of the PREVENT and EPAD datasets are discussed in the following section.

## 2.2 Data Description

In developing a transfer learning framework, we mainly relied on two data sources: (i) the EPAD LCS V.IMI dataset [39], derived from the Longitudinal Cohort Study (LCS) component of the European Prevention of Alzheimer's Dementia (EPAD) project, aimed at developing an environment for testing different interventions for secondary prevention of AD, and (ii) the PREVENT baseline dataset [37], curated as part of the PREVENT Dementia programme to aid research into mid-life risk factors of AD. As both the EPAD and PREVENT studies are related to dementia research, they share similarities in terms of the types of variables collected, including socio-demographic information, genetics, cognitive and neuroimaging outcomes, as well as lifestyle and behavioural characteristics. However, the varying rationales and objectives of both studies have resulted in differences in terms of their data distribution, making transfer learning an ideal tool for the assessment of model generalisability . Notably, the EPAD population consists of individuals drawn across Europe with a mean age of 66 years, whereas the PREVENT dataset is limited to a population drawn from the UK and Ireland with a relatively lower mean age (52 years). Furthermore, some individuals from the EPAD cohort have been diagnosed with AD while the PREVENT cohort only consists of individuals who are without an AD diagnosis.

## 2.3 Data Preprocessing

The EPAD study followed up on participants annually, covering the period between 2016 to 2019. By merging the data of individuals across each visit, we were able to expand our sample size from 2096 unique individuals to 3690 observations covering 3 years of follow-up data. In contrast, the PREVENT dataset being used is the baseline data involving 700 observations (before accounting for missing values), covering the period between February 2014 to October 2018.

### 2.3.1 Feature Engineering

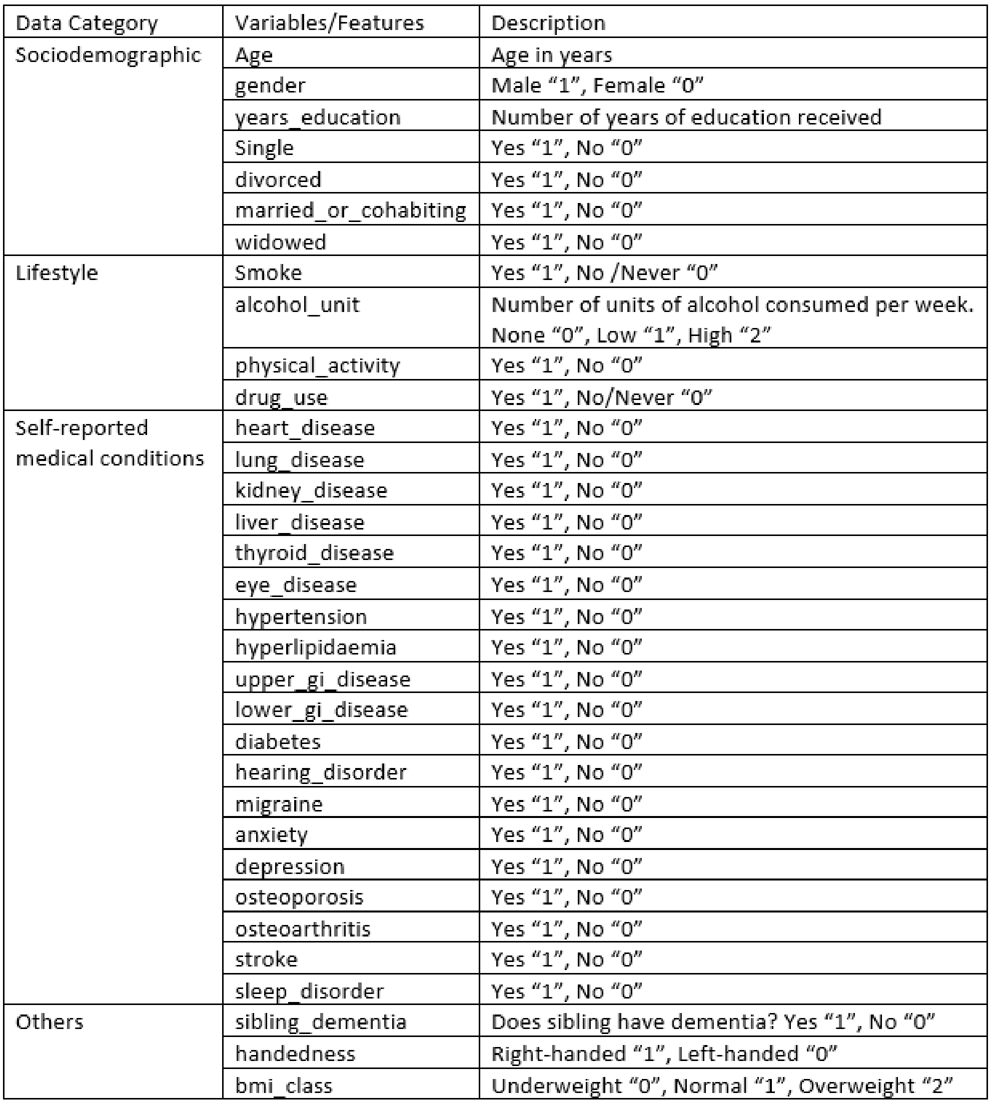
Transfer learning can be categorised depending on the similarity of the source and target domains: heterogeneous transfer learning refers to the case when the feature spaces of the source and target domains are different, whereas homogeneous transfer learning is applied when the source and target domains share the same feature space. Our study involved the latter, as the EPAD and PREVENT datasets overlap considerably in their recorded features or variables, typical of AD risk stratification studies. To enable homogeneous transfer learning, we first identified common features shared between both datasets. Excluding any neuroimaging and cognitive assessment outcomes, 33 features have been identified altogether, as summarised in Table 2.1. Furthermore, due to some differences in the format that common variables were recorded in – for example, the same variable could be encoded as categorical levels in EPAD but as numerical values in PREVENT – further preprocessing steps were employed to ensure a unified representation. Excluding the features corresponding to age and number of years of education, all other features were encoded to either have binary or ordinal feature representation. For instance, binary encoding was applied to all medical history features, with "0" and "1" indicating the absence and presence of a medical condition respectively. Similarly, features such as "handedness", "gender", "smoke", "drug use", "sibling dementia", and "physical activity" all had two categorical levels and thus processed to have binary feature representation, as shown in Table X. For the feature "marital status", there were altogether four categories including "single", "married or cohabitating", "divorced" and "widowed", with each being represented with binary encoding "1" and "0" based on a response of "yes" and "no" respectively.

Ordinal encoding was used to represent features such as "BMI class" and "Alcohol unit", so as to retain the natural ordering or ranking inherent to the categories of these features. Consequently, the Body Mass Index (BMI) was processed to have four ordinal levels "0", "1", "2", and "3", respectively representing each of the following categories in accordance with WHO classification [50]: underweight ($<$18.5 kg/m$^2$), normal (18.5–24.9 kg/m$^2$), overweight (25–29.9 kg/m$^2$), and obese ($>$30 kg/m$^2$). The feature "alcohol unit" represents the number of units of alcohol intake per week based on UK standards, and we have introduced three categorical levels: Never (0 units), Low (1-14 units), High ($\geq$14 units) which were encoded as "0", "1" and "2" respectively. This decision was made based on the recommended alcohol intake suggested by the NHS [31], whereby an alcohol intake of 14 or more units per week is considered excessive.

Finally, further preprocessing procedures were applied to derive the outcome variable "AD risk". As mentioned previously, the EPAD study collected data on whether a participant had been diagnosed with AD or not - however, this information was absent in the PREVENT dataset. Furthermore, as participants who were diagnosed with AD were subsequently dropped from the EPAD study, only 30 individuals out of the 3690 observations were found to have an AD diagnosis, thus constituting less than 1% of the dataset. Therefore, to avoid the severe class imbalance as well as ensure consistency across both datasets, we employed a classification scheme similar to that of Danso et al. [14] that did not rely on AD diagnosis, but instead classified individuals into "high AD risk" or "low AD risk". Ritchie et al. [37] previously described the stratification of individual AD risk based on the Apolipoprotein E 4 (ApoE 4) genotype of each individual as well as their parental history of dementia. As this information was collected in both studies, we were able to derive the outcome variable as follows: Individuals with a parental dementia diagnosis and ApoE 4 genotype are labelled as "High-Risk", whereas all other individuals are labelled as "Low-Risk".

### 2.3.2 Handling of Missing Data

As with any real-world dataset, both the EPAD and PREVENT data contained missing values. The proportion of missing values for each variable are provided in Table 2.2. Before training the source model, a machine-learning based data imputation technique



| Data Category | Variables/Features | Description |
|---|---|---|
| Sociodemographic | Age | Age in years |
| | gender | Male “1”, Female “0” |
| | years_education | Number of years of education received |
| | Single | Yes “1”, No “0” |
| | divorced | Yes “1”, No “0” |
| | married_or_cohabiting | Yes “1”, No “0” |
| | widowed | Yes “1”, No “0” |
| Lifestyle | Smoke | Yes “1”, No /Never “0” |
| | alcohol_unit | Number of units of alcohol consumed per week. None “0”, Low “1”, High “2” |
| | physical_activity | Yes “1”, No “0” |
| | drug_use | Yes “1”, No/Never “0” |
| Self-reported medical conditions | heart_disease | Yes “1”, No “0” |
| | lung_disease | Yes “1”, No “0” |
| | kidney_disease | Yes “1”, No “0” |
| | liver_disease | Yes “1”, No “0” |
| | thyroid_disease | Yes “1”, No “0” |
| | eye_disease | Yes “1”, No “0” |
| | hypertension | Yes “1”, No “0” |
| | hyperlipidaemia | Yes “1”, No “0” |
| | upper_gi_disease | Yes “1”, No “0” |
| | lower_gi_disease | Yes “1”, No “0” |
| | diabetes | Yes “1”, No “0” |
| | hearing_disorder | Yes “1”, No “0” |
| | migraine | Yes “1”, No “0” |
| | anxiety | Yes “1”, No “0” |
| | depression | Yes “1”, No “0” |
| | osteoporosis | Yes “1”, No “0” |
| | osteoarthritis | Yes “1”, No “0” |
| | stroke | Yes “1”, No “0” |
| | sleep_disorder | Yes “1”, No “0” |
| Others | sibling_dementia | Does sibling have dementia? Yes “1”, No “0” |
| | handedness | Right-handed “1”, Left-handed “0” |
| | bmi_class | Underweight “0”, Normal “1”, Overweight “2” |

Table 2.1: Common features between the EPAD and PREVENT datasets.

was applied on the EPAD dataset using the MissForest algorithm [40], which is available through the missingpy Python package. This step essentially applied the Random Forest algorithm on the EPAD dataset for missing data imputation in an iterative fashion. For each column with missing values, the MissForest algorithm fitted a Random Forest model using the aforementioned column as the outcome variable, and other columns as the predictors. Meanwhile, missing values in the predictor columns were imputed, either using their means (in the case of numerical features) or modes (in the case of categorical features), allowing for the Random Forest model to be trained on both observed and imputed values. The missing values in the outcome variable were then imputed with the predictions of the fitted Random Forest model. The algorithm iterates through each column in the order of increasing number of missing values, and this process of training and predicting is repeated until a stopping criterion is met, or a specified number iterations is reached.

As the PREVENT dataset was mainly used for updating the source model through transfer learning, we simply applied listwise deletion in handling missing data, i.e. observations with missing values were omitted, resulting in the remaining 361 complete observations used for developing the target model.

| Dataset | Feature with missing values | No. of rows of mising values; Proportion |
|---|---|---|
| EPAD (n=3690) | handedness | 17; 0.46% |
| | years_education | 1; 0.0003% |
| | marital_status | 1; 0.0003% |
| | physical_activity | 56; 1.52% |
| | smoke | 163; 4.42% |
| | drug_use | 163; 4.42% |
| | alcohol_unit | 90; 2.44% |
| | bmi_class | 114; 3.09% |
| PREVENT (n=700 before dropping missing values) | handedness | 239; 34.14% |
| | years_education | 1; 0.14% |
| | marital_status | 2; 0.29% |
| | physical_activity | 5; 0.71% |
| | smoke | 2; 0.29% |
| | drug_use | 3; 0.43% |
| | alcohol_unit | 158; 22.57% |
| | bmi_class | 2; 0.29% |
| | sibling_dementia | 6; 0.86% |
| | hypertension | 2; 0.29% |
| | diabetes | 3; 0.43% |
| | hyperlipidaemia | 1; 0.14% |
| | anxiety | 1; 0.14% |
| | depression | 2; 0.29% |
| | sleep_disorder | 4; 0.57% |
| | eye_disorder | 2; 0.29% |
| | hearing_disorder | 2; 0.29% |

Table 2.2: Study characteristics

### 2.3.3 Handling Class Imbalance

The class distributions for the EPAD and PREVENT datasets are as follows: “High Risk” (26.86%) and “Low-Risk” (73.14%) ; “High Risk” (23%) and “Low-Risk” (77%), respectively. We observe that in both datasets, the proportion of observations corresponding to low AD risk is significantly higher than that of high AD risk, thereby introducing a problem of class imbalance [27]. In many classification tasks, imbalanced classes pose added difficulty as most machine learning algorithms were designed to optimise performance based on an equal number of examples from each class, making it more challenging for the model to learn from the minority class. This tendency to learn from the majority class and disregard examples from the minority class – which is often of more interest to the classification task at hand – frequently results in poor predictive performance [27].

