Integrating Demographic, Clinical, and Behavioral Risk Factors for Cardiovascular Disease: A Random Forest Approach for Analysis, Prevention, and Prediction

Ai Li[1,*,†], Fanrui Yang[2,†]

{lia396504@gmail.com[1], yangfanrui07@gmail.com[2]}

College of Advertising, Communication University of China, Beijing, 100 000, China[1]
Data Science and Big Data Technology, Nanjing University of Information Science & Technology, Nanjing, 210 000, China[2]
*corresponding author
[†]These authors contributed equally and should be co-first authors.

Abstract. Cardiovascular disease (CVD) remains a critical health concern worldwide, posing a significant threat to human well-being. Previous studies have established that behavioral factors (e.g. alcohol consumption), specific clinical indicators, and demographic characteristics (e.g., CKD) are key determinants influencing the risk of CVD. To identify the most impactful predictive factors and further enhance the prevention and treatment of CVD, we analyzed two datasets containing various CVD-related factors. Following Exploratory Data Analysis (EDA), we utilized multiple models for prediction, including random forest, MLP, deepFM,XGBoost etc, using GridSearch for best performance. Our findings reveal that the best prediction model is Random Forest model. In dataset A, the primary factors are BMI, AgeCategory (age), SleepTime (sleep duration), GenHealth and PhysicalHealth. While in dataset B, which includes more clinically relevant features, the most significant predictors are HadAngina, State, AgeCategory, ChestScan and BMI. The comparative analysis of both datasets demonstrates that the dataset with more detailed clinical data (dataset B) yields more accurate predictions for CVD risk than the dataset focusing on just behavioral and demographic factors (dataset A). These findings highlight the importance of combining detailed clinical data with behavioral and demographic information to improve the precision of CVD risk prediction and management.

Keywords: Cardiovascular Disease (CVD), Risk Prediction, Random Forest, Mendelian Randomization (MR), Epidemiological Data

1 Introduction

Cardiovascular diseases (CVD) remain a leading cause of mortality worldwide, driven by complex, multifactorial etiologies. To address that, numerous studies about factors to CVD and prediction of early CVD have been launched.

Although previous studies have explored the effects of smoking, hypertension and other factors on CVD and used traditional statistical models and methods to assess the risk of these factors,

CONF-MLA 2024, November 21, Adana, Turkey

DOI 10.4108/eai.21-11-2024.2354595

there are insufficient comparisons of the combined effects of multiple factors and static limitations of models.

To solve that, our research novelly categorized various factors related to CVD into three categories: behavioral, demographic, and clinical factors. We utilized two comprehensive datasets from Centers for Disease Control and Prevention (CDC), referred to as dataset A and dataset B, which contain a range of relevant factors.

To uncover the patterns and relationships between variables, we did Exploratory Data Analysis (EDA) across both datasets. Whereafter, we compared different machine learning and deep learning models for optimal CVD prediction, including Random Forest, XGBoost, and DeepFM. The results indicated that the Random Forest model has the best performance regarding accuracy and interpretability, especially when the dataset includes thoroughly recorded clinical information. Such a comprehensive comparison offers useful reference to future researchers and practitioners, helping further to make better model choices toward similar application scenarios. Moreover, we compare the performance of the best model on the two datasets to infer which category of information is more predictive of the CVD so as to provide reference for the information categories needed in CVD prediction modeling.

In summary, our main contributions are threefold.

First, we conducted a comprehensive evaluation of multiple machine learning and deep learning models for CVD prediction. After reviewing prior studies, we performed a comprehensive comparison of the performance of various machine learning models and two deep learning models on Dataset A and Dataset B for CVD prediction, finding that the Random Forest model is the most effective for these datasets.

Additionally, we novelly categorized CVD risk factors into demographic, behavioral, and clinical categories. The exact contribution of each category to CVD risk was teased out by investigation of individual categories and their interactions. Such a fine-grained analysis will help deepen the understanding of how the interaction of host characteristics affects cardiovascular health—an aspect less explored by previous studies. Our investigation attempts to find out which—clinical, demographic, or behavioral—category has the highest predictive value for CVD.

In the end, we took insight into which type of information was most predictive of CVD for early prevention. We compared model performance across the two datasets and differences in the information they contain to ascertain which type of information gives more prediction capability of CVD for valuable insight into the early prevention of CVD in practical settings.

2 Related Work

2.1 Factors to CVD

Cardiovascular diseases are influenced by a mix of factors that include chronic diseases such as hypertension, high cholesterol, and diabetes, among others, and behavioral factors like smoking, alcohol, obesity, and physical inactivity. In addition, a wide range of comorbid conditions could also impact CVD incidence: for instance, one contributing chronic condition to increased CVD risk is chronic kidney disease [1].

In the study conducted by Yusaku Hashimoto [2], smoking is recognized as a very significant risk factor for the development of chronic kidney disease by promoting inflammation, oxidative stress, and vascular injury; it also contributes through mechanisms of atherosclerosis to an increased cardiovascular disease risk. These interlinked progressions relate CKD and CVD with smoking-related vascular injury.

