INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 03, March 2024, pp: 433-438

e-ISSN : 2583-1062

Impact Factor: 5.725

CARDIOVASCULAR RISK PREDICTION

Tejashwini P[1], Rekha D S[2], Suprith R[3]

[1,2,3]Dept. of Electronics and Communication Engineering T John Institute of Technology Karnataka, India.

DOI: https://www.doi.org/10.58257/IJPREMS32915

ABSTRACT

Cardiovascular disease (CVD) poses a major global health danger to human society. In order to identify high-risk individuals and implement early therapies, the application of machine learning techniques to predict the risk of CVD is highly relevant. Our study introduces the XGBH machine learning model for predicting cardiovascular disease (CVD) risk, incorporating data from 14,832 Chinese Shanxi CVD patients into the Kaggle dataset. By analyzing key variables such as age, systolic blood pressure, and cholesterol levels, our model offers a streamlined approach to early intervention with minimal accuracy loss. We also explored the Elastic Net model for predicting subclinical atherosclerosis (SA), integrating data from vascular ultrasonography and coronary artery calcification scores. Our findings highlight the importance of modifiable risk factors like hypertension and lifestyle choices in CVD risk assessment, emphasizing the potential of machine learning in optimizing risk prediction and improving patient outcomes. This study contributes to the growing body of research on utilizing machine learning techniques in healthcare, particularly in identifying high-risk individuals and implementing timely interventions. By leveraging advanced algorithms and comprehensive datasets, our model aims to enhance the accuracy and efficiency of CVD risk prediction, ultimately leading to better patient care and outcomes in the field of preventive cardiology.

1. INTRODUCTION

Heart and vascular conditions together referred to as cardiovascular diseases (CVD) include rheumatic, coronary, and cerebral vascular illnesses. According to reports, heart and vascular disorders claim the lives of around 17.9 million patients annually worldwide1. The Global Burden of Disease Report 20192 states that the prevalence of CVD is rising gradually, with 523 million cases reported in 2019. Of these cases, 18.6 million fatalities occurred, or one-third of all deaths. Research has shown that a range of information is needed to accurately forecast cardiovascular disease (CVD), including genetic information, symptoms, lifestyle, and risk factors in addition to the patient's medical history. Therefore, in order to find the cause of a sickness, we must look at the relationship between risk factors and the condition and employ data analysis as theoretical support inherent patterns to provide precise illness occurrence prediction.

Since there are many risk variables with nonlinear interactions, the implicit assumption in the current models of CVD risk assessment is that each risk factor is linearly related to the chance of CVD prevalence11. This assumption may oversimplify the connection. These models all show considerable geographic and population specificity due to their tight modelling assumptions and small number of predictors, and current algorithms typically do not accurately forecast CVD risk12, especially for specific subgroups13.

By minimizing the error between projected and true outcomes, machine learning models can be used to create complicated nonlinear associations between diseases and risk factors15,16. The creation of a CVD risk score model with fewer features and higher accuracy is the goal of this research. In The following summarizes this paper's contributions in general: By including a histogram technique, the XGBH model presented in this research can provide improved predictive performance and reduce memory space. In order to increase the dataset, 14,832 Chinese cardiovascular patients' data are included at the same time. In order to prove the XGBH model's superiority, it is finally contrasted with four other machine learning models. In this paper, a few feature, high accuracy CVD risk prediction model is developed by ranking the importance of features on the dataset. Decision curves are used to assess the practical value of the four models, and ultimately, only three features are required to make a more accurate risk assessment of CVD. Lastly, this essay creates an organ risk nomogram for cardiovascular disease. The three screening traits are used to assess the risk of cardiovascular disease, making it possible to estimate how likely a patient's ailment will be.

