Integrating Machine Learning into Cardiovascular Disease Risk Prediction: A Comprehensive Analysis of Cholesterol, Heart Rate, and Gender Impact on Disease Prevalence

Abdul Rahim*, Dr. Amit Chhabra, Manya, Dr. Sunil K. Singh,
Dr. Sudhakar Kumar, Hardik Gupta, and Karan Sharma

{ co20301, amitchhabra, mco21376, sksingh
sudhakar, mco21373, mco21375 }@ccet.ac.in
Chandigarh College of Engineering and Technology

Abstract. Cardiovascular disease (CVD) remains a major global health issue, requiring accurate risk prediction models for early intervention. While traditional models use established risk factors, this study leverages machine learning to improve predictive accuracy by integrating variables like gender, serum cholesterol, and resting blood pressure. A novel approach is proposed to enhance a baseline CVD risk prediction model with machine learning predictions. The performance of this enhanced model using a hybrid dataset showed superior predictive accuracy over the baseline. Feature importance analysis highlighted the significant contributions of gender, serum cholesterol, and resting blood pressure. Initial results from machine learning algorithms were Random Forest (0.83), Logistic Regression (0.77), Decision Trees (0.77), ANN(0.58) and KNN (0.71). With the hybrid dataset, improved accuracies were seen: Random Forest (0.91), Logistic Regression (0.86), Decision Tree (0.83), ANN (0.76) and KNN (0.83). This research refines CVD risk assessment, leading to personalized interventions and better public health outcomes.

Keywords: Cardiovascular disease prediction, Machine learning, Medicine and science, Models

1 Introduction

The term Cardiovascular Diseases (CVD) refers to a variety of heart and blood vessel disorders. It causes reduced blood flow to the body, brain or heart because of the fatty deposits accumulating inside an artery, causing a blood clot (thrombosis), which causes atherosclerosis (hardening and shrinking of the artery)[1].

Cardiovascular Diseases(CVD) continue to pose a substantial global health challenge, demanding innovative approaches to risk assessment and prevention. The integration of advanced data analytics techniques, particularly machine learning, has ushered in a new era of precision medicine[2][3][4][5][6]. In this

context, the amalgamation of diverse datasets has become increasingly common to leverage the power of comprehensive information sources. In the pursuit of improving the accuracy of CVD risk prediction, this study introduces a novel hybrid dataset, merging data from Kaggle[7] a well known data science community platform, with clinical data from the esteemed Cardiovascular Disease Dataset (Mendeley)[8]. This unique amalgamation not only enriches the dataset but also enhances its diversity and breadth.

This research focuses on three key variables i.e. serum cholesterol levels, fasting blood pressure (BP), and gender[9] due to their well-documented significance in cardiovascular disease (CVD) risk assessment[10][11]. Serum cholesterol is a crucial biomarker for evaluating lipid profiles, which are vital for assessing heart health[1][12]. Fasting BP serves as an important physiological metric, reflecting an individual's baseline cardiovascular condition[13]. By examining how gender interacts with these predictive factors, the study aims to develop more personalized risk assessments[14][12][15]. Understanding the combined effects of these variables can provide deeper insights into CVD risk, ultimately leading to more effective prevention strategies and improved patient outcomes[16][4][17].

Figure 1. illustrates the distribution of deaths due to various cardiovascular diseases in Australia in 2021, highlighting that coronary heart disease was the leading cause, accounting for 41% of the deaths, followed by stroke at 20%, and other conditions such as heart failure, hypertensive disease, atrial fibrillation, peripheral arterial disease, and rheumatic heart disease making up the remainder[9].

Section 2 of this paper describes Literature review related to this research. The Materials and methods are described in Section 3, and it expands on the dataset, data preparation steps, and the analysis steps. Section 4 highlights the results found using the steps and section 5 ultimately presents the conclusion.

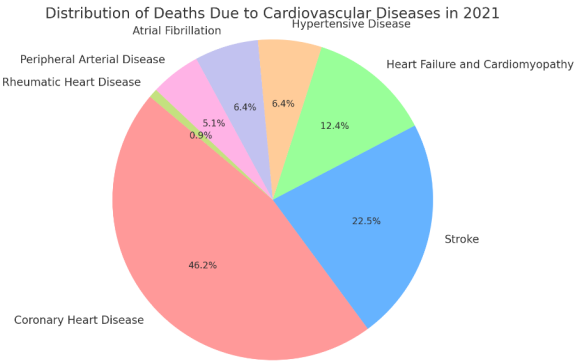

Fig. 1. Distribution of deaths due to different types of Cardiovascular Diseases (CVD) in 2021

2 Literature Review

It is imperative to conduct a comprehensive survey of existing research endeavors within this domain to develop effective machine learning models. This section of the paper summarizes numerous earlier studies on Cardiovascular diseases and various factors affecting them.