Additionally, Emily Banks and her team found that smoking greatly increases the risk of various CVD, including AMI, cerebrovascular disease, heart failure, and PAD, with risks rising in proportion to the number of cigarettes smoked. Current smokers are significantly more likely to experience these CVD events than never-smokers, particularly those smoking over 25 cigarettes per day [3, 4].

Furthermore, Alfred Pozarickij et al.'s study demonstrated that blood pressure, particularly systolic blood pressure (SBP) and diastolic blood pressure (DBP), is a primary risk factor for CVD. However, the use of antihypertensive medications can modulate the effects of genetic variations on blood pressure and CVD risk, suggesting that pharmacological interventions might alter the genetic predisposition's impact on blood pressure and CVD outcomes [5].

Moreover, long-COVID can result in various cardiovascular complications, including myocarditis, arrhythmias, and thromboembolic issues, which contribute to an elevated risk of major adverse cardiovascular events. Carme Pérez-Quilis's team investigates the prolonged effects of COVID-19 on the cardiovascular system, covering areas such as post-acute complications, autonomic dysfunction, vascular aging, and cardiovascular outcomes in patients with severe COVID-19 [6, 7]. Dipti Tripathi's team's study is similar in that they used induced pluripotent stem cell-derived cardiomyocytes (iPSC-CMs) to model hypoplastic left heart syndrome (HLHS). They investigated mechanisms such as apoptosis, oxidative stress, and mitochondrial dysfunction, while also testing therapies like sildenafil and cyclosporine A to improve mitochondrial function. This iPSC model allowed the prediction of disease progression and evaluation of treatments for early heart failure in HLHS, offering critical insights into its pathogenesis [8].

Although existing studies have found the relationship between a specific factor and CVD, few have comprehensively compared the impact of various factors on CVD in three aspects. Our datasets take into account behavioral, demographic, and clinical factors to assess their relative influence on CVD. Identifying the most predictive factors for CVD can provide valuable insights for early prevention and mid-term screening efforts.

2.2 CVD Prediction

Traditionally, CVD risk prediction has depended on epidemiological information that applies statistical models to develop associations between various risk factors and disease-related outcomes. For example, research conducted by Yusaku Hashimoto and Emily Banks quantifies the association between smoking habits and CVD incidence using the Cox proportional hazards model. For example, smoking status is considered a significant predictive factor in smoking risk assessment instruments, such as the Framingham Risk Score. For better accuracy of data, most models have multivariable adjustments, like gender, age, eGFR, hypertension, dyslipidemia, diabetes, BMI, and smoking status, to involve them in the risk estimate for more precision of data [2, 3, 5].

Indeed, in recent years there has been a growing research field for improving classical models by the integration of newer and more sophisticated methodologies with the use of blood pressure-based risk scoring systems like the Framingham Risk Score and QRISK2 in this aspect. The systems predict future cardiovascular events using static baseline measurements of both SBP and DBP. In fact, more rigorously investigating the static models' shortcomings, Pozarickij et al. used Mendelian Randomization to explore causality between blood pressure and CVD, applying genetic scores to delineate more subtle impacts of blood pressure on different types of cardiovascular disease [5].

Moreover, to further the field of prediction in CVD, Pozarickij et al. went ahead to combine genome-wide association studies (GWAS) with MR analysis to come up with the genetic score (GS) models. These use extensive genetic data to predict an individual's risk of a cardiovascular disease under various blood pressures, thereby increasing precision in risk assessment. However, they used multivariable Mendelian Randomization (MVMR) to evaluate multiple blood pressure traits at the same time and to get more precise and understandable predictions of each of their independent influences on different types of CVD. In the meanwhile, machine learning approaches, including multivariable regression analysis, were also taken up by Hashimoto et al. for the dissection of combined and independent effects of smoking in relation to CKD risk and CVD [2, 5].

In the study by Saravanan Srinivasan, they explore the application of various machine learning techniques for predicting cardiovascular heart disease using data from the UCI repository. The research evaluated several classifiers, including Learning Vector Quantization (LVQ), Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), Radial Basis Function (RBF), K-Nearest Neighbors (KNN), and XGBoost. The results indicate that LVQ is the most suitable method for heart disease prediction in this context, offering superior classification accuracy and reliability compared to other widely used techniques [9-15].

Traditional risk prediction models for cardiovascular diseases are developed traditionally based on static epidemiological data. Recent research is being done to update these models with genetic information in modeling causal inference, as well as developing more advanced techniques. It is new in the sense that our research integrates demographic, behavioral, and clinical factors for the prediction of CVD by an advanced machine learning approach. It brings comprehensive clinical indicators together with lifestyle factors, not solely focusing on one single risk factor but giving comprehensive evaluation to multiple interactions of data varieties, thus opening new avenues for better CVD risk prediction. Our current approach appropriately integrates many of the complexities of data types, in contrast to traditional static statistical models, so as to enhance the accuracy of prediction and clinical utility.