To address this class imbalance, we applied the Synthetic Minority Oversampling Technique (SMOTE) [9], an oversampling approach in which new artificial samples from the minority class were generated. In short, a sample from the minority class alongside two or more of its nearest neighbours are selected at random. This is then used to generate a synthetic sample that lies between the chosen sample and its nearest neighbours. We specifically applied a combination of SMOTE and random undersampling (as described in the original paper by Chawla et al. [9]) as it has been shown to enhance model performance, and the best resampling ratios were determined through hyperparamter tuning. Similar to the missing data imputation steps, these techniques were applied only on the EPAD dataset in developing the source model, but not on the PREVENT dataset.

## 2.4 Building the Prediction Models

In developing the prediction models, we trained four ensemble-based machine learning models with the Random Forest [8] and XGBoost algorithms [10]. In the following sections, we describe the procedures applied for optimising the model parameters and hyperparameters, which involved the use of modelling pipelines and K-fold cross validation.

### 2.4.1 Building Modelling Pipelines

It is generally considered good practice to include all preprocessing, hyperparameter tuning and model training steps as part of a unified modelling pipeline, so as to reduce data leakage, i.e. information from the validation or test sets are used for model training, thus resulting in drastically inflated or otherwise invalid predictive performance [23].

A predefined set of hyperparameter values are necessary for both the Random Forest and XGBoost classification algorithms, with the best configuration subsequently decided through hyperparameter tuning. Additionally, hyperparameter tuning was also required for the MissForest algorithm used in missing data imputation, as well as the SMOTE and random undersampling steps used for class rebalancing. Therefore, in order to enable the tuning of all hyperparameters at once, we combined these preprocessing steps and the final classification step into a single modelling pipeline, as illustrated in the figure below. This yielded two modelling pipelines, rf_pipe and xgb_pipe, each corresponding to the Random Forest and XGBoost classifier respectively, which we had optimised for determining the best source model.

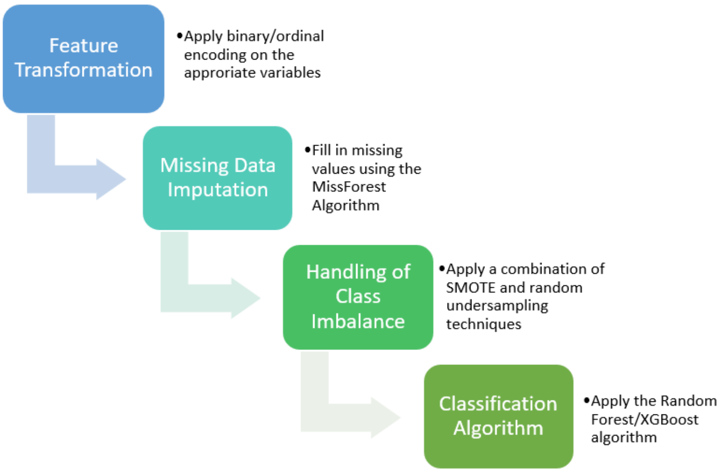

Feature Transformation
•Apply binary/ordinal encoding on the approriate variables
Missing Data Imputation
•Fill in missing values using the MissForest Algorithm
Handling of Class Imbalance
•Apply a combination of SMOTE and random undersampling techniques
Classification Algorithm
•Apply the Random Forest/XGBoost algorithm

Figure 2.2: The modelling pipeline structure of rf_pipe and xgb_pipe

### 2.4.2 Repeated K-fold Cross Validation and Hyperparameter Tuning

In selecting the optimal set of hyperparameters, we applied an approach combining cross-validation (CV) with hold-out [33] which involved randomly splitting the the full dataset into two: the training set (used for model training and hyperparameter tuning) and the test set (used for model evaluation). We chose an 80:20 train-test split ratio, whereby 80% of the observations from the EPAD dataset were part of the training set

(EPAD_train) and the remaining 20% constituted the test set (EPAD_test). The same train-test split ratio was employed for the PREVENT dataset, resulting in the training and test sets which we refer to as PREV_train and PREV_test.

In short, K-fold CV was applied as follows: the training set was further split into K folds and the model was trained K times, each time with a different fold used for validation and the remaining K-1 folds used for training and hyperparameter tuning. Hyperparameters resulting in the lowest validation error when averaged over the folds were then selected. In both cases, we opted for K=5.

A single run of K-fold CV may often lead to noisy estimates of model performance as results may vary with different splits of the training. While this could have been addressed by using a larger number of splits K, it was not possible in our study due to the limited number of samples available. We thus applied repeated K-fold CV, which has been shown to demonstrate reduced noise, increased precision, and greater reliability in performance estimates, relative to single K-fold CV. This involved simply repeating the K-fold CV procedure multiple times and selecting the best-performing set of hyperparameters which correspond to the best validation performance when averaged across all folds from all runs. Repeating our 5-fold CV 5 times resulted in 25 iterations altogether. Furthermore, all splits were stratified to ensure the proportion of class distribution were consistent across the training and test sets, as well as the CV folds.

Overall, the process of developing the source and target models can be summarised as follows:

1. To develop our source model, we first trained our two modelling pipelines, rf_pipe and xgb_pipe, using repeated 5-fold CV on the EPAD training set EPAD_train.

2. The set of hyperparameter values considered are shown in Table 2.3. By employing the random search optimisation algorithm, the optimal hyperparamaters were selected based on the evaluation metric specified for validation. In this case, we optimised the f1-score metric.

3. We then retrained both pipelines on the entire training set EPAD_train using the optimal hyperparameters, obtaining the two trained model pipelines rf_pipe_EPAD and xgb_pipe_EPAD.

4. We evaluated both rf_pipe_EPAD and xgb_pipe_EPAD on the held-out EPAD test set, EPAD_test, based on the performance metrics defined in Section 3.1.

xgb_pipe_EPAD produced the best overall performance, and was thus determined to be our source model, and was subsequently used for developing our target model.

5. We applied a parameter-based transfer learning approach, which builds on the assumption that the source and target models share parameters. Using the trained XGBoost classifier extracted from the best performing source model pipeline xgb_pipe_EPAD, we updated its parameters using the PREVENT training set, PREV_train, to obtain our target model, target_PREV.

6. A baseline model was trained on PREV_train using the XGBoost algorithm. Similar to Steps 2-3, the optimal hyperparameters (Table 2.3) were determined through repeated 5-fold CV, yielding the baseline model, baseline_PREV, after retraining was performed on the full, PREV_train training set.

7. We evaluated both the target and baseline model, target_PREV and baseline_PREV respectively, on the PREVENT test set PREV_test, and compared their performances.

| **Model** | Hyperparameter values considered | Optimal hyperparameters |
|---|---|---|
| **rf_pipe_EPAD** | **MissForest algorithm**:<br>max_features:[‘sqrt’,‘log2’],<br>max_depth:[10,12,14,16,18,20,22,24,26,28]<br>**SMOTENC/RandomUnderSampler(sampling ratio)**:<br>SMOTENC: [0.40,0.45,0.50,0.55,0.60,0.65]<br>RandomUnderSampler: [0.65,0.70,0.75]<br>**RandomForestClassifier**:<br>n_estimators: [20,30,40,50,70,100,150,200,250,300,350]<br>max_features: [’sqrt’,’log2’,6,8,10,12,14,16,18,20]<br>max_depth: [12,14,16,18,20,22,24,26,28,30,32]<br>min_samples_split: [2,3,4,5,6,7,8,9,10]<br>min_samples_leaf: [1,2,3,4,5]<br>criterion: [‘gini’,‘entropy’]<br>class_weight: [‘balanced’, ‘balanced_subsample’, None] | **MissForest algorithm**:<br>max_features:‘sqrt’,<br>max_depth:12<br>**SMOTENC/ RandomUnderSampler(sampling ratio)**:<br>SMOTENC: 0.40, RandomUnderSampler: 0.70<br>**RandomForestClassifier**:<br>n_estimators:350,<br>max_features: ‘sqrt’,<br>max_depth: 28,<br>min_samples_split:4,<br>min_samples_leaf: 1,<br>criterion:’entropy’,<br>class_weight: None |
| **xgb_pipe_EPAD** | **MissForest algorithm**: same as above<br>**SMOTENC/RandomUnderSampler(sampling ratio)**: same as above<br>**XGBClassifier**:<br>n_estimators: [20,30,40,50,70,100,150,200,250,300,350]<br>max_depth: [12,14,16,18,20,22,24,26,28,30,32]<br>learning_rate: [0.05,0.1,0.15,0.2,0.25,0.30,0.35, 0.4,0.45,0.5]<br>colsample_bytree: [0.6,0.7,0.8,0.9,1]<br>min_child_weight: [0.001,0.003,0.005,0.01,0.03]<br>scale_pos_weight: [1,1.3,1.35,1.4,1.45,1.5,1.55, 1.6,2,2.5,3,4,5,6,7,8,9,10] | **MissForest algorithm**:<br>max_features:‘log2’,<br>max_depth:10<br>**SMOTENC/ RandomUnderSampler(sampling ratio)**:<br>SMOTENC: 0.65, RandomUnderSampler: 0.65<br>**XGBClassifier**:<br>n_estimators:350,<br>max_depth: 24,<br>learning_rate:0.25,<br>colsample_bytree: 1,<br>min_child_weight: 0.001,<br>scale_pos_weight: 8 |
| **baseline_PREV** | **XGBClassifier**:<br>n_estimators: [20,30,40,50,70,100,150,200,250,300,350]<br>max_depth: [12,14,16,18,20,22,24,26,28,30,32]<br>learning_rate: [0.05,0.1,0.15,0.2,0.25,0.30,0.35, 0.4,0.45,0.5]<br>colsample_bytree: [0.6,0.7,0.8,0.9,1]<br>min_child_weight: [0.001,0.003,0.005,0.01,0.03]<br>scale_pos_weight: [1,1.3,1.35,1.4,1.45,1.5,1.55, 1.6,2,2.5,3,4,5,6,7,8,9,10] | **XGBClassifier**:<br>n_estimators:30,<br>max_depth: 22<br>learning_rate:0.05,<br>colsample_bytree: 0.8,<br>min_child_weight: 0.01<br>scale_pos_weight: 8 |

Table 2.3: The set of hyperparameters considered and optimal hyperparameters obtained. Default values were used for all other hyperparameters. We set random_state=42 for random processes.

# Chapter 3

# Model Evaluation, Feature Importance and Model Interpretability

In Section 3.1, we introduce the metrics used for model evaluation. This is followed by Section 3.2 in which we discuss how the importance of each feature can be quantified. Finally, in Section 3.3, we present the methods we used for model output interpretation.

## 3.1 Evaluation Metrics

For model evaluation on the unseen test sets, we employed a series of evaluation metrics suitable for assessing binary classification tasks. As we have access to the ground truth labels ("High-Risk" or "Low-Risk") in both our test sets, EPAD_test and PREVENT_test, we were able to compare our model outputs to the actual ground truths, thus enabling us to obtain the following information:

- True Positives (TP): "High-Risk" cases correctly predicted as "High-Risk"
- True Negatives (TN): "Low-Risk" cases correctly predicted as "Low-Risk"
- False Positives (FP): "Low-Risk" cases incorrectly predicted as "High-Risk"
- False Negatives (FN): "High-Risk" case incorrectly predicted as "Low-Risk"

We typically present the proportions of TP, TN, FP and FN in a confusion matrix, as shown in Figure 3.1.

Another common metric for assessing binary classification is accuracy, i.e. the proportion of correctly labelled instances, as expressed in Equation (1). However, while a high accuracy is often desirable, it may present an "over-optimistic" view on model performance, since accuracy often fails to take class imbalance into account [27]. For instance, in a binary classification task with a class distribution of 90:10, a model that predicts correctly for all majority class examples but wrongly for all minority class examples will still yield a high accuracy score of 90%.

