Data and Metadata. 2023; 2:65
doi: 10.56294/dm202365

ORIGINAL

Explainable machine learning for coronary artery disease risk assessment and prevention

Aprendizaje automático explicable para la evaluación y prevención del riesgo de arteriopatía coronaria

Louridi Nabaouia[1], Douzi Samira[2], El Ouahidi Bouabid[1]

[1]Faculty of Sciences, IPSS Laboratory. Mohammed V University. Rabat, Morocco.
[2]Faculty of Medicine and Pharmacy, IPSS Laboratory. Mohammed V University. Rabat, Morocco

Cite as: Nabaouia L, Samira D, Bouabid EO. Explainable machine learning for coronary artery disease risk assessment and prevention. Data and Metadata 2023; 2:65. https://doi.org/10.56294/dm202365.

Submitted: 01-08-2023 **Revised:** 23-10-2023 **Accepted:** 28-12-2023 **Published:** 29-12-2023

Editor: Javier Gonzalez-Argote

Guest Editor: Yousef Farhaoui

Note: Paper presented at the International Conference on Artificial Intelligence and Smart Environments (ICAISE'2023).

ABSTRACT

Coronary Artery Disease (CAD) is an increasingly prevalent ailment that has a significant impact on both longevity and quality of life. Lifestyle, genetics, nutrition, and stress are all significant contributors to rising mortality rates. CAD is preventable through early intervention and lifestyle changes. As a result, low-cost automated solutions are required to detect CAD early and help healthcare professionals treat chronic diseases efficiently. Machine learning applications in medicine have increased due to their ability to detect data patterns. Employing machine learning to classify the occurrence of coronary artery disease could assist doctors in reducing misinterpretation. The research project entails the creation of a coronary artery disease diagnosis system based on machine learning. Using patient medical records, we demonstrate how machine learning can help identify if an individual will acquire coronary artery disease. Furthermore, the study highlights the most critical risk factors for coronary artery disease. We used two machine learning approaches, Catboost and LightGBM classifiers, to predict the patient with coronary artery disease. We employed various data augmentation methods, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAE), to solve the imbalanced data problem. Optuna was applied to optimize hyperparameters. The proposed method was tested on the real-world dataset Z-Alizadeh Sani. The acquired findings were satisfactory, as the model could predict the likelihood of cardiovascular disease in a particular individual by combining Catboost with VAE, which demonstrated good accuracy compared to the other approaches. The proposed model is evaluated using a variety of metrics, including accuracy, recall, f-score, precision, and ROC curve. Furthermore, we used the SHAP values and Boruta Feature Selection (BFS) to determine essential risk factors for coronary artery disease.

Keywords: Coronary Artery Disease; Explainable Machine Learning; Risk Factors; Data Augmentation.

RESUMEN

La enfermedad arterial coronaria (EAC) es una dolencia cada vez más prevalente que tiene un impacto significativo tanto en la longevidad como en la calidad de vida. El estilo de vida, la genética, la nutrición y el estrés contribuyen en gran medida al aumento de las tasas de mortalidad. La EAC puede prevenirse con una intervención precoz y cambios en el estilo de vida. En consecuencia, se necesitan soluciones automatizadas de bajo coste para detectar precozmente la EAC y ayudar a los profesionales sanitarios a tratar eficazmente las enfermedades crónicas. métodos de aumento de datos, como las redes generativas adversariales (GAN) y los autocodificadores variacionales (VAE), para resolver el problema de los datos desequilibrados. Se aplicó Optuna para optimizar los hiperparámetros. El método propuesto se probó en el conjunto de datos del mundo real Z-Alizadeh Sani. Los resultados obtenidos fueron satisfactorios, ya que el modelo fue capaz de predecir la probabilidad de padecer una enfermedad cardiovascular en un individuo concreto combinando Catboost con VAE, lo que demostró una buena precisión en comparación con los otros enfoques. El modelo propuesto se evaluó utilizando diversas métricas, como la exactitud, la recuperación, la puntuación f, la precisión y la curva ROC. Además, utilizamos los valores SHAP para determinar importantes factores de riesgo de enfermedad coronaria.

Palabras clave: Enfermedad Arterial Coronaria; Aprendizaje Automático Explicable; Factores de Riesgo; Aumento de Datos.

INTRODUCTION

According to World Health Organization (WHO) statistics, heart disease is a significant concern for humans worldwide. According to the most recent WHO data published in 2020, 77,340 Moroccans perished from coronary heart disease, accounting for 33,89 % of all fatalities.(1)

When compared to other harmful conditions, including cancer, respiratory illness, tuberculosis, human immunodeficiency viruses (HIV), car accidents, and others, coronary artery disease is the most fatal. CAD is the heart malfunction induced by fatty buildup in blood vessels. The severity of cardiac disease and its increasing occurrence is a challenging threat nowadays. It is vital to establish an effective remedy for this possibly lethal circumstance.(2) Cardiovascular diseases (CVD) are divided into four categories.(3) The most frequent variant is CAD, which originates due to plaque development in coronary arteries; this leads to limited blood flow to the heart muscle, causing heart attacks, chest tightness, and shortness of breath.(4) The proposed initiative primarily focuses on the most prevalent form of cardiovascular disease, CAD. Machine learning (ML) algorithms endow academics with effective instruments. It uses statistical methodologies in immense datasets to detect connections between patient factors and outcomes and allows for objective integration of data to predict outcomes. ML has been applied in numerous medical-related disciplines, such as diagnosis, outcome prediction, treatment, and medical image interpretation,(5,6) and it has also been used to predict unfavorable outcomes in patients with HF by integrating clinical and other data in recent studies.(7,8,9) However, there is presently a lack of studies on ML for the prognosis of HF caused by CHD, notably medium and long-term mortality risk prediction.

Moreover, despite the optimistic performance of ML in past studies, evidence of its use in a real-world clinical environment and explainable risk prediction models to support disease prognosis is rare.(10,11) Because of the "black-box" nature of ML algorithms, it is difficult to articulate the reason for making specific patient predictions; that is, what particular qualities of the patient lead to a given prognosis. The lack of interpretability has so far limited the application of more powerful ML approaches in medical decision support,(12) and the lack of intuitional comprehension of ML models is one of the critical hurdles to the deployment of ML in the medical area.(13) To solve these difficulties, this research paired an advanced ML approach with a framework based on SHapley Additive exPlanations (SHAP).(14) It not only enhances the accuracy of predicting patients with CHD, it also provides intuitive explanations that help patients forecast risk, assisting physicians in better comprehending the decision-making process about disease severity and optimizing prospects for quick intervention. This is a significant stride forward for machine learning in medicine(15) and will aid in developing interpretable and tailored risk prediction algorithms. This research seeks to overcome the aforementioned limitations in the literature by addressing the following objectives:

- Preprocessing the data, which includes normalization for numerical features with the minmax algorithm, One hot encoding for binary variables, and mapping ordinal categorical variables.
- Address class imbalance and avoid overfitting in CAD prediction using GANs and VAEs to attain a balanced target class.
- Classifying the patients with cardiovascular disease using Catboost and LightGBM classifiers.
- Establish an optimal classifier with outstanding accuracy in predicting CAD at the earliest stage through employing Optuna hyperparameter tuning.
- Using SHAP Values and BFS to highlight the most essential cardiovascular risk factors.

The systematic organization of this work begins with a survey of related works in Section II. Section III thoroughly overviews the proposed work's methodology and architecture. Section IV details the proposed strategy. In Section V, the findings are reviewed and assessed. The publication closes with Section VI, which summarizes the study's key findings and implications for future research.