3 Methodology

3.1 Dataset and preprocessing

For a comprehensive analysis and exploration of factors influencing CVD, we utilize two datasets from the Centers for Disease Control and Prevention (CDC), the CDC 2020 dataset (dataset A) and the CDC 2022 dataset (dataset B). As shown in Table 1 and Table 2, dataset A mainly consists of behavioral and lifestyle indicators such as hours of sleep, activity, and

mobility, which are crucial to understand lifestyle-driven risk factors of CVD. Dataset B, on the other hand, would be more clinically detailed, including whether there is an angina complaint, history of chest scans, some clinical health parameters, and hence represents a growing shift toward data with higher diagnostic relevance.

Dataset A includes cardiovascular disease records of 319,795 patients. There are 4 numerical variables and 14 categorical variables. While dataset B includes cardiovascular disease records of 246,022 patients, with 6 numerical variables and 34 categorical variables.

Table 1. Description of Dataset A.

columns type	name	description
demographic	Sex	Gender
	AgeCategory	Age group
	Race	Ethnicity
behavioral	BMI	Body Mass Index
	SleepTime	Hours of sleep per night
	Smoking	Smoking status
	AlcoholDrinking	Alcohol consumption
	PhysicalActivity	Engagement in physical activity
clinical	Diabetic	Diabetes status
	Asthma	Presence of asthma
	Stroke	Presence of stroke (Yes/No)
	KidneyDisease	Presence of kidney disease (Yes/No)
	SkinCancer	Presence of skin cancer (Yes/No)
	PhysicalHealth	Physical health score (0-30 points)
	MentalHealth	Mental health score (0-30 points)
	GenHealth	General health condition
	DiffWalking	Difficulty in walking (Yes/No)
Target column	HeartDisease	Had Cardiovascular disease (Yes/No)

This table categorizes the columns of Dataset A into three main types: demographic, behavioral, and clinical. The demographic columns include information such as gender, age group, and ethnicity. The behavioral columns cover lifestyle factors, including SleepTime and Smoking status. The clinical columns provide data on health conditions and metrics, such as Diabetic status and Asthma. The table illustrates how these different types of factors contribute to CVD risk in Dataset A.

Table 2. Description of Dataset B.

columns type	name	description
demographic	Sex	Gender
	AgeCategory	Age group
	State	State of residence
	RaceEthnicityCategory	Race and ethnicity category (categorical data)
	BMI	Body Mass Index
	HeightInMeters	Height in meters
	WeightInKilograms	Weight in kilograms

	SleepHours	Hours of sleep per night
	PhysicalActivities	Engagement in physical activity
behavioral	ECigaretteUsage	Usage of electronic cigarettes (Yes/No)
	SmokerStatus	Smoking status
	AlcoholDrinkers	Alcohol consumption
	HadDiabetes	Diabetes status
	HadAsthma	Presence of asthma
	HadStroke	History of stroke (Yes/No)
	HadKidneyDisease	History of kidney disease (Yes/No)
	HadAngina	History of angina (Yes/No)
	HadCOPD	History of COPD (Yes/No)
	HadSkinCancer	History of SkinCancer(Yes/No)
	HadDepressiveDisorder	History of depressive disorder (Yes/No)
	HadArthritis	History of arthritis (Yes/No)
	DeafOrHardOfHearing	Hearing impairment (Yes/No)
	BlindOrVisionDifficulty	Visual impairment (Yes/No)
	DifficultyConcentrating	Difficulty with concentration (Yes/No)
	DifficultyDressingBathing	Difficulty with dressing and bathing (Yes/No)
clinical	DifficultyErrands	Difficulty with errands (Yes/No)
	CovidPos	COVID-19 positive status (Yes/No)
	LastCheckupTime	Time since last medical checkup (e.g., in months)
	RemovedTeeth	Number of teeth removed (count)
	ChestScan	Whether a chest scan was performed (Yes/No)
	HIVTesting	Whether HIV testing was abnormal (Yes/No)
	FluVaxLast12	Flu vaccination in the last 12 months (Yes/No)
	PneumoVaxEver	Ever received pneumonia vaccine (Yes/No)
	TetanusLast10Tdap	Tetanus vaccination in the last 10 years (Yes/No)
	HighRiskLastYear	High risk status in the last year (Yes/No)
	DifficultyWalking	Difficulty with walking (Yes/No)
	GeneralHealth	General health condition
	MentalHealthDays	Days of poor mental health
	PhysicalHealthDays	Days of poor physical health
Target column	HadHeartAttack	Had Cardiovascular disease (Yes/No)

Compared to Dataset A, Dataset B includes more detailed clinical information. The clinical columns provide extensive health-related data, including detailed diagnostic and treatment indicators. This table highlights the enhanced clinical detail present in Dataset B, which contributes to a more comprehensive analysis of CVD risk.

For the preprocessing, after addressing missing and extreme values, we applied oversampling techniques to balance the proportion of coronary heart disease cases and controled to a 1:1 ratio. Then we used label encoding and one-hot encoding for encoding categorical columns and min-max normalization for encoding continuous variables. For both datasets, we divided the dataset into training, test, and validation sets according to the ratio of 10:1:1.