Given that class imbalance is apparent in our datasets, additional metrics such as sensitivity (or recall TP rate), specificity (or TN rate), precision, geometric accuracy (GA), and F1-score are required to fully assess the effectiveness of our models. Sensitivity and specificity allow us to assess how good the model is at detecting positive or negative cases respectively, whereas precision allows us to measure the proportion of identified positive cases which was actually correct. There is often a trade-off between sensitivity and specificity, or between sensitivity and precision, such that a high sensitivity could result in lower specificity and precision, or vice versa [11]. As we aimed to strike a balance between both pairs of measures, we relied on the GA and F1-score – GA is the geometric mean of sensitivity and specificity, whereas the F1-score is the harmonic mean of sensitivity and precision. The equations of these measures are given below:

$$Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} \tag{1}$$

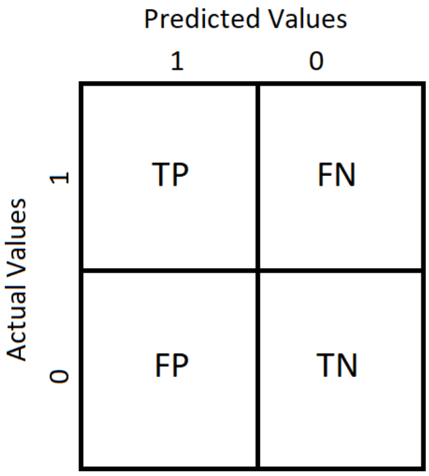
$$Sensitivity = \frac{TP}{(TP+FN)} \tag{2}$$

$$Specificity = \frac{TN}{(TP+FP)} \tag{3}$$

$$Precision = \frac{TP}{(TP+FP)} \tag{4}$$

$$GA = \sqrt{Sensitivity \times Specificity} \tag{5}$$

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \tag{6}$$



| | | Predicted Values 1 | Predicted Values 0 |
|---|---|---|---|
| Actual Values | 1 | TP | FN |
| | 0 | FP | TN |

Figure 3.1: Confusion Matrix

As recommended by the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines [12], we further evaluated our models based on measures of discrimination, i.e. the ability of a prediction model to distinguish between classes. This was achieved through observing the Receiver Operating Characteristics (ROC) curve, which is obtained by plotting the TP rate against the FP rate at different decision thresholds. Classifiers demonstrating curves closer to the top-left corner are indicative of better performance. Also important is the area under the ROC curve (AUROC), which reflects the probability that a randomly selected positive instance has a higher predicted probability of being positive than a randomly selected negative instance. Perfect models would have a maximum AUROC value of 1, whereas AUROCs lower than 0.5 would suggest an unreliable model that performs worse than random chance.

However, as alluded to previously, our datasets suffered from imbalanced classes, whereby the number of negative or “Low-Risk” cases outweighed the positive or “High-Risk” cases. As a consequence, the ROC and AUROC may be indifferent to the class distribution, resulting in an overly optimistic performance, especially when the negative class is more prevalent. To better account for imbalanced class distributions, we further employed the Precision-Recall curve (PRC), and the area under the Precision-Recall curve (AUPRC), which are tailored for applications whereby the rare positive cases are of more interest. The PRC essentially plots precision against recall (i.e. TP rate) at different decision thresholds, with a higher value of AUPRC indicating better performance. Additionally, by employing the bootstrapping technique [16], we performed one-tailed hypothesis testing to examine the significance in improvement of the AUROC and AUPRC between both models.

We then sought to assess the clinical utility of the model by applying Decision Curve Analysis (DCA) which plots the “net benefit” at varying probability thresholds – i.e. the minimum probability of disease risk at which further intervention is warranted – as proposed by Vickers et al. [48]. In brief, DCA can be used to assess the “net benefit” of strategies employed based on model predictions, as compared to default strategies of “intervention for all” or “intervention for none”. The “net benefit” can be interpreted as the rate at which TP cases are identified and treated, without an increase in the rate of false-positive (FP) cases. For instance, a net benefit of 0.10 would indicate an additional 10 patients per 100 subjected to a particular intervention (all of whom were subsequently confirmed to have AD), with no patients being falsely identified as having AD.

Finally, to evaluate the efficacy of our transfer learning approach, we employed the following metric proposed by Taylor et al. [44]: $ratio = \frac{(AUC_{with\,transfer} - AUC_{without\,transfer})}{AUC_{without\,transfer}}$.

## 3.2 Feature Importance

Feature importance allows us to score each feature based on their importance in affecting the predictions made by our model. As tree-based algorithms such as Random Forest and XGBoost were applied in our model, feature importance could easily be obtained in the form of weights, based on the effect the feature has on the mean decrease in impurity, i.e. the effectiveness of the feature at reducing uncertainty. These feature importances were then visualised using bar charts, and arranged from top to bottom in the order of decreasing importance.

While impurity-based feature importance has been a widely popular approach, it tends to suffer from unstable or biased results, whereby high-cardinality categorical features or continuous features are favoured [42]. To overcome this limitation, we employed a model-agnostic approach namely permutation-based feature importance, also known as “permutation importance” [5]. Permutation importance relies on the idea that the importance of a feature is proportional to the decrease in model accuracy after randomly shuffling the values of that particular feature, as this is indicative of how dependent the model is on the feature. As with impurity-based feature importance, permutation importance assigns a score to each feature, which we then visualised using bar charts.

## 3.3 Model Interpretability

To better explain our model outputs, we utilised the SHapley Additive exPlanation (SHAP) algorithm [26], which leverages concepts from coalitional game theory to ascertain the contribution of each feature in pushing the predicted value away from the average predicted value (i.e. the “base value”). SHAP is able to provide both global and local interpretations of model outputs. In terms of global explanations, SHAP provides quantification of feature effect sizes (similar to feature importance), with the added advantage of providing information on the directionality of the effect that each feature has on model outputs. Additionally, through local explanations of SHAP, we can characterise the magnitude and direction of each feature’s influence on the predicted value on a case-by-case basis, allowing us to distinguish the impacts of the features between individual instances. Lastly, as SHAP is able to provide mathematical guarantees for accuracy and consistency, it is preferable over alternative local interpretation techniques such as Locally Interpretable Model Agnostic Explanations (LIME) [21, 26].

# Chapter 4

# Results

## 4.1 Model Performance Analyses

In this section, we present the results of each model performance based on the evaluation methods described in Section 3.1. The confusion matrix for each model is given in Figure 4.1, whereby Figures 4.1a and 4.1b correspond to the results when rf_pipe_epad and xgb_pipe_epad were evaluated on the unseen test set of the EPAD dataset respectively, while Figures 4.1c and 4.1d represent the results obtained by evaluating target_PREV and baseline_PREV on the PREVENT test set respectively. The numbers of TP, TN, FP and FN cases found in these confusion matrices can further be used to derive the sensitivity, specificity, precision, accuracy of the models, as summarised in Table 4.1. We observe that xgb_pipe_EPAD demonstrated superior performance across all evaluation metrics present in the table, achieving a sensitivity of 71.72%, f1-score of 71.36% and GA of 80.01%, a significant improvement over the performance of rf_pipe_EPAD (sensitivity = 64.14%; f1-score = 66.15%; GA = 75.58%). xgb_pipe_EPAD also demonstrated slightly enhanced specificity, precision, and accuracy over the rf_pipe_EPAD model.

Comparing target_PREV against baseline_PREV, we notice that target_PREV outperformed baseline_PREV in most of the reported metrics – including sensitivity, precision, f1-score, as well as GA – as seen from Table 4.1. In contrast, baseline_PREV was able to achieve better performance for metrics such as specificity and accuracy.

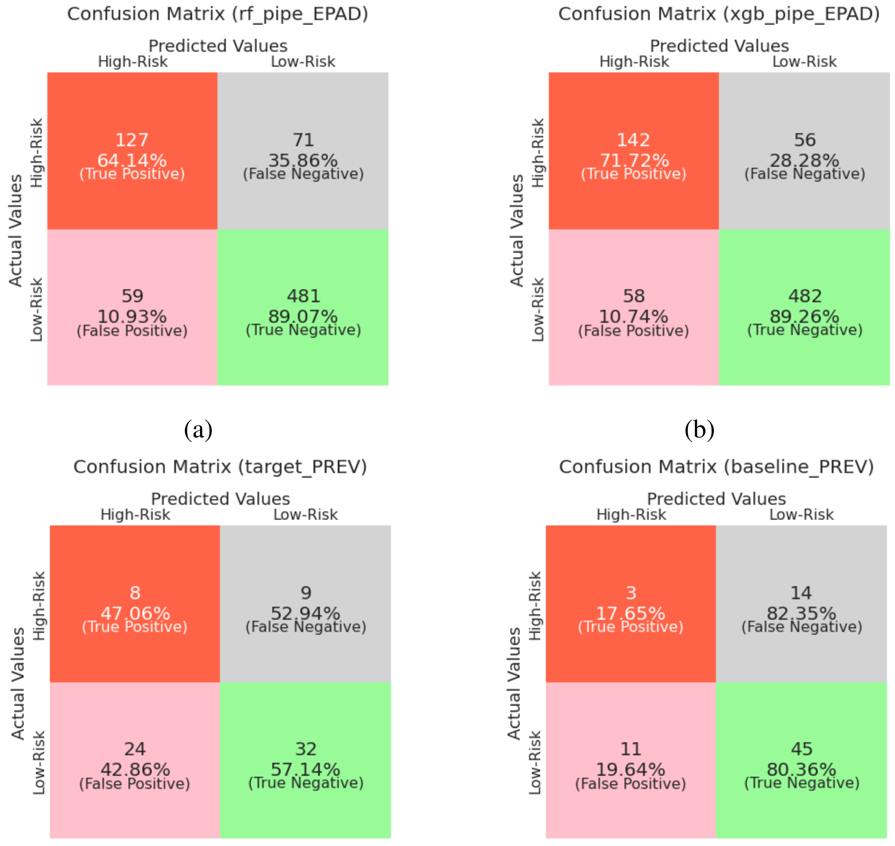
The AUROC and AUPRC of each model are given in Figure 4.2. xgb_pipe_EPAD outperformed rf_pipe_EPAD across both AUROC and AUPRC, as shown in Table 4.2. As mentioned previously, significance tests were applied to further substantiate our results, and the obtained p-values suggest a statistically-significant improvement (p-value=0.0269) in the AUPRC performance of xgb_pipe_EPAD compared to rf_pipe_EPAD, with no significant improvement in AUROC performance (p-value = 0.2260). Similarly, target_PREV was shown to achieve a higher AUROC than that of baseline_PREV. This was not the case for the AUPRC, whereby baseline_PREV demonstrated a slight improvement by 1.38%. In both cases, the obtained p-values (0.3878 for AUROC); 0.4124 for AUPRC) indicate that there were no significant differences in AUROC and AUPRC performances between target_PREV and baseline_PREV. Overall, we calculated the transfer learning efficacy rate to be 6.41%.

| **Metric** | **rf_pipe_EPAD** | **xgb_pipe_EPAD** | **target_PREV** | **baseline_PREV** |
|---|---|---|---|---|
| **Sensitivity (%)** | 64.14 | 71.72 | 47.06 | 17.65 |
| **Specificity (%)** | 89.07 | 89.26 | 57.14 | 80.36 |
| **Precision (%)** | 68.28 | 71.00 | 25.00 | 21.43 |
| **F1-Score (%)** | 66.15 | 71.36 | 32.65 | 19.35 |
| **Accuracy (%)** | 82.38 | 84.55 | 54.79 | 65.75 |
| **GA (%)** | 75.58 | 80.01 | 51.86 | 37.66 |

Table 4.1: The sensitivity, specificity, precision, f1-score, accuracy and GA of each model

| **Model** | **AUROC (%)** | **AUPRC (%)** | **p-value AUROC** | **p-value AUPRC** | **Transfer efficacy (%)** |
|---|---|---|---|---|---|
| **rf_pipe_EPAD** | 85.72 | 74.87 | | | |
| **xgb_pipe_EPAD** | 86.90 | 77.02 | 0.2260 | 0.0269 | N/A |
| **baseline_PREV** | 47.58 | 24.49 | | | |
| **target_PREV** | 50.84 | 23.11 | 0.3878 | 0.4124 | 6.41 |

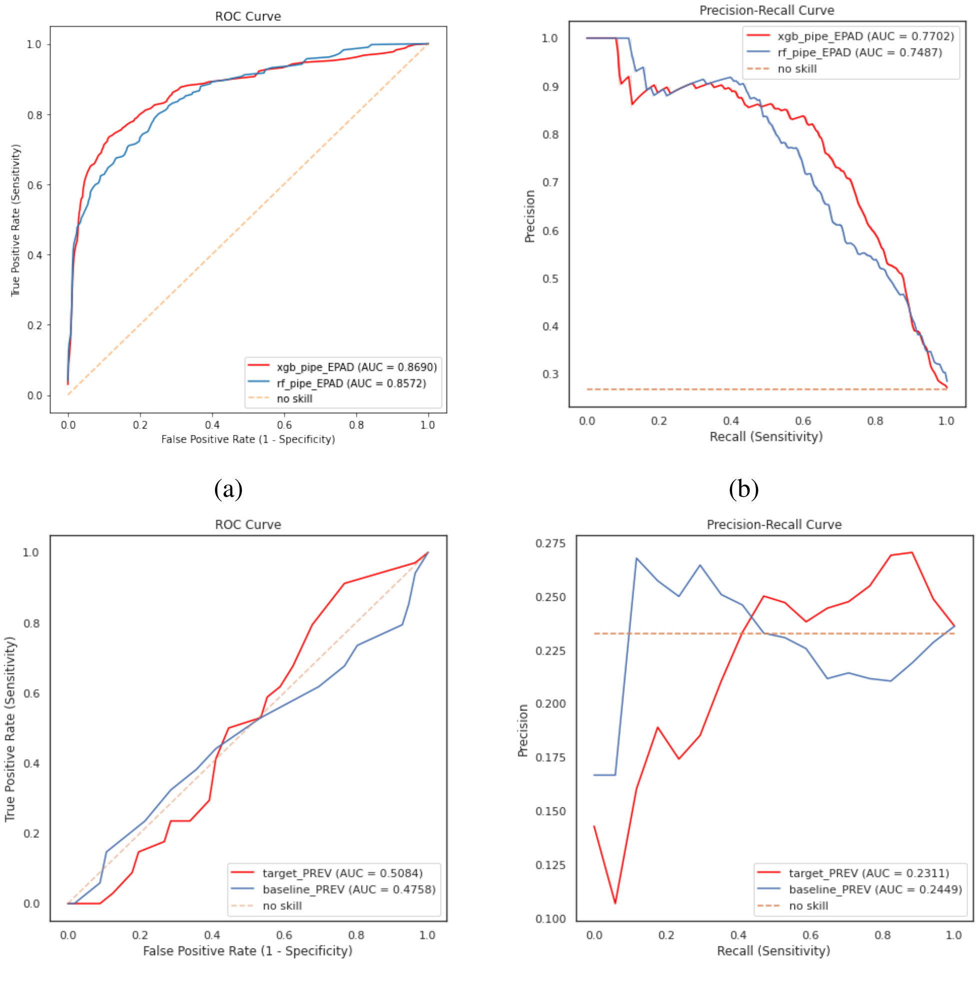
Table 4.2: The AUROCs, AUPRCs, the resulting p-values when comparing AUROCs and AUPRCs, and the transfer efficacy rate.



