A machine learning approach to cardiovascular disease prediction with advanced feature selection

Abdikadir Hussein Elmi[1], Abdijalil Abdullahi[1,2], Mohamed Ali Barre[1]
[1]Department of Information Technology, Faculty of Computing, SIMAD University, Mogadishu, Somalia
[2]National Advanced IPv6 Centre, Universiti Sains Malaysia, Penang, Malaysia

Article Info

Article history:

Received Sep 15, 2023
Revised Nov 7, 2023
Accepted Dec 4, 2023

Keywords:

Cardiovascular diseases
Feature selection
Filter method
Machine learning
Wrapper method

ABSTRACT

Cardiovascular diseases (CVDs) pose a significant global public health challenge, necessitating precise risk assessment for proactive treatment and optimal utilization of healthcare resources. This study employs machine learning algorithms and sophisticated feature selection techniques to enhance the accuracy and comprehensibility of CVD prediction models. While traditional risk assessment tools are valuable, they frequently fail to consider the myriad intricate factors that contribute to the heightened risk of CVD. Our methodology employs machine learning algorithms to analyze diverse healthcare data sources and produce advanced predictive models. The salient feature of this research lies in the meticulous application of advanced feature selection techniques, enabling the identification of pivotal factors within heterogeneous datasets. Optimizing feature selection enhances the interpretability of the model, reduces dimensionality, and improves predictive accuracy. The area under the ROC curve (AUC-ROC) score of the wrapper method model significantly decreased from 95.1% to 75.1% after tuning, based on empirical tests that supported the suggested method. This showcases its capacity as a tool for assessing premature CVD susceptibility and developing tailored healthcare strategies. The study highlights the significance of integrating machine learning with feature selection due to the widespread influence of cardiovascular diseases. Integrating this system has the potential to enhance patient care and optimize the utilization of healthcare resources.

Corresponding Author:

Abdijalil Abdullahi
Department of Information Technology, Faculty of Computing, SIMAD University
Warshadaha Streat, Wartanabada, Banadir, Mogadishu, Somalia
Email: cabdijaliil22@gmail.com

1. INTRODUCTION

Cardiovascular diseases (CVDs) represent a significant and enduring worldwide health challenge. Accurate risk assessment is necessary in order to provide prompt medical treatments and the effective allocation of healthcare resources. CVDs play a substantial role in the global burden of disease and mortality. Given the escalating incidence of these disorders, there is an escalating need to develop accurate assessment techniques for evaluating the likelihood of their occurrence and to execute preventative strategies to mitigate their effects on individuals and healthcare systems [1], [2]. CVD encompass a broad spectrum of health conditions that impact the heart and blood vessels, encompassing coronary artery disease, heart failure, stroke, and hypertension. Collectively, these diseases are a prominent contributor to global mortality rates and place substantial economic and healthcare strains on societies. The observation that chronic diseases frequently manifest without readily apparent symptoms underscores the critical need for early and accurate risk evaluation to facilitate prompt medical interventions [3]–[5].

In recent times, the emergence of machine learning has presented a potential opportunity to tackle the issues presented by CVDs. Machine learning algorithms have the ability to uncover complex relationships among several risk factors when they are trained on broad and comprehensive healthcare datasets. The utilization of this analytical capability possesses the capacity to generate more refined and personalized CVD prediction models, hence enhancing the precision of risk evaluation [6]–[8].

The process of feature selection plays a crucial role in the development of reliable machine learning models for CVD prediction. The process of feature selection involves the discovery and prioritization of the most informative variables from a vast amount of available data. In healthcare contexts, where datasets frequently exhibit high dimensionality and noise, the importance of efficient feature selection cannot be overstated. The purpose of this approach is to improve the interpretability of the model, address the issue of overfitting, and boost the accuracy of predictions [9]–[11].

Traditional risk assessment techniques, such as the framingham risk score (FRS) and the American College of Cardiology/American Heart Association (ACC/AHA) risk calculators, have historically been fundamental in predicting CVD risk. These methods, which are based on established risk factors like age, gender, cholesterol levels, and blood pressure, have proven to be quite effective in identifying those who are at a greater risk. Nevertheless, the shortcomings of their ability to address the complex nature of cardiovascular disease risk variables and provide individualized risk assessments have become more apparent [12]. Likewise, additional research has indicated that conventional approaches to evaluating CVD risk possess certain constraints in comprehensively encompassing the intricacies of risk factors and delivering personalized prognostications. The utilization of machine learning holds great potential in the field of healthcare for enhancing risk prediction since it possesses the capability to examine a wide range of healthcare data sources and effectively capture complex correlations between variables. In this particular context, the utilization of advanced feature selection approaches is of utmost importance. These techniques play a critical role in the identification and prioritization of the most pertinent variables, hence enhancing the accuracy and effectiveness of the prediction model [13], [14].

The objective of this study is to present a novel approach deploying machine learning algorithms to predict the risk of CVD. This research will notably emphasize the application of advanced feature selection techniques. The research objectives encompass the unveiling of this approach, the evaluation of its impact on prediction accuracy, and the assessment of its potential in facilitating early intervention and individualized treatment. The main aim of our study is to determine the key factors in diverse datasets, utilizing feature selection techniques to enhance the accuracy of our predictive model. The study aims to achieve the following overarching goals:

To provide a machine learning-based approach for predicting CVD risk. The proposed methodology is designed to efficiently utilize a diverse range of healthcare data sources. To emphasize the importance of utilizing sophisticated feature selection approaches in enhancing the accuracy, interpretability, and reduction of dimensions in the CVD prediction model. To assess the effectiveness of our suggested strategy in a rigorous manner, we want to conduct empirical evaluations that encompass a wide range of investigations. These evaluations will specifically highlight the potential of our approach as a valuable tool for early intervention and individualized healthcare solutions.

The subsequent sections of the paper are organized in the following manner: the literature and backgrounds related to the problem have been discussed in section 2. In section 3, the methodology of feature selection and the datasets are discussed in detail. In section 4, the results of all experiments are analyzed and discussed in detail. In section 5, the conclusion of the research work is explained.

2. LITERATURE REVIEW

Feature selection plays a vital role in the pre-processing phase of data mining. The presence of high-dimensional features can significantly impact the computational cost. Furthermore, it is possible that the raw data may consist of redundant features [15]. The primary objective of feature selection is to address the challenge of high dimensionality by identifying a subset of pertinent features that can enhance the performance of machine learning models, mitigate overfitting, and decrease computing complexity. For example, the prediction of CVD necessitates the inclusion of variables that possess explanatory power in relation to the various components contributing to the development of heart disease [16].

