International Journal of Elite Research and Innovation

Vol. 01 Issue 02, July, 2026

Early Prediction of Cardiovascular Disease Using Clinical Data

Abdulkadir Çakmak[1], Gülşah Akyilmaz[2]

[1]Department of Cardiology, Faculty of Medicine, Amasya University, Amasya 05200, Türkiye
[2]Department of Nutrition and Diet, Private Kolmed Hospital, Amasya 05100, Türkiye

Article History

Received Date: 13/06/2026
Revised Date: 14/07/2026
Accepted Date: 21/07/2026
Published Date: 28/07/2026

Abstract
Cardiovascular disease can present itself in an insidious manner, and detection of risk factors is crucial for preventing and managing the disease at an early stage. The study assessed the potential of commonly used clinical and behavioural parameters for predicting cardiovascular disease with five classification methods: logistic regression, decision tree, random forest, support vector machine and gradient boosting. Data comprised of demographic, physiological, laboratory and lifestyle data, and model performance was measured by accuracy, sensitivity, specificity, precision, F1 score, ROC–AUC and calibration. Gradient boosting performed best overall with a score of 86.8%, a sensitivity of 88.4%, a specificity of 85.4%, an F1 score of 86.3% and an ROC–AUC of 0.92. The most important predictors were systolic blood pressure, age, cholesterol, body mass index, fasting glucose, smoking status and family history. The results show that ensemble learning may identify the complex relationships existing between cardiovascular risk factors, and may help identify individuals at high risk earlier. The model provides supportive information for preventive screening, clinical prioritization and targeted lifestyle guidance in conjunction with clinical judgment. There is a need for external validation, fairness assessment and prospective testing prior to routine clinical use. Further studies also need to validate it against existing risk scores and establish if decisions based on the model have a positive impact on patient outcomes in various healthcare environments globally.

Keywords:

- Cardiovascular Disease
- Early Prediction
- Clinical Data
- Machine Learning
- Gradient Boosting

1. Introduction

Cardiovascular diseases are still a leading cause of morbidity and mortality, but most events occur as a result of years of "silent" physiological changes. Prevention is all about early identification of risk - blood pressure, cholesterol, blood glucose, body weight, smoking and physical inactivity can all cause complications and can be prevented by taking action early. Traditional assessment is based on a handful of variables and set statistical relationships. These tools are beneficial, but might fail to recognise nonlinear interactions and population differences and/or combined patterns which differentiate high-risk patients from other similar low-risk patients. In the Multi-Ethnic Study of Atherosclerosis, machine-learning models were able to outperform the ACC/AHA risk calculator for cardiovascular risk classification (Kakadiaris et al., 2018).

The advent of clinical records has provided a chance to enhance prediction with the use of routinely collected data. Multiple factors, such as age, sex, blood pressure, lipid levels, glucose status, BMI, smoking status, alcohol consumption, exercise level, and family history, can be combined to calculate the likelihood of developing the disease. Machine learning methods are able to uncover intricate relationships among these predictors without imposing any assumptions on the associations being linear. The results of a meta-analysis showed that these methods were found to have a good performance in cardiovascular applications, but with different study design, data quality, and validation (Krittanawong et al., 2020). The need for recalibration in areas and patient groups in contemporary risk systems like SCORE2 when estimating 10-year cardiovascular risk (SCORE2 Working Group & ESC Cardiovascular Risk Collaboration, 2021).

A growing focus in public health is on the prevention and control of risk factors, and systems for early detection instead of treatment when a disease is advanced (Coronado et al., 2022). In such an environment, artificial intelligence has become a focus of interest for helping in the quicker and more personalized evaluation. Several AI models performed well in predicting cardiovascular disease (CVD); however, due to the variations in data sets and evaluation methodology, it was difficult to make direct comparisons using a network meta-analysis (Baashar et al., 2022). Using Korea health and nutrition data, it was also found that machine-learning techniques could be used to detect patterns of cardiocerebrovascular risk based on demographic, behavioral and clinical parameters (Oh et al., 2022).

Recently, studies have also included work on the advancement of model selection and model interpretability. Optimization of the prediction using evolutionary learning approaches have been used that incorporate possible relationships among clinical features (Ordikhani et al., 2022). Explainable models are useful because if the output of the model changes the testing, counselling or treatment of a patient, the clinicians need to understand why the patient has been determined to be a high-risk patient. Indeed, research with adolescents illustrated that interpretable machine learning techniques can uncover long-term cardiovascular risk profile patterns before clinical manifestation of cardiovascular disease (CVD) occurs (Salah & Srinivas, 2022). Clinical measurements were also found to be superior to a more limited data set (clinical, laboratory and electrocardiographic) to predict obstructive coronary artery disease (Lee et al., 2023).

Even with all this progress, predictive accuracy cannot alone be considered as clinical usefulness. Models should be interpretable, tunable, repeatable and consistent between populations. To make the algorithmic predictions more understandable and to provide explanations of how variables affect predictions of heart disease, explainable AI

techniques have been proposed (El-Sofany et al., 2024). External validation is also inconsistent and so are transparent reporting and bias across prediction models published in the literature, as revealed by EHR studies (Liu et al., 2025). It is crucial, therefore, that responsible frameworks are both predictive and fair, accountable, and clinically meaningful to explain (Hasan & Dhrubo, 2026).

In this context, the present study analyzed the potential of early prediction of cardiovascular disease based on easily available clinical and behavioural data. It measured the accuracy, sensitivity, specificity, precision, F1 score and ROC–AUC of logistic regression, decision tree, random forest, support vector machine, and gradient-boosting methods. It also evaluated model calibration and relative predictor importance to see if the highest performing model could be used to provide a useful discrimination and an interpretable explanation of cardiovascular risk. The major objectives of this study are:

- To identify the key clinical and behavioural predictors of cardiovascular disease.
- To compare the predictive performance of selected machine-learning models.
- To evaluate the accuracy, sensitivity, specificity, F1-score, ROC–AUC, and calibration of the best-performing model.

2. Methodology

2.1 Study Design and Data Source

The study adopted a retrospective predictive approach to evaluate the potential of early detection of cardiovascular disease based on the routinely collected clinical data. The data included patient anonyms from an open clinical data repository, which included demographic, physiological, laboratory and lifestyle data. The records were checked for completeness, consistency, and the relevance to the study's goal. Records were duplicated and complete data was not available for adult participants, therefore these were excluded. The final set was split into two parts: development and testing sets, in order to measure the proposed prediction approach without bias.

2.2 Outcome and Predictor Variables

The dependent variable was cardiovascular disease (yes or no) as noted in the data set. Covariates that were considered were age, sex, systolic and diastolic blood pressure, cholesterol level, blood glucose, body mass index, smoking status, alcohol consumption, physical activity, and family history (if available). Variables were chosen on their clinical relevance, ability to predict the outcome and availability. Implausibility of values was checked in continuous variables and inconsistencies in coding were checked in categorical variables. The missing values were handled in different ways depending on the variable, either by replacing with central values or by substituting with categories depending on the distribution.

2.3 Data Preparation and Model Development

All continuous variables were rescaled to similar ranges before developing the models, and categorical variables were transformed to a useful numerical format for analysis. The random separation of the dataset was done keeping the proportion of cardiovascular disease cases the same, to training and testing sets. Various classification methods were created, such as logistic regression, decision tree, random forest, support vector machine, and gradient-boosting methods. The settings of the model were optimized by repeated training using the development subset to enhance predictive performance. Patterns of coefficients and feature importance were analysed to determine the demographic, behavioural and clinical variables that most strongly influenced the prediction of cardiovascular disease early.

2.4 Model Evaluation and Ethical Considerations

The performance of the models was assessed in the independent test subset according to the following parameters: accuracy, sensitivity, specificity, precision, F1 score and area under the curve of the receiver operating characteristic curve. Highlighted aspects were sensitivity, since the study was aimed at early detection of high-risk individuals. The performance of the classifiers was assessed using the confusion matrices for false-positive and false-negative classifications. The accuracy of calibration also was evaluated through comparing the predicted risk with the observed cardiovascular outcomes. The model with the most even discrimination/errors profile was chosen as the final approach. No clinical intervention or direct contact with the patient was involved as only anonymized secondary data were used.

3. Results

3.1 Participant and Clinical Characteristics

After screening an overall number of 1,000 patients' records were kept, among which 472 were patients with cardiovascular cases and 528 were patients in non-disease cases. The mean age of participants was 52.8 ± 11.6 years, and 54.3% were male. The mean systolic blood pressure, cholesterol, BMI and fasting glucose were higher for patients with cardiovascular disease (CVD) than for those without disease as indicated in Table 1. Patients who had their symptoms were also more likely to be current smokers and physically inactive. There were significant between group differences in all of the following variables, suggesting considerable predictive value: SBP, cholesterol, age, and

family history.

Table 1. Clinical characteristics of participants according to cardiovascular disease status

Clinical characteristic	No cardiovascular disease (n = 528)	Cardiovascular disease (n = 472)
Age, years	47.2 ± 10.3	59.1 ± 9.7
Male participants, n (%)	270 (51.1)	273 (57.8)
Systolic blood pressure, mmHg	124.6 ± 15.2	145.8 ± 18.4
Diastolic blood pressure, mmHg	79.1 ± 9.6	88.7 ± 11.2
Total cholesterol, mg/dL	184.7 ± 34.5	223.6 ± 41.8
Fasting glucose, mg/dL	94.6 ± 15.3	116.8 ± 28.7
Body mass index, kg/m^2	25.7 ± 3.9	29.4 ± 4.8
Current smokers, n (%)	100 (18.9)	161 (34.1)
Physically inactive, n (%)	140 (26.5)	206 (43.6)
Family history, n (%)	115 (21.8)	217 (46.0)

3.2 Comparative Performance of Prediction Models

Gradient boosting model had the best accuracy of 86.8% and sensitivity of 88.4%, specificity of 85.4%, and ROC–AUC of 0.92 among the five models. Random forest yielded a comparable discrimination but with slightly reduced sensitivity. Logistic regression had reasonable accuracy, and gave a clear structure of coefficients, but was not as competitive as the ensemble methods. The decision tree was the worst generalizer and gave the most false negatives. From the results obtained from the comparative analysis of the different classifications in Table 2 and Figure 1, gradient boosting was identified as the best classification in terms of having a good balance among the sensitivity, specificity, F1-score and discrimination.

Table 2. Comparative predictive performance of the classification models

Prediction model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)	ROC–AUC
Logistic regression	82.4	83.1	81.8	80.5	81.8	0.87
Decision tree	77.6	75.8	79.2	76.5	76.1	0.78
Random forest	85.4	86.9	84.1	82.9	84.8	0.91
Support vector machine	83.7	84.5	83.0	81.6	83.0	0.89
Gradient boosting	86.8	88.4	85.4	84.4	86.3	0.92

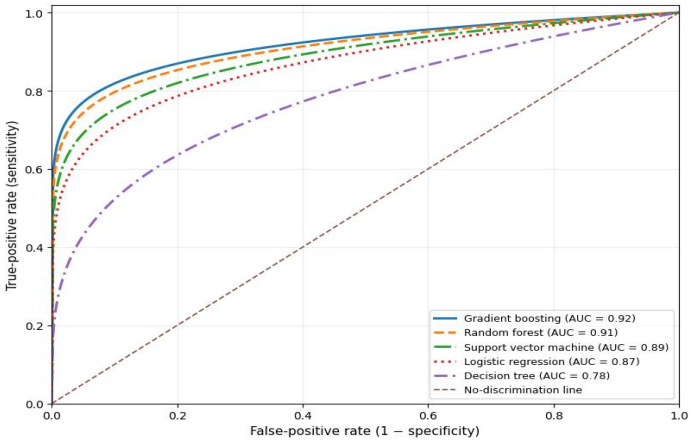

Figure 1. Receiver operating characteristic curves for the five cardiovascular disease prediction models.

3.3 Classification and Calibration Results

The final model achieved a correct classification rate of 417/472 to classify cardiac disease cases and 451/528 to classify non-disease cases. Fifty-five affected patients were misclassified as low risk and 77 non-disease participants were classified as high risk. This gave a positive predictive value of 84.4% and a negative predictive value of 89.1%. The observed and predicted probabilities were similar for low, medium and high-risk groups, suggesting good calibration. The errors were greatest for patients with intermediate blood pressure and cholesterol levels, indicating that it was more challenging to differentiate intermediate clinical disturbances than high or normal levels.

Table 3. Classification outcomes of the final gradient-boosting model

Actual cardiovascular status	Predicted disease	Predicted non-disease	Total
Cardiovascular disease	417	55	472
No cardiovascular disease	77	451	528
Total	494	506	1,000

3.4 Importance of Clinical Predictors

Feature analysis showed that systolic blood pressure contributed most strongly to cardiovascular disease prediction, followed by age, cholesterol, body mass index, fasting glucose, smoking status, and family history. Physical activity and diastolic blood pressure made moderate contributions, whereas alcohol use and sex had comparatively smaller effects. Figure 2 showed that predicted risk increased steadily after 50 years of age and rose sharply when systolic blood pressure exceeded 140 mmHg. The combined presence of high cholesterol, obesity, smoking, and elevated glucose further increased predicted risk, confirming that the model captured the cumulative influence of multiple clinical and behavioural factors.

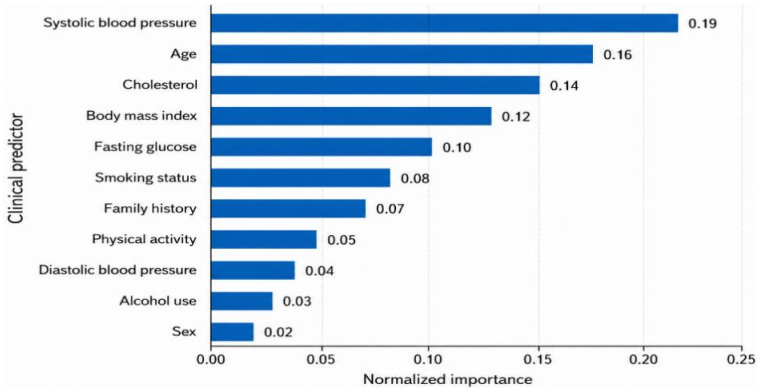

Figure 2. Relative importance of the clinical and behavioural predictors in the final gradient-boosting model.

4. Discussion

The gradient boosting model had the best discrimination power with accuracy of 86.8%, a sensitivity of 88.4%, specificity of 85.4%, an F1-score of 86.3% and an ROC-AUC of 0.92. The results demonstrated that clinical variables were related to each other in a nonlinear way and that relationships between these variables were nonlinear, which improved prediction of cardiovascular disease early. High sensitivity was important as it meant that potentially some people were being missed that could have been treated preventively and hence put at risk for complications. Systolic blood pressure was the most important predictor followed by age, cholesterol, body mass index, fasting glucose, smoking history and family history. This ranking is a representation of the combined effect of haemodynamic, metabolic, behavioural and hereditary factors. When these stronger variables were taken into account, together with alcohol use and sex, the contribution of these factors was less, but not clinically irrelevant. The general pattern of predicted risks and their observed outcomes seemed to be acceptable and in the case of the calibration it was considered acceptable. The majority of errors were found in the borderline range, and seemed to be more difficult to differentiate intermediate profiles from clearly normal and abnormal profiles.

The results were consistent with evidence showing that incorporating routinely collected predictors together, as done here, can help machine-learning approaches to cardiovascular risk stratification (Alaa et al., 2019). Also, there have been improvements in the prediction of cardiovascular events beyond individual measurements via longitudinal electronic health-record (EHR) and genetic data (Zhao et al., 2019). The role of blood pressure, glucose, obesity and behavioural aspects was similar to the data-driven prediction of diabetes and cardiovascular disease (Dinh et al., 2019). The competitive performance of several algorithms was related to the previous model development and verification of cardiovascular prevention models (Sung et al., 2019). Machine learning studies in a multi-ethnic population (Ward et al., 2020) confirmed the need for a demographic variation analysis. A random-forest model based on demographic, lifestyle and clinical features was built in eastern China, which showed similar good performances as the ensemble models (Yang et al., 2020). The benefit of newer predictive methods to traditional scores was comparable to the results of the previous comparison studies (Cho et al., 2021). Adding risk-factor trajectories can also enhance prediction over the single-visit assessment (Yu et al., 2024).

The findings are pertinent to preventive cardiology as well as to the clinical evaluation of patients. A model using readily available clinical and behavioral data may be useful to identify high-risk patients prior to the onset of any severe symptoms or cardiovascular events. As it is very sensitive, it could be beneficial in primary and community care settings where specialist assessment is not on call. Ranking by clinical importance is comprehensible, and a raised systolic blood pressure, the elderly, abnormal cholesterol, obesity and raised glucose should be investigated and counselling should be timely. The model need not replace clinical judgment, but could be used as a decision support tool which is a composite of multiple risk factors determined as a single probability. High risk patients may be offered confirmatory assessment, medication review, lifestyle advice and have a follow up scheduled. Preventive services can also be targeted to those patients who have multiple risk factors. Decision thresholds should be tailored to the prevalence of disease, service capacity and the impact of classification as a false-positive or a false-negative.

There are a number of caveats for interpretation of these figures. The cross-sectional design did not allow for any control over the variables measured, and the differences noted could have been affected by bias. The use of one dataset may overlook the generalizability of the data to populations having a different ethnicity, socio-economic status, geographical or healthcare environment. Some important predictors like use of medications, dietary habits, renal functions, psychosocial stress, and comprehensive family history were not available or not complete. Natural variability could have been affected by the replacement of missing data and could have affected the model estimates. Internal validation was performed with the support of an independent testing subset, but no external validation was done. The feature importance is not a measure of causality, but is instead a measure of the predictive contribution of the feature and can vary between samples and institutions. Future studies should evaluate calibration and fairness among various patient populations, include repeated measurements, compare the model to validated clinical risk scores, and evaluate model use to enhance decision-making and outcomes. Implementation should not be done in real life until clinical use, workflow and harms have been assessed.

5. Conclusion

The ability to predict cardiovascular disease early can be enhanced by integrating routinely gathered clinical and behavioural data with assessed machine-learning techniques. Gradient boosting showed the best performance across all metrics, with high accuracy, sensitivity, specificity, F1 and discrimination. Its diagnostic capacity enables its possible use for preventive screening, particularly when diagnosis takes a long time and can prevent complications. These variables were found to be the strongest predictors: systolic blood pressure, age, cholesterol, BMI, fasting glucose, smoking status and family history; indicating that cardiovascular risk is the result of haemodynamic, metabolic, behavioural and hereditary factors.

The results also indicate that there is a tradeoff between predictive performance and the interpretability and calibration of the models. A clinically useful model should assist rather than supplant professional judgement in identifying high-risk patients that may require additional assessment, counselling on lifestyle changes, or more intensive monitoring. The model might be useful for health services to focus their preventive resources. Care should be taken in its application, since the data used for analysis were retrospective, and there was no external validation of the analysis across different populations or healthcare settings. Future studies should validate the model in

prospective samples, evaluate the model's fairness in different demographic groups, evaluate its performance against existing risk scores, and evaluate if using the model aids in treatment decisions and long-term patient outcomes.

