Utilization of machine learning techniques to predict coronary artery Disease: A Comparative Study

استخدام خوارزميات التعلم الالي للتنبؤ بمرض الشريان التاجي: دراسة مقارنة

Sarmad muwfak Khazaal **Haitham Maarouf**
سرمد موفق خزعل **هيثم معروف**
Modern University for Business & Science, Beirut, Lebanon
Sarmad.m.khazaal@gmail.com *Abedrchehab@mubs.edu.lb*

تاريخ تقديم البحث: 2023/07/20
تاريخ قبول النشر: 2022/08/06

Abstract

Background: The coronary arteries are essential in transporting blood to the heart, as it is through the blood that the heart brings in sufficient oxygen to carry out its function. Coronary artery disease occurs due to hardening and blockage in the cavity of these arteries. It is a severe disease that affects people of old age and occurs suddenly without warning.

Objective: Employing machine learning algorithms to diagnose this disease is a severe concern that must be diagnosed early in patients, as neglecting it leads to death.

Methods: Through this study, machine learning algorithms (RF, SVM, NB, LR, SLR) were used as an appropriate solution in the early prediction of coronary heart disease, thus assisting specialists and healthcare workers in the early detection of coronary heart disease. This disease will reduce the severity of the disease and reduce the number of deaths that occur due to this disease. Clinical characteristics indices were used for a group of 303 subjects, adopted from a study conducted by R. Alizadehsani et al [1], to predict CAD, where 216 subjects are confirmed with CAD and 87 are not.

Results: This study found accuracy effects for the involved algorithms (RF, SVM, NB, LR, SLR) which are 78.69%, 72.13%, 45.90%, 81.97%, 83.61%, respectively. The AUC value was (79.3%, 67.3%, 85.9%). %, 70.5%, 80.1% (respectively).

Conclusions: It is concluded, after many comparisons between the machine learning algorithms that are operated and in the light of what is appropriate to the subject of this analysis, that the most suitable algorithm in terms of model accuracy in performance is the Simple Linear Regression (SLR) algorithm with an accuracy of 83.61% and an AUC value of 80.1%.

Keywords: AI technologies, Machine learning, Expert systems, Artificial Neural Network, Natural Language Processing, Heuristics Analysis, Big data, AI applications.

المستخلص

الخلفية العلمية: الشرايين التاجية ضرورية في نقل الدم إلى القلب ، حيث من خلال الدم ياخذ القلب كمية كافية من الأكسجين لأداء وظيفته. يحدث مرض الشريان التاجي نتيجة تصلب وانسداد في تجويف هذه الشرايين. وهو مرض خطير يصيب كبار السن ويحدث فجأة دون سابق إنذار.

الهدف: استخدام خوارزميات التعلم الآلي لتشخيص هذا المرض كونه مصدر قلق شديد يجب تشخيصه مبكرًا لدى المرضى ، حيث يؤدي إهماله إلى الوفاة.

الطرق: من خلال هذه الدراسة ، تم استخدام خوارزميات التعلم الآلي (RF، SVM، NB، LR، SLR) كحل مناسب في التنبؤ المبكر بأمراض القلب التاجية ، وبالتالي مساعدة المتخصصين والعاملين في مجال الرعاية الصحية في الكشف المبكر عن

أمراض القلب التاجية. سيؤدي ذلك إلى تقليل شدة المرض وتقليل عدد الوفيات التي تحدث بسبب هذا المرض. تم استخدام مؤشرات الخصائص السريرية لمجموعة من 303 شخصًا ، تم تبنيها من دراسة أجراها R. Alizadehsani et al [1]، للتنبؤ بمرض الشريان التاجي CAD، حيث تم تاكيد اصابة 216 شخص بـ CAD، و 87 شخص غير مصابين بـ CAD.

النتيجة: وجدت هذه الدراسة تأثيرات الدقة للخوارزميات المعنية (RF، SVM، NB، LR، SLR) والتي بلغت 78.69٪ ، 72.13٪ ، 45.90٪ ، 81.97٪ ، 83.61٪ على التوالي. كانت قيمة AUC (79.3٪ ، 67.3٪ ، 85.9٪). ٪ ، 70.5٪ ، 80.1٪ (على التوالي).

الاستنتاجات: استنتج بعد العديد من المقارنات بين خوارزميات التعلم الآلي التي تم تطبيقها وفي ضوء ما هو مناسب لموضوع هذا التحليل أن الخوارزمية الأنسب من حيث دقة النموذج في الأداء هي خوارزمية الانحدار الخطي البسيط (SLR) بدقة 83.61٪ وقيمة AUC 80.1٪.

الكلمات المفتاحية: تكنولوجيا الذكاء الاصطناعي، التعلم آلالي، الانظمة الخبيرة، الشبكات العصبية الاصطناعية، اللغة البرمجية العصبية، التحليل الارشادي، البيانات الكبيرة، تطبيقات الذكاء الاصطناعي.

Introduction

Artificial intelligence is going through continuous developments, especially as an example of this in deep learning technologies, which is one of the most widespread and developed areas of artificial intelligence. Where it is used in the field of medicine in the detection, prediction and analysis of medical images based on big data [2]. The human brain is a network of huge and complex neurons that specialists have sought to make artificial intelligence design that it mimics since it was formulated in 1950. This matter is very complex for developers and who are still seeking to develop artificial intelligence for scientific fields. The concept of artificial intelligence is the creation of machines that can carry out human tasks called automatic intelligence [3]. Also, artificial intelligence is one of the branches of computer science, which was established in 1956 one of the important things that humans have implemented by making machines that can perform a specific task through a program that was prepared in advance, with high accuracy and speed. Artificial intelligence can deal with large data with high efficiency and this has an effect in the clinical diagnosis process and medical care as a whole [4]. Artificial intelligence plays a role in the clinical characteristics of patients' health during the clinical diagnosis period, as artificial intelligence technologies help doctors working in medical care to predict accurately and faster than current traditional methods [5]. Cardiovascular Disease CVD is widely raised in the world with 17.9 million people dies annually due to this disease by 31% of the total. Coronal artery disease is one of the most prevalent heart disease and blood vessels. Clinical care workers seek to use automatic learning algorithms that help early detect CAD. It turns out that SVM algorithm is the best algorithm capable of predicting CAD disease with a resolution of 0.8947 [6]. It turns out that people who do not have clinical properties are dangerous and bad. That is why algorithms that predict the clinical risk factors that depend on data from the CTCIGRAPY (CTCA). As well as the most advanced automatic learning methods that help to give the best risk factors depending on computerized photography data. To build high -resolution algorithms models in CAD risks [7]. A Stress echocardiography is a good tool in the diagnosis of CAD disease. Identical learning algorithms have been used to detect heart disease for people who have chest pain and this may allow us to know the clinical properties of patients who are likely to have CAD. SVM algorithm showed the best performance of risk factors with a precision of 67.63% [8]. Automatic learning curricula and tools are important in clinical medical applications. TRE-Based Pipeline Optimization Tool (TPOT) was used to know and evaluate its performance in the CAD disease associated with Genomics. Where Single Nucleotide Polymorphisms (SNPS) and their contributions to Tree-Based