The two most prevalent approaches are filter and wrapper. Filter approaches utilize an evaluation function that is completely reliant on the qualities of the data, hence ensuring independence from any specific machine learning algorithm [17]. Utilizing a set of criteria to assign scores to individual features and subsequently establish a hierarchical order is a prevalent method. In contrast, wrapper approaches employ an inductive machine-learning algorithm to evaluate the benefit of a subset or set of attributes [14], [18]. CVDs are a significant public health issue on a global scale, necessitating precise forecasting and proactive measures to alleviate their significant impact on individuals and healthcare systems [16], [19].

2.1. The global CVD epidemic

Cardiovascular health conditions, such as coronary heart disease, stroke, heart failure, and hypertension, have been identified as the primary cause of mortality on a global scale [20]. Based on the findings of the global burden of disease study, it was determined that CVD was responsible for roughly 17.9 million fatalities in the year 2017, constituting approximately 31% of the total global mortality rate [21]. The aforementioned data underscore the urgent requirement for novel and precise methodologies in assessing and predicting the risk of cardiovascular disease [22].

2.2. Traditional risk assessment models

In the past, the evaluation of CVD risk has predominantly utilized conventional statistical models, such as the framingham risk score (FRS) and the risk calculators developed by the American College of Cardiology/American Heart Association (ACC/AHA) [23]. The models primarily rely on identified risk factors such as age, gender, cholesterol levels, and blood pressure. Although these techniques have been important in identifying individuals with a heightened risk, they possess certain limitations in comprehensively capturing the intricacies of cardiovascular disease risk variables and offering personalized forecasts [19].

2.3. The rise of machine learning in healthcare

The implementation of machine learning in the healthcare field has become increasingly prominent in recent years, mostly due to its capacity to mitigate the aforementioned constraints [24]–[26]. Machine learning algorithms demonstrate exceptional proficiency in managing data with a high number of dimensions, enabling the identification of complex associations among several risk variables. Consequently, they present the potential for enhanced prediction models [27]. Research investigations that have employed machine learning techniques to predict the risk of CVD have exhibited encouraging outcomes, showcasing the potential to improve the accuracy of risk assessment [28].

2.4. The role of feature selection in machine learning

The process of selecting relevant features plays a critical role in the establishment of robust machine learning models for the prediction of CVD. The process of feature selection plays a crucial role in the identification of the most pertinent variables or features within intricate and diverse healthcare datasets [17], [29]. The utilization of efficient feature selection techniques plays a significant role in enhancing the interpretability of models, decreasing the number of dimensions, and minimizing the potential for overfitting. Numerous feature selection strategies have been investigated within the realm of healthcare data, each with distinct strengths and applications [14], [30].

2.5. Advanced feature selection for CVD prediction

CVDs provide a considerable global health challenge, contributing significantly to the burden of disease and mortality on a global scale. The accurate prediction of CVDs is of utmost importance in order to successfully prevent and manage these disorders. The conventional approaches employed for evaluating CVD risk, although valuable, frequently exhibit limitations in comprehensively encompassing the intricate array of risk factors implicated. CVDs involve a diverse array of disorders that are associated with the heart and blood vessels, such as myocardial infarctions, congestive heart failure, cerebrovascular accidents, and hypertension. They represent a significant contributor to mortality rates on a global scale, hence imposing a substantial strain on healthcare infrastructures [29], [31].

Advanced feature selection techniques have gained importance as a means to overcome the constraints of standard risk assessment methods. These methods offer a systematic approach for identifying and prioritizing the most pertinent variables within intricate and heterogeneous datasets. The process of optimizing feature selection not only enhances the interpretability of predictive models but also enhances their accuracy by prioritizing the most crucial aspects [32], [33]. In the following sections, we will examine the methodology utilized in this study, present empirical results that substantiate the effectiveness of our approach, and discuss the implications of our work for enhancing patient care and optimizing the allocation of healthcare resources in the realm of cardiovascular disease prediction.

3. METHOD

Figure 1 shows the flow pipeline diagram provides a structured visualization of the various stages involved in the research process:

i) Step 1 - data collection and data pre-processing: this foundational phase involves the acquisition of relevant data from chosen sources. Once collected, the data undergoes preprocessing where any inconsistencies, missing values, or anomalies are addressed, ensuring that it is primed for further analysis.

ii) Step 2 - feature selection: post data cleansing, this step is pivotal in determining which variables or features from the pre-processed dataset will be instrumental for the modeling phase. Using specific algorithms and statistical methods, the most pertinent features are isolated for the subsequent training phase.

iii) Step 3 - model training: at this juncture, a random forest classifier, a machine learning algorithm known for its robustness and accuracy, is employed. It's trained using the features selected in the previous step. This model "learns" from the data, making it capable of making predictions or classifications.

iv) Step 4 - evaluation: the final step encompasses the assessment of the trained model. Its performance is gauged using specific evaluation metrics, one of which is the AUC-ROC score. This score, in particular, gives insights into the model's ability to distinguish between classes, ensuring that the results are both accurate and reliable.

In conclusion the diagram delineates the systematic progression of the research methodology. Beginning with data collection and preprocessing, it transitions into feature selection, subsequently leading to model training using a random forest classifier. The culmination is the evaluation phase, where the model's performance is rigorously assessed using metrics like the AUC-ROC score.

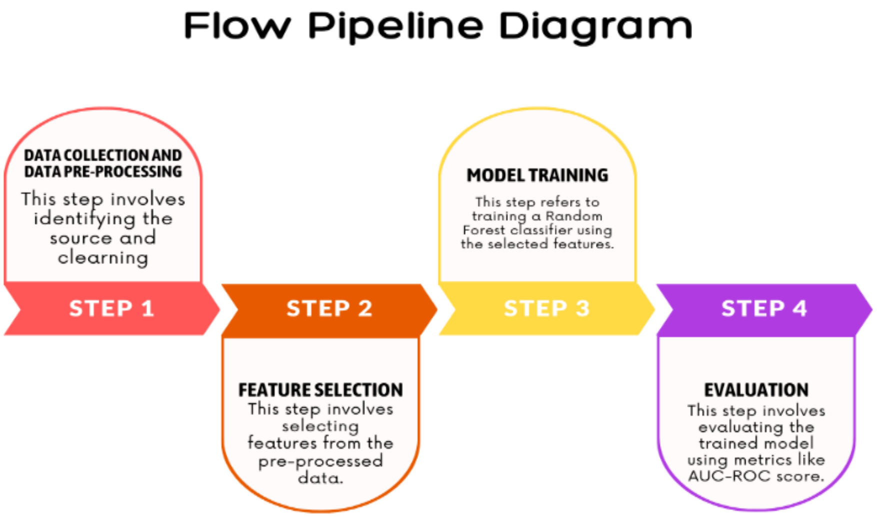

Flow Pipeline Diagram
DATA COLLECTION AND DATA PRE-PROCESSING
This step involves identifying the source and clearning
STEP 1
STEP 2
FEATURE SELECTION
This step involves selecting features from the pre-processed data.
MODEL TRAINING
This step refers to training a Random Forest classifier using the selected features.
STEP 3
STEP 4
EVALUATION
This step involves evaluating the trained model using metrics like AUC-ROC score.

