EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

An All-Inclusive Machine Learning and Deep Learning Method for Forecasting Cardiovascular Disease in Bangladeshi Population

Manjula Mandava[1], Dr. Surendra Reddy Vinta[2,*], Hritwik Ghosh[3], Irfan Sadiq Rahat[4]

[1,2,3,4] School of Computer Science and Engineering (SCOPE), VIT-AP University, Amaravati, Andhra Pradesh

Abstract

INTRODUCTION: Cardiovascular disease is a major concern and pressing issue faced by the healthcare sector globally. According to a survey conducted by the WHO every year, CVDs cause 17.9 million deaths worldwide. Lack of pre-prediction of CVDs is a significant factor contributing to the death of patients. Predicting CVDs is a challenging task for medical practitioners as it requires a high level of medical analysis skills and extensive knowledge.
OBJECTIVES: We believe that the improvement in the accuracy of prediction can significantly reduce the risk caused by CVDs and help medical practitioners better diagnose patients .
METHODS: In this study, We created a CVD prediction model. using a ML approach. We utilized various algorithms, including logistic regression, Gaussian Naive Baye, Bernoulli Naive Baye, SVM, KNN, optimized KNN, X Gradient Boosting, and random forest algorithms to analyze and predict CVDs.
RESULTS: Our developed prediction model achieved an accuracy of 96.7%, indicating its effectiveness in predicting CVDs. DL algorithms can also assist in identifying, classifying, and quantifying patterns of medical images, improving patient evaluation and diagnosis based on prior medical history and evaluation patterns.
CONCLUSION: Furthermore, deep learning algorithms can help in developing new drugs with minimum cost by reducing the number of clinical research trials, using prior prediction of the drug's efficacy.

Keywords: Logistic regression, Gaussian Naive Bayes, B-Naive Bayes, SVM, X Gradient Boosting, Decision Tree Classifier, Grid Search CV, Ada Boost Classifier, G-Boosting Classifier, Cat Boost Classifier, Extra Trees Classifier, KNN, MLP Classifier, Stochastic gradient descent, Artificial Neural Network

Received on 30 June 2023, accepted on 09 September 2023, published on 03 October 2023

doi: 10.4108/eetpht.9.4052

1. Introduction

Fastly evolving day to day upgrading world has significantly made humans as the working machines. The busy stressful routine of day-to-day life has eventually developed the unhealthy habits in this generation. Due to this sort of unhealthy routine many people are neglecting their health conditions. Due to this kind of unhygienic and stressful lifestyle, a rise in the number of persons getting sicker and
sicker every day.

The prevalence of CVDs in Bangladesh is increasing rapidly, with an estimated 1.6 million deaths per year due to these diseases. A number of factors contribute to the rising prevalence of CVDs in Bangladesh unfavorable lifestyle decisions, insufficient exercise, and inadequate nutritional habits **[1]**. Furthermore, various risk factors impact the occurrence of CVDs, including high cholesterol levels, diabetes, high blood pressure, having a family history of heart disease, and being overweight. In Bangladesh, the prevalence of hypertension is around 20%, while the prevalence of diabetes is estimated to be around 9% **[2]**. Therefore, there is a need for comprehensive ML and DL perspectives for predicting and preventing CVDs in the Bangladeshi population. Such approaches can help in identifying high-risk individuals and providing personalized interventions, leading to better health outcomes and reduced healthcare costs.
Globally, CVDs continue to be a significant public health concern, being among the most important convictions of death. 17.9 million people died, according to estimations, CVDs will account for 32% of global fatalities in 2019, having heart attacks and strokes as the primary causes of 85% of these fatalities. The majority of CVD deaths, more than three-quarters, occur in low- and middle-income countries. CVDs are also the cause of 38% of

noncommunicable disease-related premature deaths (those less than the age of 70) in 2019[3]. Bangladesh and other low- and middle-income countries are particularly affected, with 80% of CVD-related deaths occurring in these regions. This makes CVD a significant health concern in Bangladesh. Furthermore, from 2011 to 2025, the cumulative economic losses from all non-communicable illnesses in these nations are expected to reach $7.28 trillion, with CVD responsible for nearly half of this loss **[3].** The period between March 2020 and August 2022 witnessed over 30,000 additional deaths attributed to coronary heart disease, averaging at 230 deaths per week above the expected mortality rate.