These estimations, however, are based on a variety of studies with different approaches and little representation from low- and middle-income nations (LIC and MIC). Large-scale international research involving LIC and MIC using consistent sampling and measurement techniques are required to improve and validate the GBD's findings. Multi-national case-control studies are biased, even if they have produced comparative information on the risk factors for stroke and myocardial infarction (MI). Therefore, more research employing a thorough study an organ risk nomogram for cardiovascular disease. The three screening traits are used to assess the risk of cardiovascular disease, making it possible to estimate how likely a patient's ailment will be designed to expand and supplement the GBD's understanding of CVD risk factors.3,4 cardiovascular disease (CVD) continues to be the primary cause of morbidity and death

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 03, March 2024, pp: 433-438

worldwide [1]. Current risk prediction algorithms are generally based on multivariate regression models that integrate data on a small number of established risk factors. These models typically assume that all of these factors have a linear relationship with the outcomes of CVD, with little to no interaction between them.

2. SURVEY

Characteristics of the study population: The individuals' average age at baseline was 53.34 (6.77) years, with 24,470 (35%) being male, 52,385 (74.8%) having normal cholesterol, 59,479 (85%) having normal glucagon, 6169 (8.8%) smoking, 3764 (5.4%) alcoholic, and 56,261 (80.4%) engaging in regular exercise.

XGBH model validation: To train the model and cross-validate its performance, we used 80% of the datasets from the Kaggle competition and 80% of the dataset from Shanxi Baoquan Hospital. The remaining data was used as the test set. This research compares the XGBH model with four machine learning models: extreme Gradient Boosting (Boost), logistic regression, random forest, and linear classification support vector machine. The classification model is then modified based on the parameters of the classification process. Regardless of whether the BMI feature was included or not, the XGBH prediction model performed better than the other four baseline models in terms of AUC, recall, precision, and F1 score. The XGBH model's AUC and precision in the test group without BMI were 0.8059 and 0.7578, respectively, suggesting that the model is more accurate at predicting the risk of CVD. Next, as Table 2 illustrates, we attempted to incorporate a new feature, BMI, into the prediction model. The accuracy has slightly decreased after the BMI component was added, but other indicators have increased and now have a stronger predictive potential than they did previously.

Feature screening and model evaluation:

In this paper, we use the Permutation Importance method to analyse the feature importance, which is the contribution of each feature to the prediction, of four models, Logistic Reression, Random Forest, XG Boost, and XGBH, respectively. Tis is done by randomly arranging the values of a feature column in the dataset to obtain unordered feature values to train the model

Dataset	Model	AUC	Recall	Precision	F1 score
Without BMI	LinearSVC	0.6511 ⁻ (0.6432–0.6589)	0.6006	0.6636	0.6306
	LogisticRegression	0.6965 ± (0.6889–0.7041)	0.6566	0.7094	0.6820
	Random Forest	0.7129 ± (0.7055–0.72044)	0.6967	0.7170	0.7067
	XGBoost	0.8018 ± (0.7945–0.8090)	0.6860	0.7552	0.7190
	XGBH	**0.8059** ± (0.7987–0.8131)	**0.7027**	**0.7578**	**0.7293**
With BMI	LinearSVC	0.6559 ± (0.6480–0.6637)	0.6126	0.6662	0.6383
	LogisticRegression	0.7123 ± (0.7049–0.7198)	0.6749	0.7256	0.6994
	Random Forest	0.7147 ± (0.7072–0.7222)	0.7025	0.7170	0.7097
	XGBoost	0.8027 ± (0.7955–0.8099)	0.6866	0.7532	0.7184
	XGBH	**0.8069** ± (0.7997–0.8140)	**0.7043**	**0.7572**	**0.7298**

Table 2. Performance of all prediction models under various feature. Values for AUC denoete the mean ± confidence interval (CI). AUC, the area under a receiver operating characteristic cure; $Precesion = TP/(TP + FP)$, $Recall = TP/(TP + FN)$ where TP stands for true positive, TN for ture negative, FP for false positive, and FN for false negative; $F1score = 2(precision * recall)/(precision + recal)$.