Related works

Artificial intelligence (AI) has been employed in different industries due to rapid information technology improvements. Several researchers have employed machine learning technologies to forecast and analyze diseases. Goldman et al.(16) developed a new artificial neural network (ANN) model for detecting CHD. They utilized the Framingham Heart Institute dataset to confirm the experiment. The results highlighted that the ANN model had a higher degree of specificity and sensitivity in predicting results than the Framingham Risk Score (FRS) (which is utilized to estimate an individual's risk of developing CHD over the ten years that follow, depending on cholesterol levels and noncholesterol factors). However, the region under the ROC curve (AUC) was smaller than that for FRS. To evaluate the classification performance of the machine learning models, Receiver operating characteristic (ROC) curves were used. The recommended ANN model gave much better outcomes than FRS for precision-recall measures.

In 2020, Du et al.(17) identified CHD among individuals with hypertension using electronic health record data by applying machine learning technologies. They partitioned the CHD dataset into a training set and a test set; then, they utilized a range of machine learning methods to train the model on the training set and the test set to evaluate the model's performance and compared the findings with the FRS score. The experimental findings indicated that the maximum AUC value (0,943) was attained utilizing the XGBoost for the test set. They compared different machine learning algorithms. The k-nearest neighbor approach had an AUC of 0,908, the random forest algorithm had an AUC of 0,938, and the logistic regression algorithm had an AUC of 0,865. They evaluated the relevant features and discovered that time-related features increased the model's performance.

To forecast CHD, Han et al.(18) studied the predictive strength of multiple machine learning algorithms to predict the probability of rapid progress of coronary atherosclerosis. The plaque features of 983 patients have been examined, combining qualitative and quantitative computed tomography angiography. The model's score was compared to the risk score for cardiovascular atherosclerosis. The clinical factors that had the most value were compared. The authors warn, however, that detecting hidden bias in the dataset using machine learning algorithms remains challenging. Joo et al.(19) evaluated the reliability of machine learning algorithms for expecting cardiovascular disease risks. The authors performed the longitudinal cohort study on 3,6 million patients requesting readmission to hospitals in England. The discrimination and calibration performance of the 19 prediction models were studied. For example, the random forest tree prediction score varied from 2,9 to 9,2 %, but the neural network prediction score varied from 2,4 to 7,2 %. It was advised that when evaluating different models, avoid applying logistic models to predict long-term risks and that the levels considered between models be verified regularly. In data science, machine learning is applied to address a vast range of issues. Existing data enables foresee results in machine learning. The authors evaluated ensemble classification as a good machine-learning strategy for enhancing multiple classifiers. The ensemble classification only boosts prediction categorization by 7 %. The Cleveland Heart dataset was utilized for training and testing.

In 2021, Akella(20) developed a technique for predicting coronary artery disease (CAD). They employed six machine learning algorithms to predict CAD on the "Cleveland Dataset" to produce a feasible clinical tool for CAD detection. They employed machine learning techniques that were over 80 % accurate and neural network algorithms that were over 93 % accurate. This review includes retrospective studies on the prediction of CHD with machine learning and data mining techniques.

Muhammad et al.(21) constructed CAD prediction models using CAD datasets from two general health centers in Kano State, Nigeria. They used this dataset to support vector machines, K-nearest neighbors, random forests, naïve Bayes, gradient-boosting trees, and logistic regression techniques. The model's effectiveness was tested regarding accuracy, specificity, sensitivity, and ROC curve. For accuracy, the random forest algorithm was the most reliable model with 92,04 %; for specificity, the naïve Bayes algorithm was the ideal model with 92,40 %; for sensitivity, the support vector machine algorithm was the best model with 87,34 %; for the ROC curve, the random forest model was the best model with 92,20 %. The experiment results revealed that the random forest algorithm was the most accurate model in terms of accuracy and ROC curve. Hassan et al.(22) employed machine learning methods to predict CHD by using numerous features to increase the accuracy of the prediction model. Among the 11 classifiers utilized, the gradient boosting tree and the multilayer perceptron had an accuracy of 95 %, and the random forest approach had an accuracy of 96 %. The prediction outcomes revealed that applying feature combinations may successfully enhance the accuracy of the algorithms.

Louridi et al.(23) proposed a novel system to predict patients suffering from cardiovascular diseases by training the UCI heart disease dataset and Framingham datasets using a variety of machine learning algorithms. They employed various approaches to solve the problem of missing values, combining the MICE data imputation method and stacking algorithm to produce good results. Louridi et al.(24) demonstrated that filling missing variables with the mean value, rather than discarding them, produces better results with SVM in classification. Benchaji et al.(25) used a genetic algorithm to handle the problem of uneven data; the given solution yields good recall and accuracy outcomes. EL Asry et al.(26) introduced a new IDS based on PV-DM and feature selection with only four characteristics, achieving 98,92 % accuracy in multiclassification.

ML algorithms enable helpful decision-making for clinical forecasts, and with the introduction of XAI, patient therapy, and diagnosis can be optimized. Moreno-Sanchez et al.[27] have constructed an ML model employing ensemble trees ML approaches to predict heart failure survival among patients. AI applies computer algorithms to investigate complex healthcare data and employs ML to evaluate patients.[28] Peng et al.[29] employed an XAI-based framework for understanding the supplemental diagnosis of hepatitis with faith in the prediction performance. In most situations, the complicated ML models avoid describing how they came to a judgment, resulting in the skepticism of practitioners. Therefore, the researchers in their study have deployed the XAI framework and picked the transparent black-box ML models, including the eXtreme Gradient Boosting (XGBoost), SVM, random forests, etc., for predicting hepatitis worsening.

METHODOLOGY

An overview of the methodology is presented in figure 1. We employed the Z-Alizadeh dataset to identify prospective patients' risk factors. The data is initially pre-processed and balanced with data augmentation methods such as GANs and VAE, then the obtained data is divided into training and testing subsets. The training subset serves to train the ML models, while the testing set is employed to evaluate the trained algorithm. Furthermore, the results from ML models are explained using the Shapley Additive exPlanations (SHAP) method and Boruta Feature Selection.

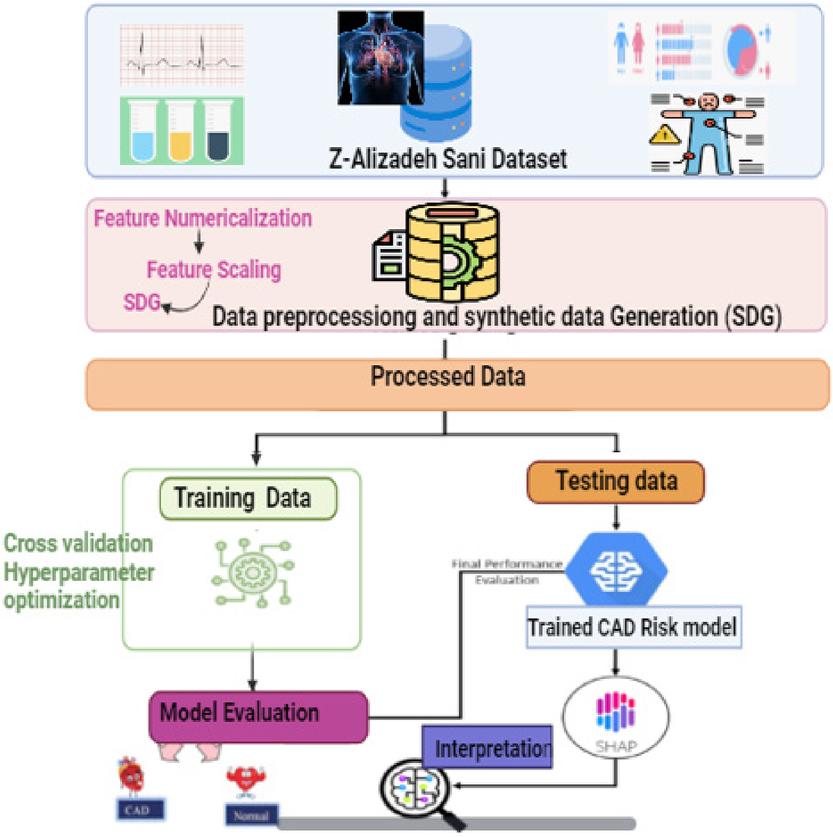

Figure 1. The proposed Architecture.

