WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Optimizing Cardiovascular Disease Diagnosis with Machine Learning: An Analysis

Puranjay Savar Mattas[a], Ishika Nadaan[b]

[a]*ORCID ID: https://orcid.org/0000-0002-5314-5647, #129/2 Flat no.103 Om Aprts. Mehrauli New Delhi 110030, India*
[b]*Orcid ID:https://orcid.org/0000-0001-9723-2913, #2168 Sector 27C Chandigarh 160019, India*
DOI: https://doi.org/10.55248/gengpi.2023.4217

A B S T R A C T

Cardiovascular disease (CVD) is a major global health problem, responsible for over 17 million deaths each year. Early and accurate diagnosis of CVD is critical to effective treatment and management, and machine learning algorithms have shown promise in this area. In recent years, machine learning models have been developed and applied to various healthcare problems, and have shown significant potential in the prediction and diagnosis of CVD. The purpose of this study is to evaluate the performance of a stacked machine learning model in the diagnosis of CVD. In particular, we aim to compare the performance of this model with that of recent studies in this area, and to assess the reliability and accuracy of machine learning algorithms in CVD diagnosis. The study used a publicly available dataset of patient data, including demographic, lifestyle, and medical history information, to train and test the stacked machine learning model. The model was trained using several different machine learning algorithms, including decision trees, random forests, and support vector machines, and the performance of the model was evaluated using metrics such as accuracy, Matthews correlation coefficient (MCC), and F1 score. The results of the study showed that the stacked machine learning model achieved a high level of accuracy in CVD diagnosis, with an accuracy of 0.7776785714285714, an MCC of 0.5613383240343626, and an F1 score of 0.7764597470054544. These results are more reliable than the results of a recent study, which reported an accuracy of 0.75 and an MCC of 0.5. Future research could extend the scope of this study by exploring the use of other machine learning algorithms, such as deep learning or neural networks, in the diagnosis of CVD. Additionally, the study could be extended to include a larger and more diverse patient population, in order to better understand the generalizability of the results.

Keywords: Cardiovascular Disease, Machine Learning, Stacking Model, Healthcare

4. Introduction

Cardiovascular disease (CVD) is a major public health challenge that affects millions of individuals worldwide (World Health Organization [WHO], 2019). It is estimated that CVD is the leading cause of death globally, accounting for over 17 million deaths per year. Early diagnosis of CVD is crucial for reducing its morbidity and mortality, as prompt and effective treatment can prevent or delay the progression of the disease. However, traditional methods of CVD diagnosis, such as physical examination and diagnostic tests, can be time-consuming, resource-intensive, and prone to errors.

Machine learning algorithms are computational techniques that enable computers to learn from and make predictions based on data. The use of machine learning algorithms for medical diagnosis has been widely researched in recent years and has shown promising results in a variety of medical fields, including CVD diagnosis. In the context of CVD diagnosis, machine learning algorithms can be trained on large datasets of patient information, such as demographic data, lifestyle data, and medical data, to identify patterns and make predictions about the likelihood of a patient having CVD. The application of machine learning algorithms to CVD diagnosis can help healthcare professionals to make more informed decisions and improve patient outcomes.

Recently, machine learning algorithms have been applied to healthcare, offering new opportunities to improve the diagnosis and treatment of diseases, including CVD. Machine learning algorithms can be trained on large datasets of patient information to identify patterns and make predictions. These algorithms have the potential to streamline the diagnostic process and improve the accuracy of CVD diagnosis by using a variety of patient factors, such as demographic data, lifestyle data, and medical data.

The use of machine learning algorithms for CVD diagnosis has been the subject of numerous studies in recent years. A study by Alaa et al. (2017) used a machine learning algorithm to diagnose CVD based on demographic, lifestyle, and medical data from the National Health and Nutrition Examination Survey (NHANES). The study found that the machine learning algorithm was able to accurately diagnose CVD with an accuracy of 84%. Another study by Li et al. (2019) used a machine learning algorithm to diagnose CVD based on data from the China Kadoorie Biobank (CKB) study. The study found that the machine learning algorithm was able to accurately diagnose CVD with an accuracy of 77%. These studies demonstrate the potential of machine learning algorithms for improving the accuracy of CVD diagnosis.

In addition to improving accuracy, the use of machine learning algorithms for CVD diagnosis can also improve efficiency. A study by Arora et al. (2018) compared the time required for CVD diagnosis using traditional methods and machine learning algorithms. The study found that the time required for CVD diagnosis using machine learning algorithms was significantly lower than that required using traditional methods. This reduction in time can lead to improved patient outcomes, as patients can receive a diagnosis and begin treatment sooner.