Pipeline Optimization Tool (TPOT) predictions (TPOT) has been studied where good CAD cases have been detected to predict [9]. The early detection of CAD can prevent the occurrence of the disease and reduce deaths. CAD coronary artery disease was predicted by generating an algorithm model and verifying the model by relying on clinical factors between two groups of people. The model was generated using the Random Forest algorithm and the model was verified through the ROC curve. The results showed that the AUC value for the algorithm is good and it is 0.948 and that the AUC value for the first verification group was good and at a value It is useful for diagnosing CAD disease [10].

On this basis, this current practical study highlighted the use of artificial intelligence techniques to predict the CAD CAD, as it is considered a fatal disease in many parts of the world. Accordingly, the automatic learning algorithms were used (RF, SVM, NB, LR, SLR), as it is used to predict CAD disease. After many comparisons between the automatic learning algorithms that were applied and in light of what is appropriate for the subject of this analysis, it was concluded that the algorithm is the most appropriate in terms of the accuracy of the model in performance is the SMR (SLR) algorithm (SLR) with Accuracy of 83.61 % and the AUC value 80.1 %.

Martials and Models

Descriptive analysis

This part will introduce data analysis, description and classification of data work to show results with high accuracy and thus achieve the highest benefit from the data of this study. Table (1) presents a complete description of the clinical factors, which are (demographic variables, symptom and examination variables, ECG variables, laboratory and echo variables) with their normal and correct level.

Table1. The attributes with their significances

S. No.	Attributes	Representative icon	Significances (Range)	Type
Demographic				
1	Age	Age in years	30 - 86	integer
2	Weight	Weight in kilograms	48 – 120 Kg	integer
3	Length	Length in centimeter	cm	integer
4	Sex	Male-Female	Male-Female	polynomial
5	BMI	Body Mass Index	18-41 Kg/m^2	real
6	DM	Diabetes Mellitus	(1=Yes, 0=No)	integer
7	HTN	Hyper Tension	(1=Yes, 0=No)	integer
8	Current smoker	Current Smoker	(1=Yes, 0=No)	integer
9	Ex-Smoker	Ex-Smoker	(1=Yes, 0=No)	integer
10	FH	Family History	(1=Yes, 0=No)	integer
11	Obesity	Obesity	(1=Yes, 0=No)	polynomial
12	CRF	Chronic Renal Failure	((1=Yes, 0=No)	polynomial
13	CVA	Cerebrovascular Accident	(1=Yes, 0=No)	polynomial
14	Airway disease	Airway Disease	(1=Yes, 0=No)	polynomial
15	Thyroid Disease	Thyroid Disease	(1=Yes, 0=No)	polynomial
16	CHF	Chronic Renal Failure	(1=Yes, 0=No)	polynomial
17	DLP	Dyslipidemia	(1=Yes, 0=No)	polynomial

S. No.	Attributes	Representative icon	Significances (Range)	Type
Symptom and Examination				
18	BP	Blood Pressure	90 – 190 mmHg	integer
19	PR	Pulse Rate	50 - 110ppm	integer
20	Edema	Edema	(1=Yes, 0=No)	integer
21	Weak peripheral pulse	Weak Peripheral Pulse	(1=Yes, 0=No)	polynomial
22	Lung rales	Lung Rales	(1=Yes, 0=No)	polynomial
23	Systolic murmur	Systolic Murmur	(1=Yes, 0=No)	polynomial
24	Diastolic murmur	Diastolic Murmur	(1=Yes, 0=No)	polynomial
25	Typical Chest Pain	Typical Chest Pain	(1=Yes, 0=No)	integer
26	Dyspnea	Dyspnea	(1=Yes, 0=No)	polynomial
27	Function class	Function Class	0, 1, 2, 3	integer
28	Atypical	Atypical	(1=Yes, 0=No)	polynomial
29	Nonanginal	Nonanginal	(1=Yes, 0=No)	polynomial
30	Exertional CP	Exertional Chest Pain	(1=Yes, 0=No)	polynomial
31	Low Th Ang	Low Threshold Angina	(1=Yes, 0=No)	polynomial
ECG				
32	Q Wave	Q Wave	(1=Yes, 0=No)	integer
33	ST Elevation	ST Elevation	(1=Yes, 0=No)	integer
34	ST Depression	ST Depression	(1=Yes, 0=No)	integer
35	T inversion	T Inversion	(1=Yes, 0=No)	integer
36	LVH	Left Ventricular Hypertrophy	(1=Yes, 0=No)	polynomial
37	PRWP	poor R Wave Progression	(1=Yes, 0=No)	polynomial
38	BBB	Bundle Branch Block	(0=N, 1=RBBB, 2=LBBB)	polynomial
Laboratory and Echo				
39	FBS	Fasting Blood Sugar	62 – 400 mg/dl	integer
40	CR	Creatine	0.5 – 2.2 mg/dl	real
41	TG	Triglyceride	37 – 1050 mg/dl	integer
42	LDL	Low Density Lipoprotein	18 – 232 mg/dl	integer
43	HDL	High Density Lipoprotein	15 – 111 mg/dl	integer
44	BUN	Blood Urea Nitrogen	6 – 52 mg/dl	integer
45	ESR	Erythrocyte Sedimentation Rate	1 – 90 mm/h	integer
46	HB	Hemoglobin	8.9 – 17.6 g/dl	real
47	K	Potassium	3.0 – 6.6 mEq/lit	real
48	Na	Sodium	128 – 156 mEq/lit	integer
49	WBC	White Blood Cell	3700 – 18.000 cells/ml	integer
50	Lymph	Lymphocyte	7 – 60 %	integer
51	Neut	Neutrophil	32 – 89 %	integer
52	PLT	Platelet	25 – 742 1000/ml	integer
53	EF	Ejection Fraction	15 – 60 %	integer
54	Region with RWMA	Regional Wall Motion Abnormality	0, 1, 2, 3, 4	integer
55	VHD	Valvular Heart Disease	(0=Normal, 1=mild, 2=moderate, 3=severe)	polynomial
56	CAD	Coronary Artery Disease	(yes, no)	polynomial

Analysis Mechanism

Correlation Matrix

The main purpose of this tool is to show the strength and correlation between the variables (Attributes) in the dataset in order to find the correlation between the most effective biomarker, which will be inferred by determining the presence or absence of disease.

