Cardiac and Cardiovascular Research 2024, 5(1), 2696.
https://doi.org/10.54517/ccr.v5i1.2696

Article

A study of cardio vascular disease prediction and optimization of health care data

L. Pushpalatha*, R. Durga

Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai 600 117, India
*** Corresponding author:** L. Pushpalatha, pushpa2779@gmail.com

CITATION

Pushpalatha L, Durga R. A study of cardio vascular disease prediction and optimization of health care data. Cardiac and Cardiovascular Research. 2024; 5(1): 2696. https://doi.org/10.54517/ccr.v5i1.2696

ARTICLE INFO

Received: 23 April 2024
Accepted: 16 May 2024
Available online: 31 May 2024

COPYRIGHT

Copyright © 2024 by author(s). *Cardiac and Cardiovascular Research* is published by Asia Pacific Academy of Science Pte. Ltd. This work is licensed under the Creative Commons Attribution (CC BY) license. https://creativecommons.org/licenses/by/4.0/

Abstract: Cardiovascular disease (CVD) is a leading cause of morbidity and mortality worldwide, accounting for a significant proportion of healthcare costs. It is a major public health concern, and early detection and prevention are critical for reducing its burden. Several risk factors for CVD have been identified, including age, sex, genetics, hypertension, diabetes, smoking, and physical inactivity. Despite the identification of several risk factors, there is still a lack of understanding of the underlying mechanisms that contribute to CVD development. This study aimed to investigate the potential predictors of CVD and identify novel biomarkers that could be used for early detection and prevention. The study explored pre-processing strategies, including Synthetic Minority Oversampling Technique (SMOTE), Z-Score Normalization, and Adaptive Synthetic Sampling (ADASYN), to address class imbalance and enhance model performance. The dataset consisted of medical images labeled with different cardiovascular diseases. By integrating the strengths of Support Vector Machines (SVM) classification and Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), and PCA with ReliefF feature retrieval methods, the study investigated various feature extraction approaches for classifying cardiovascular diseases.

Keywords: cardio vascular; prediction; pre-processing; classification; feature extraction; principal component analysis; ReliefF; support vector machine

1. Introduction

One of the most difficult diseases to diagnose has developed from it. Heart diseases are becoming more common, according to a recent WHO study. This causes 17.9 million deaths annually [1]. According to Ghosh et al., CVDs are among the most prevalent serious diseases that impair human health. Proper detection may allow for CVD control or prevention, which could lower mortality rates [2]. Huge amount of data is collected by the healthcare sector, some of which include diagnosis-related information for heart disease and are helpful for decision-making [3]. With an accuracy rating 86.60% the RF model was the most precise. An algorithmic technique to forecast the risk of heart disease was created by Rustam et al. [4]. When the predictor's accuracy, recall, precision, and F1-score metrics are determined, the results show that this model achieves, respectively, 98.56%, 99.35%, 97.84%, and 0.983. The strategy suggested for predicting heart-related illness is shown to be effective and reliable by the average AUC score of the model, which reaches 0.983 [5].

2. Related work

The background research on the prediction of CVD offers a thorough examination of the prevalence, contributing variables, and public health importance of

CVD. The background learning determines to set the scene for CVD prediction research and to pinpoint information gaps that the current study tries to fill. The background study starts by outlining the prevalence of CVD globally and highlighting the fact that it is a primary contributor to mortality and death globally.

All disorders impacting the circulation through blood circulation—which comprises the circulatory system and heart that move and carry blood, respectively—are referred to as CVD [6]. Cardiovascular diseases (CVD) encompass a range of disorders affecting the heart and blood vessels, with types including coronary artery disease, stroke, and congenital heart defects, among others [7]. Prediction techniques for CVD have evolved significantly, incorporating machine learning (ML) algorithms to analyze clinical data for early detection and prevention. These algorithms, such as ensemble learning models and oversampling methods, have shown promise in improving prediction accuracy [8]. Additionally, the inclusion of non-traditional risk factors like mental health markers has been explored to enhance predictive models [9]. Interestingly, while ML methods are at the forefront of predictive techniques, challenges such as class imbalance in datasets have been noted, which can be mitigated through oversampling [10]. Moreover, the integration of genetic information, specifically polygenic risk scores, is gaining attention for its potential to refine CVD risk prediction [11]. Traditional risk factors and lifestyle modifications remain integral to primary prevention strategies, as highlighted by global risk score models like the Framingham Risk Score [12]. In summary, the prediction of CVD utilizes a multifaceted approach that includes advanced ML techniques, genetic risk assessment, and traditional risk factor analysis. The combination of these methods aims to enhance the accuracy and early detection of CVD, thereby contributing to better prevention and management strategies [8].

