IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 24, 2021, accepted February 11, 2021, date of publication February 15, 2021, date of current version February 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3059469

Risk Factors and Comorbidities Associated to Cardiovascular Disease in Qatar: A Machine Learning Based Case-Control Study

HAMADA R. H. AL-ABSI[1], MAHMOUD AHMED REFAEE[1,2,3], ATIQ UR REHMAN[1], (Member, IEEE), MOHAMMAD TARIQUL ISLAM[4], SAMIR BRAHIM BELHAOUARI[1], (Senior Member, IEEE), AND TANVIR ALAM[1]

[1]College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar
[2]Geriatric Department, Hamad Medical Corporation, Doha 3050, Qatar
[3]Faculty of Medicine, Ain Shams University, Cairo 11566, Egypt
[4]Computer Science Department, Southern Connecticut State University, New Haven, CT 06515, USA

Corresponding author: Tanvir Alam (talam@hbku.edu.qa)

The open access publication of this article is funded by the Qatar National Library (QNL), Doha, Qatar. This work is in part supported by the Qatar Biobank (QBB) under Grant QF-QBB-RES-ACC-0164.

ABSTRACT Cardiovascular disease (CVD) is reported to be the leading cause of mortality in the middle eastern countries, including Qatar. But no comprehensive study has been conducted on the Qatar specific CVD risk factors identification. The objective of this case-control study was to develop machine learning (ML) model distinguishing healthy individuals from people having CVD, which could ultimately reveal the list of potential risk factors associated to CVD in Qatar. To the best of our knowledge, this study considered the largest collection of biomedical measurements representing the anthropometric measurements, clinical biomarkers, bioimpedance, spirometry, VICORDER readings, and behavioral factors of the CVD group from Qatar Biobank (QBB). CatBoost model achieved 93% accuracy, thereby outperforming the existing model for the same purpose. Interestingly, combining multimodal datasets into the proposed ML model outperformed the ML model built upon currently known risk factors for CVD, emphasizing the importance of incorporating other clinical biomarkers into consideration for CVD diagnosis plan. The ablation study on the multimodal dataset from QBB revealed that physio-clinical and bioimpedance measurements have the most distinguishing power to classify these two groups irrespective of gender and age of the participants. Multiple feature subset selection techniques confirmed known CVD risk factors (blood pressure, lipid profile, smoking, sedentary life, and diabetes), and identified potential novel risk factors linked to CVD-related comorbidities such as renal disorder (e.g., creatinine, uric acid, homocysteine, albumin), atherosclerosis (intima media thickness), hypercoagulable state (fibrinogen), and liver function (e.g., alkaline phosphate, gamma-glutamyl transferase). Moreover, the inclusion of the proposed novel factors into the ML model provides better performance than the model with traditional known risk factors for CVD. The association of the proposed risk factors and comorbidities are required to be investigated in clinical setup to understand their role in CVD better.

INDEX TERMS Cardiovascular disease, coronary heart disease, cerebrovascular disease, risk factor, machine learning, Qatar Biobank (QBB), Qatar.

I. INTRODUCTION

Cardiovascular disease (CVD) is a group of diseases that includes hypertension, coronary heart disease (CHD), heart failure, cerebrovascular accident (CVA-also known as stroke), peripheral vascular disease (PVD), ischemic heart disease, etc. and is taking a massive toll on the public health ([1], [2]). The World Health Organization (WHO) considers CVD as a primary cause of death worldwide [3]. In its 2018 report on noncommunicable diseases (NCD), the WHO

The associate editor coordinating the review of this manuscript and approving it for publication was Xianzhi Wang.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

stated that NCDs accounted for 71% of worldwide death in 2016 in which CVDs alone contributed to 44% (31%) of the total NCDs (overall global deaths) [4]. In the East Mediterranean region, this rate is even higher and CVD caused 54% of the total NCDs deaths [5]. In Gulf Cooperation Countries (GCC), including Qatar, CVD is reported to be the primary death cause [6]. According to the Ministry of Public Health (MoPH) in Qatar, CVD is the primary cause of death among Qataris and is now considered an economic burden for this small wealthy country of the gulf [7].

Risk factor identification for CVD is an active and open research area, and many studies exist along this line [8], [9] [10], [11] [12]. The INTERHEART study focused on risk factors identification for acute myocardial infarction (AMI) across 52 countries, considering more than 15000 participants from AMI and control group [13]. The study reported smoking, lipid profile, diabetes, hypertension, psychosocial factors (stress, depression, life events, and locus of control), abdominal obesity, fruits and vegetables daily intake, physical activity, and alcohol as potential risk factors associated with AMI. The first phase of the INTERSTROKE study focused on risk factors identification for intracerebral haemorrhagic and ischaemic from 22 countries [14]. In the INTERSTROKE study's second phase [15], the number of countries increased to 32 regions worldwide, and the sample size also increased to cover more than 13000 stoke cases and additional 13000 cases as control. In the second phase, hypertension, diet, diabetes, physical activity, smoking habit, alcohol consumption, apolipoprotein ratio, waist-to-hip ratio, and psychosocial factors (stress, depression) were identified as being associated with stroke. It is important to stress that both the INTERSTROKE and INTERHEART studies highlighted nine common risk factors, i.e., diabetes, hypertension, obesity, lipid profile, diet, alcohol consumption, smoking, physical activity, and psychosocial factors that, collectively, could be attributed to 86% of the CVD cases [16]. The Prospective Urban Rural Epidemiology (PURE) is another large scale study that focused on recruiting participants from low, middle, and high-income countries to study the association of social influence on lifestyle, CVD risk factors and NCD incidents [17]. Yusuf *et al.* studied modifiable risk factors on CVD and mortality based on 21 countries from the PURE study and mentioned lipid profile, blood pressure (BP), diabetes, obesity and behavioral factors (e.g., use of tobacco, diet, physical activity, alcohol, and sodium intake) were significantly associated with CVD [9]. There exist only a few studies along this line focusing on CVD risk factors identification from the Middle East countries. Alhabib *et al.* investigated a CVD cohort from Saudi Arabia and reported obesity, diabetes, hypertension, hypercholesterolemia, smoking, physical inactivity, and diabetes as potential CVD risk factors [10]. Al-Shamsi *et al.* presented a retrospective cohort study on CVD within 9-year (2009-2018) incidents and its related risk factors in the United Arab Emirates [11]. The findings from their study suggested that men, compared to women, were at greater risk of getting CVD. For both genders, elevated level of systolic BP (SBP), estimated glomerular filtration rate, Hemoglobin A1c (HbA1c) were observed. The study also suggested that age and smoking in men and total cholesterol to high-density lipoprotein (HDL) cholesterol ratio among women could be considered as potential risk factors. Recently machine learning (ML) based models have gained a lot of attention from the scientific community to predict CVD-associated risk factors. Hu *et al.* presented a study that considered health behaviors, sociodemographic factors, prevention and environmental data to predict CVD in US [12]. The authors utilized Bayesian Additive Regression Trees (BART) to rank those factors and showed that obesity, physical activity, cholesterol, binge drinking, and age (> 65 years) are the top predictors for CVD. Many CVD risk score calculators also exist in clinical setups to determine the susceptibility of CVD for the population. Framingham risk calculator considered age, gender, BP, cholesterol amount, and smoking as risk factors for CVD ([18], [19]). Atherosclerotic Cardiovascular (ASCVD) risk calculator considered age, gender, BP, cholesterol amount, diabetes, race, and smoking as risk factors for CVD [20]. In QRISK calculator, atrial fibrillation, body mass index (BMI), kidney disease, rheumatoid arthritis, ethnicity, and family history of CVD were considered along with the traditional CVD risk factors [16]. Table 1 presents a summary of the literature focusing on the identification of associated risk factors for CVD.

Considering the above backdrop, it is essential to emphasize that there exist no unanimous agreement of the risk factors for CVD and it may vary across cohorts from different ethnicity, countries and risk calculators [16]. Moreover, most of the studies conducted on CVD were based on population from North America or Europe [23], and none of the large scale comprehensive studies like INTERSTROKE, or PURE considered the Qatari cohort as part of their analysis. Based on our literature review, we found only two studies focusing on the risk factor identification for CVD considering the Qatari population. Rehman *et al.* adopted ML-based approach to report potential group of CVD risk factors in Qatar [22]. The authors designed a case-control study based on a Qatar Biobank (QBB) cohort to differentiate CVD group from the control group. Several ML models were applied, and Decision Tree achieved the highest accuracy of 82.98% on the testing set. Based on that study, Bioimpedance measurements and clinical biomarkers contributed to the better classification accuracy of the proposed model. Zainel *et al.* designed a cross-sectional study to investigate CVD risk factors in Qatar [2]. The study conducted on electronic medical records collected from the Primary Health Care Corporation health centers in Qatar, reported that modifiable CVD risk factors (e.g., diet and lifestyle) can be controlled by raising awareness among the Qatari population. The reported risk factors for CVD in the Qatari population were age (old people have high risk of developing CVD), gender (men have a high risk), nationality (Northern African, West and South Asian nationalities have high CVD risk), high BP, insulin resistance and

TABLE 1. Related studies focusing on the CVD risk factor identification in chronological order of publication year.