The top five critical features shared by the two models with the highest prediction accuracy, XG Boost and XGBH, are age, body mass index (BMI), systolic blood pressure (Aphid), cholesterol (Chol), and diastolic blood pressure (Alpo). Additionally, out of all the projected results, systolic blood pressure (Aphi) has the highest feature weight, suggesting that it is the most predictive feature.

The findings demonstrate that the XGBH model can outperform the other four models when it employs the top five features for prediction. Moreover, the AUC only drops from 0.803 to 0.7999 when the number of features is reduced from five to three, suggesting that ap_lo and BMI have less of an impact on the improvement of model accuracy. Taking everything into account, we simply utilized three questions: 1. the systolic value blood pressure? 2. Is a normal cholesterol level? 3. What is the age?) of the survey will provide a precise evaluation of the risk of CVD.

3. LITERATURE EMBEDDING MODEL:

1. Intrinsic Evaluation Results:

We assessed the literature embedding model in three different ways in order to evaluate it using the similarity score: (a) Identification of words linked with CVD risk; (b) Identification of genetic risk factors; and (c) Identification of risk

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 03, March 2024, pp: 433-438

factors based on cohorts. We chose the top-15 words for (a) and (c), and the top-10 words for (b), in order to assess the quality of the embedding model. We examined the captured genes and phrases using the reference values connected with the input queries in order to confirm the accuracy of the risk factors and the data they were associated with for the intrinsic evaluation.

2. The Extrinsic Evaluation's Results:

Using the suggested methods, we performed CVD prediction tasks (CVD versus non CVD) using MESA data. Please take note that the goal of this section is to predict CVD utilizing the 564 parameters (age, sex, BMI, etc.) in the MESA data. The objective was to reduce the number of features from 564 to a sufficient number, and to do the same for the dimension. Then, the goal was to determine whether, in comparison to using the full feature set and dimensions, the reduced features and dimension could still result in accurate prediction between the two classes (CVD versus non-CVD). As previously mentioned in Section 3, we utilized the inner product of a label embedding vector and the variable embedding vectors of the MESA data for FS to determine the comparison

THE PERFORMANCE OF FEATURE SELECTION AND DIMENSIONALITY REDUCTION APPROACHES FOR CVD PREDICTION USING MESA DATA

(a) Performance with Feature Selection Approach								(b) Performance with Dimensionality Reduction Approach							
Method	Predictor	Acc.	Pre	Re.	F1	Input Size	Computation time of 6 predictors	Method	Predictor	Acc.	Pre.	Re.	F1	Input Size	Computation time of 6 predictors
Original	SVC	0.760	0.672	0.719	0.685	564 (all features)	675 sec	Original	SVC	0.760	0.672	0.719	0.685	564 (all features)	675 sec
	DT	0.736	0.638	0.684	0.647				DT	0.736	0.638	0.684	0.647		
	RF	0.762	0.672	0.724	0.685				RF	0.762	0.672	0.724	0.685		
	LR	0.761	0.671	0.721	0.684				LR	0.761	0.671	0.721	0.684		
	DNN	0.747	0.699	0.663	0.673				DNN	0.747	0.699	0.663	0.673		
	CNN	0.757	0.713	0.681	0.691				CNN	0.757	0.713	0.681	0.691		
Our method	SVC	0.758	0.659	0.722	0.672	128	**517 sec**	**Our method**	SVC	0.757	0.644	0.732	0.656	128	**1389 sec**
	DT	0.743	0.636	0.703	0.645				DT	0.737	0.607	0.702	0.609		
	RF	**0.768**	**0.684**	**0.730**	**0.697**				RF	0.743	0.641	0.696	0.652		
	LR	0.758	0.660	0.721	0.673				**LR**	**0.753**	**0.647**	**0.718**	**0.659**		
	DNN	0.740	0.692	0.639	0.650				**DNN**	**0.751**	**0.711**	**0.647**	**0.658**		
	CNN	0.755	0.711	0.671	0.681				CNN	0.742	0.696	0.641	0.650		
DT	SVC	0.748	0.636	0.710	0.648	128	597 sec	UMAP	SVC	0.433	0.465	0.289	0.320	128	1496 sec
	DT	0.747	0.626	0.717	0.633				DT	0.609	0.503	0.407	0.396		
	RF	0.756	0.669	0.712	0.680				RF	0.348	0.516	0.234	0.291		
	LR	0.749	0.642	0.709	0.654				LR	0.400	0.461	0.308	0.315		
	DNN	0.740	0.691	0.642	0.652				DNN	0.420	0.323	0.512	0.322		
	CNN	0.747	0.701	0.653	0.665				CNN	0.698	0.349	0.500	0.411		
RF	SVC	0.756	0.653	0.719	0.666	128	1789 sec	PCA	SVC	0.660	0.514	0.534	0.494	128	1409 sec
	DT	0.745	0.630	0.712	0.637				DT	0.607	0.514	0.520	0.488		
	RF	0.764	0.680	0.725	0.693				RF	0.616	0.514	0.533	0.510		
	LR	0.759	0.661	0.723	0.674				LR	0.662	0.516	0.536	0.496		
	DNN	0.752	0.706	0.662	0.674				DNN	0.640	0.505	0.502	0.486		
	CNN	0.758	0.719	0.665	0.678				CNN	0.699	0.525	0.518	0.458		