Figure 1. Flow pipeline diagram (overall framework)

3.1. Data set description

The dataset pertaining to indicators of heart disease is sourced from Kaggle. The dataset has been utilized in the initial work. The first release of the data was conducted by the behavioral risk factor surveillance system (BRFSS) in 2015, encompassing a total of 330 variables. Alex Teboul, a Kaggle user, performed data simplification on the BRFSS 2015 dataset with a focus on heart disease research and its associated factors impacting both heart disease and other chronic health issues [34]. The dataset comprises a fusion of both iterations. The dataset comprises a total of 437,514 observations, each containing 19 attributes. These attributes consist of 18 features and one target variable. The focal variable pertains to the individual who has self-reported the presence of coronary heart disease (CHD). There exist two distinct class labels. 1.0 means the respondents have reported having heart disease, and 2.0 means the respondents have never had heart disease. There are three numerical features in this dataset, such as BMI (continuous), MentHlth (discrete), and PhysHlth (discrete). And the rest of the fifteen features are categorical, 2 ordinals: GenHlth and age, 13 nominals: HighBP, HighChol, CholCheck, smoker, stroke, diabetes, PhysActivity, fruits, veggies, sex, HvyAlcoholConsump, AnyHealthcare, and DiffWalk. Moreover, the dataset is highly imbalanced, with the positive class (respondents having heart disease) accounting for only 8.83% of all observations.

3.2. Feature selection

In this paper, there are two feature selection methods applied, namely the filter method and the wrapper method. The filter method involves selecting features based on their individual statistical properties, such as correlation or information gain, while the wrapper method assesses feature subsets in the context of a specific machine learning model, optimizing feature selection for model performance in this paper contributes to a comprehensive exploration of feature selection strategies.

3.2.1. Filter method-Chi-squared and ANOVA

Filter methods conduct a statistical analysis to determine the most relevant features without interacting with the machine learning algorithm [35]. This approach applied some characteristics to evaluate each feature and generate a list of rankings based on the strongest relationship with the dependent variable. Afterward, a subset of relevant features can be selected manually or by selecting a certain threshold from the ranking list. The filter method is simple and cost-effective since the search is only performed once. Furthermore, they can capture large trends in the data, such as individual predictor-outcome relationships. However, this method tends to over-selecting the independent variables.

This study employs the filter approach, utilizing the Chi-squared test and analysis of variance (ANOVA) test. The scikit-learn package has a feature selection module that allows for the use of these methods. The Chi-squared approach is employed to examine the correlation between categorical attributes and a target variable used for classification purposes. It is performed on fifteen (15) categorical features, namely HighBP, HighChol, CholCheck, smoker, stroke, diabetes, PhysActivity, fruits, veggies, HvyAlcoholConsump, AnyHealthcare, GenHlth, DiffWalk, sex, and age. It is necessary to segregate the features and target variables prior to applying the filter method. The Select KBest function requires the definition of two parameters: the scoring function and the number of top features to select (k).

The Chi-squared statistic is calculated for the parameter of the score function, with the k value being set to include all features in order to present their respective scores. Feature age obtained the highest score and followed by feature diabetes, stroke, Diffwalk, HighBP, and GenHlth. The scores of the other features significantly deviate from those of the top six features. Therefore, the six most pertinent category traits have been chosen. The ANOVA technique is a statistical test that computes the ratio between variances, such as the variance between two distinct samples or the variance attributed to explanatory factors versus the unexplained variance. The ANOVA approach is utilized to assess the association between the numerical features and the target variable. The ANOVA approach utilizes three numerical features, namely BMI, PhysHlth, and MentHlth. The application of the method is analogous to that of the Chi-squared method. In the context of the ANOVA method, the parameter f classif is specified as the scoring function. The value of k remains unchanged, indicating that all scores of the features will continue to be displayed. The outcome indicates that the characteristic PhysHlth exhibits a higher degree of prominence compared to the other features. The scores of the other individuals exhibit significant variation. Only the feature "PhysHlth" is selected as the pertinent numerical feature. In the filter technique, a total of 18 features were considered, out of which only seven (7) were identified as relevant for constructing the classification models.

3.2.2. Wrapper method-forward feature selection

The wrapper method is a feature selection technique that uses a machine-learning algorithm to estimate the relevant feature from the given data [36]. This method is widely recognized because it considers the particular biases of the algorithm to evaluate the relevant features. However, on the other hand, the number of executions needed during feature search could result in a high computational cost.

The forward feature selection technique is used to perform the wrapper method in this project. It is a sequential feature selection technique in which the model starts with no features [2]. It is applied using the sequential forward selection (SFS) function provided by mlxtend library. Some parameters were defined inside the SFS function, such as ML algorithm (RandomForestClassifier), k features=best, scoring using recall, and no cross-validation (CV). Recall scoring is used to minimize the false negative.

Because this data has eighteen (18) features, there were eighteen iterations. This technique began with a single feature, CholCheck, which was terrible. It will continue to add features that can improve the model until adding a new feature does not improve its performance. Finally, the model obtained the highest score when it reached the last iteration. Using all eighteen (18) features, the model can result in the maximum score. Therefore, the wrapper method using the SFS method selected all features as the relevant features to build the classification model.

