Informatics in Medicine Unlocked 23 (2021) 100538

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

Prediction of secondary testosterone deficiency using machine learning: A comparative analysis of ensemble and base classifiers, probability calibration, and sampling strategies in a slightly imbalanced dataset

Monique Tonani Novaes [a], Osmar Luiz Ferreira de Carvalho [b], Pedro Henrique Guimarães Ferreira [b], Taciana Leonel Nunes Tiraboschi [a], Caroline Santos Silva [a], Jean Carlos Zambrano [a], Cristiano Mendes Gomes [c], Eduardo de Paula Miranda [d], Osmar Abílio de Carvalho Júnior [e,*], José de Bessa Júnior [f,**]

[a] Department of Public Health and Epidemiolgy, Universidade Estadual de Feira de Santana, Avenida Transnordestina, S/n - Novo Horizonte, 44036-900, Feira de Santana, Bahia, Brazil
[b] Departament of Electrical Engineering, University of Brasília, University Campus Darcy Ribeiro, Asa Norte, University of Brasília, DF, 70910-900 Brasília, Brazil
[c] Division of Urolgy, Universidade de São Paulo, São Paulo, São Paulo, Brazil
[d] Division of Urolgy, Universidade Federal Do Ceara, Fortaleza, Ceara, Brazil
[e] Departament of Geografia, University of Brasília, University Campus Darcy Ribeiro, Asa Norte, University of Brasília, DF, 70910-900 Brasília, Brazil
[f] Division of Urology, Universidade Estadual de Feira de Santana, Feira de Santana, Bahia, Brazil

ARTICLE INFO

Keywords:
Machine learning
Imbalanced data
Testosterone deficiency
Ensemble classifier

ABSTRACT

Testosterone is the most important male sex hormone, and its deficiency brings many physical and mental harms. Efficiently identifying individuals with low testosterone is crucial prior to starting proper treatment. However, routine monitoring of testosterone levels can be costly in many regions, resulting in an underreporting of cases, especially in developing countries. Moreover, there are few studies that employ machine learning (ML) in prognosticating testosterone deficiency. This research, therefore, aims to offer a coherent comparative analysis of machine learning methods that can predict testosterone deficiency without having patients undergo costly medical tests. In doing so, we seek to provide to the urological community a publicly available dataset (https://github.com/osmarluiz/Testosterone-Deficiency-Dataset) to increase research in this yet untapped field. For this analysis, we used ten base classifiers (optimized with grid search stratified K-fold cross-validation); three ensemble methods; and eight sampling strategies to analyze a total of 3397 patients. The analysis was based on six features (age; abdominal circumference; triglycerides; high-density lipoprotein; diabetes; and hypertension), all of which were obtained by low-cost exams. We compared the sampling strategies and the classifiers' performance on an independent test set using ranking (PR-AUC), probabilistic (Brier score), and threshold metrics. We found that: (1) within the ranking metrics, sampling strategies did not enhance results in this slightly imbalanced (4:1 ratio) dataset; (2) the ensemble classifier using weighted average presented the best performance; (3) the best base classifier was XGBoost; (4) calibration showed significant improvement for the sampling strategies and slight improvements for the no sampling strategy; (5) the McNemar's test presented statistically similar results among all classifiers; and (6) abdominal circumference (AC) had by far the highest feature importance, followed by triglycerides (TG). Age showed very little significance in predicting testosterone deficiency.

* Corresponding author.
** Corresponding author.
E-mail addresses: moniquetonani@yahoo.com.br (M.T. Novaes), osmarcarvalho@ieee.org (O.L. Ferreira de Carvalho), pedroferreira@ieee.org (P.H. Guimarães Ferreira), tacianaleonel@hotmail.com (T.L. Nunes Tiraboschi), s.carolinne5@gmail.com (C.S. Silva), zambranojeancarlos@gmail.com (J.C. Zambrano), crismgomes@uol.br (C.M. Gomes), mirandaedp@gmail.com (E. de Paula Miranda), osmarjr@unb.br (O. Abílio de Carvalho Júnior), bessa@uefs.br (J. de Bessa Júnior).

https://doi.org/10.1016/j.imu.2021.100538
Received 19 December 2020; Received in revised form 5 February 2021; Accepted 10 February 2021
Available online 16 February 2021

1. Introduction

Testosterone is the most important sex hormone among males and significantly impacts men's physical and psychological well-being [1,2]. Patients with Testosterone Deficiency Syndrome (TDS) may experience hypogonadism, a condition defined by low serum testosterone levels combined with clinical symptoms. This condition is associated with various comorbidities, such as metabolic syndrome, cardiovascular diseases, erectile dysfunction, atherosclerosis, respiratory problems, depression, and other complications that reduce overall health indicators [2–8].

Hypogonadism has two leading causes: primary and secondary [9,10]. Primary hypogonadism (hypergonadotropic hypogonadism) is often associated with primary testicular failure, resulting in an inability to produce physiological levels of testosterone, leading to an androgen deficiency and increase in gonadotropin concentration. Primary hypogonadism is much less common and may be due to congenital (Klinefelter syndrome, Y-chromosome microdeletions, mutations in luteinizing hormone and follicle-stimulating hormone receptors, myotonic dystrophy, and cryptorchidism) or acquired causes (testicular trauma or torsion, testicular radiation, orchitis, chemotherapy with alkylating agents, treatment with ketoconazole, autoimmune testicular failure, infiltrative disease, varicocele, sickle-cell disease, cirrhosis, and excessive alcohol intake) [11].

Secondary hypogonadism (hypogonadotropic hypogonadism) results in low or inappropriately normal gonadotropin levels, which impacts the secretion of testosterone by the testicles and prompts a negative feedback on the hypothalamus-pituitary unit. Secondary hypogonadism is more common and can also result from several congenital disorders (e.g., Kallmann syndrome; Prader-Willi syndrome; or mutations in LH and FSH receptors) or acquired causes (e.g., hyperprolactinemia; pituitary damage from tumors; apoplexy; infection or infiltrative disease; head trauma; acute systemic illness; medications; sickle-cell disease; morbid obesity or diabetes; eating disorders; excessive exercise; cirrhosis; or idiopathic hypogonadotropic hypogonadism) [11].

Secondary hypogonadism has therapeutic implications [12]. It can result from functional causes (e.g., obesity, type 2 diabetes, opioids, or systemic disease) and can be reversible with treating or preventing of these conditions. Several reviews and meta-analyses demonstrate the relationship between obesity and male hypogonadism [13–18]. Studies suggest that obese men can increase their testosterone levels upon weight loss, either by supervised diet, exercise, or, in more extreme cases, bariatric surgery [19–21]. Other treatments make use of hormone replacement therapy or testosterone gel supplementation [22–25], in which testosterone treatment reduced visceral adiposity and waist circumference [26–28]. In sum, a number of studies have shown that the increase in testosterone levels can lead to weight loss and vice versa [14].

Several studies point to an association between testosterone levels, triglycerides (TG) [29,30] and hypertension (HT) [31–33]. Moreover, low testosterone is strongly associated with type 2 diabetes (T2D), given that one-third of men with T2D have secondary hypogonadism [34]. Yao et al. [35] perform a systematic meta-analysis and conclude that higher testosterone levels in men can significantly decrease the risk of T2D. Long-term testosterone therapy in hypogonadal men prevents the progression of prediabetes [36] and achieves T2D remission [37]. However, the benefit-risk balance of long-term testosterone treatment is unclear in men with T2D [38]. These associations suggest that testosterone is a component of the metabolic syndrome, a cluster of risk factors, including abdominal obesity, dyslipidemia, HT, and insulin resistance [39–41]. Hence, testosterone replacement therapy is sometimes an adjunctive therapeutic option for metabolic syndrome [42].

The diagnosis of TDS requires biochemical evaluation of total testosterone (TT) (<300 ng/dl) [12] or free testosterone (FT) (<6.5 ng/dl) [43] levels via blood test. However, men in the general population do not carry out routine monitoring of TT and FT levels due to high costs. This results in a high rate of unidentified and untreated patients suffering from low testosterone levels [44]. Unlike women's health care service (which includes mammograms, cervical cancer screening, gynecological services, etc.), men's are generally not gender specific [45]. The lack of diagnosis and control of TDS is one of the reasons life expectancy in men is shorter than that of women [45]. Studies highlight that decreasing TT levels increase the Charlson comorbidity index (CCI) [46,47]. Testosterone deficiency is also a factor in increased cardiovascular and all-cause mortality, as shown by systematic reviews and meta-analyses [48–53]. Testosterone replacement therapy in hypogonadal men provides a 9–10% increase in five-year survival rate, like eugonadal men [54]. Yet, a significant obstacle is that few men seek health care despite treatment options [55]. This problem is even more acute in developing countries confronting income inequality and unequal access to health services [56]. In Brazil, the cost of testosterone dosing is six to eight times higher than that of blood glucose, TG, and high-density lipoprotein (HDL) cholesterol.

Predictive analysis using Artificial Intelligence (AI) algorithms is of great interest to those working in medical diagnosis since it provides indispensable resources for data analysis [57]. Clinical prediction rules combine several predictors based on the weights assigned to each predictor, obtaining a risk or probability. The likelihood of having the disease can be used to prompt urological referral for further testing based on the risk of a particular health condition [58]. When applied to many medical specialties, ML have already shown promising results in predicting a variety of clinical illnesses and conditions, such as cardiovascular risks [59–61], diabetes mellitus [62,63], cancer [64,65], kidney diseases [66], metabolic syndrome [67], and appendicitis [68].

However, a search for scientific articles in English in the Web of Science database using the keywords "hypogonadism" and "machine learning" in the period between January 1945 and November 2020 yielded only a single article, by Lu et al. [44]. A search with the words "testosterone" and "machine learning" yielded 11 more articles, mostly about cancer, which do not directly address the issue of TDS [69–72]. Accordingly, no articles were found in the database addressing the use of meta-classifiers or ensemble classifiers in either detection or prediction of TDS – a gap this research aims to fill.

We assessed that it is difficult to apply predictive algorithms (i.e., ML and deep learning) to hypogonadism, especially when the condition is caused by external factors. Nevertheless, the fact that TDS caused by secondary causes are often associated with comorbidities such as obesity, metabolic syndrome, and systemic illnesses offer ML algorithms ample data, which may boost its predictive ability.

Prediction studies has achieved high performance using methods based on ML and deep learning (DL). Nevertheless, two factors make traditional ML adequate for several investigations [73]: (a) DL does not work well with small amounts of data, making it more suitable with big data; and (b) reliance on DL hardware, which requires the Graphics Processing Unit (GPU). However, defining the best ML configuration for a particular clinical prediction should test a set of procedures: (a) different base or ensemble classifiers; (b) strategies for dealing with imbalanced learning; (c) calibration for reliable risk predictions; and (d) the use of different metrics for performance analysis in classification, considering independent validation.

A large number of ML techniques have been used and compared in various medical fields [74,75]. However, some challenges are difficult to solve using a single ML classifier, and the optimal solution may be outside the scope of a single model. Therefore, we assessed that one way to overcome this deficiency was to use the ensemble-based classifier that combines models to improve predictive performance [76]. The ensemble algorithm has two stages [77]: (a) the first stage applies several classifiers independently, and (b) the second stage uses the outputs of the individual classifiers as input to perform a new prediction. Besides, ensemble algorithms usually yield [78]: (a) increased performance, especially when applied to small amounts of data, due to the

greater propensity to find different hypotheses in the prediction of training data; (b) reducing the due to a greater tendency to find different hypotheses in the prediction of training data; (b) reduced risk of obtaining a local minimum and choosing an incorrect hypothesis; and (c) an increased analysis among methods due to a wide combination of models. Ensemble-based classifiers have been used successfully in various biomedical research such as bioinformatics [79–81], breast cancer diagnosis [82], diabetes prediction [83,84], and monitoring of the intensive care unit [85].