There are twelve risk factors for CVD at the individual or household level:

Table 2 shows that among the behavioural risk variables, tobacco smoking had the strongest correlation with cardiovascular disease (CVD), followed by physical inactivity and poor diet. The highest correlation between metabolic risk variables and CVD was seen in hypertension, which was followed by raised non-HDL cholesterol, higher WHR, and diabetes. A increased risk of CVD was also linked to low grip strength, low education levels, depressive symptoms, and indoor air pollution. The risk of diabetes was highest in HIC and LIC, the risk of low education was highest in LIC, and the risk of tobacco use was highest in HIC. Compared to myocardial infarction, hypertension was a greater risk factor for stroke; however, diabetes, non-HDL cholesterol.

High sodium versus CVD and mortality:

Compared to a reference of 4–6g/day, excretion of >6g/day of sodium was associated with a 1·12(95% CI 1·03, 1·22) risk of CVD, 1·16(1·00, 1·34) of MI, 1·09 (0·98, 1·21) of stroke, and 1·18(1·07, 1·29) of death. Elevated sodium intake accounted for 3·2% of the PAF for CVD, 2·7% for MI, 3·3% for stroke, and 3·9% for death.

Ambient PM2.5 air pollution vs CVD and mortality:

For each 10 unit increase in outdoor PM2·5 there was a HR of 1·05 (95% CI 1·02–1·08) in the risk of CVD, with a larger effect with stroke (HR = 1·08 (95% CI 1·05–1·11) than with MI (HR = 1·03 (95% CI 1·00–1·06))

Prediction accuracy in individuals with history of diabetes:

The variable ranking for the diabetic sub-population is provided in Table 5. We note that the list of important variables in the diabetic subgroup is substantially different from that of the overall population. One major difference is that for

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 03, March 2024, pp: 433-438

diabetic patients, microalbuminuria appeared to be strongly linked to an elevated CVD risk. In the diabetic population (17,908 participants), participants with no CVD events had an average microalbumin in urine of 61.0 mg/L, whereas for those with a CVD event, the average microalbumin in urine was 128.76 mg/L.

Variable ranking for the diabetic population.

Variable	Score
Age	0.207
Microalbumin in urine	0.110
Usual walking pace	0.078
Smoking status	0.064
Systolic blood pressure	0.034
Red blood cell distribution width	0.027
Neutrophill count	0.018
Number of Treatments	0.018
High blood pressure	0.014
Urinary sodium concentration	0.014

4. DISCUSSION

Xgbh model: This study presents the XGBH model, which introduces the concept of a histogram to reduce the number of samples and features without sacrificing accuracy. The XGBH model outperforms XG Boost in terms of prediction performance while offsetting its longer training time and higher memory usage. The XGBH model uses treatment data from 70,000 patients in Europe and Asia along with the ML algorithm to forecast the chance of CVD incidence. We may apply the ML method for more straightforward and effective CVD prediction because the prediction model simply uses retrospective patient data. This strategy, as opposed to conventional CVD prediction models, saves the extra expense and hassle of gathering baseline data. This study presents the XGBH model, which introduces the concept of a histogram to reduce the number of samples and features without sacrificing accuracy. The XGBH model outperforms XG Boost in terms of prediction performance while offsetting its longer training time and higher memory usage. The XGBH model uses treatment data from 70,000 patients in Europe and Asia along with the ML algorithm to forecast the chance of CVD incidence.

We may apply the ML method for more straightforward and effective CVD prediction because the prediction model simply uses retrospective patient data. This strategy, as opposed to conventional CVD prediction models, saves the extra expense and hassle of gathering baseline data.

Literature embedding model: We examined related terms for every query in three scenarios for the intrinsic evaluations: (a) the identification of the connected words with CVD risk, (b) the identification of the genetic risk factor, and (c) the identification of the cohort-based risk factor. In the (a)–(c) scenarios, our method yielded an average accuracy of greater than 96%. These findings confirm that finding risk factor terms and genes related to the input query can be done quickly and accurately with our method. According to the findings of the extrinsic evaluation, our embedding model well represented the genotype and phenotypic variables for CVD prediction using the MESA dataset. The excellent CVD prediction performance with the chosen/dimension-reduced variables indicates that our embedding model evaluated CVD-associated risk factors and related data with accuracy. These outcomes demonstrate the viability of our method for precisely identifying CVD risk factors and related data. One little group of modifiable risk variables accounts for almost 70% of CVD cases. Globally, metabolic risk factors were responsible for the majority of PAF for CVD, stroke, and MI. The biggest risk factor for CVD, hypertension, accounted for slightly more than one-fifth of the PAF for CVD. Compared to MI, stroke was more impacted by hypertension. In the worldwide cohort, the PAFs of excessive salt consumption (i.e., >6 g/day) for CVD and death were comparatively minor (approximately 3.0%), which is in line with the majority of research that have looked at the direct relationship between sodium excretion and CVD or mortality.13, 19–21 A increased risk was mostly linked to ambient air pollution. of CVD, whereas home air pollution was linked to increased risks of death and CVD, possibly as a result of the higher pollution levels when using solid fuels for cooking. A 3% increase in the risk of CVD fatalities, a 5% increase in CVD occurrences, a 3% increase in MI, and a 7% increase in stroke are linked to a 10 microgram rise in PM 2.5. PURE suggests that focusing on a small number of modifiable risk variables could prevent a significant amount of CVD and premature mortality. The relevance of various risk factors differs throughout countries at different economic levels, underscoring the need for extra context-specific goals for global policies, even while some risk factors (such as smoke control, hypertension control, or improved education) merit worldwide policies.prevention of fatalities and early CVD. When compared to established scoring methods based on conventional risk variables and as currently recommended by primary prevention guidelines (Framingham score), Auto Prognosis dramatically improved the accuracy of CVD risk prediction. Second, Auto Prognosis found novel CVD risk predictors agnostically. Non-laboratory characteristics that are quite easy to gather through questionnaires, like the individuals' self-reported health ratings and typical walking pace, were found to be

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 03, March 2024, pp: 433-438

e-ISSN : 2583-1062

Impact Factor: 5.725

among the predictors. Third, Auto Prognosis revealed intricate relationships between various personal traits, resulting in the identification of risk factors unique to particular subpopulations for which current recommendations were yielding inaccurate forecasts.

5. METHODS

Model XGBH : Tree-based learning algorithm21, XGBH is a rapid high-performance gradient enhancement framework. The base model was selected to be the highest performing XG Boost model. Because XG Boost handled node splitting in earlier research by using a pre-sorting technique, XGBH presents a histogram approach (Histogram)26 that improves the accuracy of the split points that are computed. Nevertheless, there is a lengthy learning curve and significant memory use during operation. The fundamental principle of the Histogram method is to divide the continuous data of each feature into k boxes, or discrete box data, with each box being separated into a specific number of data points. Next, the k distinct boxes are utilized to utilized to create a k feature histogram. Consequently, the initial requirement of going through every sample point to identify the segmentation sites is reduced to a search across boxes, increasing throughput and decreasing memory usage. Additionally, discretizing the feature values using the histogram will not reduce accuracy; rather, it will regularize the data and strengthen the algorithm's capacity to be used generally.

External Assessment Embedding characteristics are used as the input for supervised machine/deep learning models in the extrinsic assessment.

To assess the validity of our embedding model representation, we applied our suggested embedding model to cohort data using (a) feature selection (FS) and (b) dimensionality reduction (DR) methodologies. We then performed CVD prediction using the features altered by FS or DR. As We used the Multi-Ethnic Study of Atherosclerosis (MESA) phenotypic data (Shemesh et al., 2020) that was gathered between 2000 and 2002 as the cohort data for CVD prediction tasks. 6,814 male and female participants ranging in age from 45 to 84 years old, representing a variety of phenotypic characteristics including age, gender, and racial/ethnic groupings, make up the data sets. This is a prospective cohort research that monitored the participants' status—that is, occurrences related to CVD—multiple times. In 2015, the participating subjects were reassessed to create the CVD and non-CVD labels. For CVD prediction tasks, we used the MESA datasets, which were gathered between 2000 and 2002 and had updated CVD labels in 2015. Additional MESA data details are given in (Shemesh et al., 2020).

EXAMINATION OF MULTIPLE LABELS: To distinguish between people with GD and those who did not exhibit any symptoms of the illness, four machine learning techniques were used: distributed random forest, naive Bayes, Elastic Net (EN), and gradient boosting machine. For exploration, we took a basic approach and divided the data into 4 groups: 3 for training and 1 for testing. Supplemental Table 3 reports the means of the four test sets for the proper (log-loss, Brier Score) and improper (area under the curve [AUC], F-scores, balanced accuracy) regulations. By all measures, EN was the top performer. The Supplemental Appendix (24–26) contains more information regarding the definition of the EN-PESA score and how EN is being implemented.

CARDIOVASCULAR RISK SCORES: To test the ENPESA score, we compared its performance with that of well-established cardiovascular risk scores designed for the prediction of cardiovascular events (27). The cardiovascular risk scores used included the

European Society of Cardiology SCORE (Systematic Coronary Risk Evaluation), which calculates 10-year risk of fatal cardiovascular disease (28), and the atherosclerotic cardiovascular disease algorithm for 10-year risk based on Pooled Cohort Equations (ASCVD) (29). For the ASCVD risk score, ACC/AHA guidelines suggest the following risk categories: low risk (LR) (<5%); borderline risk (BR) (5% to 7.4%); intermediate risk (IR) (7.5% to 20%); and high risk (HR) ($20%). For the SCORE, we used European Heart Association categories: LR (<1%); medium risk (1% to 5%); and HR ($5%). We also considered the 10- and 30-year risk of coronary heart disease (FH10Y, FH30Y) from the Framingham Heart Study (2).

Measurement of Risk Factors: Supplementary Appendix B, Table 2 provides a comprehensive overview of each risk factor, including its measurement technique and classification for the purpose of determining population attributable fractions (PAFs). The procedures used to collect the data were standardized. At the neighbourhood, home, and individual levels, baseline data were gathered. We assessed the risk at the individual and population levels for this analysis based on 14 potentially changeable risk factors. We employed a composite diet score for overall diet quality, which is at least as excellent as, or better than, prior diet risk assessments (unpublished data), and has been repeated in five independent trials. Since non-HDL-C exhibited the highest correlation with CVD, it was selected as our primary lipid value (Supplementary Appendix B, Table 3). urine fasting Excretion was calculated in 101,609 persons for whom data were available, using the Kawasaki formula as a proxy for sodium consumption.