Year	Reference	Data used	Focused Countries in the study	Leveraged ML	Proposed CVD Risk Factors	Summary and Limitations
2004	Yusuf et al. [13]	A combination of clinical and behavioral measurements, such as smoking and physical activity.	52 countries from Europe, Asia, Middle East, Africa, Australia, North America and South America[a,b,c]	No	Smoking, lipid profile, diabetes, hypertension, psychological factors, abdominal obesity, fruits and vegetables daily intake, physical activity, and alcohol.	This study, also known as INTERHEART study, aimed at identifying risk factors associated with AMI from low- and medium-income countries. Since the focus was on AMI, this could have affected the accuracy of the collected data for participants with AMI if they have a history of a previous CVD condition which could have led to changes in their lifestyle or the risk factors' level prior to AMI.
2014	Gehani et al. [21]	Leveraged data collected from the INTERHEART study	Eight Middle Eastern Countries	No	Age, apolipoprotein (Apo)B/ApoA1 ratio, smoking, diabetes, hypertension, abdominal obesity, fruit and vegetable intake and depression.	The study used data collected in the INTERHEART study to find risk factor associated with AMI. Though the study included Qatar, yet the sample size was very small (total number of participants from three Middle eastern countries (including Qatar) was 257)
2016	O'Donnell et al. [15]	A combination of clinical and behavioral measurements such as smoking and physical activity.	32 countries from Europe, Asia, Africa, Middle East, North America and South America[a,b]	No	Hypertension, diet, diabetes, physical activity, smoking habit, alcohol consumption, apolipoprotein ratio, waist-to-hip ratio, cardiac causes, and psychosocial factors.	The study, also known as INTERSTROKE study, aimed at measuring risk factors associated with stroke from multiple regions around the world. Although several limitations of the study were addressed, however, for cardiac causes such as atrial fibrillation in low- and middle-income countries, affected cases had more measurements compared to controls which indicates that atrial fibrillation could be prevalent higher than what was recorded.
2019	Al-Shamsi et. al [11]	Patients data from Tawam hospital, UAE between April-December 2008 and yearly follow-ups till 2018.	UAE	No	Men (age, smoking), Women (Total Cholesterol to HDL-cholesterol ratio), Both sexes (SBP, estimated glomerular filtration rate, HbA1c).	This study aimed at estimating the 9-year CVD incident rate and identify associated risk factors in the UAE. Though some risk factors were identified, the study did not include important measurements such as, family history, diet, waist circumference, menopausal status, and physical activity due to their unavailability.
2020	Yusuf et al. [9]	A combination of behavioral (such as smoking, alcohol, etc.) and metabolic measurements (such as BP, obesity, etc.).	21 countries from Europe, Asia, Middle East, Africa, North America, and South America[a,b]	No	Lipid profile, blood pressure (BP), diabetes, obesity, and behavioral factors (e.g., use of tobacco, diet, physical activity, alcohol, and sodium intake).	The study, also known as PURE study, explored modifiable risk factors associated with CVD in Low-, middle- and high-income countries. However, Australia, west and north Africa was not included, and the number of middle eastern participants was small which make the study outcomes not generalizable.
2020	Hu et al. [12]	Dataset used from three different data sources: (a) e-Centers for Disease Control and Prevention, (b) American Community Survey 5-Year Estimates from the US Census Bureau and (c) Environmental Justice Screening database from the Environmental Protection Agency	USA	Yes: root-mean-squared-error was 0.47 for CHD and 0.97 for stroke	Top predictors for stroke were cholesterol, binge drinking, no physical activity, age 65 years, and dental checkup. Top predictors for CHD were cholesterol, no physical activity, binge drinking, aged 65 years, and low income.	The study focused on ranking and investigating the effect of different factors on cardiovascular health from 500 major USA cities. Sociodemographic factors, environmental factors, prevention measures, and health behaviors were considered. The study did not include a important measurements related to CVD such as BMI, family history, hip circumference etc.
2020	Alhabib et al. [10]	Demographics, health behavior, socioeconomic, medical history and data on family members living with participants from the central province of KSA	KSA	No	Hypercholesterolemia, obesity, hypertension, diabetes, smoking, physical inactivity, and diabetes	This study focused on KSA participants as part of the PURE study and aimed at finding CVD risk factors from demographic and behavioral factors. The participants of this study were recruited from the central province of KSA making it difficult to generalize the outcome to the whole population.
2020	Rehman et al. [22]	Clinical biomarkers, Bioimpedance, Spirometry and VICORDER	Qatar	Yes; 82.98% accuracy was achieved with decision tree classifier	Clinical markers and Bioimpedance show better performance, but no suggestion on the associated factor	The study aimed at differentiating CVD from non-CVD groups using different biomedical measurements from a Qatari cohort. But no specific clinical factors were suggested as part of this study.
2020	Zainel et al. [2]	Blood sugar, Blood pressure, Waist circumference, BMI, fasting triglyceride, Fasting high-density lipoprotein	Qatar	No	age, gender (men), nationality, high blood pressure, insulin resistance and low serum HDL.	This study considered data from medical records in the State of Qatar to find associated risk factors. The study is not nationally representative since Qataris and non-Qataris participants were considered.

[a] MENA region was included.
[b] At least one member of the Gulf Cooperation Countries was included
[c] Qatar was included.

low serum high-density lipoprotein (HDL). The lack of study focusing on risk factor identification for the Qatari population motivated us to explore if the current known traditional risk factors of CVD are prevalent in Qatar and, moreover, to find any other factors that could be considered as novel risk factors for the Qatari population. To fulfill this goal, we collected the largest collection of biomedical measurements from QBB to identify the associated risk factors and comorbidities for Qatari CVD cohort. The objective of this case-control study was to develop an ML model distinguishing CVD group from the control group focusing on a Qatari adult cohort to identify CVD-associated risk factors and comorbidities. The contributions of this study can be summarized as follows:

- We introduced a novel clinical dataset describing lifestyle and behavioral factors, carotid artery related measurements, vicorder, bioimpedance, spirometry and physio-clinical biomarkers from a Qatari CVD group covering more than 150 measurements.
- We proposed a highly accurate ML model to distinguish the CVD group from the control group within the dataset. To the best of our knowledge, this case-control study is the first, which considered such a large variety of clinical measurements to distinguish CVD from the control group leveraging ML techniques for Qatari nationals.
- We identified a list of potential novel risk factors which are linked to CVD comorbidities such as renal disorder, hypercoagulable state and liver function and recommended additional clinical attributes to be integrated in real life clinical setup.

The rest of this article is organized as follows: Section II provides the details of the dataset we used, data pre-processing steps, and the development of ML models. Section III presents the details of the results from different experimental setups; Section IV compares the proposed model against the existing model and the possibility of incorporating additional clinical parameters in risk score calculation. Additionally, the association of novel risk factors to other comorbidities are also highlighted. In Section V we describe the limitations of this study, and finally, Section VI concludes the article.

II. MATERIALS AND METHODS

A. ETHICAL APPROVAL

We conducted this study under the regulation of the Ministry of Public Health, Qatar. The Institutional Review Board, Hamad Medical Corporation, Qatar approved this study and only de-identified dataset was collected from QBB.

B. DATA COLLECTION

The methodology of the data collection by QBB can be found in [24], [25]. Briefly, participants were invited at QBB and they were interviewed by the clinicians and nurses to collect data on background, family history, dietary habit, and other lifestyle-related information. Then bloods samples were collected from the participants to measure different biochemical markers. For bioimpedance measurements analyzer from Seca GmbH & Co. KG, Hamburg, Germany was considered. Anthropometric measurements (e.g. hip circumference, waist circumference, height, weight, etc.) were captured using Seca Stadiometer. Pneumotrac Vitalograph from Vitalograph Ltd, Ireland was used to assess the respiratory function of the participants. VICORDER device from SMT medical GmbH & Co. KG; Bristol, UK was utilized for assessing participants' arterial stiffness. A total of 153 measurements were obtained from each participant. The dataset is not publicly available in accordance with the Qatar Biobank data-sharing policy. Table 2 summarizes the list of clinical measurements that were incorporated as part of this study.

TABLE 2. Summary of the information obtained from each participant at QBB.