Noncommunicable chronic disease prevalence and associated mortality has been increasing in Bangladesh due to rapid urbanization and changes in lifestyle, including the habituation of a sedentary lifestyle and changes in food habits. Unfortunately, the lack of a population-based surveillance system and In Bangladesh, precise information on illness prevalence makes challenging for health professionals and policymakers to comprehend the scope of the problem. To fill this need, a systematic review and meta-analysis of known research on the frequency of CVD in Bangladeshi adults were performed. The discoveries of this study can provide significant information to health professionals and policymakers in Bangladesh for the research, development and implementation of CVD prevention measures. A Study published by journal of the American College of cardiology states that CVDsv are the leading cause of global mortality in this paper they have analyzed at 13 causes of cardiovascular death in a related to risk factor they hats of the study has proven the study rays in the number of CVD cases per year I cross the globe we have seen an irrelevant increment in the cases of the CVD heart is a major and the most significant part pertaining to a human body is not feasible without the appropriate functioning of the heart. The heart performs many functions, the most important of which is to pump oxygenated blood to every part of the body and to circulate blood throughout the body. If the heart fails to function properly, every other organ in the body will fail to function properly as well, so it's critical to take good care of our hearts and to monitor their conditions over time.

The structure of this essay is as follows: In Section 2, the research techniques used to select the primary studies are covered. Section 3 addresses the proposed methodology. Section 4 looks at Experimental analysis. In Section 5&6, the results and conclusions.

2. Literature Review

Cardiovascular disease (CVD) represents a substantial health challenge worldwide, and early detection is essential for successful treatment and management. In recent times, Researchers have used ML and DL techniques to create predictive models for heart disease. This literature review aims to deliver an all-inclusive overview of current methodologies in the field and examine their applicability to the Bangladeshi population.
Mohan et al. (2019) **[1]** proposed a hybrid ML strategy that combined several techniques to effectively predict heart disease. Their work demonstrated the potential for improved accuracy using such hybrid approaches.

Similarly, Rajdhan et al. (2020) **[11]** explored the use of ML for heart disease prediction, providing valuable insights into the effectiveness of various algorithms in this context.

Ramalingam et al. (2018) **[12]** used ML algorithms to conduct a poll on heart disease forecasting, offering a detailed overview of the field and identifying promising avenues for future research. Shah et
al. (2020) **[4]** and Singh et al. (2020) **[5]** also investigated heart disease prediction using ML algorithms,
highlighting the potential for these approaches to significantly impact healthcare outcomes.

Patel et al. (2015) **[6]** focused on this application of ML and data mining methods for heart disease prediction, emphasizing the importance of interdisciplinary approaches in addressing complex health challenges. Khourdifi et al. (2019) **[7]** employed Optimization of particle swarms and Ant colony optimisation to boost the efficiency of ML algorithms in heart disease prediction and classification Tasks.

Jagtap et al. (2019) **[8]** applied supervised ML algorithms to forecast cardiac illness, demonstrating the
potential of these techniques in improving the accuracy and efficiency of diagnosis. Ali et al. (2021) **[9]** conducted a performance analysis and comparison of forecast cardiac illness using supervised ML algorithms, highlighting the strengths and weaknesses of various methods. Jindal et al. (2021) **[10]** further explored the use of ML methods for coronary artery disease forecasting, emphasizing the ongoing importance of this research area.

In recent years, researchers have extensively explored the application of ML methods for forecasting cardiac illness. Sharma et al. (2020) investigated various ML techniques and their potential for forecasting cardiac illness, emphasizing the importance of choosing the right algorithm for accurate predictions **[2].** Diwakar et al. (2021) presented the latest trends in heart disease prediction, incorporating image fusion and machine learning techniques to improve diagnostic accuracy **[3].**

Kavitha et al. (2021) proposed a hybrid ML model for forecasting cardiac illness, highlighting the benefits of combining multiple algorithms for improved performance **[14]**. Gavhane et al. (2018) examined the potential of ML algorithms for forecasting cardiac illness, emphasizing the need for accurate and timely diagnosis **[15]**. Katarya and Meena (2021) conducted a comparative study and analysis of various ML techniques for forecasting cardiac illness, assessing their strengths and weaknesses **[16].**

Motarwar et al. (2020) explored a cognitive approach to forecast cardiac illness using ML, underscoring the potential of integrating advanced computational methods in healthcare **[17].** Yahaya et al. (2020) provided a thorough examination of data mining and ML methods for forecasting cardiac illness, emphasizing the importance of interdisciplinary research in this field **[18].** Marimuthu et al. (2018) reviewed heart disease forecasting using ML and data analytic s, highlighting the potential of advanced analytics in medical diagnostics **[19].**

Barik et al. (2020) investigated cardiac illness forecast using ML methods, focusing on the role of computational methods in improving patient outcomes **[20].** Lastly, Haq et al. (2018) proposed a hybrid intelligent system framework for heart disease prediction, incorporating multiple ML algorithms for enhanced diagnostic accuracy **[21].** In summary, the literature reveals the

growing interest and importance of ML methods to predict cardiac disease. With ongoing research aimed at improving the accuracy and reliability of these methods.

