Clinical Synthetic Data Generation to Predict and Identify Risk Factors for Cardiovascular Diseases

García-Vicente, C., Chushig-Muzo, D., Mora-Jimenez, I., Fabelo, H., Gram, I. T., Løchen, M.-L., Granja, C. & Soguero-Ruiz, C.

Published in: Lecture Notes in Computer Science (LNCS)
DOI: 10.1007/978-3-031-23905-2_6

Available online: 21 Jan 2023

Citation:
Clinical Synthetic Data Generation to Predict and Identify Risk Factors for Cardiovascular Diseases. Lecture Notes in Computer Science (LNCS). 13814. doi: 10.1007/978-3-031-23905-2_6

Clinical synthetic data generation to predict and identify risk factors for cardiovascular diseases

Clara García-Vicente[1 [0000−0001−9805−0011]], David Chushig-Muzo[1 [0000−0001−5585−2305]], Inmaculada Mora-Jiménez[1 [0000−0003−0735−367X]], Himar Fabelo[2,3 [0000−0002−9794−490X]], Inger Torhild Gram[4,5 [0000−0002−0031−4152]], Maja-Lisa Løchen[5 [0000−0002−8532−6573]], Conceição Granja[4,6 [0000−0002−3028−8899]], and Cristina Soguero-Ruiz[1 [0000−0001−5817−989X]]

[1] Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, Madrid 28943, Spain
{clara.garcia.vicente,david.chushig,
inmaculada.mora,cristina.soguero}@urjc.es
[2] Research Institute for Applied Microelectronics, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain
[3] Fundación Canaria Instituto de Investigación Sanitaria de Canarias (FIISC), Las Palmas de Gran Canaria, Spain
hfabelo@iuma.ulpgc.es
[4] Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø 9019, Norway
{inger.gram,conceicao.granja}@ehealthresearch.no
[5] Department of Community Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø 9019, Norway
maja-lisa.lochen@uit.no
[6] Faculty of Nursing and Health Sciences, Nord University

Abstract. Noncommunicable diseases are among the most significant health threats in our society, being cardiovascular diseases (CVD) the most prevalent. Because of the severity and prevalence of these illnesses, early detection and prevention are critical for reducing the worldwide health and economic burden. Though machine learning (ML) methods usually outperform conventional approaches in many domains, class imbalance can hinder the learning process. Oversampling techniques on the minority classes can help to overcome this issue. In particular, in this paper we apply oversampling methods to categorical data, aiming to improve the identification of risk factors associated with CVD. To conduct this study, questionnaire data (categorical) obtained by the Norwegian Centre for E-health Research associated with healthy and CVD patients are considered. The goal of this work is two-fold. Firstly, evaluating the influence of combining oversampling techniques and linear/nonlinear supervised ML methods in binary tasks. Secondly, identifying the most relevant features for predicting healthy and CVD cases. Experimental results show that oversampling and FS techniques help to improve CVD prediction. Specifically, the use of Generative Adversarial Networks and linear models usually achieve the best performance (area under the curve of

67%), outperforming other oversampling techniques. Synthetic data generation has proved to be beneficial for both identifying risk factors and creating models with reasonable generalization capability in the CVD prediction.

Keywords: Non-communicable diseases · Cardiovascular diseases · Generative adversarial networks · SMOTE · Synthetic data generation · Feature selection · Risk factor identification

1 Introduction

Non-communicable diseases (NCDs) are among the significant health threats in our society due to their impact and severity, affecting all age ranges and countries [46]. According to the World Health Organization, NCDs are responsible for the most significant number of deaths worldwide, with approximately 41 million people dying yearly [35]. The leading NCDs are cardiovascular diseases (CVDs), cancer, chronic respiratory diseases, and diabetes mellitus, with 17.9, 9, 3.9 and 1.6 million deaths per year, respectively [18]. This accounts for 71% of all deaths globally, and although NCDs tend to be associated with old people, 15 million of these deaths occur in people within the range of 30 and 69 years. Therefore, young people, adults and the elderly may be vulnerable to risk factors related to the development of NCDs.

NCDs are caused by different factors, including genetic, physiological, behavioral and environmental factors [2]. Among them, the lack of physical activity, unhealthy diets, alcohol/tobacco use, and obesity play an important role in the onset of these diseases [36]. The main priority for NCDs prevention is linked to lifestyle change as well as early intervention. The main problem with NCDs is that they are often diagnosed at an advanced stage, making it challenging to deal with them. In this scenario, the availability of models to support decision-making would help in the early diagnosis of NCDs, as well as to identify high-risk patients and reduce mortality rates [14, 30].

Over the last years, different machine learning (ML) methods have been developed to support health practitioners in decision-making, providing remarkable advances in different knowledge domains. ML techniques use data to build models capable of making predictions and identifying patterns [9]. In the clinical setting, a variety of works have applied ML in different applications, including support to disease diagnosis, extraction of hidden patterns and analysis of health statuses, among others [9]. Data availability is crucial to the success of ML classifiers which, in general, are built under the assumption of a similar number of observations per class [20]. Class imbalance usually causes that data-driven models capture a better representation of the observations in the majority class, leading to poor model performance for the minority class [45]. In real-world scenarios, when dealing with medical databases, as in our case study, the class imbalance is one of the main challenges for designing data-driven models, usually due to the limitation in the number of samples.

To cope with the class imbalance problem, a variety of methods have been proposed in the literature [21]. In this paper, we focus on oversampling techniques, with the synthetic minority oversampling technique (SMOTE) [7] being one of the most extensively used. Other approaches are increasingly coming into use, such as the Generative Adversarial Networks (GANs), which have changed findings in a variety of fields by providing high performance when generating synthetic data [1]. Although GANs have been tested in a variety of domains, they have not been thoroughly investigated when it comes to electronic health records (EHRs) [1]. In the literature, different GANs have been presented in the clinical domain to generate synthetic patient samples from real-world data, addressing the challenge of restricted data sources in healthcare applications [8]. The GAN-based model called medGAN [8] was recently presented to generate synthetic categorical data related to EHRs using the clinical code-based MIMIC dataset [28]. Most previous studies refer to the generation of synthetic data from numerical databases rather than categorical databases, even when most healthcare applications handle categorical data.