Z-Alizadeh Sani Dataset

The Z-Alizadeh Sani Dataset is collected from the UCI repository.[30] The Z-Alizadeh Sani dataset is the latest dataset that includes several examination indicators. This dataset comprises 303 medical records from 303 cases of patients who visited Shaheed Rajaei Hospital to treat chest pain. Every record comprises 55 features belonging to four categories. Four categories are demographic features, symptoms, and physical examination; ECG; and laboratory findings and echocardiography features. The patients' ages in the dataset vary from 30 to 86, and there are no duplicate or absent values in the dataset. A record is a sample. These 303 samples belong to two classes, namely, CAD class and conventional class. When the stenosis of the coronary artery lumen of a sample is greater than or equal to 50 %, this sample is classified as CAD class; otherwise, it belongs to the normal class. Accordingly, in 303 samples, 216 instances, accounting for 71,29 %, are CAD class, and 87 instances, accounting for 28,71 %, are normal class.[31] Details of the Z-Alizadeh Sani dataset are shown in table

Table 1. Description of dataset features

Category	Feature Name	Range	Type
Demographic features	Age	30-86	continuous
	Weight	48-120	continuous
	Length	140-188	continuous
	Sex	Male, Female	categorical
	BMI	18,12-40,90	continuous
	DM	0-1	categorical
	HTN	0-1	categorical
	Current smoker	0-1	categorical
	Ex-smoker	0-1	categorical
	FH	0-1	categorical
	Obesity (Yes (BMI > 25), else No)	Y-N	categorical
	CRF	Y-N	categorical
	CVA	Y-N	categorical
	Airway disease	Y-N	categorical
	Thyroid disease	Y-N	categorical
	CHF	Y-N	categorical
	DLP	Y-N	categorical
Symptoms and Physical examination	BP	90-190	continuous
	PR	50-110	continuous
	Edema	0-1	categorical
	Weak peripheral pulse	Y-N	categorical
	Lung rales	Y-N	categorical
	Systolic murmur	Y-N	categorical
	Diastolic murmur	Y-N	categorical
	Typical chest pain	0-1	categorical
	Dyspnea	Y-N	categorical
	Function class	1-4	categorical
	Atypical	Y-N	categorical
	Nonanginal	Y-N	categorical
	Exertional CP	N	categorical
	LowTH Ang	Y-N	categorical
ECG	Q Wave	0-1	categorical
	St elevation	0-1	categorical
	St depression	0-1	categorical
	T inversión	0-1	categorical
	LVH	Y-N	categorical
	Poor R progression	Y-N	categorical
	BBB	N, LBBB, RBBB	categorical
Laboratory Tests and Echocardiography	FBS	62-400	continuous
	CR	0,5-2,2	continuous
	TG	37-1050	continuous
	LDL	18-232	continuous
	HDL	15,9-111	continuous
	BUN	6-52	continuous
	ESR	1-90	continuous

	HB	8,9-17,6	continuous
	K	3-6,6	continuous
	Na	128-156	continuous
	WBC	3700-18000	continuous
	Lymph	7-60	continuous
	Neut	32-89	continuous
	PLT	25-742	continuous
	EF-TTE	15-60.	continuous
	Region RWMA	0, 1, 2, 3, 4	categorical
	VHD	Mild, N, moderate, severe	categoric

Dataset preprocessing

Label conversion and feature numericalization:

- One Hot encoding: Is an approach to transforming data to prepare it for an algorithm and obtain a better forecast. With one-hot, we transform each categorical value into a new categorical column and attribute a binary value of 1 or 0 to those columns. Each integer value can be depicted as a binary vector.
- Mapping with dictionary: We modified the variable by utilizing a dictionary mapping each category to a corresponding integer.
- Feature scaling: In this study we MinMaxScaler. The MinMaxScaler is a method of normalizing data such the th transformed feature has 0 mean and has a standard deviation of 1. The converted features shows us how many standard deviation the original feature is distant from the feature's mean value also called a z-score in statistics.(32)

$$x_{Normalized}=(x_j-x_{min})/(x_{max}-x_{min}) \quad (1)$$

where j in {1,....n} and n is the number of features.

Balancing the data

Machine learning algorithms have proven their capacity to handle complex data structures, providing excellent results in different disciplines, including health care. Nevertheless, an enormous quantity of data must be collected to train these frameworks.(33) This is incredibly challenging in this research because the available dataset is limited (303 records and 55 attributes) such that limited information is unable to be utilized to analyze and establish models.

To address this problem, Synthetic Data Generation (SDG) is one of the most promising approaches and offers several possibilities for collaborative study, including making prediction models and recognizing patterns.

Synthetic Data is artificial data produced through a model trained or constructed to replicate the distributions (i.e., shape and variance) and structure (i.e., correlations among the variables) of real data.(34,35) It has been examined across multiple modalities inside medical care involving biological signals,(36) medical images,(37) and electronic health records (EHR).(38)

In this reasearch we used Generative Adversarial Networks to generate 129 samples from the minority class (Normal), GANs have achieved progress with machine learning models.(39,40) This artificial intelligence strategy discriminates networks to discriminate between synthetic and accurate data. In contrast to classical machine learning algorithms, GANs operate in such a manner that they can learn the joint distribution of the entire data set. GANs use two neural networks: the Generator (G) and the Discriminator (D) networks.(41) The function of the G is to input a random noise vector into synthetic data that approximately represents the actual data.

Conversely, the objective of the D is to take real samples and serve as a trainer who can assess the performance of the output and verify if the data are real.(42) G and D are trained in a manner in which way—through the Min-Max game—the losses of G are minimized, and the losses of D get maximized. The architecture of GAN is described in figure 2 below.

The formula of GAN is:

$$V(D,G)=Ex\sim pr[\log D(x)]+Ez\sim pz(z)[\log(1-D(G(z))) \quad (2)$$

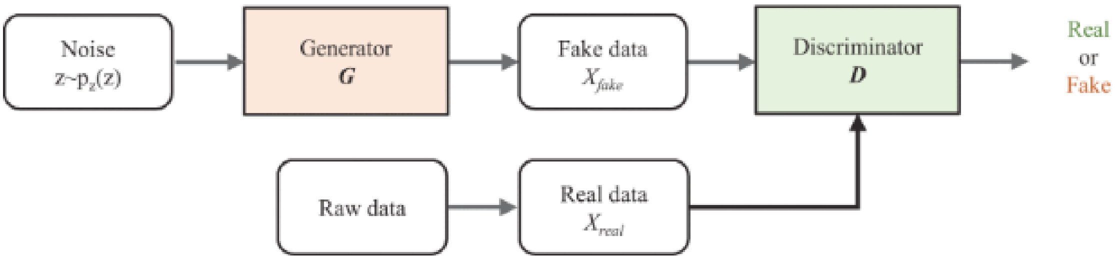

Figure 2. GAN Architecture

Moreover, a VAE network-based method is provided to produce 129 samples of synthetic data from minority class of real data. VAE's approach consists of giving labeled sample data (X) to the Encoder, which captures the distribution of the deep feature (z), and the Decoder, which creates data from the deep feature (z). The architecture of VAE is detailed in figure 3. The VAE design kept each sample's probability and linked the column means to the real data.