Measurement Type	Total Variables	Examples
Lifestyle and habitual factors	3	Smoking habit, Sedentary Pattern
Carotid artery	4	Mean and standard deviation of intima media thickness on left and right side.
VICORDER	6	Heart rate, Heart beats, Pulse pressure index, Pulse wave.
Bioimpedance	31	Body composition measurements.
Spirometry	33	Forced expiratory flow, Forced vital capacity, Forced expiratory time, etc.
Physio-clinical Biomarkers	76	Hip waist circumference, Systolic blood pressure (SBP), Diastolic blood pressure (DBP), Cholesterol, Glucose, White blood cells, Red blood cells, Lymphocyte, Hemoglobin, Platelet count, Urea, Sodium, etc.
All	153	Smoking and sedentary habitual factors, Spirometry, Physio-clinical Biomarkers, Bioimpedance, VICORDER, Carotid artery

C. COHORT DESCRIPTION

From the QBB cohort, the CVD group was comprised of adult participants, aged above 18 years, and having self-reported history of heart attack, stroke, angina, high BP, transient cerebral ischemic attacks and abnormalities of heartbeat or heart revascularization (bypass, angioplasty, coronary atherectomy). All the CVD group participants were free from self-reported diabetes, cancer, and other diseases. As control, a group of participants with no history of CVD, stroke, diabetes, hypertension, sleep disorder, or cancer were selected. These groups were determined by QBB medical practitioners and nurses. All the participants were Qatari nationals. The cohort consisted of 500 participants where 250 participants

were from the CVD group and 250 were from the control group.

D. DATA PRE-PROCESSING

For all participants, we collected the clinical measurements (will be considered as variables/features for downstream computational analysis) from QBB. Variables with more than 30% missing values were discarded. For the remaining variables, we replaced missing values by the mean value which was calculated from the respective group. CVD and control groups were considered as the positive and negative classes respectively in the proposed classification model. The remaining variables were then normalized using the min-max method (also known as range normalization) [26]. Since all variables have values with various scale differences, we normalized them to be within the range [0,1] using the following equation:

$$x_i' = \frac{x_i - min_i\{x_i\}}{max_i\{x_i\} - min_i\{x_i\}} \tag{1}$$

where for each feature (x_i), the normalized feature (x_i') is calculated using its minimum ($min_i\{x_i\}$) and maximum ($max_i\{x_i\}$) values.

E. STATISTICAL SIGNIFICANCE OF THE CLINICAL VARIABLES

We considered the Anderson-Darling test [27] to verify if variables are normally distributed. Student's t-test [28] was applied on the normally distributed variables to determine the statistical significance of each variable (p-value<0.05) when comparing both groups (i.e., CVD vs. control). Mann-Whitney [29] test was applied on the other variables for the same purpose.

F. MACHINE LEARNING MODEL DEVELOPMENT

We considered six different ML algorithms: Decision Tree (DT) [30], Artificial Neural Network (ANN) ([31], [32]), Random Forest (RF) [33], Extreme Gradient Boosting (XGBoost) [34], CatBoost [35] and Logistic Regression (LR) [36] to separate the CVD group from the control group. For all models, we used GridSearchCV from Python's Scikit-learn package for hyperparameter tuning with 10-fold nested cross validation [37]. GridSearchCV finds the parameters that contribute to the best performance from a set of given ranges of values for some parameters [38]. The details of parameter optimization for the models are provided in Supplementary File 1. Figure 1 shows the workflow adopted for this study to select the features and compare ML models' performances.

G. MODEL EVALUATION

The training and validation of the ML models were performed considering 10-fold nested cross validation approach [37]. Therefore, 90% of the available data was used for training and 10% thereof were used for independent testing. This was repeated for ten times following the principle of nested cross validation. The performance of different machine learning models was analyzed with several metrics for performance evaluation (Eq.2 – Eq.6): accuracy, sensitivity, precision, F_1 score and Matthews Correlation Coefficient (MCC).

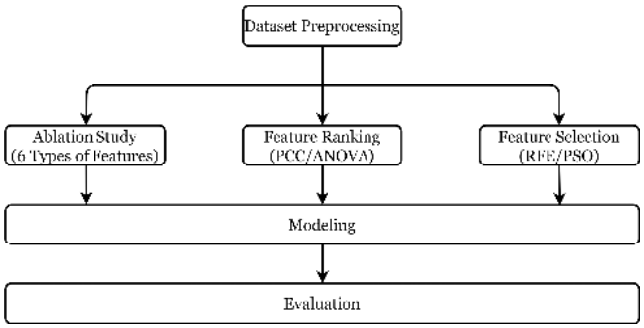

FIGURE 1. Workflow adopted in this study to discover the important features and ML model development.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{2}$$

$$Sensitivity(recall) = \frac{tp}{tp + fn} \tag{3}$$

$$Precision = \frac{tp}{tp + fp} \tag{4}$$

$$F_1 score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5}$$

$$MCC = \frac{tp * tn - fp * fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{6}$$

Here tp, tn, fp, and fn stand for the number of true positive, true negative, false positive, and false negative samples, respectively. As our dataset was balanced, we considered accuracy as the prominent metric for performance comparison of different ML models.

H. FEATURE SUBSET SELECTION

We applied various feature subset selection techniques to filter out redundant or un-correlated features. We applied both the Pearson correlation coefficient and Spearman correlation coefficient on the features against the class label (positive class: CVD, negative class: control). The range of correlation values from Pearson correlation coefficient and Spearman correlation coefficient were (−0.49, 0.46) and (−0.48, 0.48), respectively (Supplementary File 2). This indicated that none of the individual variable had high correlation with the class label. Therefore, we considered all the features in the downstream computational analysis and ranked them based on Pearson correlation coefficient and Spearman correlation coefficient. We considered ANOVA F-test to rank the features and determine their relative importance in distinguishing CVD group from the control group. We also applied recursive feature elimination (RFE) and particle swarm optimization (PSO) based method with cross validation to identify a subset of features from the pool of all features. For RFE, we used RF as an estimator and evaluated the model on all

features one by one to select only features that contributed best to achieve the highest accuracy. For PSO based feature subset selection, we evaluated the performance with different classifiers (i.e., RF, DT, XGBoost, and ANN) on the whole dataset [39].

III. RESULTS

A. BASELINE CHARACTERISTICS OF THE COHORT

The baseline characteristics as well as their smoking habit and sedentary lifestyle pattern of the cohort are summarized in Table 3. In the CVD group, 54.8% (45.2%) of participants were male (female); where in the control group, the samples were distributed evenly (i.e., 50% each). The mean of the age of the CVD group was 42 where it was 30.276 for the control group. Furthermore, all participants used in this study were Qatari nationals. The mean BMI for the CVD group was 26.172 where it is 23.29 for the control group (p-value <0.001); this indicates that the CVD group participants were having higher weight in comparison with the control participants [40]?. Waist and hip sizes were elevated in the CVD group compared to the control group. CVD group was spending more time in sports, but alarmingly, they are still smoking more than the control group. This is in perfect align with the other Qatari CVD cohort that we studied earlier [41].

TABLE 3. Baseline characteristics for the CVD group and the control group. (FMI stands for fat mass index).

Attribute	Unit	CVD		Control		p-value
		Mean	Standard Deviation	Mean	Standard Deviation	
Age	year	42	13.144	30.276	8.338	<.001
BMI	kg/m2	26.127	2.867	23.29	2.817	<.001
Weight	kg	72.375	11.347	64.235	10.774	<.001
Z-FMI	-	0.536	0.78	-0.243	0.709	<.001
Waist circumference	cm	0.839	0.113	0.75	0.086	<.001
Hip circumference	cm	84.268	10.369	75.024	8.599	<.001
HbA1c	mmol/mol	5.278	0.397	5.146	0.27	<.001
Number of cigarettes/ cigars/pipes per day	-	17.367	10.961	14.747	9.269	0.004
Moderate Sport (swimming, yoga, gym, etc.)	Minutes	71.132	182.406	101.128	217.137	0.03
Heavy Sport (running, fast swimming, weightlifting, etc.)	Minutes	47.184	142.476	70.888	163.497	0.017

B. PERFORMANCE OF ML MODELS BASED ON ABLATION STUDY

To check the effectiveness of different types of clinical measurements, we developed machine learning models considering individual data categories as well as their combinations. Based on ablation study, the contribution of features representing sedentary and smoking habits, Spirometry, VICORDER were not high (Table 4). Considering carotid artery related measurements (e.g., intima medium thickness), the models achieved nearly 75% accuracy. But the most dominant set of features were the set of physio-clinical biomarkers and Bioimpedance achieving over 84% and 91% accuracy, respectively (Table 4). Interestingly, when the models were developed based on all clinical features, sedentary lifestyle and smoking habit related features, we obtained the best performance with over 92% accuracy (using DT and XGBoost) (Table 4). We also developed ML models with the known CVD risk factors. However, the highest achieved accuracy was slightly over 82% (82.6% with XGBoost; 82.4% with DT and CatBoost) (Table 4). This clearly indicates the superiority of our model developed by integrating variety of clinical measurements, rather than relying upon the known risk factors that are heavily used in current clinical setup for the diagnosis plan of CVD.

TABLE 4. Performance of ML models for ablation study. Numbers in bold highlight the highest value of evaluation metric considering the specific experimental setup.