In order to achieve class balance, we created synthetic samples by focusing on the following types of data augmentation schemes: SMOTEN, a variant of SMOTE for categorical data, Tabular Variational Autoencoder (TVAE) [47], Gaussian Copula (GC) [33] and medGAN [8]. Once the classes were balanced, we applied several ML approaches to extract the most predominant risk factors and perform classification tasks. Finally, the performance of the resulting model was evaluated using a subset of real observations, independent from those considered during the model design.

The rest of the paper is organized as follows. Section 2 describes the dataset and pre-processing stage. Section 3 introduces the theoretical foundations of the oversampling and classification methods used. Next, Section 4 shows experimental results related to CVD classification performance and model interpretability outcomes when considering linear and nonlinear methods. Finally, Section 5 presents the discussion and main conclusions.

2 Dataset description and pre-processing

The dataset considered in this work is part of the contribution to a three-year project carried out by the Norwegian Centre for E-health Research, UiT The Arctic University of Norway and Healthcom, who designed the "Health and Disease" of NCDs [19]. A smartphone-based method was used to collect the data by a series of survey questions to a population group in Norway. This study was developed for monitoring the modifiable risk factors of four NCDs: diabetes, cancer, CVD and chronic respiratory diseases. The dataset was composed of 2303 individuals, but in the preprocessing stage we eliminated 10 individuals who had not completed the questionnaire, resulting in a dataset with a total of 2293 individuals.

The survey was designed with a total of 26 questions (variables): 7 questions related to socioeconomic factors, 7 questions related to alcohol and drug use,

4 questions related to physical activity, 7 questions related to the type of diet, and 1 question indicating current/previous NCDs. In particular, the following NCDs were considered: high cholesterol, atrial fibrillation, myocardial infarction, heart failure, stroke, chronic respiratory disease, cancer and diabetes. Finally, by studying the disease groups separately, we observed that all those who suffered or had suffered heart failure, cardiovascular accident, atrial fibrillation and/or myocardial infarction had also responded that they suffered cardiovascular disease. For this reason, we decided to group these four variables into a new variable indicating only whether the patient had CVD. Individuals who did not respond to the question related to NCDs were considered as healthy individuals. Thus, there were 465 people with CVD, 72 people with cancer, 46 people with both NCDs and 1578 people who do not suffer from any disease (associated with the healthy population group). Considering the low number of patients with cancer and both diseases, we decided to focus only on the study of CVD patients, which according to the literature it is also the predominant disease within NCDs [18]. Furthermore, we created a new category called 'NA' to indicate that the answer to a question is not available. Finally, our dataset consists of 2043 individuals, with 1578 healthy respondents and 465 individuals with CVD. Regarding the variables, we have organized them into the following six groups:

- **Socioeconomic background factors**: year of birth (age), sex, body mass index (BMI) and level of education (education).
- **Substance use**: cigarette consumption (smoking), snuff and e-cigarette use, and alcohol consumption. Concerning alcohol, there are specific variables extracted from the *Alcohol Use Disorders Identification Test* (AUDIT) [3], with information about frequency, the number of units usually consumed and the frequency of occasions of consumption of more than 6 units of alcohol.
- **Physical activity**: it was extracted from the *International Physical Activity Questionnaire* (IPAQ) [10]. There are data about: number of days of strenuous physical activity in the last 7 days, number of days of moderate physical activity in the last 7 days, number of days of walking for more than 10 minutes in the previous 7 days, and hours spent sitting (excluding sleeping hours) on a regular weekday in the previous 7 days.
- **Dietary intake**: servings of fruits and berries per day, lettuce and vegetable intake per day, sugary drinks and number of glasses per day, fish and number of times consumed per week, red meat and number of times consumed per week, processed meat and number of times consumed per week, and frequency with which extra salt is added to food before eating.
- **Income**: number of persons in the household over and under 18, and gross household income the previous year.
- **Clinical**: presence of high cholesterol.

Since we are dealing with categorical variables, we considered the one-hot-encoding strategy [6] for the majority of ML approaches. This type of coding creates one additional feature for each category in the variable (excepting for binary variables) and sets to '1' just the feature linked to the active category for

each observation. After this encoding, the dataset dimension [23] increased to a total of 153 features.

3 Methodology for predicting cardiovascular diseases

The workflow is sketched in Figure 1. First, an exploratory analysis and preprocessing stage was performed to clean the data and ensure that the database is curated for use by ML methods. Next, the dataset was split into two independent sets (training and test sets), and the bootstrap resampling-based test was used to perform FS. Subsequently, we used different oversampling methods to balance the classes and design CVD classifiers with balanced datasets. This section describes the FS approach, the oversampling techniques and the ML methods.

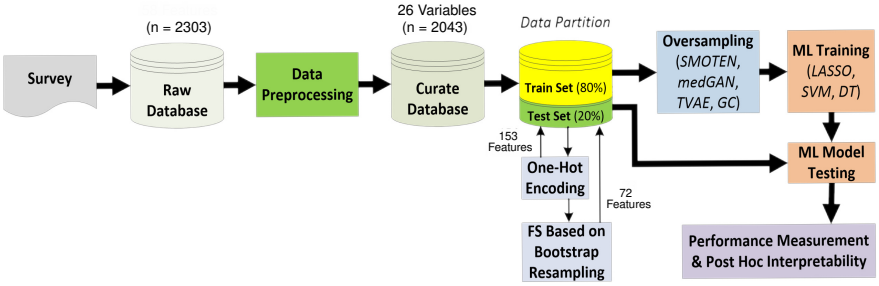

Fig. 1: Workflow using CVD data, oversampling methods and ML classifiers.