Attributes	Age	Weight	Length	Sex	BMI	DM	HTN	Current...	EX-Smoker	FH	Obesity	CRF	CVA	Airway ...	Thyroid...	CHF	DLP
Age	1	-0.265	-0.164	0.046	-0.161	0.073	0.247	-0.144	0.077	-0.184	0.126	0.127	0.026	0.070	-0.096	-0.022	-0.128
Weight	-0.265	1	0.461	-0.235	0.725	-0.004	-0.029	0.157	0.069	0.022	-0.547	-0.026	0.052	-0.058	0.033	0.030	0.080
Length	-0.164	0.461	1	-0.700	-0.269	-0.052	-0.154	0.335	0.079	0.004	0.172	-0.034	-0.007	0.004	-0.042	0.014	0.173
Sex	0.046	-0.235	-0.700	1	0.284	0.194	0.149	-0.336	-0.157	0.071	-0.212	-0.025	-0.005	-0.022	0.092	-0.049	-0.278
BMI	-0.161	0.725	-0.269	0.284	1	0.045	0.092	-0.089	0.005	0.014	-0.713	0.009	0.067	-0.063	0.069	0.020	-0.047
DM	0.073	-0.004	-0.052	0.194	0.045	1	0.218	-0.208	-0.120	-0.064	-0.021	0.115	0.029	0.028	-0.052	-0.037	-0.250
HTN	0.247	-0.029	-0.154	0.149	0.092	0.218	1	-0.169	0.041	-0.098	-0.136	0.118	0.055	0.054	0.039	-0.069	-0.109
Current S...	-0.144	0.157	0.335	-0.336	-0.089	-0.208	-0.169	1	-0.095	0.090	0.051	0.044	-0.003	0.074	-0.079	-0.029	0.190
EX-Smoker	0.077	0.069	0.079	-0.157	0.005	-0.120	0.041	-0.095	1	-0.080	-0.042	0.106	-0.024	-0.036	-0.028	-0.011	0.103
FH	-0.184	0.022	0.004	0.071	0.014	-0.064	-0.098	0.090	-0.080	1	-0.011	0.068	0.015	-0.084	0.054	-0.025	-0.061
Obesity	0.126	-0.547	0.172	-0.212	-0.713	-0.021	-0.136	0.051	-0.042	-0.011	1	0.009	-0.029	0.025	-0.006	-0.038	0.074
CRF	0.127	-0.026	-0.034	-0.025	0.009	0.115	0.118	0.044	0.106	0.068	0.009	1	0.168	-0.028	-0.022	-0.008	0.011
CVA	0.026	0.052	-0.007	-0.005	0.067	0.029	0.055	-0.003	-0.024	0.015	-0.029	0.168	1	-0.025	-0.020	-0.007	0.046
Airway di...	0.070	-0.058	0.004	-0.022	-0.063	0.028	0.054	0.074	-0.036	-0.084	0.025	-0.028	-0.025	1	-0.030	-0.011	0.039
Thyroid D...	-0.096	0.033	-0.042	0.092	0.069	-0.052	0.039	-0.079	-0.028	0.054	-0.006	-0.022	-0.020	-0.030	1	-0.009	0.027
CHF	-0.022	0.030	0.014	-0.049	0.020	-0.037	-0.069	-0.029	-0.011	-0.025	-0.038	-0.008	-0.007	-0.011	-0.009	1	-0.075
DLP	-0.128	0.080	0.173	-0.278	-0.047	-0.250	-0.109	0.190	0.103	-0.061	0.074	0.011	0.046	0.039	0.027	-0.075	1

Figure1. Correlation matrix of clinical demographic characteristics of peoples

Attributes	Age	BP	PR	Edema	Weak ...	Lung ral...	Systolic Murmur	Diastolic ...	Typical ...	Dyspnea	Functio...	Atypical	Nonang...	Exertional CP	LowTH ...
Age	1	0.216	0.024	0.132	0.154	0.106	0.045	0.030	0.138	0.059	0.051	-0.142	-0.089	-0.066	0.087
BP	0.216	1	0.183	0.085	0.017	-0.051	0.037	-0.022	0.115	0.061	0.018	-0.115	-0.022	-0.090	0.088
PR	0.024	0.183	1	0.061	-0.011	0.120	0.179	0.135	0.080	0.008	0.053	-0.138	-0.019	0.031	-0.047
Edema	0.132	0.085	0.061	1	0.107	0.322	0.068	-0.036	-0.017	0.126	0.050	0.012	0.028	-0.012	-0.017
Weak Peripheral Pulse	0.154	0.017	-0.011	0.107	1	-0.025	0.024	-0.023	0.067	0.093	0.067	-0.086	-0.031	-0.007	-0.011
Lung rales	0.106	-0.051	0.120	0.322	-0.025	1	0.284	0.174	-0.105	0.111	0.063	0.024	0.033	-0.011	-0.016
Systolic Murmur	0.045	0.037	0.179	0.068	0.024	0.284	1	0.329	-0.101	0.172	0.148	0.009	-0.007	-0.023	-0.032
Diastolic Murmur	0.030	-0.022	0.135	-0.036	-0.023	0.174	0.329	1	-0.151	0.157	0.152	-0.032	0.133	-0.010	-0.014
Typical Chest Pain	0.138	0.115	0.080	-0.017	0.067	-0.105	-0.101	-0.151	1	-0.194	0.072	-0.723	-0.256	0.053	-0.007
Dyspnea	0.059	0.061	0.008	0.126	0.093	0.111	0.172	0.157	-0.194	1	0.420	-0.031	0.146	0.065	0.009
Function Class	0.051	0.018	0.053	0.050	0.067	0.063	0.148	0.152	0.072	0.420	1	-0.158	-0.080	-0.037	-0.052
Atypical	-0.142	-0.115	-0.138	0.012	-0.086	0.024	0.009	-0.032	-0.723	-0.031	-0.158	1	-0.157	-0.038	0.034
Nonanginal	-0.089	-0.022	-0.019	0.028	-0.031	0.033	-0.007	0.133	-0.256	0.146	-0.080	-0.157	1	-0.014	-0.019
Exertional CP	-0.066	-0.090	0.031	-0.012	-0.007	-0.011	-0.023	-0.010	0.053	0.065	-0.037	-0.038	-0.014	1	-0.005
LowTH Ang	0.087	0.088	-0.047	-0.017	-0.011	-0.016	-0.032	-0.014	-0.007	0.009	-0.052	0.034	-0.019	-0.005	1

