IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.1120000

Interpretable data-driven approach based on feature selection methods and GAN-based models for cardiovascular risk prediction in diabetic patients

DAVID CHUSHIG-MUZO[1], HUGO CALERO-DÍAZ[1], FRANCISCO J. LARA-ABELENDA[1], VANESA GÓMEZ-MARTÍNEZ[1], CONCEIÇÃO GRANJA[2] and CRISTINA SOGUERO-RUIZ[1]

[1]Department of Signal Theory and Communications, Telematics and Computing, Rey Juan Carlos University, Madrid, Fuenlabrada 28943 Spain (e-mail: david.chushig@urjc.es; h.calero.2017@alumnos.urjc.es; francisco.lara@urjc.es; vanesa.gomez@urjc.es; cristina.soguero@urjc.es)

[2]Norwegian Centre for E-health Research, University Hospital of North Norway

Corresponding author: David Chushig-Muzo (e-mail: david.chushig@urjc.es).

This work was supported by the European Commission through the H2020-EU.3.1.4.2, European Project WARIFA (Watching the risk factors: Artificial intelligence and the prevention of chronic conditions) under the Grant Agreement 101017385; by the Spanish Government by the Grant AAVis-BMR PID2019-107768RA-I00/AEI/10.13039/50110 0011033; by the Community of Madrid (YEI grant TIC-11649); and by Rey Juan Carlos University (2023/SOLCON-132212).

ABSTRACT Noncommunicable diseases (NCDs) are the leading cause of morbidity and mortality worldwide. Cardiovascular diseases (CVDs) and diabetes are the most prevalent NCDs, causing 1.9 and 1.5 million deaths yearly. Individuals diagnosed with type 1 diabetes (T1D) are at high risk of developing CVDs. Machine learning (ML) models have provided outstanding results in different domains, including healthcare, allowing to obtain models with high predictive performance. The aim of this study was to develop an interpretable data-driven approach to predict the 10-year CVD risk for T1D older individuals, aiming to provide both reasonable predictive performance and the identification of risk factors associated with CVDs. Data from T1D individuals at the Steno Diabetes Center Copenhagen were used. Different ML-based models were considered, including KNN, decision tree, random forest, and multilayer perceptron (MLP). To enhance the predictive performance of ML models, the conditional tabular generative adversarial network (CTGAN) was used to create synthetic data and increase the size of the training data. Several filter and wrapper feature selection (FS) techniques were considered for identifying the most relevant features involved in CVD risk and enhancing the performance of the ML-based models used. To gain interpretability on predictive models, we used the post-hoc methods: SHAP and accumulated local effects. The experimental results showed a great performance of FS and ML-based models for predicting CVD risk. In particular, the MLP obtained the best results, with a mean absolute error of 0.0088 and mean relative absolute error of 0.0817. Regarding risk factors, age, Hba1c, and albuminuria were identified as crucial in CVD risk prediction, which is in line with recent clinical evidence. Our study contributes to identifying CVD risk and associated risk factors in a data-driven manner, helping to make early interventions and adequate treatments to prevent CVDs.

INDEX TERMS Cardiovascular risk prediction, type 1 diabetes, machine learning, interpretable methods, feature selection, generative adversarial networks, accumulated local effects, post-hoc interpretability, ctgan

I. INTRODUCTION

NONCOMMUNICABLE diseases (NCDs) have become a global health and economic issue in modern society. Recent reports from the World Health Organization identified NCDs as the leading cause of disability and morbidity worldwide [1]. Cardiovascular diseases (CVDs) and diabetes are among the most prevalent NCDs, causing 1.9 and 1.5 million deaths per year, respectively [2]. According to the International Diabetes Federation [3] approximately 700 million individuals will develop diabetes by 2045 [4]. Previous studies have shown that the risk of developing cardiovascular events is higher in prediabetic cohorts than in cohorts of

healthy individuals [5]. Furthermore, epidemiological studies have shown that the risk of developing CVD is higher in individuals with type 1 diabetes (T1D) [6], [7]. Although T1D is frequently diagnosed in children and youth, many cases have been reported in adulthood [8]. People with NCDs significantly increase the cost and demand for healthcare services owing to multiple hospitalizations, adverse events, and frequent visits to primary and specialized care [4]. Early identification of CVD cases and effective interventions are crucial for reducing both health and economic burden [1], [9].

Risk calculators have supported public health stakeholders in the identification of individuals at high risk of CVD, fostering early clinical interventions and reducing acute events and associated mortality risk [10]. Several clinical guidelines recommend the use of risk models as the first step in decision-making, primary prevention and the design of risk-reducing strategies [10], [11]. Over the last few years, various CVD risk calculators have been developed, including the Framingham risk score [12], the systematic coronary risk evaluation [13], the Reynolds risk score [14], the PROCAM calculator [15] among others [16], [17]. Regarding CVD risk prediction models for diabetic cohorts, three approaches have been extensively employed [18]: *(i)* considering diabetes as a CVD risk factor and treating diabetic patients as high-risk patients; *(ii)* applying risk models developed using cohorts of healthy individuals to diabetic populations; and *(ii)* developing diabetes-specific risk prediction models. Of the 45 CVD risk prediction models identified for patients with diabetes [19], 33 corresponded to the second approach, and only 12 were developed using data from cohorts diagnosed with diabetes [20]. Additionally, the first and second approaches are not robust because the pathogenesis of CVD in diabetic patients is multifactorial and presents significant heterogeneity owing to the presence of other comorbidities [21]. While the high prevalence of CVD among people with type 2 diabetes (T2D) has been widely studied and recognized over the past several decades, the link between T1D and CVD has been less studied.

Several CVD risk models have been developed for T2D patients, but a few risk calculators have been created and validated using data from T1D individuals [22]. In the literature, some studies have proposed risk engines focused on T1D, such as the Swedish T1D risk score (SWT1RS) [23], the Scottish T1D risk score (SCT1RE) [24] and the Danish Steno T1 Risk Engine (ST1RE) [22]. The SWT1RS considered eight features: diabetes mellitus (DM) duration, age at onset of T1D, log ratio of total cholesterol, high-density lipoprotein (HDL), glycosylated hemoglobin (HbA1c), systolic blood pressure (SBP), smoking, macroalbuminuria, and if the patient had previous CVD. The SCT1RS used nine features: age, sex, HbA1c, EGFR, HDL, DM duration, smoking status, antihypertensive treatment, and statin therapy. The ST1RE considered age, sex, diabetes duration, SBP, low-density lipoprotein (LDL), HbA1c, albuminuria, estimated glomerular filtration rate (EGFR), and lifestyle habits such as smoking and exercise. In this study, ST1RE was used to obtain the CVD risk for diabetic cohorts, since it considers different types of albuminuria, lifestyle and clinical features.

In the clinical setting, several studies have explored the use of machine learning (ML) models in a range of applications such as disease prediction, identification of risk factors, prediction of adverse events among others [25]–[28]. ML field has not been only limited to the development of predictive models, and have been successfully used to identify disease risk factors [29]. In critical domains such as healthcare, understanding how models reach predictions is of paramount importance for the implementation and adoption of ML-based models in clinical practice [30]. The goal is not only to create models with high predictive performance but also to obtain transparent and interpretable models [31]. Recently, methods that provide post-hoc explanations of model predictions, such as Shapley additive explanations (SHAP) [32] and accumulated local effects (ALE) [33], have received considerable attention [30], [31]. Despite the great benefits of ML, in several applications and domains, the generalization and performance of models are limited by the number of samples in the datasets. To address this, several resampling approaches have been proposed for generating synthetic data and increasing the amount of data used for model training. Among them, generative adversarial networks (GANs) have provided remarkable results for generating high-quality data in computer vision [34]. In this study, the GAN-based model named conditional tabular GAN (CTGAN) [35], which has shown excellent performance in previous studies [36], [37], has been used to create synthetic tabular data that help to enhance predictive results.

In this study, we developed an interpretable data-driven approach to predict the 10-year CVD risk in T1D older individuals, aiming to provide both reasonable predictive performance and interpretability in the identification of risk factors associated with CVDs. To conduct this study, we used data collected from patients diagnosed with T1D at the *Steno Diabetes Center Copenhagen* [22]. Different ML-based models were considered, including the K-nearest neighbors (KNN), decision tree (DT), random forest (RF), and multilayer perceptron (MLP). To enhance the predictive performance of these models, the oversampling model CTGAN was used to create synthetic data and combine them with real patient data. Several filter and wrapper feature selection (FS) techniques were considered for identifying most relevant features involved in the development of CVD, and thus enhancing the performance of ML models. To identify the most relevant features and gain interpretability on ML-based models trained for CVD risk prediction, we used the post-hoc methods: SHAP [32] and ALE [33]. To the best of our knowledge, this paper is one of the first that explores GAN-based models for tabular data augmentation in combination with filter and wrapper FS methods for enhancing CVD risk prediction in T1D older individuals.

The rest of the paper is organized as follows. Section II describes the dataset and preprocessing stage. Section III presents the methods employed in this work. Section IV

shows the experimental setup, the results achieved by ML models for predicting CVD risk, and the analysis of CVD risk factors identified through FS and post-hoc methods. Finally, Section V and Section VI presents discussion and conclusions, respectively.

II. DATASET DESCRIPTION AND PREPROCESSING

This section presents the dataset used and the preprocessing stage. In this study, we employed data collected from 1,000 Danish adults diagnosed with T1D and treated at the *Steno Diabetes Center Copenhagen* [22]. Patients with previous CVD events were excluded, resulting in a dataset of 677 individuals. A total of 10 features were considered, including age, sex, smoking, exercise, DM duration (in years), SBP (in mmHg), LDL (in mmol/l), Hba1c (in mmol/mol), EGFR (in ml/min/1.72m^2), and albuminuria. All features were continuous except for the binary features of sex, smoking, and exercise, and the categorical feature albuminuria. Three categories of albuminuria, differing in the urinary albumin-to-creatinine ratio, were available: normoalbuminuria ($<$30 mg/g), microalbuminuria (30–299 mg/g), and macroalbuminuria ($\geq$ 300 mg/g). The one-hot encoding [38] was used for transforming the original feature (albuminuria) into three new binary features named: normoalbuminuria, microalbuminuria, and macroalbuminuria. Smoking was coded as '0' (absence) and '1' (presence). Regularly exercise was coded as '1' and '0' otherwise. Regarding sex, men and women were coded as '0' and '1', respectively. None of the features in the dataset contained outliers or missing data.