3.2 Exploratory Data Analysis (EDA)

We did Exploratory Data Analysis (EDA) on dataset A and dataset B to understand how features interact with each other, including pie plot for target column, line plot for comparison between two datasets, correlation heat map, the results are described in section 4 (See section 4).

3.3 models

To find the best model for two datasets, we tried different Machine Learning models on dataset A and dataset B, including random forest, decision tree, logistic regression [16], XGBoost, Gradient Boosting [17], AdaBoost [18], CatBoost [19], Ridge [20]. For each model, we used GridSearch to make the best performance.

Random Forest is an ensemble method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. Logistic Regression models binary outcomes using a logistic function. XGBoost is a fast and efficient gradient boosting implementation that incorporates regularization to prevent overfitting, popular in competitive settings. Gradient Boosting sequentially combines weak learners, like decision trees, to correct errors and enhance predictive accuracy. While AdaBoost focuses on misclassified instances, adapting the model to improve robustness. CatBoost [21-24] is designed for categorical features, requiring minimal preprocessing and handling overfitting well. Ridge Regression adds L2 regularization to linear regression.

Moreover, we utilized two neural network models Multi-layer Perceptron (MLP) and deepFM. deepFM is a factorization machine model based on neural network [21-24], originally used for CTR prediction in the field of recommending systems. It combines linear and nonlinear features, and can automatically learn high-order and low-order feature interactions, which is more interpretable than other deep learning models. In CVD datasets, feature interactions are often complex and difficult to identify, and DeepFM can naturally model these interactions.

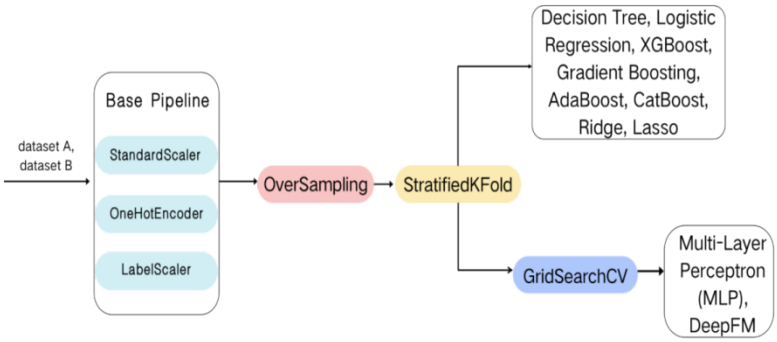


Fig. 1. Architecture of the Model Pipeline.

This figure shows the model pipeline architecture, detailing the use of different machine learning models, including Multi-Layer Perceptron (MLP) and DeepFM. It also illustrates the GridSearch process used for automated hyperparameter tuning.

3.4 Evaluation Metrics

Given the specificity of predicting CVD, we employ three evaluation metrics in our experiments: Accuracy, Area Under the Receiver Operating Characteristic Curve (AUC), and the F1 score.

4 Experimental results

4.1 EDA

The proportion of individuals with CVD is 5.5% in Dataset A, compared to 8.6% in Dataset B. The majority of individuals in both datasets do not have CVD, although there is a minor variation in the percentage of affected individuals. The similarity in the overall trends suggests consistent patterns in the prevalence of CVD across both datasets. However, slight differences in data representation, such as the categorization of smoking status, may indicate potential variations in data collection methods or demographic characteristics.

As illustrated in Figure 2, both datasets exhibit a strong positive correlation between age and the incidence of CVD, confirming that older populations have a higher risk of developing CVD. This trend identifies age as a consistent risk factor across different years and datasets. Moreover, In Dataset A, the correlation between age and CVD is more pronounced.

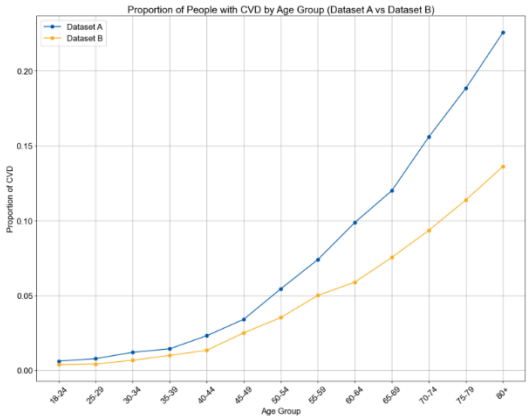


Fig. 2. CVD Percentage by Age Category

The chart displays the percentage of CVD across different age categories in Dataset A and Dataset B. Both datasets show a positive correlation between age and CVD rate, indicating a higher incidence of CVD among older populations.

As illustrated in Figure 3 and 4, differences in the distributions of smoking and alcohol consumption between the datasets were observed. The categories for smoking status displayed slight variations in granularity, which could potentially affect the comparative analysis. In contrast, the trends in alcohol consumption showed significant differences, raising concerns about the reliability and impact of this variable on CVD prediction. Further investigation revealed that, while smoking status remained a relatively stable predictor, alcohol consumption

did not independently influence the prediction results, suggesting a complex interaction with other variables, e.g. age.

