Towards Predicting the Risk of Cardiovascular Disease Using Machine Learning Approach

Hanna Teshager Mekonnen[1(✉)] and Michael Melese Woldeyohannis[2]

[1] University of Gondar, Gondar, Ethiopia
[2] Addis Ababa University, Addis Ababa, Ethiopia
michael.melese@aau.edu.et

Abstract. Cardiovascular disease (CVDs) is one of the leading causes of mortality in the world taking around 18 million lives every year. As a result of the silent progression of CVDs, the rate of mortality is increasing at a higher rate than communicable, maternal, and neonatal diseases in a country like Ethiopia. The early stage detection and treatment, in turn, reduces the rate of mortality as well as the health care cost. For this, a total data set consisting of 10,029 unlabeled instances were analyzed from the Ethiopian Public Health Institution (EPHI). The data were collected by NCD STEPS survey. The population's demographic and behavioral characteristics and also each participant's physical and medical measurement data included in this dataset. Thus the given dataset doesn't have a target variable. Therefore, in order to identify the hidden patterns from unsupervised learning, we use the k-means clustering algorithm and specify the number of clusters to $k = 3$ and cluster the patient condition into high risk, medium risk, and low risk. The data is further experimented with five different machine learning (ML) algorithm to build a predictive model for the risk of CVDs. The result obtained from the experiment using an artificial neural network (ANN) shows a promising result which is 99.4% accuracy. This result shows it's possible to build an effective and efficient model for predicting the risk of having cardiovascular disease.

Keywords: Cardiovascular disease · Risk prediction · Machine learning techniques

1 Introduction

According to World Health Organization (WHO), health refers to a state of complete emotional and physical well-being. It's one of the most important aspects of thing in our day-to-day activities [1]. A health technology is the application of organized knowledge's in the form of devices, medicines, vaccines, procedures and systems developed to solve a health problem and improve quality of lives [2]. Thus, in the healthcare sector a lot of data are being generated every day. However, as data analytics come into existence, the hospitals and non-governmental

M. L. Berihun (Ed.): ICAST 2021, LNICST 411, pp. 489–499, 2022.
https://doi.org/10.1007/978-3-030-93709-6_33

organizations (NGOs) are making use of this data to generate useful information from the available resource [3]. Diseases might classify as communicable and non-communicable diseases. Non-communicable diseases (NCD's) are the leading cause of death globally, and one of the major health challenges of the 21st century [4]. According to WHO report each year, 15 million people die from NCD between the ages of 30 and 69 years; over 85% of these "premature" deaths occur in low- and middle-income countries [5]. In Ethiopia, NCDs cause 42% of deaths, of which 27% are premature deaths before 70 years of age. With no action, Ethiopia will be the first among the most populous nations in Africa to experience dramatic burden of premature deaths and disability from NCDs by 2040 [6].

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. CVD are the number one cause of death globally [1]. However, most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet, physical inactivity and harmful use of alcohol [7]. The major modifiable risk factors are responsible for about 80% of coronary heart disease and cerebrovascular disease in emerging public health (PH) problems [8]. For instance, the major comorbidity of diabetes is CVD, which is estimated to affect about one-third (32.2%) of all people with diabetes. Besides cardiovascular complications also contribute substantially to the costs for managing diabetes [9]. CVDs are responsible for the majority of morbidity and mortality among NCDs in Ethiopia. Despite the rise of NCDs, the country's full attention was in combating the communicable diseases such as HIV, tuberculosis, and malaria [10]. According to a systemic review, CVD (24%), cancer (10 percent), diabetes (5%), and chronic obstructive pulmonary disease (3%) were found to be important causes of death in different parts of Ethiopia and managing multiple risk factors that are associated with CVD is difficult but could prevent numerous deaths [11]. The practice of medicine is changing with the development of new AI methods of machine learning [12]. Early identification of persons with higher risk of CVD is useful for timely strategies on preventing cardiac incidents that lead to death or disabilities.

Thus, there is a need to design and develop a machine learning model that detect and classify the CVD risk at earliest stage. Therefore, the main aim of this study is, to investigate the possibility to design and develop efficient predictive model that classifies the risk of CVD using machine learning techniques.

2 Related Works

Risk prediction on cardiovascular disease has been conducted over the past two decades with validated predictive models. Numerous multivariable risk scores have been developed to estimate a patient's risk of CVD based on certain key known risk factors using traditional approach [13]. For machine learning based researches, researcher [14] compared the Cox PH model that uses all variables, neural networks, AdaBoost, gradient boosting, and Auto-Prognosis; all achieved a significantly higher AUC-ROC compared to all other standard ML models which is 95%. These study conducted in the absence of cholesterol and other blood-based biomarkers which is used as a predictor. In addition, a prospective

cohort study made using 378,256 patients on a routine clinical data in UK [15]. The study compares the established risk prediction using machine-learning technique and found a better prediction of the absolute number of cardiovascular disease cases correctly with neural network algorithm.

Besides this, heart disease diagnosis by utilizing Echocardiography (ECG) report attributes using hybrid data mining technology [16] conducted on a total of 6,987 patient's records using J48 decision tree which provides an accuracy of 96.72% in the absence of socio-demographic attributes on their dataset. Furthermore, institutional based cross sectional study conducted on a total of 416 study participants which are greater than or equal to 18 years from February to March 2017 [17]. These study reveals that Hypertension, dyslipidemia, and physical inactivity were common CVD risk factors in individuals with Type 1 and 2 diabetes malaises (DM). However, the study only refers to associations without inferring the causality.

In recognition of CVD risk prediction, different mechanisms were applied by many researchers for preventing and early detect the risk of having CVD through different algorithms. From the experimental result of these papers, the machine learning approaches resulted in good performance rather than the traditional or statistical approaches. Therefore, in this study, a wide ranging experiment is done to fill some gaps identified while reviewing the related works such as; integrating socio-demographic records, clinical measurement, associations and causal information variables of the patients to detect the risk of CVD's.

3 Methods

In this study, we employed experimental research to design and develop a model that detect and classify the risk of CVD after a set of step by step procedure from data collection to model development. The following section discuss the detail of data preparation, pre-processing, data engineering and feature engineering in attempt to develop the model.

3.1 Dataset

The dataset collected takes the 9 regional and 2 administrative cites in Ethiopia. The source is NCDs STEPS Survey made by Ethiopian public health institution (EPHI). The collected dataset is consisting of 437 attributes and 10,029 instances saved in Statistical Package for the Social Sciences (SPSS) software of unlabeled class. Each attribute is a potential risk factor containing demographic, behavioral and medical risk factors. The demographic attribute comprises of age, sex, educational background including address information. Similarly, behavioral attributes such as current smoking status, alcohol drinking status, and nutritional intakes of the patients. Beside the demographic and behavioral attributes, the survey data has clinical measurement histories like diabetes, high blood pressure, fastbloodglucose, total cholesterol results etc.

3.2 Pre-processing

In machine learning, data preprocessing could be a significant step that makes a difference to enhance and advance the extraction of significant insights from the data [18]. It alludes to the procedure of cleaning and organizing the raw data to make it fitting for building a model. The dataset collected is pre-processed using different techniques such as identifying missing value, encoding categorical data, feature scaling, removing inconsistent and duplicated data. Among the given attributes, a number of duplicated attributes such as town, woreda and region has been removed from the total dataset. Consequently, to make the data appropriate for the machine learning process, we use alternate ways to remove missing data by re-placing with the most frequented data value for categorical variables and attributes with missing value above 60% has been removed. Similarly, for the missing value of continuous data, mean imputation method is used. Then we perform data transformation, normalization and scale categorical variables in to machine readable format. In the feature engineering phase, data transformation activities such as smoothing, feature construction, normalization and aggregation of the data were done to handle categorical variables using label encoder.

3.3 Feature Selection

Feature selection used to select the most relevant attributes and enable learning algorithms to operate faster and more effectively by reducing overfitting problem [19]. Once the redundant attributes are excluded during data cleaning tasks, again the remaining attributes were given to domain experts (internal medicine doctors) to identify relevant attributes which are highly recommended risk factors used for identifying if the patient is having cardiovascular disease risk warning signs. While we reduce the original dataset besides the domain expert's recommendation we consider the Framingham risk factor predictor features and American heart association (AHA/ACC) risk calculator. In addition to these 20 features were selected as important features by ExtraTreesClassifier feature selection algorithm. Figure 1 presents the selected 20 most important attributes.

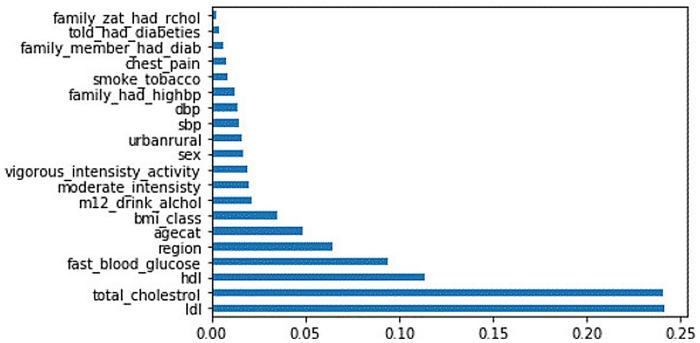
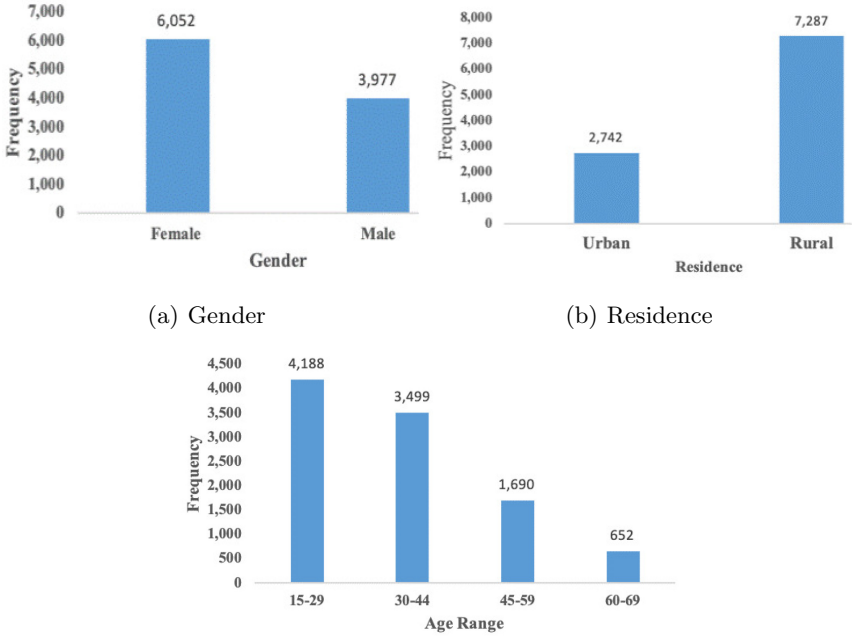


Fig. 1. Top 20 most important variables

However, for building best performing model five more relevant attributes (triglyceride, raised blood pressure, raised blood pressure medicated, raised glucose and raised glucose medicated) are added to the existing features. This relevant attributes are highly recommended risk factors for CVD risk prediction. Accordingly, out of the original dataset presented, we used 25 features and 10,029 instances.

3.4 Data Exploration

Exploratory data analysis is concerned with understanding the nature of data that we have to work with, try to find correlations between each variables, the distribution of the data [20]. To achieve this, mainly we use descriptive statistics, visual techniques, and modeling. These insights help in selecting the right ML algorithm based on our data. In this study for the data exploration and comparison with different variables we apply variety of techniques and we use Matplotlib and Seaborn library. Figure 2(a), 2(b) and 2(c) presents the detail distribution of patient age, gender and residence against the frequency in Fig. 2.



(a) Gender

(b) Residence

(c) Age

Fig. 2. Distribution of age, gender and residence

As depicted in Figure 2 out of 10,029 patients, 6,052 (60.35%) were female and 3,977 (39.65%) were male's. The majority of the patients were from rural

areas of the country which accounts 7,287. In addition to this, from the total patients most of them are at the youngest age category which is from 15–29.

4 System Architecture

The system architecture includes various procedures such as labeling the unlabeled data, pre-processing, feature engineering and building a risk prediction model. Figure 3 presents the architecture of CVD risk prediction using different ML classification algorithms.

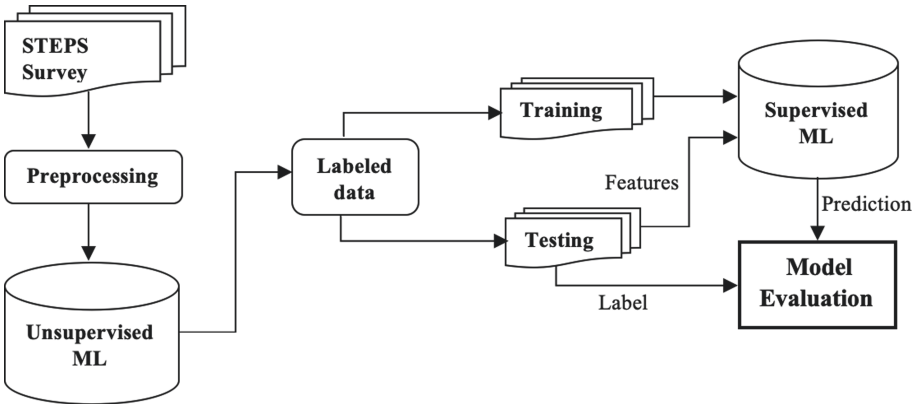


Fig. 3. Architecture of the system

As shown in Fig. 3 the flow of the architecture goes first by processing the unlabeled survey data and apply data preprocessing tasks using different ML data cleaning, transformation, and feature selection techniques and prepare the data for model building stage. After pre-processing, since our dataset doesn't have given target variable in order to find the hidden patterns, from unsupervised learning techniques k-means clustering algorithm were used. Afterwards, the labeled dataset is split for model building using five ML classification algorithms and then the performance of the model is evaluated with a standard evaluation metric's using accuracy.

5 Experiment

The experimental step includes both supervised and unsupervised learning techniques through the process ML model is selected, trained, and validated. From unsupervised learning we use K-Means clustering algorithm to partition and cluster our data based on attributes or features in to a K number of clusters [21]. Which is in our case three as high risk, medium (moderate) risk and low

risk. The three class labels are selected based on previous CVD risk estimator studies and we also evaluate the clustering algorithm using silhouette score which is used to evaluate how well samples are clustered with other samples that are similar to each other. Giving this we get 0.45 silhouette score therefore, k=3 is the optimal numbers of the cluster. Figure 4 presents the dataset after applying k-means clustering.

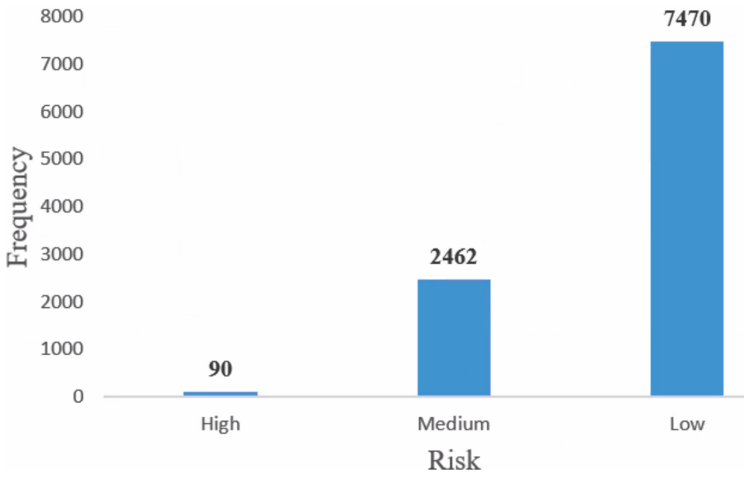


Fig. 4. Class distribution of target variable

After applying K-means clustering we get the clustered variables which is our target variable called risk which is classified as high, medium and low. Furthermore, the result of a cluster values has been evaluated with the help of domain experts, giving that from the high risk class which has 90 patients only 10 patients were misclassified as having high risk of CVD. Correspondingly for medium risk class which accounts 2469 patients the clustering model misclassified 3 of them as high risk class patients and 2 of them as low risk class patients. And as for the low risk form 7470 patients 23 patients were misclassified as medium risk class and 2 as high risk class. According to this the overall accuracy is 99.5%. As depicted in Fig. 4 the risk class levels are imbalance therefore, for improving the accuracy of the model and get accurate prediction result the Synthetic Minority Over-sampling Technique (SMOTE) [22] were used since these sampling method alleviates overfitting problems and we will not loss any information during sample creation. After making the data suitable for building a model we have done fifteen experiments using five supervised learning algorithms. Since all these five algorithms are well known classification algorithms we compare their performance by giving data obtained from k-means clustering algorithm. In model training we split our dataset using three methods, with single holdout random subsampling method the dataset split into two for training set which is used for model building 8,023 (80%), and the rest 2,006 (20%) is

used to test or validate the models after training is complete. Correspondingly, in 70/30 for training 7,020 and 3,009 for testing additionally we also used Kfold splitting method. From supervised machine learning classification techniques five algorithms namely Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM) and k-nearest neighbor's (KNN) and Artificial Neural Network (ANN) were used. Since our experimental method is multi class classification here for SVM algorithm we set the decision function shape argument to one vs one(ovo). Additionally, as for ANN we use Keras Classifier, the first procedure were to reshape the output attribute of each class by converting the vector of integers to a one hot encoding. Then we pass the parameters by specifying the number of layers and nodes for the classifier as hidden layer sizes = (150,100,50), also the number of epoches were set to 300, and the activation function for the hidden layers were set to Rectified Linear Unit (ReLu) since it induces a sense of non-linearity in the network. Another parameter Adam gradient descent weight optimization were given as a solver. Besides, the random state were also set to the default value which is 42. finally we fit our model to train the algorithm on our training data then make a prediction on our test data. Table 1 presents the results obtained from the experiments.

Table 1. Classification results

Algorithm	Split			Evaluation		
	80/20	70/30	k-fold	f-score	Precision	Recall
Decision Tree	88.3	86.4	85.5	92.0	88.0	89.0
Naive Bayes	83.0	84.0	68.8	90.0	83.0	86.0
SVM	98.0	97.5	98.2	98.0	98.0	98.0
KNN	98.8	97.8	98.4	98.0	98.0	98.0
ANN	99.4	99.2	99.0	99.0	99.0	99.0

Table 1 presented the classification results of each algorithms performance in terms of accuracy, precision, recall and f-measure. Giving that ANN achieves highest accuracy than the other algorithms and correctly predicts 1994 instances and missed 12 instances. When we see the performance of other four algorithms, DT algorithm correctly predicts 1772 instances and missed 234 instances. The SVM algorithm also correctly predict 1967 instances and missed 39 instances and the KNN algorithm correctly predicts 1966 instances and missed 40 instances. As for NB model it has less performance accuracy than the other algorithms, this algorithm correctly predicts 1659 instances and missed 345 instances.

6 Discussion of the Results

As discussed in the experimental setup we use five algorithms for the prediction model. Giving that the result shows DT, NB, SVM (One Vs One), KNN and

ANN classifier algorithms had performance accuracy of 88.3%, 83.0%, 98.0%, 98.8% and 99.4%. In order to evaluate the performance of all the models, specificity, sensitivity and average classification accuracy of each model was compared. From our experiments the obtained results show a good performance and this shows that it's possible to predict cardiovascular disease risk effectively and get good accuracy using ML algorithms. Figure 5 presents the experimental result of the five machine learning algorithm using three splitting techniques.

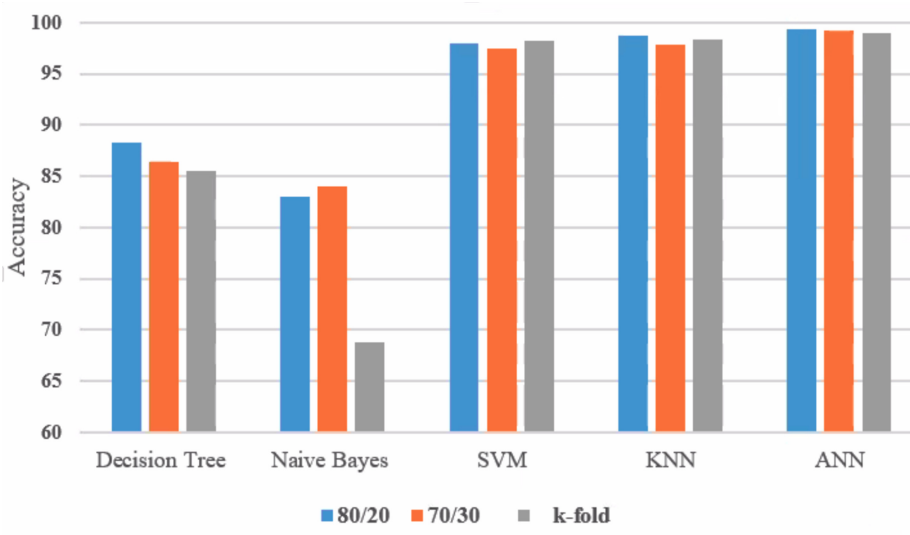


Fig. 5. Comparison of classifier algorithms

Based on the above comparison from the five algorithms used in this paper artificial neural network algorithm is the best performing model and Naïve Bayes got the least performance accuracy. Since neural networks able to deal with complex relations between variables, highly tolerance to noisy data, as well as their ability to classify patterns on which they have not been trained, makes the accuracy of the classifier better than other classifiers. In case of Naïve Bayes algorithm it has less performance level, we perform most of feature engineering techniques to handle categorical variables in our dataset yet, the NB algorithm seem to have difficulties on handling those variables since this algorithm assigns a zero probability and unable to make a prediction on those kind of variable data types. Therefore, this shows that this is the reason why we have less performance accuracy since most of our data include associations and causal information of patients. Additionally, as presented in Fig. 5 from the three splitting techniques the 80/20 splitting method get the highest score.

7 Conclusion and Recommendations

In this study, a wide ranging experimentation is done to fill some gaps identified while reviewing the related works and create better predictive model by integrating different socio-demographic, behavioral and clinical measurement variables. Besides, we also comprise different well known classification techniques to develop best performed classification model for predicting the risk of cardiovascular disease. As the result shown from all five algorithms used in this study ANN algorithm gets the highest accuracy score followed by KNN and SVM (One Vs One). Therefore, based on this results ANN is chosen as the best model at predicting the risk of cardiovascular disease. For future researchers in addition to this variable used in this study if echocardiography report variables have been added to this investigation we can extract more significant and useful information in the classification of each patient's risk. Similarly, additional deep learning algorithms with much larger data size needs to be taken to optimize the prediction of cardiovascular disease risk classification.

References

1. Organization Prevention of Cardiovascular Disease Guidelines for assessment cardiovascular risk (2020). https://www.who.int/about/who-we-are/constitution
2. Organization health-technology-assessment (2020). https://www.who.int/health-technology-assessment/about/healthtechnology/en
3. Dinesh, K., Arumugaraj, K., Santhosh, K., Mareeswari, V.: Prediction of cardiovascular disease using machine learning algorithms. In: 2018 International Conference on Current Trends Towards Converging Technologies (ICCTCT), pp. 1–7 (2018)
4. Organization, W., et al.: Noncommunicable diseases country profiles 2018. World Health Organization (2018)
5. Organization noncommunicable-diseases (2020). https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases
6. Shiferaw, F., et al.: Non-communicable Diseases in Ethiopia: disease burden, gaps in health care delivery and strategic directions. Ethiopian J. Health Dev. **32** (2018)
7. Organization health topics (2020). https://www.who.int/health-topics/cardiovascular-diseasestab=tab
8. Baye, M.: Prevalence of overweight, obesity among urban civil servants in Southern Ethiopia. Ethiopian Med. J. **57** (2019)
9. Einarson, T., Acs, A., Ludwig, C., Panton, U.: Economic burden of cardiovascular disease in type 2 diabetes: a systematic review. Value Health **21**, 881–890 (2018)
10. Tefera, Y., Abegaz, T., Abebe, T., Mekuria, A.: The changing trend of cardiovascular disease and its clinical characteristics in Ethiopia: hospital-based observational study. Vasc. Health Risk Manag. **13**, 143 (2017)
11. Sung, J., et al.: Development and verification of prediction models for preventing cardiovascular diseases. PLoS One. **14**, e0222809 (2019)
12. Ahuja, A.: The impact of artificial intelligence in medicine on the future role of the physician. PeerJ **7**, e7702 (2019)
13. Tsao, C., Vasan, R.: The Framingham Heart Study: Past. Present and Future. Oxford University Press, Oxford (2015)

14. Alaa, A., Bolton, T., Di Angelantonio, E., Rudd, J., Schaar, M.: Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants. PloS One **14**, e0213653 (2019)
15. Weng, S., Reps, J., Kai, J., Garibaldi, J., Qureshi, N.: Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One **12**, e0174944 (2017)
16. Abrha, T.: School of Graduate Studies Faculty of Informatics. Addis Abeba University (2012)
17. Abdosh, T., Weldegebreal, F., Teklemariam, Z., Mitiku, H.: Cardiovascular diseases risk factors among adult diabetic patients in eastern Ethiopia. JRSM Cardiovasc. Dis. **8**, 2048004019874989 (2019)
18. Goyal, K.: Undefined. https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/
19. Shrivastava, H., Sridharan, S.: Conception of data preprocessing and partitioning procedure for machine learning algorithm. Int. J. Recent Adv. Eng. Technol. (IJRAET) **1**, 2347–2812 (2013)
20. Education, I.: Undefined (2020). https://www.ibm.com/cloud/learn/exploratory-data-analysis
21. Greene, D., Cunningham, P., Mayer, R.: Unsupervised learning and clustering. In: Machine Learning Techniques For Multimedia, pp. 51–90 (2008)
22. Rendon, E., Alejo, R., Castorena, C., Isidro-Ortega, F., Granda-Gutierrez, E.: Data sampling methods to deal with the big data multi-class imbalance problem. Appl. Sci. **10**, 1276 (2020)