In Figure 1, histograms and bar plots were used to visualize the distribution of continuous and binary features. It was observed that patients were adults with a mean age of 45 years. Lifestyle information showed that most patients did not exercise regularly and had a high smoking rate. Regarding the types of albuminuria, most of individuals presented normoalbuminuria, and a few patients had microalbuminuria and macroalbuminuria. ST1RE [22] was used as risk calculator to obtain the 10-year CVD risk in T1D patients. Contrary to traditional CVD risk calculators that only considered age, sex, SBP, LDL-cholesterol, EGFR, ST1RE included information on DM duration, HbA1c, albuminuria, and patients' lifestyle (smoking and physical activity). The resulting CVD risk ranged between $[0, 1]$, with 0 and 1 denoting a low and high risk, respectively. In the current work, the 10-year CVD risk was considered as the target variable and used to train predictive ML-based models.

III. METHODS

In this subsection, we first introduced the ML models used to predict the CVD risk in T1D individuals. Then, we detailed the data augmentation methods and discuss how they are applied to the current study to enhance the models' performance for predicting CVD risk. Finally, we presented the FS and post-hoc interpretability methods to identify the most relevant CVD risk features and provide model interpretability. A schematic of the interpretable and data-driven workflow proposed in this study is shown in Figure 2.

A. NOTATION

Let an input dataset $\mathcal{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ consisting of N samples, with the i-th sample represented by a vector $\mathbf{x}^{(i)} = [x_1^{(i)}, \ldots, x_D^{(i)}] \in \mathbb{R}^D$, where D is the number of features. The corresponding target (10-year CVD risk by ST1RE) is identified by $\mathbf{y} = [y_1, \ldots, y_N]$. In this work, we estimated the CVD risk (defined as $\hat{y}_i$) using several ML-based models. We split the input dataset $\mathcal{X}$ into train subset $\mathcal{X}_{train}$ and test subset $\mathcal{X}_{test}$, with 70% and 30% of the samples, respectively. The training subset was only used for training the models, whereas the test subset for evaluating the trained models. Five different partitions of train and test subsets were considered to evaluate the generalization capability of predictive models. The mean absolute error (MAE) and the mean relative absolute error (MRAE) were considered as figures of merit, defined as follows: $MAE = \frac{1}{N_t}\sum_{i=1}^{N_t}|y_i - \hat{y}_i|$, $MRAE = \sum_{i=1}^{N_t}|y_i - \hat{y}_i|/\hat{y}_i$, being N_t the size of the test subset, x_i the i-th test sample, and y_i and $\hat{y}_i$ the true CVD risk and the predicted risk, respectively.

B. ML-BASED MODELS TO PREDICT CVD RISK

In this study, due to the flexibility and high performance of ML-based models compared to traditional statistical techniques, the KNN, DT, RF, and MLP models were used to predict the 10-year CVD risk for T1D adults.

KNN is a nonparametric and nonlinear model that uses dissimilarity measures to make predictions [39]. Unlike parametric models, KNN does not make any assumptions regarding the underlying data distribution, making it highly flexible and suitable for a wide range of applications [40]. Formally, given a sample $\mathbf{x}_i$ belonging to the test subset, KNN computes the similarity measure between $\mathbf{x}_i$ and all samples in the training subset [41]. These measures are then sorted to find smaller values and thus find the corresponding K-nearest neighbors. The prediction of $\mathbf{x}_i$ is the mean of the outputs of its K nearest neighbors. In the algorithm, both the distance measure and the number of neighbors K are crucial for achieving reasonable predictive results.

DT is a nonparametric and nonlinear model that divides complex decisions into simpler ones and organizes them hierarchically with a tree-like structure [42]. The feature space is iteratively partitioned into regions containing homogeneous sets of samples. Each partition (split) in the feature space is represented as a new node in a tree-like structure [42]. DTs are very popular in the clinical field because of their interpretability, which provides visualization of decision-making processes [43]. During the tuning phase, several hyperparameters, such as the splitting criterion, the minimum number of samples for splitting, and the maximum depth of the tree, need to be assigned. This is particularly relevant for this algorithm because DT tends to cause overfitting when the tree becomes overly complex, posing a challenge for achieving generalization [42].

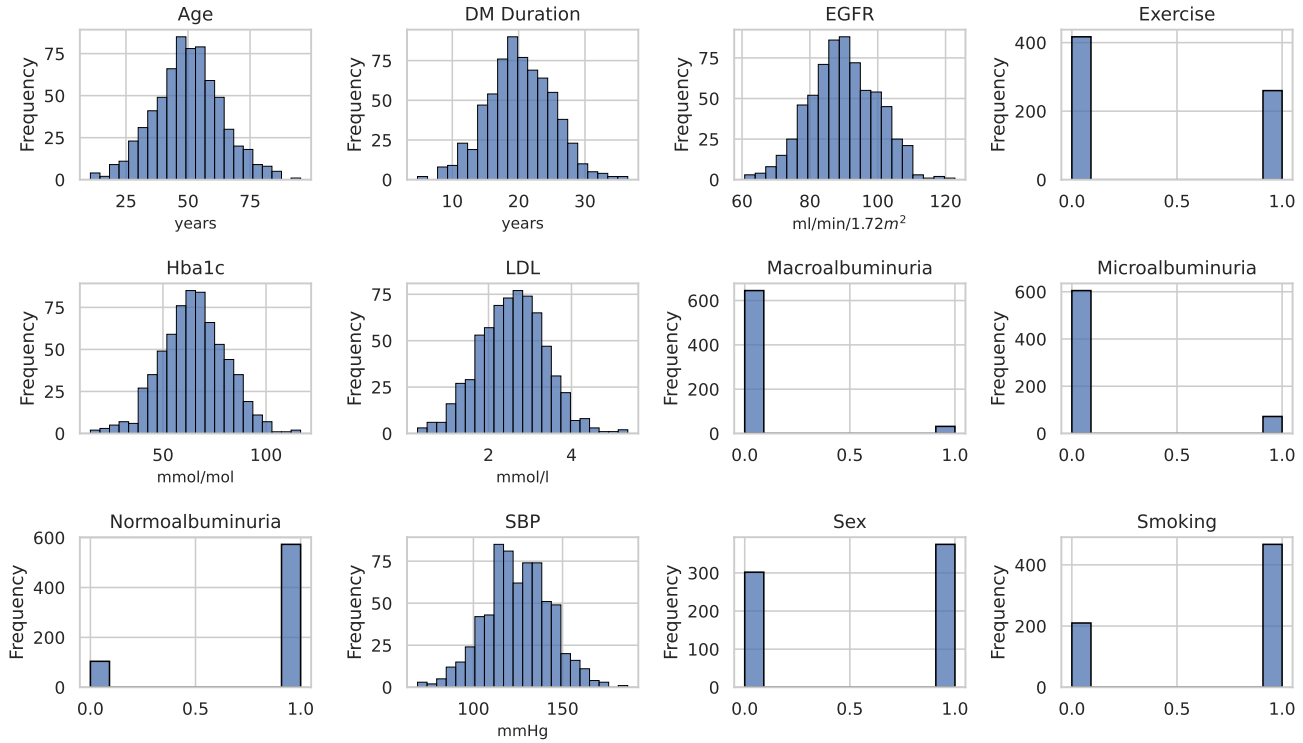

FIGURE 1: Histograms and bar plots associated with features of the dataset collected by the *Steno Diabetes Center Copenhagen*.

RF is a nonparametric ensemble model that combines multiple DTs to make predictions [44]. Initially, from the training subset, RF employs a bagging sampling method to generate M training sets, each containing a similar number of samples [45]. Subsequently, using these M training subsets, RF constructs an ensemble of M DTs, defined as $\{p_{r1}(\mathbf{x}), p_{r2}(\mathbf{x}), \ldots, p_{rM}(\mathbf{x})\}$, obtaining M predictions $\{\hat{y}_1 = p_{r1}(\mathbf{x}), \hat{y}_2 = p_{r2}(\mathbf{x}), \ldots, \hat{y}_M = p_{rM}(\mathbf{x})\}$. Finally, the prediction is determined by averaging the results of M DTs [45]. The hyperparameters typically explored for RF include the number of estimators (trees) and the number of features considered at each split.

MLP is a feed-forward Artificial Neural Network (ANN) consisting of an input layer, one or more hidden layers, and an output layer, which are interconnected by processing units called neurons [46]. Each neuron within a layer is connected to the other neurons in successive layers through weighted connections. During training, the weights are randomly initialized, and the objective is to learn the optimal weight values that minimize the error between the estimated output of MLP and the real target. This is achieved using the back-propagation algorithm combined with stochastic gradient descent [47], which adjusts the weights of the ANN in a supervised manner to minimize the error [48]. Furthermore, MLP has several hyperparameters that need to be carefully tuned to optimize its performance. In this study, we explored different numbers of neurons in hidden layers, activation functions, the optimizer among others.