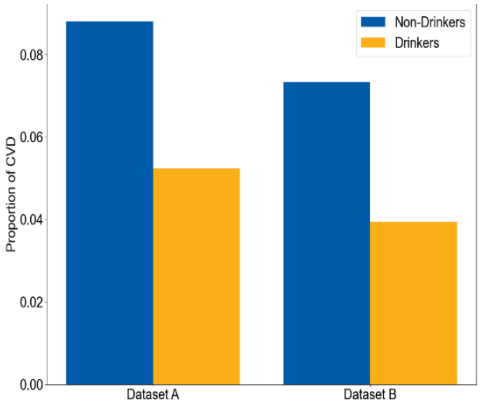


Fig. 3. Bar Chart of CVD Percentage by Alcohol Drinking Status (Left)

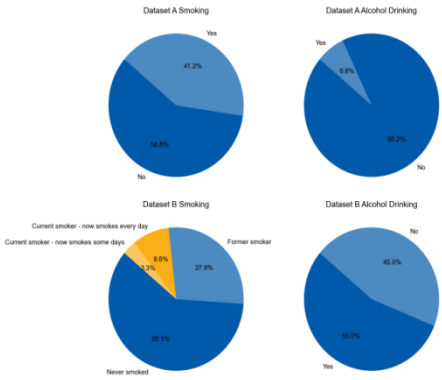


Fig. 4. Pie Chart of Smoking and Alcohol Consumption Distribution (Right)

Figure 3 illustrates the percentage of cardiovascular diseases (CVD) categorized by alcohol drinking status. Figure 4 depicts the distribution of smoking and alcohol consumption across the datasets, highlighting variations between them.

Correlation heat maps were utilized to investigate the relationships between various values. As shown in Figure 5 and 6 below, Dataset A identified factors such as AgeCategory, Stroke, Diabetic, PhysicalHealth and KidneyDisease as the primary predictors of CVD. In contrast, Dataset B highlighted more clinically relevant features, including HadAngina, HadStroke, AgeCategory, ChestScan and DifficultyWalking, as the principal predictors. This transition from behavioral to clinical indicators between datasets suggests an increased emphasis on diagnostic and medical history data in the more recent dataset, which enhances the accuracy of CVD risk prediction.

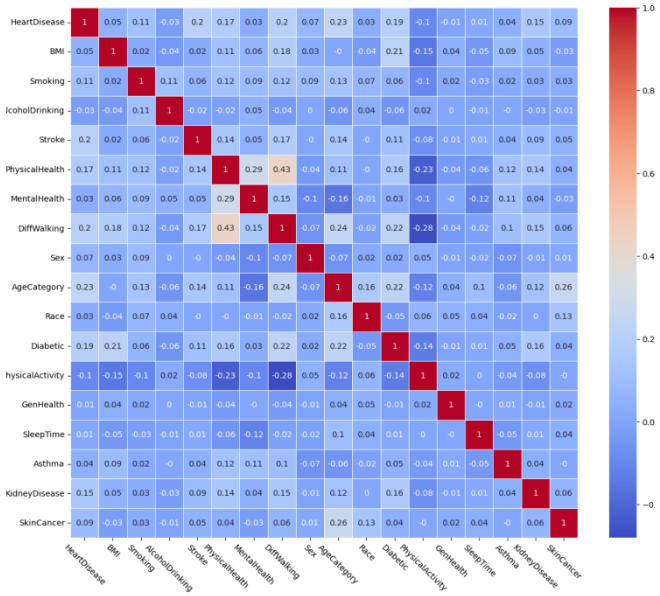


Fig. 5. Heatmap of the Correlation Matrix for Variables in Dataset A

This heat map displays the correlation matrix for the variables in Dataset A. The five variables most strongly correlated with CVD are, in order of significance, AgeCategory, Stroke, Diabetic, PhysicalHealth, KidneyDisease.

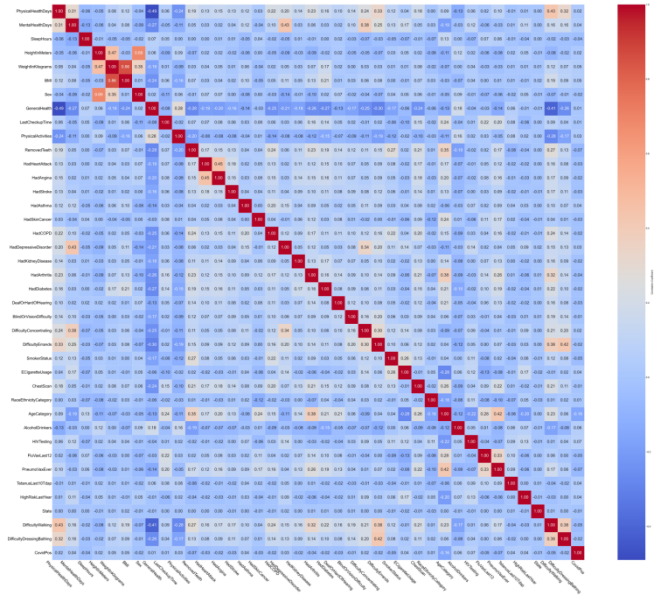


Fig. 6. Heatmap of the Correlation Matrix for Variables in Dataset B

For dataset B,the most five relevant features are HadAngina, HadStroke, AgeCategory, ChestScan and DifficultyWalking.