3.1 Test based on bootstrap resampling for feature selection

High-dimensional data could lead to irrelevant and redundant features, which can cause overfitting and worsen the model performance. To cope with these issues, FS techniques aim to find a subset of the input variables that best describes the underlying data structure [4]. According to the literature, FS techniques can be categorized in filter, wrapper and embedded methods [4]. This paper focuses on filter methods due to their simplicity and ease of implementation. In particular, we use the non-parametric test named *bootstrap resampling* to estimate the distribution of one statistics (*e.g.,* the mean) taking samples without replacement from a population [16, 27, 31].

In our work, we compute the difference Δ between the proportion of a specific binary feature in the CVD population and in the healthy population. To estimate the distribution of Δ, a bootstrap resampling is performed. Thus, each class is resampled 3,000 times, with the size of each resample being the same for both classes. Then, the difference in proportions between the populations for each resampling and the 95% confidence interval (CI_Δ) for each feature is computed.

The null hypothesis H_0 is true if $0 \in CI_\Delta$, while the alternative hypothesis H_1 is considered true if $0 \notin CI_\Delta$ (*e.g.,* no overlapping with 0). When H_1 is true, it indicates a significant difference between the proportion of the same feature in both populations (healthy and CVD individuals). Thus, features fulfilling H_1 are selected as relevant for subsequent analysis.

3.2 Oversampling techniques for categorical variables

We applied the following oversampling strategies to generate synthetic samples of the minority class for categorical data:

SMOTE is one of the most popular methods for oversampling. It performs a linear interpolation of the variables associated with random samples of the minority class through the k-Nearest Neighbours scheme [17]. Since SMOTE only deals with continuous variables, variants such as SMOTEN [7, 32] have been proposed for dealing with categorical variables. In SMOTEN, the nearest neighbors are found through a modified version of the value difference measure [42] proposed by Cost and Salzberg [11].

VAE is a deep generative model based on artificial neural networks [22]. Like most autoencoders, VAE is comprised of one encoder compressing the input data into a lower dimensional latent space and one decoder reconstructing the input by only using the latent space. VAE is similar in architecture to AE, but includes a generative part to learn underlying probability distribution from data and generate synthetic samples.

VAEs are applied in fields such as image/text classification, anomaly detection or image generation [38, 48]. To generate tabular data, a variant of VAE called TVAE [47] for handling numerical and categorical variables was considered.

GC is a probabilistic method based on *copula functions*. A *copula* is a joint probability distribution built from marginal univariate probability distributions [33]. In other words, a copula function allows us to describe the joint probability distribution of several random variables through the dependencies among their marginal distributions.

GAN-based oversampling methods make use of neural networks. They are composed of two parts: a generative model, which we train to generate new samples, and a discriminative model which attempts to classify samples as real or synthetic (generated). Both models are trained together until the generator model produces believable examples [12]. Despite their popularity and performance, most of GAN-based methods have been used for unstructured data and in applications related to image generation. However, just a few studies have analysed GAN-based approaches for oversampling structured and tabular data. One of the most remarkable works in this line is medGAN [8], orientated to the generation of synthetic patients from EHR data.

3.3 Classification methods

Different ML classifiers have been employed in the health-care literature [34]. To support interpretability by medical professionals as well as the identification of the most predominant risk factors for CVD, we considered the following approaches [5, 26]: Least Absolute Shrinkage and Selection Operator (LASSO), Linear Support Vector Machine (LSVM), and Decision Trees (DTs).

LASSO is a technique for obtaining the best linear model for a data set by minimizing the sum of squared residuals among real and predicted values. L1 regularization is a strategy used in LASSO to penalize the previous cost function by including a term computed as the sum of the absolute values of the model coefficients [39]. As a consequence, the less relevant variables are set to zero, implicitly discarding the less relevant features in the model. The hyperparameter λ regulates the degree of penalty: the higher the λ value, the greater the penalty and the greater the number of coefficients that will be set to 0.

LSVM is a kernel-based method [29] which transforms the input space into a higher dimensional space where a hyperplane with good generalization capability is created. The main idea of SVM is to minimize the classification error by finding the maximum margin hyperplane splitting the observations into two classes. The hyperparameter C in SVM quantifies the degree of importance given to misclassifications. Different kernels have been proposed in the literature, including linear, radial basis function, sigmoid or polynomial. We used a linear kernel since its weights allows us to characterize the feature importance of the variables in an easy way [43].

DT allows to create nonlinear models in a non-parametric way. DT is composed by a set of nodes recursively divided into branches according to some criteria such as entropy [41]. When the DT just consider one feature per node, every time a node is created the associated region in the feature space is split into two parts by a linear boundary. In classification tasks, a label is assigned to each part according to the majority class among the training samples in that region. The root node (placed at the top of the tree) indicates the most relevant feature for classification, from which the first partition is performed. Below the root node are the intermediate nodes, which continue subdividing the input space. The terminal nodes indicate the final classification [41].

4 Experimental results

This section analyzes the influence of several oversampling techniques on the classifier performance when tackling a binary classification task (healthy versus CVD individuals), and conduct a post-hoc interpretability stage.

4.1 Experimental setup

Data-driven classifiers are designed and validated using two independent subsets, the training set and the test set, by allocating 80% and 20% of the samples, respectively. In order to better characterize the model performance, 5 independent

training and test partitions have been considered in this work. For hyperparameter selection in the classifiers, the 3-fold cross validation (CV) [40] approach was considered just with the training set: λ for LASSO, C for LSVM, and both the minimum number of samples for splitting a node and the maximum tree depth for DT.

Since we are dealing with an imbalanced dataset, the area under the receiver operating characteristic (AUC) was chosen as a figure of merit for assessing the classification performance [20]. It is commonly used in the medical domain since it provides a trade-off between sensitivity and specificity. Note that, as we work with 5 partitions, the classifier performance is shown in terms of the average AUC and the standard deviation on the AUC values over the five test partitions (which remain imbalanced for evaluation purposes).

4.2 CVD classification performance

We present the AUC statistics when classifying healthy and CVD patients in two scenarios: (1) considering all features, and (2) only using the most relevant features according to the bootstrap resampling test. Firstly, considering only the training subset, we generate synthetic data of the minority class (CVD), aiming to balance the dataset and improve the CVD prediction. Note that for evaluating classification performance, the test subset does not present synthetic samples, and only real samples are used for getting the figures of merit. Different imbalance ratios (IRs) are considered to balance the number of samples of the dataset classes and evaluate the influence of synthetic samples on the classifier performance. IR is defined as the ratio between the number of samples of the minority class (N_{min}) and the number of samples of the majority classes (N_{maj}). Note that IR is directly related to N_{min}, with greater IR indicating more synthetic samples of the minority class (when keeping a constant value in N_{maj}) and with IR=1 indicating a balanced dataset. We compare the following oversampling techniques: SMOTEN, TVAE, GC and medGAN. We analyze the influence of N_{min} (the minority class size) in terms of AUC when varying the IR and considering different oversampling techniques (see Figure 2 for details). Since we consider five partitions, the mean (indicated with $\circ, \triangle, \times$) and the standard deviation (represented by a shaded color) of the AUC values obtained in the test sets are shown.

We can observe in Figure 2 that varying the IR does not substantially change the AUC values when considering TVAE, GC and SMOTEN. This means that a more significant number of synthetic samples from the minority class does not make the classification models perform better. However, in the case of medGAN, the effect of the size of the synthetic dataset (by varying IR in the interval $[0.5, 1.0]$) improves the AUC values. Secondly, medGAN seems to be the oversampling technique providing the better performance, reaching AUC $= 0.65$ (see Figure 2 (d)). Also, in general, we can observe that linear models (LASSO and LSVM) provide the best AUC values compared with DT. As stated, handling categorical data with one-hot encoding increases the dimensionality of the dataset considerably, with a total of $D = 153$ features. Bootstrap resampling

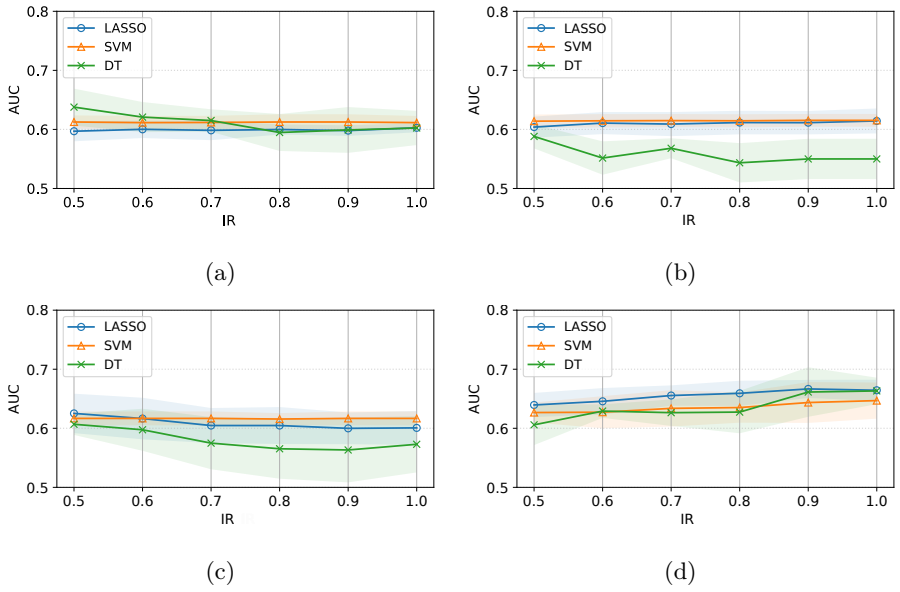

Fig. 2: Classification performance (AUC) in the test set by varying IR when considering all features and several oversampling techniques: (a) SMOTEN; (b) TVAE; (c) GC; and (d) medGAN. Solid lines refer to average AUC values, while shaded areas are associated with the corresponding standard deviation in AUC.

was used to remove those features that were non-relevant and uninformative for predicting the target variable. By applying bootstrap resampling, we take $\hat{D}$ features, being $\hat{D} < D$. With this, it is sought to evaluate if there are improvements in the classification performance and compare them.

Figures 3 and 4 show the CI_{Δ} for each feature when the hypothesis test based on bootstrap resampling (see subsection 3.1) is performed. Each CI_{Δ} shown in blue in Figure 3 refers to one selected feature, with 72 out of the 153 initial features. In contrast, the CI_{Δ} shown in red in Figure 4 indicates the non-selected features. It is important to highlight that features such as high cholesterol, age, BMI, household adults, house income, alcohol drink frequency, and smoking were selected. According to the literature [13, 15, 44], BMI, high cholesterol and age are documented as risk factors in CVD. The AUC values when considering the selected features are shown in Figure 5. As in the previous scenario, the linear models (LASSO and LSVM) provide better AUC values than those obtained with DT.

To compare the differences in the binary classification performance using all features and the selected ones, we show in Table 1 the AUC values when considering the different test subsets (different partitions). From this table, note that the best AUC values (about 0.65 or 0.66) using all the features were obtained with medGAN. Also, focusing on this oversampling technique, the best AUC

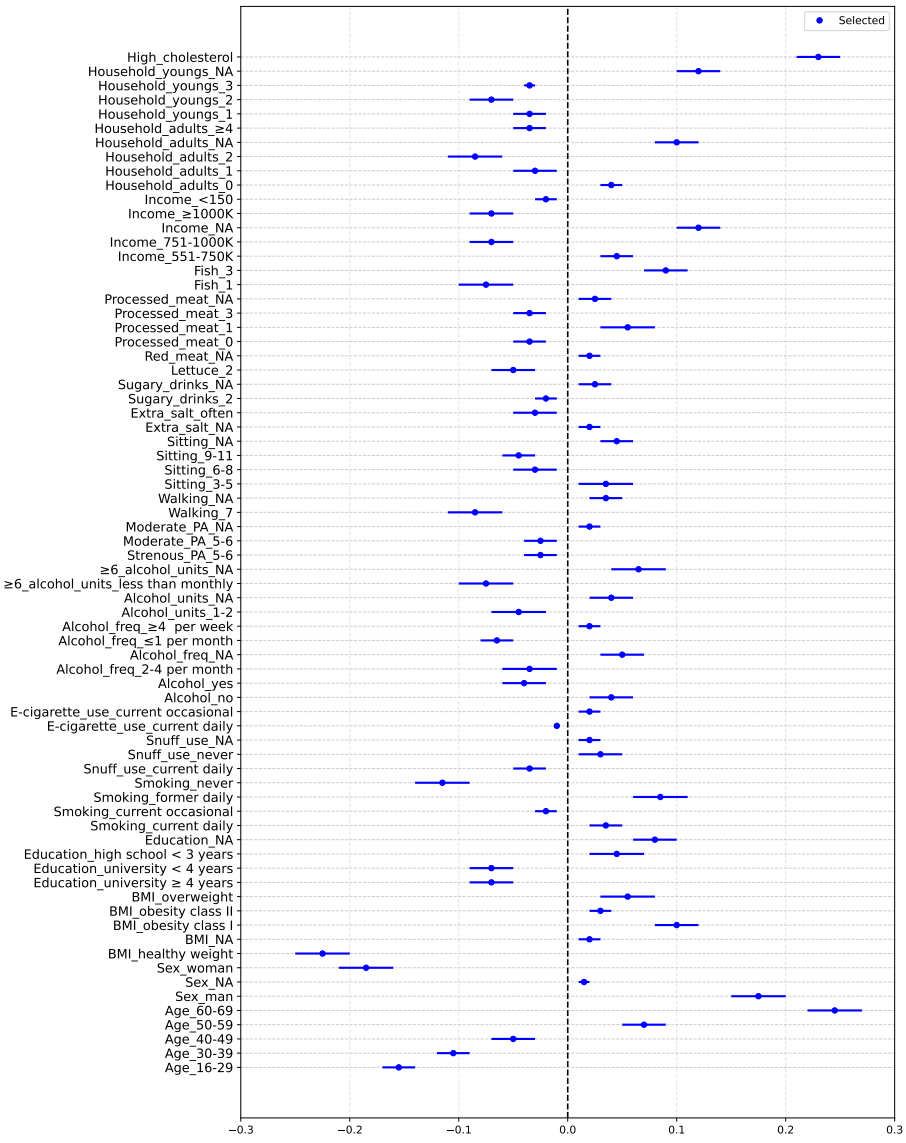

Fig. 3: CI_{Δ} for the selected features when considering the hypothesis test based on bootstrap resampling.

value (0.66) was achieved considering LASSO. DT presents a very similar performance, with a score of 0.66, although with a highest standard deviation (0.02 versus 0.01 of the LASSO model). From Table 1, we can conclude that better performance are obtained after FS and the best AUC values are obtained using the medGAN technique and the LASSO model.

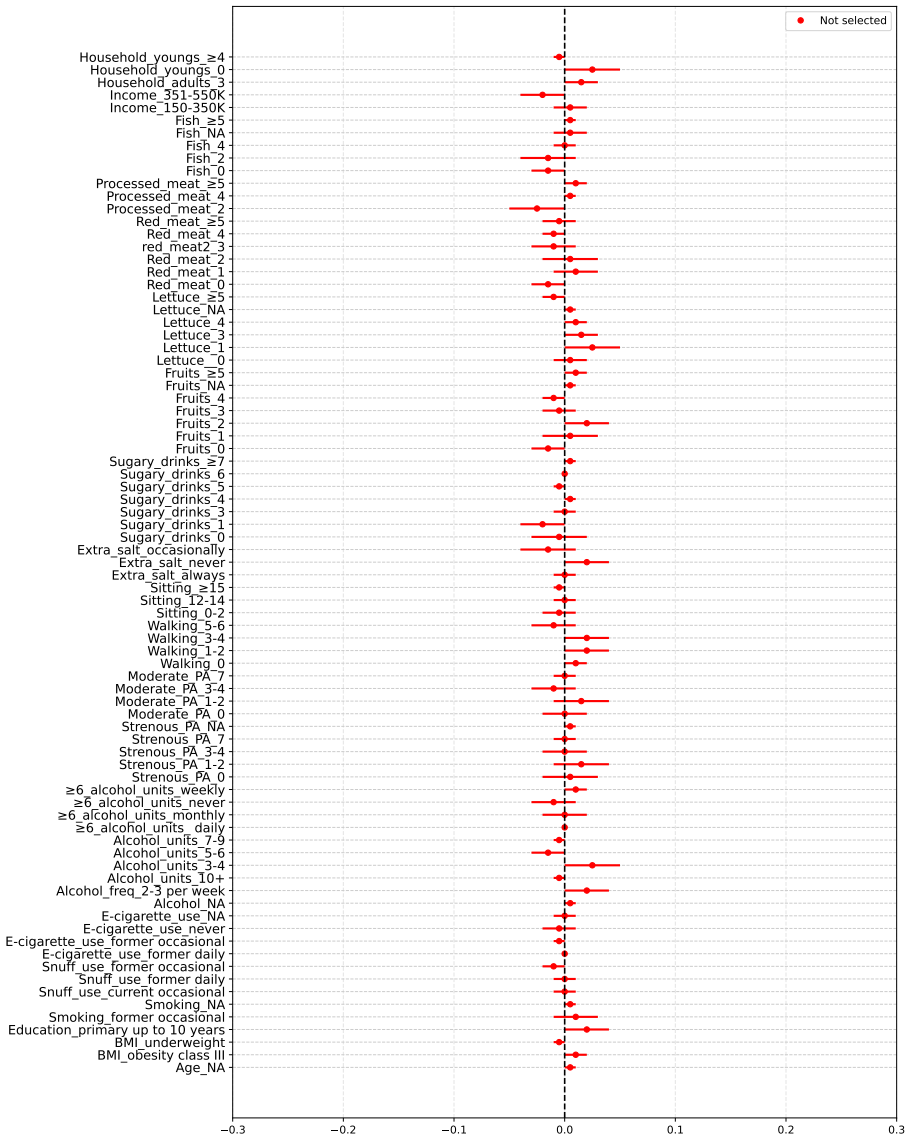

Fig. 4: CI_{Δ} for the non-selected features when considering the hypothesis test based on bootstrap resampling.

4.3 Post hoc interpretability

An important challenge in ML is the model interpretation, which refers to the reasoning behind the model decision in a way that humans can understand. In healthcare, interpretability is key to extracting knowledge and, above all, sup-

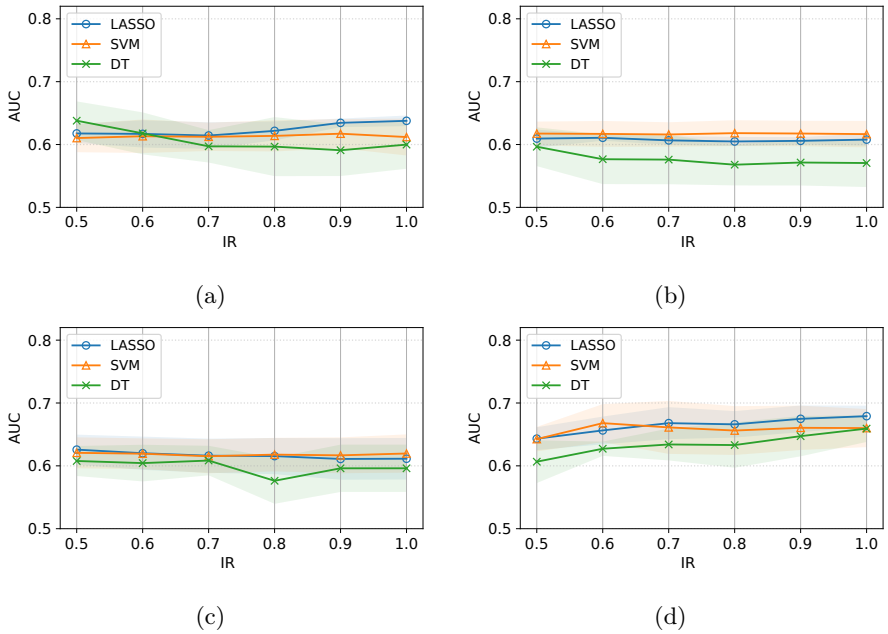

Fig. 5: Classification performance by considering the selected features with bootstrap resampling test and for: (a) SMOTEN; (b) TVAE; (c) GC; and (d) medGAN.

Table 1: AUC values (mean ± standard deviation) for different oversampling techniques and classifiers when considering all features and those selected with bootstrap resampling.

Oversampler	Classifier	All features	Selected features
SMOTEN	LASSO	0.60±0.01	**0.63±0.01**
	SVM	0.61±0.01	0.61±0.02
	DT	0.62±0.02	0.63±0.03
TVAE	LASSO	0.61±0.01	**0.61±0.01**
	SVM	0.61±0.01	0.61±0.02
	DT	0.58±0.02	0.59±0.03
GC	LASSO	0.60±0.00	**0.63±0.01**
	SVM	0.61±0.01	0.62±0.02
	DT	0.60±0.01	0.60±0.02
medGAN	LASSO	0.66±0.01	**0.67±0.01**
	SVM	0.64±0.03	0.66±0.03
	DT	0.66±0.02	0.65±0.02

porting physicians in decision-making. Three techniques to create interpretable models (see Section 3.3) were used in this work. On the one hand, two linear models (LASSO and SVM) were analyzed since the coefficient weighting each fea-

ture provides us with information on their relevance in the class prediction [29]. On the other hand, a nonlinear classifier (DT) has been studied. DTs also allow us to identify the importance of the features, assigning a score to each feature according to their usefulness in predicting the output class [24].

The feature importance values associated with DT are shown in Figure 6 when considering four oversampling techniques. Note that age, BMI, high cholesterol (by denoting the presence/absence) and gender were the more representative features in DT. According to the literature, excessive adiposity is a major cause of hypertension, dyslipidemia and type 2 diabetes, which is one of the primary precursors of CVDs. BMI is a key indicator of overall adiposity [44]. Our data-driven approach could identify this feature as relevant to distinguishing CVD cases. Note that in all panels of Figure 6, the features linked to BMI are those with the highest importance values. We highlighted features related to high cholesterol and smoking among other relevant features. Evidence suggests that high cholesterol levels and smoking are two predominant risk factors for CVD, which are also two of the leading causes of death in industrialized countries [13]. The literature also supports that all of these modifiable risk factors are prevalent in all age groups and both genders, but increase when people get older [15, 25]. Another relevant aspect is that individuals with low socioeconomic status seem to have a higher risk of CVD. According to the literature, people with higher socioeconomic backgrounds and higher educational level have more access to nutritionally balanced food [37], supporting that diet is considered one of the most crucial risk factors.

5 Conclusion and discussion

In this paper, we studied the effectiveness and feasibility of using oversampling techniques on categorical data for CVD prediction. Several state-of-the-art methods were evaluated by varying the size of the minority class subset in a binary classification scenario (healthy versus CVD patients). Experimental results showed medGAN outperformed SMOTEN, TVAE and GC when generating new samples from real data in our dataset. Also in favour of medGAN, note that the AUC obtained when using the other three oversamplers did not improve with IR as N_{maj} is constant.

Further research in this line may explore a quantitative and qualitative framework analysis related to the quality of synthetic data, measuring and comparing, for instance, the joint probability distribution of features associated with real and synthetic data.

As mentioned in the paper, model interpretability is crucial for practitioners in the healthcare domain. To address this issue, we identified the most representative features for both linear (LASSO and LSVM) and nonlinear classifiers (DT). According to the coefficient values and the feature importance indexes, the most relevant risk factors associated with CVD were age, BMI, high cholesterol, gender and smoking. These findings are in line with the state-of-the-art [13, 15, 25, 44, 49].

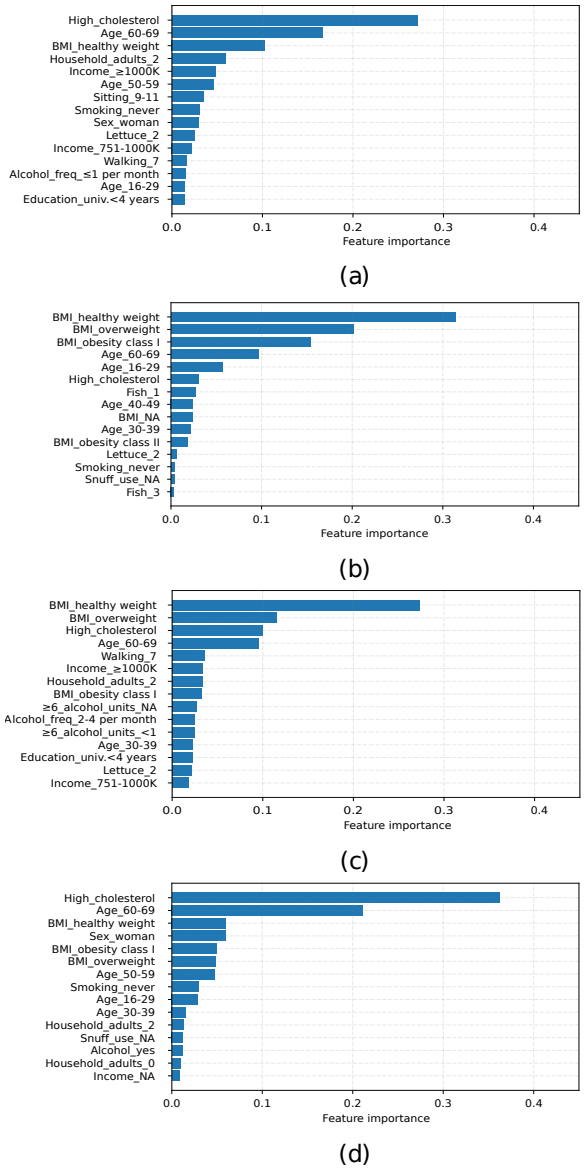

Fig. 6: Feature importance provided by DT when considering: (a) SMOTEN; (b) TVAE; (c) GC; and (d) medGAN. Note that features in each panel are sorted according to their feature importance value.

Our work reveals that the combination of medGAN and LASSO achieve the best classification performance, reaching an average AUC value of about 67.90%. Furthermore, using an FS technique allows us to improve the CVD prediction,

obtaining higher AUC values and identifying the most representative features for CVD. In summary, the results in this study show that the combination of FS and oversampling strategies improve the prediction efficiency of healthy and CVD cases, allowing their extrapolation to more complex scenarios.

Acknowledgements This work has been partly supported by 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 Grant Agreement 101017385; and by the Spanish Government by the Spanish Grants BigTheory (PID2019-106623RB-C41), and AAVis-BMR PID2019-107768RA-I00); Project Ref. 2020-661, financed by Rey Juan Carlos University and Community of Madrid; and by the Research Council of Norway (HELSE-EU-project 269882).

References

1. Aggarwal, A., et al.: Generative adversarial network: An overview of theory and applications. International Journal of Information Management Data Insights **1**(1), 100004 (2021)
2. Budreviciute, A., et al.: Management and prevention strategies for non-communicable diseases (ncds) and their risk factors. Frontiers in Public Health **8**, 788 (2020)
3. Bush, K., et al.: The audit alcohol consumption questions (audit-c): an effective brief screening test for problem drinking. Archives of Internal Medicine **158**(16), 1789–1795 (1998)
4. Cai, J., et al.: Feature selection in machine learning: A new perspective. Neurocomputing **300**, 70–79 (2018)
5. Carvalho, D.V., et al.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019)
6. Cerda, P., et al.: Similarity encoding for learning with dirty categorical variables. Machine Learning **107**(8), 1477–1494 (2018)
7. Chawla, N.V., et al.: Smote: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research **16**, 321–357 (2002)
8. Choi, E., et al.: Generating multi-label discrete patient records using generative adversarial networks. In: Machine Learning for Healthcare Conference. pp. 286–305. PMLR (2017)
9. Chushig-Muzo, D., et al.: Interpreting clinical latent representations using autoencoders and probabilistic models. Artificial Intelligence in Medicine **122**, 102211 (2021)
10. Cleland, C., et al.: Validity of the international physical activity questionnaire (ipaq) for assessing moderate-to-vigorous physical activity and sedentary behaviour of older adults in the united kingdom. BMC medical research methodology **18**(1), 1–12 (2018)
11. Cost, S., Salzberg, S.: A weighted nearest neighbor algorithm for learning with symbolic features. Machine Learning **10**(1), 57–78 (1993)
12. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: An overview. IEEE Signal Processing Magazine **35**(1), 53–65 (2018)