Property (No. of features)	Evaluation Parameter	DT	ANN	RF	XGBoost	CatBoost	LR
Sedentary lifestyle and Smoking (3)	Accuracy	0.52	0.546	0.522	0.492	0.516	**0.548**
	Precision	0.510	0.499	0.519	0.490	0.510	0.517
	Recall	0.728	0.708	0.648	0.720	0.668	0.692
	F1-score	0.598	0.555	0.575	0.583	0.575	0.576
	MCC	0.051	0.051	0.045	-0.012	0.037	0.114
Spirometry (33)	Accuracy	0.584	0.602	0.616	0.594	0.604	**0.682**
	Precision	0.583	0.543	0.633	0.599	0.625	0.685
	Recall	0.596	0.696	0.564	0.584	0.544	0.688
	F1-score	0.585	0.600	0.593	0.591	0.578	0.686
	MCC	0.171	0.208	0.236	0.189	0.212	0.366
VICORDER (6)	Accuracy	0.646	0.624	0.648	0.636	0.640	**0.664**
	Precision	0.685	0.440	0.667	0.648	0.667	0.683
	Recall	0.580	0.396	0.608	0.608	0.564	0.616
	F1-score	0.616	0.415	0.632	0.625	0.608	0.646
	MCC	0.304	0.251	0.300	0.274	0.285	0.331
Carotid artery (4)	Accuracy	**0.742**	0.696	0.730	0.724	0.738	0.732
	Precision	0.827	0.684	0.811	0.819	0.827	0.758
	Recall	0.616	0.632	0.612	0.588	0.612	0.688
	F1-score	0.705	0.646	0.695	0.68	0.699	0.718
	MCC	0.503	0.398	0.478	0.471	0.497	0.469
Physio-clinical Biomarker (76)	Accuracy	0.790	0.750	0.822	0.844	**0.846**	0.800
	Precision	0.847	0.707	0.859	0.888	0.901	0.816
	Recall	0.716	0.744	0.772	0.792	0.784	0.780
	F1-score	0.773	0.715	0.812	0.835	0.834	0.793
	MCC	0.591	0.503	0.65	0.697	0.703	0.605
Bioimpedance (31)	Accuracy	0.878	0.812	0.902	0.892	**0.910**	0.892
	Precision	0.920	0.868	0.929	0.927	0.968	0.938
	Recall	0.832	0.852	0.872	0.852	0.848	0.840
	F1-score	0.871	0.838	0.899	0.888	0.903	0.886
	MCC	0.762	0.632	0.807	0.787	0.827	0.790
All Clinical + Sedentary + Smoking (150+2+1=153)	Accuracy	**0.926**	0.872	0.916	0.922	0.914	0.882
	Precision	0.986	0.898	0.947	0.977	0.977	0.916
	Recall	0.864	0.840	0.884	0.864	0.848	0.844
	F1-score	0.921	0.867	0.915	0.918	0.907	0.878
	MCC	0.859	0.747	0.835	0.850	0.836	0.767
Known risk factors: Age, Gender, SBP, DBP, BMI, lipid profile, HbA1c, Sedentary lifestyle and smoking (13)	Accuracy	0.824	0.672	0.812	**0.826**	0.824	0.740
	Precision	0.989	0.537	0.904	0.946	0.972	0.761
	Recall	0.656	0.500	0.704	0.692	0.668	0.708
	F1-score	0.785	0.517	0.789	0.797	0.79	0.731
	MCC	0.689	0.347	0.642	0.679	0.683	0.483

C. PERFORMANCE OF ML MODELS BASED ON AGE- AND GENDER-STRATIFIED SAMPLES

Considering all the features, we developed ML models on age- and gender-stratified samples. Among the ML models, CatBoost achieved the highest accuracy of 93.9% and 88.2% for distinguishing CVD group form the control group for male and female participants, respectively (Figure 2). This indicates that the ML models were able to perform slightly better for males than females considering gender-stratified samples.

Based on the age-stratified samples, ML models achieved the highest MCC of 0.82, and 0.86 for the middle-aged (30 to 39 yrs), and senior (40 and above) groups, respectively (Figure 3). This is very close the performance of the models considering the full dataset achieving 0.85 MCC (Table 4). But the performance of the ML models for the young adult group (18 to 29 yrs) were relatively lower (MCC = 0.66).

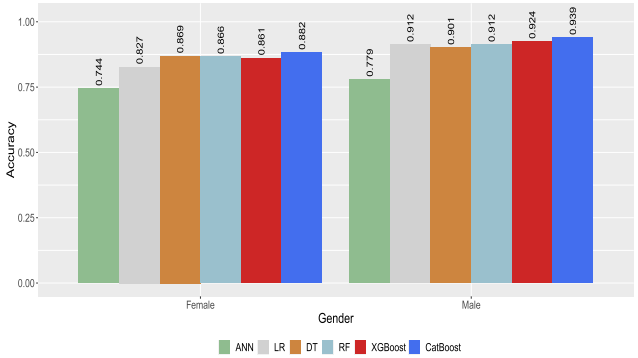

FIGURE 2. Performance of different ML models based on gender-stratified groups.

This might be due to the few CVD cases (only 43 in CVD compared to 134 in the control group) in this age group (Supplementary File 3), and ML models could not learn from such a small number of samples. We may conclude the ML models were more effective in detecting middle-aged and senior CVD participants than the small number of young CVD participants.

D. PERFORMANCE OF ML MODELS BASED ON SELECTED SUBSET OF FEATURES

To filter out redundant features, we used multiple feature subset selection techniques to identify a smaller number of features which can represent the whole dataset. Considering the ANOVA F-test based feature ranking, we developed ML models using top ranked 1%, 5%, 10%, 15%, 20%, 25% and 30% features. Using the top 5% and 10% features, we achieved 92.8% accuracy using DT model (Figure 4). When RFE and PSO was applied on the dataset, a subset of 10 and 9 features were selected, respectively (Supplementary File 2). Considering the selected features from RFE and PSO, we achieved the highest accuracy of 92.4% and 92.6%, respectively (Table 5). After combining the feature subsets returned by RFE and PSO we got in total 16 features and slightly better level of accuracy (93.0%) was achieved with RF and CatBoost (Table 5).

TABLE 5. Performance of ML models on the selected features by different methods.

Method	Evaluation Parameter	DT	ANN	RF	XGBoost	CatBoost	LR
ANOVA (Top 5 %)	Accuracy	**0.928**	0.800	0.918	0.926	0.916	0.864
	Precision	0.991	0.799	0.964	0.986	0.973	0.897
	Recall	0.864	0.748	0.868	0.864	0.856	0.828
	F1-score	0.922	0.758	0.912	0.920	0.909	0.858
	MCC	0.865	0.609	0.842	0.860	0.840	0.734
RFE (10 Features)	Accuracy	0.922	0.770	**0.924**	0.916	0.920	0.872
	Precision	0.982	0.700	0.975	0.974	0.980	0.916
	Recall	0.860	0.688	0.872	0.856	0.860	0.828
	F1-score	0.915	0.682	0.918	0.908	0.913	0.864
	MCC	0.852	0.546	0.855	0.840	0.849	0.754
PSO (9 Features)	Accuracy	0.918	0.880	0.924	0.918	**0.926**	0.816
	Precision	0.974	0.924	0.971	0.953	0.976	0.836
	Recall	0.860	0.828	0.876	0.880	0.876	0.792
	F1-score	0.911	0.872	0.919	0.915	0.922	0.811
	MCC	0.844	0.765	0.855	0.840	0.859	0.636
RFE & PSO (16 Features)	Accuracy	0.924	0.820	**0.930**	0.926	**0.930**	0.874
	Precision	0.983	0.860	0.971	0.979	0.991	0.918
	Recall	0.864	0.876	0.888	0.872	0.868	0.828
	F1-score	0.919	0.850	0.927	0.922	0.925	0.868
	MCC	0.856	0.645	0.865	0.859	0.868	0.755

Among the 16 features selected by RFE and PSO, three features, namely BMI, SBP and bioimpedance hydration are well known risk factors of CVD, and were selected by both methods (Supplementary File 2). In addition to those three features, PSO selected bicarbonate, beats, spirometry forced expiratory time as well as bioimpedance vector analysis (which measures body composition) and bioimpedance phase angle. On the other hand, RFE also selected free triiodothyronine, mean intima media thickness (left and right) and bioimpedance measurements for fat mass index, visceral adipose, and energy. Figure 5 shows the biplot of the first two components of principal component analysis (PCA) based on the 16 features selected by RFE and PSO. The two principal components explained nearly 47% of the variance from the selected features (Figure 5). The vectors (direction) in Figure 5 indicates the high correlation between BMI and intima media thickness-right; bioimpedance energy