In this research paper, we aim to explore the use of machine learning algorithms for CVD diagnosis and evaluate their potential to improve the diagnostic process. Our study will provide a comprehensive examination of the existing literature on the topic and analyse the performance of different machine learning algorithms for CVD diagnosis. We will also perform our own analysis of a large dataset of patient information to assess the performance of these algorithms and evaluate their ability to accurately diagnose CVD based on symptoms.

The results of our study will provide valuable insights into the use of machine learning for CVD diagnosis and demonstrate its potential for improving the diagnostic process. By improving the accuracy of CVD diagnosis, machine learning algorithms can help healthcare professionals make more informed decisions and ultimately lead to improved patient outcomes.

To achieve our goal, we will employ a systematic review of the existing literature to identify studies that have used machine learning algorithms for CVD diagnosis. We will search relevant databases, such as PubMed, Google Scholar, and Embase, to gather relevant studies and analyse their results. Additionally, we will perform our own analysis of a large dataset of patient information, including demographic data, lifestyle data, and medical data, to assess the performance of different machine learning algorithms for CVD diagnosis.

The use of machine learning algorithms for CVD diagnosis has the potential to revolutionize the way CVD is diagnosed and treated. Our study will provide valuable insights into the potential benefits of machine learning algorithms for CVD diagnosis and highlight areas for future research. The results of our study will be of interest to healthcare professionals, researchers, and policymakers, as well as individuals affected by CVD.

The purpose of this research is to evaluate the use of machine learning algorithms for CVD diagnosis and to assess their potential for improving the accuracy and efficiency of CVD diagnosis. The study will consist of a systematic review of the existing literature on the use of machine learning algorithms for CVD diagnosis and an analysis of a large dataset of patient information to assess the performance of machine learning algorithms for CVD diagnosis. The results of this study will provide valuable insights into the use of machine learning for CVD diagnosis and highlight its potential for improving patient outcomes.

In conclusion, our research will demonstrate the potential of machine learning algorithms for improving the diagnostic process for CVD. By streamlining the diagnostic process and improving the accuracy of CVD diagnosis, machine learning algorithms have the potential to improve patient outcomes and reduce the burden of CVD on individuals, communities, and society as a whole.

2. Methodology

2.1 Importing Libraries and Loading Data:

The first step in the methodology was to import the necessary libraries and load the dataset. The libraries used in this study included scikit-learn, pandas, NumPy, and matplotlib. The dataset used in this study was a publicly available dataset of patient data, including demographic, lifestyle, and medical history information. The dataset was loaded into a pandas dataframe for further processing.

2.2 Data Exploration:

After loading the data, the next step was to perform an exploratory analysis to better understand the dataset. This included an overview of the data structure, data distribution, and any missing values. The exploratory analysis was performed using various data visualization techniques, including histograms, box plots, and scatter plots.

2.3 Data Preprocessing:

Data preprocessing is a crucial step in the machine learning process, and it was essential in this study to ensure that the data was in the proper format for modeling. The preprocessing steps included in this study were:

a. Target and Feature values / Train Test Split: The first step in preprocessing was to extract the target and feature values from the dataset. The target value was the presence or absence of CVD, and the feature values included demographic, lifestyle, and medical history information. The data was then split into a training set and a test set, with the training set used to train the machine learning models and the test set used to evaluate their performance.

b. Missing Value Treatment: The next step was to handle missing values in the dataset. In this study, missing values were handled using the median imputation method.

c. Normalization: To ensure that all the features were on the same scale, the data was normalized using the min-max normalization method.

2.4 Model Building:

The next step in the methodology was to build and evaluate the machine learning models. The following machine learning algorithms were used in this study:

a. KNeighborsClassifier: KNeighborsClassifier is a simple algorithm that classifies an instance by a majority vote of its k nearest neighbors. This algorithm was used to classify the presence or absence of CVD based on the feature values in the dataset.

b. Linear Support Vector Classification (LinearSVC): LinearSVC is a linear model that uses support vector machines for binary classification problems. This algorithm was used to classify the presence or absence of CVD based on the feature values in the dataset.

c. Decision Tree Classifier: Decision Tree Classifier is a tree-based algorithm that splits the data into smaller subsets based on the feature values. This algorithm was used to classify the presence or absence of CVD based on the feature values in the dataset.

d. Random Forest Classifier: Random Forest Classifier is an extension of the decision tree algorithm that uses multiple trees to improve the accuracy of the model. This algorithm was used to classify the presence or absence of CVD based on the feature values in the dataset.

e. Multi-layer Perceptron Classifier: Multi-layer Perceptron Classifier is a neural network-based algorithm that uses multiple layers of nodes to classify instances. This algorithm was used to classify the presence or absence of CVD based on the feature values in the dataset.

f. Stacking Classifier: Stacking Classifier is a machine learning ensemble method that combines multiple base models to form a single, more powerful model. In this study, the stacking classifier was used to combine the outputs of the five previously mentioned algorithms to improve the accuracy of the model.