Commonly, medical data contains an uneven distribution of observations [86,87], in which only a small portion of patients experiences a health problem. Therefore, depending on the proportion of negative and positive samples, pre-processing the imbalanced data may be necessary because conventional algorithms are prone to consider minority observation as noise [88,89]. In this regard, imbalanced data may introduce biased results in predictive modeling. The methods for resolving the class imbalance problem at the data level are subdivided into undersampling (US), oversampling (OS), and hybrid sampling (HS).

The present study offers an analysis of the use of ML in predicting testosterone deficiency, and aims to make its dataset publicly available (https://github.com/osmarluiz/Testosterone-Deficiency-Dataset). Our investigation uses several trends in the field of ML. We compared ten traditional ML classifiers, optimized using grid search and stratified K-fold cross-validation, and three ensemble classifiers. We evaluated different class imbalance treatment methods, including undersampling, oversampling, and hybrid techniques. Therefore, we compared multiple classification and sampling techniques, extracting the best quality from each procedure in a wide set of analyses, in the effort to properly address this important medical issue.

2. Material and methods

The methodology is divided into the following steps (Fig. 1): (2.1) dataset acquisition and split; (2.2) base classifiers; (2.3) ensemble classifiers (second level classifiers); (2.4) sampling strategies; (2.5) stratified K-fold cross-validation; (2.6) grid search; (2.7) probability calibration; and (2.8) accuracy analysis.

2.1. Dataset acquisition and split

We gathered data from a sample of 3397 patients between the ages of 40 and 85 drawn from a urology clinic in Feira de Santana, Brazil. Participants with primary hypogonadism or undergoing treatment were excluded from the analysis. The features were obtained by low-cost routine exams: age, diabetes, HT, HDL, and AC (Table 1). TG was the feature that presented the highest standard deviation (88.84). HAS and diabetes are categorical features, represented by absence (0) or presence (1). The medical literature suggests that normal testosterone levels

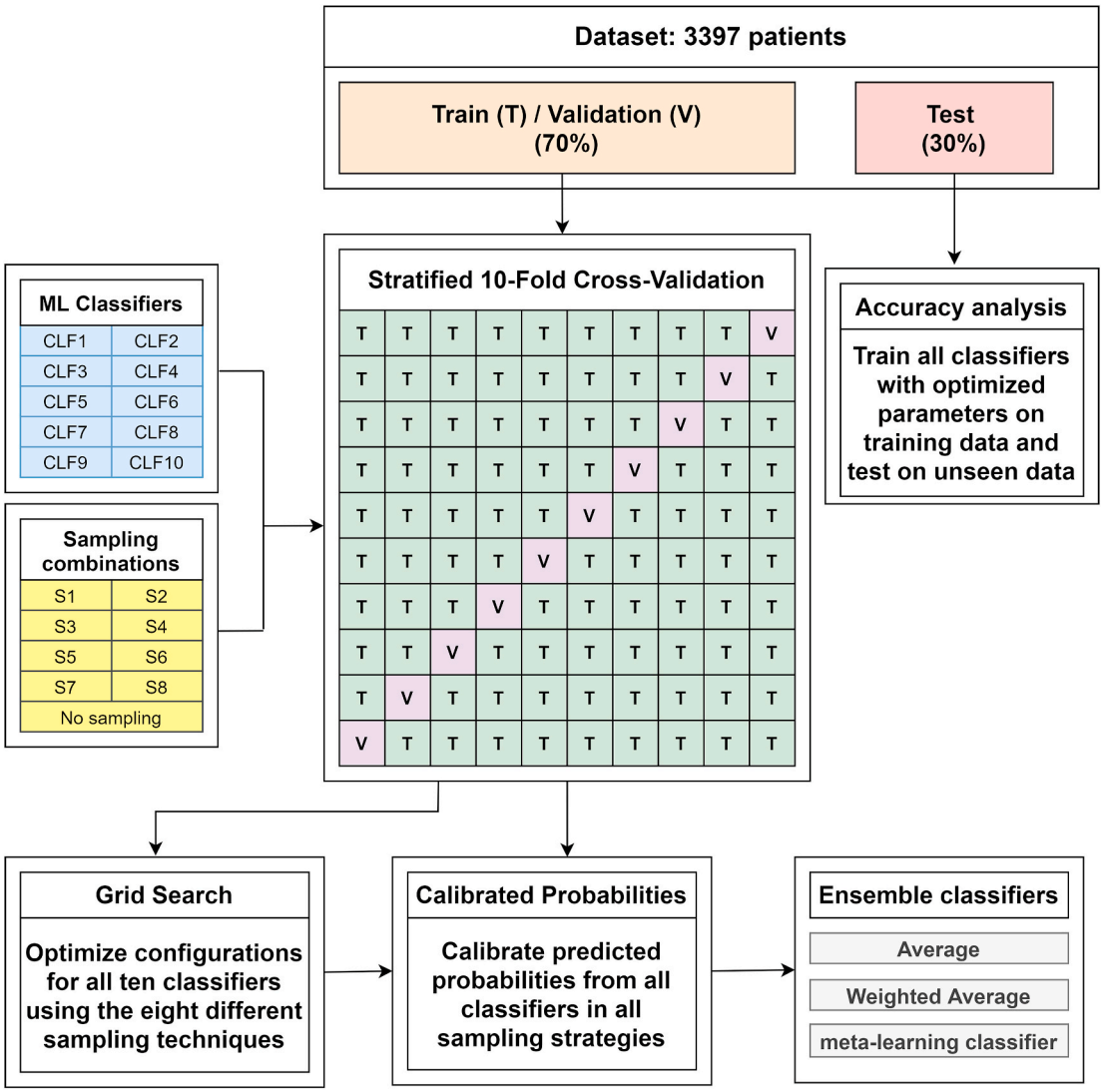

Fig. 1. Flowchart of the training procedure to obtain optimal parameters for each single classifier and the weights for the ensemble-classifier.

Table 1
Descriptive analysis from the seven features used in this experiment: Age, Diabetes, Triglycerides (TG), Hypertension (HT), High-density lipoprotein (HDL), Abdominal Circumference (AC), and Testosterone (T).

Input	Description	Range	Descriptive Statistics
Age	Age in years	40–85	$\mu = 61.33$; $\sigma = 10.07$
Diabetes	Diabetes	Yes/No	Yes = 39%; No = 61%
TG	Triglycerides (mg/dl)	20–809	$\mu = 155.27$; $\sigma = 88.84$
HT	Hypertension	Yes/No	Yes = 51%; No = 49%
HDL	High-density lipoprotein (mg/dl)	20–116	$\mu = 46.33$; $\sigma = 10.96$
AC	Abdominal Circumference (cm)	66–145	$\mu = 98.92$; $\sigma = 10.63$
T	Testosterone (ng/dl)	25–1375	$\mu = 449.19$; $\sigma = 172.52$

range from 300 to 1200 (ng/dl) [90]. For this reason, we separated testosterone in two classes: (a) 0 ($T < 300$ ng/dl) and (b) 1 ($T \geq 300$ ng/dl). Fig. 2 shows the class distribution, where the class imbalance ratio (i.e., number of samples from the majority class divided by number of samples from the minority class) is approximately 4:1 (slightly imbalanced). Regarding data partition, we separated 30% of the data to the testing stage only, and then implemented stratified K-fold cross-validation ($k = 10$) in the remaining 70% data. The test set provides an independent validation, demonstrating the model's ability to generalize unseen data.

2.2. Base classifiers (first level classifiers)

The application of several classifiers has two main advantages: (a) it enables a vast comparison between ML algorithms, and (b) the use of different classifiers bearing low correlation between one another presents better results for the ensemble classifiers. Therefore, we used ten classifiers compatible with the scikit-learn library: Artificial Neural Networks (ANN) [91], Supporting Vector Machine (SVM) [92]; Random Forest (RF) [93], Extremely Randomized Trees (ERT) [94], AdaBoost [95], XGBoost [96], Gradient Boosting (GB) [97], k-Nearest Neighbors (k-NN) [98], Naïve Bayes (NB) [99], and Logistic Regression (LR) [100].

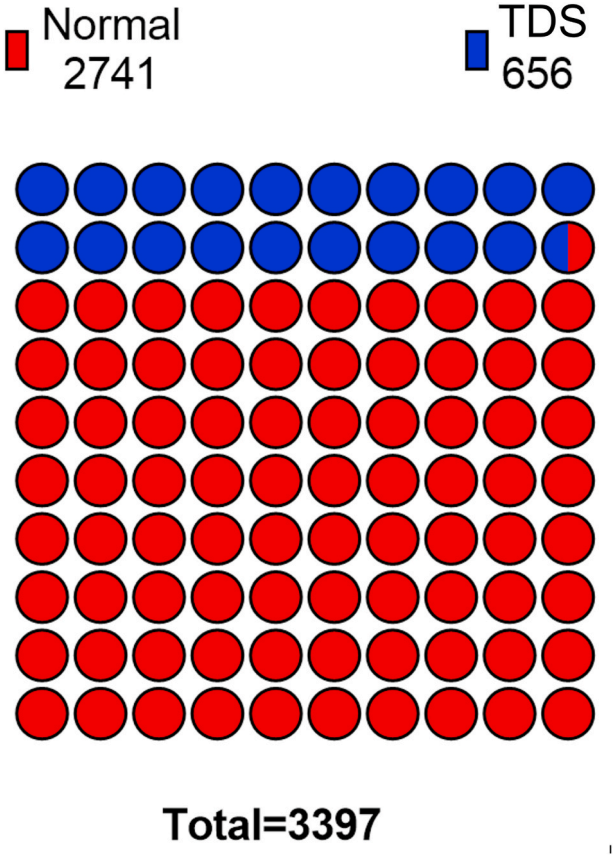

Fig. 2. Representation of class imbalance, where blue represents patients with Testosterone Deficiency Syndrome (TDS) ($T < 300$ ng/dl) and red illustrates patients with normal levels of testosterone ($T \geq 300$ ng/dl). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

2.3. Ensemble classifiers (second level classifiers)

The ensemble classifiers usually outperform the best single classifier because it combines the other methods' strengths to integrate a more complex and powerful learning approach. The ensemble predictions use the probability outputs from the base classifiers as input to predict new data. There are different ensemble strategies, such as average (Avg), weighted average (wAvg), majority voting, rank average, and stacking with learning algorithms. This research used three ensemble classifiers: (a) Avg; (b) wAvg; and (c) Stacking Meta-Classifier with Logistic Regression (meta-classifier).

The Avg classifier is the most straightforward approach, in which the final probability is the average probabilities from all classifiers. The wAvg gives higher weights to the better classifiers. Defining the best set of weights for the ensemble classifier can be an exhaustive procedure since the number of possible combinations may result in millions of iterations. To overcome this problem, we applied a randomized grid search, in which each classifier may have weights in the range of 0–1 with 0.1 steps. We established 10,000 as the maximum number of iterations. The meta-classifier uses the probability scores from the base classifiers as features. In our research, we used LR as the second level classifier.