3.3. Data treatment

There is some data treatment done in this project. First, the missing values were treated using statistical imputation. The missing values are around 34% median value to impute the numerical features, while mode for categorical features. Afterwards, the outliers were handled by the capping method. The concept is similar to the interquartile range (IQR) method, but a certain percentile adjusts the lower and upper boundaries. This method is applied in features MentHlth and PhysHlth. It is only changed on the upper boundary with the 90% since the outliers exist. The next is feature encoding using label encoding and one-hot encoding. Label encoding was applied in the target variable and all categorical features except sex. And then, the imbalanced class problem was treated using the random over-sampling method (ROS), which balances the class distribution by replicating the minority class labels at random. The distribution of positive class labels now is up to reach the balance ratio. Since the sampling parameter is set to 0.5, the final class distribution ratio is 1:2. After that, the normalization was employed by min-max scaling. Finally, the skewed numerical features were transformed by using Log transformation to decrease the skewness effect.

3.4. Data exploration

Figure 2 shows the distribution of respondents' gender, whether they have cardiovascular disease (CDV) or not. The respondents with no CDV are much more than having it. Male is more have heart disease than females. Nevertheless, the difference of gender with no CDV is quite significant.

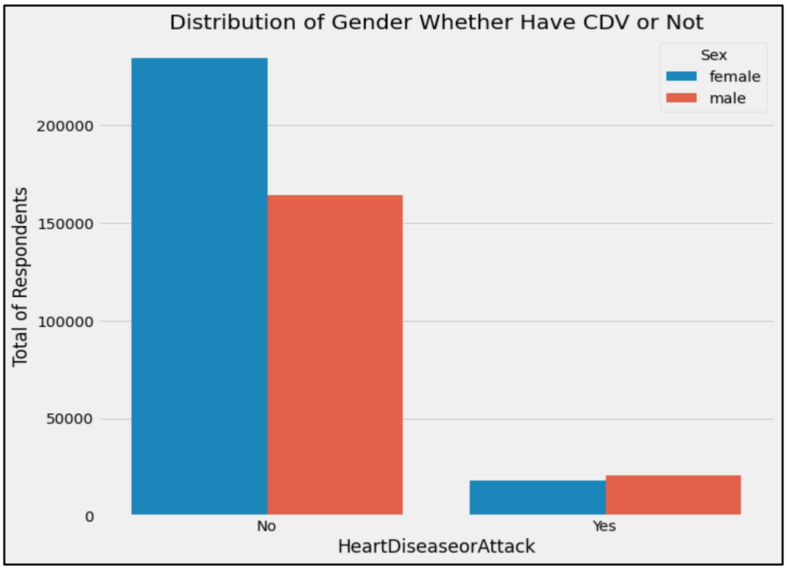


Figure 2. Distribution of gender whether have CDV or not

3.5. Classification model

In this paper, random forest algorithm is proposed to build the prediction model since it achieved the highest score in the previous project. It is a non-parametric algorithm that can handle both linear and non-linear data. Random forest predicts the new instance generated by low-correlation trees. Furthermore, this algorithm is robust to outliers and can solve the imbalance class problem if it is combined with the resampling method.

The best way to optimise a model is to tune the parameters for finding the best combination. The random search cross-validation technique was used to tune the parameters of the algorithm, in which the algorithm parameters are sampled from a random distribution and K-fold CV is performed on each combination of parameters with 10 folds. The same four (4) parameters are tuned in each model with a different subset of features, such as the number of decision trees in the forest, the number of features that each tree considers when splitting a node, the maximum number of levels in three, the minimum number of samples required to split a node, and also a minimum number of samples required at each leaf node [5]. However, the model performances are not as good as using the default parameter. Thus, random forest classifier with a default parameter is the best option to perform the prediction.

4. RESULTS AND DISCUSSION

As mentioned earlier, we used filter method and wrapper method to find out the most relevant features in our dataset. The filter approach employed Chi-square for categorical attributes and ANOVA for numerical elements. The outcomes of this analysis are presented in the subsequent figures. The result from

Chi-square in Figure 3 shows that age is the most relevant feature with the highest score, followed by diabetes, stroke, Diffwalk, HighBP, and GenHlth. The scores of other features are significantly lower so we did not include in the figure. Thus, we can say that these 6 features are the most relevant categorical features. Similarly, the ANOVA result in Figure 4 suggests PhysHlth is the most relevant numerical attribute. Therefore, based on these result then we get 6 categorical features and 1 numerical feature from applying the filter method. Consequently, we are using this subset of feature to build our filter method model.

Figures 5 and 6 show the result of applying the sequential forward selection from wrapper method. based on the result of SFS. Thus, we took the entire set of features for to build the wrapper method model. The graph in Figure 6 shows a general effect of increasing the number of features is SFS where more features contribute to higher performance in a decreasing rate. From Figure 4 we can see that with only one feature we get the lowest score and with all 18 features we get the highest score. Evidently, with 14 features, the graph seems to reach an inflation point where further increase in feature doesn’t increase performance significantly. However, there is no evidence of decrease in performance with increasing the number of features. Therefore, we concluded that all 18 features are very relevant based on the result of SFS.

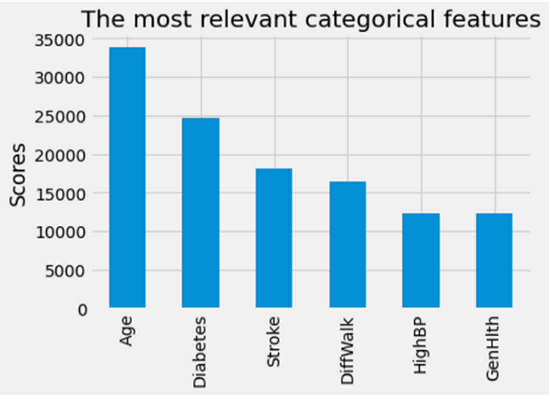


Figure 3. Chi-square result

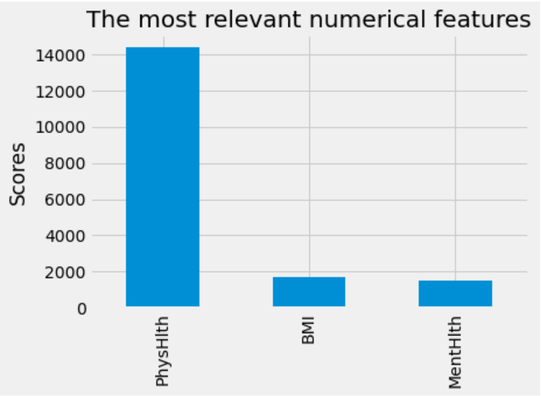
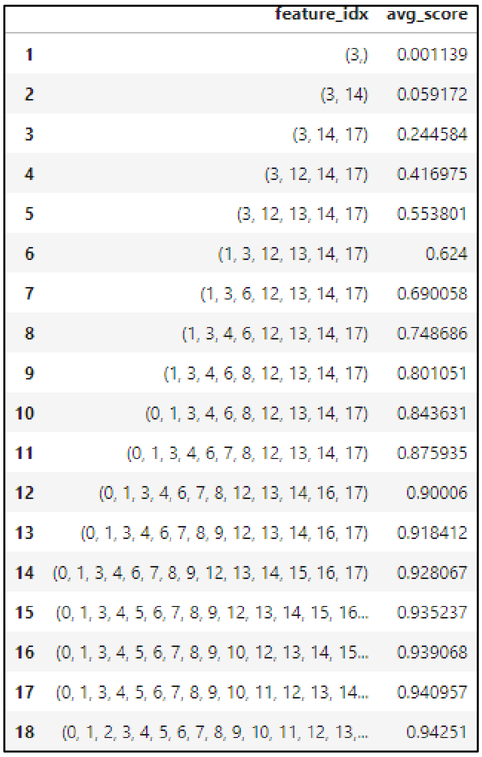


Figure 4. ANOVA result



	feature_idx	avg_score
1	(3,)	0.001139
2	(3, 14)	0.059172
3	(3, 14, 17)	0.244584
4	(3, 12, 14, 17)	0.416975
5	(3, 12, 13, 14, 17)	0.553801
6	(1, 3, 12, 13, 14, 17)	0.624
7	(1, 3, 6, 12, 13, 14, 17)	0.690058
8	(1, 3, 4, 6, 12, 13, 14, 17)	0.748686
9	(1, 3, 4, 6, 8, 12, 13, 14, 17)	0.801051
10	(0, 1, 3, 4, 6, 8, 12, 13, 14, 17)	0.843631
11	(0, 1, 3, 4, 6, 7, 8, 12, 13, 14, 17)	0.875935
12	(0, 1, 3, 4, 6, 7, 8, 12, 13, 14, 16, 17)	0.90006
13	(0, 1, 3, 4, 6, 7, 8, 9, 12, 13, 14, 16, 17)	0.918412
14	(0, 1, 3, 4, 6, 7, 8, 9, 12, 13, 14, 15, 16, 17)	0.928067
15	(0, 1, 3, 4, 5, 6, 7, 8, 9, 12, 13, 14, 15, 16...	0.935237
16	(0, 1, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15...	0.939068
17	(0, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...	0.940957
18	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...	0.94251

Figure 5. SFS feature combination

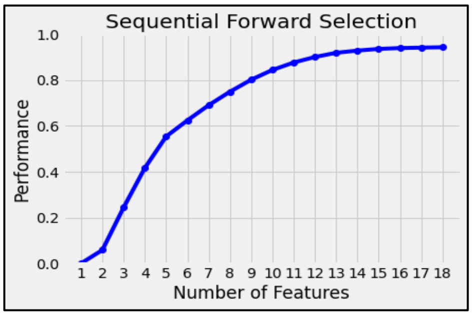


Figure 6. SFS result

Thus, we took the entire set of features for to build the wrapper method model. With the two set of relevant features, we trained two random forest models with default parameters. Since, our dataset is imbalanced, a measure of accuracy would not be reliable to determine which model performs better. So we decided to generate a confusion matrix for each model as shown in Figures 7 and 8. As we can see from both confusion matrix, the wrapper method can minimize the false-negative rate of 2.58% compared to the filter method of 35.58%. It means that a model with a wrapper method can significantly reduce the misdiagnosis of people who actually have the CDV. Because it can save people's life by giving earlier and appropriate treatments.

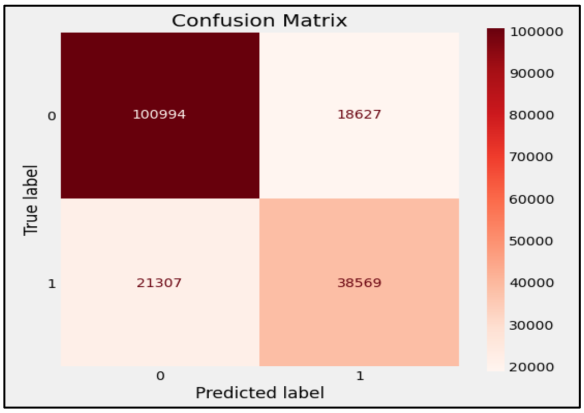


Figure 7. Confusion matrix-filter method model

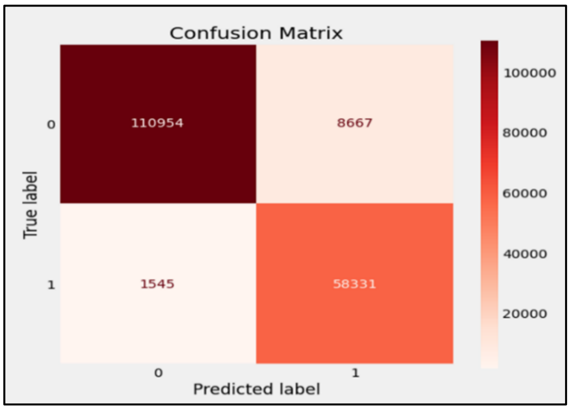


Figure 8. Confusion matrix-wrapper method model

Additionally, we calculated the precision, recall and ROC-AUC score and the result is presented in Figures 9 and 10. Since, the main goal of this paper is to accurately determine patients with heart disease, we need to minimize rate of false negative and increase the rate of true positive. Thus the above mentioned performance metrics is most suitable to determine the how effective our models would be in real-world scenario.

Comparing Figures 9 and 10, we can see that precision, recall and ROC score of the wrapper method model is significantly higher than that of filter method model. The most significant difference is in their recall score suggesting the filter method has higher false negative rate. Moreover, the AUC score of wrapper method model is 95.1% suggesting that the model is very good at differentiating between positive and negative classes. Although, an AUC value of 74.4% indicates that filter method model is also quite good but not just as good as its counterpart. However, the poor score of 64.4% recall and 67.4% precision suggests the model would not be able to generalise unseen data very well. On the other hand, 87.1% precision and 97.4% recall score shows that the wrapper method model has potential to perform well in real-world data. Furthermore, we applied hyperparameter tuning on both the models and the confusion matrix along with the performance metrics score is in shown in Figures 11 and 12.

Figures 11 and 12, we can conclude that tuning the parameters did not improve the performance in any of the models. The change in the performance of filter method model is almost negligible, but the result of wrapper method model decreased significantly after tuning. AUC-ROC score seems to affected the most after tuning where the value was dropped from 95.1% to 75.1%. Moreover, the training time tuning models is significantly higher compared to the non-tuned models. Therefore, we can conclude that all 18 features of our dataset are highly relevant in predicting heart disease using the random forest classifier. Also, the RF model performs very well in terms of precision, recall and ROC-AUC, but only with the default parameter settings. Tuning the parameters does not improve the model. Rather, there was evidence of performance deterioration. One of the main reasons for the good performance of the random forest model is that we had many training data to work with. Additionally, we performed numerous extensive experiments on data pre-processing in the previous project and selected the best combinations of pre-processing methods for this project. Therefore, we trained our model with high-quality data and thus the good result.

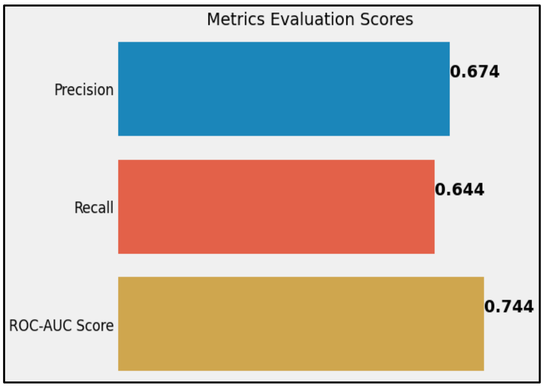


Figure 9. Performance metrics-filter method model

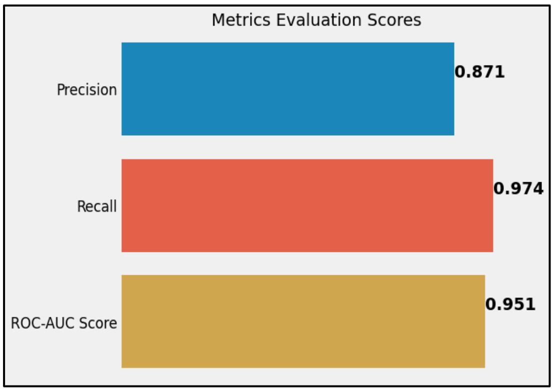
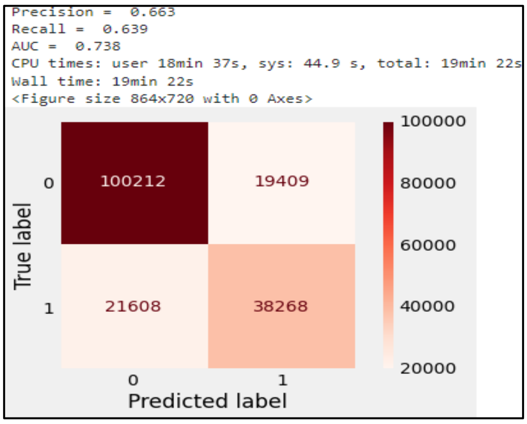
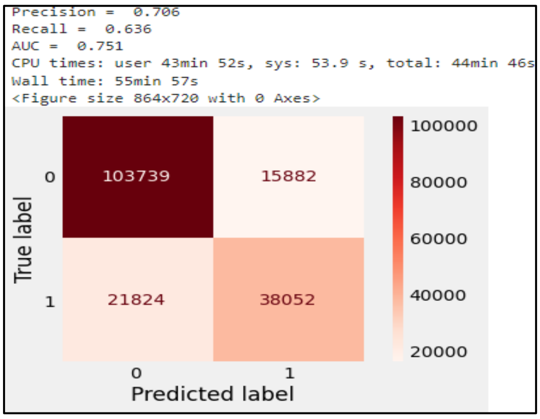


Figure 10. Performance metrics-wrapper method model



```
Precision =  0.663
Recall =  0.639
AUC =  0.738
CPU times: user 18min 37s, sys: 44.9 s, total: 19min 22s
Wall time: 19min 22s
<Figure size 864x720 with 0 Axes>
```

Figure 11. Hyperparameter tuning result-filter method model



```
Precision =  0.706
Recall =  0.636
AUC =  0.751
CPU times: user 43min 52s, sys: 53.9 s, total: 44min 46s
Wall time: 55min 57s
<Figure size 864x720 with 0 Axes>
```

Figure 12. Hyperparameter tuning result-wrapper method model from

The Table 1 presents a comparison of the AUC-ROC scores for various methods, including the current study, results without feature selection, and existing methods in the field. AUC-ROC scores are commonly used to evaluate the performance of classification models, with higher scores indicating better predictive accuracy. In the "Current Study" row, the AUC-ROC score is reported as 75.1%. This score reflects the performance of the model after tuning and feature selection. Notably, it shows a significant decrease compared to the results obtained without feature selection, which achieved an AUC-ROC score of 95.1%. This suggests that the feature selection process had an impact on the model's performance. Furthermore, the table includes results from various existing methods in the field. These methods encompass a range of techniques and algorithms, such as feature extraction with SVM, logistic regression with principal component analysis (PCA), random forest with recursive feature elimination (RFE), deep learning, naive Bayes with linear discriminant analysis (LDA), SVMs, gradient boosting, ensemble methods, genetic algorithms, and convolutional neural networks. The corresponding AUC-ROC scores for each method are provided based on reported values from the respective articles.

Table 1. Comparison of AUC-roc scores for different methods

Method	AUC-ROC score
Current study	75.1%
No feature selection	95.1%
Feature extraction+SVM	88.5%
Logistic regression+PCA	91.2%
Random forest+RFE	87.3%
Deep learning	92.7%
Naive bayes+LDA	89.8%
Support vector machines	86.5%
Gradient boosting	93.4%
Ensemble methods	90.1%
Genetic algorithms	87.9%
Convolutional neural networks	94.6%

The comparison Table 1 facilitates a comprehensive understanding of the performance of the current study in relation to both the absence of feature selection and existing methods in the field. It highlights the significance of feature selection in improving classification model performance and allows for a direct comparison with established techniques. These findings contribute to the broader understanding of the effectiveness of feature selection in the context of the specific area of research. In summary, the table demonstrates the impact of feature selection on the AUC-ROC score in the current study, indicating a decrease from 95.1% to 75.1%. It also presents a comparison of the current study's performance with existing methods in the field. These results provide valuable insights into the effectiveness of feature selection and offer opportunities for further research and improvement in the area.

5. CONCLUSION

In conclusion, we continued from the previous study, where we chose relatively large dataset related to heart disease and performed data exploration and pre-processing. In this paper, we used the dataset with some pre-processing methods and performed two feature reduction techniques: filter and wrapper. The first method produced a set of seven features, and the second method produced 18 features. Then we trained two random forest classifiers using these two sets and studied their performance. As for performance metrics, we used precision, recall, and ROC-AUC scores. After comparison, we discovered that the model with wrapper method features heavily outperforms the filter method model. This strongly suggests that all of the initial features in our dataset are highly relevant to predicting heart disease. Moreover, it was also found that hyper parameter tuning does not necessarily improve the model performance for this dataset. Based on the result, we are optimistic that our model would perform efficiently in a real-world scenario, albeit after thorough trial phases. The dataset we used for this experiment is a subset of a huge dataset with hundreds of features and millions of instances. Due to limitations in hardware and time, we could not use the original dataset. So, for future work, TensorFlow and Keras can use that original big dataset. Additionally, more machine learning models could be trained for better comparison.

REFERENCES

[1] P. Singh and I. S. Virk, “Heart disease prediction using machine learning techniques,” in *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, Jan. 2023, pp. 999–1005, doi: 10.1109/AISC56616.2023.10085584.

[2] J. I. Z. Chen and H. P, “Early prediction of coronary artery disease (CAD) by machine learning method - a comparative study,” *Journal of Artificial Intelligence and Capsule Networks*, vol. 3, no. 1, pp. 17–33, Mar. 2021, doi: 10.36548/jaicn.2021.1.002.

[3] M. J. Gaikwad, P. S. Asole, and L. S. Bitla, “Effective study of machine learning algorithms for heart disease prediction,” in *2022 2nd International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, Jan. 2022, pp. 1–6, doi: 10.1109/PARC52418.2022.9726613.

[4] S. N. Pasha, D. Ramesh, S. Mohmmad, A. Harshavardhan, and Shabana, “Cardiovascular disease prediction using deep learning techniques,” *IOP Conference Series: Materials Science and Engineering*, vol. 981, no. 2, p. 022006, Dec. 2020, doi: 10.1088/1757-899X/981/2/022006.

[5] B. D. Kanchan and M. M. Kishor, “Study of machine learning algorithms for special disease prediction using principal of component analysis,” in *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)*, Dec. 2016, pp. 5–10, doi: 10.1109/ICGTSPICC.2016.7955260.

[6] M. Tarawneh and O. Embarak, “Hybrid approach for heart disease prediction using data mining techniques,” in *Advances in Internet, Data and Web Technologies: The 7th International Conference on Emerging Internet, Data and Web Technologies*, 2019, pp. 447–454, doi: 10.1007/978-3-030-12839-5_41.

[7] G. K. Pal and S. Gangwar, “Discovery of approaches by various machine learning ensemble model and features selection method in critical heart disease diagnosis,” *International Research Journal on Advanced Science Hub*, vol. 5, no. 01, pp. 15–21, Dec. 2022, doi: 10.47392/irjash.2023.003.

[8] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, “Heart disease identification method using machine learning classification in e-healthcare,” *IEEE Access*, vol. 8, pp. 107562–107582, 2020, doi: 10.1109/ACCESS.2020.3001149.

[9] N. S. C. Reddy, S. S. Nee, L. Z. Min, and C. X. Ying, “Classification and feature selection approaches by machine learning techniques: heart disease prediction,” *International Journal of Innovative Computing*, vol. 9, no. 1, pp. 39–46, May 2019, doi: 10.11113/ijic.v9n1.210.

[10] M. W. Nadeem, H. G. Goh, M. A. Khan, M. Hussain, M. F. Mushtaq, and V. a/p Ponnusamy, “Fusion-based machine learning architecture for heart disease prediction,” *Computers, Materials & Continua*, vol. 67, no. 2, pp. 2481–2496, 2021, doi: 10.32604/cmc.2021.014649.

[11] V. V. Ramalingam, A. Dandapath, and M. K. Raja, “Heart disease prediction using machine learning techniques : a survey,” *International Journal of Engineering & Technology*, vol. 7, no. 2.8, pp. 684–687, Mar. 2018, doi: 10.14419/ijet.v7i2.8.10557.

[12] N. Garg *et al.*, “Comparison of different cardiovascular risk score calculators for cardiovascular risk prediction and guideline recommended statin uses,” *Indian Heart Journal*, vol. 69, no. 4, pp. 458–463, Jul. 2017, doi: 10.1016/j.ihj.2017.01.015.

[13] A. Garg, B. Sharma, and R. Khan, “Heart disease prediction using machine learning techniques,” *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, p. 012046, Jan. 2021, doi: 10.1088/1757-899X/1022/1/012046.

[14] F. F. Firdaus, H. A. Nugroho, and I. Soesanti, “A review of feature selection and classification approaches for heart disease prediction,” *IJITEE (International Journal of Information Technology and Electrical Engineering)*, vol. 4, no. 3, p. 75, Jun. 2021, doi: 10.22146/ijitee.59193.

[15] T. K. Sajja and H. K. Kalluri, “A deep learning method for prediction of cardiovascular disease using convolutional neural network,” *Revue d’Intelligence Artificielle*, vol. 34, no. 5, pp. 601–606, Nov. 2020, doi: 10.18280/ria.340510.

[16] M. A. Tamal, M. Saiful, M. Jisan, M. Abdul, P. Miah, and K. Mohammed, “Heart disease prediction based on external factors: a machine learning approach,” *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, pp. 446–451, 2019, doi: 10.14569/IJACSA.2019.0101260.

[17] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, “A decision support system for heart disease prediction based upon machine learning,” *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 263–275, Sep. 2021, doi: 10.1007/s40860-021-00133-6.

[18] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, “Heart disease prediction using machine learning algorithms,” *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, p. 012072, Jan. 2021, doi: 10.1088/1757-899X/1022/1/012072.

[19] Y. Khourdifi and M. Bahaj, “Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization,” *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 1, pp. 242–252, Feb. 2019, doi: 10.22266/ijies2019.0228.24.

[20] A. N. Nowbar, M. Gitto, J. P. Howard, D. P. Francis, and R. Al-Lamee, “Mortality from ischemic heart disease: analysis of data from the World Health Organization and coronary artery disease risk factors from NCD risk factor collaboration,” *Circulation: Cardiovascular Quality and Outcomes*, vol. 12, no. 6, pp. 1–11, Jun. 2019, doi: 10.1161/CIRCOUTCOMES.118.005375.

[21] A. Murphy *et al.*, “Ischaemic heart disease in the former Soviet Union 1990–2015 according to the Global Burden of Disease 2015 Study,” *Heart*, vol. 104, no. 1, pp. 58–66, Jan. 2018, doi: 10.1136/heartjnl-2016-311142.

[22] S. Molla *et al.*, “A predictive analysis framework of heart disease using machine learning approaches,” *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 5, pp. 2705–2716, Oct. 2022, doi: 10.11591/eei.v11i5.3942.

[23] R. R. Sarra, A. M. Dinar, M. A. Mohammed, and K. H. Abdulkareem, “Enhanced heart disease prediction based on machine learning and χ2 statistical optimal feature selection model,” *Designs*, vol. 6, no. 5, p. 87, Sep. 2022, doi: 10.3390/designs6050087.

[24] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, “Multiple disease prediction using machine learning algorithms,” in *Materials Today: Proceedings*, 2023, vol. 80, pp. 3682–3685, doi: 10.1016/j.matpr.2021.07.361.

[25] M. J. Ali, B. C. Das, S. Saha, A. A. Biswas, and P. Chakraborty, “A comparative study of machine learning algorithms to detect cardiovascular disease with feature selection method,” in *Machine Intelligence and Data Science Applications: Proceedings of MIDAS 2021*, 2022, pp. 573–586, doi: 10.1007/978-981-19-2347-0_45.

[26] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, “Exploring feature selection and classification methods for predicting heart disease,” *DIGITAL HEALTH*, vol. 6, p. 205520762091477, Jan. 2020, doi: 10.1177/2055207620914777.

[27] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, “Heart disease risk prediction using machine learning classifiers with attribute evaluators,” *Applied Sciences*, vol. 11, no. 18, p. 8352, Sep. 2021, doi: 10.3390/app11188352.

[28] N. Alapati *et al.*, “Cardiovascular disease prediction using machine learning,” in *2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP)*, Nov. 2022, pp. 60–66, doi: 10.1109/ICFIRTP56122.2022.10059422.

[29] R. Ahmed, M. Bibi, and S. Syed, “Improving heart disease prediction accuracy using a hybrid machine learning approach: a comparative study of SVM and KNN algorithms,” *International Journal of Computations, Information and Manufacturing (IJCIM)*, vol. 3, no. 1, pp. 49–54, Jun. 2023, doi: 10.54489/ijcim.v3i1.223.

[30] M. R. Sajid *et al.*, “Nonclinical features in predictive modeling of cardiovascular diseases: a machine learning approach,” *Interdisciplinary Sciences: Computational Life Sciences*, vol. 13, no. 2, pp. 201–211, Jun. 2021, doi: 10.1007/s12539-021-00423-w.

[31] A. Al Ahdal, M. Rakhra, S. Badotra, and T. Fadhaeel, “An integrated machine learning techniques for accurate heart disease prediction,” in *2022 International Mobile and Embedded Technology Conference (MECON)*, Mar. 2022, pp. 594–598, doi: 10.1109/MECON53876.2022.9752342.

[32] S. J. Pasha and E. S. Mohamed, “Advanced hybrid ensemble gain ratio feature selection model using machine learning for enhanced disease risk prediction,” *Informatics in Medicine Unlocked*, vol. 32, p. 101064, 2022, doi: 10.1016/j.imu.2022.101064.

[33] S. Shilaskar and A. Ghatol, “Feature selection for medical diagnosis : evaluation for cardiovascular diseases,” *Expert Systems with Applications*, vol. 40, no. 10, pp. 4146–4153, Aug. 2013, doi: 10.1016/j.eswa.2013.01.032.

[34] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, “Data level preprocessing methods,” in *Learning from Imbalanced Data Sets*, Cham: Springer International Publishing, 2018, pp. 79–121.

[35] Y. B. Wah, N. Ibrahim, H. A. Hamid, S. Abdul-Rahman, and S. Fong, “Feature selection methods: case of filter and wrapper approaches for maximising classification accuracy,” *Pertanika Journal of Science and Technology*, vol. 26, no. 1, pp. 329–340, 2018.

[36] N. Nnamoko, F. Arshad, D. England, J. Vora, and J. Norman, “Evaluation of filter and wrapper methods for feature selection in supervised machine learning,” in *The 15th Annual Postgraduate Symposium on the convergence of Telecommunication, Networking and Broadcasting*, 2014, no. JUNE, pp. 63–67.

BIOGRAPHIES OF AUTHORS

Abdikadir Hussein Elmi is a highly skilled IT and Computer Science professional with a wealth of experience in various data-related roles. Over the past several years, he has successfully transitioned towards data science and machine learning, honing his expertise in these cutting-edge fields. He holds a Master of Science in Data Science and Analytics from the prestigious University Science Malaysia (USM), where he gained a strong foundation in advanced data analytics and machine learning techniques. Currently, Abdikadir serves as a Lecturer at SIMAD University, where he passionately shares his knowledge and expertise with aspiring data scientists. Known for his dynamic and engaging teaching style. His research interests revolve around machine learning, artificial intelligence, natural language processing, and neural networks. For any inquiries, collaborations, or opportunities related to data science, machine learning, or artificial intelligence, He can be contacted at xayeeysi77@gmail.com.

Abdijalil Abdullahi received a B.Sc. degree in information technology and an M.Sc. degree in networking and data communication from SIMAD University, Mogadishu, Somalia, in 2014 and 2018, respectively, where he is currently pursuing the Ph.D. degree with the National Advanced IPv6 Center (NAv6), Universiti Sains Malaysia (USM). His research interests include software-defined networking, inter-domain routing, and the internet ecosystem. He can be contacted at email: cabdijaliil22@gmail.com.

Mohamed Ali Barre is a highly experienced system administrator with seven-plus years of expertise in managing advanced network systems and teachings of computer science courses. He has strong skills in troubleshooting technical issues related to hardware, network infrastructure, operating systems, CCTV systems, and software installations. He received a Master of Science in Networking and Data Communication and a Bachelor of Science in Information Technology from SIMAD University. His research interests include network security, internet of things, mechine learning related to computer networks. He can be contacted at email: eng.barre1@gmail.com.