3. Proposed model

When handling the traditional dataset, using various algorithms to find CVD Prediction, performing pre-processing and optimization techniques. Pre-processing is a crucial step in the detection of CVD that aims to develop the superiority and performance of predictive models by streamlining and optimizing datasets.

Z-Score normalization: A popular pre-processing method for scaling characteristics in a dataset to have a mean of 0 and an SD of 1 is called Z-score normalization, sometimes referred to as standardization.

3.1. SMOTE

One well-known method for building an unbalanced dataset, a classification algorithm, is SMOTE. The underlying output classes are distributed unevenly in an imbalanced dataset. SMOTE is frequently utilized in problems with classification with unbalanced datasets.

3.2. ADASYN

A method for addressing class imbalance in ML datasets is called ADASYN. It creates artificial samples for the minority class, emphasizing cases that are challenging to categorize. Another closest neighbor-based method that resembles the SMOTE

technique is ADASYN. The primary distinction between them is that, as opposed to data samples that are simpler to learn, ADASYN concentrates more on minority data samples that are more difficult to learn.

3.3. Evaluation metrics

Accuracy analysis:

Accuracy=Number of correctly predicted individuals with CVD+Number of correctly predicted individuals without CVD/ Total Number of individuals in the dataset—(3.1).

Precision and recall analysis:

Precision=Number of correctly predicted individuals with CVD/Total number of predicted individuals with CVD—(3.2).

Recall=Number of correctly predicted individuals with CVD/Total number of individuals with CVD—(3.3).

By discovering instructive patterns apparent in the information that has been identified as pertinent for classification, feature extraction plays a critical part in CVD prediction. This provides background information for a thorough analysis of feature extraction techniques used in the suggested CVD prediction system by examining approaches including LDA, PCA, and PCA combined with Relief.

A crucial aspect to take into account is the amount and caliber of features employed. Using as many features as possible is the current trend in machine learning approaches.

PCA with ReliefF:

Using a hybrid approach that combines the strengths of PCA and ReliefF, this method applies ReliefF to the principal components generated by PCA. This approach efficiently selects significant variables for CVD classification by considering both inter-class and intra-class variances. This hybrid methodology, which incorporates PCA's dimensionality reduction capabilities and ReliefF's ability to identify key features, was used to obtain features. By applying ReliefF to the principal components generated by PCA, we were able to rapidly identify relevant characteristics while accounting for both inter-class and intra-class differences.

3.4. SVMHA

Research on CVD prediction is essential to improve early identification and preventative methods. To optimize predictive performance, the suggested approach combines SVMHA (Support Vector Machine Heuristic approach) with a variety of pre-processing and feature extraction techniques.

To produce precise and trustworthy predictions, this methodology integrates pre-processing, feature selection, classification, and optimization techniques. In total, it provides a methodical approach to CVD prediction. Enhancing early identification and preventive interventions requires research on CVD prediction. The effective classification of individuals based on many clinical characteristics and risk factors, including age, gender, bp, fate rate, and lifestyle choices, is facilitated by ML methods, which are crucial in this field. These methods make it possible to classify people based on their level of risk of developing CVD—low, moderate, or high—which helps with

focused intervention techniques. SVM in conjunction with KHO is one of the sophisticated ML approaches used to maximize predicted accuracy. SVMHA is an integrated strategy that combines many pre-processing and feature extraction approaches with advanced model optimization strategies.

4. Result and discussion

This research article provides an in-depth examination of the effectiveness of machine learning (ML) algorithms in predicting cardiovascular disease (CVD). These algorithms use various techniques to analyze large sets of medical data and identify patterns or features that are indicative of cardiovascular disease. It details the accuracy, precision, and recall rates of decision tree (DT), random forest (RF), support vector machine (SVM), and SVM high-dimensional analysis (SVMHA) algorithms. These insights are crucial for enhancing diagnostic capabilities and creating reliable predictive models that aid in early diagnosis and intervention in CVD management.

The study's findings reveal the relative performance of ML algorithms across different datasets, pre-processing techniques, and evaluation criteria. It shows that SVMHA outperforms other algorithms on Kaggle, UCI, and Data world datasets.

5. Conclusion

The study's conclusions emphasize the potential of the proposed method to revolutionize CVD prediction and enhance patient outcomes by implementing personalized healthcare plans and early intervention. This work is in line with the growing body of research that uses predictive analytics to address significant healthcare challenges by employing advanced ML techniques and rigorously evaluating their performance. This will eventually lead to more effective CVD risk assessment and management in clinical practice. Future research should explore additional preprocessing methods to improve the quality and robustness of predictive models for CVD prediction. Additionally, by integrating domain-specific knowledge with expert-driven feature engineering techniques, new biomarkers and risk factors for CVD may be identified, expanding the feature space and improving prediction accuracy.

Author contributions: Conceptualization, LP and RD; methodology, LP; software, LP; validation, LP and RD; formal analysis, LP; investigation, LP; resources, LP; data curation, LP; writing—original draft preparation, LP; writing—review and editing, LP; visualization, LP; supervision, LP; project administration, LP; funding acquisition, RD. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Jebaseeli TJ, M NK, Subagar A, et al. Cardio Vascular Disease Prediction and Classification Report Generation using Data Mining Technique. In: Proceedings of the 3rd International Conference on Smart Data Intelligence (ICSMDI).
2. Ghosh P, Azam S, Jonkman M, et al. Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques. IEEE Access. 2021; 9: 19304-19326. doi: 10.1109/access.2021.3053759

3. Alalawi HH. Detection of Cardiovascular Disease using Machine Learning Classification Models. International Journal of Engineering Research & Technology (IJERT). 2021; 10(7).
4. Rustam F, Ishaq A, Munir K, et al. Incorporating CNN Features for Optimizing Performance of Ensemble Classifier for Cardiovascular Disease Prediction. Diagnostics. 2022; 12(6): 1474.
5. Zhang D, Chen Y, Chen Y, et al. Heart Disease Prediction Based on the Embedded Feature Selection Method and Deep Neural Network. Journal of Healthcare Engineering. 2023; 6260022. doi: 10.1155/2021/6260022
6. Thiriet M. Cardiovascular Disease: An Introduction. Springer; 2018.
7. Sale Elsedawy HF. Overview of the Most Prevalent Pediatric Congenital Heart Diseases: A Literature Review. Asian Journal of Pediatric Research. 2023.
8. Erkan AKKUR. Prediction of Cardiovascular Disease Based on Voting Ensemble Model and SHAP Analysis. Sakarya University Journal of Computer and Information Sciences. 2023.
9. Dorraki M. Cardiovascular disease risk prediction via machine learning using mental health data. European Heart Journal—Digital Health. 2022.
10. García-Vicente C, Soguero-Ruiz C, Mora-Jiménez I, et al. Clinical Synthetic Data Generation to Predict and Identify Risk Factors for Cardiovascular Diseases. In: Lecture Notes in Computer Science. Springer; 2022.
11. Abraham G, Rutten-Jacobs L, Inouye M. Risk Prediction Using Polygenic Risk Scores for Prevention of Stroke and Other Cardiovascular Diseases. Stroke. 2021; 52(9). doi: 10.1161/strokeaha.120.032619
12. Rodriguez F, Foody JM. Primary Prevention of Cardiovascular Disease. In: Stergiopoulos K, Brown D (editors). Evidence-Based Cardiology Consult. Springer; 2014. doi: 10.1007/978-1-4471-4441-0_12