Krittanawong et al. [18] evaluated machine learning algorithm's performance in predicting cardiovascular diseases using diverse datasets from March 2019, employing the AUC metric to assess conditions like coronary artery disease, arrhythmias, heart failure, and stroke. However, finding the optimal algorithm remains challenging due to algorithm diversity. Lippi et al. [19] examined the impact of COVID-19 lockdowns on cardiovascular health, noting increased risks despite WHO guidelines on physical activity. Adverse health outcomes post-lockdown led to a recommendation for continued exercise during quarantine, though the study mainly focuses on physical inactivity rather than all contributing factors to cardiovascular diseases.

Han et al. [20] evaluated machine learning algorithms for predicting rapid coronary atherosclerosis progression using plaque data from 983 CT angiography scans, comparing model performance to atherosclerosis risk scores and key clinical variables. The study highlighted challenges in detecting hidden dataset biases. Anjan N. Repaka et al. [21] introduced a model evaluating the predictive performance of two classification models, finding that their proposed method outperforms others in accuracy for predicting risk percentage.

Lapague et al. [22] used a hybrid dataset from BRFSS and WHO to develop ML models for CVD risk assessment, addressing class imbalance and identifying Logistic Regression as the best model. Feature importance highlighted attributes like sex, diabetes, and general health. Suman et al. [12] explored gender disparities in CVD, noting higher post-acute event mortality rates in women due to genetic and hormonal factors, emphasizing the need for gender-specific prevention, diagnosis, and treatment. These studies highlight the potential of ML approaches and the importance of gender differences in enhancing CVD risk prediction and prevention.

The limitations of the approaches outlined in the studies include a heavy reliance on traditional metrics such as the area under the curve (AUC) for algorithm evaluation, which may not fully capture the complexity of cardiovascular disease (CVD) prediction. Additionally, challenges persist in selecting the most accurate algorithm due to inherent biases within datasets and the difficulty in integrating diverse variables effectively. Furthermore, previous models often overlook key contributors to CVD risk, such as gender disparities and variations in serum cholesterol and resting blood pressure levels.

We propose an innovative approach to address these limitations by leveraging machine learning advancements to augment baseline CVD risk prediction models. By integrating a hybrid dataset containing gender, serum cholesterol, and resting blood pressure variables, the enhanced model demonstrates superior predictive accuracy compared to traditional methods. Specifically, the Random

Forest and Logistic Regression models exhibit the highest accuracies, highlighting the potential of machine learning in refining CVD risk assessment.

Table 1. Literature Review

Authors	**Key Findings**	**Technology Used**	**Limitations**
Krittanawong et al. [18]	Utilized diverse datasets available as of March 2019; assessed efficacy in predicting various cardiovascular conditions	Diverse datasets, AUC metric	Challenging to determine the optimal algorithm
Lippi et al.[19]	Explored potential impact of pandemic on cardiovascular health; suggested importance of continuing physical exercise during lockdowns	Analysis of Pandemic Impact	Adverse health outcomes post-lockdown
Han et al.[20]	Analyzed qualitative and quantitative plaque characteristics; compared model performance to cardiovascular atherosclerosis risk score	Machine Learning, plaque characteristics	Detecting dataset biases
Anjan N. Repaka et al.[21]	Introduced a model comparing predictive performance of two classification models: outperformed other models in accuracy	Classification models	Comparison with previous research needed
Lapuage et al. [22]	Leveraged hybrid dataset from BRFSS and WHO; addressed class imbalance through sampling techniques; identified Logistic Regression as best-performing model	Hybrid dataset, Logistic Regression	Addressing class imbalance
Suman et al. [12]	Explored gender disparities in CVD prevalence, mortality rates, and disease onset; summarized variations in CVDs by gender	Gender disparity analysis	Under-recognized CVD risk in women

3 Methods and Materials

The methodology is divided into two sections, data acquisition part which describes dataset details and the approach part which includes all the steps followed to get output desired.

3.1 Data Acquisition

In this research, a hybrid dataset has been assembled, comprising two primary sources. The first dataset, "Cardiovascular Disease Risk Prediction Dataset"[7]

was obtained from Kaggle and is a component of the 2021 BRFSS (Behavioral Risk Factor Surveillance System) Dataset provided by the CDC [23]. BRFSS is a prominent nationwide system for conducting health-related telephone surveys, collecting information on the health-related risk behaviors, chronic health conditions, and utilization of preventive services of residents in the United States. The second dataset, "Cardiovascular Disease Dataset" was obtained from Mendeley Data[8].

3.2 Approach

A systematic approach is followed here to improve the performance and the accuracy of the CVD prediction using a hybrid dataset. This process was meticulously designed and implemented in several stages and Figure 2 shows the same.

We start by importing essential Python libraries like NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, and Imbalanced-learn for tasks such as numerical operations, data manipulation, visualization, preprocessing, modeling, and evaluation. Custom functions are created for data exploration, visualization, and generating classification reports. The final dataset is loaded into a Pandas DataFrame. Comprehensive exploratory data analysis (EDA) includes target variable analysis with count plots for 'Heart_Disease' classes, univariate analysis with count plots for categorical features and histograms for numerical ones, and bivariate analysis to examine relationships between features and the target variable.

Data preprocessing involves identifying categorical and numerical features, creating preprocessing pipelines for each, and integrating these steps using ColumnTransformer. Categorical features are one-hot encoded, numerical features are log-transformed and standardized, and ordinal features are encoded with OrdinalEncoder. The machine learning pipeline includes data preprocessing, SMOTE for oversampling the minority class, and training models: Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors and ANN. KNeighborsClassifier uses default parameters, while other models are trained with specified parameters. Models with custom training parameters are given in Table 2. Stratified 10-fold cross-validation evaluates model performance using the F1 score. After training, classification reports for each model include precision, recall, F1-score, and support for both heart disease classes.

Model Evaluation

3.3 Algorithm

Algorithm 1 highlights the systematic approach used to improve the accuracy and performance of cardiovascular disease prediction. Comparative analysis of model outcomes yields classification reports, culminating in a composite mean score for comprehensive evaluation. This algorithmic framework underscores a methodical approach to predictive modeling in cardiovascular health, emphasizing both robustness and interpretability in its predictive outcomes.

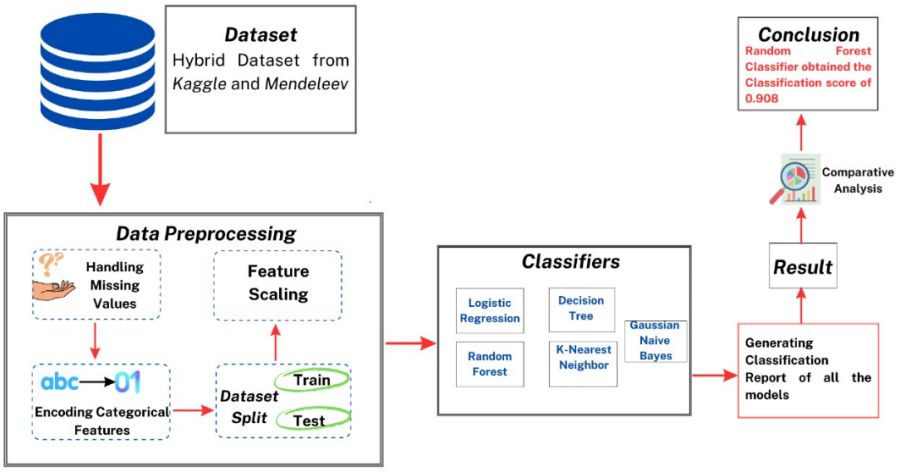

Fig. 2. Flow-chart for the approach

Table 2. PARAMETERS USED FOR TUNING

	Parameter Used	**Parameter Value**
Logistic Regression	max_iter	10000
	random_state	22
Decision Tree Classifier	random_state	22
ANN	epochs	100
	callbacks	early stopping
	batch size	8
Random Forest Classifier	n_estimators	100
	random_state	22

4 Results

In this study, various factors were incorporated into an existing dataset to convert it into a hybrid dataset. The features considered include resting blood pressure, serum cholesterol, and gender, among others[24][11][25].

When examining the relation of resting blood pressure with heart diseases, the data shows that as resting blood pressure increases, the risk of heart disease also rises. Individuals with a blood pressure range of 120-140 mmHg usually do not have heart problems, but as the range increases from 140-190 mmHg, a higher prevalence of heart diseases is observed. This relationship is illustrated in Figure 3.

Algorithm 1 Proposed architecture algorithm

begin
load dataset:
    load the required dataset
data preprocessing:
    check the data for missing values if any and then encode the various features of the dataset for a proper result and to reduce biases
EDA:
    create various univariate and bivariate graphs to get the features relation with each other
    then plot the correlation matrix between the features
pipeline generation:
    create ordered pipeline for both categorical and numerical data for easier case to run the program all together in the correct order
model training:
    use the models logistic regression, decision tree, random forest, k-nearest neighbour, gaussianNB
    the models are applied on two datasets and then the results are compared
    one of the datasets is original and the other is hybrid dataset with more features
return:
    generate classification reports for all and thus create a mean score from them to compare
end

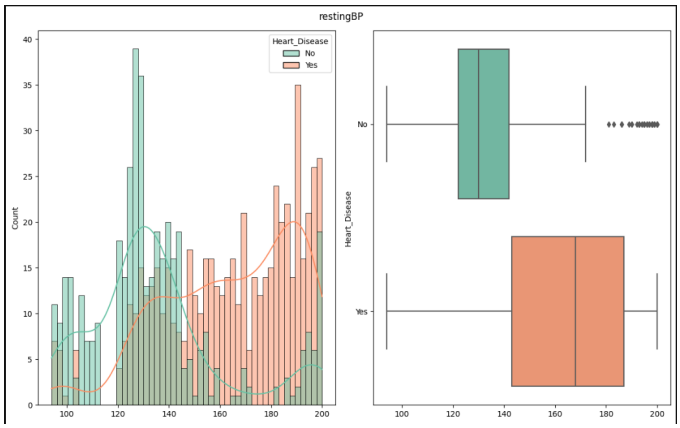

Fig. 3. Resting Blood Pressure relation with heart disease

Similarly, the relation between serum cholesterol and heart disease risk indicates that higher serum cholesterol levels are associated with increased risk. In the range of 220-350 mg/dL, there are mixed cases, but as cholesterol levels rise up to 470 mg/dL, the number of individuals with heart diseases increases significantly, as shown in Figure 4.

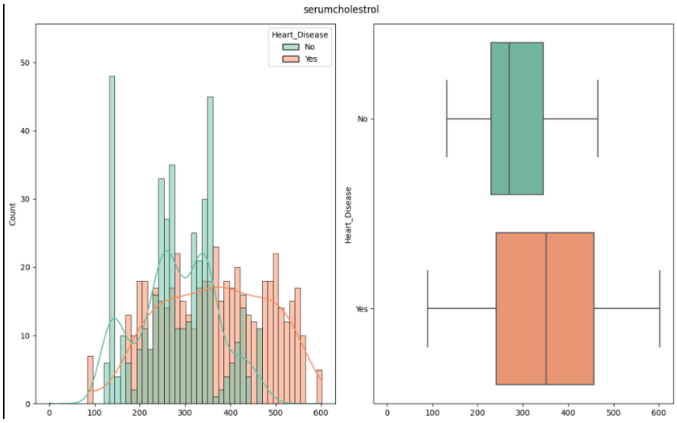

Fig. 4. Serum Cholesterol relation with heart disease

There is also a notable relation between gender and the risk of heart disease. From the dataset, it is observed that women have a 45.3% chance of having heart diseases, while men have a 74.6% chance of having a heart disease. This comparison is visualized in Figure 5.

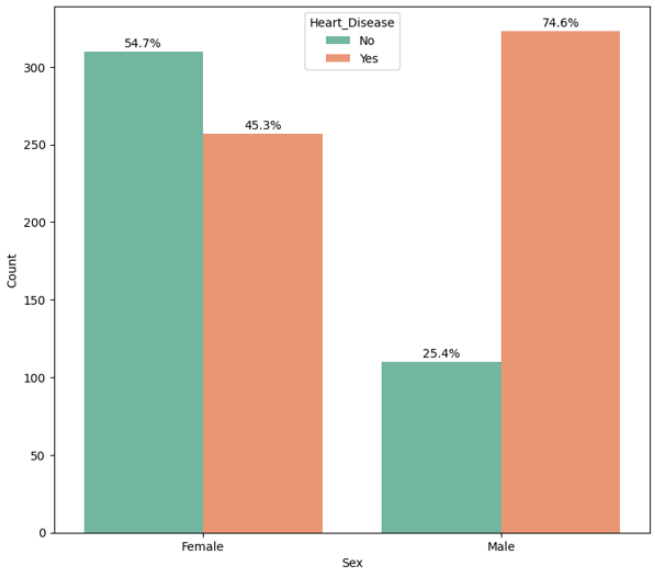

Fig. 5. Gender relation with heart disease

Table 3. ML MODELS APPLIED ON BASIC DATASET

S. No.	ML Model	Mean Score
1	Logistic Regression	0.7771
2	Decision Tree	0.7775
3	Random Forest	0.83558
4	K-Nearest Neighbor	0.71069
5	ANN	0.5800

Table 3 shows the mean scores of different ML models applied to the basic dataset. In contrast, Table 4 displays the mean scores of the ML models applied to the hybrid dataset. The enhanced feature set resulted in significantly improved performance across all models.

Table 4. ML MODELS APPLIED ON THE HYBRID DATASET

S. No.	ML Model	Mean Score
1	Logistic Regression	0.86838
2	Decision Tree	0.83755
3	Random Forest	0.91194
4	K-Nearest Neighbor	0.83335
5	ANN	0.75006

Random Forest achieved the highest accuracy among the models, with a mean score of 0.91 on the hybrid dataset. This superior performance can be attributed to several factors inherent to the Random Forest algorithm. Firstly, its ensemble nature, which combines the predictions of multiple decision trees, significantly reduces the risk of overfitting and enhances predictive accuracy. Secondly, Random Forest is highly effective in handling a diverse set of features, both categorical and numerical, and is robust to missing values. Additionally, its ability to provide insights into feature importance helps in understanding which factors most significantly impact the prediction of heart disease, thus leading to more informed and accurate models. In summary, the study demonstrates that the hybrid dataset, enriched with additional features, significantly improves the performance of various ML models, with Random Forest achieving the highest accuracy due to its robust ensemble learning capabilities.

5 Conclusion

This research highlights the transformative impact of data diversity and machine learning on cardiovascular disease (CVD) risk prediction. By integrating multiple

data sources and examining key factors such as serum cholesterol, fasting blood pressure, and gender, we achieve a more comprehensive understanding of CVD risk[25][15]. The application of machine learning models has significantly improved predictive accuracy, with Random Forest emerging as the top performer, as shown in Figure 6.

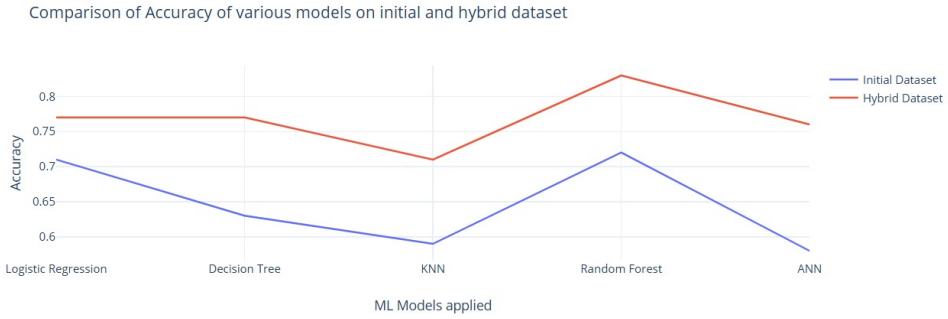

Fig. 6. Graph comparing the scores of both datasets

The study underscores the potential for personalized interventions and enhanced public health outcomes in CVD prevention. The findings offer promise for more effective CVD risk assessment and intervention strategies, representing a crucial step in combating this global health challenge.

Looking ahead, future research can build on this foundation by expanding data sources to include genetic information, lifestyle factors, and socio-economic variables to further refine the precision of CVD risk prediction models[26][27]. Additionally, employing advanced machine learning techniques, such as deep learning and ensemble methods, holds potential to enhance predictive accuracy even more. Integrating real-time data through wearable technology could enable dynamic risk assessment and timely interventions, significantly boosting personalized health monitoring and proactive CVD management[28].

By continually advancing data integration and machine learning methodologies, future work can significantly enhance CVD prediction and prevention strategies. This will not only contribute to better public health outcomes but also help in reducing the global burden of cardiovascular disease, ultimately saving lives and improving quality of life worldwide.

References

1. E.B. Komilovich, EUROPEAN JOURNAL OF MODERN MEDICINE AND PRACTICE **3**(12), 81–87 (2023). URL https://inovatus.es/index.php/ejmmp/article/view/2186

2. H. Mitani, K. Suzuki, J. Ako, K. Iekushi, R. Majewska, S. Touzeni, S. Yamashita, Journal of Atherosclerosis and Thrombosis **30**(11), 1622 (2023). DOI 10.5551/jat.63940. Epub 2023 Mar 16
3. S. Dalal, P. Goel, E.M. Onyema, A. Alharbi, A. Mahmoud, M.A. Algarni, H. Awal, Computational Intelligence and Neuroscience **2023**(1), 9418666 (2023). DOI https://doi.org/10.1155/2023/9418666. URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2023/9418666
4. M.M. Yaqoob, M. Nazir, M.A. Khan, S. Qureshi, A. Al-Rasheed, Applied Sciences **13**(3) (2023). DOI 10.3390/app13031911. URL https://www.mdpi.com/2076-3417/13/3/1911
5. T. Saini, A. Chhabra, in *Artificial Intelligence of Things*, ed. by R.K. Challa, G.S. Aujla, L. Mathew, A. Kumar, M. Kalra, S.L. Shimi, G. Saini, K. Sharma (Springer Nature Switzerland, Cham, 2024), pp. 258–276
6. G. Singh, A. Chhabra, A. Mittal, in *Communication and Intelligent Systems*, ed. by H. Sharma, V. Shrivastava, A.K. Tripathi, L. Wang (Springer Nature Singapore, Singapore, 2024), pp. 1–18
7. Alphiree. Cardiovascular diseases risk prediction dataset (2021). URL https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset
8. B.P. Doppala, D. Bhattacharyya. Cardiovascular disease dataset (2021). Mendeley Data, V1, doi: 10.17632/dzz48mvjht.1
9. AIHW. Heart, stroke and vascular disease: Australian facts (2024). URL https://www.aihw.gov.au/reports/heart-stroke-vascular-diseases/hsvd-facts/contents/disease-types. Accessed: 17 July 2024
10. X. Wang, H. Ma, X. Li, Z. Liang, V. Fonseca, L. Qi, Diabetes, Obesity and Metabolism **26**(4), 1421 (2024). DOI https://doi.org/10.1111/dom.15443. URL https://dom-pubs.onlinelibrary.wiley.com/doi/abs/10.1111/dom.15443
11. A.J. Nelson, N.J. Pagidipati, H.B. Bosworth, Nature Reviews Cardiology **21**(6), 417 (2024). DOI 10.1038/s41569-023-00972-1. © 2024. Springer Nature Limited.
12. S. Suman, J. Pravalika, P. Manjula, U. Farooq, Current Problems in Cardiology **48**(5), 101604 (2023). DOI https://doi.org/10.1016/j.cpcardiol.2023.101604. URL https://www.sciencedirect.com/science/article/pii/S014628062300021X
13. H. Moradi, A. Al-Hourani, G. Concilia, F. Khoshmanesh, F.R. Nezami, S. Needham, S. Baratchi, K. Khoshmanesh, Biophysical Reviews **15**(1), 19 (2023). DOI 10.1007/s12551-022-01040-7. © International Union for Pure and Applied Biophysics (IUPAB) and Springer-Verlag GmbH Germany, part of Springer Nature 2022, Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.
14. A. Isath, K.J. Koziol, M.W. Martinez, C.E. Garber, M.N. Martinez, M.S. Emery, A.L. Baggish, S.S. Naidu, C.J. Lavie, R. Arena, C. Krittanawong, Progress in Cardiovascular Diseases **79**, 44 (2023). DOI 10.1016/j.pcad.2023.04.008. URL https://doi.org/10.1016/j.pcad.2023.04.008. Copyright © 2023 Elsevier Inc. All rights reserved.
15. A. Meloni, C. Cadeddu, L. Cugusi, M.P. Donataccio, M. Deidda, S. Sciomer, S. Gallina, C. Vassalle, F. Moscucci, G. Mercuro, S. Maffei, International Journal of Molecular Sciences **24**(2), 1588 (2023). DOI 10.3390/ijms24021588. URL https://doi.org/10.3390/ijms24021588. The authors declare no conflict of interest.

16. P. Raggi, M. Becciu, E. Navarese, Current opinion in lipidology **35** (2024). DOI 10.1097/MOL.0000000000000921
17. D. Frank, A. Johnson, L. Hausmann, W. Gellad, E. Roberts, R. Vajravelu, Annals of internal medicine **176** (2023). DOI 10.7326/M23-0720
18. C. Krittanawong, H.U.H. Virk, S. Bangalore, Z. Wang, K.W. Johnson, R. Pinotti, H. Zhang, S. Kaplin, B. Narasimhan, T. Kitai, U. Baber, J.L. Halperin, W.H.W. Tang, Scientific Reports **10**(1), 16057 (2020). DOI 10.1038/s41598-020-72685-1. URL https://doi.org/10.1038/s41598-020-72685-1
19. G. Lippi, B.M. Henry, F. Sanchis-Gomar, European Journal of Preventive Cardiology **27**(9), 906 (2020). DOI 10.1177/2047487320916823
20. D. Han, K.K. Kolli, S.J. Al'Aref, L. Baskaran, A.R. van Rosendael, H. Gransar, D. Andreini, M.J. Budoff, F. Cademartiri, K. Chinnaiyan, J.H. Choi, E. Conte, H. Marques, P. de Araújo Gonçalves, I. Gottlieb, M. Hadamitzky, J.A. Leipsic, E. Maffei, G. Pontone, G.L. Raff, S. Shin, Y. Kim, B.K. Lee, E.J. Chun, J.M. Sung, S. Lee, R. Virmani, H. Samady, P. Stone, J. Narula, D.S. Berman, J.J. Bax, L.J. Shaw, F.Y. Lin, J.K. Min, H. Chang, Journal of the American Heart Association **9**(5), e013958 (2020). DOI 10.1161/JAHA.119.013958. URL https://www.ahajournals.org/doi/abs/10.1161/JAHA.119.013958
21. A.N. Repaka, S.D. Ravikanti, R.G. Franklin, in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (2019), pp. 292–297. DOI 10.1109/ICOEI.2019.8862604
22. R. Marcus, J.M. Lupague, R.C. Mabborang, A.G. Bansil, European Journal of Computer Science and Information Technology **11**(3), 44 (2023)
23. Centers for Disease Control and Prevention. 2021 brfss survey data and documentation. {Centers for Disease Control and Prevention} (2021). URL https://www.cdc.gov/brfss/annual_data/annual_2021.html. {Accessed: 2023-07-15}
24. L.S. Mehta, G.P. Velarde, J. Lewey, G. Sharma, R.M. Bond, A. Navas-Acien, A.M. Fretts, G.S. Magwood, E. Yang, R.S. Blumenthal, R.M. Brown, J.H. Mieres, on behalf of the American Heart Association Cardiovascular Disease, S. in Women, U.P.C. of the Council on Clinical Cardiology; Council on Cardiovascular, S.N.C. on Hypertension; Council on Lifelong Congenital Heart Disease, H.H. in the Young; Council on Lifestyle, C.H.C. on Peripheral Vascular Disease;, S. Council, Circulation **147**(19), 1471 (2023). DOI 10.1161/CIR.0000000000001139. URL https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000001139
25. A. Razavi, V. Jain, G. Grandhi, P. Patel, A. Karagiannis, N. Patel, D. Dhindsa, C. Liu, S. Desai, Z. Almuwaqqat, Y. Sun, V. Vaccarino, A. Quyyumi, L. Sperling, A. Mehta, The Journal of clinical endocrinology and metabolism **109** (2023). DOI 10.1210/clinem/dgad406
26. S.A. Claas, S. Aslibekyan, D.K. Arnett, *Genetics of Cardiovascular Disease* (Springer International Publishing, Cham, 2015), pp. 117–127. DOI 10.1007/978-3-319-22357-5_13. URL https://doi.org/10.1007/978-3-319-22357-5_13
27. A.M. Clark, M. DesMeules, W. Luo, A.S. Duncan, A. Wielgosz, Nature Reviews Cardiology **6**(11), 712 (2009). DOI 10.1038/nrcardio.2009.163. URL https://doi.org/10.1038/nrcardio.2009.163
28. A. Mizuno, S. Changolkar, M.S. Patel, Annual Review of Medicine **72**, 459 (2021). DOI 10.1146/annurev-med-050919-031534. URL https://doi.org/10.1146/annurev-med-050919-031534