13. Dahlöf, B.: Cardiovascular Disease Risk Factors: Epidemiology and Risk Assessment. The American Journal of Cardiology **105**(1), 3A–9A (2010)
14. Davagdorj, K., et al.: Explainable artificial intelligence based framework for non-communicable diseases prediction. IEEE Access **9**, 123672–123688 (2021)
15. Díez, J.M.B., et al.: Cardiovascular disease epidemiology and risk factors in primary care. Revista Española de Cardiología (English Edition) **58**(4), 367–373 (2005)
16. Efron, B., Tibshirani, R.J.: An introduction to the bootstrap. CRC Press (1994)
17. Fernández, A., et al.: Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary **61**, 863–905 (2018)
18. Forouzanfar, M.H., et al.: Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the global burden of disease study 2015. The lancet **388**(10053), 1659–1724 (2016)
19. Gram, I.T., et al.: A smartphone-based information communication technology solution for primary modifiable risk factors for noncommunicable diseases: Pilot and feasibility study in norway. JMIR formative research **6**(2), e33636 (2022)
20. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263–1284 (2009)
21. Japkowicz, N., et al.: Learning from imbalanced data sets: a comparison of various strategies. In: AAAI Workshop on Learning from Imbalanced Data Sets. vol. 68, pp. 10–15. AAAI Press Menlo Park, CA (2000)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
23. Kunanbayev, K., et al.: Complex encoding. In: International Joint Conference on Neural Networks. pp. 1–6. IEEE (2021)
24. Lavanya, D., Rani, K.U.: Performance evaluation of decision tree classifiers on medical datasets. International Journal of Computer Applications **26**(4), 1–4 (2011)
25. Maas, A.H., Appelman, Y.E.: Gender differences in coronary heart disease. Netherlands Heart Journal **18**(12), 598–603 (2010)
26. Marchese Robinson, R.L., et al.: Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. Journal of Chemical Information and Modeling **57**(8), 1773–1792 (2017)
27. Martínez-Agüero, S., et al.: Interpretable clinical time-series modeling with intelligent feature selection for early prediction of antimicrobial multidrug resistance. Future Generation Computer Systems **133**, 68–83 (2022)
28. Meng, C., et al.: Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. Scientific Reports **12**(1), 1–28 (2022)
29. Meyer, D., Wien, F.T.: Support vector machines. The Interface to libsvm in Package e1071 **28** (2015)
30. Mohd Noor, N.A., et al.: Consumer attitudes toward dietary supplements consumption. International Journal of Pharmaceutical and Healthcare Marketing **8**(1), 6–26 (2014)
31. Mora-Jiménez, I., et al.: Artificial intelligence to get insights of multi-drug resistance risk factors during the first 48 hours from icu admission. Antibiotics **10**(3), 239 (2021)
32. Naim, F.A., et al.: Effective rate of minority class over-sampling for maximizing the imbalanced dataset model performance. In: Proceedings of Data Analytics and Management, pp. 9–20. Springer (2022)
33. Nelsen, R.B.: An introduction to copulas. Springer Science & Business Media (2007)

34. Ngiam, K.Y., Khor, W.: Big data and machine learning algorithms for health-care delivery. The Lancet Oncology **20**(5), e262–e273 (2019)
35. Organization, W.H., et al.: Noncommunicable diseases country profiles 2018 (2018)
36. Organization, W.H., et al.: Noncommunicable diseases: Progress monitor 2020 (2020)
37. Psaltopoulou, T., Hatzis, G., et al.: Socioeconomic status and risk factors for cardiovascular disease: impact of dietary mediators. Hellenic Journal of Cardiology **58**(1), 32–42 (2017)
38. Pu, Y., et al.: Variational autoencoder for deep learning of images, labels and captions. Advances in Neural Information Processing Systems **29**(1), 295–308 (2019)
39. Ranstam, J., Cook, J.: Lasso regression. Journal of British Surgery **105**(10), 1348–1348 (2018)
40. Refaeilzadeh, P., et al.: Cross-validation. Encyclopedia of Database Systems **5**, 532–538 (2009)
41. Safavian, S.R., Landgrebe, D.: A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics **21**(3), 660–674 (1991)
42. Stanfill, C., Waltz, D.: Toward memory-based reasoning. Communications of the ACM **29**(12), 1213–1228 (1986)
43. Steinwart, I., Christmann, A.: Support vector machines. Springer Science & Business Media (2008)
44. Taylor Jr, H.A., et al.: Relationships of bmi to cardiovascular risk factors differ by ethnicity. Obesity **18**(8), 1638–1645 (2010)
45. Van Rijsbergen, C.J.: The geometry of information retrieval. Cambridge University Press (2004)
46. Wagner, K.H., Brath, H.: A global view on the development of non communicable diseases. Preventive Medicine **54**, S38–S41 (2012)
47. Xu, L., et al.: Modeling tabular data using conditional gan. Advances in Neural Information Processing Systems **32** (2019)
48. Xu, W., Tan, Y.: Semisupervised text classification by variational autoencoder. IEEE Transactions on Neural Networks and Learning Systems **31**(1), 295–308 (2019)
49. Yusuf, H.R., et al.: Impact of multiple risk factor profiles on determining cardiovascular disease risk. Preventive Medicine **27**(1), 1–9 (1998)