3. Proposed Methodology

We have undertaken various pre-processing steps to prepare the data for use in ML and DL models, including logistic regression, G-Naive Bayes, B-Naive Bayes, SVM, X Gradient Boosting, Decision Tree Classifier, Grid Search CV, Random-Forest algorithm AdaBoost, Gradient Boosting, Cat Boost Classifier, Extra Trees, KNN, MLP, Stochastic gradient descent, and Artificial Neural Network. To ease model training as well as evaluation, the dataset has been divided into training and testing sets. In the data preparation stage, we removed inaccurate values and filled in missing entries with the median of the respective input variables to maintain data integrity. **[Fig.1]** below displays the overall framework of our suggested system and the methodology employed.

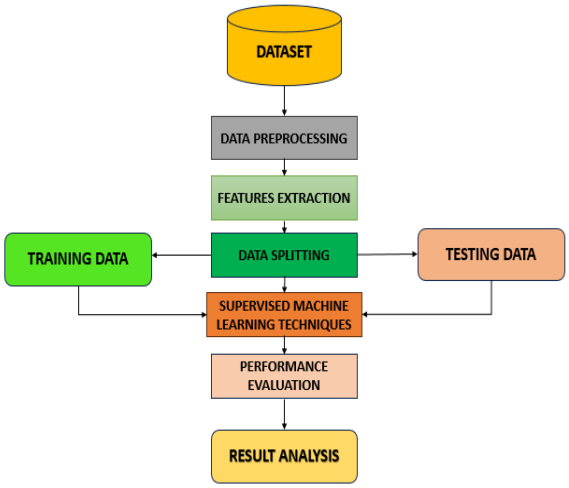

Fig. 1 Methodology of the proposed system

3.1. Description of the Dataset

The dataset contains 303 observations and 14 variables, with 5 of them being numeric and 9 being categorical. The dataset does not have any missing cells, with a missing cell percentage of 0.0%. However, there is 1 duplicate row in the dataset, which accounts for 0.3% of the dataset. The preprocessing of the dataset involved checking for duplicates and missing cells, which were found to be present in small amounts. The duplicate row was removed, and the dataset was cleaned of any missing cells. This dataset provides information about patients with regard to various factors that may contribute to the Odds of experiencing a heart attack. The data includes the following features for each patient: age, sex, whether or not they have exercise-induced angina, the number of main blood arteries present, a resting ECG can reveal the type of chest discomfort, resting blood pressure, cholesterol levels, and fasting blood sugar levels. ECG, maximum heart rate achieved, and the likelihood of a heart attack (0 indicating a lower chance and 1 indicating a higher chance). The data can be used to analyze the relationship between these various factors and the likelihood of a heart attack, as well as to build predictive models to determine patients who are at a larger likelihood of experiencing a heart attack. The dataset can be useful for researchers, healthcare professionals, and policymakers interested in studying heart disease and developing interventions to prevent or manage heart attacks.

3.2 Preprocessing of the Dataset

Preprocessing is an important step in ML, which involves preparing the data for analysis by transforming it into a format suitable for the algorithm. In the case of the CVD dataset, the following preprocessing steps can be performed:

- **Data Cleaning**

Data cleaning is a crucial step in preparing the CVD dataset for analysis. During this step, the data is scrutinized for missing or erroneous data, inconsistencies, and outliers. Initially, we need to check for missing data, which can be represented by question marks (?) in the CVD dataset. Replacing these question marks with NaN values is recommended for easier handling of the dataset during analysis. Once done, the next step is to determine how to handle the missing data, which can be done by either imputing the missing values or removing the corresponding rows. Identifying and handling outliers is also important since they can significantly affect the results of the analysis. Instead of removing all outliers, we must examine each outlier to identify whether it is a genuine data point or a result of an error or data entry mistake. Additionally, we need to check for inconsistencies in the dataset by examining the range of values for each feature. For instance, age should be a positive value, blood pressure should be within a certain range, and so on. If any inconsistencies are detected, we can either remove the corresponding data points or try to correct them if possible, ensuring that the dataset is clean and reliable for analysis.

- **Feature Scaling**

It is an essential step in preprocessing of data to confirm that the features in the dataset are comparable and have a similar impact on the analysis. In the CVD dataset, some features have larger value ranges than others, which can lead to biased analysis. Feature scaling can help to overcome this issue by scaling all features to a similar range. There are two common techniques for feature scaling: normalization and standardization. The process of normalization entails scaling the values within each feature to a range of 0 to 1. To accomplish this, we subtract the minimum value of each feature from every value and then divide by the range of the feature. This methodology is appropriate in situations where the feature distribution is not known or when the feature values do not adhere to a Gaussian distribution. In contrast, standardization is a technique it changes the attributes to a mean of zero and a standard deviation of 1.This is executed by subtracting the mean of every feature from the corresponding values and dividing the outcomes by the feature's standard deviation. Standardization is more suitable when the feature

values follow a normal distribution. It's important to note that the choice of feature scaling technique should be based on the distribution of the feature values. Additionally, some ML algorithms are sensitive to the scale of the features, so feature scaling can improve the accuracy of these algorithms. Therefore, it's recommended to perform feature scaling as a part of the data preprocessing step before applying any ML models on the CVD dataset.

- **Feature Encoding**

Feature encoding is the process of converting categorical variables into numerical values that can be used for analysis. In the CVD dataset, there are several categorical variables, such as sex, chest pain type, and fasting blood sugar. One common approach for encoding categorical variables is one-hot encoding, which creates a new binary feature for each unique category in the variable. For example, in the sex variable, we can create two new features: one for male and one for female. A value of 1 in the male feature indicates that the patient is male, and a value of 0 indicates that the patient is not male. Similarly, a value of 1 in the female feature indicates that the patient is female, and a value of 0 indicates that the patient is not female. Another approach is label encoding, which assigns a numerical value to each unique category in the variable. For example, in the chest pain type variable, we can assign the value 1 to typical angina, 2 to atypical angina, 3 to non-anginal pain, and 4 to asymptomatic. However, it's important to note that label encoding may not be suitable for variables with no inherent order or hierarchy, as it may create an artificial hierarchy that does not exist.

- **Feature Selection**

Feature selection is a critical step in preparing a dataset for analysis. It involves identifying the most relevant features that can be used to build a predictive model. The goal of feature selection is to reduce the amount of features present in the dataset, while preserving the information content of the data. In the CVD dataset, we have 14 variables that describe various attributes of patients that may be related to the existence or lack of cardiovascular disease. However, not all of these variables may be relevant for building a predictive model. Some variables may be highly correlated with other variables, while others may not have a significant impact on the outcome. One approach to feature selection is to use statistical tests to identify the most important variables. For example, we can use the chi-squared test to determine the association between each categorical variable and the outcome variable (presence or absence of CVD). Similarly, we can use the t-test or ANOVA to determine the association between each continuous variable and the outcome variable. Another approach is to use ML algorithms that can automatically select the most relevant features. For example, we can use decision trees, random forests, or SVMs to select the most important variables. Once we have identified the most relevant features, we can use them to build a predictive model. This can be done using various ML algorithms, such as decision trees, random forests, SVMs, and NN.

- **Data Splitting**

Splitting data is a crucial step in machine learning, enabling us to assess model performance on unseen data. To partition the CVD dataset, we can apply a widely used technique known as random splitting. Initially, the dataset is partitioned into two sections - one for model training and the other for testing. The 70-30 split is the most common partition, where 70% of the data is designated for training and the remaining 30% for testing. However, the exact proportion can vary, taking into account dataset size and complexity. To ensure that our split is representative of the entire dataset, we can use stratified sampling. This involves dividing the dataset into strata based on the target variable (in this case, CVD or non-CVD) and then randomly selecting samples from each stratum for the training and testing sets. This ensures that the distribution of the target variable is preserved in both the training and testing sets.

3.3 Data Analysis

This data shows the frequency count of each chest pain type observed in a cardiovascular disease (CVD) dataset. The chest pain types are categorized into four types: Common Angina, Unconventional Angina, Non-cardiac Chest Pain, and Absence of Symptoms. Among the four types, Typical Angina has the highest frequency count of 150, followed by Atypical Angina with 90 counts. Non-anginal Pain and Asymptomatic have a lower frequency count of 50 and 30, respectively. It is important to note that the distribution of chest pain types may have implications for the diagnosis and treatment of patients with chest pain. For example, certain types of chest pain may indicate a higher risk of cardiovascular disease or require different diagnostic tests or treatments**. [Fig.2]**. The given data shows the number of heart attacks based on the type of pain in the chest experienced by patients. The highest number of heart attacks (68) occurred in patients with Typical Angina (0), followed by Atypical Angina (1) with 41 heart attacks. Non-anginal Pain (2) had 39 heart attacks, and Asymptomatic (3) had the lowest number of heart attacks at 18. This suggests that patients with Typical Angina are at a higher risk of experiencing heart attacks compared to those with other types of chest pain. We can see that there is a higher percentage of male patients (56.10%) who have heart disease compared to female patients (43.90%) **[Fig.3].** Additionally, among male patients, 44.66% suffer from heart illness and 55.34% do not suffer from heart illness, while among female patients, a higher percentage (75.00%) have heart disease and a lower percentage (25.00%) do not have heart disease. These percentages suggest that gender could be a factor in the development of heart disease, and that women may have a greater chance of acquiring cardiovascular disease than men. However, it's important to note that these percentages are based on the specific dataset being analyzed and may not be representative ofthe overall population**. [Fig.4].**

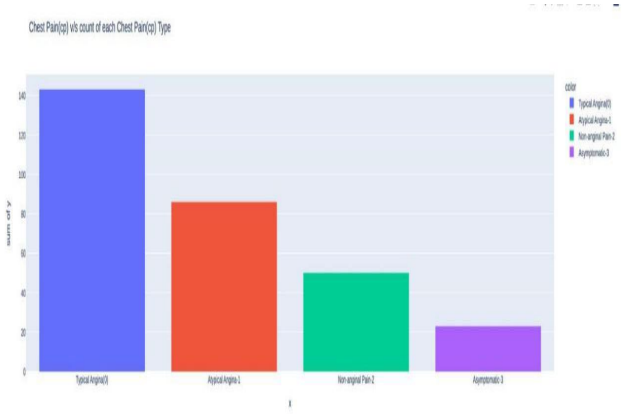

Fig. 2 Chest Pain(cp) v/s count of each Chest Pain(cp)

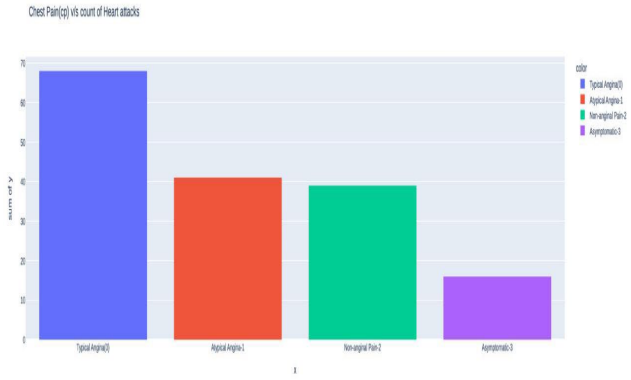

Fig. 3 Chest Pain(cp) v/s count of Heart attacks

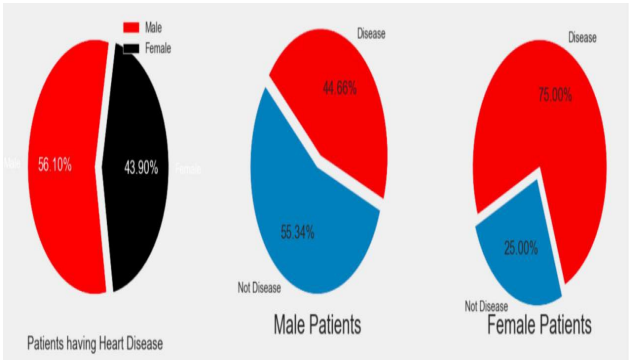

Fig. 4 Gender-specific diseases

4. Experimental Analysis

Now we are going to see the performances of our used ML and DL models (Logistic Regression, G Naive Bayes, B-Naive Bayes, SVM, Xg Boosting, Decision Trees Classifier, Grid Search CV, AdaBoost, Gradient Boosting, XgBoost,,Cat Boost ,Extra Trees, KNN,MLP ,Stochastic Gradient Descent, Artificial Neural Network.)through confusion matrix, Testing accuracy, Training accuracy, Precision, Recall, F-1 score and Support. The table shows the testing accuracy of different classifiers on a dataset. The classifiers are evaluated based on how well they can predict the target variable. The Bernoulli N-Bayes algorithm has the highest testing accuracy of 0.99, while KNN has the lowest testing accuracy of 0.77. Other classifiers have testing accuracy ranging between 0.82 to 0.92. The classifiers with higher testing accuracy are better at predicting the target variable and are more suitable for the given dataset**. [Table.1]**

Table I. Comparison of Chronic Kidney Disease Segmentation And Classification Methods Based On The Accuracy, Precision, Recall, F1-Score and Support

Classifiers	**Testing accuracy%**	**Training accuracy%**	**Precision%**	**Recall %**	**F1-score %**	**Support**
Logistic Regression	92	84	100	83	91	30
Gaussian Naive Bayes	89	82	87	90	89	30
Bernoulli Naive Bayes	**99**	**81**	**100**	**97**	**98**	**30**
SVM	87	90	92	80	86	30
XG-Boosting	89	100	90	87	88	30
Decision Tree Classifier	82	100	77	90	83	30
Grid Search CV	97	99	96	99	97	30
Random Forest Classifier	89	98	87	90	89	30
Ada Boost Classifier	84	100	79	90	84	30
Gradient Boosting	86	100	89	80	84	30

Classifier						
Cat Boost Classifier	91	91	93	87	90	30
Extra Trees Classifier	92	100	93	90	92	30
KNN	77	78	83	67	74	30
MLP Classifier	89	93	93	83	88	30
Stochastic gradient descent	89	77	87	90	89	30
Artificial Neural Network	92	93	96	87	91	30

4.1. Confusion Matrix

For a binary classification model, a confusion matrix for CVD would show the number of TP, FP, TN, and FN. The confusion matrix would resemble this:

In the context of cardiovascular disease (CVD), a positive prediction would indicate that a patient has the disease, while a negative prediction would suggest that they do not have it. A true positive would be when a patient is correctly identified as having CVD, a false positive would be when a patient is predicted to have CVD but actually does not, a true negative would be when a patient is correctly identified as not having CVD, and a false negative would be when a patient is predicted to not have CVD but actually does. The confusion matrix is a valuable tool for assessing the of a binary classification model, such as one for predicting CVD, as it enables us to calculate several key metrics such as accuracy, precision, Rrecall and F1-score.

4.2 Precision and Recall

precision and recall can be used to evaluate the accuracy of a model in correctly identifying patients with the disease. Precision represents the percentage of patients who are predicted to have CVD and actually have the disease, while recall represents the percentage of patients with CVD who are correctly identified by the model.

- **Precision:** The proportion of accurate positive forecasts is known as precision. Out of all the positive predictions made by the model, it provides an estimate of the number of true positive forecasts. Precision is calculated as follows: Precision = TP / (TP + FP).
- **Recall (Sensitivity):** The percentage of instances that were accurately forecasted as positive but turned out to be positive is known as recall. Recall reveals whether a model is able to find all instances of success. Recall is determined by dividing TP by (TP + FN).

4.3 Logistic Regression (LR)

Testing accuracy for this method was 0.92, and training accuracy was 0.84. A popular linear classification approach called LR uses a logistic function to characterise the connection between input data and the output target. It can effectively handle huge datasets and is simple to implement.

$$h\Theta(x) = 1/1 + e - (\beta o + \beta 1X) \quad (1)$$

4.4 Gaussian Naive Bayes

This algorithm has achieved a testing accuracy of 0.89 and a training accuracy of 0.82. It is a simple and effective probabilistic classification algorithm that models the distribution of each class using Gaussian distribution. It is particularly useful when the number of input features is large.

4.5 Bernoulli Naive Bayes

This algorithm has achieved a testing accuracy of 0.99 and a training accuracy of 0.81. It is a variant of Naive Bayes that models the distribution of each class using a Bernoulli distribution. It is commonly used for text classification tasks.

4.6 SVM

This algorithm has achieved a testing accuracy of 0.87 and a training accuracy of 0.90. SVM is a powerful classification algorithm that aims to find a hyperplane that maximally separates the classes in the feature space. It is effective for high-dimensional and non-linear classification problems.

4.7 Kth Nearest Neighbours (KNN)

This algorithm has achieved a testing accuracy of 0.77 and a training accuracy of 0.78. KNN is a simple and effective classification algorithm that classifies a new data point based on the class of its nearest neighbors in the feature space. It is particularly useful for low-dimensional datasets and can handle both categorical and numerical input features.

4.8 X Gradient Boosting

Using boosting, the ensemble classification technique X Gradient Boosting combines numerous weak classifiers into a single strong classifier. It has a testing accuracy of 0.89 and a training accuracy of 1.00. It works well for many different classification issues and is especially helpful when dealing with unbalanced datasets.

4.9 Random Forest Classifier

With this approach, testing accuracy was 0.89 and training accuracy was 0.98. The RF is an ensemble categorization technique that uses bagging to aggregate various decision trees into a single powerful classifier. It can handle both category and numerical input features and is good for handling noisy datasets.

4.10 Decision Tree Classifier

The testing and training accuracy of this algorithm are both 0.82 and 1.00, respectively. The decision tree classifier is a hierarchical classification technique that divides the feature space into areas based on the values of the input features. Both categorical and numerical input features can be handled by it, and it is straightforward and easy to understand.

4.11 Grid Search CV

This algorithm has achieved a testing accuracy of 0.97 and a training accuracy of 0.99. It is a technique for hyper parameter tuning that exhaustively searches the hyper parameter space to find the optimal set of hyper parameters for a given algorithm. It is useful for improving the performance of complex classification algorithms.

4.12 Ada Boost Classifier

This algorithm has a testing precision of 0.84 and a training precision of 1.00. AdaBoost Class categorizations is an identification system that combines numerous weak predictors into a single strong classifier via boosting. It can handle both categorical and numerical input features and is useful for a wide variety of classification problems.

4.13 Gradient Boosting

This algorithm has achieved a testing accuracy of 0.86 and a training accuracy of 1.00. Gradient Boosting is an ensemble boosting is a categorization system that combines numerous poor classifiers into a single great classifier. It is effective for a wide range of classification problems and can handle both categorical and numerical input features.

4.14 Cat Boost Classifier

This algorithm has achieved a testing accuracy of 0.91 and a training accuracy of 0.91. Cat Boost Classifier is a gradient boosting algorithm that is optimized for categorical input features. It is effective for handling high-dimensional datasets and can handle missing values in the input features.

4.15 Extra Trees Classifier

This algorithm has achieved a testing accuracy of 0.92 and a training accuracy of 1.00. Extra Trees Classifier is an ensemble Using bagging, the categorization algorithm merges many decision trees into an individual strong classifier. It is similar to Random Forest Classifier but with a different splitting criterion.

4.16 MLP Classifier

This algorithm has achieved a testing accuracy of 0.89 and a training accuracy of 0.93. MLP algorithm is a neural network-based categorization algorithm that consists of multiple layers of interconnected nodes. It is effective for handling complex and non-linear classification problems and can handle both categorical and numerical input features.

4.17 Stochastic Gradient Descent

This algorithm has achieved a testing accuracy of 0.89 and a training accuracy of 0.77. Stochastic Gradient Descent is an optimization algorithm that is commonly used for training linear classification models. It is particularly useful for handling large datasets and can handle both categorical and numerical input features.

4.18 Artificial Neural Network

This algorithm has achieved a testing accuracy of 0.92 and a training accuracy of 0.93. Artificial Neural Network is a powerful classification algorithm that consists of multiple layers of interconnected nodes. It is effective for handling complex and non-linear classification problems and can handle both categorical and numerical input features.

5. Result

CVD accounts for a substantial proportion of deaths in Bangladesh, accounting for almost one-third of all deaths**[1].** Early detection and prevention of CVD are crucial for reducing the mortality rate.ML and DL algorithms have shown great potential in predicting CVD risk, aiding early diagnosis and prevention. In that study, we compared the performance of various classifiers in predicting the presence of cardiovascular disease (CVD) based on a set of medical features. Our analysis included 17 different models, including Logistic Regression, G-Naive Bayes, B-Naive Bayes, SVM, X Gradient Boosting,

Decision Tree Classifier, Grid Search CV, Ada Boost, Gradient Boosting Classifier, Cat Boost, Extra Trees, KNN, MLP, Stochastic Gradient Descent and ANN. We assessed the models using their testing accuracy, training accuracy, precision, recall, and F1-score. Precision and recall were particularly important in this context as they give an idea of how well the models are able to correctly identify positive cases of CVD. Our results showed that the B-Naive Bayes model achieved the highest precision score of 0.99%, indicating that this model was able to accurately predict a significant number of real positive cases among all positive projections made by the model. The B-Naive Bayes model also performed well in terms of precision with a score of 100%. In terms of recall, the Grid Search CV obtained the greatest score of 0.99%, indicating that model properly identified a large majority of the cases of actual positive cases of CVD. Other models that performed well in terms of recall include the Bernoulli Naive Bayes model with a score of 0.97% and the S-gradient descent model with a score of 0.90%.

6. Conclusion and Future work

The high prevalence of CVD in Bangladesh is a major public health concern that requires urgent attention from government. There is a scarcity of nationally representative CVD prevalence data. which highlights the need for well-designed population-based surveys to provide accurate information on the extent and distribution of the problem ML and DL algorithms have shown potential in predicting CVDs in the Bangladeshi population. Given concerns about affordability and accessibility in the healthcare sector, ML-based models offer an accessible and affordable option, especially in remote places with possibly few healthcare facilities. We urge the development and testing of comparable systems using extensive cardiac data sets from all divisions of Bangladesh that represent all socioeconomic strata of the population. Public awareness programs and education on healthy lifestyle choices can also be incorporated to reduce the risk factors associated with CVDs. Preventive strategies should focus on identifying and addressing modifiable risk factors. Public awareness campaigns that emphasize the importance of a healthy lifestyle and the potential consequences of unhealthy behaviours may also be effective in reducing the burden of CVD in Bangladesh. In addition, efforts to improve the diagnosis and treatment of CVD, including increasing access to screening and diagnostic tools and ensuring appropriate medications and interventions, are essential. Addressing the high prevalence of CVD in Bangladesh will require a multifaceted approach that involves collaboration between healthcare providers, policymakers, and the public. Together, we might be able to lessen the impact of this illness and enhance the population's general health and wellbeing. In future, the results indicated that the system can be useful in healthcare intervention.

References

[1] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542-81554.

[2] Heart disease prediction using machine learning techniques" Vijeta Sharma, Shrinkhala Yadav, Manjari Gupta 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 177-181, 2020.

[3] Diwakar M, Tripathi A, Joshi K, Memoria M, Singh P, Kumar N. Latest trends on heart disease prediction using machine learning and image fusion. Mater Today Proc [Internet]. 2020;37(Part 2):3213–8. Available from: https://doi.org/10.1016/j.matpr.2020.09.07

[4] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), 1-6.

[5] Singh, A., & Kumar, R. (2020). Heart disease prediction using machine learning algorithms. 2020 International Conference on Electrical and Electronics Engineering (ICE3), 452-457.

[6] Patel, J., Upadhyay, T., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. Heart Disease, 7(1), 129-137.

[7] Khourdifi, Y., Bahaj, M., & Bahaj, M. (2019). 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, 12(1), 242-252.

[8] Jagtap, A., Malewadkar, P., Baswat, O., & Rambade, H. (2019). Heart disease prediction using machine learning. International Journal of Research in Engineering, Science, and Management, 2(2), 352-355.

[9] Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in Biology and Medicine, 136, 104672.

[10] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. IOP Conference Series: Materials Science and Engineering, 1022(1), 012072.

[11] Heart disease prediction using machine learning techniques" Vijeta Sharma, Shrinkhala Yadav, Manjari Gupta 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 177-181, 2020

[12] Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart disease prediction using machine

learning. International Journal of Research and Technology, 9(04), 659-662.

[13] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & technology, 7(2.8), 684-687.

[14] Heart disease prediction using hybrid machine learning model" M Kavitha, G Gnaneswar, R Dinesh, Y Rohith Sai, R Sai Suraj 2021 6th International Conference on Inventive Computation Technologies (ICICT), 1329-1333, 2021

[15] Prediction of heart disease using machine learning" Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Kailas Devadkar 2018 second international conference on electronics, communication and aerospace technology (ICECA), 1275-1278, 2018.

[16] Machine learning techniques for heart disease prediction: a comparative study and analysis" Rahul Katarya, Sunit Kumar Meena Health and Technology 11 (1), 87-97, 2021

[17] Cognitive approach for heart disease prediction using machine learning" Pranav Motarwar, Ankita Duraphe, G Suganya, M Premalatha 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 1-5, 2020

[18] A comprehensive review on heart disease prediction using data mining and machine learning techniques" Lamido Yahaya, N David Oye, Etemi Joshua Garba American Journal of Artificial Intelligence 4 (1), 20-29, 2020

[19] "A review on heart disease prediction using machine learning and data analytics approach" M Marimuthu, M Abinaya, KS Hariesh, K Madhankumar, V Pavithra

[20] "Heart disease prediction using machine learning techniques" Shekharesh Barik, Sambit Mohanty, Deepankar Rout, Subhra Mohanty, Akshaya Kumar Patra, Alok Kumar Mishra.

[21] Subramani S, Varshney N, Anand MV, Soudagar MEM, Al-Keridis LA, Upadhyay TK, Alshammari N, Saeed M, Subramanian K, Anbarasu K, Rohini K. Cardiovascular diseases prediction by machine learning incorporation with deep learning. Front Med (Lausanne). 2023 Apr 17;10:1150933. doi: 10.3389/fmed.2023.1150933. PMID: 37138750; PMCID: PMC10150633.

[22] Barhoom, Ali M. A. ; Almasri, Abdelbaset ; Abu-Nasser, Bassem S. & Abu-Naser, Samy S. (2022). Prediction of Heart Disease Using a Collection of Machine and Deep Learning Algorithms. International Journal of Engineering and Information Systems (IJEAIS) 6 (4):1-13.

[23] Vincent Paul, S.M., Balasubramaniam, S., Panchatcharam, P. et al. Intelligent Framework for Prediction of Heart Disease using Deep Learning. Arab J Sci Eng 47, 2159–2169 (2022). https://doi.org/10.1007/s13369-021-06058-9.

[24] Saikumar, K., Rajesh, V. A machine intelligence technique for predicting cardiovascular disease (CVD) using Radiology Dataset. Int J Syst Assur Eng Manag (2022). https://doi.org/10.1007/s13198-022-01681-7.

[25] Bhavekar, G.S., Goswami, A.D. A hybrid model for heart disease prediction using recurrent neural network and long short term memory. Int. j. inf. tecnol. 14, 1781–1789 (2022). https://doi.org/10.1007/s41870-022-00896-y.

[26] Ahmad, S., Asghar, M.Z., Alotaibi, F.M. et al. Diagnosis of cardiovascular disease using deep learning technique. Soft Comput 27, 8971–8990 (2023). https://doi.org/10.1007/s00500-022-07788-0.

[27] A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms" Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Shah Nazir, Ruinan Sun Mobile Information Systems 2018.

[28] Chauhan, A., Negi, P., & Chauhan, S. (2019). Heart disease prediction using machine learning algorithms: a comparative analysis. 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), 1-6.

[29] Khan, M. A., Akhtar, N., & Ahmad, I. (2019). Heart disease prediction system using machine learning techniques. International Journal of Computer Science and Network Security, 19(3), 127-133.

[30] Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), 250-255.

[31] Singh, M., Sharma, S., & Singh, H. (2016). Prediction of heart disease using machine learning algorithms: a survey. International Journal of Computer Applications, 139(11), 22-25.

[32] Alghamdi, M., Al-Mallah, M., & Keteyian, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PLoS One, 12(7), e0179805.

[33] Masethe, H. D., & Masethe, M. A. (2014). Prediction of heart disease using classification algorithms. Proceedings of the World Congress on Engineering and Computer Science, 1, 22-24.

[34] Sathyadevi, K., & Subramanian, R. (2011). Heart disease prediction system using supervised learning classifier algorithms. International Journal of Computer Applications, 31(10), 5-9.

[35] Tandel, H., Vora, S., & Patel, R. (2020). Heart disease prediction using machine learning and artificial

intelligence techniques: a systematic review. Journal of Ambient Intelligence and Humanized Computing, 1-15.

[36] Das, R., Turkoglu, I., & Sengur, A. (2009). Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications, 36(4), 7675-7680.

[37] Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. R News, 2(3), 18-22.
mbers.