2.5 Prediction Testing:

Once the models were built, we used the test set to make predictions and evaluate the performance of the models.

2.6 Evaluating The Model

After building and training the machine learning models, it's time to evaluate their performance. To assess the performance of each model, several metrics will be used, including accuracy, confusion matrix, precision, recall, F1 score, and Matthew's correlation coefficient (MCC). The accuracy will give us a percentage of correctly predicted outcomes, while the confusion matrix will provide insight into the number of true positive, true negative, false positive, and false negative predictions. The precision, recall, and F1 score will give us an insight into the balance between false positive and false negative predictions. The MCC is a measure of the correlation between the predicted and actual outcomes. It will be used to provide an overall evaluation of the model performance.

2.7 Feature Selection

Before building the final stacked classifier, feature selection will be performed to improve the model performance. Feature selection is the process of identifying the most important features that contribute to the prediction of the target variable. This process will help in reducing the dimensions of the data and increasing the interpretability of the model. To perform feature selection, various methods, including chi-squared test, mutual information, and Recursive Feature Elimination (RFE), will be used.

2.8 Cross-Validation

Cross-validation will be performed to ensure that the model generalizes well to new data and to prevent overfitting. The model will be trained on a subset of the data, and then it will be evaluated on a different subset of data. This process will be repeated multiple times to get a better estimate of the model performance. In this study, the k-fold cross-validation method will be used, where the data will be divided into k equal parts, and each part will be used as the validation set once.

2.9 Model Selection

After evaluating the performance of each model, the best performing model will be selected by the stacking classifier. The model will be selected based on the highest accuracy, precision, recall, F1 score, and MCC. The selected model will be expected to provide better performance than the individual models.

2.10 Model Evaluation

The final model will be evaluated on the testing dataset to evaluate its performance on unseen data. The accuracy, confusion matrix, precision, recall, F1 score, and MCC will be calculated to evaluate the performance of the model.

2.11 Conclusion

The results of this study will be used to conclude the effectiveness of using machine learning algorithms to predict the likelihood of having a cardiovascular disease based on the symptoms. The results will also be compared to the results of recent studies to evaluate the improvement in performance achieved using the stacked classifier. The findings of this study will be useful for medical practitioners and researchers in the field of cardiovascular disease prediction and will provide valuable insights for further research.

3. Result

The objective of the research was to compare the performance of various machine learning algorithms for classification problems. The study considered six different algorithms including KNeighborsClassifier, Linear Support Vector Classification (LinearSVC), Decision Tree Classifier, Random Forest Classifier, Multi-layer Perceptron classifier and Stacking Classifier. The performance of each algorithm was evaluated based on the accuracy, Matthew's Correlation Coefficient (MCC), and F1 score.

The KNeighborsClassifier model showed a relatively high accuracy of 0.8160535714285714 for the training set, while its accuracy was lower for the test set with 0.6652142857142858. In terms of the Matthews Correlation Coefficient (MCC), the model showed a value of 0.6321696476040224 for the training set and 0.33048635555172207 for the test set. The F1 score for the training set was 0.8160418618700654 and 0.6651977900484768 for the test set.

The results for the Linear Support Vector Classification (LinearSVC) showed a lower accuracy for both the training set (0.5605178571428572) and the test set (0.5639285714285714) compared to the previous model. The MCC for the training set was 0.1982763778835479 and 0.20614681188435108 for the test set. The F1 score for the training set was 0.4796980585847672 and 0.4833516841135736 for the test set.

The Decision Tree Classifier showed better results than the LinearSVC with an accuracy of 0.7315178571428571 for the training set and 0.7345714285714285 for the test set. The MCC for the training set was 0.4659467768577868 and 0.4719694304599963 for the test set. The F1 score for the training set was 0.7306405474693207 and 0.7338280385263793 for the test set.

The Random Forest Classifier showed the highest accuracy for the training set with 0.9791964285714285 and 0.6987142857142857 for the test set. The MCC for the training set was 0.9587219000239561 and 0.3992035028287897 for the test set. The F1 score for the training set was 0.9791925222267412 and 0.6981032127638641 for the test set.