C. OVERSAMPLING METHODS

In the literature, a variety of techniques have been proposed to increase the size of training subsets and improve predictive results [49]. Resampling methods that create synthetic samples for minority classes [50] have received considerable attention because of their computational efficiency and versatility [51]. However, in the clinical setting, datasets are generally characterized by a high degree of heterogeneity, and present mixed-type data with numerical and categorical features [35]. Most oversampling techniques are designed to work with numerical features and do not perform adequately when mixed-type data are used. Recently, generative adversarial networks (GANs) [52] have gained great popularity due to their impressive results in generating synthetic data, especially in computer vision applications [53]. GANs are generative models that train two networks simultaneously through an adversarial process: a *generator* G and *discriminator* D. While G aims to produce synthetic samples, D strives to differentiate between real and synthetic samples. Despite the benefits of GANs in multiple applications [54], these models present several challenges for generating tabular and mixed-type data. Recently, a novel GAN-based model named CTGAN [35] has been proposed to address these limitations. CTGAN uses a mode-specific normalization to solve the problem of non-Gaussian distributions in numerical variables, and a conditional generator to address imbalanced categorical features [35]. Wasserstein divergence and the weight clipping with a gradient penalty have also been used

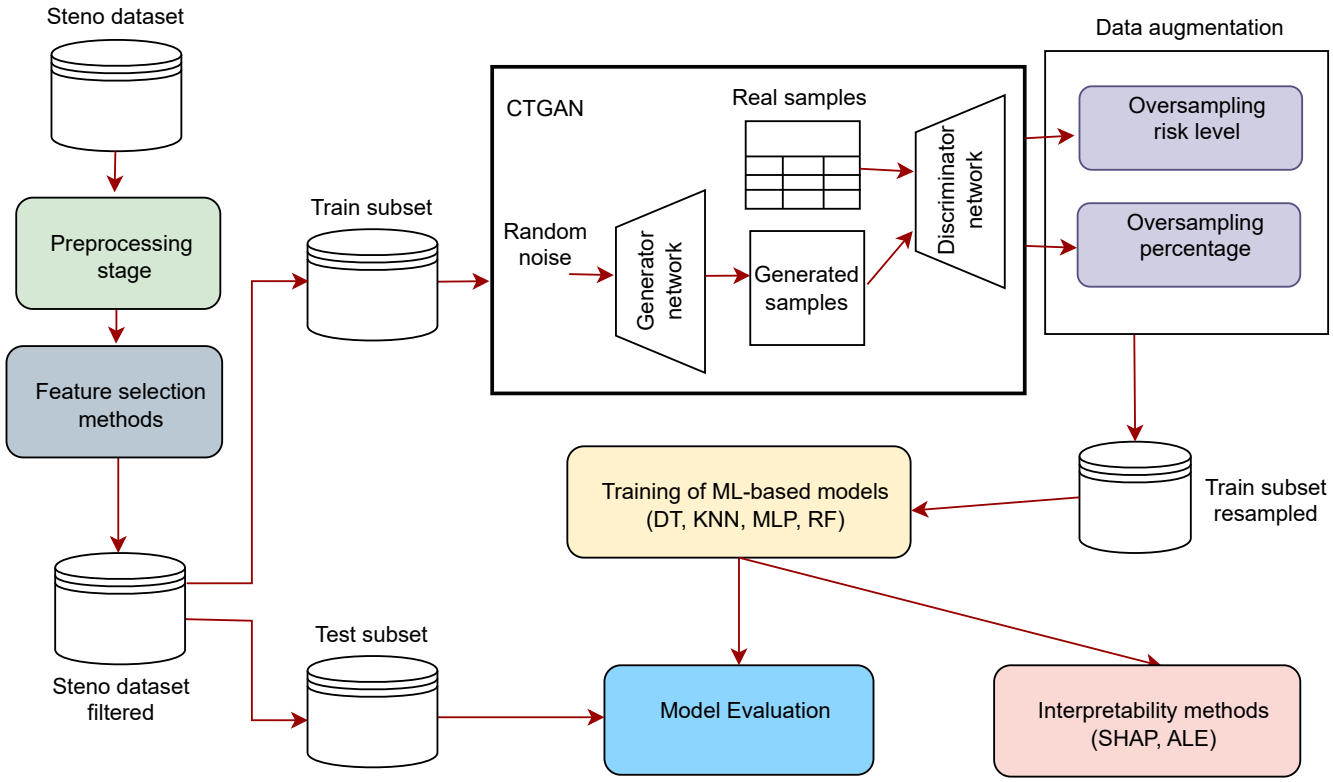

FIGURE 2: Workflow of the interpretable and data-driven approach for CVD risk prediction.

to enhance synthetic data [35]. In a previous study [36], the authors assessed the performance of GAN-based models to create synthetic data with numerical and categorical features. CTGAN exhibited the best performance to create synthetic data by maintaining intrinsic characteristics from the original data, leading to improvements in subsequent predictive tasks.

In this study, two data augmentation strategies were used: *over-per* and *over-level*. In *over-per*, synthetic samples were generated using a fixed percentage of all samples from the training subset $\mathcal{X}_{train}$. The percentage of samples was selected from several values within the range $[1, 20]$, being 5% which provided an improvement in CVD-risk prediction. In *over-level*, the number of synthetic samples was based on different CVD risk levels. First, we categorized the 10-year CVD risk (provided by ST1RE) into three levels: low, intermediate and high. These levels were identified using the risk stratification guidelines from the *National Institute for Health and Care Excellence* [55] and by setting specific risk cut-off values, thus distinguishing: *(i)* low-risk patients (CVD risk <0.1), *(ii)* moderate-risk patients (CVD risk in the range $[0.1, 0.2)$); and *(iii)* high-risk patients (CVD risk ≥ 0.2). We split the individuals in $\mathcal{X}_{train}$ into three groups (CVD risk levels), identifying the one with the most samples (moderate-risk patients). Then, the number of new samples for the low-risk and high-risk groups was created by taking the number of the moderate-risk group as reference.

D. FEATURE SELECTION METHODS

FS methods choose a subset of features and aim to achieve several objectives [56]–[58]: *(i)* overcoming the curse of dimensionality; *(ii)* reducing the computational cost for training models; *(iii)* improving generalization capacity and predictive performance in subsequent tasks; and *(iv)* enhancing interpretability. FS methods are classified into three categories: filter, embedded and wrapper methods [56]. Since no single FS method can guarantee optimal results in terms of both predictive performance and stability of selection, this study explored several FS methods. Specifically, a variety of filter and wrapper FS techniques are considered for: *(i)* selecting the most relevant features that help to improve model performance in predicting CVD; and *(ii)* identifying those features that play a significant role in the development of CVD.

Filter methods select features that present a strong relationship with the target and work independently of any predictive model [59]. They evaluate features based on specific scoring criteria, such as statistical tests, mutual information (MI), or dissimilarity measures, and subsequently select a subset of features with the highest scores, discarding those deemed irrelevant [59]. In this study, we employed the minimal redundancy and maximal relevance (mRMR) [60] and Relief [61] methods as filter-based FS methods. These methods were chosen because of their computational efficiency and their ability to identify relevant features that contribute to enhancing performance in subsequent predictive tasks.

mRMR method performs FS by minimizing the re-

dundancy among features and maximizing the relevance of the features to the target [60]. To compute redundancy and relevance, mRMR uses MI, which quantifies the amount of information that one random variable encompasses about another [60]. Formally, given two random variables $\mathcal{X}$ and $\mathcal{Y}$, the MI is computed as follows [62]: $\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\rho(x,y)log(\frac{\rho(x,y)}{\rho(x)\rho(y)})$, where $\rho(x,y)$ denotes the joint probability density function (PDF), $\rho(x)$, and $\rho(y)$ represent the PDFs of $\mathcal{X}$ and $\mathcal{Y}$, respectively. Relief measures the relevance of features by uncovering the dependencies between the features and target [61]. The algorithm is as follows. First, a sample is randomly selected, and then the feature vectors of the nearest samples from both the same class and different classes are identified. This process allows to rank the importance of each feature individually, similar to a univariate approaches, while consider dependencies among other features. Relief has been extensively used because of its simplicity and effectiveness, particularly in high-dimensional feature spaces [61].

Wrapper methods iteratively train prediction models by searching for the best feature subset [57]. They use a predictive model to assess the effectiveness of feature subsets using a search strategy, making them computationally complex and time consuming [59]. Despite these drawbacks, they benefit from their interaction with predictive algorithms to identify the best performing feature subsets. This study considers the following wrapper methods: permutation importance (PI) [63] and particle swarm optimization (PSO) [64]. The PI was originally proposed for the RF algorithm [65], and further research was developed to create a model-agnostic FS method [66]. It quantifies feature importance by measuring the change in a specific figure of merit (*e.g.,* accuracy for classification, MAE for regression) when a feature is excluded as an input to obtain model predictions [63]. The importance is assessed through the feature importance difference (FID), which is calculated as the difference between a reference score and a corrupted score [67]. The reference score is derived using the original features, whereas the corrupted score is the average after shuffling features a fixed number of times [67]. A feature is considered significant if shuffling significantly affects the score (high FID value), indicating its strong impact on model predictions. MRAE was chosen to evaluate the impact of permuting features on the predictive performance of ML-based models [67].

PSO is a metaheuristic optimization algorithm inspired by the collective behavior of swarms in nature, such as bird flocking [64]. In the context of optimization problems, a swarm is conceptualized as a group of particles, where each particle represents a potential solution [68]. Similar to how a flock of birds collectively searches for the best landing spot, PSO iteratively seeks an optimal solution by simulating the movement of particles within a search space. Regarding FS, each particle in PSO represents a potential solution, characterized by a multidimensional position vector and a multidimensional velocity vector [69]. In the former, dimensionality is equal to the number of features, and each dimension represents the probability of a feature to be selected. The velocity of the particle is updated after each iteration, depending on the particle's best position and the global best position, which are determined using a fitness function [69]. PSO finds the optimal regions of complex search spaces through the interaction of individuals in a population [70]. In contrast to other optimization algorithms, PSO presents global search capability, high computational efficiency, fast convergence rate, minimal parameter tuning requirements, and avoids local minima [71].

E. POST-HOC INTERPRETABILITY METHODS FOR IDENTIFYING CVD RISK FACTORS

In the clinical setting, obtaining models with high predictive performance is not sufficient for physicians and clinical researchers, and it is crucial to understand why models provide a particular outcome. Owing to ever-increasing advances in ML for healthcare, it is paramount to provide interpretability to trained models [30]. Interpretability is defined as the process of generating human-understandable explanations of outcomes provided by computational models [43]. The interpretability in supervised approaches aims to explain how predictions are achieved for any given input [72]. Several methods have been developed for model interpretability, being post-hoc and model-agnostic approaches the most used [73]. These techniques can be categorized into global and local approaches [74]. Global approaches describe the overall behavior of a model, whereas local approaches aim to explain how the models reached a prediction for a specific input. In this study, two post-hoc and global methods were considered: SHAP [32] and ALE [33].

SHAP is a post-hoc interpretability method that identifies the features that significantly impact on the model's predictions [32]. SHAP uses Shapley values from coalitional game theory, combining optimal credit allocation and local explanations [32]. Each feature value of a data sample is conceptualized as a player in a game, where the prediction of the sample minus the average prediction for the dataset is considered the payout [32]. Shapley values ensure fair distribution of this payout among players based on their contribution to the output, thus explaining the average marginal contribution of a feature value across all possible coalitions [74]. Summary plots are commonly used to visualize the Shapley values and feature importance [32]. These plots combine the feature importance for a prediction task, with each point representing the Shapley value given for a feature in a particular sample. The features are organized in decreasing order of importance for model prediction on the vertical axis, the horizontal axis shows the Shapley value, and a color bar is used to show the value of the feature for each sample. In the following section, we provide an example of a SHAP plot using the results from the trained ML models.

ALE is a post-hoc interpretability method proposed as an alternative to partial dependence plots (PDPs) [33]. Although both PDPs and ALE aim to visualize how features impact on

the model's predictions [33], [74], PDPs present two main disadvantages. PDPs are computationally expensive, and when features present high correlations, they are unreliable because their development involves including artificial data samples that are not representative of the original data [75]. By contrast, ALE computes the differences in predictions averaged over the conditional distribution of each feature [33]. This approach avoids the need for artificial data samples and provides more reliable interpretations, particularly in the presence of correlated features [33]. ALE allows the visualization of the effect of feature interactions and provides insights into how the model's predictions change with variations in feature values [33]. For example, ALE can be used to analyze the effect of the interaction between two features by averaging the changes in model's predictions.

IV. EXPERIMENTAL RESULTS

In this section, we analyze the effectiveness of combining different FS methods and ML-based models for predicting CVD risk in T1D adults. We first present the experimental setup, and then an extended comparison of the predictive performance of ML-based models by using all features and those selected by FS methods. Finally, we identified the risk factors involved in the development of CVD using FS and post-hoc interpretability methods.

A. EXPERIMENTAL SETUP

In this study, ML models DT, KNN, MLP, and RF are used to predict 10-year CVD risk in individuals diagnosed with T1D. Although DT and KNN were used in a previous work [76], the overall results were not the highest because of the scarcity of samples in the dataset. We extend and evaluate the effectiveness of tabular data augmentation models based on GANs for generating synthetic mixed-type data that leads to improved predictive performance, thus achieving better CVD risk prediction. The source code for reproducibility of results is available in github.com/ai4healthurjc/cvd-risk-fs-ctgan.

To find the best hyperparameters of ML models, k-fold cross-validation (CV) [77] was performed, with $k = 5$ and the MRAE as the figure of merit. The following hyperparameters were explored: for DT, the split criterion (Gini, entropy), the maximum depth in the range $[2, 12]$ and the minimum samples per split in the range $[2, 20]$; for KNN, K values between $[1, 15]$; for RF, the number of samples per split between $[2, 6]$ and the number of estimators in the range $[10, 40]$; and for MLP, the number of neurons, and the weight initialization approach (random, uniform, Glorot) were examined. Specifically, we selected an architecture composed of m_n inputs (the same size as the input features, D), a single hidden layer with h neurons, and a single neuron in the output layer. Different numbers of neurons in the hidden layers were explored, including $\{2, 4, 6, 8, 10, 12\}$. ReLu was considered as the activation function for neurons in the input and hidden layers and the sigmoid activation function for the output layer. Adam optimization was considered, and the mean squared error was used as the loss function.

In this study, we analyzed and compared the impact of different feature subsets (FES) of the original dataset provided by the *Steno Diabetes Center Copenhagen* [22] to identify 10-year CVD risk. Several subsets were selected using filter FS (mRMR, Relief) and wrapper FS (PI, PSO) methods, which are described as follows.

1) *FES1* contains all features of the dataset;
2) *FES2* considers demographics (age and sex) and lifestyle features (exercise and smoking);
3) *FES3* contains selected features by the PI method;
4) *FES4* contains selected features by the PSO;
5) *FES5* contains selected features by mRMR;
6) *FES6* contains selected features by Relief;

B. PREDICTING CVD RISK FOR T1D PATIENTS

Table 1 shows the predictive results (measured by MAE and MRAE) using different FES, ML models and oversampling strategies. As stated, both *FES1* and *FES2* are independent of FS methods, with the former considering all features, whereas the latter only selects demographic and lifestyle features (age, sex, smoking, and exercise). For *FES1*, MLP achieved the best predictive performance, obtaining the highest values for MAE and MRAE, with 0.0112 ± 0.0011 and 0.1072 ± 0.0056, respectively. Most ML-based models trained using real and synthetic samples generated by CTGAN (considering *over-per* and *over-level*) achieved a slight improvement in MAE and MRAE. The lowest predictive results were obtained using *FES2*, demonstrating that the clinical features play a crucial role in CVD prediction.

Regarding *FES3* and *FES4*, it can be observed that the selection of features by the wrapper FS methods improved the predictive results compared to *FES1* (all features). As argued, the features selected in these approaches depend directly on the ML algorithm. Overall, the features selected using PSO (*FES4*) achieved the best MAE and MRAE values, with 0.0088 ± 0.0006 and 0.0817 ± 0.0129, respectively, outperforming the results of the different FES. The feature subset obtained by PI (*FES3*) also provided reasonable results, performing the same as when using the entire set of features. Regarding the filter methods (*FES5* and *FES6*), the features selected by mRMR and Relief led to obtain low MAE and MRAE values. In Figure 3, we visually compare the CVD risk obtained with ST1RE versus the estimated CVD risk obtained using ML models. Note that models trained with *FES1* and *FES4* were considered. As shown, MLP was the model most effective, providing less error in the predicted risk for each patient. DT and KNN worked correctly for CVD risk < 0.2, but they presented difficulties otherwise. RF and MLP obtained similar results, and although the errors between real and estimated CVD risks were small in RF, these errors were less marked in MLP.

C. IDENTIFYING CVD RISK FACTORS USING FEATURE SELECTION AND INTERPRETABILITY METHODS

In this subsection, we showed the most informative features selected by FS methods, which led to the identification of risk

TABLE 1: MAE and MRAE obtained by combining different FS methods and ML models. The results without oversampling (WO) and considering *over-per* and *over-level* are shown. The best results for MAE and MRAE are marked in bold.

	Model	MAE			MRAE		
		WO	*over-per*	*over-level*	WO	*over-per*	*over-level*
FES1	DT	0.0361±0.0034	0.0369±0.0031	0.0367±0.0036	0.2354±0.0130	0.2368±0.0072	0.2244±0.0100
	KNN	0.0378±0.0028	0.0349±0.0038	0.0358±0.0032	0.2662±0.0104	0.2541±0.0268	0.2435±0.0207
	MLP	0.0112±0.0011	0.0095±0.0010	0.0107±0.0010	0.1072±0.0056	0.0923±0.0035	0.1037±0.0029
	RF	0.0250±0.0018	0.0238±0.0020	0.0246±0.0019	0.1701±0.0113	0.1639±0.0108	0.1625±0.0106
FES2	DT	0.0449±0.0055	0.0445±0.0022	0.0475±0.0044	0.2993±0.0293	0.2910±0.0182	0.3097±0.0167
	KNN	0.0446±0.0029	0.0439±0.0019	0.0437±0.0038	0.3175±0.0226	0.3106±0.0132	0.3013±0.0123
	MLP	0.0428±0.0020	0.0409±0.0029	0.0431±0.0030	0.3350±0.0168	0.3117±0.0169	0.3380±0.0089
	RF	0.0412±0.0027	0.0405±0.0019	0.0424±0.0028	0.2768±0.0235	0.2746±0.0141	0.2788±0.0197
FES3	DT	0.0331±0.0026	0.0341±0.0024	0.0328±0.0027	0.2197±0.0107	0.2175±0.0042	0.2080±0.0071
	KNN	0.0349±0.0019	0.0328±0.0025	0.0323±0.0003	0.2363±0.0125	0.2246±0.0147	0.2123±0.0067
	MLP	0.0101±0.0012	0.0102±0.0020	0.0098±0.0011	0.0101±0.0004	0.0097±0.0021	0.0091±0.0012
	RF	0.0252±0.0019	0.0238±0.0017	0.0243±0.0016	0.1707±0.0114	0.1623±0.0103	0.1613±0.0099
FES4	DT	0.0338±0.0022	0.0322±0.0030	0.0338±0.0036	0.2155±0.0183	0.2041±0.0209	0.2147±0.0127
	KNN	0.0292±0.0018	0.0275±0.0024	0.0291±0.0008	0.1917±0.0118	0.1844±0.0108	0.1904±0.0072
	MLP	0.0096±0.0012	**0.0088±0.0006**	0.0099±0.0012	0.0971±0.0011	**0.0817±0.0129**	0.0905±0.0032
	RF	0.0240±0.0010	0.0224±0.0005	0.0238±0.0006	0.1610±0.0074	0.1524±0.0080	0.1575±0.0064
FES5	DT	0.0331±0.0023	0.0336±0.0026	0.0333±0.0021	0.2165±0.0089	0.2214±0.0075	0.2089±0.0054
	KNN	0.0303±0.0041	0.0300±0.0033	0.0283±0.0024	0.2038±0.0141	0.2049±0.0143	0.1871±0.0154
	MLP	0.0103±0.0012	0.0101±0.0012	0.0104±0.0020	0.0102±0.0016	0.0103±0.0012	0.0102±0.0012
	RF	0.0253±0.0017	0.0236±0.0015	0.0245±0.0013	0.1715±0.0116	0.1629±0.0114	0.1594±0.0088
FES6	DT	0.0345±0.0036	0.0352±0.0045	0.0316±0.0014	0.2177±0.0075	0.2193±0.0182	0.2042±0.0160
	KNN	0.0335±0.0055	0.0362±0.0053	0.0345±0.0038	0.2255±0.0257	0.2509±0.0301	0.2326±0.0315
	MLP	0.0087±0.0020	0.0092±0.0006	0.0084±0.0019	0.0741±0.0150	0.0856±0.0049	0.0764±0.0158
	RF	0.0245±0.0013	0.0234±0.0016	0.0234±0.0010	0.1624±0.0050	0.1586±0.0107	0.1581±0.0129

factors associated with the development of CVD in T1D patients. Subsequently, we employed two post-hoc interpretability methods to identify the most relevant features that impact in the predictions of ML models.

Figure 4 shows the features selected by the wrapper FS methods (PI and PSO), indicating the frequency of selection of each feature (number of times that features were selected) for five train partitions. The x-axis and y-axis represent the ML model and feature name, respectively. As five partitions were considered, the maximum number of votes could last up to five. We selected only wrapper methods to analyze how the selection of ML models affects the selected features and to measure the stability and robustness of the selection. A consensus on the features selected using different partitions and ML models increase the reliability of the predictive results. By analyzing the features selected in *FES3* (see Figure 4 (a-c)), we can observe that both age and Hba1c were chosen in all cases (unanimity voting), considering data augmentation strategies (*over-per* and *over-level*), and data without oversampling (WO). Normoalbuminuria was the feature that received the third-most votes, reaching five votes by all ML models trained with data with *over-per* and *over-level*, whereas for data WO, the votes were five (MLP and RF) and four (DT and KNN). Regarding the features selected by the PSO (*FES4*) (see Figure 4 (d-f)), the most frequently selected were age, Hba1c, normoalbuminuria, sex, and smoking. Notably, the best predictive results were achieved using the features selected by PSO, which may be due to the variability in the selection of features for different partitions. Note also that age, Hba1c and normoalbuminuria were the features with the most votes, followed by sex and smoking. There was a significant difference with PI, where fewer votes were in favor of sex and smoking features.

To distinguish the impact of features on the CVD risk prediction and determine potential risk factors, post-hoc interpretability methods SHAP and ALE were used. Figure 5 presents the SHAP summary plot that shows the SHAP mean values obtained over five partitions and using *FES4*. In this plot, features are sorted in the y-axis in decreasing order of importance for the predictive task and the x-axis represents the SHAP mean values. As shown, age presented the highest SHAP values, standing out above the rest of the features for all the models (DT, KNN, MLP and RF). The second and third features with the highest SHAP values were Hba1c and normoalbuminuria, respectively. These findings were also supported by the previous analysis, where age, Hba1c and normoalbuminuria were the features most frequently selected by FS methods (see Figure 4).

The results obtained using SHAP were also validated using ALE. Figure 6 shows the ALE plots associated with the features of *FES4* and using the MLP model. The ALE plots for MLP are shown because this model achieved the highest predictive performance (the best MAE and MRAE in Table 1. For continuous features (age, DM duration, EGFR, Hba1c, LDL, SBP), ALE shows the effect of the feature values (x-axis) on the predicted outcome (y-axis). The confidence interval (CI) of the estimated effect on predictions is depicted in gray. For binary variables (exercise, macroalbuminuria, microalbuminuria, normoalbuminuria, sex, and smoking), a bar plot with a line representing the estimated effects and error bars showing the CI are depicted. The number of patients for each feature value is depicted in violet, and the difference in the impact on predictions is represented by a dashed line connecting the average impact of each feature value. Figure 6

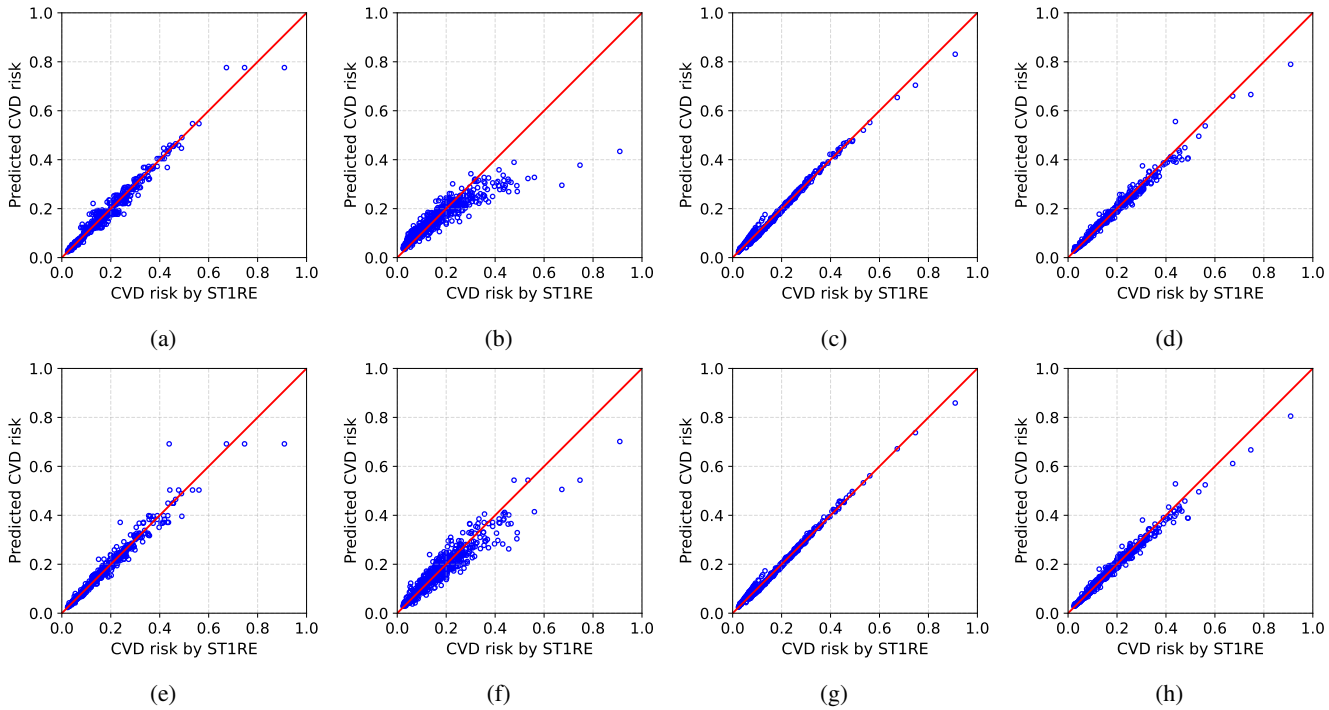

FIGURE 3: Scatter plots that compare the 10-year CVD risk provided by ST1RE (x-axis) and the estimated CVD risk (y-axis) by ML-based models, considering: (a-d) *FES1*; and (e-h) *FES4* using *over-per* and MLP as predictive model since it provided the best results in MAE and MRAE (see Table 1. First column (DT), second column (KNN), third column (MLP) and fourth column (RF).

shows that age has a monotonically increasing effect on the CVD risk. As expected, older patients were at higher risk of developing CVD. After age, Hba1c had a strong effect on the predictions following a growing curve. Note that the impact of presenting age was approximately four-fold higher than Hba1c. SBP and LDL also showed ascending curves, where greater values implied higher CVD risk. Note that SBP, LDL, and DM duration had less impact on the predictions (see the range on the left of the plots), with a maximum effect of prediction between $[0.02, 0.05]$. Normoalbuminuria had a high effect on prediction (0.1), indicating that this type of albuminuria is key for CVD risk prediction, and in line with the results of SHAP feature importance and selected features by FS methods. Macroalbuminuria presented a low effect on prediction. Regarding demographic and lifestyle features (sex, exercise, and smoking), both exercise and sex had a moderate effect on predictions (with a maximum effect estimation of 0.02).

V. DISCUSSION

In this work, we analyzed the effectiveness of ML models for predicting 10-year CVD risk in T1D individuals. First, we evaluated the predictive performance of several ML models, specifically DT, KNN, MLP, and RF. Subsequently, several filter and wrapper FS techniques were used to identify the features most relevant in CVD risk prediction, with the aim of extracting relevant risk factors and improving predictive performance. We also evaluated two oversampling strategies (*over-per* and *over-level*) and using CTGAN to create synthetic samples and improve subsequent predictive tasks. The best predictive results were achieved using the MLP model, employing *over-per* and the features selected by PSO, obtaining a MAE of 0.0088 ± 0.0006 and a MRAE of 0.0817 ± 0.0129.

The post-hoc interpretability methods SHAP and ALE were used to gain interpretability in the trained ML models. By analyzing the results of the wrapper FS techniques and SHAP values, we obtained several valuable insights. Age, Hba1c and normoalbuminuria were mostly selected by the wrapper FS methods (see Figure 4) as the most significant features involved in the prediction of CVD risk. These features have been extensively studied as relevant risk factors for CVD development in clinical studies [78], [79]. Previous research has identified that CVD risk increases with aging, with age being the most important non-modifiable risk factor for the development of CVDs [6]. Regarding Hba1c, several clinical studies [7], [80] have recognized that a high HbA1c level is associated with increased CVD risk. HbA1c provides a measure of average glucose levels over time, reflecting the average plasma glucose level over the previous 8–12 weeks. Despite its benefits, it only provides an approximate measure of glucose control and does not consider short-term glycemic variability, which can indicate its lower impact on the 10-year CVD risk in this study. Although the effect of Hba1c

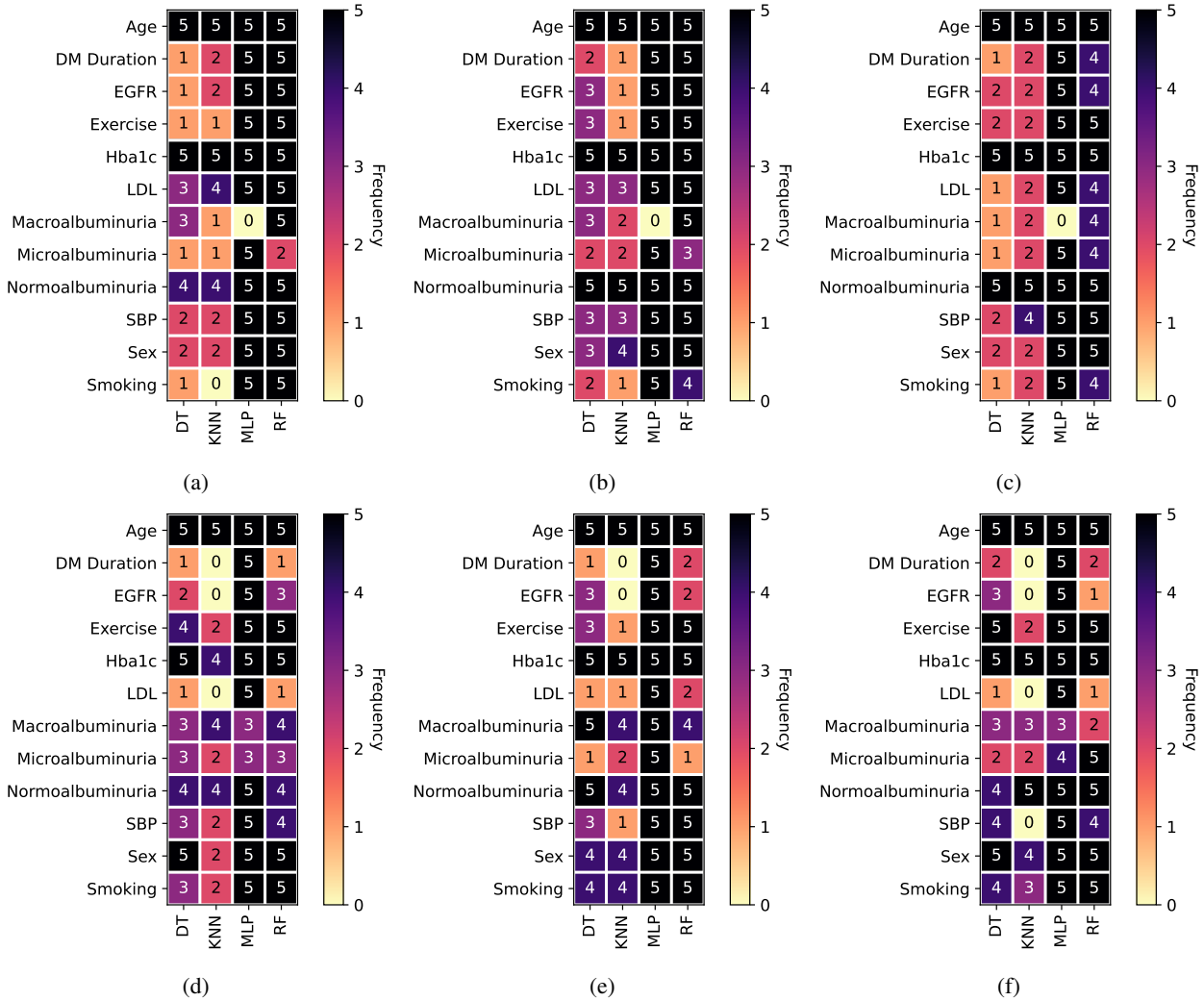

FIGURE 4: Heatmaps that indicate the frequency of selection of features by wrapper FS methods and considering: (a-c) *FES3* (PI); and (d-f) *FES4* (PSO). We show the selected features considering data WO (left panels), and using the oversampling strategies: *over-level* (middle panels) and *over-per* (right panels).

on CVD prediction is lower than that of age, when comparing the corresponding ALE plot (see Figure 6), it can be observed that higher values of Hba1c increase CVD risk.

Several clinical studies have reported that elevated albumin levels can be associated with the onset of several CVDs, such as ischemic heart disease, heart failure, atrial fibrillation, and stroke [81], [82]. In contrast to normoalbuminuria, the relevance of macroalbuminuria and microalbuminuria was significantly lower in all ML models. In the case of SBP and LDL (see Figure 6), models identified a positive relationship with the output, but the magnitude of the effect on prediction was low compared to age or Hba1c. Regarding demographic and lifestyle features, both physical exercise and smoking had a slight impact on model predictions. Several clinical studies have examined behavioral risk factors associated with CVD, the most important being tobacco use followed by physical activity, which are widely recognized as risk factors for different chronic diseases [83], [84]. As stated, our study used a cohort of patients diagnosed with T1D and lifestyle features (*e.g.,* smoking, exercise) have been less associated with CVDs.

VI. CONCLUSION

In this study, we analyzed the performance of several ML-based models for predicting the 10-year CVD risk in older adults with T1D. To improve the performance of these models, we combined filter and wrapper FS methods and tabular data augmentation with the GAN-based model, CTGAN. CTGAN was effective in creating synthetic data for mixed-type data and helped to improve the results of CVD risk prediction. Our methodology, which leverages the advantages of FS methods and data augmentation approaches, provided significant predictive results for identifying CVD risk, with the best figures of merit achieved using MLP and *over-per*

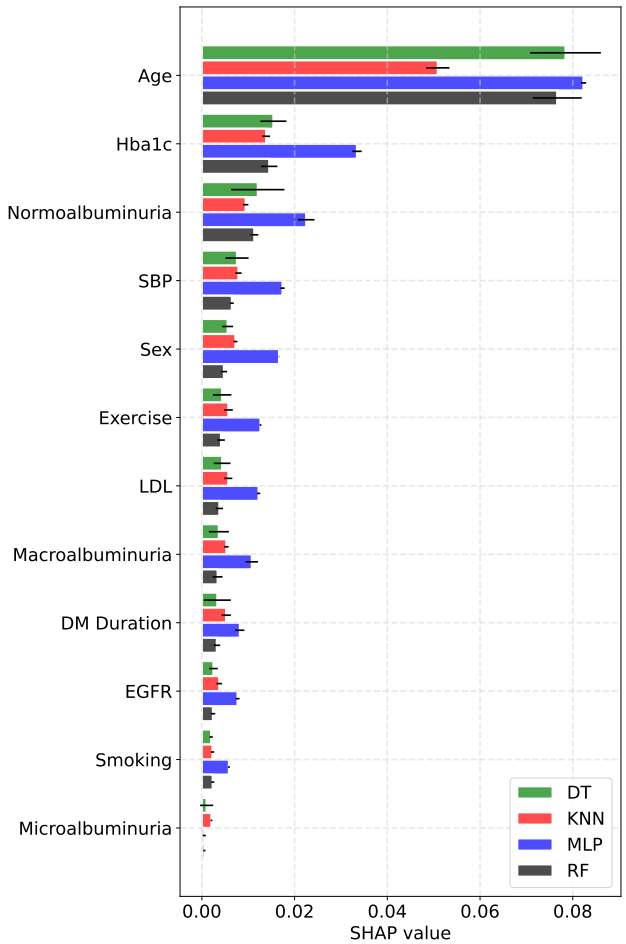

FIGURE 5: SHAP mean values (mean and standard deviation) obtained over the 5 partitions considering features of FES4.

with a MAE and MRAE of 0.0088 and 0.0817, respectively. Our work has also shown that FS methods and post-hoc interpretability methods are capable of identifying risk factors involved in the development of CVD risk, highlighting the importance of non-modifiable factors such as age, Hba1c and albuminuria over 300 mg/g (normoalbuminuria). Among modifiable risk factors, the physical activity was recognized as one of the most important. This study highlights the significance of ML in the clinical setting, particularly for predicting CVD risk in T1D individuals, supporting the creation of automated prediction systems and identification of disease risk factors. ML models are promising for CVD risk assessment and support the identification of high-risk individuals and prevention of the onset of acute clinical events.

ACKNOWLEDGMENT

This work was supported by the European Commission through the H2020-EU.3.1.4.2, European Project WARIFA (Watching the risk factors: Artificial intelligence and the prevention of chronic conditions) under the Grant Agreement 101017385; by the Spanish Government by the Grant AAVis-BMR PID2019-107768RA-I00/AEI/10.13039/501100011033; by the Community of Madrid (YEI grant TIC-11649); and by Rey Juan Carlos University (2023/SOLCON-132212).

REFERENCES

[1] W. H. Organization *et al.*, *Noncommunicable diseases: progress monitor 2022*. World Health Organization, 2022.

[2] R. E. Harris, *Epidemiology of chronic disease: global perspectives*. New York: Jones & Bartlett Learning, 2019.

[3] H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, K. Ogurtsova, B. B. Duncan, C. Stein, A. Basit, J. C. Chan, J. C. Mbanya, *et al.*, "Idf diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes research and clinical practice*, vol. 183, p. 109119, 2022.

[4] D. R. Whiting, L. Guariguata, C. Weil, and J. Shaw, "Idf diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030," *Diabetes Research and Clinical Practice*, vol. 94, no. 3, pp. 311–321, 2011.

[5] R. Garcia-Carretero, L. Vigil-Medina, I. Mora-Jimenez, C. Soguero-Ruiz, R. Goya-Esteban, J. Ramos-Lopez, and O. Barquero-Perez, "Cardiovascular risk assessment in prediabetic patients in a hypertensive population: The role of cystatin c," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 12, no. 5, pp. 625–629, 2018.

[6] B. Vergès, "Cardiovascular disease in type 1 diabetes: A review of epidemiological data and underlying mechanisms," *Diabetes & Metabolism*, vol. 46, no. 6, pp. 442–449, 2020.

[7] I. Bebu, D. Schade, B. Braffett, M. Kosiborod, M. Lopes-Virella, E. Z. Soliman, W. H. Herman, D. A. Bluemke, A. Wallia, T. Orchard, *et al.*, "Risk factors for first and subsequent cvd events in type 1 diabetes: the dcct/edic study," *Diabetes Care*, vol. 43, no. 4, pp. 867–874, 2020.

[8] A. Katsarou, S. Gudbjörnsdottir, A. Rawshani, D. Dabelea, E. Bonifacio, B. J. Anderson, L. M. Jacobsen, D. A. Schatz, and Å. Lernmark, "Type 1 diabetes mellitus," *Nature Reviews Disease Primers*, vol. 3, no. 1, pp. 1–17, 2017.

[9] K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, and P. P. Sengupta, "Machine learning in cardiovascular medicine: are we there yet?," *Heart*, vol. 104, no. 14, pp. 1156–1164, 2018.

[10] A. D. Jamthikar, D. Gupta, L. Saba, N. N. Khanna, K. Viskovic, S. Mavrogeni, J. R. Laird, N. Sattar, A. M. Johri, G. Pareek, *et al.*, "Artificial intelligence framework for predictive cardiovascular and stroke risk assessment models: A narrative review of integrated approaches using carotid ultrasound," *Computers in Biology and Medicine*, vol. 126, p. 104043, 2020.

[11] S.-Y. Cho, S.-H. Kim, S.-H. Kang, K. J. Lee, D. Choi, S. Kang, S. J. Park, T. Kim, C.-H. Yoon, T.-J. Youn, *et al.*, "Pre-existing and machine learning-based models for cardiovascular risk prediction," *Scientific reports*, vol. 11, no. 1, pp. 1–10, 2021.

[12] K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel, "Cardiovascular disease risk profiles," *American Heart Journal*, vol. 121, no. 1, pp. 293–298, 1991.

[13] R. M. Conroy, K. Pyörälä, A. e. Fitzgerald, S. Sans, A. Menotti, G. De Backer, D. De Bacquer, P. Ducimetiere, P. Jousilahti, U. Keil, *et al.*, "Estimation of ten-year risk of fatal cardiovascular disease in europe: the score project," *European Heart Journal*, vol. 24, no. 11, pp. 987–1003, 2003.

[14] P. M. Ridker, J. E. Buring, N. Rifai, and N. R. Cook, "Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the reynolds risk score," *Jama*, vol. 297, no. 6, pp. 611–619, 2007.

[15] G. Assmann, P. Cullen, and H. Schulte, "Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular munster (procam) study," *Circulation*, vol. 105, no. 3, pp. 310–315, 2002.

[16] J. W. Stephens, G. Ambler, P. Vallance, D. J. Betteridge, S. E. Humphries, and S. J. Hurel, "Cardiovascular risk and diabetes. are the methods of risk prediction satisfactory?," *European Journal of Preventive Cardiology*, vol. 11, no. 6, pp. 521–528, 2004.

[17] D. M. Lloyd-Jones, L. T. Braun, C. E. Ndumele, S. C. Smith Jr, L. S. Sperling, S. S. Virani, and R. S. Blumenthal, "Use of risk assessment tools to guide decision-making in the primary prevention of atherosclerotic

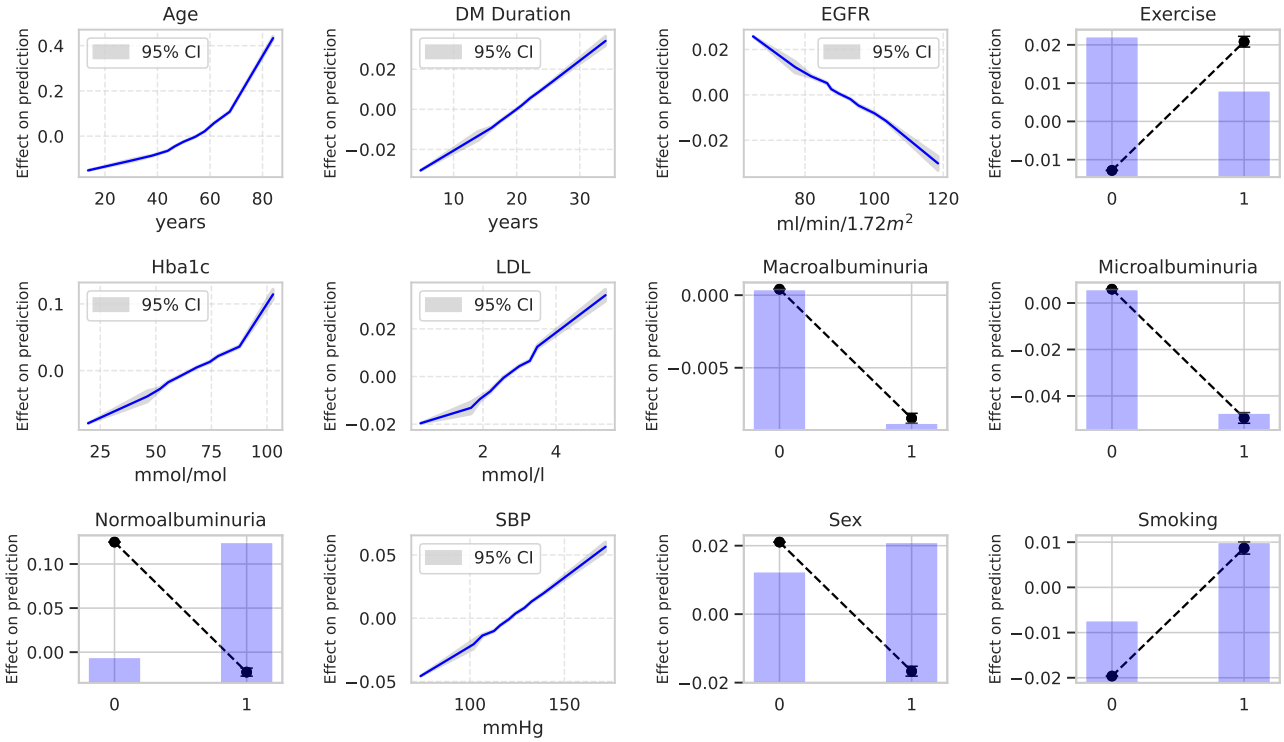

FIGURE 6: ALE plots obtained for the best predictive model (MLP model using *FES4* and data augmentation with *over-per*).

cardiovascular disease: a special report from the american heart association and american college of cardiology," *Circulation*, vol. 139, no. 25, pp. e1162–e1177, 2019.

[18] K. N. Bachmann and T. J. Wang, "Biomarkers of cardiovascular disease: contributions to risk prediction in individuals with diabetes," *Diabetologia*, vol. 61, no. 5, pp. 987–995, 2018.

[19] S. Van Dieren, J. Beulens, A. Kengne, L. Peelen, G. Rutten, M. Woodward, Y. Van der Schouw, and K. Moons, "Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review," *Heart*, vol. 98, no. 5, pp. 360–369, 2012.

[20] M. Petretta, W. Acampa, G. Fiumara, and A. Cuocolo, "Cardiovascular risk stratification in diabetic patients," *Clinical and Translational Imaging*, vol. 1, no. 5, pp. 325–339, 2013.

[21] G. Goliasch, G. Silbernagel, M. E. Kleber, T. B. Grammer, S. Pilz, A. Tomaschitz, P. E. Bartko, G. Maurer, W. Koenig, A. Niessner, *et al.*, "Refining long-term prediction of cardiovascular risk in diabetes–the vildia score," *Scientific reports*, vol. 7, no. 1, pp. 1–9, 2017.

[22] D. Vistisen, G. S. Andersen, C. S. Hansen, A. Hulman, J. E. Henriksen, H. Bech-Nielsen, and M. E. Jørgensen, "Prediction of first cardiovascular disease event in type 1 diabetes mellitus: the steno type 1 risk engine," *Circulation*, vol. 133, no. 11, pp. 1058–1066, 2016.

[23] B. Zethelius, B. Eliasson, K. Eeg-Olofsson, A.-M. Svensson, S. Gudbjörnsdottir, J. Cederholm, *et al.*, "A new model for 5-year risk of cardiovascular disease in type 2 diabetes, from the swedish national diabetes register (ndr)," *Diabetes Research and Clinical Practice*, vol. 93, no. 2, pp. 276–284, 2011.

[24] S. J. McGurnaghan, P. M. McKeigue, S. H. Read, S. Franzen, A.-M. Svensson, M. Colombo, S. Livingstone, B. Farran, T. M. Caparrotta, L. A. Blackbourn, *et al.*, "Development and validation of a cardiovascular risk prediction model in type 1 diabetes," *Diabetologia*, vol. 64, no. 9, pp. 2001–2011, 2021.

[25] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," *IEEE Reviews in Biomedical Engineering*, vol. 14, pp. 156–180, 2020.

[26] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Scientific Reports*, vol. 10, no. 1, p. 11981, 2020.

[27] D. Chushig-Muzo, C. Soguero-Ruiz, P. d. Miguel Bohoyo, I. Mora-Jiménez, *et al.*, "Learning and visualizing chronic latent representations using electronic health records," *BioData Mining*, vol. 15, no. 1, pp. 1–27, 2022.

[28] M. T. Jurado-Camino, D. Chushig-Muzo, C. Soguero-Ruiz, P. de Miguel-Bohoyo, and I. Mora-Jiménez, "On the use of generative adversarial networks to predict health status among chronic patients.," in *HEALTHINF*, pp. 167–178, 2023.

[29] B. Ambale-Venkatesh, X. Yang, C. O. Wu, K. Liu, W. G. Hundley, R. McClelland, A. S. Gomes, A. R. Folsom, S. Shea, E. Guallar, *et al.*, "Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis," *Circulation Research*, vol. 121, no. 9, pp. 1092–1101, 2017.

[30] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in *Proceedings of the 2018 ACM Intl Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 559–560, 2018.

[31] W. Saeed and C. Omlin, "Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities," *Knowledge-Based Systems*, vol. 263, p. 110273, 2023.

[32] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777, 2017.

[33] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 82, no. 4, pp. 1059–1086, 2020.

[34] Y.-J. Cao, L.-L. Jia, Y.-X. Chen, N. Lin, C. Yang, B. Zhang, Z. Liu, X.-X. Li, and H.-H. Dai, "Recent advances of generative adversarial networks in computer vision," *IEEE Access*, vol. 7, pp. 14985–15006, 2018.

[35] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[36] C. García-Vicente, D. Chushig-Muzo, I. Mora-Jiménez, H. Fabelo, I. T. Gram, M.-L. Løchen, C. Granja, and C. Soguero-Ruiz, "Evaluation of syn-

thetic categorical data generation techniques for predicting cardiovascular diseases and post-hoc interpretability of the risk factors," *Applied Sciences*, vol. 13, no. 7, p. 4119, 2023.

[37] O. Habibi, M. Chemmakha, and M. Lazaar, "Imbalanced tabular data modelization using ctgan and machine learning to improve iot botnet attacks detection," *Engineering Applications of Artificial Intelligence*, vol. 118, p. 105669, 2023.

[38] P. Cerda, G. Varoquaux, and B. Kégl, "Similarity encoding for learning with dirty categorical variables," *Machine Learning*, vol. 107, no. 8, pp. 1477–1494, 2018.

[39] M. Svensén and C. M. Bishop, "Pattern recognition and machine learning," 2007.

[40] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[41] Y. Song, J. Liang, J. Lu, and X. Zhao, "An efficient instance selection algorithm for k nearest neighbor regression," *Neurocomputing*, vol. 251, pp. 26–34, 2017.

[42] Y.-Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015.

[43] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 5, p. e1379, 2020.

[44] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?," in *International workshop on machine learning and data mining in pattern recognition*, pp. 154–168, Springer, 2012.

[45] G. Biau, "Analysis of a random forests model," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1063–1095, 2012.

[46] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[47] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[48] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric environment*, vol. 32, no. 14-15, pp. 2627–2636, 1998.

[49] T. Zhu, C. Luo, Z. Zhang, J. Li, S. Ren, and Y. Zeng, "Minority oversampling for imbalanced time series classification," *Knowledge-Based Systems*, vol. 247, p. 108764, 2022.

[50] T. Li, Y. Wang, L. Liu, L. Chen, and C. P. Chen, "Subspace-based minority oversampling for imbalance classification," *Information Sciences*, vol. 621, pp. 371–388, 2023.

[51] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.

[52] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[53] V. Sampath, I. Maurtua, J. J. Aguilar Martin, and A. Gutierrez, "A survey on generative adversarial networks for imbalance problems in computer vision tasks," *Journal of Big Data*, vol. 8, pp. 1–59, 2021.

[54] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, "Gans for medical image analysis," *Artificial Intelligence in Medicine*, vol. 109, p. 101938, 2020.

[55] M. Duerden, N. O'Flynn, and N. Qureshi, "Cardiovascular risk assessment and lipid modification: Nice guideline," *British Journal of General Practice*, vol. 65, no. 636, pp. 378–380, 2015.

[56] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and information systems*, vol. 34, no. 3, pp. 483–519, 2013.

[57] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data classification: Algorithms and applications*, p. 37, 2014.

[58] C.-W. Chen, Y.-H. Tsai, F.-R. Chang, and W.-C. Lin, "Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results," *Expert Systems*, vol. 37, no. 5, p. e12553, 2020.

[59] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 1200–1205, Ieee, 2015.

[60] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[61] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *Journal of Biomedical Informatics*, vol. 85, pp. 189–203, 2018.

[62] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[63] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *Journal of Machine Learning Research*, vol. 20, no. 177, pp. 1–81, 2019.

[64] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4, pp. 1942–1948, IEEE, 1995.

[65] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[66] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[67] A. Hassan, J. H. Paik, S. Khare, and S. A. Hassan, "Ppfs: Predictive permutation feature selection," *arXiv preprint arXiv:2110.10713*, 2021.

[68] G. Castillo-García, L. Morán-Fernández, and V. Bolón-Canedo, "Feature selection for domain adaptation using complexity measures and swarm intelligence," *Neurocomputing*, p. 126422, 2023.

[69] D. Wang, D. Tan, and L. Liu, "Particle swarm optimization algorithm: an overview," *Soft computing*, vol. 22, pp. 387–408, 2018.

[70] X. Wang, J. Yang, X. Teng, W. Xia, and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," *Pattern Recognition Letters*, vol. 28, no. 4, pp. 459–471, 2007.

[71] R. Sharkawy, K. Ibrahim, M. Salama, and R. Bartnikas, "Particle swarm optimization feature selection for the classification of conducting particles in transformer oil," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 18, no. 6, pp. 1897–1907, 2011.

[72] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: definitions, methods, and applications," *arXiv preprint arXiv:1901.04592*, 2019.

[73] C. Moreira, Y.-L. Chou, M. Velmurugan, C. Ouyang, R. Sindhgatta, and P. Bruza, "Linda-bn: An interpretable probabilistic approach for demystifying black-box predictive models," *Decision Support Systems*, vol. 150, p. 113561, 2021.

[74] C. Molnar, "Interpretable machine learning,"

[75] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[76] H. Calero-Diaz, D. Chushig-Muzo, H. Fabelo, I. Mora-Jiménez, C. Granja, and C. Soguero-Ruiz, "Data-driven cardiovascular risk prediction and prognosis factor identification in diabetic patients," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 01–04, IEEE, 2022.

[77] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Switzerland: Springer, 2006.

[78] D. Control, C. Trial, E. of Diabetes Interventions, C. D. R. Group, *et al.*, "Risk factors for cardiovascular disease in type 1 diabetes," *Diabetes*, vol. 65, no. 5, p. 1370, 2016.

[79] J. Schofield, J. Ho, and H. Soran, "Cardiovascular risk in type 1 diabetes mellitus," *Diabetes Therapy*, vol. 10, pp. 773–789, 2019.

[80] Y.-Y. Chen, Y.-J. Lin, E. Chong, P.-C. Chen, T.-F. Chao, S.-A. Chen, and K.-L. Chien, "The impact of diabetes mellitus and corresponding hba1c levels on the future risks of cardiovascular disease and mortality: a representative cohort study in taiwan," *PloS one*, vol. 10, no. 4, p. e0123116, 2015.

[81] S. Arques, "Human serum albumin in cardiovascular diseases," *European Journal of Internal Medicine*, vol. 52, pp. 8–12, 2018.

[82] M. V. Fangel, P. B. Nielsen, J. K. Kristensen, T. B. Larsen, T. F. Overvad, G. Y. Lip, and M. B. Jensen, "Albuminuria and risk of cardiovascular events and mortality in a general population of patients with type 2 diabetes without cardiovascular disease: a danish cohort study," *The American Journal of Medicine*, vol. 133, no. 6, pp. e269–e279, 2020.

[83] P. T. Katzmarzyk, C. Friedenreich, E. J. Shiroma, and I.-M. Lee, "Physical inactivity and non-communicable disease burden in low-income, middle-income and high-income countries," *British Journal of Sports Medicine*, vol. 56, no. 2, pp. 101–106, 2022.

[84] R. Ng, R. Sutradhar, Z. Yao, W. P. Wodchis, and L. C. Rosella, "Smoking, drinking, diet and physical activity—modifiable lifestyle risk factors and their associations with age to first chronic disease," *International journal of epidemiology*, vol. 49, no. 1, pp. 113–130, 2020.

DAVID CHUSHIG-MUZO received the Ph.D. degree in machine learning in healthcare from the Rey Juan Carlos University in 2022. He works as a post-doctoral researcher in the WARIFA project by building interpretable risk prediction models. He has co-authored several research papers in international journals and conferences. He has participated as a researcher in public funding projects, mainly related to machine learning in healthcare. His main research interests include statistical learning theory, machine learning, data mining, and computer vision.

HUGO CALERO-DÍAZ received the B.Sc. in Biomedical Engineering by the Rey Juan Carlos University and the M.Sc. degree in Data Science and Artificial Intelligence by Newcastle University. Hugo is currently an Artificial Intelligence Engineer at IBM. His research interests include machine learning, explainable artificial intelligence, and multimodal fusion methods.

FRANCISCO J. LARA-ABELENDA received the B.Sc. in Biomedical Engineering from Rey Juan Carlos University and the M.Sc. in Machine Learning for Health from Carlos III University. He is currently working towards a Ph.D. in Explainable and Multimodal Artificial Intelligence with applications in healthcare at Rey Juan Carlos University. His research interests encompass signal processing, natural language processing, explainable artificial intelligence, and multimodal fusion methods.

VANESA GÓMEZ-MARTÍNEZ received a B.Sc. in Biomedical Engineering and M.Sc. in Computer Vision at Rey Juan Carlos University. She is currently pursuing a Ph.D. in Explainable and Multimodal Artificial Intelligence in healthcare. Her Ph.D. is being conducted at the Department of Signal Theory and Communications, Telematics, and Computing. Her main research interests include machine learning, deep learning, computer vision, data mining, and natural language processing.

CONCEIÇAO GRANJA received her Ph.D. in 2013 at the Faculty of Engineering, University of Porto. Her Ph.D. thesis focused on process optimization in healthcare providers, in the particular case of diagnostic imaging, achieved through the implementation of modeling and simulation techniques. She is a postdoctoral fellow at the Norwegian Centre for Integrated Care and Telemedicine (NST) at the University Hospital of North Norway, in the department for eHealth and ICT.

CRISTINA SOGUERO-RUIZ got the Ph.D. degree in machine learning with applications in healthcare, in 2015, with the Joint Doctoral Program in Multimedia and Communications in conjunction with University Rey Juan Carlos and University Carlos III. She won the Orange Foundation Best Ph.D. Thesis Award by the Spanish Official College of Telecommunication Engineering. She has published several papers in JCR journals and international conferences. She has participated in several research projects (with public and private funding) related to healthcare data-driven machine learning systems. Her current research interests include machine learning, data science, and statistical learning theory.