2.4. Sampling strategies

Datasets are imbalanced when the distribution of classes is uneven [101], and they are prevalent in real-world problems covering many scientific fields [102,103]. ML algorithms often present bad classification results in imbalanced data. There are many ways to address class-imbalance [104]. At the data level, there are three main sampling methods: (a) undersampling (reduction of samples from the majority class), (b) oversampling (enlargement of samples from the minority class), and (c) hybrid sampling (a combination of oversampling and undersampling) [105]. This research compared eight sampling strategy combinations using the open-source python toolbox Imbalanced-Learn [106]: Random undersampling (RUS), Repeated Edited Nearest Neighbors (RENN), Random Over-Sampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE), RENN + SMOTE, RUS + ROS, RENN + RUS, and RUS + SMOTE.

RUS is the most straightforward approach, where the removal of elements from the majority class is random [107]. However, the RUS may remove essential elements from the majority class, decreasing the model's functionality. Other solutions aim to minimize this effect, removing redundant elements instead of randomly. In this context, RENN is a more sophisticated undersampling technique [108], where the removed elements tend to be redundant with their nearest neighbors.

Like RUS, the ROS technique consists of duplicating random rows. This method uses a substantial amount of data from the majority class, but the duplicated elements may introduce overfitting [109]. SMOTE [110] is very popular in imbalanced data problems because it reduces the overfitting effect compared to random oversampling since it generates new information.

2.5. Stratified K-Fold cross-validation

The stratified K-fold cross-validation separates the dataset into k bins of equal size, maintaining the same ratio of positive and negative

instances from the original dataset, using k-1 bins for training and one bin for testing, varying the testing position bin. The result is the average of all testing bins for a specific accuracy metric. This approach gives more realistic results than the standard train/test, especially when the objective is to optimize parameters since it uses many facets of the data, reducing variance. This method has three utilities in our study: (a) providing better hyperparameters through grid search; and (b) obtaining the wAvg classifier weights; and (c) obtaining the training data (predictions from the base classifiers) to the meta-classifier.

2.6. Grid search

We used the grid search with cross-validation to obtain the optimal hyperparameter values for all classifier-sampling combinations to improve PR AUC. The only exception was the NB classifier, which does not have parameters to tune. Thus, we obtained optimal values for the ten classifiers considering each sampling combination, resulting in 90 optimized classifiers (eight sampling strategies and a no sampling strategy multiplied by ten single classifiers). Table 2 lists the grid values within each iteration.

Table 2
Grid Search values for every single classifier: Random Forest (RF), Extremely Randomized Trees (ET), Gradient Boosting (GB), Supporting Vector Machine (SVM), k-Nearest Neighbor (k-NN), Logistic Regression (LR), AdaBoost (ADA), and Artificial Neural Networks (ANN).

Model	Parameter	Values
RF	bootstrap	(True, False)
	oob_score	(True, False)
	max_depth	3, 4, 5, 6, 7
	n_estimators	50, 100, 150, 200, 250
	min_samples_split	2, 3, 4, 5
	max_leaf_nodes	None, 2, 3, 4
ET	Bootstrap	(True, False)
	oob_score	(True, False)
	n_estimators	100, 150, 200, 250, 300
GB	n_estimators	100, 200, 300, 400, 500, 600
	learning_rate	0.01, 0.05, 0.1
	subsample	0.3, 0.4, 0.5, 0.6
	max_depth	3, 4, 5, 6, 7
	min_samples_split	2, 3, 4
SVM	kernel	Linear, rbf, poly
	degree	2, 3, 4
	C	0.5, 1, 2, 3, 4, 5, 6
	Class_weight	Balanced, None
k-NN	N_neighbors	5, 10, 15, 20, 25, 30
	Weights	Uniform, distance
	Algorithm	Ball_tree, kd_tree, brute
	P	1, 2, 3
LR	C	0.5, 1, 2, 3, 4, 5, 6
	penalty	L1, L2, elasticnet
	solver	Newton-cg, lbfgs, saga
	max_iter	50, 100, 200
	class_weight	Balanced, None
ADA	DT max_depth	None, 2, 3, 4
	DT min_samples_split	2, 3, 4
	DT max_leaf_nodes	None, 2, 4
	DT max_features	None, 3, 5
	n_estimators	300, 400, 500, 600
	learning_rate	0.1, 0.01
XGBoost	min_child_weight	1, 3, 5, 7, 10
	gamma	1, 3, 5, 7, 10
	colsample_bytree	0.4, 0.5, 0.6
	reg_alpha	0, 0.2, 0.3
	max_depth	4, 5, 6
	subsample	0.6, 0.7, 0.8
	n_estimators	100, 200, 300, 400, 500
	learning_rate	0.01, 0.05, 0.1
ANN	hidden_layer_sizes	(10, 10), (15, 15), (20, 10), (20, 15)
	Activation	Logistic, tanh, relu
	learning_rate	0.01, 0.001
	max_iter	200, 400, 600
	Solver	Lbfgs, sgd, adam

2.7. Probability calibration

The probability or probability-like score from ML algorithms may not represent the observed proportions in real-world scenarios [111] because they are often uncalibrated. Besides, the ML classifiers may present different probability distributions. The class-imbalance correction techniques (undersampling, oversampling, and hybrid sampling) produce consistently biased probability estimates. Uncalibrated probabilities can show imprecise risk predictions, which may prove highly consequential in medical diagnoses [112,113]. Besides, comparing algorithms with uncalibrated probabilities may induce unrealistic results, especially in threshold metrics. In sum, uncalibrated estimates present three significant problems: (a) risk predictions are not reliable; (b) comparing classifiers using threshold metrics is difficult, as the same threshold can present significantly different results between methods; and (c) the ensemble-classifier will have unrealistic biases from each classifier. Therefore, we applied two calibration techniques: (a) prior probability correction (for the sampling strategies) [114], and (b) isotonic regression (for the no sampling procedure) [115] using scikit-learn. The prior correction used the following expression (Equation (1)):

$$p_{new} = \frac{p_{old} * \frac{r1}{r2}}{p_{old} * \frac{r1}{r2} + (1 - p_{old}) * \frac{(1-r1)}{(1-r2)}} \quad (1)$$

Where:

p_{new}: the new calibrated probability,
p_{old}: the prior probability from the classifier,
r1: the ratio of positive samples in the original dataset,
r2: the ratio of positive samples in the sampled dataset.

2.8. Accuracy analysis

Accuracy analysis is a fundamental component in comparing ML models. We used three commonly used performance metrics: (2.5.1) ranking metrics, (2.5.2) probabilistic metrics, and (2.5.3) threshold metrics.

2.8.1. Ranking metrics

Ranking metrics are very efficient in understanding the classifier's abilities to differentiate classes. There are two main metrics: (a) Receiving Operating Characteristic Area Under the Curve (ROC AUC), and (b) Precision-Recall Area Under the Curve (PR AUC). Both metrics consider the probabilities of the classifiers (the model calibration does not affect these results). The ROC curve uses the four quadrants from the confusion matrix (Fig. 3): True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN). The axis is referent to the True Positive rates (TP/(TP + FN)) and the False Positive rates (FP/(FP + TN)). In its turn, the PR Curve does not use the TN quadrant, and it compares the axis Precision (TP/(TP + FN)) and Recall (TP/(TP + FP)) at different thresholds. In imbalanced datasets, the ROC AUC scores may give over-optimistic results, whereas the PR AUC scores gives more attention to the minority class [116,117]. Thus, to evaluate the best classifier, we used PR AUC score. However, we also present the results of the ROC curve as it is the most used metric performance in medical studies with ML [75].

2.8.2. Probabilistic metrics

Probabilistic metrics enable an evaluation of how well-calibrated each model's output is. In this research, when applying sampling strategies, the classifiers' training data has different a priori probabilities than the test set's probabilities, since they have different proportions of positive and negative samples (healthy and unhealthy patients), resulting in lower probabilistic metric scores. Thus, we used the Brier Score Loss before and after the calibration procedure (Equation (1)). The Brier Score Loss (Equation (2)) is given by:

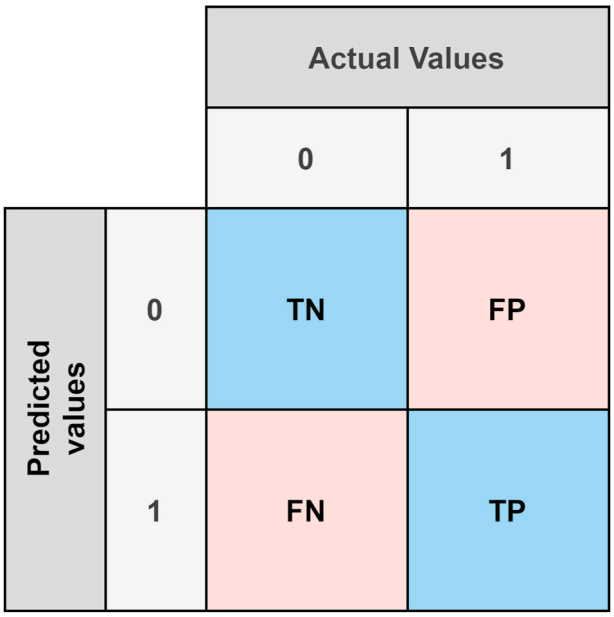

Fig. 3. Confusion Matrix, where TN is True Negatives, FP is False Positives, FN is False Negatives, and TP is True Positives.

$$\text{Brier Score} = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \quad (2)$$

Where:

N: number of elements (in our case, the number of patients in the test set),

f_t: the classifier output,

o_t: the actual outcome.

2.8.3. Threshold metrics

After selecting the best model, we evaluated other metrics obtained from the confusion matrix (Fig. 3, Table 3): (a) accuracy; (b) precision; (c) recall; (d) specificity; (e) F-score; (f) positive predictive value; and (g) negative predictive value. The threshold cutoff point may vary according to the clinical scenario. Whenever testing many patients is viable, a threshold with high recall is preferred. When it is not (e.g., due to financial constraints), a threshold with a high precision is more adequate. Thus, we chose the cutoff point within every classifier by selecting the highest F-score.

We applied the McNemar's test [118] to verify differences and similarities between the models within the threshold metrics. The analysis is pairwise, and when comparing different classifiers, the chi-squared statistic (χ^2) has one degree of freedom, as follows (Equation (3)):

Table 3
Threshold metrics.

Accuracy Metric	Equation
Overall *Accuracy*	$\frac{TP+TN}{TP+FP+TN+FN}$
Specificity	$\frac{TN}{TN+FP}$
Sensitivity	$\frac{TP}{TP+FN}$
F-Score	$2\times\frac{P\times R}{P+R}$
Positive Predivtive Value	$\frac{TP}{TP+FN}$
Negative Predictive Value	$\frac{\text{TN}}{TN+FN}$

$$\chi 2 = \frac{(b-c)^2}{(b+c)} \quad (3)$$

Where:

$\chi 2$: chi-squared statistics,

b and c: elements from the secondary diagonal from the 2x2 contingency table.

Moreover, we can reject the null hypothesis (assuming a different behavior within the classifiers) if the obtained χ^2 is higher than the values of the χ^2 distribution table.

3. Results

3.1. Ranking metrics

We used the PR AUC and ROC AUC scores to evaluate the different sampling strategies (Table 4). The wAvg classifier with no sampling presented the best results. Besides, no sampling presented the best results among most base classifiers, except for LR, NB, GB, SVM, and k-NN. Among those five classifiers, k-NN benefited from an undersampling technique (RUS), whereas all other classifiers presented better results using ROS or SMOTE. Commonly, SMOTE is preferred over ROS because it avoids overfitting, since it generates new data instead of random duplicating rows [110]. However, in our dataset, there was no sign of overfitting using ROS. Nevertheless, SMOTE presented better results within the ensemble classifiers, being the best sampling strategy apart from no sampling. The similar values between different sampling techniques suggest that class imbalance is not skewed enough to benefit from those methods. The application of an exhaustive grid search for each classifier in each sampling strategy trims down differences in results.

Apart from NB and k-NN, the base classifiers presented good overall results, providing good predictions with a slight variance within each sampling strategy. The XGBoost classifier presented the best overall results among the base classifiers. The ensemble classifiers (Avg, wAvg, and meta-classifier) presented more stable results (smaller variance). The wAvg classifier presented PR AUC values surpassing 44% among all sampling strategies.

The ROC AUC score is widely used within the medical community to compare ML algorithms. Since we optimized the base classifiers on their PR AUC scores (because the minority class is more relevant in this scenario), the ROC AUC scores may not represent the highest possible values [117]. Nevertheless, the ROC AUC scores reinforce the PR AUC conclusions, in which the base classifiers (apart from k-NN and NB) display overall good results. The sampling strategies do not present better results than no sampling, reinforcing that it is crucial to make a wide analysis and consider simpler methods, especially in slightly imbalanced datasets. Among, the sampling strategies, ROS and SMOTE presented slightly better results than the rest.

3.2. Probabilistic metrics

Table 5 lists the Brier score before and after calibration for all sampling strategies. The prior probability technique (eight first columns) presented improvements within all base classifiers' sampling strategies. Nevertheless, calibration for the no sampling strategy using isotonic regression presented less improvement but better score values. The sampling strategies' application introduces an additional (expected) probabilistic bias shown by the Brier score before the calibration application. RENN was the sampling strategy with the worst calibration. Within the models using isotonic regression, SVM, LR, ANN, and k-NN did not improve using this method. We also evaluated the sigmoid scaling within those models and it still presented no improvement, showing that these classifiers are already well-calibrated for this task.

The Brier score loss analysis gives useful insight into how well-calibrated the models are and enables more in-depth information on

Table 4
Average Precision from the different undersampling and oversampling combinations: Repeated Edited Nearest Neighbors (RENN), Synthetic Minority Oversampling Technique (SMOTE), Random Undersampling (RUS), and Random Oversampling (ROS), and ten Machine Learning Models: Random Forest, Extremely Randomized Trees, Gradient Boosting, Supporting Vector Machine (SVM), k-Nearest Neighbor (k-NN), Naïve Bayes (NB), Logistic Regression (LR), AdaBoost, XGBoost, Artificial Neural Networks (ANN), Average Ensemble Classifier (Avg), Weighted Average Ensemble Classifier (wAvg), and Meta-learning classifier with Logistic Regression (Meta).

PR AUC Score									
	Undersampling		Oversamping		Hybrid Sampling				
Model	RUS	RENN	ROS	SMOTE	RUS + ROS	RENN + ROS	RUS + SMOTE	RENN + SMOTE	No Samp
RF	42.3	42.6	43.8	42.5	42.2	42.1	42.5	41.5	44.1
ERT	41.9	41.6	41.4	41.5	41.9	41.3	41.0	42.2	42.2
GB	41.6	43.5	43.9	41.7	43.4	43.0	40.6	42.8	43.1
SVM	42.4	41.6	43.3	43.1	42.1	41.5	42.4	41.8	43.4
k-NN	40.8	35.6	34.1	31.7	35.9	36.2	37.7	35.1	36.8
NB	38.8	38.1	40.0	40.6	38.7	38.1	38.6	38.1	40.1
LR	41.9	41.9	42.8	43.2	41.7	41.8	42.1	42.2	42.9
AdaBoost	39.8	41.7	43.1	42.9	40.5	42.3	40.8	40.1	43.9
XGBoost	42.3	43.4	43.1	42.2	43.6	41.0	42.2	43.6	44.7
ANN	41.7	41.8	43.2	42.9	40.9	42.9	41.6	42.2	43.0
Avg	42.6	44.0	43.2	44.2	42.5	43.7	42.3	44.1	44.2
wAvg	44.4	44.2	44.7	45.3	44.5	44.6	44.0	44.9	45.4
Meta	42.6	43.0	43.8	44.6	42.8	42.6	42.9	42.6	44.2
ROC AUC Score									
	Undersampling		**Oversampling**		**Hybrid Sampling**				
Model	**RUS**	**RENN**	**ROS**	**SMOTE**	**RUS + ROS**	**RENN + ROS**	**RUS + SMOTE**	**RENN + SMOTE**	**No Samp**
RF	74.9	75.7	75.7	74.0	74.3	75.3	75.3	74.7	76.2
ERT	74.7	73.8	73.6	73.2	74.9	73.9	74.2	73.8	74.8
GB	75.0	75.5	76.0	73.5	75.1	75.3	73.9	76.0	75.5
SVM	74.9	74.8	75.9	75.4	75.2	74.7	75.2	75.0	75.9
k-NN	70.7	69.5	68.4	66.8	68.7	69.4	69.5	69.2	69.4
NB	73.3	73.6	73.7	73.3	73.4	73.5	73.1	73.5	74.1
LR	75.2	75.0	75.6	75.7	74.9	74.9	75.2	75.3	75.7
AdaBoost	74.5	73.3	75.3	74.4	74.7	75.9	74.5	71.2	75.7
XGBoost	74.9	75.3	75.5	75.1	74.7	74.8	74.7	75.5	76.3
ANN	74.5	73.7	75.8	75.7	74.2	74.6	74.9	74.9	75.4
Avg	75.3	75.3	75.9	75.3	75.1	75.2	75.0	75.4	75.9
wAvg	75.2	75.6	76.2	76.0	75.0	75.5	75.4	75.6	76.0
Meta	75.7	76.1	76.3	75.8	75.4	75.7	75.6	75.9	76.3

Table 5
Brier Score before and after calibration from the ten single classifiers: Random Forest, Extremely Randomized Trees, Gradient Boosting, Supporting Vector Machine (SVM), k-Nearest Neighbor (k-NN), Naïve Bayes (NB), Logistic Regression (LR), AdaBoost, XGBoost, Artificial Neural Networks (ANN), Average Ensemble Classifier (AE), Weighted Average Ensemble Classifier (WAE), and Meta-learning classifier with Logistic Regression (Meta). Sampling strategies used prior probability correction, and no sampling used isotonic regression (ISO).

	Brier Score									
		Prior probability								**ISO**
		Undersampling		**Oversampling**		**Hybrid Sampling**				
Model		**RUS**	**RENN**	**ROS**	**SMOTE**	**RUS + ROS**	**RENN + ROS**	**RUS + SMOTE**	**RENN + SMOTE**	**No Samp**
RF	before	0.209	0.176	0.202	0.195	0.201	0.230	0.211	0.237	0.133
	after	0.135	0.135	0.135	0.136	0.135	0.138	0.138	0.141	0.132
ERT	before	0.219	0.185	0.217	0.217	0.217	0.240	0.219	0.243	0.141
	after	0.141	0.141	0.142	0.142	0.142	0.138	0.142	0.139	0.136
GB	before	0.207	0.222	0.199	0.200	0.204	0.263	0.206	0.243	0.134
	after	0.135	0.173	0.132	0.134	0.135	0.164	0.136	0.143	0.134
SVM	before	0.209	0.219	0.206	0.206	0.209	0.273	0.211	0.274	0.134
	after	0.136	0.165	0.134	0.134	0.135	0.164	0.135	0.167	0.134
k-NN	before	0.213	0.216	0.214	0.229	0.214	0.273	0.219	0.282	0.142
	after	0.14	0.176	0.147	0.151	0.142	0.188	0.142	0.194	0.136
NB	before	0.202	0.244	0.194	0.207	0.196	0.270	0.203	0.279	0.146
	after	0.156	0.212	0.147	0.145	0.158	0.210	0.158	0.214	0.137
LR	before	0.211	0.217	0.205	0.207	0.210	0.266	0.211	0.263	0.134
	after	0.136	0.165	0.134	0.134	0.136	0.157	0.136	0.155	0.134
AdaBoost	before	0.235	0.239	0.224	0.187	0.225	0.244	0.225	0.242	0.172
	after	0.149	0.157	0.146	0.144	0.146	0.143	0.145	0.136	0.133
XGBoost	before	0.227	0.199	0.194	0.167	0.203	0.240	0.221	0.257	0.139
	after	0.146	0.141	0.135	0.139	0.134	0.137	0.143	0.152	0.133
ANN	before	0.218	0.234	0.230	0.215	0.222	0.265	0.204	0.261	0.135
	after	0.136	0.183	0.134	0.133	0.137	0.179	0.136	0.167	0.133
AE	After	0.135	0.150	0.134	0.134	0.135	0.145	0.135	0.147	0.132
WAE	After	0.136	0.172	0.132	0.167	0.175	0.198	0.179	0.193	0.132
Meta	After	0.134	0.134	0.133	0.132	0.134	0.134	0.134	0.133	0.132

model selection. The wAvg classifier (which presented the best-ranking scores) does not provide well-calibrated values when all classifiers are not well calibrated. The no sampling strategy (which has well-calibrated models) provides good Brier Score values to the wAvg classifier. Meanwhile, the meta-classifier presented good results within all sampling strategies.

3.3. Threshold metrics

The threshold metrics enable a good understanding of how the model performs at a specific threshold point. Since the model outputs probabilities, we must choose a specific cutoff point where all values above the chosen threshold will be considered 1 and all values below it will be considered 0. Adjusting this threshold may lead to different strategies (e.g., a low threshold point would assume more patients have the condition, which would signal a need for treating or examining more patients). In contrast, a higher threshold would imply a scenario with a limited amount of testing. The choice of the cutoff point may vary according to the problem specification. When analyzing these metrics, there is a trade-off between metrics (e.g., a higher sensitivity will often imply a lower specificity and vice-versa). The same applies to precision and recall. Besides, calibrated probabilities are critical to compare different classifiers at the same threshold cutoff point. For this reason, we assumed F-score as the most critical metric in this scenario since it is the harmonic average between two metrics (precision and recall).

Table 6 lists the values for the best threshold for each classifier based on their F-score. Even though the classifiers present very similar calibration results, the best threshold point for each one varies. Also, threshold metrics evaluate a single cutoff point, which may present diverging metrics when compared to ranking metrics. In this way, classifiers with lower-ranking metrics may present higher threshold metrics for some points. Nevertheless, ranking metrics are still a safer choice to choose the best classifiers since they tend to present high values for a broader spectrum of threshold values.

Table 7 lists six threshold metrics (accuracy; sensitivity; specificity; positive predictive value; negative predictive value; and F-score) for two scenarios with distinct threshold values: (i) the average from the best thresholds (0.23) and (ii) conventional threshold (0.5). The best classifier for each metric varies substantially, mostly due to a trade-off between the sensitivity and specificity metrics. ANN and weighted average ensemble presented the highest F-score for the 0.23 threshold, whereas RF had the best results with a threshold value of 0.5. There is a significant difference between F-score at both threshold values, highlighting the importance of choosing wisely the operation point and analyzing different thresholds. Although accuracy is the most intuitive metric, analyzing both thresholds make it more evident why it is not appropriate for imbalanced datasets, since even in a scenario with low F-scores (threshold value of 0.5), we achieve higher accuracy values than those with higher F-scores (threshold value of 0.23). The McNemar's test shows statistically equal values at the 1% significance level for all classifiers, given that the differences are tight.

Table 6
Best threshold value for each classifier based on their higher value of F-score.

Classifier	Best threshold	F-score
RF	0.18	49.9
ERT	0.3	48.5
GB	0.17	48.9
SVM	0.26	50.7
k-NN	0.21	43.0
NB	0.22	48.6
LR	0.27	50.6
AdaBoost	0.22	49.2
XGBoost	0.21	50.1
ANN	0.28	50.6
Avg	0.23	50.3
wAvg	0.25	50.9
Meta	0.21	50.2

Table 7
Accuracy analysis for all the ten classifiers: Random Forest (RF), Extremely Randomized Trees (ERT), Gradient Boosting (GB), Supporting Vector Machine (SVM), k-Nearest Neighbor (k-NN), Naïve Bayes (NB), AdaBoost, eXtreme Gradient Boosting (XGBoost), Artificial Neural Networks (ANN), Average Ensemble Classifier (Avg), Weighted Average Ensemble Classifier (wAvg), and Meta-learning classifier with Logistic Regression (Meta). Where, PPV and NPV represent Positive Predictive Value and Negative Predictive Value, respectively.

Threshold at 0.23						
Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F-score
RF	75.5	60.4	79.1	40.9	89.3	48.8
ERT	74.2	59.9	77.6	39.1	89.0	47.3
GB	74.3	59.9	77.8	39.2	89.0	47.4
SVM	75.5	60.9	79.0	41.0	89.4	49.0
k-NN	75.9	45.7	83.1	39.3	86.5	42.3
NB	72.5	61.4	75.2	37.2	89.1	46.4
LR	76.0	62.9	79.1	41.9	89.9	50.3
AdaBoost	74.4	64.0	76.9	39.9	89.9	49.1
XGBoost	73.5	66.0	76.9	39.0	90.2	49.1
ANN	74.5	64.0	77.0	44.0	89.9	49.2
Avg	75.3	64.5	77.9	41.1	90.2	50.2
wAvg	74.5	66.0	76.5	40.2	90.4	50.0
Meta	76.2	59.4	80.2	41.8	89.2	49.1
Threshold at 0.5						
Model	**Accuracy**	**Sensitivity**	**Specificity**	**PPV**	**NPV**	**F-score**
RF	81.5	20.3	96.1	55.6	83.4	29.7
ERT	81.0	9.6	98.1	54.3	81.9	16.4
GB	81.4	15.7	97.1	56.4	82.8	24.6
SVM	81.5	12.2	98.1	60.0	82.3	20.3
k-NN	80.7	5.6	98.7	50.0	81.4	10.0
NB	81.0	15.2	96.7	52.6	82.7	23.6
LR	81.2	10.7	98.1	56.8	82.1	17.9
AdaBoost	81.3	17.8	96.5	54.7	83.1	26.8
XGBoost	81.1	15.7	96.7	53.4	82.7	24.3
ANN	81.6	14.2	97.7	59.6	82.6	23.0
AE	81.8	12.7	98.3	64.1	82.5	21.2
WAE	81.8	14.7	97.8	61.7	82.7	23.8
Meta	81.3	14.7	97.2	55.8	82.6	23.3

3.4. Feature importance

When analyzing medical data, it is critical to understand how each feature impacts the model. RF, AdaBoost, GB, ERT, and XGBoos present feature importance, a value from zero to one corresponding to the significance of each feature. Fig. 4 shows the average feature importance from the five classifiers. In this context, AC, TG, Diabetes, and HDL show a considerable impact, in which AC is by far the most relevant feature. Age was found to have little relevance. These findings enable a better understanding of the more relevant causes associated with TDS.

4. Discussion

In recent years, the employment of machine learning algorithms in the medical field has increased substantially [119–121]. We compared ten ML classifiers (RF, ERT, ANN, SVM, RF, LR, NB, XGBoost, AdaBoost, and k-NN), three ensemble methods (Avg, wAvg, and meta-classifier), eight sampling strategies (RUS, ROS, RENN, SMOTE, RUS + ROS, RUS + SMOTE, RENN + ROS, and RENN + SMOTE), and two calibration techniques (prior correction and isotonic regression) in order to find the best approach to help patients perform expensive TT and FT tests.

The results of the performance measures considering different combinations of sampling and ML algorithms showed similar results, demonstrating that advances in ML paired with optimization algorithms narrows the differences among ML algorithms. Ferri et al. [122] compared metrics to evaluate classifiers and concluded that one ML method being better than another by one metric may not be comparable and extensible to the other metrics, even within the same family

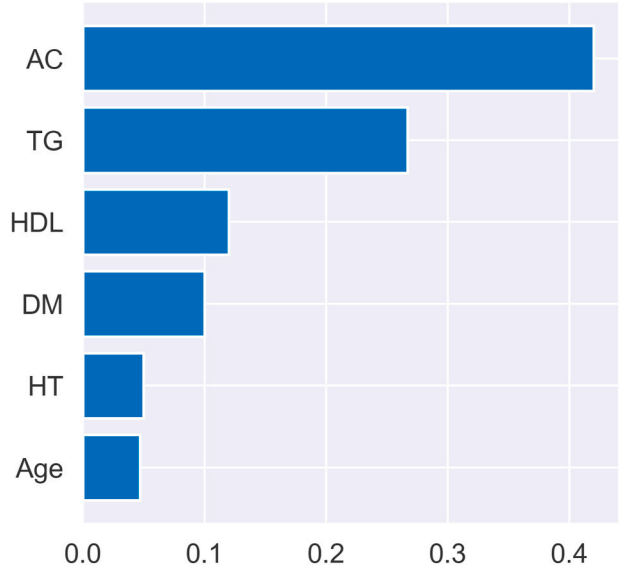

Fig. 4. Average Feature importance from five single ensemble classifiers: Random Forest, AdaBoost, Gradient Boosting, Extremely Randomized Trees, and XGBoost. Where the order of the features was: Abdominal Circumference (AC), Triglycerides (TG), High-density lipoprotein (HDL), Diabetes Mellitus (DM), Hypertension (HT), and Age, respectively.

(ranking, probabilistic, threshold). Therefore, good ranking measures do not guarantee good threshold measures. Usually, the AUC measures (ranking) are preferable in the analyses with imbalanced data [122,123]. Among the AUC metrics, the PR AUC is preferable in case of low-prevalence diseases, since ROC due to overly optimistic ROC AUC scores in skewed data [116,117]. However, medical studies use ROC AUC more frequently. Ozzene et al. [116] made a comparative analysis of different imbalance ratios of disease and non-disease patients, varying from a prevalence of 0.0099 up to 0.5. The authors analyzed the correlation between PR AUC and ROC AUC scores, in which by increasing the imbalance data, the correlation decreased. At an imbalanced ratio of 0.17 (closest to ours), the research presented high correlations (0.93). Since we optimized the values according to the PR AUC scores, ROC AUC scores may not present the most optimal values. Many challenges involving Artificial Intelligence, such as computer vision tasks, consider ranking measures (AP) as the primary metrics, and the differences are tight since classifiers present high scores. For example, increasing 1% in the COCO dataset's average precision score would be considered a great novelty [124,125]. This characteristic may be a tendency in other areas, where the state-of-the-art methods are slightly better than previous classifiers.

In our study, the three ensemble methods tested are among the first four in PR AUC scores. The wAvg classifier was the best classifier (45.4 PR AUC), followed by XGBoost (44.7), Meta (44.2), and Avg (44.2). he ensemble classifier presented the best results, which shows that using many base classifiers increases performance. XGBoost was the best single classifier, presenting the highest impact in constructing the ensemble classifier. In contrast, k-NN displayed the worst PR AUC metrics (a 6.4% difference in performance when compared to XGBoost).

Our research also showed that combining base classifiers increases accuracy. Analyzing the ROC curve, the results show a higher variance within the best ML classifiers. Nevertheless, among the different sampling strategies the ensemble classifiers present the best results, which reinforces conclusions from the PR AUC score, highlighting the importance of the ensemble classifiers.

A prevalent problem in medical data is imbalanced datasets. The lack of balanced data may alter the forecast due to the majority class's bias, which can cause the loss of the intended event. Resampling is a popular strategy for imbalanced data, being versatile and independent from the classification stage [126]. Christodoulou et al. [75] made a systematic review of ML applied in medicine, showing that few studies evaluated calibration, which is essential for risk assessment prediction. Also, the authors stated that adjusting class imbalance (sampling techniques) yields inadequate predictions. We noticed this behavior in our research, but this can be adjusted using prior probability correction, shown by better Brier Score values before and after calibration. Moreover, calibration is fundamental to adequately evaluate threshold metrics and produce better ensemble classifiers. Our data has an IR of approximately 4, which represents a slight imbalance. No sampling presented better results than all eight sampling strategies. Among the sampling strategies, oversampling strategies (ROS and SMOTE) had slightly better results. Loyola-Gonzalez et al. [127] established a guide to the best sampling approach according to the IR range, where subsampling (NCL) is suitable for the 1.820–5.3 range and hybrid sampling (SMOTE-ENN), in the 5.300–9.175 range. Our results did not show this behavior, possibly due to a small number of features (six) and an exhaustive grid search to all classifiers, resulting in similar results. Although some research establishes some indication to start the training, experimentation within methods and classifiers is always necessary. No sampling and SMOTE had the best top score (45.3), both using the wAvg classifier. Nevertheless, no sampling had a better overall score among most classifiers.

The threshold metrics did not show conclusive results regarding the best classifiers. Each classifier presented different best threshold cutoff points. The F-score for the best threshold values did not vary a lot, ranging from 43.0 to 50.9. The k-NN algorithm was an outlier, being nearly 4% worse than the second-worst classifier (NB). The wAvg classifier had the best F-Score for a single threshold value (50.9%). Furthermore, all ensemble classifiers had results greater than 50%. This proximity within the values makes it very difficult to compare threshold metrics, mainly because we need an exact cutoff point (which may not be optimal for all classifiers). Our research is also in hand with Ferri's research findings [122], in which the best ranking metrics do not always guarantee the best threshold metrics and vice-versa. For this reason, we believe ranking metrics are more suitable for comparing algorithms since they tend to present better results among a broader range of thresholds. Moreover, statistical comparisons within threshold metrics may be misleading or inconclusive. Other researchers advocate that ML findings should be based upon their predictive values, aiming to optimize predictive accuracy, which may not always correspond to statistical differences [128].

In our study, the McNemar's paired test concluded that all classifiers are statistically equal at a 1% significance level. Other studies point to similar results using different ML algorithms. Yadav et al. [129] performed a significant comparative analysis of nine ensemble classifiers in the medical disease diagnosis combining the prediction of eleven single classifiers for ten medical datasets. Yadav's study concluded that the ensemble model was not generally more effective than the best single classifier, demonstrating that the best base classifier outperformed the ensemble classifier in five of the ten datasets and was equally accurate in two other datasets. Christodoulou et al. [75] compared ML performance for the clinical prediction modeling of 71 articles (selected from 927) and found no evidence that ML has a better performance than LR.

Specifically for TDS, Lu et al. [44] were pioneers in employing ML to analyze late-onset hypogonadism in China. The authors used a dataset comprised of 772 patients with 16 features and applied four ML algorithms (Decision Tree, AdaBoost + Decision Tree, LR, and AdaBoost + LR), achieving 85% accuracy, 86% sensitivity, and 84% specificity. To overcome class imbalance, the authors implemented random resampling in the entire dataset, which lead to a considerable boost in the accuracy metric. In our study, we applied sampling methods only in the training data, to avoid biases. Nevertheless, applying the same methodology as Lu et al. (i.e., sampling on the entire dataset) we obtained nearly the same values (85.71% accuracy, 85.29% sensitivity, and 86.13% specificity) with less features (six) using the XGBoost classifier.

In metabolic syndrome, Karimi-Alavijeh et al. [130] employed a predictive model using decision tree and SVM algorithms. Their analysis used features also present in our research (i.e., age, HDL, WC, and hypertriglyceridemia), and SMOTE to overcome class imbalance. As with our findings, the authors verified a strong correlation between TG and body max index (BMI), as well as between metabolic syndrome and testosterone deficiency. When analyzing TDS, our study suggests that AC is the best predictor for hypogonadism, in line with findings by Yassin et al. [131].

5. Conclusion

Testosterone Deficit Syndrome (TDS) significantly impairs men's quality of life, but its timely and accurate diagnosis poses a challenge in areas with poor access to public health services. TT and FT levels are not measured during routine inspections and their tests are often expensive. This leads to an effort to obtain cheaper predictive methods to filter and guide patients in undergoing further, more specific, and high-priced exams. Secondary hypogonadism occurs in conjunction with other disorders, which facilitates its detection. But despite medical evidence connecting TDS with several factors (congenital or otherwise), there is a noticeable lack of research comparing and combining different ML techniques applied to either prediction or detection of TDS. This paper presented a broad comparison of ten well-known ML classifiers, eight sampling methods for imbalanced datasets, and the calibration of the predicted probabilities to identify a proper method to improve accuracy in TDS diagnosis using only features obtained from low-cost routine exams.

Overall, we can consider some critical points in the data processing. For data with a slight imbalance, tests should be performed with and without sampling techniques. A fair comparison between classifiers should consider extensive grid search optimization and the reliability of risk predictions. The use of these techniques significantly improves ML performance, which tends to become more similar. The classifier performance should consider an independent dataset to avoid overestimated measures.

Advancements in machine learning techniques have narrowed the differences between the methods, which tends to consider ranking metrics instead of threshold metrics. Results show that the wAvg classifier improves predictive performance with a PR AUC 2% higher than XGBoost (best base classifier). In some scenarios, the XGBoost classifier may be advantageous because of its low computational cost when compared to an ensemble classifier. Apart from k-NN and Naïve Bayes, all base classifiers presented satisfactory results.

We also analyzed metrics obtained from the confusion matrix at 0.23 and 0.5 threshold cutoff points, which demonstrated similarities in values without an evident prevalent method. McNemar's statistical significance test showed no difference between the methods. When analyzing feature importance, the results obtained with ML algorithms converge with the scientific literature, highlighting the importance of obesity, diabetes, TG, and diverge in the monitoring of age, showing little correlation and impact on the ML models. The applicability of TDS predictions through AI is very important, especially in developing countries, due to the costs of diagnostic tests.

Ethical statement

The authors declare that research used all ethical practices for its development, writing, and publication. The Research Ethics Committee of the State University of Feira de Santana, Bahia, Brazil, approved the study with the ethical approval code: 3,057,301.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors are grateful for financial support from the CNPq fellowship (Osmar Abílio de Carvalho Júnior) and from FAPESB doctoral grants (Monique Tonani Novaes and Caroline Santos Silva). This study was financed in part by the Coordination for the Improvement of Higher Education Personnel (CAPES) – Finance Code 001. Finally, the authors acknowledge the contribution of anonymous reviewers.

References

[1] Kanakis GA, Tsametis CP, Goulis DG. Measuring testosterone in women and men. Maturitas 2019;125:41–4. https://doi.org/10.1016/j.maturitas.2019.04.203.
[2] Shabsigh R, Kaufman JM, Steidle C, Padma-Nathan H. Randomized study of testosterone gel as adjunctive therapy to sildenafil in hypogonadal men with erectile dysfunction who do not respond to sildenafil alone. J Urol 2004;172:658–63. https://doi.org/10.1097/01.ju.0000132389.97804.d7.
[3] Laouali N, Brailly-Tabard S, Helmer C, Ancelin M-L, Tzourio C, Singh-Manoux A, Dugravot A, Elbaz A, Guiochon-Mantel A, Canonico M. Testosterone and all-cause mortality in older men: the role of metabolic syndrome. J. Endocr. Soc. 2018;2:322–35. https://doi.org/10.1210/js.2018-00005.
[4] Snyder PJ, Ellenberg SS, Cunningham GR, Matsumoto AM, Bhasin S, Barrett-Connor E, Gill TM, Farrar JT, Cella D, Rosen RC, Resnick SM, Swerdloff RS, Cauley JA, Cifelli D, Fluharty L, Pahor M, Ensrud KE, Lewis CE, Molitch ME, Crandall JP, Wang C, Budoff MJ, Wenger NK, Mohler ER, Bild DE, Cook NL, Keaveny TM, Kopperdahl DL, Lee D, Schwartz AV, Storer TW, Ershler WB, Roy CN, Raffel LJ, Romashkan S, Hadley E. The Testosterone Trials: seven coordinated trials of testosterone treatment in elderly men. Clin. Trials J. Soc. Clin. Trials. 2014;11:362–75. https://doi.org/10.1177/1740774514524032.
[5] Traish AM, Guay A, Feeley R, Saad F. The dark side of testosterone deficiency: I. metabolic syndrome and erectile dysfunction. J Androl 2009;30:10–22. https://doi.org/10.2164/jandrol.108.005215.
[6] Corona G, Rastrelli G, Vignozzi L, Mannucci E, Maggi M. Testosterone, cardiovascular disease and the metabolic syndrome. Best Pract Res Clin Endocrinol Metabol 2011;25:337–53. https://doi.org/10.1016/j.beem.2010.07.002.
[7] Elisabeth Hak A, Witteman JCM, De Jong FH, Geerlings MI, Hofman A, Pols HAP. Low levels of endogenous androgens increase the risk of atherosclerosis in elderly men: the Rotterdam Study. J Clin Endocrinol Metab 2002;87:3632–9. https://doi.org/10.1210/jc.87.8.3632.
[8] Zarrouf FA, Artz S, Griffith J, Sirbu C, Kommor M. Testosterone and depression: systematic review and meta-analysis. J Psychiatr Pract 2009;15:289–305. https://doi.org/10.1097/01.pra.0000358315.88931.fc.
[9] Aversa A, Morgentaler A. The practical management of testosterone deficiency in men. Nat Rev Urol 2015;12:641–50. https://doi.org/10.1038/nrurol.2015.238.
[10] Dandona P, Rosenberg MT. A practical guide to male hypogonadism in the primary care setting. Int J Clin Pract 2010;64:682–96. https://doi.org/10.1111/j.1742-1241.2010.02355.x.
[11] Basaria S. Male hypogonadism. Lancet 2014;383:1250–63. https://doi.org/10.1016/S0140-6736(13)61126-5.
[12] Bhasin S, Brito JP, Cunningham GR, Hayes FJ, Hodis HN, Matsumoto AM, Snyder PJ, Swerdloff RS, Wu FC, Yialamas MA. Testosterone therapy in men with hypogonadism: an endocrine society. J Clin Endocrinol Metab 2018;103:1715–44. https://doi.org/10.1210/jc.2018-00229.
[13] Kelly DM, Jones TH. Testosterone and obesity. Obes Rev 2015;16:581–606. https://doi.org/10.1111/obr.12282.
[14] Carrageta DF, Oliveira PF, Alves MG, Monteiro MP. Obesity and male hypogonadism: tales of a vicious cycle. Obes Rev 2019;20:1148–58. https://doi.org/10.1111/obr.12863.
[15] Pizzol D, Smith L, Fontana L, Caruso MG, Bertoldo A, Demurtas J, et al. Associations between body mass index, waist circumference and erectile dysfunction: a systematic review and META-analysis. Rev Endocr Metab Disord 2020;21:657–66. https://doi.org/10.1007/s11154-020-09541-0.
[16] Lamm S, Chidakel A, Bansal R. Obesity and hypogonadism. Urol Clin 2016;43:239–45. https://doi.org/10.1016/j.ucl.2016.01.005.
[17] Salas-Huetos A, Maghsoumi-Norouzabad L, James ER, Carrell DT, Aston KI, Jenkins TG, Becerra-Tomás N, Javid AZ, Abed R, Torres PJ, Luque EM, Ramírez ND, Martini AC, Salas-Salvadó J. Male adiposity, sperm parameters and reproductive hormones: an updated systematic review and collaborative meta-analysis. Obes Rev 2020:1–33. https://doi.org/10.1111/obr.13082.
[18] Saad F, Aversa A, Isidori AM, Gooren LJ. Testosterone as potential effective therapy in treatment of obesity in men with testosterone deficiency: a review. Curr Diabetes Rev 2012;8:131–43. https://doi.org/10.2174/157339912799424573.
[19] Giagulli VA, Castellana M, Murro I, Pelusi C, Guastamacchia E, Triggiani V, De Pergola G. The role of diet and weight loss in improving secondary hypogonadism in men with obesity with or without type 2 diabetes mellitus. Nutrients 2019;11. https://doi.org/10.3390/nu11122975.
[20] Allan CA, McLachlan RI. Androgens and obesity. Curr Opin Endocrinol Diabetes Obes 2010;17:224–32. https://doi.org/10.1097/MED.0b013e3283398ee2.

[21] Corona G, Rastrelli G, Monami M, Saad F, Luconi M, Lucchese M, Facchiano E, Sforza A, Forti G, Mannucci E, Maggi M. Body weight loss reverts obesity-associated hypogonadotropic hypogonadism: a systematic review and meta-analysis. Eur J Endocrinol 2013;168:829–43. https://doi.org/10.1530/EJE-12-0955.

[22] Zitzmann M, Nieschlag E. Hormone substitution in male hypogonadism. Mol Cell Endocrinol 2000;161:73–88. https://doi.org/10.1016/S0303-7207(99)00227-0.

[23] Elliott J, Kelly SE, Millar AC, Peterson J, Chen L, Johnston A, Kotb A, Skidmore B, Bai Z, Mamdani M, Wells GA. Testosterone therapy in hypogonadal men: a systematic review and network meta-analysis. BMJ Open 2017;7:1–10. https://doi.org/10.1136/bmjopen-2016-015284.

[24] Ponce OJ, Spencer-Bonilla G, Alvarez-Villalobos N, Serrano V, Singh-Ospina N, Rodriguez-Gutierrez R, Salcido-Montenegro A, Benkhadra R, Prokop LJ, Bhasin S, Brito JP. The efficacy and adverse events of testosterone replacement therapy in hypogonadal men: a systematic review and meta-analysis of randomized, placebo-controlled trials. J Clin Endocrinol Metab 2018;103:1745–54. https://doi.org/10.1210/jc.2018-00404.

[25] Steidle C, Schwartz S, Jacoby K, Sebree T, Smith T, Bachand R. AA2500 testosterone gel normalizes androgen levels in aging males with improvements in body composition and sexual function. J Clin Endocrinol Metab 2003;88:2673–81. https://doi.org/10.1210/jc.2002-021058.

[26] Saad F, Haider A, Doros G, Traish A. Long-term treatment of hypogonadal men with testosterone produces substantial and sustained weight loss. Obesity 2013;21:1975–81. https://doi.org/10.1002/oby.20407.

[27] Yassin AA, Doros G. Testosterone therapy in hypogonadal men results in sustained and clinically meaningful weight loss. Clin. Obes. 2013;3:73–83. https://doi.org/10.1111/cob.12022.

[28] Corona G, Giagulli VA, Maseroli E, Vignozzi L, Aversa A, Zitzmann M, Saad F, Mannucci E, Maggi M. Testosterone supplementation and body composition: results from a meta-analysis of observational studies. J Endocrinol Invest 2016;39:967–81. https://doi.org/10.1007/s40618-016-0480-2.

[29] Chung TH, Kwon YJ, Lee YJ. High triglyceride to HDL cholesterol ratio is associated with low testosterone and sex hormone-binding globulin levels in Middle-aged and elderly men. Aging Male 2020;23:93–7. https://doi.org/10.1080/13685538.2018.1501015.

[30] Agledahl I, Skjærpe PA, Hansen JB, Svartberg J. Low serum testosterone in men is inversely associated with non-fasting serum triglycerides: the Tromsø study. Nutr Metabol Cardiovasc Dis 2008;18:256–62. https://doi.org/10.1016/j.numecd.2007.01.014.

[31] Jiang Y, Ye J, Zhao M, Tan A, Zhang H, Gao Y, Lu Z, Wu C, Hu Y, Wang Q, Yang X, Mo Z. Cross-sectional and longitudinal associations between serum testosterone concentrations and hypertension: results from the fangchenggang area male health and examination survey in China. Clin Chim Acta 2018;487:90–5. https://doi.org/10.1016/j.cca.2018.08.027.

[32] Torkler S, Wallaschofski H, Baumeister SE, Völzke H, Dörr M, Felix S, Rettig R, Nauck M, Haring R. Inverse association between total testosterone concentrations, incident hypertension and blood pressure. Aging Male 2011;14:176–82. https://doi.org/10.3109/13685538.2010.529194.

[33] Yang Q, Li Z, Li W, Lu L, Wu H, Zhuang Y, Wu K, Sui X. Association of total testosterone, free testosterone, bioavailable testosterone, sex hormone-binding globulin, and hypertension. Medicine (Baltim) 2019;98:e15628. https://doi.org/10.1097/MD.0000000000015628.

[34] Hackett G. Type 2 diabetes and testosterone therapy. World J. Men?S Heal. 2019;37:31–44. https://doi.org/10.5534/wjmh.180027.

[35] Yao QM, Wang B, An XF, Zhang JA, Ding L. Testosterone level and risk of type 2 diabetes in men: a systematic review and meta-analysis. Endocr. Connect. 2018;7:220–31. https://doi.org/10.1530/EC-17-0253.

[36] Yassin A, Haider A, Haider KS, Caliber M, Doros G, Saad F, Timothy Garvey W. Testosterone therapy in men with hypogonadism prevents progression from prediabetes to type 2 diabetes: eight-year data from a registry study. Diabetes Care 2019;42:1104–11. https://doi.org/10.2337/dc18-2388.

[37] Haider KS, Haider A, Saad F, Doros G, Hanefeld M, Dhindsa S, Dandona P, Traish A. Remission of type 2 diabetes following long-term treatment with injectable testosterone undecanoate in patients with hypogonadism and type 2 diabetes: 11-year data from a real-world registry study. Diabetes Obes Metabol 2020:14122. https://doi.org/10.1111/dom.14122.

[38] Gianatti EJ, Grossmann M. Testosterone deficiency in men with Type 2 diabetes: pathophysiology and treatment. Diabet Med 2020;37:174–86. https://doi.org/10.1111/dme.13977.

[39] Corona G, Monami M, Rastrelli G, Aversa A, Tishova Y, Saad F, Lenzi A, Forti G, Mannucci E, Maggi M. Testosterone and metabolic syndrome: a meta-analysis study. J Sex Med 2011;8:272–83. https://doi.org/10.1111/j.1743-6109.2010.01991.x.

[40] Bianchi VE, Locatelli V. Testosterone a key factor in gender related metabolic syndrome. Obes Rev 2018;19:557–75. https://doi.org/10.1111/obr.12633.

[41] Muraleedharan V, Jones TH. Review: testosterone and the metabolic syndrome. Ther. Adv. Endocrinol. Metab. 2010;1:207–23. https://doi.org/10.1177/2042018810390258.

[42] Anaissie J, Roberts NH, Wang P, Yafi FA. Testosterone replacement therapy and components of the metabolic syndrome. Sex. Med. Rev. 2017;5:200–10. https://doi.org/10.1016/j.sxmr.2017.01.003.

[43] Rosner W, Auchus RJ, Azziz R, Sluss PM, Raff H. Position statement: utility, limitations, and pitfalls in measuring testosterone: an endocrine society position statement. J Clin Endocrinol Metab 2007;92:405–13. https://doi.org/10.1210/jc.2006-1864.

[44] Lu T, Hu YH, Tsai CF, Liu SP, Chen PL. Applying machine learning techniques to the identification of late-onset hypogonadism in elderly men. SpringerPlus 2016;5. https://doi.org/10.1186/s40064-016-2531-8.

[45] Tharakan T, Bettocchi C, Carvalho J, Corona G, Joensen UN, Jones H, Kadioglu A, Martínez Salamanca JI, Serefoglu EC, Verze P, Darraugh J, Plass K, N'Dow J, Salonia A, Minhas S. Male sexual and reproductive health—does the urologist have a role in addressing gender inequality in life expectancy? Eur. Urol. Focus. 2020;6:791–800. https://doi.org/10.1016/j.euf.2019.10.009.

[46] Eisenberg ML, Li S, Behr B, Pera RR, Cullen MR. Relationship between semen production and medical comorbidity. Fertil Steril 2015;103:66–71. https://doi.org/10.1016/j.fertnstert.2014.10.017.

[47] Ventimiglia E, Capogrosso P, Boeri L, Serino A, Colicchia M, Ippolito S, Scano R, Papaleo E, Damiano R, Montorsi F, Salonia A. Infertility as a proxy of general male health: results of a cross-sectional survey. Fertil Steril 2015;104:48–55. https://doi.org/10.1016/j.fertnstert.2015.04.020.

[48] Araujo AB, Dixon JM, Suarez EA, Murad MH, Guey LT, Wittert GA. Endogenous testosterone and mortality in men: a systematic review and meta-analysis. J Clin Endocrinol Metab 2011;96:3007–19. https://doi.org/10.1210/jc.2011-1137.

[49] Daka B, Langer RD, Larsson CA, Rosén T, Jansson PA, Råstam L, Lindblad U. Low concentrations of serum testosterone predict acute myocardial infarction in men with type 2 diabetes mellitus. BMC Endocr Disord 2015;15:1–7. https://doi.org/10.1186/s12902-015-0034-1.

[50] Haring R, Völzke H, Steveling A, Krebs A, Felix SB, Schöfl C, Dörr M, Nauck M, Wallaschofski H. Low serum testosterone levels are associated with increased risk of mortality in a population-based cohort of men aged 20-79. Eur Heart J 2010;31:1494–501. https://doi.org/10.1093/eurheartj/ehq009.

[51] Muraleedharan V, Marsh H, Kapoor D, Channer KS, Jones TH. Testosterone deficiency is associated with increased risk of mortality and testosterone replacement improves survival in Men with type 2 diabetes. Eur J Endocrinol 2013;169:725–33. https://doi.org/10.1530/EJE-13-0321.

[52] Ruige JB, Mahmoud AM, De Bacquer D, Kaufman JM. Endogenous testosterone and cardiovascular disease in healthy men: a meta-analysis. Heart 2011;97:870–5. https://doi.org/10.1136/hrt.2010.210757.

[53] Oskui PM, French WJ, Herring MJ, Mayeda GS, Burstein S, Kloner RA. Testosterone and the cardiovascular system: a comprehensive review of the clinical literature. J. Am. Heart Assoc. 2013;2:1–22. https://doi.org/10.1161/JAHA.113.000272.

[54] Comhaire F. Hormone replacement therapy and longevity. Andrologia 2016;48:65–8. https://doi.org/10.1111/and.12419.

[55] Galdas PM, Cheater F, Marshall P. Men and health help-seeking behaviour: literature review. J Adv Nurs 2005;49:616–23. https://doi.org/10.1111/j.1365-2648.2004.03331.x.

[56] Peters DH, Garg A, Bloom G, Walker DG, Brieger WR, Hafizur Rahman M. Poverty and access to health care in developing countries. Ann N Y Acad Sci 2008;1136:161–71. https://doi.org/10.1196/annals.1425.011.

[57] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 2001;23:89–109. https://doi.org/10.1016/S0933-3657(01)00077-X.

[58] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. Circulation 2015;131:211–9. https://doi.org/10.1161/CIRCULATIONAHA.114.014508.

[59] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PloS One 2017;12:e0174944. https://doi.org/10.1371/journal.pone.0174944.

[60] Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol 2017;69:2657–64. https://doi.org/10.1016/j.jacc.2017.03.571.

[61] Masetic Z, Subasi A. Congestive heart failure detection using random forest classifier. Comput Methods Progr Biomed 2016;130:54–64. https://doi.org/10.1016/j.cmpb.2016.03.020.

[62] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet 2018;9:1–10. https://doi.org/10.3389/fgene.2018.00515.

[63] Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 2017;15:104–16. https://doi.org/10.1016/j.csbj.2016.12.005.

[64] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Canc Inf 2006;2:59–77. https://doi.org/10.1177/117693510600200030.

[65] Aličković E, Subasi A. Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Comput Appl 2017;28:753–63. https://doi.org/10.1007/s00521-015-2103-9.

[66] Vijayarani S, Dhayanand S, Professor A, Research Scholar MP. Kidney disease prediction using svm and ann algorithms. Int. J. Comput. Bus. Res. ISSN (Online. 2015;6:2229–6166.

[67] Worachartcheewan A, Shoombuatong W, Pidetcha P, Nopnithipat W, Prachayasittikul V, Nantasenamat C. Predicting metabolic syndrome using the random forest method. Sci World J 2015;2015:1–10. https://doi.org/10.1155/2015/581501.

[68] Hsieh C-H, Lu R-H, Lee N-H, Chiu W-T, Hsu M-H, C Y, (Jack) Li. Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks. Surgery 2011;149:87–93. https://doi.org/10.1016/j.surg.2010.03.023.

[69] Suh J, Yoo S, Park J, Cho SY, Cho MC, Son H, et al. Development and validation of explainable AI-based decision-supporting tool for prostate biopsy. BJU Int 2020; 126(6):694–703. https://doi.org/10.1111/bju.15122.

[70] Snow O, Lallous N, Ester M, Cherkasov A. Deep learning modeling of androgen receptor responses to prostate cancer therapies. Int J Mol Sci 2020;21:5847. https://doi.org/10.3390/ijms21165847.

[71] Deng K, Li H, Guan Y. Treatment stratification of patients with metastatic castration-resistant prostate cancer by machine learning. IScience 2020;23: 100804. https://doi.org/10.1016/j.isci.2019.100804.

[72] Albini A, Bruno A, Bassani B, D'Ambrosio G, Pelosi G, Consonni P, Castellani L, Conti M, Cristoni S, Noonan DM. Serum steroid ratio profiles in prostate cancer: a new diagnostic tool toward a personalized medicine approach. Front Endocrinol (Lausanne) 2018;9. https://doi.org/10.3389/fendo.2018.00110.

[73] Schulz MA, Yeo BTT, Vogelstein JT, Mourao-Miranada J, Kather JN, Kording K, Richards B, Bzdok D. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. Nat Commun 2020; 11. https://doi.org/10.1038/s41467-020-18037-z.

[74] Cabitza F, Locoro A, Banfi G. Machine learning in orthopedics: a literature review. Front. Bioeng. Biotechnol. 2018;6. https://doi.org/10.3389/fbioe.2018.00075.

[75] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110: 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004.

[76] Rokach L. Ensemble-based classifiers. Artif Intell Rev 2010;33:1–39. https://doi.org/10.1007/s10462-009-9124-7.

[77] Ting KM, Witten IH. Issues in stacked generalization. J Artif Intell Res 1999;10: 271–89. https://doi.org/10.1613/jair.594.

[78] Sagi O, Rokach L. Ensemble learning: a survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2018;8:1–18. https://doi.org/10.1002/widm.1249.

[79] Chen C, Zhang Q, Yu B, Yu Z, Lawrence PJ, Ma Q, Zhang Y. Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. Comput Biol Med 2020;123:103899. https://doi.org/10.1016/j.compbiomed.2020.103899.

[80] Mishra A, Pokhrel P, Hoque MT. StackDPPred: a stacking based prediction of DNA-binding protein from sequence. Bioinformatics 2019;35:433–41. https://doi.org/10.1093/bioinformatics/bty653.

[81] Xiong Y, Wang Q, Yang J, Zhu X, Wei DQ. PredT4SE-stack: prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. Front Microbiol 2018;9:1–9. https://doi.org/10.3389/fmicb.2018.02571.

[82] Abdar M, Zomorodi-Moghadam M, Zhou X, Gururajan R, Tao X, Barua PD, Gururajan R. A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recogn Lett 2020;132:123–31. https://doi.org/10.1016/j.patrec.2018.11.004.

[83] El-Sappagh S, Elmogy M, Ali F, Abuhmed T, Islam SMR, Kwak KS. A comprehensive medical decision–support framework based on a heterogeneous ensemble classifier for diabetes prediction. Electron 2019;8. https://doi.org/10.3390/electronics8060635.

[84] Tama BA, Rhee KH. Tree-based classifier ensembles for early detection method of diabetes: an exploratory study. Artif Intell Rev 2019;51:355–70. https://doi.org/10.1007/s10462-017-9565-3.

[85] El-Rashidy N, El-Sappagh S, Abuhmed T, Abdelrazek S, El-Bakry HM. Intensive care unit mortality prediction: an improved patient-specific stacking ensemble model. IEEE Access 2020;8:133541–64. https://doi.org/10.1109/ACCESS.2020.3010556.

[86] Idri A, Benhar H, Fernández-Alemán JL, Kadi I. A systematic map of medical data preprocessing in knowledge discovery. Comput Methods Progr Biomed 2018;162: 69–85. https://doi.org/10.1016/j.cmpb.2018.05.007.

[87] Han W, Huang Z, Li S, Jia Y. Distribution-sensitive unbalanced data oversampling method for medical diagnosis. J Med Syst 2019;43. https://doi.org/10.1007/s10916-018-1154-8.

[88] Jain A, Ratnoo S, Kumar D. Addressing class imbalance problem in medical diagnosis: a genetic algorithm approach. In: Int. Conf. Information, commun. Instrum. Control. Indore, India: IEEE; 2017. p. 1–8. https://doi.org/10.1109/ICOMICON.2017.8279150. 2017.

[89] Zhu M, Xia J, Jin X, Yan M, Cai G, Yan J, Ning G. Class weights random forest algorithm for processing class imbalanced medical data. IEEE Access 2018;6: 4641–52. https://doi.org/10.1109/ACCESS.2018.2789428.

[90] Mulhall JP, Trost LW, Brannigan RE, Kurtz EG, Redmon JB, Chiles KA, Lightner DJ, Miner MM, Murad MH, Nelson CJ, Platz EA, Ramanathan LV, Lewis RW. Evaluation and management of testosterone deficiency: AUA guideline. J Urol 2018;200:423–32. https://doi.org/10.1016/j.juro.2018.03.115.

[91] Basheer I, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. J Microbiol Methods 2000;43:3–31. https://doi.org/10.1016/S0167-7012(00)00201-3.

[92] Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Network 1999;10:988–99. https://doi.org/10.1109/72.788640.

[93] Breiman L. Random forests. Mach Learn 2001;45:5–32.

[94] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63: 3–42. https://doi.org/10.1007/s10994-006-6226-1.

[95] Rätsch G, Onoda T, Müller KR. Soft margins for AdaBoost. Mach Learn 2001;42: 287–320. https://doi.org/10.1023/A:1007618119488.

[96] Chen T, He T. Xgboost: extreme gradient boosting. R Lect; 2014. p. 1–84.

[97] Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002;38: 367–78. https://doi.org/10.1016/S0167-9473(01)00065-2.

[98] Fix E, Hodges J. Performance, Discriminatory analysis. Nonparametric discrimination: small sample. Texas: Randolph Field; 1952.

[99] Lowd D, Domingos P. Naive Bayes models for probability estimation. ICML 2005 - Proc. 22nd Int. Conf. Mach. Learn 2005:529–36. https://doi.org/10.1145/1102351.1102418.

[100] Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol 2001;54:979–85. https://doi.org/10.1016/S0895-4356(01)00372-9.

[101] He Haibo, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009;21:1263–84. https://doi.org/10.1109/TKDE.2008.239.

[102] Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. Neural Network 2008;21:427–36. https://doi.org/10.1016/j.neunet.2007.12.031.

[103] Awoyemi JO, Adetunmbi AO, Oluwadare SA. Credit card fraud detection using machine learning techniques: a comparative analysis. In: Proc. IEEE int. Conf. Comput. Netw. Informatics. ICCNI; 2017. p. 1–9. https://doi.org/10.1109/ICCNI.2017.8123782. 2017-Janua (2017).

[104] Chawla NV. Data mining for imbalanced datasets: an overview. In: Data min. Knowl. Discov. Handb. Boston, MA: Springer US; 2009. p. 875–86. https://doi.org/10.1007/978-0-387-09823-4_45.

[105] Shelke MS, Deshmukh PR, Shandilya PVK. A review on imbalanced data handling using undersampling and oversampling technique. Int. J. Recent Trends Eng. Res. 2017;3:444–9. https://doi.org/10.23883/IJRTER.2017.3168.0UWXM.

[106] Lemaître G, Nogueira F, West WS, Mv O, Aridas CK. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 2017;40:1–5.

[107] Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newsl. 2004; 6:20–9. https://doi.org/10.1145/1007730.1007735.

[108] Wilson DL. Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans. Syst. Man. Cybern. SMC- 1972;2:408–21. https://doi.org/10.1109/TSMC.1972.4309137.

[109] Han H, Wang W-Y, Mao B-H, Borderline-Smote. A new over-sampling method in imbalanced data sets learning. Adv. Intell. Syst. Comput. 2005:878–87. https://doi.org/10.1007/11538059_91.

[110] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57. https://doi.org/10.1613/jair.953.

[111] Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets. Cham: Springer International Publishing; 2018. https://doi.org/10.1007/978-3-319-98074-4.

[112] Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167–76. https://doi.org/10.1016/j.jclinepi.2015.12.005.

[113] Wallace BC, Dahabreh IJ. Class probability estimates are unreliable for imbalanced data (and how to fix them). In: IEEE 12th int. Conf. Data min. Brussels, Belgium: IEEE; 2012. p. 695–704. https://doi.org/10.1109/ICDM.2012.115. 2012.

[114] Saerens M, Latinne P, Decaestecker C. Adjusting the outputs of a classifier to new a Priori probabilities: a simple procedure. Neural Comput 2002;14:21–41. https://doi.org/10.1162/089976602753284446.

[115] Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In: Proc. Eighth ACM SIGKDD int. Conf. Knowl. Discov. Data min. - KDD '02. New York, USA: ACM Press; 2002. p. 694. https://doi.org/10.1145/775047.775151. New York.

[116] Ozenne B, Subtil F, Maucort-Boulch D. The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol 2015;68:855–9. https://doi.org/10.1016/j.jclinepi.2015.02.010.

[117] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proc. 23rd int. Conf. Mach. Learn. - ICML '06. New York, New York, USA: ACM Press; 2006. p. 233–40. https://doi.org/10.1145/1143844.1143874.

[118] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12:153–7. https://doi.org/10.1007/BF02295996.

[119] Awaysheh A, Wilcke J, Elvinger F, Rees L, Fan W, Zimmerman KL. Review of medical decision support and machine-learning methods. Vet Pathol 2019;56: 512–25. https://doi.org/10.1177/0300985819829524.

[120] Beam AL, Kohane IS. Big data and machine learning in health care. JAMA, J Am Med Assoc 2018;319:1317–8. https://doi.org/10.1001/jama.2017.18391.

[121] Chen JH, Asch SM. Machine learning and prediction in medicine — beyond the peak of inflated expectations. N Engl J Med 2017;376:2507–9. https://doi.org/10.1056/NEJMp1702071.

[122] Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. Pattern Recogn Lett 2009;30:27–38. https://doi.org/10.1016/j.patrec.2008.08.010.

[123] Rosset S. Model selection via the AUC. In: Twenty-first int. Conf. Mach. Learn. - ICML '04. New York, New York, USA: ACM Press; 2004. p. 89. https://doi.org/10.1145/1015330.1015400.

[124] Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: common objects in context. In: Fleet D, Tomas P, Schiele B, Tuytelaars T, editors. Comput. Vis. – ECCV 2014. Lect. Notes comput. Sci., vol. 8693. Zurich, Switzerland: Springer Cham; 2014. p. 740–55. https://doi.org/10.1007/978-3-319-10602-1_48.

[125] Huang Z, Huang L, Gong Y, Huang C, Wang X. Mask scoring R-CNN. Long Beach. In: IEEE/CVF conf. Comput. Vis. Pattern recognit. CA, USA, USA: IEEE; 2019. p. 6402–11. https://doi.org/10.1109/CVPR.2019.00657. 2019.
[126] Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl 2017;73:220–39. https://doi.org/10.1016/j.eswa.2016.12.035.
[127] Loyola-González O, Martínez-Trinidad JF, Carrasco-Ochoa JA, García-Borroto M. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. Neurocomputing 2016;175:935–47. https://doi.org/10.1016/j.neucom.2015.04.120.
[128] Breiman L. Statistical modeling: the two cultures. Stat Sci 2001;16:199–215. https://doi.org/10.1214/ss/1009213726.
[129] Yadav SS, Kadam VJ, Jadhav SM. Comparative analysis of ensemble classifier and single base classifier in medical disease diagnosis. In: Bansal J, Gupta M, Sharma H, Agarwal B, editors. Commun. Intell. Syst. ICCIS 2019. Lect. Notes Networks syst. Singapore: Springer; 2020. p. 475–89. https://doi.org/10.1007/978-981-15-3325-9_37.
[130] Karimi-Alavijeh F, Jalili S, Sadeghi M. Predicting metabolic syndrome using decision tree and support vector machine methods. ARYA Atheroscler 2016;12.
[131] Yassin AA, Nettleship JE, Salman M, Almehmadi Y. Waist circumference is superior to weight and BMI in predicting sexual symptoms, voiding symptoms and psychosomatic symptoms in men with hypogonadism and erectile dysfunction. Andrologia 2017;49:1–6. https://doi.org/10.1111/and.12634.