4.2 Prediction models

We employed multiple machine learning methods and deep learning models on Datasets A and B, with the results summarized in Table 2. The Random Forest model demonstrated superior performance for both datasets, achieving an accuracy of 0.9688, AUC of 0.9688, and F1 score of 0.9697 on Dataset A, and an accuracy of 0.9901, AUC of 0.9901, and F1 score of 0.9902 on Dataset B. Notably, Random Forest exhibited better performance on Dataset B compared to Dataset A.

Table 2. Performance of Different Models on Dataset A and Dataset B

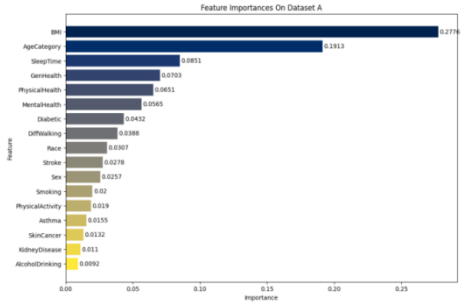
	dataset A			dataset B		
	accuracy	AUC	F1 score	accuracy	AUC	F1 score
Decision tree	0.9551	0.9551	0.9570	0.9706	0.9706	0.9715
Logistic regression	0.7639	0.7639	0.7676	0.8034	0.8034	0.7971
XGBoost	0.7860	0.7860	0.7942	0.8569	0.8569	0.8577
Gradient Boosting	0.7669	0.7669	0.7747	0.8069	0.8069	0.8029
AdaBoost	0.7623	0.7623	0.7655	0.8010	0.8010	0.7933
CatBoost	0.7963	0.7963	0.8045	0.8888	0.8888	0.8903
Ridge	0.7640	0.7640	0.7693	0.8028	0.8028	0.7933
MLP	0.6975	0.6975	0.7011	0.7111	0.7124	0.7312
deepFM	0.7015	0.7010	0.7199	0.7129	0.7363	0.7489
Random Forest	**0.9688**	**0.9688**	**0.9697**	**0.9901**	**0.9901**	**0.9902**

The table presents the performance metrics of various machine learning models on Dataset A and Dataset B. The Random Forest model achieved the highest accuracy on both datasets, with improved performance on Dataset B compared to Dataset A.

4.3 Feature importance

The features which were most significant in both datasets A and B that cause CVD have been reflected in the feature importance table by their analysis and visualization. Random forest provides the best performance for both datasets, as described above. The five most important features are BMI, AgeCategory, SleepTime, GenHealth, and PhysicalHealth in Dataset A. On the other hand, for Dataset B, the top five features are HadAngina, State, AgeCategory, ChestScan, and BMI. The above results show an apparent difference in the importance of demographic, behavioral, and clinical factors in the prediction of cardiovascular disease between the two datasets. In Dataset A, more prominence is being found to be on demographic and behavioral factors, such as BMI, AgeCategory, and SleepTime, evidently underlining more significant importance on general health and lifestyle metrics for CVD prediction. However, in Dataset B, the importance of clinical factors like HadAngina and ChestScan becomes dominant,

tending towards more specific medical data in the attempt to increase accuracy in predicting CVD risks.



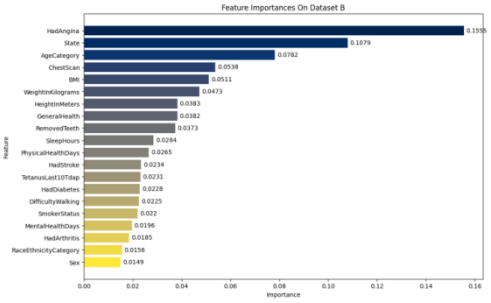


Fig. 7. Feature Importances for Dataset A (Left)
Fig. 8. Feature Importances in Dataset B(Right)

For Dataset A, the five most important features identified are BMI, AgeCategory, SleepTime, GenHealth, and PhysicalHealth, indicating a greater significance of demographic and behavioral information in the prediction model. While for Dataset B, the five most important features are HadAngina, State, AgeCategory, ChestScan, and BMI, highlighting the greater significance of clinical information in comparison to Dataset A.

5 Conclusion and future work

The results indicate that demographic factors, especially BMI and AgeCategory (age), and behavioral factors, especially SleepTime (sleep duration), significantly impact CVD prediction. Broader clinical indicators, including DiffWalking (difficulty walking), GenHealth, PhysicalHealth, and MentalHealth, also contribute to CVD prediction, though their effects are less pronounced compared to demographic and behavioral data. We developed a Random Forest model incorporating extensive demographic and behavioral data, along with a limited set of broad clinical indicators, achieving an accuracy exceeding 95%.