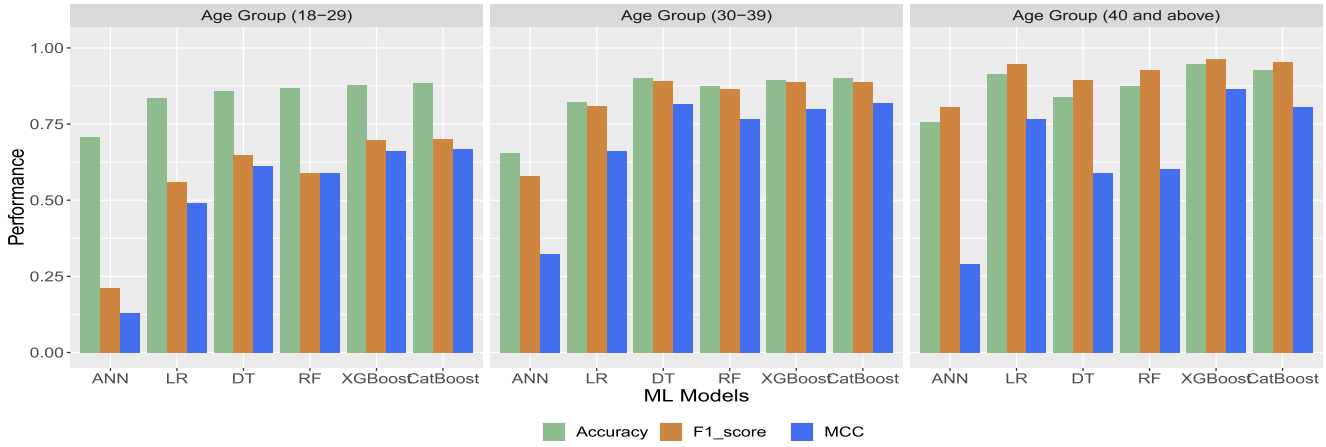

FIGURE 3. Performance of different ML models based on age-stratified groups.

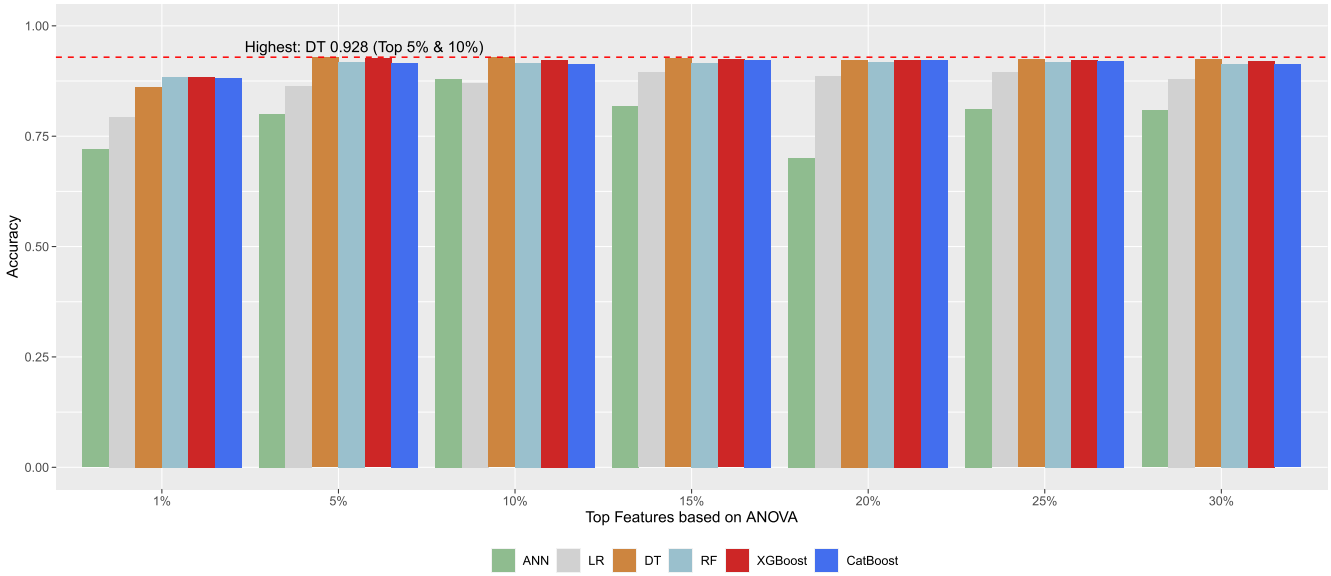

FIGURE 4. Top ranked (1% to 30%) feature selected by ANOVA F-test. Best accuracy (92.8%) was achieved by DT considering top ranked 5% and 10%.

content and fat mas index; SBP and bioimpedance visceral adipose tissue. On the other hand, the direction of free triiodothyronine is nearly opposite to BMI and intima media thickness-right. There was little or no correlation between free triiodothyronine and bioimpedance phase angle; bioimpedance hydration and fat mass index.

E. STATISTICAL SIGNIFICANCE OF THE FEATURES BY COMPARING CVD GROUP AGAINST CONTROL GROUP

Based on statistical analysis, 38 physio-clinical biomarkers were found to be statistically significant (p-value < 0.05) (Supplementary File 2). This list includes known risk factors for CVD such as: BMI, HbA1c, total cholesterol, low-density lipoprotein (LDL)-cholesterol, high-density lipoprotein (HDL)-cholesterol, SBP, and DBP. The total cholesterol and the HDL-cholesterol (good cholesterol) readings form CVD and the control group indicate that they were within the standard reference range (i.e., reference range for total cholesterol < 5.2 mmol/L; HDL-cholesterol >1.04 mmol/L (male) and > 1.30 mmol/L (female)) [42]. However, comparing CVD group against the control group (CVD: control) total cholesterol (4.93:4.70) and LDL-cholesterol(2.96: 2.78) was higher, but HDL-cholesterol (1.41:1.47) was lower in CVD group. Though the BP was within a reference range for both groups (SBP > 140 and DBP > 90 is considered as high [43]), SBP (119:104.58) and DBP (71.36:62.38) were higher in CVD group compared to the control group. Also, we found novel factors such as triglycerides, potassium, bicarbonate, C-peptide, fibrinogen, uric acid, homocysteine, free triiodothyronine, total protein, albumin, etc. were significantly (p-value< 0.05) different between two groups.

We also discovered in total 26 statistically significant (p-value< 0.05) features from bioimpedance measurement (Supplementary File 2). A closer look at the features revealed that almost all the bioimpedance measurements, except hydration, were higher in the CVD group. Average fat mass (24.49:18.38), relative fat mass (33.76: 28.573), visceral adipose tissue (1.95:1.18), waist circumference (0.839: 0.75), phase angle (27.36:23.58) were higher in the CVD group. Almost all of the spirometry measurements related to forced expiratory volume (FEV) were low in the CVD group (Supplementary File 2) and this is in alignment with the outcome of a recent study where FEV1 was linked with CVD and mortality [44].

In summary, the results indicate that CatBoost was the best performing model of all models tested considering MCC (.868) and accuracy (93%) as the evaluation metrics. We also found that the tested ML models attained a reasonable performance on the task when applied to age- and gender-stratified samples. Additionally, the ablation study we performed was insightful in revealing the most dominant feature sets for the CVD risk factor prediction are physio-clinical biomarkers and Bioimpedance, which is also corroborated by our analysis revealing statistically significant (p-value<0.05) 38- and 26-count of these features.

IV. DISCUSSION

A. COMPARISON OF THE PROPOSED ML MODEL's PERFORMANCE AGAINST THE EXISTING MODEL

A previous study proposed a decision tree based ML model to differentiate the Qatari CVD group from the control group with 82.98% accuracy [22]. In this study, we outperformed the previous model by achieving 93% accuracy (Table 5). Among the ML models we investigated, CatBoost

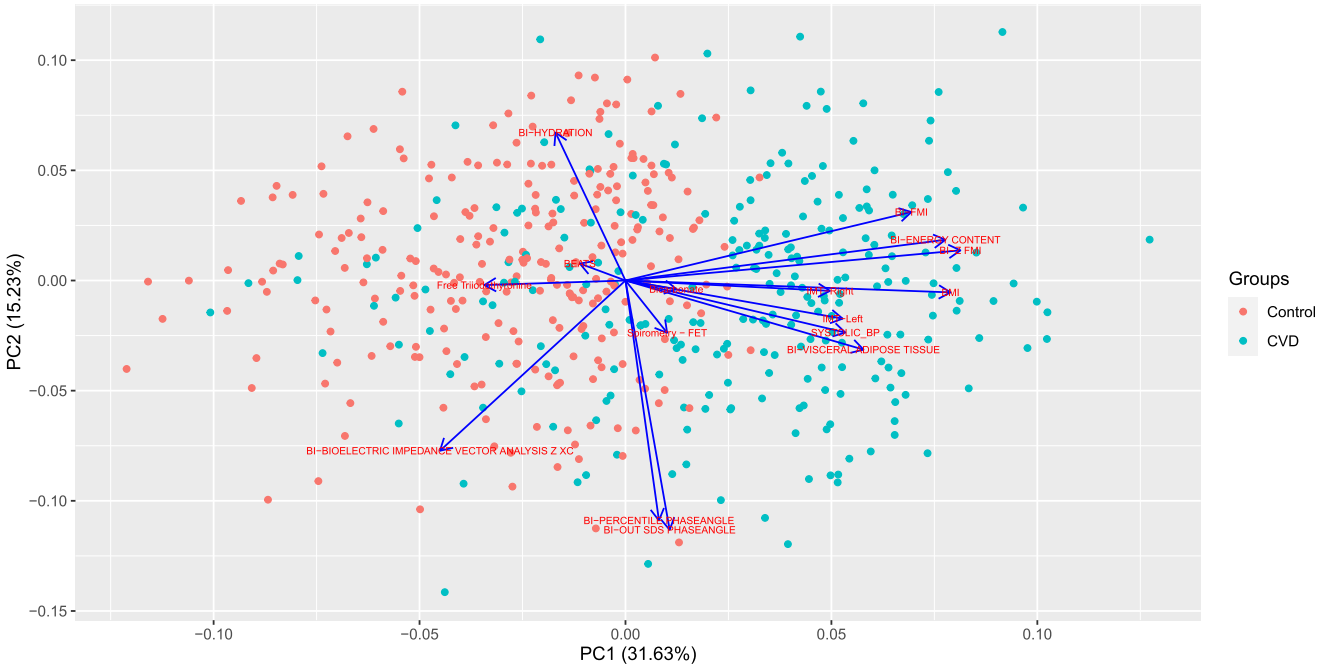

FIGURE 5. PCA biplot for the selected features by RFE and PSO. BI: Bioimpedance. IMT: Intima media thickness.

model performed the best considering multiple evaluation metrics such as accuracy and MCC. We also showed that the performance of the proposed ML models for age- and gender-stratified samples were very similar to the model developed without stratification, and this clearly indicates that proposed methodology fits well independent of the gender and age of the CVD participants. Unlike the previous study [22], we applied feature subset selection on the dataset. Application of multiple feature ranking methods (e.g., Pearson correlation coefficient, Spearman correlation coefficient, and ANOVA) and feature subset selected methods (e.g., RFE and PSO) revealed a shorter list of 16 features from the pool of all features.

B. APPLICABILITY OF EXISTING CVD RISK SCORE CALCULATION METHODS FOR QATARI CVD PATIENTS

Both the Framingham scale ([18], [19]) and ASCVD scale ([20]) considered age, gender, smoking habit, diabetes status, BP, cholesterol value to calculate the risk score for CVD. Interestingly, statistically significant difference was observed for all these variables in the Qatari CVD group compared to the control group (Supplementary File 2). Diabetes status (in terms of HbA1c), BP, cholesterol levels all were high in the Qatari CVD group compared to the control group (Supplementary File 2). This indicates that the existing CVD risk score calculation mechanism could be applied for Qatari CVD groups as well, though the effectiveness of different multiple risk score calculation scales is out of the scope of this study. To check the effectiveness of the proposed clinical measurements from this study, we also compared the proposed ML model against the ML model with traditional known risk factors for CVD such as: smoking habit, level of physical activity, diabetes (HbA1c), BMI, cholesterol and BP. ML models considering traditional know risk factors achieved 82.6% accuracy (Table 4). On the other hand, after the inclusion of the clinical measurements proposed in this study, CatBoost model achieved 93% accuracy (Table 5). This clearly indicates the superiority of the proposed model as well as the importance of integrating other measurements in clinical setup for CVD diagnosis plan.

C. ASSOCIATION OF POTENTIAL RISK FACTORS AND COMORBIDITIES FOR CVD IN QATARI POPULATION

Our study confirmed the known risk factors (e.g., BMI, SBP, diastolic BP (DBP), low-density lipoproteins (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, diabetes marker (HbA1c)) for CVD that are used in multiple risk score calculators (e.g., Framingham and ASCVD) [16]. SBP, DBP, total cholesterol, LDL-cholesterol were higher in CVD compared to the control group. In addition, the triglycerides outcome (CVD: Control = 1.23:0.99) indicates desirable results (<1.7 mmol/L) [42] for both groups?. This, in summary, indicates that CVD group in Qatar needs to improve their lipid profile and it is in perfect alignment with the other CVD groups [45].

BMI, cholesterol, bioimpedance, hip and waist circumference, measurements for absolute fat mass, visceral adipose tissue, etc., were higher in CVD and these all can be well linked to fat disposition in the body and clear indicator of their role in the atherosclerosis. Due to the high body fat content

and low level of hydration, clot could be formed easily in the CVD group [46]. Moreover, high level of HbA1c and C-peptide (an indirect measurement of the insulin level in the body and has longer half-life in comparison to insulin) reflects insulin resistance among CVD group which is a risk factor for atherosclerosis and CVD. Low level of free tri-iodothyronine observed in the CVD group can be associated with a lower metabolic rate and this may affect the people to be overweight or obese as well.

It is well known that activated partial thromboplastin time, fibrinogen, and platelet contribute to the formation of clot [47]. Our results showed that fibrinogen was higher but activated partial thromboplastin time and platelet were lower in the CVD group compared to the control group. Therefore, for the Qatari population, it could be hypothesized that the role of activated partial thromboplastin time and fibrinogen towards leading to CVD is higher than the role of platelet, however further study is required in this regard. Hypercoagulable state indicated by higher fibrinogen level and lower activated partial thromboplastin time may increase the clotting likelihood in the blood vessel. The intima media thickness for both the left and right coronary artery was higher in CVD group. This clearly indicates that thickness of intima will shrink the lumen area for carotid artery. This reflects the atherosclerotic condition of whole body, including renal arteries; and may, ultimately, affect the renal function as well [48]. Uric acid, homocysteine and creatinine all were high in the CVD group. Creatinine is one of the known biomarkers for renal function. Homocysteine is known to be associated with CVD [49] and uric acid may cause gout which is also a risk factor of CVD [50]. As total protein was high and albumin was low in the CVD group compared to the control group, this may due to the possible loss of albumin in urine as a result of chronic kidney disease complications [51]. All these factor, in summary, may indicate the comorbidity of CVD group for chronic kidney disease [52].

Potassium, bicarbonate, alkaline phosphate, gamma-glutamyl transferase level were higher in the studied CVD group compared to the control group. High potassium level tend to lower the risk of CVD, such as stroke [53], indicating that either the food intake by the CVD group was rich in potassium or they were taking CVD medications which are known to increase serum potassium. Higher bicarbonate level ($>$ 26 which is similar to this study) is associated with CVD mortality rate [54]. Elevated alkaline phosphate and gamma-glutamyl transferase in the studied CVD group may deranges the liver function ([55], [56]), however further study is required in this regard.

In summary, our findings revealed that while the standard CVD risk score calculation scales such as Framingham, ASCVD, etc. could be used for effective CVD prediction in Qatari cohort, incorporating the novel risk factors discovered in this study into these schemes could significantly improve the performance of the CVD detection task. Additionally, the proposed novel risk factors from this study could be shown to be linked to multiple CVD-related comorbidities.

V. LIMITATIONS

Data set was relatively small having 500 participants from Qatar only. So, the obtained results might be specific to Qatari population only and this was one of the goals of this study. We did not consider deep learning-based techniques such as convolutional neural network, recursive neural networks etc. for this study because the dataset we had from QBB is tabular in nature, whereas the most popular deep learning-based models, such as convolutional neural network and recursive neural networks, are mostly suitable for image datasets and sequence datasets, respectively. However, the interpretability easiness of the models and more than 92% accuracy level motivated us to rely on the models that we proposed. In the future, this work will be extended on a larger cohort size from Qatar.

VI. CONCLUSION

This work presented a case-control study to develop ML models to differentiate CVD group from the control group. We used a multimodal dataset from QBB covering more than 150 attributes from various clinical measurements for the Qatari population. A 93% accuracy was achieved by the proposed CatBoost model which outperformed the existing model developed for the same purpose. Multiple feature ranking methods confirmed known risk factors of CVD such BP, lipid profile, smoking, sedentary life, diabetes and proposed novel risk factors linked to CVD-related comorbidities such as renal disorder, liver function, atherosclerosis etc.

SUPPLEMENTARY FILES

1) Supplementary File 1: List of parameters that were optimized for the development of ML models.
2) Supplementary File 2: List of features, their mean, standard deviation and statistical significance. We also highlighted the features selected by different feature subset selection techniques
3) Supplementary File 3: ML model performance details for age-stratified samples.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS's CONTRIBUTIONS

Conceived and designed the experiments: HRHA,TA. Performed the experiments: HRHA. Analyzed the data: HRHA. Wrote the manuscript: HRHA, MAR, TA with the support from other authors. All authors read and approved the final manuscript.

ACKNOWLEDGMENT

The authors thank Dr. Rezaul Karim for his valuable comments on the manuscript.

LIST OF ABBREVIATIONS

Acute Myocardial Infarction	AMI
Artificial Neural Network	ANN
Atherosclerotic Cardiovascular	ASCVD
Body Mass Index	BMI
Blood Pressure	BP
Cerebrovascular Disease	CBVD
Coronary Heart Disease	CHD
Cerebrovascular Accident	CVA
Cardiovascular disease	CVD
Diastolic Blood Pressure	DBP
Decision Tree	DT
Gulf Cooperation Countries	GCC
Hemoglobin A1c	HbA1c
High-Density Lipoprotein	HDL
Institutional Review Board	IRB
Low-Density Lipoprotein	LDL
Logistic Regression	LR
Matthews Correlation Coefficient	MCC
Machine Learning	ML
Noncommunicable Diseases	NCD
Particle Swarm Optimization	PSO
Prospective Urban Rural Epidemiology	PURE
Peripheral Vascular Disease	PVD
Qatar Biobank	QBB
Random Forest	RF
Recursive Feature Elimination	RFE
Systolic Blood Pressure	SBP
World Health Organization	WHO
Extreme Gradient Boosting	XGBoost

REFERENCES

[1] D. F. López-Cevallos, G. Escutia, Y. Gonzalez-Pena, and L. I. Garside, "Cardiovascular disease risk factors among Latino farmworkers in Oregon," *Ann. Epidemiol.*, vol. 40, pp. 8–12, Dec. 2019.

[2] A. J. A. L. Zainel, A. S. A. Nuaimi, and M. A. Syed, "Risk factors associated with cardiovascular diseases among adults attending the primary health care centers in Qatar a cross sectional study," *J. Community Med. Public Health*, vol. 4, no. 1, 2020.

[3] WH Organization. (2020). *Cardiovascular Diseases*. Accessed: Sep. 21, 2020. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases/

[4] "Noncommunicable diseases country profiles 2018," World Health Org., Geneva, Switzerland, Tech. Rep., 2018.

[5] WHO R Office for the Eastern Mediterranean. (2020). *Cardiovascular Diseases*. Accessed: Sep. 24, 2020. [Online]. Available: http://www.emro.who.int/health-topics/cardiovascular-diseases/index.html

[6] S. K. Al-Kaabi and A. Atherton, "Impact of noncommunicable diseases in the state of Qatar," *ClinicoEconomics Outcomes Res.*, vol. 7, p. 377, Jul. 2015.

[7] M Public Health. *Qatar Public Health Strategy 2017-2022*. Accessed: Sep. 22, 2020. [Online]. Available: https://www.moph.gov.qa/english/strategies/Supporting-Strategies-and-Frameworks/QatarPublicHealthStrategy/Pages/default.aspx

[8] J. C. Brown, T. E. Gerhardt, and E. Kwon, "Risk factors for coronary artery disease," in *StatPearls*. 2020.

[9] S. Yusuf *et al.*, "Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): A prospective cohort study," *Lancet*, vol. 395, no. 10226, pp. 795–808, Mar. 2020.

[10] K. F. Alhabib, M. A. Batais, T. H. Almigbal, M. Q. Alshamiri, H. Altaradi, S. Rangarajan, and S. Yusuf, "Demographic, behavioral, and cardiovascular disease risk factors in the Saudi population: Results from the prospective urban rural epidemiology study (PURE-Saudi)," *BMC Public Health*, vol. 20, no. 1, pp. 1–14, Dec. 2020.

[11] S. Al-Shamsi, D. Regmi, and R. D. Govender, "Incidence of cardiovascular disease and its associated risk factors in at-risk men and women in the United Arab Emirates: A 9-year retrospective cohort study," *BMC Cardiovascular Disorders*, vol. 19, no. 1, p. 148, Dec. 2019.

[12] L. Hu, B. Liu, and Y. Li, "Ranking sociodemographic, health behavior, prevention, and environmental factors in predicting neighborhood cardiovascular health: A Bayesian machine learning approach," *Preventive Med.*, vol. 141, Dec. 2020, Art. no. 106240.

[13] S. Yusuf, S. Hawken, S. Ôunpuu, T. Dans, A. Avezum, F. Lanas, M. McQueen, A. Budaj, P. Pais, J. Varigos, and L. Lisheng, "Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): Case-control study," *Lancet*, vol. 364, no. 9438, pp. 937–952, Sep. 2004.

[14] M. J. O'Donnell *et al.*, "Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): A case-control study," *Lancet*, vol. 376, no. 9735, pp. 112–123, Jul. 2010.

[15] M. J. O'Donnell *et al.*, "Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): A case-control study," *Lancet*, vol. 388, no. 10046, pp. 761–775, Aug. 2016.

[16] N. Garg, S. K. Muduli, A. Kapoor, S. Tewari, S. Kumar, R. Khanna, and P. K. Goel, "Comparison of different cardiovascular risk score calculators for cardiovascular risk prediction and guideline recommended statin uses," *Indian Heart J.*, vol. 69, no. 4, pp. 458–463, Jul. 2017.

[17] K. Teo, C. K. Chow, M. Vaz, S. Rangarajan, and S. Yusuf, "The prospective urban rural epidemiology (pure) study: Examining the impact of societal influences on chronic noncommunicable diseases in low-, middle-, and high-income countries," *Amer. Heart J.*, vol. 158, no. 1, pp. 1–7, 2009.

[18] P. A. Lotufo, "O escore de risco de Framingham para doenças cardiovasculares," *Revista de Medicina*, vol. 87, no. 4, pp. 232–237, 2008.

[19] E. J. Cesarino, A. L. G. Vituzzo, J. M. C. Sampaio, D. A. S. Ferreira, H. A. F. Pires, and L. D. Souza, "Assessment of cardiovascular risk of patients with arterial hypertension of a public health unit," *Einstein (São Paulo)*, vol. 10, no. 1, pp. 33–38, Mar. 2012.

[20] AC Cardiology. *ASCVD Risk Estimator*. Accessed: Oct. 2020. [Online]. Available: http://tools.acc.org/ASCVD-Risk-Estimator-Plus/#!/calculate/estimate/

[21] A. A. Gehani, A. T. Al-Hinai, M. Zubaid, W. Almahmeed, M. R. M. Hasani, A. H. Yusufali, M. O. Hassan, B. S. Lewis, S. Islam, S. Rangarajan, S. Yusuf, and The INTERHEART Investigators in Middle East, "Association of risk factors with acute myocardial infarction in middle eastern countries: The INTERHEART middle east study," *Eur. J. Preventive Cardiol.*, vol. 21, no. 4, pp. 400–410, Apr. 2014.

[22] A. U. Rehman, T. Alam, and S. B. Belhaouari, "Investigating potential risk factors for cardiovascular diseases in adult Qatari population," in *Proc. IEEE Int. Conf. Informat., IoT, Enabling Technol. (ICIoT)*, Feb. 2020, pp. 267–270.

[23] K. Steyn, K. Sliwa, S. Hawken, P. Commerford, C. Onen, A. Damasceno, S. Ounpuu, and S. Yusuf, "Risk factors associated with myocardial infarction in Africa: The INTERHEART Africa study," *Circulation*, vol. 112, no. 23, pp. 3554–3561, Dec. 2005.

[24] H. Al Kuwari, A. A. Thani, A. A. Marri, A. Al Kaabi, H. Abderrahim, N. Afifi, F. Qafoud, Q. Chan, I. Tzoulaki, P. Downey, H. Ward, N. Murphy, E. Riboli, and P. Elliott, "The Qatar biobank: Background and methods," *BMC Public Health*, vol. 15, no. 1, p. 1208, Dec. 2015.

[25] A. A. Thani, E. Fthenou, S. Paparrodopoulos, A. A. Marri, Z. Shi, F. Qafoud, and N. Afifi, "Qatar biobank cohort study: Study design and first results," *Amer. J. Epidemiol.*, vol. 188, no. 8, pp. 1420–1433, Aug. 2019.

[26] M. J. Zaki and W. Meira, *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[27] T. W. Anderson and D. A. Darling, "A test of goodness of fit," *J. Amer. Stat. Assoc.*, vol. 49, no. 268, pp. 765–769, 1954.

[28] Student, "The probable error of a mean," *Biometrika*, vol. 6, no. 1, pp. 1–25, Mar. 1908.

[29] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, Mar. 1947.