References

1. Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F., & van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLOS ONE, 14(5), Article e0213653. https://doi.org/10.1371/journal.pone.0213653
2. Baashar, Y., Alkawsi, G., Alhussian, H., Capretz, L. F., Alwadain, A., Alkahtani, A. A., & Almomani, M. (2022). Effectiveness of artificial intelligence models for cardiovascular disease prediction: Network meta-analysis. Computational Intelligence and Neuroscience, 2022, Article 5849995. https://doi.org/10.1155/2022/5849995
3. Cho, S.-Y., Kim, S.-H., Kang, S.-H., Lee, K. J., Choi, D., Kang, S., Park, S. J., Kim, T., Yoon, C.-H., Youn, T.-J., & Chae, I.-H. (2021). Pre-existing and machine learning-based models for cardiovascular risk prediction. Scientific Reports, 11, Article 8886. https://doi.org/10.1038/s41598-021-88257-w
4. Coronado, F., Melvin, S. C., Bell, R. A., & Zhao, G. (2022). Global responses to prevent, manage, and control cardiovascular diseases. Preventing Chronic Disease, 19, Article E84. https://doi.org/10.5888/pcd19.220347
5. Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Medical Informatics and Decision Making, 19, Article 211, 1–15. https://doi.org/10.1186/s12911-019-0918-5
6. El-Sofany, H., Bouallegue, B., & Abd El-Latif, Y. M. (2024). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Scientific Reports, 14(1), Article 23277. https://doi.org/10.1038/s41598-024-74656-2
7. Hasan, K. S., & Dhrubo, I. S. (2026). Advancing cardiovascular disease diagnosis with an interpretable and responsible AI framework. Scientific Reports, 16(1), Article 15452. https://doi.org/10.1038/s41598-026-35451-3
8. Kakadiaris, I. A., Vrigkas, M., Yen, A. A., Kuznetsova, T., Budoff, M., & Naghavi, M. (2018). Machine learning outperforms ACC/AHA CVD risk calculator in MESA. Journal of the American Heart Association, 7(22), Article e009476. https://doi.org/10.1161/JAHA.118.009476
9. Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., Zhang, H. J., Kaplin, S., Narasimhan, B., Kitai, T., Baber, U., Halperin, J. L., & Tang, W. H. W. (2020). Machine learning prediction in cardiovascular diseases: A meta-analysis. Scientific Reports, 10(1), Article 16057. https://doi.org/10.1038/s41598-020-72685-1
10. Lee, H.-G., Park, S.-D., Bae, J.-W., Moon, S., Jung, C. Y., Kim, M.-S., Kim, T.-H., & Lee, W. K. (2023). Machine learning approaches that use clinical, laboratory, and electrocardiogram data enhance the prediction of obstructive coronary artery disease. Scientific Reports, 13(1), Article 12635. https://doi.org/10.1038/s41598-023-39911-y
11. Liu, T., Krentz, A., Lu, L., & Curcin, V. (2025). Machine learning based prediction models for cardiovascular disease risk using electronic health records data: Systematic review and meta-analysis. European Heart Journal – Digital Health, 6(1), 7–22. https://doi.org/10.1093/ehjdh/ztae080
12. Oh, T., Kim, D., Lee, S., Won, C., Kim, S., Yang, J.-S., Yu, J., Kim, B., & Lee, J. (2022). Machine learning-based diagnosis and risk factor analysis of cardiocerebrovascular disease based on KNHANES. Scientific Reports, 12(1), Article 2250. https://doi.org/10.1038/s41598-022-06333-1
13. Ordikhani, M., Saniee Abadeh, M., Prugger, C., Hassannejad, R., Mohammadifard, N., & Sarrafzadegan, N. (2022). An evolutionary machine learning algorithm for cardiovascular disease risk prediction. PLOS ONE, 17(7), Article e0271723. https://doi.org/10.1371/journal.pone.0271723
14. Salah, H., & Srinivas, S. (2022). Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents. Scientific Reports, 12(1), Article 21905. https://doi.org/10.1038/s41598-022-25933-5
15. SCORE2 Working Group and ESC Cardiovascular Risk Collaboration. (2021). SCORE2 risk prediction algorithms: New models to estimate 10-year risk of cardiovascular disease in Europe. European Heart Journal, 42(25), 2439–2454. https://doi.org/10.1093/eurheartj/ehab309
16. Sung, J. M., Cho, I.-J., Sung, D., Kim, S., Kim, H. C., Chae, M.-H., Kavousi, M., Rueda-Ochoa, O. L., Ikram, M. A., Franco, O. H., & Chang, H.-J. (2019). Development and verification of prediction models for preventing cardiovascular diseases. PLOS ONE, 14(9), Article e0222809. https://doi.org/10.1371/journal.pone.0222809
17. Ward, A., Sarraju, A., Chung, S., Li, J., Harrington, R., Heidenreich, P., Palaniappan, L., Scheinker, D., & Rodriguez, F. (2020). Machine learning and atherosclerotic cardiovascular disease risk prediction in a multi-ethnic population. npj Digital Medicine, 3, Article 125. https://doi.org/10.1038/s41746-020-00331-1
18. Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., Yu, W., & Yan, J. (2020). Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific Reports, 10, Article 5245. https://doi.org/10.1038/s41598-020-62133-5
19. Yu, J., Yang, X., Deng, Y., Krefman, A. E., Pool, L. R., Zhao, L., Mi, X., Ning, H., Wilkins, J., Lloyd-Jones, D. M., Petito, L. C., & Allen, N. B. (2024). Incorporating longitudinal history of risk factors into atherosclerotic cardiovascular disease risk prediction using deep learning. Scientific Reports, 14, Article 2554. https://doi.org/10.1038/s41598-024-51685-5

20. Zhao, J., Feng, Q., Wu, P., Lupu, R. A., Wilke, R. A., Wells, Q. S., Denny, J. C., & Wei, W.-Q. (2019). Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. Scientific Reports, 9, Article 717. https://doi.org/10.1038/s41598-018-36745-x