However, precise clinical data related to CVD could largely enhance the performance of cardiovascular disease prediction models, especially chest scan status, whether angina pectoris. Although the prediction model performed well on the mainly behavioral and demographic information dataset A, the prediction improved significantly when more clinically relevant features were added, achieving an accuracy exceeding 99%. The combination of detailed CVD related diagnostic test results significantly improved the accuracy of CVD risk prediction. This provides a more reliable predictive basis for doctors to make diagnoses in combination with clinical characteristics and daily behaviors and demographics.

It should be noted that although several preprocessing steps were undertaken to enhance data quality as much as possible, the influence of data quality on the prediction performance still cannot be entirely eliminated in this study, leading to a little inevitable bias.

For the future, we wish to continue predicting CVD. After this study, based on its results, further multimodal data sources could be integrated for enhancing the robustness and interpretability of

our models. Recent studies have further shown such modalities, especially electrocardiograms, to have good interpretability with respect to genetic data and environmental data in assessing risk for CVD [21-24]. Developing on our current work in demographic, behavioral, and clinical factors, we will include more specialized imaging data, such as coronary angiography, MRI, CT, and echocardiography, for example, chest X-rays (ChestScan). Furthermore, we aim to include more genetic inputs, for example, family history, specific genes associated with CVD risk, such as APOE, PCSK9. We will additionally take into account environmental data, like long-term exposure to air pollution, for example, PM2.5, and socioeconomic characteristics including income, level of education, occupation, social support networks, and the geographical and climatic characteristics of the place of residence.

We will explore CVD prediction at the same time under scenarios with incomplete input data. The results show that, due to the inclusion of relevant clinical factors combined with demographic and behavioral variables, the model has greatly improved and it outperforms 99% in comparison to the testing dataset. However, in practice, for prevention and treatment of CVD, not all patients are subject to comprehensive diagnosis, hence having only a part of the input data in the model. Future research will focus on optimizing the CVD prediction performance, especially when only partial data are available, which actually makes the model most effective and flexible in real clinical and public health interventions.

Acknowledgments

Ai Li and Fanrui Yang contributed equally to this work and should be considered co-first authors.

References

[1] A. I. o. H. a. Welfare. "Heart, stroke and vascular disease: Australian facts." https://www.aihw.gov.au/reports/heart-stroke-vascular-diseases/hsvd-facts/contents/about

[2] Y. Hashimoto *et al.*, "Effect of body mass index on the association between alcohol consumption and the development of chronic kidney disease, " *Scientific Reports 2021 11:1,* vol. 11, no. 1, 2021-10-14, doi: 10.1038/s41598-021-99222-y.

[3] E. Banks *et al.*, "Tobacco smoking and risk of 36 cardiovascular disease subtypes: fatal and non-fatal outcomes in a large prospective Australian study, " *BMC Medicine 2019 17:1,* vol. 17, no. 1, 2019-07-03, doi: 10.1186/s12916-019-1351-4.

[4] D. G. Cook, S. J. Pocock, A. G. Shaper, S. J. Kussick,"GIVING UP SMOKING AND THE RISK OF HEART ATTACKS: A report from The British Regional Heart Study, " *The Lancet,* vol. 328, no. 8520, 1986/12/13, doi: 10.1016/S0140-6736(86)92017-9.

[5] A. Pozarickij *et al.*, "Causal relevance of different blood pressure traits on risk of cardiovascular diseases: GWAS and Mendelian randomisation in 100, 000 Chinese adults, " *Nature Communications 2024 15:1,* vol. 15, no. 1, 2024-07-24, doi: 10.1038/s41467-024-50297-x.

[6] M. E. E. Hussein,A. Sharma, "Long-term impact of COVID-19 on the cardiovascular system, " *COVID-19's Consequences on the Cardiovascular System,* 2024/01/01, doi: 10.1016/B978-0-443-19091-9.00018-4.

[7] S. Nicolò,V. Serafina,M. G. Elena,S. Ciro,S. Carlotta,F. Federico,C. Paolo,M. Sergio,C. Matteo, "What is the association of COVID-19 with heart attacks and strokes? - PubMed, " *Lancet (London, England),* vol. 398, no. 10300, 08/14/2021, doi: 10.1016/S0140-6736(21)01071-0.

[8] D. Tripathi,S. Reddy"iPSC model of congenital heart disease predicts disease outcome," *Cell stem cell,* vol. 29, no. 5, 05/ 05/ 2022, doi: 10. 1016/ j. stem. 2022. 04. 010.

[9] A. Darolia,R. S. Chhillar,M. Alhussein,S. Dalal,K. Aurangzeb,U. K. Lilhore, "Enhanced cardiovascular disease prediction through self-improved Aquila optimized feature selection in quantum neural network & LSTM model - PubMed, " *Frontiers in medicine,* vol. 11, 06/20/2024, doi: 10.3389/fmed.2024.1414637.