(a) (b)

(c) (d)

Figure 4.1: Confusion matrix of each model



(a) (b)

(c) (d)

Figure 4.2: The ROC and PRC of each model

## 4.2 Feature Importance Analyses

### 4.2.1 Impurity-Based Feature Importance

The impurity-based feature importances of rf_pipe_EPAD and xgb_pipe_EPAD are given in Figures 4.3a and 4.3b respectively. At first glance, the ranking of the most important features appear to be significantly different between the two models, e.g. the top 5 features (in descending order of importance) of rf_pipe_EPAD are “age”, “years_education”, “bmi_class”,“alcohol_unit”, and “gender”, whereas the top 5 features for xgb_pipe_EPAD is “hypertension”, “alcohol_unit”, “heart_disease”, “gender” and “osteoarthritis”. Despite these differences, we observe that there is significant overlap between the top 15 most important features in both models. For instance, features such as “age”, “gender”, “years_education”, “smoke”, “alcohol_unit”, “hypertension”, “hyperlipidaemia”, “eye_disease”, “heart_disease”, “lung_disease”, and “osteoarthritis” are among the top 15 most importance features shared across both models. Additionally, both models rank “drug_use” and “lower_gi_disease” among their 5 least important features.

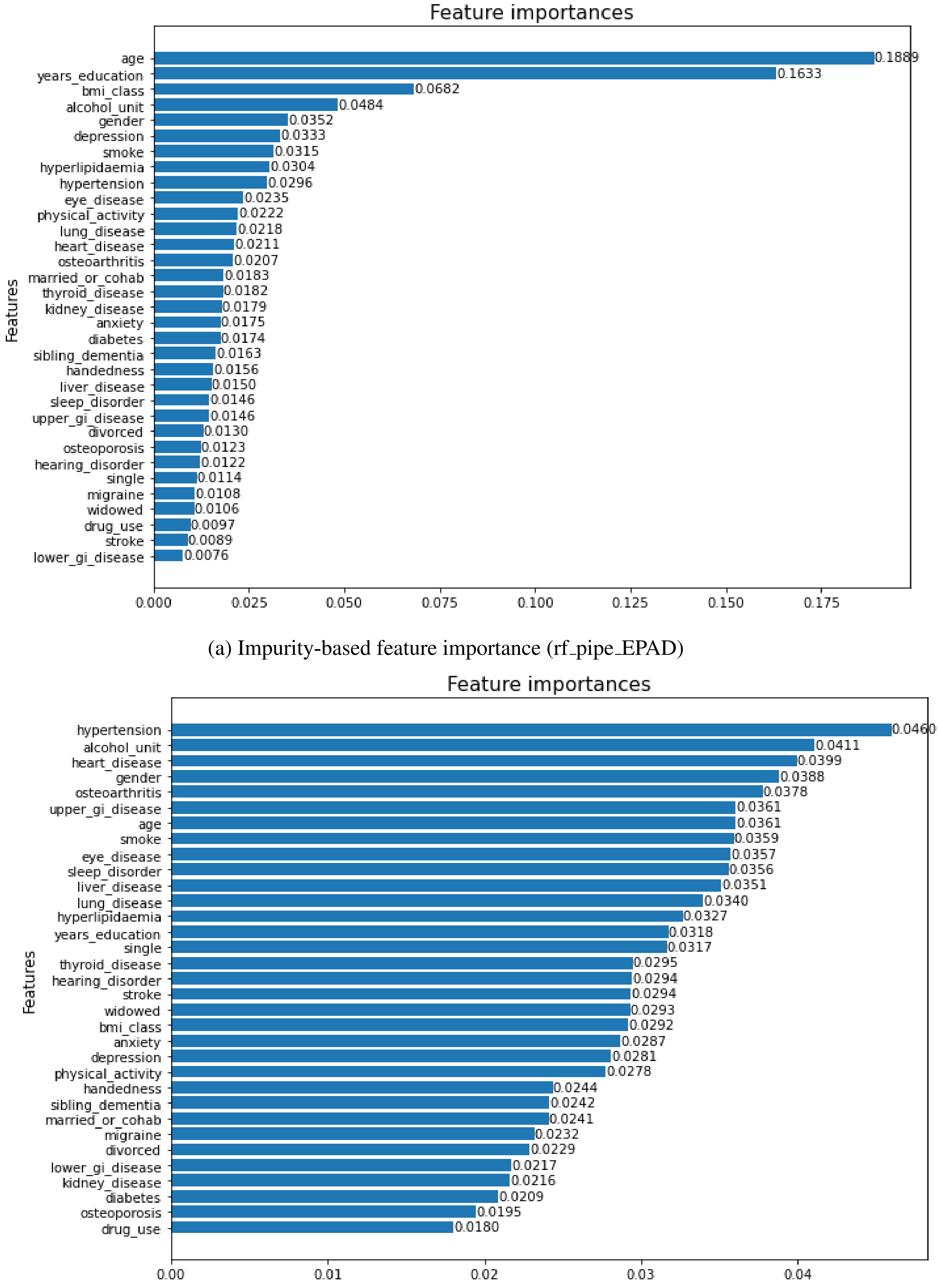
Figures 4.4a and 4.4b describe the feature importances of target_PREV and baseline_PREV respectively. In comparing the top 15 most important features for both models, 7 common features were identified, including: “age”, “years_education”, “single”,“heart_disease”, “hyperlipidaemia”, “lower_gi_disease”, and “sleep_disorder”. While similar features were identified as being important in both models, the ranked order of features by importance differs between models. For instance, target_PREV ranks “sleep_disorder” and “single” as the 1st and 5th most important features respectively, whereas baseline_PREV ranks both variables as the 3rd and 4th most important features respectively. Furthermore, while “upper_gi_disease was found to be the most important feature for baseline_PREV, the exact same variable is ranked as the least important for target_PREV.

Further comparisons can be drawn between the feature importances of the source xgb_pipe_EPAD and target target_PREV models. Upon updating the source model, about half of the original features remain in the top 15 most important features for the target model, namely “age”, “years_education”, “sleep_disorder”, “single”, “heart_disease”, “eye_disease”, “hypertension” and “hyperlipidaemia”. Other features, such as “alcohol_unit”, “gender”, “smoke”, “osteoarthritis”, “liver_disease”, “lung_disease”, and “upper_gi_disease”, are supplanted.

### 4.2.2 Permutation-Based Feature Importance

As mentioned in Section 3.2, we also implemented permutation-based feature importance as an alternative approach to the impurity-based feature importancee. The permutation importance analysis results for each model can be found in Figures 4.5 and 4.6. A comparison between the permutation importances of rf_pipe_EPAD and xgb_pipe_EPAD demonstrates that there is a considerable overlap (over 70%) among the top 15 most important features. In particular, “age”, “year_education”, and “hyperlipidaemia” are ranked as the top 3 most important features respectively in both models. Both models also include “gender”, “bmi_class”,“hypertension”,“alcohol_unit”,“smoke”, “thyroid_disease”, “kidney_disease”, “liver_disease” as part of their top 15 most important features, albeit with some differences in the order of their rankings. Further similarities are observed across both models, whereby “widowed” and “sleep_disorder” are respectively ranked as the second least important and least important features, with negative values for their permutation importances indicating that they may have a negative impact on predictions.

In examining the permutation importances of target_PREV and baseline_PREV, we observe that both models rank “bmi_class, “drug_use”, “lower_gi_disease”, and “liver_disease” among their top 15 most important features, as seen in Figures 4.6a and 4.6b. Furthermore, only the top 11 most important features are observed to have positive scores in both models, whereas all remaining features have a score equal or less than 0. For instance, “years_education” has been ranked as the least important feature in both models – in sharp contrast to the results of previously reported feature importances whereby “years_education” has always been among the top 10 or top 15 most important features. Finally, we can examine how the permutation importances change as a result of updating the source model, xgb_pipe_EPAD, for the development of the target model, target_PREV.. In this case , features such as “age”, “gender”,“bmi_class”, “depression”, “kidney_disease”, and “liver_disease” remain among the top 15 most important features, whereas all other features that were originally among the top 15 have dropped in rank, to the extent of having negative permutation importance scores.

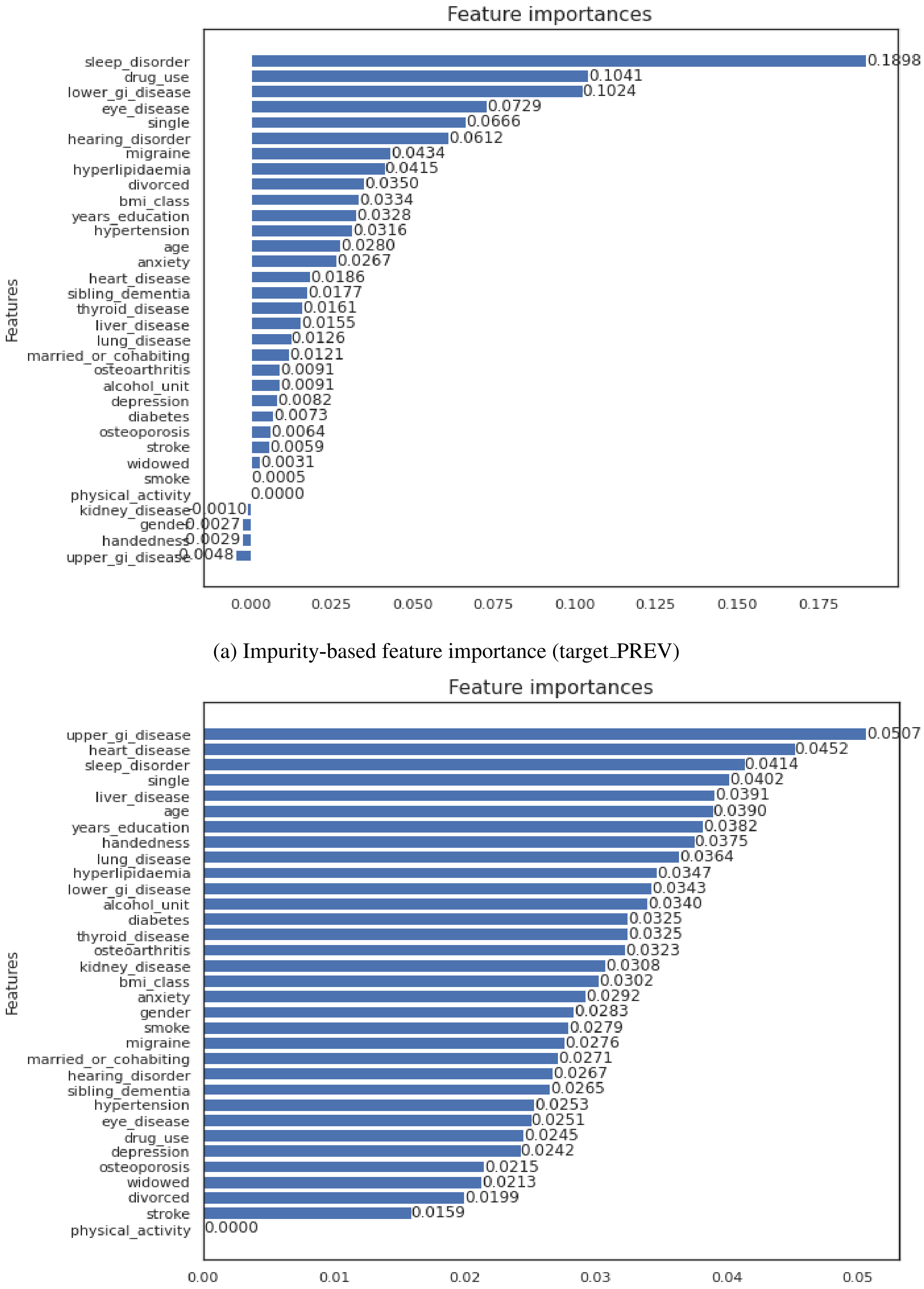


(a) Impurity-based feature importance (rf_pipe_EPAD)

B

(b) Impurity-based feature importance (xgb_pipe_EPAD)

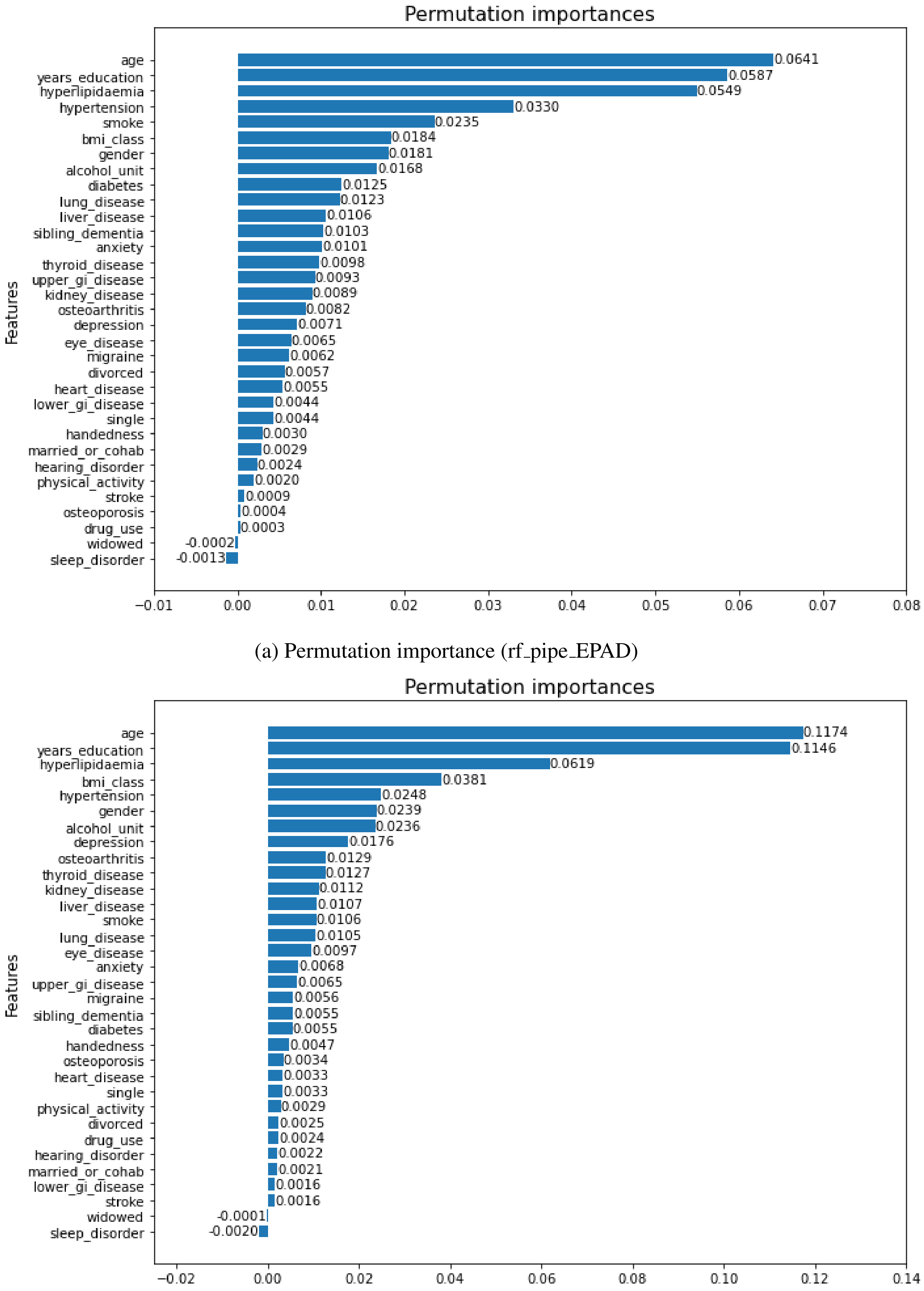
Figure 4.3: Impurity-based feature importance for rf_pipe_EPAD and xgb_pipe_EPAD



(a) Impurity-based feature importance (target_PREV)

(b) Impurity-based feature importance (baseline_PREV)

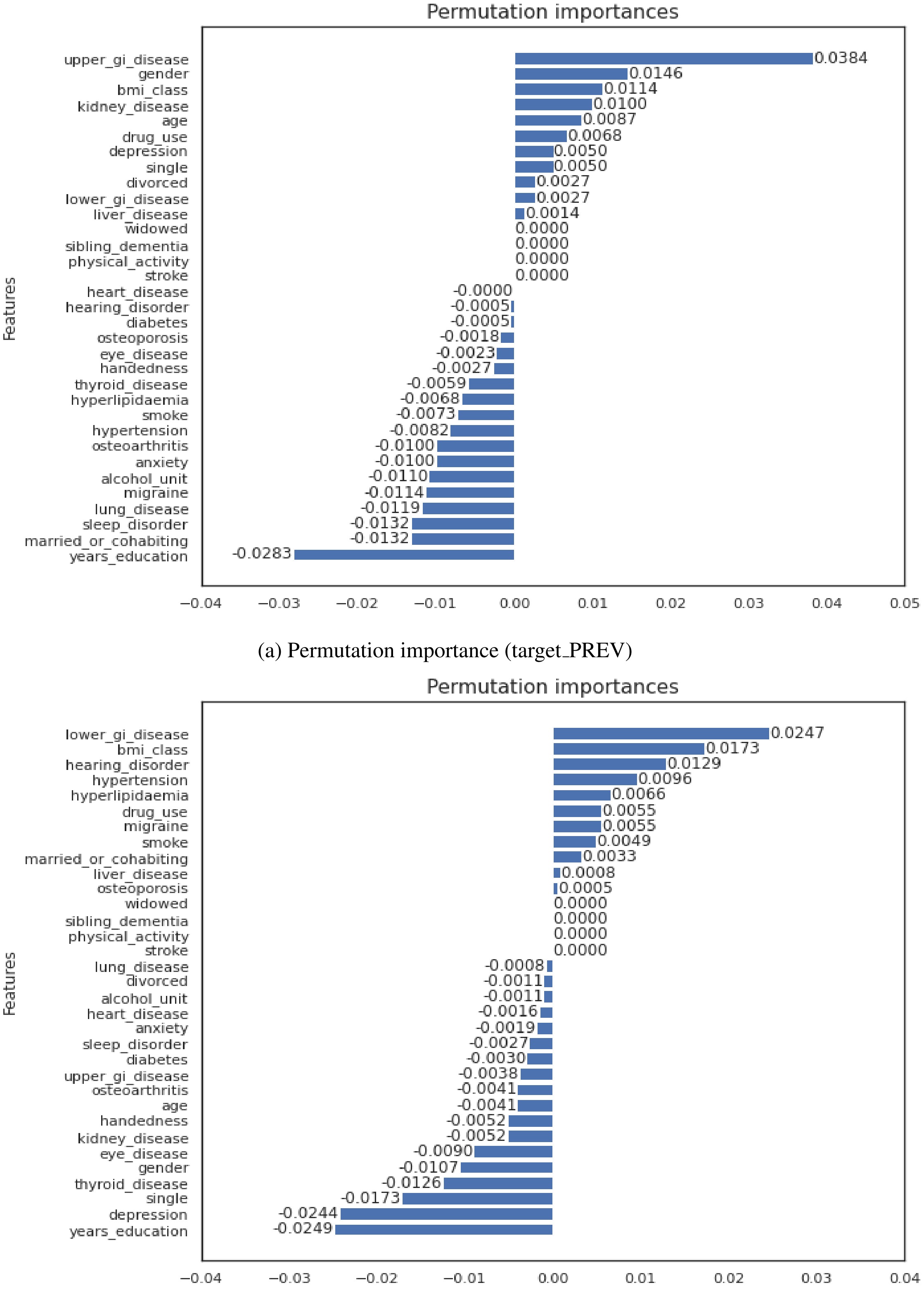
Figure 4.4: Impurity-based feature importance for target_PREV and baseline_PREV



(a) Permutation importance (rf_pipe_EPAD)

(b) Permutation importance (xgb_pipe_EPAD)

Figure 4.5: Permutation importance for rf_pipe_EPAD and xgb_pipe_EPAD



(a) Permutation importance (target_PREV)

(b) Permutation importance (baseline_PREV)

Figure 4.6: Permutation importance for target_PREV and baseline_PREV

## 4.3 Model Interpretability with SHAP

While the feature and permutation importances allow us to quantify the significance of each feature for our models on a global level, we can further examine how each feature contributes to a prediction at an individual level through SHAP values. Given that the xgb_pipe_EPAD model outperformed the rf_pipe_EPAD model, our analysis will mainly be based on the former. Figure 4.7 illustrate the SHAP explanations of four randomly selected predictions obtained through applying the xgb_pipe_EPAD model on the unseen EPAD test set, with each of Figure 4.7a, 4.7b, 4.7c, 4.7d corresponding to a TP, FN, TN, and FP case respectively. Through the SHAP explanations, we can determine both the direction and magnitude of each feature's effect, based on the colour and the length of the bar respectively. For instance, features labelled in blue correspond to protective factors, whereas those labelled in red indicate risk factors. A longer bar length would suggest a larger impact of the feature in driving the prediction.

### 4.3.1 SHAP explanations for xgb_EPAD_pipe

In Figure 4.7a, we observe that the individual is correctly predicted as "High-Risk", with a probability of 86%. For this individual, the effects of protective factors – including a high number of years of education and a normal BMI – are offset by risk factors such as hyperlipidaemia, hypertension, an age of 71, not being single, smoking, the absence of lung disease and a sibling with dementia, resulting in a prediction of "High-Risk".

In Figure 4.7b, we examine a 70 year-old individual with relatively fewer years of education who has been incorrectly predicted as being "Low-Risk". Consistent with the previous individual, "years_education", "age", "sibling_dementia", "hypertension" and "bmi_class" appear to be significantly impacting the prediction outcome, with the main differences being that age and the absence of sibling dementia are now protective factors, whereas the lower number of years of education is now a risk factor.

Figure 4.7c represents an individual who is correctly predicted as being "Low-Risk", with a probability of only 6% for being "High-Risk". Again, a relatively high number of years of education, normal BMI, and an absence of hyperlipidaemia are protective factors against AD risk, offsetting the effects of risk factors such as smoking, alcohol consumption and physical activity.

Finally, in Figure 4.7d, risk factors such as heart disease, hyperlipidaemia, an abnormally high BMI, and low number of years of education are seen to drive the prediction of this particular individual, resulting in a high AD risk probability of 89%

being predicted despite being a “Low-Risk” individual in reality.

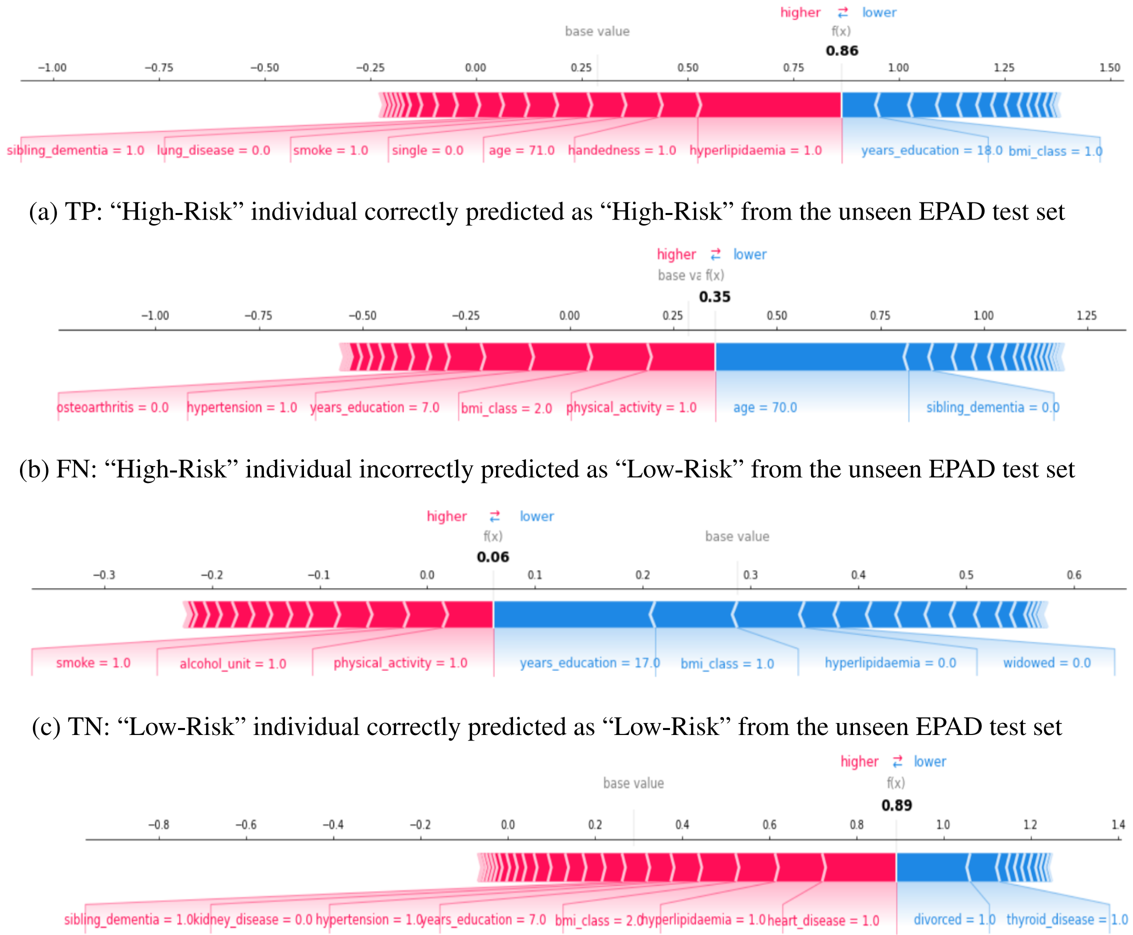
### 4.3.2 SHAP explanations for target_PREV

We applied the same approach in the interpretation of predicted outcomes, when the target model,target_PREVENT, was applied on the unseen PREVENT dataset. Similarly, four random samples, each of which corresponding to a TP,FN,TN, and FP case, are selected and their SHAP explanations are provided in Figures 4.8a, 4.8b, 4.8c, and 4.8d respectively. For all cases, we observe that “years_education” appears to be the most important protective feature. Additionally, the absence of anxiety (Figures 4.8a and 4.8d) appears to be protective, whereas the presence of anxiety (Figures 4.8b and 4.8c) appears to increase AD risk. We further note that “married_or_cohabiting” appears to be a risk factor in all cases, particularly in Figures 4.8b and 4.8c where it is the risk factor exerting the greatest impact on the predictions.

### 4.3.3 SHAP summary plots

While the above examples provide local interpretations of a randomly sampled instances from the test sets, a global interpretation based on aggregations of SHAP values of each feature is given in the form of the SHAP summary plots (Figure 4.9), allowing us to directly compare the feature effects and importances between the source model xgb_pipe_EPAD and the target model target_PREV. Each point on the summary plot represents an instance from the test set. The SHAP values are given on the x-axis, whereas the features are arranged along the y-axis in order of decreasing importance from top to bottom. The value of each feature is colour-coded along a red-blue spectrum, with red representing higher feature values and blue representing lower feature values. In this case, binary features may only have two colours (either red or blue) whereas continuous or ordinal features can take any colour along the spectrum.

The summary plot for the xgb_pipe_EPAD model is shown in Figure 4.9a. We observe that “age” is the most important feature, followed by “year_education”, “bmi_class” and “hyperlipidaemia”, findings that are consistent with the local interpretations discussed previously. Furthermore, the effects of some features are extremely evident, based on the distribution of red and blue points. For instance, the presence of hyperlipidaemia (denoted in red) is associated with a increased predicted AD risk, as suggested by their association with positive SHAP values. In contrast, the absence of hyperlipidaemia (denoted in blue) is associated with negative SHAP values, and thus a decreased



(a) TP: "High-Risk" individual correctly predicted as "High-Risk" from the unseen EPAD test set

(b) FN: "High-Risk" individual incorrectly predicted as "Low-Risk" from the unseen EPAD test set

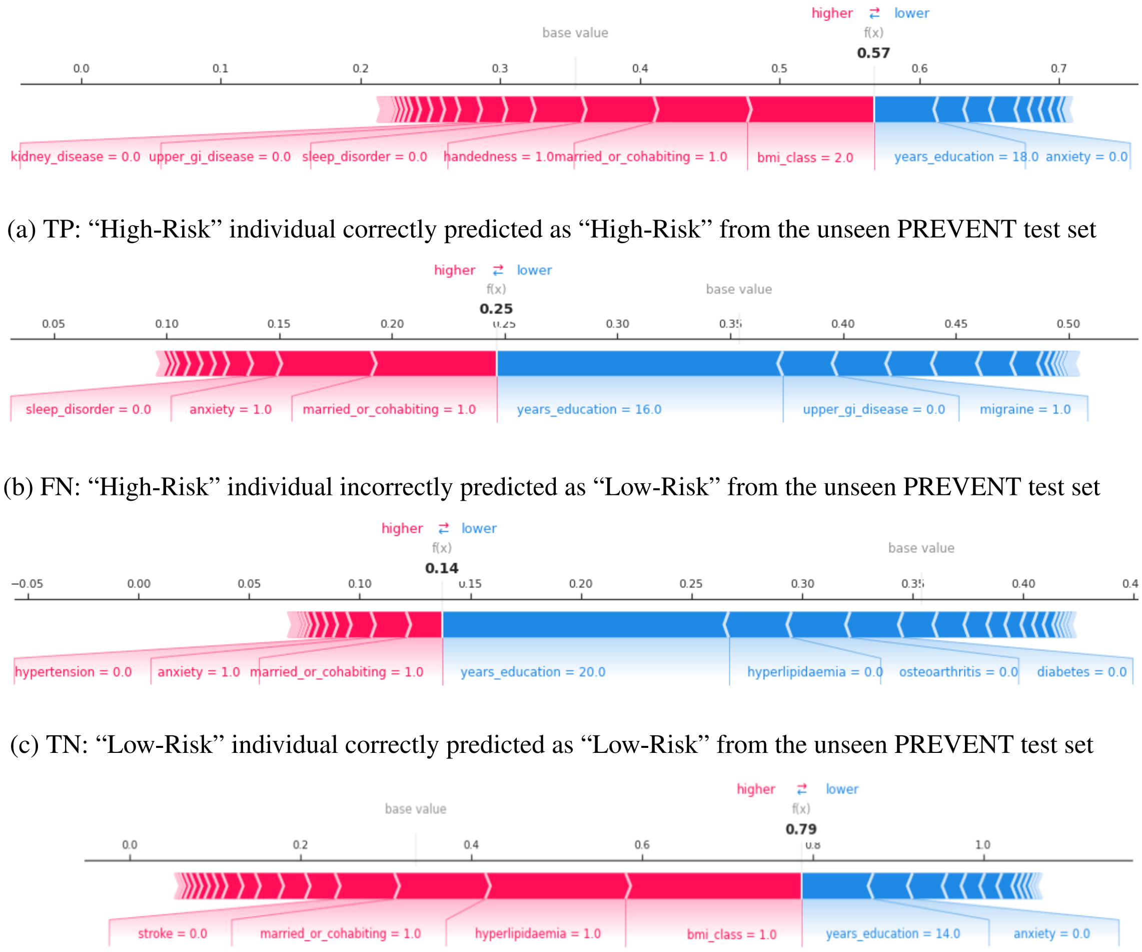
(c) TN: "Low-Risk" individual correctly predicted as "Low-Risk" from the unseen EPAD test set

(d) FP: "Low-Risk" individual incorrectly predicted as "High-Risk" from the unseen EPAD test set

Figure 4.7: SHAP visualisations for xgb_pipe_EPAD on four instances sampled from the unseen EPAD test set

predicted AD risk. The same pattern can be observed for other binary features such as "married_or_cohabiting", "liver_disease", "heart_disease", and "sibling_dementia".

A similar summary plot for the target_PREV model is shown in Figure 4.9b. We note that "years_education", "osteoarthritis", "married_or_cohabiting", "age", and "bmi_class" are among the top 5 most significant features, displaying considerable similarity to that of xgb_pipe_EPAD. In examining the effects of each feature, it appears that the presence of anxiety is associated with increased AD risk, consistent with our findings from the local SHAP explanations (Figure 4.8). Furthermore, binary features such as "hypertension", "kidney_disease", "thyroid_disease" display the same effects, in which a value of 0 (i.e. not present) is associated with increased AD risk.



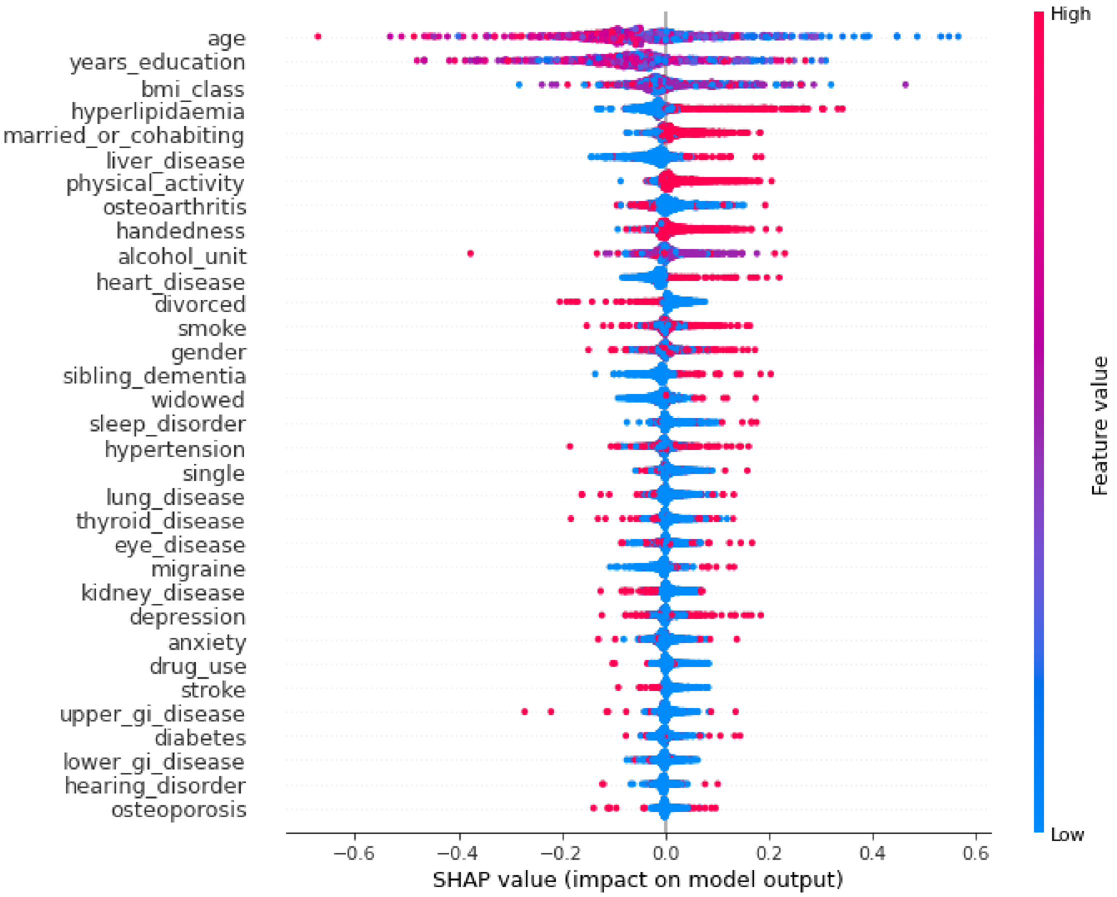
(a) TP: "High-Risk" individual correctly predicted as "High-Risk" from the unseen PREVENT test set

(b) FN: "High-Risk" individual incorrectly predicted as "Low-Risk" from the unseen PREVENT test set

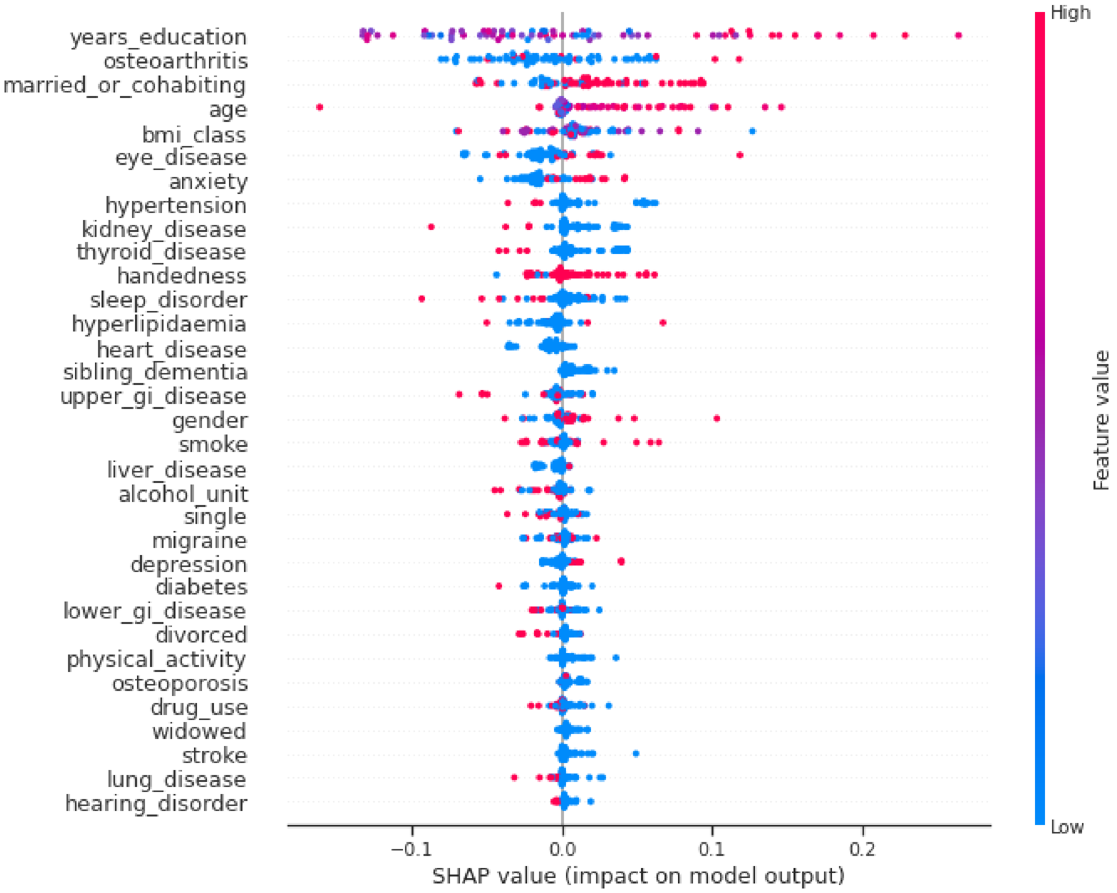
(c) TN: "Low-Risk" individual correctly predicted as "Low-Risk" from the unseen PREVENT test set

(d) FP: "Low-Risk" individual incorrectly predicted as "High-Risk" from the unseen PREVENT test set

Figure 4.8: SHAP visualisations for target_PREVENT on four instances sampled from the unseen PREVENT test set



(a) SHAP summary plot for xgb_pipe_EPAD



(b) SHAP summary plot for target_PREV

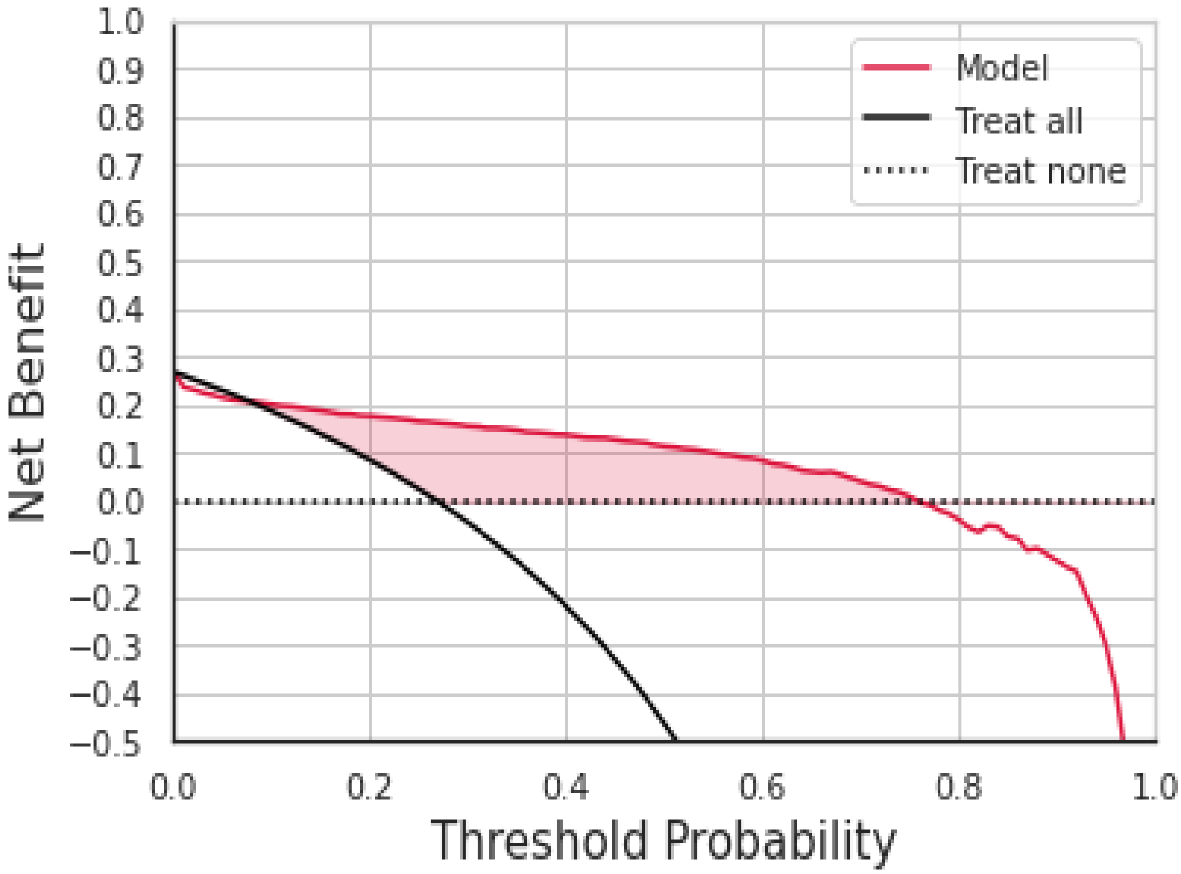
Figure 4.9: SHAP summary plots

## 4.4 Decision Curve Analysis

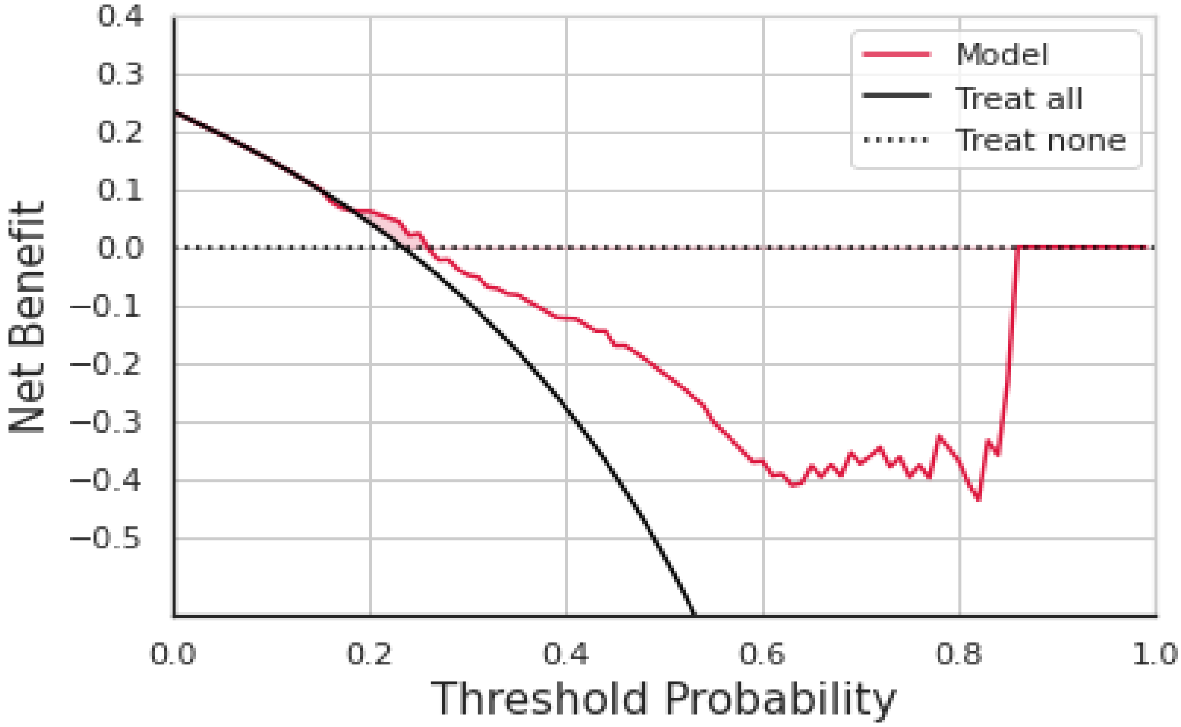
We applied Decision Curve Analysis (DCA) to our best-performing models, including the best source model, xgb_pipe, and the best target model, target_PREV. The results from DCA are presented in Figure 4.10. The x-axis represents the range of probability thresholds, essentially reflecting the varying preference of a medical professional to intervene. A value towards the left end of the x-axis would represent a preference that weighs the relative harm of missing an AD diagnosis greater than the harm or cost of unnecessary intervention, whereas the a value towards the right end would suggest the opposite. On the y-axis, we are given the net benefit in units of TP per patient.

In Figure 4.10a, we examine the net benefit of the xgb_pipe_EPAD model, which is represented as a solid red curve. The solid black curve represents the net benefit conferred by assuming that all patients will develop AD and treating them all (i.e. "treat all"), whereas the dashed line represents the net benefit of assuming that no patients will develop AD and therefore not treating anyone (i.e. "treat none"). Curves at higher values of the y-axis would imply that following that particular strategy will lead to greater benefit. In this case, we observe that between a probability threshold of about 0.1 to 0.75, the net benefit of our model is greater than that of the default strategies of "treat all" and "treat none". For instance, at a probability threshold of 0.4 (indicating that a doctor will potentially intervene if AD risk is 40% greater), the xgb_pipe_EPAD model has a net benefit of about 0.14 over the strategy of treating no one, further suggesting that 14 additional true positives per 100 patients are treated, with no additional increase in false positives. For probability threholds below 0.1, the net benefit of xgb_pipe_EPAD appears to be slightly lower than the "treat all" strategy but higher than the "treat none" strategy, whereas for probability thresholds above 0.75, the net benefit of our model is lower than that of the "treat none" strategy.

Similarly, we compare the net benefit of our target model, target_PREV against the net benefits of default strategies, "treat all" and "treat none" in Figure 4.10b. In this case, we observe that the net benefit of target_PREV is only marginally higher than that of both default strategies for a small range of probability thresholds, somewhere between 0.18 and 0.25. For larger probability thresholds, the target_PREV model performs worse than the "treat none" eventually confers negative benefit, though it does demonstrate superior performance over the "treat all" strategy.



(a) SHAP summary plot for xgb_pipe_EPAD



(b) SHAP summary plot for target_PREV

Figure 4.10: SHAP summary plots

# Chapter 5

# Discussion

This study developed an ML-based AD risk prediction model based on a transfer learning framework, as inspired by the approach adopted by Danso et al. [14]. Using data drawn from two different populations, we developed a source model on the EPAD dataset, and subsequently updated this model using the PREVENT dataset. Our best-performing source model was able to achieve a sensitivity of 71.72%, specificity of 89.26%, AUROC of 86.90% and GA of 80.01 %. While at first glance these results appear to be lower than the model performance reported in Danso et al., we note that our model was developed using a significantly smaller data set (n=3690), in contrast to the number of samples available (n=84856) in the dataset used by Danso et al.. For our target model, we achieved a sensitivity of 47.06%, specificity of 57.14%, GA of 51.86%, and AUROC of 50.84%. In this case, we were able to obtain better sensitivity compared to the target model developed by Danso et al. (sensitivity = 38.1%, specificity = 84.7%, GA = 56.5%, AUROC=63%), but lower specificity, GA and AUROC. This can be explained by the sensitivity-specificity trade-off – the increased sensitivity of our target model is at the expense of the lower specfcity, whereas Danso et al. achieved a high specificity at the cost of a lower sensitivity.

To improve on the AD risk prediction algorithm proposed by Danso et al., we applied additional preprocessing procedures. In handling missing data, Danso et al. applied listwise deletion, essentially discarding all observations for which there are missing values on any of the variables. Despite being a straightforward and popular approach, listwise deletion could potentially result in a significant loss of information, leading to biased analysis in subsequent modelling. To account for this problem, we explored several missing data imputation techniques, including mean/mode imputation, K-nearest neighbours (KNN) imputation and the MissForest algorithm, eventually opting for the MissForest algorithm [40], given the different limitations associated with mean/mode imputation and KNN imputation. While mean/mode imputation tends to bias the variance and standard errors of the imputed variables, KNN imputation suffers from the need for additional data preprocessing for since distance metrics such as the Euclidean distance is used, in addition to being sensitive to outliers and noise (). In contrast, not only can the MissForest algorithm be easily applied on mixed data types, it is robust to noise and also able to capture nonlinear relationships or complex interactions between variables. While the use of the MissForest algorithm is known to be computationally expensive due to its added complexity compared to other missing data imputation approaches, given the relatively small sample size of our datasets this issue was less apparent in our study. However, we acknowledge that this may pose a challenge for larger datasets.

In most binary classification tasks, class imbalance is often observed where the class of interest (i.e. the positive class) is in the minority. As discussed previously, this reduces predictive performance, especially on the minority class. It is common to apply resampling techniques to rebalance the class distribution, either by increasing the number of samples from the minority class or decreasing the number of samples from the majority class. Common techniques to achieve this include random oversampling or random undersampling, which involve duplicating random observations from the minority class or removing random observations from the majority class respectively. As random oversampling simply adds exact copies of examples from the minority class to the dataset, this increases the risk of a model overfitting. Therefore, instead of naive random oversampling, we applied the more advanced SMOTE technique [9] which simulated artificial samples from the minority class, and obtained enhanced performance by combining it with random oversampling. Additionally, as class imbalance could contribute to misleading or overly optimistic results for some standard evaluation metrics such as accuracy and AUROC, we reported additional performance metrics such as sensitivity, specificity, precision, GA, f1-score and AUPRC to ensure that performance estimates of our models were realistic and objective. Additionally, decision curve analysis was applied for our source and target models, with the results we obtained justifying the use of our model in clinical settings, as it is able to better inform shared decision-making strategies surrounding AD that take both clinician and patient preferences into account, while also providing key information to healthcare financiers performing cost-benefit analyses into secondary prevention strategies for AD.

Our work examined both the feature importance and the permutation importance of each model in some detail. We acknowledge the challenge in pinpointing the relative differences in terms the features' impact sizes and rankings between models. However, given that the xgb_pipe_EPAD model demonstrated superior performance over all other models, we will focus our discussion on the top 15 most important features identified for the xgb_pipe_model, specifically the results from permutation importance analysis for this model, given its enhanced reliability over impurity-based feature importance (the reasons for which are discussed in Section 3.2).

Consistent with the results of Danso et al., age was identified to be the most important risk factor for our model. Of the Top 15 most important risk factors, only "age" and "gender" were non-modifiable risk factors, with the remaining 13 features all considered to be modifiable. Of the 13 modifiable risk factors that were identified, 6 of these (education level, BMI, hypertension, hyperlipidaemia, osteoarthritis, smoking) corresponded with the top 10 most important risk factors for the model developed by Danso et al. Interestingly, more than half of the top 15 modifiable risk factors identified through our study concur with the top 12 modifiable risk factors by the 2020 Lancet commission report on dementia [25], including less education, high BMI, hypertension, excessive alcohol consumption, depression and smoking.

As our model relied on a larger number of features, we were able to identify novel risk factors that were important for risk prediction but not included for model development by Danso et al.. For instance, while alcohol consumption, thyroid disease, kidney disease and eye disease were ranked among the top 15 most important risk factors in our study, these features were not present in the study of Danso et al.. Of these, lung disease, thyroid disease , kidney disease and eye disease were not included as established risk factors in the Lancet report [25]. However, emerging evidence appears to support an association between the presence of certain thyroid [17], kidney [51], eye [24] and lung diseases [35, 46], and the pathophysiology of AD, lending credence to our findings from feature importance analysis. For example, a systematic review by Figueroa et al. concluded that there is association between thyroid dysfunction and AD [17], though a cause-effect relationship is as yet not fully established. Another review examined the relationship between chronic kidney disease and AD, via a variety of mechanisms including impaired clearance of uraemic toxins and renin-angiotensin-aldosterone system dysfunction [51].

While feature importance is able to provide us with valuable information, there are certain limitations or pitfalls that we should be aware of when interpreting feature importance results. For instance, when two features are highly correlated, this would result in a lower importance value for both features as the importance weights are "shared" across both features, even if they might actually be very important [22]. For instance, by looking at the features used in our models, it may be possible for "smoke" and "lung_disease" , or for "physcial_activity" and "bmi_class" to be correlated, resulting in misleading feature importance weights. Therefore, it may be worthwhile considering alternative methods such as drop-column importance [28], or conditional variable importance, as proposed by Strobl et al. [41]. Additionally, it may be beneficial to enhance model interpretability by establishing causal relationships between risk factors and predicted outcomes. It is important to note that the results of feature importance and SHAP explanations do not imply any cause-effect relationships, and further investigations based on causal inference [13] techniques are required.

Although we have attempted to avoid loss of information through the application of missing data imputation methods, the problem of information loss was inevitable, as a result of the need to group certain categorical features when harmonising the feature representations across the EPAD and PREVENT datasets, necessary for the purposes of homogeneous transfer learning. For instance, due to differences of how the "physical_activity" variable was recorded across both datasets, we were unable to retain the intensity and exact frequency of physical activity that was performed, eventually resorting to a binary encoding of the variable which only corresponds to a response of "Yes" or "No". In contrast, this level of granularity with respect to the physical activity feature was retained in the model developed by Danso et al.. This may be a potential cause for "physical_activity" not being ranked as high in the feature importance for our model. The same applies for variables such as "lung_disease", which was derived by collapsing several different types of diseases such as asthma, pneumonia, and Chronic obstructive pulmonary disease (COPD) into a single category. Furthermore, in developing our source model, the outcome was not based on actual AD diagnosis, but derived according to parental dementia diagnosis and APoE4 carrier status, which are proxy outcome measures. It would be beneficial to develop a model using actual AD diagnosis as an outcome measure, so as to enable direct clinical correlation. Finally, we note that the target model target_PREV developed in our study was able to outperform the baseline model baseline_PREV across most of the measures considered – including AUROC, sensitivity, precision, f1-score, and GA, – therefore substantiating that a transfer learning framework was indeed more efficient and effective than the traditional machine learning framework of retraining a model from scratch. However, we acknowledge that the performance of our target model could be further improved. We attribute the low transfer efficacy rate of 6.41% to the limited sample size of the PREVENT dataset (n=361). While more powerful algorithms such as support vector machines or artificial neural networks could be considered, this may lead to an increase in computational cost in addition to a loss of model explainability, as there is often a trade-off between applying more complex model for enhanced performance and model interpretability.

# Chapter 6

# Conclusions

Our work demonstrates ensemble-based ML models that are able to predict high AD risk in asymptomatic individuals drawn from a population with a lower mean age, with promising results. We were also successful in generating data visualisations that intuitively convey the relative importances of each feature in driving the predictions made by our models, to enhance model explainability alongside increasing its clinical utility. Through our work, we identified a number of putative modifiable risk factors for AD, which could lend themselves to suggesting future avenues for research. Considering the highly implementable nature of our model, we predict that, with further improvements and external validation, our predictive model has the potential to revolutionise existing diagnostic and management protocols surrounding AD, enabling high-risk patients to be screened and identified by clinicians prior to the onset of clinically-overt symptoms, without relying on expensive neuroimaging or biomarker-based screening techniques. In the same vein, the application of our AD risk prediction model could also have positive implications from a healthcare financing perspective, enabling the mass screening of asymptomatic individuals for AD risk as part of an integrative secondary prevention program.

# Bibliography

[1] Dementia: Who fact sheet. *WHO Website*, 2021.

[2] 2022 alzheimer's disease facts and figures. *Alzheimers Dement*, 18(4):700–789, 2022.

[3] Alzheimer's disease. *www.alzheimers.org.uk*, 2022.

[4] Md Manjurul Ahsan and Zahed Siddique. Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, page 102289, 2022.

[5] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[6] Kaj Blennow. Cerebrospinal fluid protein biomarkers for alzheimer's disease. *NeuroRx*, 1(2):213–225, 2004.

[7] Kaj Blennow and Henrik Zetterberg. The past and the future of alzheimer's disease fluid biomarkers. *Journal of Alzheimer's Disease*, 62(3):1125–1140, 2018.

[8] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[9] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[10] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[11] Kevin Chu. An introduction to sensitivity, specificity, predictive values and likelihood ratios. *Emergency Medicine*, 11(3):175–181, 1999.

[12] Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel GM Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *Journal of British Surgery*, 102(3):148–158, 2015.

[13] Peng Cui, Zheyan Shen, Sheng Li, Liuyi Yao, Yaliang Li, Zhixuan Chu, and Jing Gao. Causal inference meets machine learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3527–3528, 2020.

[14] Samuel O Danso, Zhanhang Zeng, Graciela Muniz-Terrera, and Craig W Ritchie. Developing an explainable machine learning-based personalised dementia risk prediction model: A transfer learning approach with ensemble learning algorithms. *Frontiers in big Data*, 4:21, 2021.

[15] Rahul C Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, 2015.

[16] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[17] Paulina Belén Sepulveda Figueroa, Ana Flávia Fernandes Ferreira, Luiz Roberto Britto, Arlette Patricia Doussoulin, and Andrea da Silva Torrao. Association between thyroid function and alzheimer's disease: A systematic review. *Metabolic Brain Disease*, 36(7):1523–1543, 2021.

[18] Serge Gauthier, Pedro Rosa-Neto, José A. Morais, and Claire Webster. Adi - world alzheimer report 2021, Sep 2021.

[19] Jantje Goerdten, Iva Čukić, Samuel O. Danso, Isabelle Carrière, and Graciela Muniz-Terrera. Statistical methods for dementia risk prediction and recommendations for future work: A systematic review. *Alzheimer's Dementia: Translational Research Clinical Interventions*, 5:563–569, 2019.

[20] Sergio Grueso and Raquel Viejo-Sobera. Machine learning methods for predicting progression from mild cognitive impairment to alzheimer's disease dementia: a systematic review. *Alzheimer's research & therapy*, 13(1):1–29, 2021.

[21] RibeiroMT SinghS GuestrinC. Why should i trust you? In *Explaining the predictions of any classifier. Paper presented at: Proceedings of the 22nd ACM*

*SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[22] Giles Hooker, Lucas Mentch, and Siyu Zhou. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6):1–16, 2021.

[23] Julius M Kernbach and Victor E Staartjes. Foundations of machine learning-based clinical prediction modeling: Part ii—generalization and overfitting. *Machine Learning in Clinical Neuroscience*, pages 15–21, 2022.

[24] Elżbieta Kuźma, Thomas J Littlejohns, Anthony P Khawaja, David J Llewellyn, Obioha C Ukoumunne, and Ulrich Thiem. Visual impairment, eye diseases, and dementia risk: a systematic review and meta-analysis. *Journal of Alzheimer's Disease*, 83(3):1073–1087, 2021.

[25] Gill Livingston, Jonathan Huntley, Andrew Sommerlad, David Ames, Clive Ballard, Sube Banerjee, Carol Brayne, Alistair Burns, Jiska Cohen-Mansfield, Claudia Cooper, et al. Dementia prevention, intervention, and care: 2020 report of the lancet commission. *The Lancet*, 396(10248):413–446, 2020.

[26] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[27] Fadel M Megahed, Ying-Ju Chen, Aly Megahed, Yuya Ong, Naomi Altman, and Martin Krzywinski. The class imbalance problem. *Nat Methods*, 18(11):1270–7, 2021.

[28] Allison E Miller, Emily Russell, Darcy S Reisman, Hyosub E Kim, and Vu Dinh. A machine learning approach to identifying important features for achieving step thresholds in individuals with chronic stroke. *Plos one*, 17(6):e0270105, 2022.

[29] Golrokh Mirzaei and Hojjat Adeli. Machine learning techniques for diagnosis of alzheimer disease, mild cognitive disorder, and other types of dementia. *Biomedical Signal Processing and Control*, 72:103293, 2022.

[30] Alexis Moscoso, Jesús Silva-Rodríguez, Jose Manuel Aldrey, Julia Cortés, Anxo Fernández-Ferreiro, Noemí Gómez-Lado, Álvaro Ruibal, Pablo Aguiar,

Alzheimer's Disease Neuroimaging Initiative, et al. Prediction of alzheimer's disease dementia with mri beyond the short-term: Implications for the design of predictive models. *NeuroImage: Clinical*, 23:101837, 2019.

[31] NHS National Health Service. The risks of drinking too much, 2019. Accessed from `https://www.nhs.uk/live-well/alcohol-advice/the-risks-of-drinking-too-much/`.

[32] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[33] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[34] Enrico Pellegrini, Lucia Ballerini, Maria del C Valdes Hernandez, Francesca M Chappell, Victor González-Castro, Devasuda Anblagan, Samuel Danso, Susana Muñoz-Maniega, Dominic Job, Cyril Pernet, et al. Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:519–535, 2018.

[35] Yi-Hao Peng, Biing-Ru Wu, Ching-Hua Su, Wei-Chih Liao, Chih-Hsin Muo, Te-Chun Hsia, and Chia-Hung Kao. Adult asthma increases dementia risk: a nationwide cohort study. *J Epidemiol Community Health*, 69(2):123–128, 2015.

[36] Martin Prince, Renata Bryce, and Cleusa Ferri. World alzheimer report 2011: The benefits of early diagnosis and intervention. 2018.

[37] Craig W Ritchie and Karen Ritchie. The prevent study: a prospective cohort study to identify mid-life biomarkers of late-onset alzheimer's disease. *BMJ open*, 2(6):e001893, 2012.

[38] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85, 2022.

[39] A. Solomon, M. Kivipelto, J. L. Molinuevo, B. Tom, and C. W. Ritchie. European prevention of alzheimer's dementia longitudinal cohort study (epad lcs): study protocol. *BMJ Open*, 8(12):e021017, 2019.

[40] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[41] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):1–11, 2008.

[42] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):1–21, 2007.

[43] E. Y. Tang, S. L. Harrison, L. Errington, M. F. Gordon, P. J. Visser, G. Novak, C. Dufouil, C. Brayne, L. Robinson, L. J. Launer, and B. C. Stephan. Current developments in dementia risk prediction modelling: An updated systematic review. *PLoS One*, 10(9):e0136181, 2015.

[44] Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

[45] Manuela Tondelli, Gordon K Wilcock, Paolo Nichelli, Celeste A De Jager, Mark Jenkinson, and Giovanna Zamboni. Structural mri changes detectable up to ten years before clinical alzheimer's disease. *Neurobiology of aging*, 33(4):825–e25, 2012.

[46] Giacomo Tondo, Fabiola De Marchi, Emanuela Terazzi, Paolo Prandi, Marta Sacchetti, Cristoforo Comi, and Roberto Cantello. Chronic obstructive pulmonary disease may complicate alzheimer's disease: a comorbidity problem. *Neurological Sciences*, 39(9):1585–1589, 2018.

[47] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[48] Andrew J Vickers and Elena B Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006.

[49] Gabriel Wardi, Morgan Carlile, Andre Holder, Supreeth Shashikumar, Stephen R Hayden, and Shamim Nemati. Predicting progression to septic shock in the emergency department using an externally generalizable machine-learning algorithm. *Annals of emergency medicine*, 77(4):395–406, 2021.

[50] WHO World Health Organization. A healthy lifestyle - who recommendations, 2010. Accessed from `https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations`.

[51] Chun-Yun Zhang, Fang-Fang He, Hua Su, Chun Zhang, and Xian-Fang Meng. Association between chronic kidney disease and alzheimer's disease: an update. *Metabolic Brain Disease*, 35(6):883–894, 2020.