Statistical analysis: To prevent overfitting, we used area under the receiver operating characteristic curve (AUC-ROC) to assess the prediction accuracy of each model under consideration using 10-fold stratified cross-validation. A training sample (381,244 participants) was utilized in each cross-validation fold to create the Cox PH models, conventional ML models, and our model (Auto Prognosis). A held-out sample (42,360 participants) was then used to assess performance. For each model, we include the 95% confidence intervals (Wilson score intervals) and the mean AUC-ROC. The Brier score was used to assess our model's calibration performance.

6. CONCLUSION

In conclusion, this research demonstrates that the XGBH model put forth in this work only needs three parameters: 1. systolic blood pressure 2. Typical cholesterol 3. age) to offer a more precise risk evaluation for CVD. Furthermore, it performs better than the prior baseline model in terms of computational time and model performance. In summary, this work presents a strategy that, by requiring only three patient indicators for good prediction, is more accurate than the current XG Boost model and better suited for predicting CVD risk in a broader variety of patients.

More accurate CVD prediction is made possible by our literature embedding model, which was trained using PubMed's literature data to identify related symptoms, processes, genes, and other risk variables for an input query. We assessed the performance of our model using both intrinsic and extrinsic assessments for impartial validation. Our method outperformed other widely used approaches on both FS and DR tasks for CVD prediction on cohort data (MESA data) with faster execution time for the extrinsic evaluation. It also provided accurate CVD risk factors, related genes, and associated information (average accuracy of >96%) for the intrinsic evaluation. Additionally, we created brand-new extrinsic assessment techniques that included both FS and DR. The application of extrinsic approaches to incorporate models in CVD literature has not been reviewed in any prior work. Because it depends, the intrinsic evaluation is subjective. Our literature embedding algorithm, trained on PubMed literature data, identifies relevant genes, processes, symptoms, and other risk indicators for an input query, enabling more precise prediction of CVD. We evaluated our model's performance using both internal and external evaluations for unbiased confirmation. With a faster execution time for the extrinsic evaluation, our method outperformed other popular algorithms on both FS and DR tasks for CVD prediction on cohort data (MESA data). Additionally, it supplied precise genes, information, and risk factors for CVD (with an average accuracy of about 96%) for the intrinsic evaluation. We also developed novel extrinsic assessment methods incorporating both FS and DR. There hasn't been any use of extrinsic methods to include models in CVD literature.

7. REFERENCES

[1] Moon, J., Posada-Quintero, H. F. & Chon, K. H. A literature embedding model for cardiovascular disease prediction using risk factors, symptoms, and genotype information. Expert Syst. Appl. 213, 118–930 (2023).

[2] Yusuf, S. et al. Modifable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middleincome, and low-income countries (PURE): A prospective cohort study . Te Lancet 395(10226), 795–808. https://doi.org/10.1016/ S0140-6736(19)32008-2 (2020).

[3] Yang, X. et al. Predicting the 10-year risks of atherosclerotic cardiovascular disease in Chinese population: Te China-PAR project (prediction for ASCVD risk in China). Circulation 134, 1430–1440. https://doi.org/10.1161/CIRCULATIONAHA.116.022367 (2016).

[4] Sánchez-Cabo, F. et al. Machine learning improves cardiovascular risk defnition for young, asymptomatic individuals. J. Am. Coll. Cardiol.https://doi.org/10.1016/j.jacc.2020.08.017 (2020).

[5] Roh, E. et al. Total cholesterol variability and risk of atrial fbrillation: A nationwide population-based cohort study. PLoS ONE 14, e0215687. https://doi.org/10.1371/journal.pone.0215687 (2019).

[6] Prediction of cardiovascular disease risk based on major contributing features Mengxiao Peng1 Fan Hou1, Zhixiang Cheng1, Tongtong Shen1 www.nature.com/scientificreports.