[10] J. Fang, C. Luncheon, C. Ayala,E. Odom,F. Loustalot, "Awareness of Heart Attack Symptoms and Response Among Adults - United States, 2008, 2014, and 2017 - PubMed, " *MMWR. Morbidity and mortality weekly report,* vol. 68, no. 5, 02/ 08/ 2019, doi: 10. 15585/ mmwr. mm6805a2.

[11] Srinivasan, S. et al. "An active learning machine technique based prediction of cardiovascular heart disease from UCI- repository database, " *Scientific reports,* vol. 13, no. 1, 08/ 21/ 2023, doi: 10. 1038/ s41598- 023- 40717- 1.

[12] K. L. Lakshmi, M. Umadevi, and L. P. Bellamkonda, "Optimized truncated singular value decomposition and hybrid deep neural network with random forest for automated disease prediction, " *Biomedical Signal Processing and Control,* vol. 100, 2025/02/01, doi: 10.1016/j.bspc. 2024.107010.

[13] T. S. Priyadarshini and M. A. Hameed, "Developing heart stroke prediction model using deep learning with combination of fixed row initial centroid method with Navie Bayes, Decision Tree, and Artificial Neural Network, " *Measurement: Sensors,* vol. 34, 2024/08/01, doi: 10.1016/j. measen.2024.101237.

[14] J. Botros, F. Mourad-Chehade, and D. Laplanche, "Explainable multimodal data fusion framework for heart failure detection: Integrating CNN and XGBoost, " *Biomedical Signal Processing and Control,* vol. 100, 2025/02/01, doi: 10.1016/j.bspc.2024.106997.

[15] H. A. Al-Alshaikh,P, Prabu,R. C. Poonia,A. K. J. Saudagar,M. Yadav,H. S. Alsagri,A. A. Alsanad, "Comprehensive evaluation and performance analysis of machine learning in heart disease prediction - PubMed, " *Scientific reports,* vol. 14, no. 1, 04/ 03/ 2024, doi: 10. 1038/ s41598- 024- 58489- 7.

[16] Z. Gao, S. Cheng, E. Wittrup, J. Gryak, and K. Najarian, "Learning using privileged information with logistic regression on acute respiratory distress syndrome detection, " *Artificial Intelligence in Medicine,* vol. 156, 2024/10/01, doi: 10.1016/j.artmed.2024.102947.

[17] R. Huang, C. McMahan, B. Herrin, A. McLain, B. Cai, and S. Self, "Gradient boosting: A computationally efficient alternative to Markov chain Monte Carlo sampling for fitting large Bayesian spatio-temporal binomial regression models, " *Infectious Disease Modelling,* vol. 10, no. 1, 2025/03/01, doi: 10.1016/j.idm.2024.09.008.

[18] C. Rao,M. Li,T. Huang,F. Li, "Stroke Risk Assessment Decision-Making Using a Machine Learning Model: Logistic-AdaBoost, " *CMES - Computer Modeling in Engineering and Sciences,* vol. 139, no. 1, 2023/12/30, doi: 10.32604/cmes.2023.044898.

[19] X. Wei, C. Rao, X. Xiao, L. Chen, and M. Goh, "Risk assessment of cardiovascular disease based on SOLSSA-CatBoost model, " *Expert Systems with Applications,* vol. 219, 2023/06/01, doi: 10. 1016/j.eswa.2023.119648.

[20] J. M. K. Aheto, H. O. Duah, P. Agbadi, and E. K. Nakua, "A predictive model, and predictors of under-five child malaria prevalence in Ghana: How do LASSO, Ridge and Elastic net regression

approaches compare?, " *Preventive Medicine Reports,* vol. 23, 2021/09/01, doi: 10.1016/j.pmedr.2021.101475.

[21] L. Hu, B. Liu, Y. Li, "Ranking sociodemographic, health behavior, prevention, and environmental factors in predicting neighborhood cardiovascular health: A Bayesian machine learning approach, " *Preventive Medicine,* vol. 141, 2020/12/01, doi: 10.1016/j.ypmed.2020.106240.

[22] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, "DeepFM: A Factorization-Machine based Neural Network for CTR Prediction, " *arXiv preprint arXiv:1703.04247,* 2017, doi: https://doi.org/10.48550/arXiv.1703.04247.

[23] Z. Jin, Y. Sun, and A. C. Cheng, "Predicting cardiovascular disease from real-time electrocardiographic monitoring: An adaptive machine learning approach on a cell phone | IEEE Conference Publication | IEEE Xplore, " doi: 10.1109/IEMBS.2009.5333610.

[24] J. Zhao *et al.*, "Learning from Longitudinal Data in Electronic Health Record and Genetic Data to Improve Cardiovascular Event Prediction, " *Scientific Reports 2019 9:1,* vol. 9, no. 1, 2019-01-24, doi: 10.1038/s41598-018-36745-x.