Autoencoders are neural networks trained to reduce the reconstruction error between inputs and outputs.[43] The most popular constraint consists of lowering the dimensionality of the hidden layers so that the neural network preserves just the most essential information required to reconstruct the input.

Variational autoencoder (VAE) preserves the design of autoencoder but applies extra constraints on the bottleneck, consequently changing conventional deterministic autoencoder into a potent probabilistic model.[44]

As illustrated in figure 3, the VAE is composed of an encoder E and a decoder D; the distributions of the encoder and decoder are appropriately denoted by qϕ and pθ. The typical VAE assumes that both X and z adhere to Gaussian distributions; therefore, the encoder does not output z directly, but instead, the distribution parameters of z, i.e., the mean and variance of the Gaussian distribution, and z is subsequently reconstructed using reparameterization. The decoder then outputs the mean of the Gaussian distribution using z as the input and sets the variance to a constant value. As the probability function of X, this Gaussian distribution is applied. The model is optimized by calculating the Kullback-Leibler divergence among the prior and posterior distributions of z and the log-likelihood function of X. In a nutshell, VAE learns the distribution of z based on X and rebuilds the distribution of X based on z. The following formula explains the procedure of encoding and decoding:

$$\begin{cases} \mu, \sigma = E(X, \varphi) \\ z = \mu + \sigma\varepsilon \text{ where } \varepsilon \sim N(0,1) \\ X^{\sim} = D(z, \theta) \end{cases} \qquad (3)$$

where μ and σ are the mean and standard deviation of z, $X \in R^L$ is the reconstructed output, ϕ and θ represent

encoder and decoder parameters.

VAE[45] was created to make the distributions learned by the encoder and decoder identical as feasible. Normally, Kullback-Leibler (KL) divergence is employed to characterize the proximity of two distributions. The objective function of the VAE, hence, starts with the KL divergence of the two variational distributions:

$$\log p\theta\,(X) - KL\, q\varphi(z|X)\text{-}||\theta\,(z|X) = -Eq\varphi(z|X)[\log q\varphi(z|X) - \log p\theta\,(X|z) - \log p\theta\,(z)] \quad (4)$$

VAE's loss function is:

$$\begin{aligned} L &= -Eq\varphi(z|X) \log q\varphi(z|X) - \log p\theta\,(X|z) - \log p\theta\,(z)\text{-} \\ &= KL\,[q\varphi(z|X)\,||p\theta\,(z)] - Eq\varphi(z|X)\, \log p\theta\,(X|z) \quad (5) \end{aligned}$$

The initial term is the regularization error of the posterior (|X) and prior pθ(z) distributions. Its aim is to decrease the difference among the posterior and prior distributions.

The second term is the log-likelihood function of X with regard to (|X), pθ(X|z) indicates the distribution of X generated by z, and this term calculates the diference between X and the regenerated output X. Thus, reconstruction error is expressed as:

$$Eq\varphi(zX)[\log p_\theta\,(X|\,z\,)]=(X - X^{\sim})^2 \; (6)$$

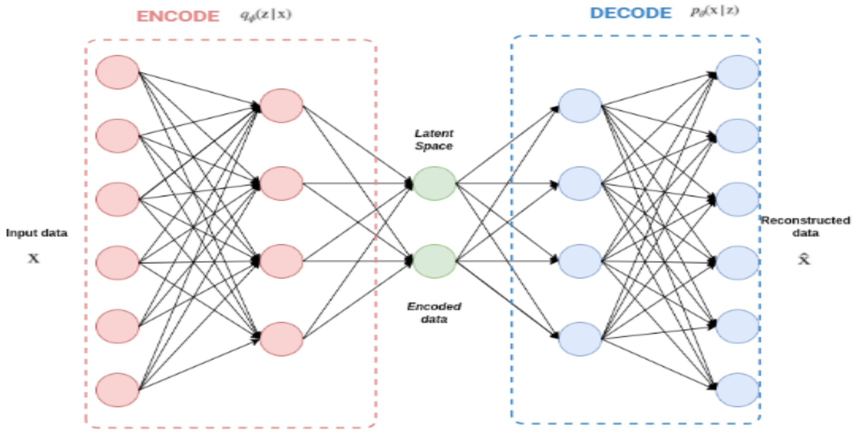

Figure 3. VAE Architecture

Table 2 summurizes the proportion of each class (Normal, CAD), before and after using SDG technics.

Table 2. Proportion of balancing data

Class	Unbalanced	VAE	CopulaGAN
0 Normal	87	216	216
1 CAD	216	216	216

Model Training

In this phase, we splitted our data into 80 % of train data and 20 % test data. We trained our model using Catboost and LightGBM classifiers.

Hyperparameters optimization

In this step we employed Optuna in order to attain optimal performances by determiningthe best parameters of each algorithm. The process of determining the appropriate hyper-parameters for a specific model is automated with the latest machine-learning framework known as Optuna. This open-source tool applies Bayesian optimization, which selects hyper-parameters by balancing exploration and exploitation to effectively research the high-dimensional space of hyper-parameters.[46] Table 3 summurizes the hyperparameters of each method.

Table 3. Results of Hyperparameters optimization

Algorithms	Best Parameters
Catboost	random_state=seed, verbose=0, iterations = 848, objective = 'CrossEntropy', bootstrap_type = 'Bernoulli', od_wait = 1785, learning_rate = 0,09967519851171666, reg_lambda = 14,530290657887146, random_strength = 48,95669246924815, depth = 4,
LightGBM	random_state=seed, max_depth=8, subsample=0,7685741440530103, colsample_bytree=0,33159946082189473, learning_rate = 0,262816149189883, min_child_weight = 7

Cross Validation

Cross-validation is used to avoid concerns like overfitting and underfitting and assess how the model will adapt to an independent dataset.[47] This is accomplished by splitting the entire dataset into two sets: training and testing. The strafied K-fold cross-validation method with k = 5 is applied in the present study. As a result, the entire dataset is split into five folds and iterated five times.

RESULTS

Performance metrics

We applied four different performance metrics, namely accuracy, precision, recall, and F-measures, to assess our model's performance.

- Precision is a metric that measures the ratio of accurately predicted positive observations to all predicted positive observations.

$$\text{precision}= T_p/(T_p+F_p)\ (7)$$

- Accuracy: quantifies the proportion of accurately predicted observations to total observations.

$$\text{accuracy}= (T_p+T_n)/(T_p+F_p+T_n+F_n)$$

- Recall: measures how well the model identifies True Positives.

$$\text{recall}=T_p/(T_p+F_n)\ (9)$$

- F-measure : measures the probability that a positive prediction is correct

$$\text{F-measure}=(2*\text{precision}*\text{recall})/(\text{precision}+\text{recall})\ (10)$$

Performance results

To construct prediction models, 55 crucial attributes are used, and modeling approaches are used to calculate the performance of each model. Three distinct methods are used to generate synthetic data, and two models are trained on synthetic data (Catboost and LightGBM).

As shown in Table 4 and Table 5, the performance of the Catboost combined with VAE proved to be superior to that of other models, with higher precision, recall, and F1 scores.

Consequently, we expect that SHAP Values and Boruta Feature Selection interpretation of the Catboost combined with VAE algorithm.

Additionally, Figure 4 demonstrates that the VAE algorithm achieved an AUC value of 0,82, indicating that it is capable of improving coronary artery disease diagnosis

Table 4. Results of CatBoost

Algorithms	*VAE*	*GAN*	*CopulaGAN*
accuracy	*0,922*	*0,893*	*0,91*
Precision	*0,92*	*0,895*	*0,91*
Recall	*0,924*	*0,889*	*0,912*
F score	*0,921*	*0,889*	*0,909*

Table 5. Results of LightGBM

Algorithms	VAE	GAN	CopulaGAN
accuracy	0,92	0,872	0,91
Precision	0,92	0,863	0,908
Recall	0,93	0,883	0,912
F score	0,92	0,87	0,909

Comparison against other approaches

We analyzed the results produced by our suggested method with those obtained on the Z-AlizadehSani dataset stated in previous literature. As seen in Table 6, our proposed technique is quite competitive. Our proposed technique provides better results than existing studies on the Z-Alizadeh Sani dataset. It is worth mentioning that several cell values are labeled as ‘N’ in Table 6, which signifies that the relevant indicators have not been published in the literature. However, these metrics are essential for evaluating the performance and stability of medical models, especially recall, F1 score, and AUC indicators. Moreover, we implemented another method combined with SMOTE NC oversampling and Catboost to prove the effectiveness of our approach we compared the obtained performance using VAE against performance when using SMOTE NC.

Model Explanation and Interpretation with SHap Values and BFS

Explaining a prediction means the display of either written or visual artifacts that provide a qualitative understanding of the relation between the instance’s characteristics and the model’s prediction. We argue that provided the explanations are accurate and intelligible, explaining predictions is crucial to convincing humans to accept and apply machine learning effectively.(50) When explanations are supplied, a doctor can better determine how to utilize a model. Catboost predicts if a patient has an acute case of CAD, whereas BFS and

SHAP emphasize the features that contributed to this prediction.

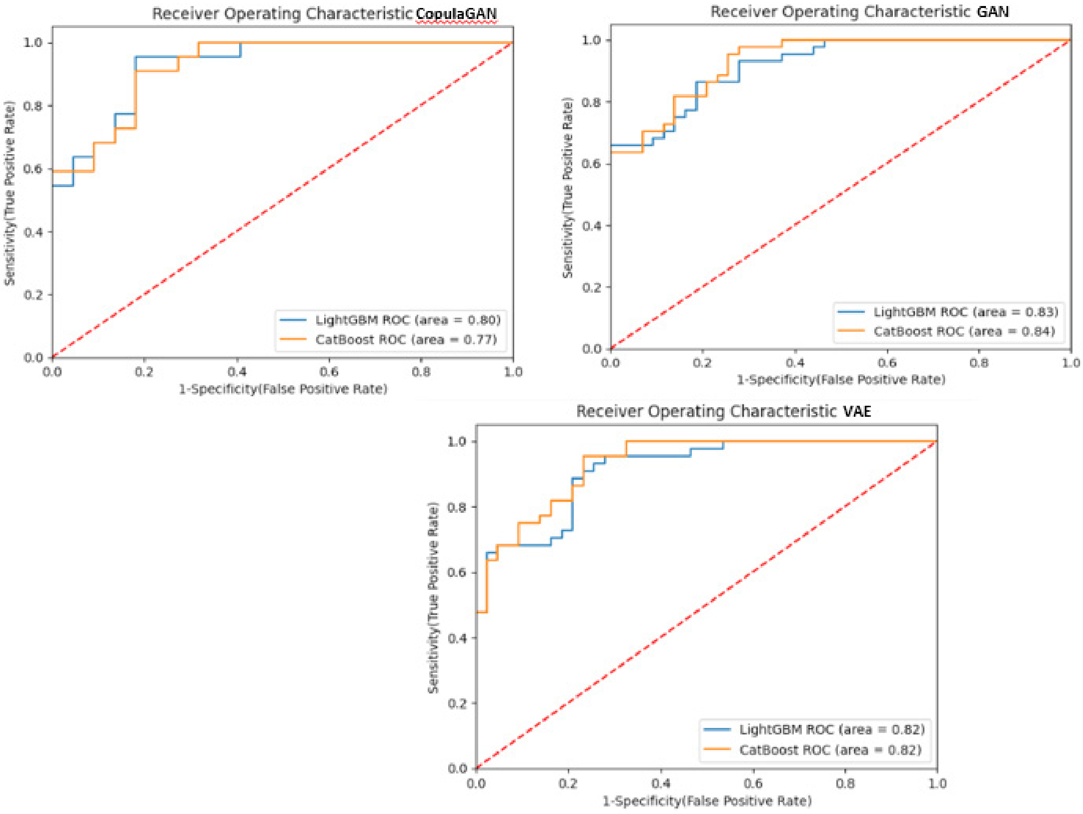

Figure 4. ROC AUC results

Table 6. Comparison against other apporaches

Metrics	Our approach (VAE+Catboost)	SMOTE NC + Catboost	Previous work [48]	Previous work [49]
accuracy	0,92	0,91	0,88	0,89
Precision	0,92	1	0,92	N
Recall	0,93	0,94	0,91	0,80

SHap Values: The Shapley Additive exPlanations (SHAP) framework was employed to understand and explain the model results.[51] As a result, the SHAP python package was used to compute and illustrate the relevance of every characteristic in the prediction. The computation of SHAP values is the basis of this framework. A SHAP value is a feature contribution indicator designed to enhance the interpretability of machine learning methods. SHAP values show how to get from the predicted or base value E[f(x)] to the actual output f if the features are unknown (x). These values also demonstrate how features influence prediction by identifying the direction of the link between the features and the target variable. A feature with a SHAP value closer to 1 or −1 has a substantial positive or vital negative contribution to predicting a specific data point. In contrast, a feature with a SHAP value closer to 0 has a minimal contribution to the prediction.[51] Several graphs that help to comprehend the contributions of features can be derived using this framework.

Boruta Feature Selection: Boruta is a feature selection method built around a random forest classifier. In contrast to the aim of a general feature selection algorithm, the objective of the Boruta feature selection algorithm is to determine the set of characteristics most pertinent to the dependent variable instead of selecting the minimum compact set of properties for which a specific model is appropriate. The specific stages of the Boruta feature selection algorithm are outlined below.[52]

(1) Create a novel feature matrix. Every real feature matrix M element is randomly disordered to acquire the shadow feature matrix M_S. Then, we combine the shaded feature matrix M_S with the original feature matrix M to produce a new feature matrix N, N = [M, M_S].

(2) employ the feature matrix N as input, train the algorithm, and output the Feature_Importances model.

(3) compute the Z_Score metric for the true feature matrix M and the shadow feature matrix M_S. Find the Z_Score metric with the highest shadow feature, denoted as.
(4) Real features with Z_Score that are more significant than are noted as" necessary," and real features with Z_Score that are less than are labeled as "rejected" and excluded from the feature set.
(5) eliminate all shadow features.
(6) Perform steps 1-5 until importance is ascribed to all features or the algorithm has achieved the previous set number of random forest runs.

Figure 5 (a) depicts the topmost 20 SHAP value features for each class in the Z-Alizadeh Sani data prediction model (CAD, Normal). A bar plot diagram presents the distribution of SHAP values for every attribute. In this instance, the listed characteristics are sorted by their highest SHAP value. Figure 5 (b) demonstrates the distribution of SHAP values for every characteristic. Here, the displayed characteristics are ordered by their highest SHAP value. The horizontal axis presents the SHAP value. The larger the positive SHAP value, the higher the positive effect of the feature, and vice versa. The color indicates the magnitude of a characteristic value. The color shifts from red to blue as the feature's value increases and decreases.

After using Boruta Feature Selection (BFS) on the initial dataset including 55 features; seven features including Age, TG, Region RWMA, Typical chest pain, Atypical, FBS and EF-TTE are select as important risk factors.

According to the obtained results using Shap Values and BFS, we can limit the top risk factors leading to CAD to seven risk factors: age, TG, Region RWMA, Typical chest pain, Atypical, FBS, and EF-TTE.

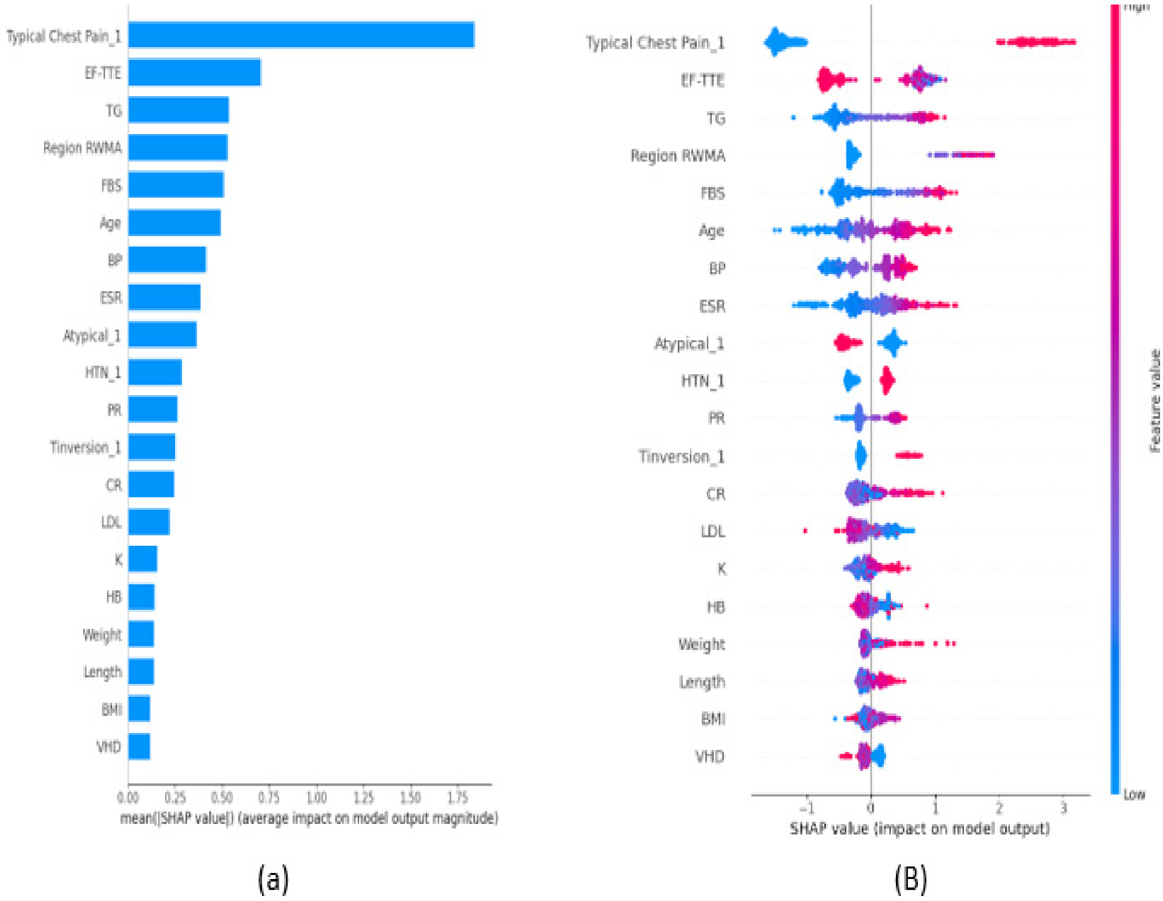

Figure 5. Shap Values results.

DISCUSSION

Early detection of risk factors and primary prevention have considerably lowered mortality and morbidity related to CAD. Risk assessment and preventative care are combined deliberations and decisions that ought to take place between the patient and their doctor.

The CAD defining patients are age, TG, Region RWMA, Typical chest pain, Atypical, FBS, and EF-TTE.

Age appears as non-modifiable CAD risk factor, CAD prevalence rises after 35 years of age among men and

https://doi.org/10.56294/dm202365

women. The cumulative risk of developing CAD in men and women after 40 is 49 % and 32 %, respectively.(53) In both sexes, the risk of CAD increased significantly with age. In the majority of individuals, serum total cholesterol grows as age increases. In men, this increase generally balances off around 45 to 50 years, while for women, the increase continues significantly until the age of 60 to 65.(54)

Epidemiological research demonstrated, that coronary artery disease (CAD) patients possess higher triglyceride (TG) levels over the general population.(55) An elevated level in the serum concentration of triglycerides presents a risk factor for coronary artery disease (CAD), the reason triglycerides are used in calculating concentrations of LDL cholesterol, which undoubtedly is a risk factor for CAD.(56) RWMA can be an indicator of indeterminate embolic stroke. In this sense, RWMA were substantially connected with individuals with imaging showed an embolic source in comparison to those missing this radiographic pattern.(57) Typical chest pain is not itself an illness but a sign of an underlying cardiac condition. Medical problems, such as coronary heart disease, can induce angina. The sort of angina patients suffer may depend on the ailment producing it.(58)

Inadequate fasting blood glucose levels are connected with increased risk of coronary heart disease and strokes.(59,60,61,62,63) Type 1 diabetes mellitus is hypothesized to originate from defective tolerance to self-antigens in vulnerable persons after exposure to incompletely understood environmental variables. Immune-mediated destruction of ß-cells in pancreatic islets results in reduced or absent insulin production and hyperglycemia. By the Brownlee unifying mechanism, elevated cellular glucose oxidation causes excessive mitochondrial production of superoxide along with other reactive oxygen species (oxidative stress) and results in increased flux of glucose through the hexosamine and polyol pathways, generation of advanced glycation end products (AGEs), and activation of protein kinase C (PKC). Via other mediators (e.g., transforming growth factor-ß, endothelin, reduced endothelial nitric oxide synthase; not shown), these pathways contribute to microcirculatory damage, tissue hypoxia, and low-grade inflammation. As a result, microvascular (retinopathy, nephropathy, and neuropathy) and macrovascular problems (coronary heart disease, stroke, and heart failure) are increased by hypertension, significantly in individuals who suffer nephropathy.(64,65,66,67,68,69)

With every heartbeat, the heart contracts and pumps blood out of the heart's primary pumping chamber, the left ventricle. EF refers to the percentage of blood pushed out of your left ventricle with each heartbeat. If the heart muscle has been injured by a heart attack, heart failure, or cardiac valve trouble the EF may be low.The average EF is 50-65 percent. If the EF is below 35 percent, the risk for sudden cardiac arrest increases considerably.(70,71)

CONCLUSIONS

In a nutshell, ML can potentially enhance preventive care for CAD, as proved via models that enhance CAD risk prediction using classical risk factors, clinical and laboratory measurements, imaging, ECG, and omics. Cleaning and balancing data is a vital phase in machine learning since it leads to improved outcomes. The purpose of this research is to employ machine learning to identify the most relevant risk variables for CAD and explain the model predictions. Various machine learning algorithms were examined using distinct performance criteria in attempt to enhance their accuracy by balancing data in diverse ways. To detect risk variables, we applied VAE with Catboost , Boruta and SAHP values. The findings imply that this strategy succeeds effectively. RF was applied to attain the high score accuracy. In order to increase medical diagnosis, this study serves to analyze the topic of cardiovascular problems applying machine learning methodologies. In the future, our purpose is to develop innovative and performant approaches to improve this field and to establish a system of recommendation based on machine learning that helps to avoid CVD and allow individuals to take care of their health. There are various possibilities for supplementary exploration that will considerably boost the functioning of the current research.

REFERENCES

1. Coronary Heart Disease in Morocco [Internet]. World Life Expectancy. Available from: https://www.worldlifeexpectancy.com/morocco-coronary-heart-disease.

2. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. Journal of the American College of Cardiology 2020;76:2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010.

3. Edgardo Olvera Lopez, Jan A. Cardiovascular Disease. National Library of Medicine 2019. https://www.ncbi.nlm.nih.gov/books/NBK535419/.

4. H. Yang, Z. Chen, H. Yang and M. Tian . Predicting coronary heart disease using an improved LightGBM model: Performance analysis and comparison. IEEE Access, vol. 11, pp. 23366-23380, 2023.

5. Rahaman A, Ashit Kumar Dutta. Developing a Deep-Learning-Based Coronary Artery Disease Detection Technique Using Computer Tomography Images. Diagnostics 2023;13:1312–2. https://doi.org/10.3390/diagnostics13071312.

6. A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, N. Engl. J. Med. 380 (14) (2019) 1347–1358.

7. M. Motwani, D. Dey, D.S. Berman, et al., Machine learning for prediction of allcause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis, Eur. Heart J. 38 (2016) 500–507.

8. C. Frederic, P.J. Slomka, G. Markus, et al., Machine learning to predict the longterm risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: a prospective study, Cardiovasc. Res. 116 (14) (2019) 2216–2225.

9. B. Saa, C. Bjm, D. Ag, et al., Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction, JACC (J. Am. Coll. Cardiol.): Heart Fail. 8 (1) (2020) 12–21.

10. E. Zihni, V.I. Madai, M. Livne, et al., Opening the black box of artificial intelligence for clinical decision support: a study predicting stroke outcome, PloS One (2020) 15.

11. M. Athanasiou, K. Sfrintzeri, K. Zarkogianni, et al., An Explainable XGBoost-Based Approach towards Assessing the Risk of Cardiovascular Disease in Patients with Type 2 Diabetes Mellitus[C]//2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, 2020.

12. S.M. Lundberg, B. Nair, M.S. Vavilala, et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nature Biomedical Engineering 2 (10) (2018) 749–760.

13. F. Cabitza, R. Rasoini, G.F. Gensini, Unintended consequences of machine learning in medicine, J. Am. Med. Assoc. 318 (2017) 517–518.

14. S. Lundberg, S.I. Lee, A Unified Approach to Interpreting Model Predictions[C]// Nips, 2017, pp. 4765-4774.

15. Danso SO, Zeng Z, Muniz-Terrera G, Ritchie CW. Developing an Explainable Machine Learning-Based Personalised Dementia Risk Prediction Model: A Transfer Learning Approach With Ensemble Learning Algorithms. Frontiers in Big Data 2021;4. https://doi.org/10.3389/fdata.2021.613047.

16. O. Goldman, O. Raphaeli, E. Goldman, and M. Leshno, ''Improvement in the prediction of coronary heart disease risk by using artificial neural networks,'' Qual. Manage. Health Care, vol. 30, no. 4, pp. 244–250, Jul. 2021, doi: 10.1097/qmh.0000000000000309.

17. Z. Du, Y. Yang, J. Zheng, Q. Li, D. Lin, Y. Li, J. Fan, W. Cheng, X.-H. Chen, and Y. Cai, ''Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: Model development and performance evaluation,'' JMIR Med. Informat., vol. 8, no. 7, Jul. 2020, Art. no. e17257, doi: 10.2196/17257.

18. D. Han, K. K. Kolli, S. J. Al'Aref et al., "Machine learning framework to identify individuals at risk of rapid progression of coronary atherosclerosis: from the PARADIGM registry," Journal of American Heart Association, vol. 9, no. 5, Article ID e013958, 2020.

19. G. Joo, Y. Song, H. Im, and J. Park, "Clinical implication of machine learning in predicting the occurrence of cardiovascular disease using big data (Nationwide Cohort Data in Korea)," IEEE Access, vol. 8, pp. 157643-157653, 2020.

20. A. Akella and S. Akella, ''Machine learning algorithms for predicting coronary artery disease: Efforts toward an open source solution,'' Future Sci. OA, vol. 7, no. 6, Jul. 2021, Art. no. FSO698, doi: 10.2144/fsoa-2020- 0206.

21. L. J. Muhammad, I. Al-Shourbaji, A. A. Haruna, I. A. Mohammed, A. Ahmad, and M. B. Jibrin, "Machine learning predictive models for coronary artery disease," Social Netw. Comput. Sci., vol. 2, no. 5, p. 350, Sep. 2021, doi: 10.1007/s42979-021-00731-4.

22. C. A. U. Hassan, J. Iqbal, R. Irfan, S. Hussain, A. D. Algarni, S. S. H. Bukhari, N. Alturki, and S. S. Ullah, "Effectively predicting the presence of coronary heart disease using machine learning classifiers," Sensors, vol. 22, no. 19, p. 7227, Sep. 2022, doi: 10.3390/s22197227.

23. Louridi, Nabaouia Douzi, Samira Ouahidi, Bouabid. (2021). Machine learning-based identification of patients with a cardiovascular defect. Journal of Big Data. 8. 10.1186/s40537-021-00524-9.

24. Louridi, Nabaouia Amar, Meryem Ouahidi, Bouabid. (2019). Identification of Cardiovascular Diseases Using Machine Learning. 1-6. 10.1109/CMT.2019.8931411

25. Benchaji, Ibtissam Douzi, Samira Ouahidi, Bouabid. (2019). NOVEL LEARNING STRATEGY BASED ON GENETIC PROGRAMMING FOR CREDIT CARD FRAUD DETECTION IN BIG DATA. 3-10. 10.33965/bigdaci2019201907L001.

26. El Asry, Chadia Douzi, Samira Ouahidi, Bouabid.Toward a new IDS based on PV-DM (Paragraph Vector-Distributed Memory Approach).

27. Moreno-Sanchez, P.A. Development of an Explainable Prediction Model of Heart Failure Survival by Using Ensemble Trees. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 4902–4910.

28. Graham, S.A.; Lee, E.E.; Jeste, D.V.; Van Patten, R.; Twamley, E.W.; Nebeker, C.; Depp, C.A. Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review. Psychiatry Res. 2020, 284, 112732.

29. Peng, J.; Zou, K.; Zhou, M.; Teng, Y.; Zhu, X.; Zhang, F.; Xu, J. An Explainable Artificial Intelligence Framework for the Deterioration Risk Prediction of Hepatitis Patients. J. Med. Syst. 2021, 45, 61.

30. https://archive.ics.uci.edu/dataset/412/z+alizadeh+sani

31. Alizadehsani, R.; Habibi, J.; Hosseini, M.J.; Mashayekhi, H.; Boghrati, R.; Ghandeharioun, A.; Bahadorian, B.; Sani, Z.A. A Data Mining Approach for Diagnosis of Coronary Artery Disease. Comput. Methods Programs Biomed. 2013, 111, 52–61. [Google Scholar] [CrossRef] [PubMed]

32. https://databasecamp.de/en/ml/minmax-scaler-en

33. Plesovskaya, E. & Ivanov, S. An empirical analysis of KDE-based generative models on small datasets. Procedia Comput. Sci. 193, 442–452 (2021).

34. Hernandez-Matamoros, A., Fujita, H. & Perez-Meana, H. A novel approach to create synthetic biomedical signals using BiRNN. Inf. Sci. 541, 218–241 (2020).

35. Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., Nakayama, H. et al. GAN-based synthetic brain MR image generation, in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 734–738 (IEEE, 2018).

36. Guan, J., Li, R., Yu, S., & Zhang, X. Generation of synthetic electronic medical record text, in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 374–380 (IEEE, 2018).

37. Xu, L., Skoularidou, M., Cuesta-Infante, A., &Veeramachaneni, K. Modeling tabular data using conditional gan. Adv. Neural Inform. Process. Syst 32 (2019).

38. Kellner, L., Stender, M., Polach, F. V. B. & Ehlers, S. Predicting compressive strength and behavior of ice and analyzing feature importance with explainable machine learning models. Ocean Eng. 255, 111396 (2022).

39. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative Adversarial

Networks: An Overview. IEEE Signal Process. Mag. 2018, 35, 53–65.

40. Park, S.-W.; Ko, J.-S.; Huh, J.-H.; Kim, J.-C. Review on Generative Adversarial Networks: Focusing on Computer Vision and Its Applications. Electronics 2021, 10, 1216.

41. Cauli, N.; Recupero, D.R. Survey on Videos Data Augmentation for Deep Learning Models. Futur. Internet 2022, 14, 93.

42. Ali-Gombe, A.; Elyan, E.; Savoye, Y.; Jayne, C. Few-shot classifier GAN. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8.

43. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. Science 313(5786), 504–507 (2006).

44. Higgins, I. et al. Beta-vae: Learning basic visual concepts with a constrained variational framework (2016).

45. Kingma, D. P., Max, W. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

46. Srinivas and R. Katarya, "hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost", Biomed. Signal Process. Control, vol. 73, Mar. 2022.

47. https://machinelearningmastery.com/repeated-k-fold-cross-validation-withpython

48. 48. Shahid, A.H.; Singh, M.P. A Novel Approach for Coronary Artery Disease Diagnosis using Hybrid Particle Swarm Optimization based Emotional Neural Network. Biocybern. Biomed. Eng. 2020, 40, 1568–1585.

49. Zhang, S.; Yuan, Y.; Yao, Z.; Wang, X.; Lei, Z. Improvement of the Performance of Models for Predicting Coronary Artery Disease Based on XGBoost Algorithm and Feature Processing Technology. Electronics 2022, 11, 315.

50. Molnar, C. Interpretable Machine Learning. Lulu.com (2020).

51. S. Lundberg and S. Lee, “A Unified approach to interpreting model predictions,” in 31st Conf. on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, pp. 1-10, 2017.

52. Kursa MB, Rudnicki WR. Feature selection with the boruta package. J Stat Softw. 2010;36:1–13.

53. Brown JC, Gerhardt TE, Kwon E. Risk Factors For Coronary Artery Disease. PubMed 2023. https://www.ncbi.nlm.nih.gov/books/NBK554410/.

54. Jousilahti P, Vartiainen E, Tuomilehto J, Puska P. Sex, Age, Cardiovascular Risk Factors, and coronary heart disease. Circulation 1999;99:1165–72. https://doi.org/10.1161/01.cir.99.9.1165.

55. Xia T, Li Y, Huang F, Chai H, Huang B, Li Q, et al. The triglyceride paradox in the mortality of coronary artery disease. Lipids in Health and Disease 2019;18. https://doi.org/10.1186/s12944-019-0972-0.

56. Dzoyem JP, Kuete V, Eloff JN. 23 - Biochemical Parameters in Toxicological Studies in Africa: Significance, Principle of Methods, Data Interpretation, and Use in Plant Screenings. ScienceDirect 2014:659–715. https://www.sciencedirect.com/science/article/abs/pii/B9780128000182000236.

57. Kim Y, Kim TJ, Lee S-H. Cardiac wall motion abnormality as a predictor for undetermined stroke with embolic lesion-pattern. Clinical Neurology and Neurosurgery 2020;191:105677. https://doi.org/10.1016/j.clineuro.2020.105677.

58. National Heart, Lung and Blood Institute . Angina (Chest Pain) - Causes and Risk Factors | NHLBI, NIH. Wwwnhlbinihgov 2022. https://www.nhlbi.nih.gov/health/angina/causes.

59. Moezi A, Soltani M, Kazemi T, Bizahem SK, Amirabadizadeh N, Hanafi N, et al. Risk Factors Associated

With the Extent of Coronary Vessel Involvement Across the Spectrum of Coronary Artery Disease. Modern Care Journal 2020;17. https://doi.org/10.5812/modernc.104261.

60. Petrie JR, Sattar N. Excess Cardiovascular Risk in Type 1 Diabetes Mellitus. Circulation 2019;139:744–7. https://doi.org/10.1161/circulationaha.118.038137.

61. Ye Z, Lu H, Li L. Reduced Left Ventricular Ejection Fraction Is a Risk Factor for In-Hospital Mortality in Patients after Percutaneous Coronary Intervention: A Hospital-Based Survey. Biomed Res Int. 2018 Dec 5;2018:8753176. doi: 10.1155/2018/8753176.

62. Gaviño Contreras, J., Ultreras Rodríguez, A., & Sánchez Gaviño, A. (2023). Organizational Behavior for the Integral Human Balance since NOM-035 in post-COVID-19 pandemic scenario. Revista Científica Empresarial Debe-Haber, 1(2), 41-57.

63. Rodríguez-Pérez JA. Strengthening the Implementation of the One Health Approach in the Americas: Interagency Collaboration, Comprehensive Policies, and Information Exchange. Seminars in Medical Writing and Education 2022;1:11-11. https://doi.org/10.56294/mw202211.

64. Farhaoui, Y., “Intrusion prevention system inspired immune systems” Indonesian Journal of Electrical Engineering and Computer Science 2016; 2(1):168–179.

65. Inastrilla CRA. Big Data in Health Information Systems. Seminars in Medical Writing and Education 2022;1:6-6. https://doi.org/10.56294/mw20226.

66. Farhaoui, Y. and All, Big Data Mining and Analytics, 2022, 5(4), pp. I IIDOI: 10.26599/BDMA.2022.9020004

67. Alaoui, S.S., and all. "Hate Speech Detection Using Text Mining and Machine Learning", International Journal of Decision Support System Technology, 2022, 14(1), 80. DOI: 10.4018/IJDSST.286680

68. Alaoui, S.S., and all. ,"Data openness for efficient e-governance in the age of big data", International Journal of Cloud Computing, 2021, 10(5-6), pp. 522-532, https://doi.org/10.1504/IJCC.2021.120391

69. El Mouatasim, A., and all. "Nesterov Step Reduced Gradient Algorithm for Convex Programming Problems", Lecture Notes in Networks and Systems, 2020, 81, pp. 140–148. https://doi.org/10.1007/978-3-030-23672-4_11

70. Tarik, A., and all."Recommender System for Orientation Student" Lecture Notes in Networks and Systems, 2020, 81, pp. 367–370. https://doi.org/10.1007/978-3-030-23672-4_27

71. Sossi Alaoui, S., and all. "A comparative study of the four well-known classification algorithms in data mining", Lecture Notes in Networks and Systems, 2018, 25, pp. 362–373. https://doi.org/10.1007/978-3-319-69137-4_32

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Louridi Nabaouia, Douzi Samira, El Ouahidi Bouabid.
Research: Louridi Nabaouia, Douzi Samira, El Ouahidi Bouabid.
Drafting - original draft: Louridi Nabaouia, Douzi Samira, El Ouahidi Bouabid.
Writing - proofreading and editing: Louridi Nabaouia, Douzi Samira, El Ouahidi Bouabid.