Figure2. Correlation matrix of clinical characteristics of ECG of peoples

Attributes	Age	Q Wave	St Elev...	St Depr...	Tinvers...	LVH	Poor R ...	BBB
Age	1	-0.062	-0.057	0.177	0.042	0.126	0.004	0.024
Q Wave	-0.062	1	0.440	-0.061	0.040	-0.063	0.046	-0.062
St Elevati...	-0.057	0.440	1	-0.122	0.132	-0.059	-0.039	-0.058
St Depre...	0.177	-0.061	-0.122	1	0.322	-0.116	-0.097	-0.127
Tinversion	0.042	0.040	0.132	0.322	1	-0.144	-0.071	-0.154
LVH	0.126	-0.063	-0.059	-0.116	-0.144	1	-0.047	-0.070
Poor R P...	0.004	0.046	-0.039	-0.097	-0.071	-0.047	1	-0.046
BBB	0.024	-0.062	-0.058	-0.127	-0.154	-0.070	-0.046	1

Figure3. Correlation matrix of clinical characteristics of ECG of peoples

Attributes	Age	FBS	CR	TG	LDL	HDL	BUN	ESR	HB	K	Na	WBC	Lymph	Neut	PLT	EF-TTE	Region ...	VHD	cad
Age	1	0.015	0.227	-0.111	-0.034	-0.036	0.301	0.183	-0.161	0.154	-0.072	0.020	-0.172	0.173	-0.049	-0.141	0.109	0.111	-0.357
FBS	0.015	1	0.070	0.098	-0.102	-0.054	0.230	0.144	-0.164	0.103	-0.059	0.160	-0.004	0.032	0.020	-0.057	0.037	0.021	-0.206
CR	0.227	0.070	1	-0.038	-0.115	-0.123	0.512	0.024	-0.020	-0.010	-0.075	0.145	-0.067	0.097	-0.092	-0.115	0.031	0.051	-0.087
TG	-0.111	0.098	-0.038	1	0.189	-0.035	0.029	-0.045	0.124	0.023	0.060	0.012	0.090	-0.082	-0.049	-0.028	0.035	-0.020	-0.141
LDL	-0.034	-0.102	-0.115	0.189	1	0.305	-0.120	-0.013	0.064	0.038	0.168	0.019	0.118	-0.085	0.013	0.159	-0.027	0.000	0.024
HDL	-0.036	-0.054	-0.123	-0.035	0.305	1	-0.139	-0.084	-0.048	-0.074	0.089	-0.064	0.028	-0.025	0.001	0.104	-0.062	-0.062	0.043
BUN	0.301	0.230	0.512	0.029	-0.120	-0.139	1	0.127	-0.085	0.099	-0.136	0.088	-0.045	0.024	0.041	-0.117	0.018	0.135	-0.089
ESR	0.183	0.144	0.024	-0.045	-0.013	-0.084	0.127	1	-0.390	0.007	-0.069	0.161	-0.158	0.139	0.247	-0.057	0.055	0.063	-0.178
HB	-0.161	-0.164	-0.020	0.124	0.064	-0.048	-0.085	-0.390	1	0.033	0.139	-0.001	0.084	-0.075	-0.106	0.006	-0.045	0.005	0.042
K	0.154	0.103	-0.010	0.023	0.038	-0.074	0.099	0.007	0.033	1	0.011	0.119	-0.009	-0.003	0.023	-0.160	0.229	0.030	-0.181
Na	-0.072	-0.059	-0.075	0.060	0.168	0.089	-0.136	-0.069	0.139	0.011	1	-0.094	0.141	-0.134	-0.022	0.138	-0.023	-0.055	0.085
WBC	0.020	0.160	0.145	0.012	0.019	-0.064	0.088	0.161	-0.001	0.119	-0.094	1	-0.322	0.378	0.291	-0.138	0.175	0.140	-0.071
Lymph	-0.172	-0.004	-0.067	0.090	0.118	0.028	-0.045	-0.158	0.084	-0.009	0.141	-0.322	1	-0.923	-0.012	0.240	-0.079	-0.093	0.127
Neut	0.173	0.032	0.097	-0.082	-0.085	-0.025	0.024	0.139	-0.075	-0.003	-0.134	0.378	-0.923	1	0.004	-0.229	0.113	0.082	-0.124
PLT	-0.049	0.020	-0.092	-0.049	0.013	0.001	0.041	0.247	-0.106	0.023	-0.022	0.291	-0.012	0.004	1	0.068	-0.011	0.069	0.095
EF-TTE	-0.141	-0.057	-0.115	-0.028	0.159	0.104	-0.117	-0.057	0.006	-0.160	0.136	-0.138	0.240	-0.229	0.068	1	-0.451	-0.361	0.234
Region RWMA	0.109	0.037	0.031	0.035	-0.027	-0.062	0.018	0.055	-0.045	0.229	-0.023	0.175	-0.079	0.113	-0.011	-0.451	1	0.162	-0.316
VHD	0.111	0.021	0.051	-0.020	0.000	-0.062	0.135	0.063	0.005	0.030	-0.055	0.140	-0.093	0.082	0.069	-0.361	0.162	1	-0.019
cad	-0.357	-0.206	-0.087	-0.141	0.024	0.043	-0.089	-0.178	0.042	-0.181	0.085	-0.071	0.127	-0.124	0.095	0.234	-0.316	-0.019	1

Figure 4. Correlation matrix of clinical characteristics of Laboratory and echo of peoples

Machine Learning Algorithms

Machine learning is one of the most meaningful branches of artificial intelligence, which is distinguished by the ability of its algorithms to predict the behavior of data. these algorithms are used in the literature to study the work behaviors of these algorithms by executing them on a dataset, comparing their performance, knowing the prediction effects, and reaching the most suitable performance among these algorithms. In all-around, machine learning algorithms are divided into three main sections: Supervised Learning Algorithms, Unsupervised Learning Algorithms, Reinforcement learning Algorithms.

Random Forest (RF)

It is one of the most well-known classical supervised machine learning algorithms and is employed in regression and classification processes. In this algorithm, the word random means that the algorithm randomly takes a sample from the dataset. On the other hand, the word forest indicates that many decision trees are made on the data sample instead of using a single tree

Support Vector Machine Algorithm (SVM)

It is one of the most famous supervised machine learning algorithms and is considered the most widely involved in many domains, especially in the medical domain. This algorithm is characterized by its ability to achieve classification and regression tasks with a small dataset with a complicated series. In addition, this algorithm aims to change the dataset into a new space in which this data diverges in a way that can be classified and sorted, where this is done by partitioning the data employing the hyperplane.

Naive Bayes(NB)

It is one of the classification methods algorithms. This algorithm is based on the Bayes theorem. This algorithm is characterized by including statistical equations, where the machine is trained on dataset variables. These variables determine the machine, as these variables effectively help the classification and prediction approach.

Logistic Regression (LR)

It is considered one of the most suitable binary classification algorithms, as it is an algorithm that includes equations that can be used to separate the data and divide it into two categories (0, 1). It is an algorithm that is utilized in a lot of literature in analyzing the behaviors of the medical dataset. It has the ability to identify an individual who has a specific disease or not.

Linear regression

It is an algorithm that is considered one of the influential and necessary algorithms in its application. In this algorithm, a scattered dataset that does not have an exact order (random order) is arranged, and this data is dealt with by finding the most suitable way to describe this data and the ease of dealing with it by finding the best line that passes between the dataset and is diverged to perform prediction operations.

Confusion Matrix

This work mainly relied on the confusion matrix, which is a crucial tool in evaluating the performance and work of the five algorithms, as through this matrix, it is possible to know the number of patients who are actually sick and who are not genuinely sick. The confusion matrix is a relationship between the actual dataset and the predicted dataset, as it contains four natural sections (TP,FP,TN,and FN), as illustrated in Figure (5).

Where:

True positive (TP): The number of true positive predictions.
False positive (FP): The number of false positive predictions.
True negative (TN): The number of true negative predictions.
False negative (FN): The number of false negative predictions.

These four axes can calculate mathematical equations from 1 to 6. Figure 4 displays the confusion matrix map, as it contains the one representing patient and the zero representing non-patients. Mathematical equations are used to know the capabilities of the five applied algorithms, which are (Accuracy, Sensitivity, Specificity, Precision, and F1-score):

$\boldsymbol{Accuracy}$: It is the ratio of correct data extracted by the algorithm which are the values (correct positive, correct negative).

$\boldsymbol{Sensitivity}$: is the ratio of correct positive values that are predicted by the algorithm divided by the predicted data (correct positive values, false negative values).

$\boldsymbol{Specificity}$: is the ratio of correct negative values that are predicted by the algorithm divided by the predicted data (correct negative values, false positive values).

$\boldsymbol{Precision}$: is the actual positive values that are predicted correctly by the algorithm and become predicted values divided by actual data (correct positive values, false positive values).

$\boldsymbol{F1-measure}$: find the relationship between Precision and Sensitivity.

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

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

$$Specificity = \frac{TN}{(TN + FP)} \quad \text{(3)}$$

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

$$F1 - measure = \frac{2\,(precision * Sensitivity\,)}{(precision + Sensitivity\,)} \quad \text{(5)}$$

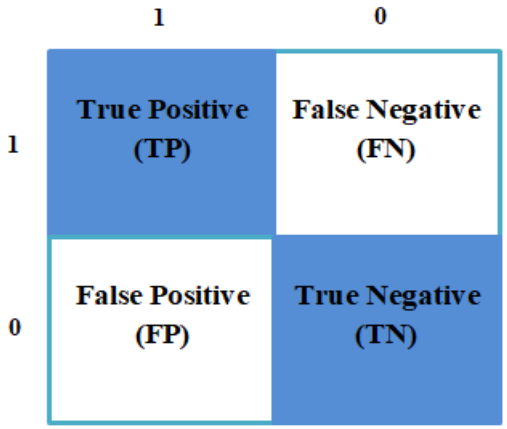
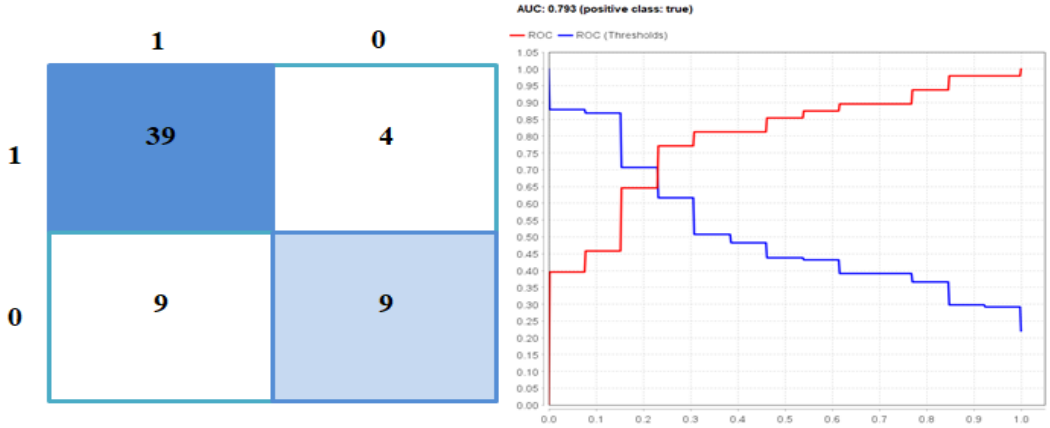


Figure5. Confusion matrix map

Results

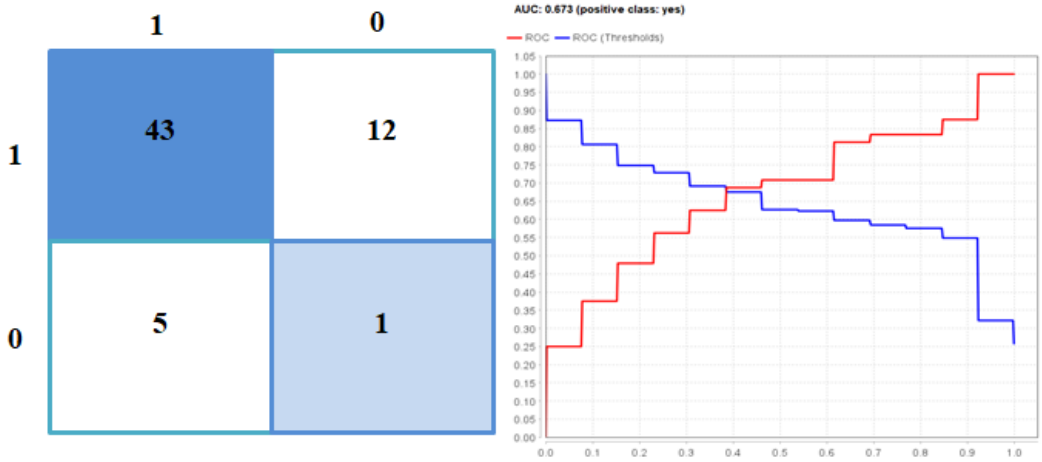
Figures 6-10 indicate the results obtained through the confusion matrix and the AUC for each algorithm. Figure 6 illustrates the results of the confounding matrix of the random forest algorithm through TP = 39, TN = 9, FP = 4 and FN = 9. At the same time, the AUC-ROC is more than 79%, which indicates the algorithm's ability to separate between patients and non-patients.



(a) (b)

Figure6. Random Forest (a) Confusion matrix (b) AUC-ROC Curve.

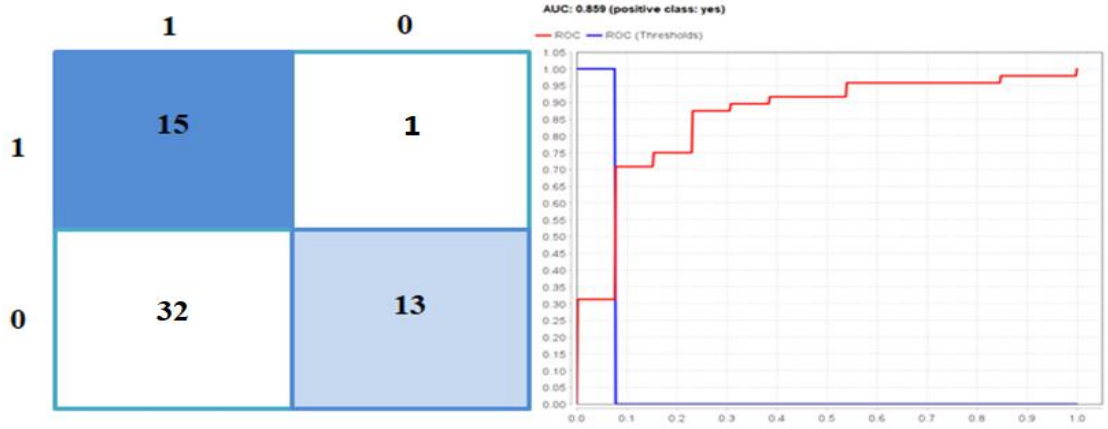
Figure 7 illustrates the results of the confusion matrix of the support vector machine algorithm through TP = 43, TN = 1, FP = 12, and FN = 5. The AUC-ROC is more than 67%, which indicates the algorithm's ability to separate between patients and non-patients.



(a) (b)

Figure7. Support Vector Machine (a) Confusion matrix (b) AUC-ROC Curve

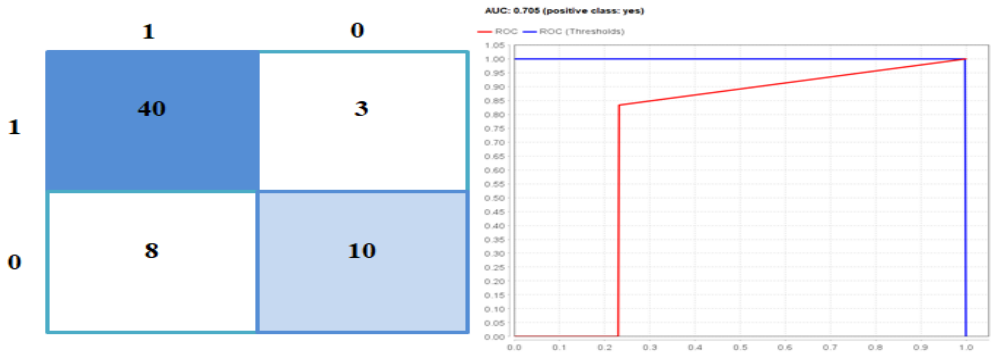
Figure 8 illustrates the results of the confusion matrix of the Naïve Bayes algorithm through TP = 15, TN = 13, FP = 1 and FN = 32. At the same time, the AUC-ROC is more than 85%, which indicates the algorithm's ability to separate between patients and non-patients.



(a) (b)

Figure8. Naïve Bayes (a) Confusion matrix (b) AUC-ROC Curve.

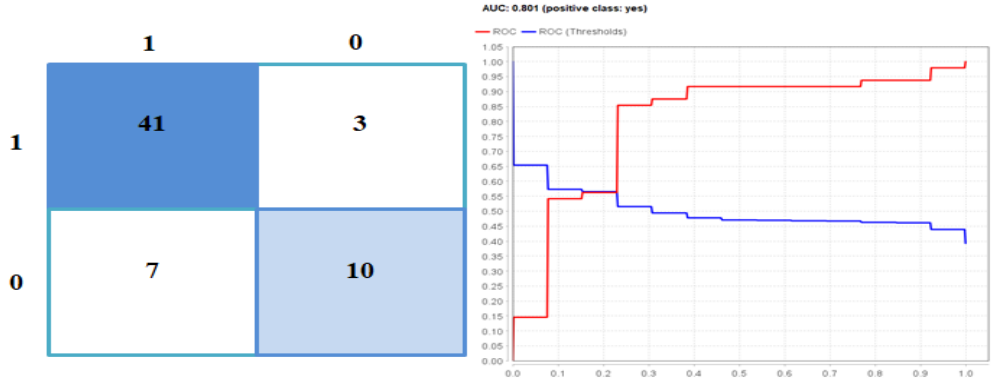
Figure 9 illustrates the results of the confounding matrix of the logistic regression algorithm through TP = 40, TN = 10, FP = 3 and FN = 8. The AUC-ROC is more than 70%, which indicates the algorithm's ability to separate patients and non-patients.



(a) (b)

Figure9. Logistic Regression (a) Confusion matrix (b) AUC-ROC Curve

Figure 10 illustrates the results of the confounding matrix of the simple linear regression algorithm through TP = 41, TN = 10, FP = 3 and FN = 7. At the same time, the AUC-ROC is more than 80%, which indicates the algorithm's ability to separate patients and non-patients.



(a) (b)

Figure10. Simple Linear Regression (a) Confusion matrix (b) AUC-ROC Curve

Table 2 displays all the received results. This table consists of eight metrics (accuracy, sensitivity, specificity, precision, F1-score, training time, and predication time), which compares all the algorithms' effects and determines the performance status for each. In addition, the most suitable and imperfect performance obtained will be determined with the determination of the best path through the algorithms applied in analyzing the behaviors of the coronary artery disease dataset. Figure 11 illustrates a comparison between the applied algorithms through two main metrics (accuracy and AUC), which measure the performance of the algorithms and identify the best practice carried out by the high-performance algorithm.

Table 2. Algorithm results

Algorithms	**Effects Metrics**							
	Accuracy %	Sensitivity %	Specificity %	Precision %	F1-Score %	AUC %	Training Time	Prediction Time
Random Forest	78.69	81.25	69.23	90.70	85.71	79.3	32 sec	28 sec
Support Vector Machine	72.13	89.58	7.69	78.18	83.50	67.3	38 sec	34 sec
Naive Base	45.90	31.25	100	100	47.62	85.9	36 sc	30 sec
Logistic Regression	81.97	83.33	76.92	93.2	87.91	70.5	43 sec	39 sec
Simple Linear regression	**83.61**	**85.42**	**76.92**	**93.18**	**89.13**	**80.1**	**30 sec**	**25 sec**
Bold indicates that these effects are the most useful.								

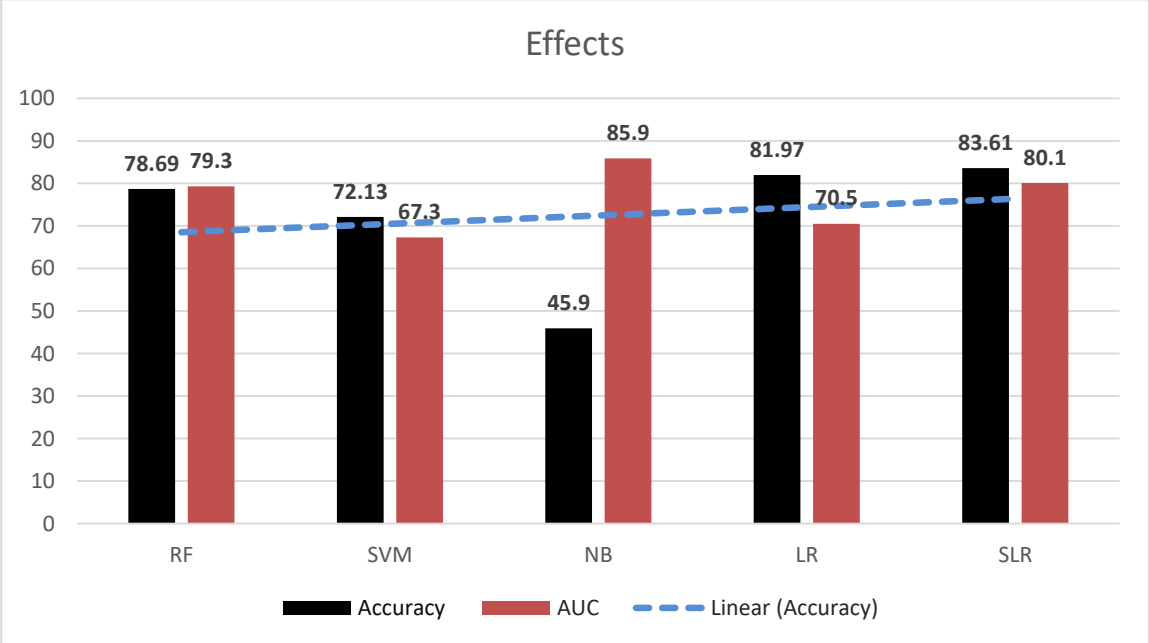


Figure11. Compare the applied algorithms through two main measures (Accuracy and AUC).

Discussion

Figure (1), it is clear that the demographic clinical factors correlation matrix shows that there is a correlation between the factor (age) and the clinical factors: length, BMI, HTN, current smoking, FH, Obesity, CRF, DLP weight, respectively and the value of this correlation is (-0.265, -0.164, -0.161, 0.247, -0.144, -0.184, 0.126, 0.127, -0.128), respectively. Figure (2) demonstrates a correlation between clinical factors, symptoms and examination, indicating that there is a correlation between the factor (age) and the clinical factors (BP, Edema, weak peripheral pulse, lung rales, typical Chest pain, Atypical), respectively, the value of this correlation (0.216, 0.132, 0.154, 0.106, 0.138, -0.142), respectively. For the Exertional CP clinical factor, it is found through the correlation matrix operator that its data contains missing values. It is found by examining the data that this factor has one category, and this generates an error of missing values when applying the matrix. To solve this, a second category was added by transferring a person's diagnosis From (N) to (y). Figure (3) is a correlation matrix between the clinical factors. The ECG indicates that there is a correlation between the factor (age) and the clinical factors (St Depression, LVH), respectively. The value of this correlation is (0.177, 0.126), respectively. As for the clinical factor BBB, it is found through the correlation matrix operator that its data contains missing values. It is found by examining the data that this factor's data is one letter and two words; the length of the word is four letters, which generates an error of missing values when applying the matrix. To solve this, a weight was given to ((0=0=N, 1=RBBB, 2=LBBB). Figure (4) is a correlation matrix between clinical factors, laboratory and echo indicates that there is a correlation between factor (Age) and clinical factors (CR, TG, BUN, ESR, HB, K, Lymph, Neut, EF-TTE, Region RWMA , VHD, CAD) respectively, the value of this correlation is (0.227, -0.111, 0.30, 0.183, -0.161, 0.154, -0.172, 0.173, -0.141, 0.109, 0.111, -0.367), respectively. As for the clinical factor VHD, it is found through the correlation matrix operator that its data contains missing values, and it is found by examining the data that this factor has four different words and long words, and this generates an error of missing values when applying the matrix. To solve this, a weight is given to these words (0 =Normal, 1=mild, 2=moderate, 3=severe).

In the random forest algorithm, it is found that the algorithm correctly predicted 39 people to be patient, and they are actually patient. That is, they are classified correctly. It turned out that nine people are correctly predicted by the algorithm that they are not patient, and they are actually not sick. That is, they were classified correctly. It turned out that the algorithm incorrectly predicted four people as not sick, and they were in fact, sick, i.e., misclassified. This algorithm showed that the algorithm incorrectly predicted nine people to be patient, but they were not patient, i.e., they are misclassified by this algorithm. The accuracy value of this algorithm was 78.69%, reflecting the accuracy of the model as a whole in classification.

In the SVM algorithm, eight clinical factors are determined, which showed that there is a correlation between them and the variable age, which are (Weight, HTN, BP, CR, BUN, ESR, and St Depression).Where the algorithm predicted that 43 people are predicted by the algorithm correctly that they are patient, and they are actually patient, i.e., they are classified correctly, and this is based on the variables that are chosen in this algorithm, through which the naming variable is classified into those with the disease and not infected with the disease. The algorithm correctly predicts one person to be not patient, and he is actually not patient; that is, it is classified correctly. In addition, it is found that 12 people incorrectly predicted by the algorithm to be not patient are in fact sick, i.e., incorrectly misclassified by this algorithm. Likewise, five people are mispredicted by the algorithm that they are patient, and they are in fact not patient; that is, they need to be correctly

classified as misclassification. The accuracy value of this algorithm is 72.13%, and it reflects the accuracy of the model as a whole in classification.

In the Naïve Bayes algorithm, it is found that the algorithm correctly predicts 15 people to be patient, and they are actually patient; that is, they are classified correctly. It turned out that 13 people are correctly predicted by the algorithm that they are not patient and actually not patient; that is, they are classified correctly. It turned out that the algorithm incorrectly predicts one person as not patient, and he is in fact patient; that is, this algorithm misclassifies it. Too, 32 people are mispredicted by the algorithm that they are patient, and they are in fact not patient; that is, they are classified incorrectly by this algorithm. The accuracy value of this algorithm was 45.90%, reflecting the accuracy of the model as a whole in classification.

In the logistic regression algorithm, it is found that 40 people are correctly predicted that they were patient, and they were actually patient; that is, they were classified correctly. Also, it was found that ten people are correctly predicted that they were not sick, and they were actually not sick; that is, they were classified correctly. At the same time, three people were incorrectly predicted to be unwell when they were patient, i.e., misclassified, and eight people were wrongly predicted to be sick when they were not patient, i.e., misclassified. The accuracy of this algorithm is 81.97%, reflecting the accuracy of the model in classification.

The Simple Linear Regression algorithm found that 41 people were correctly predicted to be patient and actually patient, meaning they were classified correctly. It also concluded that ten people were correctly predicted that they were not patient, and they were actually not patient, i.e. they were classified correctly. In addition, it was found that three people were incorrectly predicted by the algorithm as not sick when they were actually sick, i.e. they were misclassified, and seven people were incorrectly predicted as sick when they were not patient, i.e. they were misclassified misclassification. The accuracy of this algorithm is 80.1% and reflects the accuracy of the model as a whole in classification.

Conclusion

Through machine learning techniques, we can predict the CAD coronary artery disease and thus help improve the clinical diagnosis of patients by doctors. In addition, the SLM algorithm is the best algorithm in terms of dealing with data and accuracy by predicting clinical data and thus helping doctors to diagnose the disease accurately.

Future work

Work on more tests that include more detailed data on patients such as data on patient lifestyles and exercise, may provide the use of SLM algorithm with electronic medical records in generating diagnostic results that can help doctors to predict the coronary artery disease to reach the final medical diagnostic decision for patients and thus and thus Using it as an effective technique that enables doctors to manage coronary disease correctly.

References

1. R. Alizadehsani, et al., A data mining approach for diagnosis of coronary artery disease, Comput. Methods Programs Biomed. (2013), vol.111, no.1, pp.52-61, http://dx.doi.org/10.1016/j.cmpb.2013. 03 .004,.
2. Suganyadevi, S., Seethalakshmi, V., & Balasamy, K. (2022). A review on deep learning in medical image analysis. International Journal of Multimedia Information Retrieval, 11(1), 19–38. https://doi.org/10.1007/s13735-021-00218-1.
3. Tandon, D., Rajawat, J., 2020. Present and future of artificial intelligence in dentistry. Journal of Oral Biology and Craniofacial Research. doi:10.1016/j.jobcr.2020.07.015.
4. Li, H., 2020. Impact of Artificial Intelligence Based on Big Data on Medical Care, in: Journal of Physics: Conference Series. Institute of Physics Publishing. doi:10.1088/1742-6596/1533/3/032077.
5. Ramakrishnan, R., Rao, S., He, J.R., 2021. Perinatal health predictors using artificial intelligence: A review. Women's Health. doi:10.1177/17455065211046132.
6. Dahal, K. R., & Gautam, Y. (2020). Argumentative Comparative Analysis of Machine Learning on Coronary Artery Disease. Open Journal of Statistics, 10(04), 694–705. https://doi.org/10.4236/ojs.2020.104043.
7. Adikari, D., Gharleghi, R., Zhang, S., Jorm, L., Sowmya, A., Moses, D., … Beier, S. (2022). A new and automated risk prediction of coronary artery disease using clinical endpoints and medical imaging-derived patient-specific insights: protocol for the retrospective GeoCAD cohort study. BMJ Open, 12(6), e054881. https://doi.org/10.1136/bmjopen-2021-054881.
8. Bennasar, M., Banks, D., Price, B. A., & Kardos, A. (2020). Minimal Patient Clinical Variables to Accurately Predict Stress Echocardiography Outcome: Validation Study Using Machine Learning Techniques. JMIR Cardio, 4(1). https://doi.org/10.2196/16975.
9. Manduchi, E., Le, T. T., Fu, W., & Moore, J. H. (2022). Genetic Analysis of Coronary Artery Disease Using Tree-Based Automated Machine Learning Informed By Biology-Based Feature Selection. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(3), 1379–1386. https://doi.org/10.1109/TCBB.2021.3099068.
10. Wang, C., Zhao, Y., Jin, B., Gan, X., Liang, B., Xiang, Y., … Zheng, F. (2021). Development and Validation of a Predictive Model for Coronary Artery Disease Using Machine Learning. Frontiers in Cardiovascular Medicine, 8. https://doi.org/10.3389/fcvm.2021.614204.