[30] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.
[31] W. Huang and R. P. Lippmann, "Neural net and traditional classifiers," in *Proc. Neural Inf. Process. Syst.*, 1987, pp. 387–396.
[32] G. E. Hinton, "Connectionist learning procedures," in *Machine Learning*. Amsterdam, The Netherlands: Elsevier, 1990, pp. 555–610.
[33] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[34] J. Brownlee, *XGBoost With Python: Gradient Boosted Trees With XGBoost and Scikit-Learn*. Australia: Machine Learning Mastery, 2016.
[35] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: Gradient boosting with categorical features support," 2018, *arXiv:1810.11363*. [Online]. Available: http://arxiv.org/abs/1810.11363
[36] J. I. Hoffman, "Logistic regression," in *Basic Biostatistics for Medical and Biomedical Practitioners*, J. I. Hoffman, Ed., 2nd ed. New York, NY, USA: Academic, 2019, ch. 33, pp. 581–589. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128170847000334
[37] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, Jul. 2010.
[38] B.-M. Hsu, "Comparison of supervised classification models on textual data," *Mathematics*, vol. 8, no. 5, p. 851, May 2020.
[39] M. Ashikur, M. Arifur, and J. Ahmed, "Automated detection of diabetic retinopathy using deep residual learning," *Int. J. Comput. Appl.*, vol. 177, no. 42, pp. 25–32, Mar. 2020. [Online]. Available: https://www.ijcaonline.org/archives/volume177/number42/31185-2020919927
[40] C. for Disease Control and Prevention. *Body Mass Index (BMI)*. Accessed: Sep. 7, 2020. [Online]. Available: http://tools.acc.org/ASCVD-Risk-Estimator-Plus/#!/calculate/estimate/
[41] H. R. Al-Absi, M. A. Refaee, A. Nazeemudeen, M. Househ, Z. Shah, and T. Alam, "Cardiovascular diseases in Qatar: Smoking, food habits and physical activities perspectives," *Stud. Health Technol. Informat.*, vol. 272, pp. 465–469, 2020.
[42] *Lab Guide–2019 Central Clinical Section Lab Guide*, H. M. Corp., Qatar, 2019.
[43] C. D. Samuel-Hodge, Z. Gizlice, S. D. Allgood, A. J. Bunton, A. Erskine, J. Leeman, and S. Cykert, "Strengthening community-clinical linkages to reduce cardiovascular disease risk in rural NC: Feasibility phase of the CHANGE study," *BMC Public Health*, vol. 20, no. 1, pp. 1–10, Dec. 2020.
[44] S.-M. Ching, Y.-C. Chia, M. A. H. Lentjes, R. Luben, N. Wareham, and K.-T. Khaw, "FEV1 and total cardiovascular mortality and morbidity over an 18 years follow-up population-based prospective EPIC-NORFOLK study," *BMC Public Health*, vol. 19, no. 1, p. 501, Dec. 2019.
[45] K. M. Ali, A. Wonnerth, K. Huber, and J. Wojta, "Cardiovascular disease risk reduction by raising HDL cholesterol–current therapies and future opportunities," *Brit. J. Pharmacol.*, vol. 167, no. 6, pp. 1177–1194, Nov. 2012.
[46] B. C.-H. Kwan, C.-C. Szeto, K.-M. Chow, M.-C. Law, M. S. Cheng, C.-B. Leung, W.-F. Pang, V. W.-K. Kwong, and P. K.-T. Li, "Bioimpedance spectroscopy for the detection of fluid overload in chinese peritoneal dialysis patients," *Peritoneal Dialysis Int., J. Int. Soc. Peritoneal Dialysis*, vol. 34, no. 4, pp. 409–416, Jun. 2014.
[47] J. P. Antovic and M. Blombäck, *Essential Guide to Blood Coagulation*. Hoboken, NJ, USA: Wiley, 2013.
[48] V. Kumar, K. Kumar, A. Lakshmi, P. V. L. N. S. Rao, and G. Das, "Carotid intima-media thickness in patients with end-stage renal disease," *Indian J. Nephrol.*, vol. 19, no. 1, p. 13, 2009.
[49] T. Wang, Z.-W. Sun, L.-Q. Shao, X.-B. Xu, Y. Liu, M. Qin, X. Weng, and Y.-X. Zhang, "Diagnostic values of serum levels of homocysteine and uric acid for predicting vascular mild cognitive impairment in patients with cerebral small vessel disease," *Med. Sci. Monitor, Int. Med. J. Exp. Clin. Res.*, vol. 23, p. 2217, May 2017.
[50] E.-H. Park, S. Choi, and J.-S. Srong, "The relationship between serum homocysteine, uric acid and renal function in chronic gouty patients: 2 year follow-up results.: 1216," *Arthritis Rheumatol.*, vol. 66, Nov. 2014.
[51] A. C. Webster, E. V. Nagler, R. L. Morton, and P. Masson, "Chronic kidney disease," *Lancet*, vol. 389, no. 10075, pp. 1238–1252, Mar. 2017.
[52] L. F. Fried, M. G. Shlipak, C. Crump, R. A. Kronmal, A. J. Bleyer, J. S. Gottdiener, L. H. Kuller, and A. B. Newman, "Renal insufficiency as a predictor of cardiovascular outcomes and mortality in elderly individuals," *J. Amer. College Cardiol.*, vol. 41, no. 8, pp. 1364–1372, Apr. 2003.
[53] N. J. Aburto, S. Hanson, H. Gutierrez, L. Hooper, P. Elliott, and F. P. Cappuccio, "Effect of increased potassium intake on cardiovascular risk factors and disease: Systematic review and meta-analyses," *BMJ*, vol. 346, p. f1378, Apr. 2013.
[54] S. G. Al-Kindi, A. Sarode, M. Zullo, S. Rajagopalan, M. Rahman, T. Hostetter, and M. Dobre, "Serum bicarbonate concentration and cause-specific mortality: The national health and nutrition examination survey 1999-2010," *Mayo Clinic Proc.*, vol. 95, no. 1, pp. 113–123, 2020.
[55] M. Kabootari, M. R. Raee, S. Akbarpour, S. Asgari, F. Azizi, and F. Hadaegh, "Serum alkaline phosphatase and the risk of coronary heart disease, stroke and all-cause mortality: Tehran lipid and glucose study," *BMJ open*, vol. 8, no. 11, 2018, Art. no. e023735.
[56] G. Ndrepepa, R. Colleran, and A. Kastrati, "Gamma-glutamyl transferase and the risk of atherosclerosis and coronary heart disease," *Clinica Chimica Acta*, vol. 476, pp. 130–138, Jan. 2018.

HAMADA R. H. AL-ABSI received the Bachelor of Technology degree (Hons.) in information and communication technology and the Master of Science degree in information technology from Universiti Teknologi PETRONAS, Malaysia, in 2009 and 2010, respectively. He is currently pursuing the Ph.D. degree with Hamad Bin Khalifa University, Qatar. His research interest includes machine learning applications to diseases detection and diagnosis. He is a member of the Australian Computer Society (ACS) and a Certified Professional by ACS.

MAHMOUD AHMED REFAEE received the M.B.B.Ch. degree, in 2003, and the master's degree in geriatric medicine, in 2008, and the Ph.D. degree in geriatric medicine, in 2013. He completed his residency training in geriatric medicine, in 2008. He was appointed as a Lecturer in geriatric medicine with Ain Shams University, Cairo, Egypt, in 2014. He joined Hamad Medical Corporation (HMC), Qatar, as a specialist Geriatric medicine, in 2014. He finished Geriatric Fellowship Training Program 2019. He led several projects in sepsis, medications use in elderly and long-term care. His research interests include cognitive impairment in elderly and management of complex geriatric patients, big data use, and artificial intelligence in health care services.

ATIQ UR REHMAN (Member, IEEE) received the master's degree in computer engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2013, and the Ph.D. degree in computer science and engineering from Hamad Bin Khalifa University, Qatar, in 2019. He is currently working as a Postdoctoral Researcher with the College of Science and Engineering, Hamad Bin Khalifa University. His research interests include the development of algorithms for evolutionary computation, pattern recognition, machine learning, and image/video processing.

MOHAMMAD TARIQUL ISLAM is currently an Assistant Professor with the Computer Science Department, Southern Connecticut State University. He has published at notable peer-reviewed conferences and journals, such as *Computer Vision*, *Pattern Recognition*, International Conference on Image Processing, International Conference on Bioinformatics and Biomedicine, *International Journal on Image and Video Processing*, and so on. His research interests include computer vision, deep learning, and applied bioinformatics. He was a recipient of several grants on the application of deep learning in computer vision and has supervised multiple graduate students in their pursuit of a master's degree.

SAMIR BRAHIM BELHAOUARI (Senior Member, IEEE) received the master's degree in telecommunications from the National Polytechnic Institute (ENSEEIHT), Toulouse, France, in 2000, and the Ph.D. degree in applied mathematics from the Federal Polytechnic School of Lausanne (EPFL), in 2006. He is currently an Associate Professor with the Division of Information and Computing Technology, College of Science and Engineering, HBKU. He also holds and leads several academic and administrator positions, the Vice Dean for Academic & Student Affairs with the College of Science and General Studies and University Preparatory Program, ALFAISAL University (KSA), University of Sharjah, UAE, Innopolis University, Russia, Petronas University, Malaysia, and EPFL Federal Swiss School, Switzerland. He is also working actively on developing algorithms in machine learning applied to visual surveillance, sensing technologies and biomedical data, with the support of several international fund for research in Russia, Malaysia, and in GCC. His main research interests include stochastic processes, machine learning, and number theory.

TANVIR ALAM is currently an Assistant Professor with the College of Science and Engineering, Hamad Bin Khalifa University. Among his notable research works are on the transcription regulation of non-coding RNAs and their roles in different diseases. His research work also centered around the application of artificial intelligence (AI) on the diagnosis and prognosis of communicable and non-communicable diseases. He is a member of FANTOM Consortium. He also served as a Reviewer in a number of international conferences and reputed journals.