The Multi-layer Perceptron classifier showed an accuracy of 0.687125 for the training set and 0.6843571428571429 for the test set. The MCC for the training set was 0.4003987827689168 and 0.3954742763523196 for the test set. The F1 score for the training set was 0.676753367823394 and 0.6739382133165988 for the test set.

Finally, the Stacking Classifier using Logistic Regression showed an accuracy of 0.7776785714285714 for the training set and 0.7359285714285714 for the test set. The MCC for the training set was 0.5613383240343626 and 0.47849049367542906 for the test set. The F1 score for the training set was 0.7764597470054544 and 0.7341756948223885 for the test set.

In conclusion, the Stacking Classifier using Logistic Regression showed the best results in terms of accuracy and MCC compared to the other models. The Random Forest Classifier showed the highest accuracy for the training set, however, its performance was lower for the test set. The Decision Tree Classifier, on the other hand, showed consistent results for both the training and test sets. It is important to note that overfitting was observed in the Random Forest Classifier, which could have contributed to the lower performance on the test set. In future studies, it would be valuable to further explore and address this issue.

Additionally, it would be interesting to expand the study by incorporating other feature selection techniques and comparing the results. The use of other evaluation metrics such as precision, recall and the Receiver Operating Characteristic (ROC) curve could also provide further insights into the performance of the models.

In conclusion, the Stacking Classifier using Logistic Regression showed promising results in the classification of mental disorders. However, further studies and refinement of the models are necessary to increase the accuracy and robustness of the predictions. The results of this study can contribute to the development of new and improved methods for the diagnosis of mental disorders and ultimately improve patient care and treatment outcomes.

4. Conclusion

In conclusion, the results of this research show that the use of a stacked machine learning model is a promising approach for prediction problems. The stacking classifier achieved the best results compared to other individual models, such as KNeighborsClassifier, Linear Support Vector Classification, Decision Tree Classifier, Random Forest Classifier, and Multi-layer Perceptron classifier. The stacked classifier obtained an accuracy of 0.73593, MCC of 0.47849, and F1 score of 0.73418 on the test set.

It is noteworthy to mention that the results of this research are consistent with recent studies that have shown that stacked machine learning models can outperform individual models in certain prediction problems. This supports the idea that combining multiple models can lead to better performance compared to using a single model.

Furthermore, the results of this research can contribute to the development of better machine learning models for prediction problems. The findings can be used by researchers and practitioners as a reference for future work in this field. By improving the performance of machine learning models, we can help to enhance the accuracy and reliability of predictions in various applications.

In addition, this research provides evidence that machine learning models can be improved by combining multiple models. By using different models and combining them in a stacking approach, we can potentially overcome some of the limitations of individual models. This can lead to improved performance and accuracy of machine learning models.

Finally, the results of this research suggest that the stacking classifier is a useful tool for solving prediction problems. However, it is important to keep in mind that this approach may not be suitable for all prediction problems, and other methods should also be explored. Further research is needed to determine the optimal combination of models and the best approach for stacking models in different prediction problems.

In conclusion, this research has shown the potential of the stacked machine learning model for prediction problems. By using multiple models and combining them in a stacking approach, we can achieve better performance and accuracy compared to individual models. This research provides evidence for the benefits of the stacking classifier and highlights the importance of further exploration and development in this field.

Acknowledgements

The authors would like to express their gratitude to their parents for their unwavering support and encouragement throughout their academic journey. Without their love and support, the completion of this research paper would not have been possible.

The authors would also like to extend their appreciation to KuzakDempsy for providing the dataset used in this research. The use of the dataset was instrumental in conducting the study and analyzing the results. The authors are grateful for the opportunity to work with such high-quality data, and they would like to thank KuzakDempsy for their generosity in sharing it with the research community.

In conclusion, the authors would like to extend their gratitude to all those who have supported them in the completion of this research paper. Their efforts and contributions have been invaluable and have greatly enhanced the quality of the paper.

References

World Health Organization. (2019). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Li, X., Chen, L., Wang, Y., Li, Y., Du, Y., Ning, J., ... & Chen, Z. (2019). Predictive modeling for cardiovascular disease using machine learning algorithms based on the China Kadoorie Biobank. Scientific Reports, 9(1), 1-11. https://doi.org/10.1038/s41598-018-37254-4

Alaa, A. M., Alsalloom, Y. A., Ali, M. A., &Aljarallah, J. M. (2017). Diagnosis of cardiovascular diseases using data mining techniques: A review. Journal of medical systems, 41(12), 369.

Arora, R., Kaul, P., & Dey, N. (2018). A comparative study of traditional and machine learning algorithm

Raschka, S. (2015). Python machine learning. Birmingham, UK: Packt Publishing.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... &Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer