# BioLearner: A Machine Learning-Powered Smart Heart Disease Risk Prediction System Utilizing Biomedical Markers

Syed Saad Amer
School of Electrical Engineering and Computer Science, Queen Mary University of London, London, UK

Gurleen Wander
Chelsea and Westminster Hospital NHS Trust London, London, UK

Manmeet Singh
Jackson School of Geosciences, University of Texas at Austin, Texas, USA
Centre for Climate Change Research, Indian Institute of Tropical Meteorology (IITM), Pune, India

Rami Bahsoon
School of Computer Science, University of Birmingham, Birmingham, UK

Nicholas R. Jennings
Department of Computing, Imperial College London, London, UK

Sukhpal Singh Gill
School of Electronic Engineering and Computer Science, Queen Mary University of London, Mile End Rd, Bethnal Green, London E1 4NS, UK
s.s.gill@qmul.ac.uk

**Abstract** - Heart disease kills more people around the world than any other disease, and it is one of the leading causes of death in the UK, triggering up to 74,000 deaths per year. An essential part in the prevention of deaths by heart disease and thus heart disease itself is the analysis of biomedical markers to determine the risk of a person developing heart disease. Lots of research has been conducted to assess the accuracy of detecting heart disease by analyzing biomedical markers. However, no previous study has attempted to identify the biomedical markers which are most important in this identification. To solve this problem, we proposed a machine learning-based intelligent heart disease prediction system called BioLearner for the determination of vital biomedical markers. This study aims to improve upon the accuracy of predicting heart disease and identify the most essential biological markers. This is done with the intention of composing a set of markers that impacts the development of heart disease the most. Multiple factors determine whether or not a person develops heart disease. These factors are thought to include Age, history of chest pain (of different types), fasting blood sugar of different types, heart rate, smoking, and other essential factors. The dataset is analyzed, and the different aspects are compared. Various machine learning models such as K Nearest Neighbours, Neural Networks, Support Vector Machine (SVM) are trained and used to determine the accuracy of our prediction for future heart disease development. BioLearner is able to predict the risk of heart disease with an accuracy of 95%, much higher than the baseline methods.

***Index Terms—Artificial Intelligence, Machine Learning, Heart Disease, Disease Detection, Biomedical Markers***

## I. INTRODUCTION

Heart Diseases is an umbrella term covering a range of cardiovascular diseases such as heart attacks, heart defects, heart failure, and more. The patient suffering from these diseases may exhibit certain symptoms such as chest pain. Additional testing may show other biomedical markers outside of the range of expected values. These biomedical markers can thus be used as features in determining the risk a person faces of developing heart disease and the individual weight and impact of each feature on the development of heart disease. Heart disease is the most significant cause of death globally and Britain [1]. The incidents of heart disease have increased with more and more people adopting modern lifestyles, with changes in diet and physical activity. The most critical factors behind heart disease are thought to be obesity, diabetes, blood cholesterol levels and smoking [2]. These factors have seen a marked shift as societies have adopted more sedentary lifestyles and a diet rich in carbohydrates. This has led to an increased risk of heart disease and death. The rapid adoption of Artificial Intelligence and Machine Learning for prediction and detection in healthcare has made it possible to detect diseases before they occur with

the efficiency that is not possible without them [3]. Heart disease has traditionally been considered a very tricky disease to detect and known to act as a silent killer, with patients unaware that they were affected [4]. The use of machine learning and artificial intelligence has made it possible to calculate a risk of patient developing heart disease with a reasonable degree of accuracy [5]. This significant aid in detection could lead to proper early diagnosis of heart disease and, therefore, adequate treatment before the patient reaches a critical stage or death—deficiencies in the correct detection of heart disease cost health services millions [6]. With the correct determination of the risk of developing heart disease, healthcare providers could look out for specific biomedical markers and tailor treatment accordingly [7].

Recent campaigns to limit obesity and a better understanding of the risks behind heart disease have led to a fall in heart disease rates in the last decades, but it seems to be rising again [8]. Multiple treatments such as Bariatric surgery and Gastric Bypass surgery and Liposuction, are available for the treatment of the factors that cause heart disease. However, the impetus to perform these procedures is not apparent in all but the most extreme cases [9]. Previous work has been done in trying to determine the risk of heart disease using Machine Learning and Artificial Intelligence. However, it does not focus on identifying the most essential biomedical markers in making that prediction. This study uses new and improved algorithms and techniques to further improve that accuracy and attempts to create a set of biomedical markers ranked by their importance. Numerous features are considered in the data provided to the machine learning algorithms, but the value of individual features is never the same.

An example is considering the colour of the car while trying to determine the average speed of the car. The colour is simply a happenstance and has no impact on the rate the driver drives. Any influences in factors such as these can be discarded as circumstantial. Other factors may indeed impact the variable that is being determined. However, the impact may be so minute that the effects of including that feature without properly weighing it may instead lead to misleading conclusions [10]. In this case, care must be taken to either pay special attention to outliers or ignore them completely while trying to minimize overfitting.

A thorough analysis of the features in the dataset shows that not all features are equal in the determination of heart disease. Powerful machine learning techniques including k-Nearest Neighbor, Neural Networks, Support Vector Classifiers (SVC), Random Forest Classifiers and Logistic Regression have been used to get the highest accuracy for determining the risk of heart disease, with the Random Forest algorithm giving the highest accuracy of 95%. It is not possible to determine the risk for individual heart diseases such as heart attacks or cardiac arrest themselves, as a much larger and more comprehensive dataset would be needed than is currently available. The general risk for developing heart disease, however, can be calculated. Following this, the biomedical markers are listed in descending order of importance, with chest pain, Thallium tests, sex, and age being the most important markers.

### 1.1 Motivation and Our Contributions

The economic impact of this system if used by health services like the NHS, can be significant. Deficiencies in detecting heart disease cost millions (avoidable shortcomings in heart failure prediction cost the NHS £21m in 2019) [11]. This can potentially be significantly reduced through the wide-scale deployment of prediction systems. Socially, the patients with heart disease will receive an earlier diagnosis and timely medical attention, resulting in better treatment and facefewer adverse effects such as pain and severe complications that could result in death. Heart disease currently has notoriously low rates of detection and diagnosis, especially in women. This system depends upon an extensive, and accurate dataset of medical information. Legal requirements mean that patient information must be protected, and efforts must be made to ensure that no patient is identifiable from the data to avoid issues such as privacy [26]. The key contributions of this work are:

- We propose a machine learning-based intelligent heart disease prediction system for risk determination using biomedical markers called **BioLearner**.
- We develop multiple models for the detection of heart disease and indicate the model with the highest accuracy. This also involves analysis of the dataset regarding the features and the corresponding occurrence of heart disease.
- We determine which biomedical markers serve as useful features in assessing the risk of heart disease, useful for medical practitioners in future.

Further research can be carried out to determine why the chosen features provide a higher risk assessment than the set of all biomedical markers while studying their impact in more detail. This information can then be used to diagnose at-risk patients and help them to manage their situation, saving lives and public funds. This study has not been conducted to develop a tool that serves as an alternative to trained medical practitioners but as a

complementary tool to assist them in determining which patients are more in need of swift and appropriate medical attention. The rest of the paper is structured as follows. Section 2 presents the related work. Section 3 describes the methodology. Section 4 presents performance evaluation and experimental results. Section 5 presents conclusions and highlights future directions.

## II. Related Work

There has recently been a push to try to get an accurate risk assessment for heart disease using machine learning and artificial intelligence. All but one study mentioned ahead involves use of the same UCI Heart Disease dataset [12]. The techniques range from using algorithms such as k-means clustering, DNA based learning, Ensemble learning, Fuzzy logic, and more. A short comparison shows that ensemble learning using Decision Trees gives a high accuracy for risk prediction. Neural Networks also give a high accuracy [8]. Latest and relevant related works have been discussed in this section.

Algorithm to assign different weights to different features in order to improve accuracy. The highest accuracy they achieved was 89% [8]. However, the basic algorithms seem to be from a code library with little tailoring for the situation. This study [13] looks at the possibility of detecting Heart Disease using a dataset similar to UCI dataset in conjunction with a custom-built system (MAPO) to calculate heart rate by video imagery. This data represents a sample from the North Indian patients. We call this data as UCII. The algorithms used are Logistic Regression, Naïve Bayes, and Artificial Neural Networks. They achieved a maximum accuracy of 89% using Naïve Bayes. A study on a dataset from a South African heart disease dataset using conventional techniques to identify risk of Coronary Heart Disease reported relatively high accuracies for risk detection [14]. This is the only study mentioned that does not use the UCI dataset, but is important so that the generality of our predictions can be ascertained. Some works have attempted to form a hybrid of different techniques to give the highest accuracy of risk detection for heart disease. An example of this is a system that is a simple combination of all well-known algorithms to prove that this is a viable method of using Data Mining techniques for detection of heart disease [15]. Another study in this direction has used Optimal Multi-nominal Logistic Regression (OMLR) to detect if the condition of the heart is severe [16]. This has yielded an accuracy of 92%. A study using a fuzzy based system and genetic algorithms [17] made a hybrid system that is very fast and identifies heart disease fairly accurately. A recent study makes a hybrid system by combining genetic algorithms (to assign suitable weights in the hidden layers), fuzzy logic, PCA (Principle Component Analysis) used to reduce feature dimensions and increase speed along with open source tools WEKA and KEEL to diagnose and predict heart disease using the UCI dataset [18].

All of the previously mentioned studies have had the objective of predicting heart disease, and the accuracies have shown a general trend of improving. However, none of these studies have even had the objective determining the most important biomedical markers, a significant step if Heart Disease is to be caught and treated earlier on. This study intends to both improve the accuracy of prediction and identify the most Biomedical markers in the dataset used. It is known that all features in a dataset ate never equally important. Redundant features may cause overfitting, decrease accuracy and increase processing time due to increased volume of data. In addition, most of the studies mentioned use implementations of machine learning algorithms as they are in existing code libraries, or use existing software such as RapidMiner for machine learning tasks, with very little customizing for this specific situation. Little work has been done on actually selecting the ideal set of features for the detection of heart disease. A number of features present in the UCI dataset could be considered to be irrelevant or redundant when performing machine learning. A very recent study looks into this a bit [19]. Their efforts involve using a chi-squared feature evaluator to identify certain features and use them for predicting heart disease. In another study [30], authors used genetic algorithm based trained recurrent fuzzy neural networks for the diagnosis of heart disease without considering biomedical markers.

However, different sets of features sometimes give contradictory results. Khourdifi and Bahaj [20] attempted to use fast feature selection to remove redundant features before performing machine learning techniques such as KNN, Decision Trees, Multi-layered perceptrons, Random Forests, and SVMs. While they do manage to increase their accuracy, they do not have a way of identifying the most valuable features for further study. An early attempt at utilizing machine learning techniques to datasets for feature selection by Polat and Güneş [21] involves converting the feature space into kernel space using Linear or Radial Basis Functions. The F-score values of medical datasets with a large number of features is calculated. A mean F-score value is used to filter the datasets. They call this the Kernel F-score Feature Selection (KFFS) and it is used to remove the irrelevant or redundant features so that they can no longer adversely impact the accuracy if the results. While it is possible to use PCA (Principle Component Analysis) to speed up the task of classification and to reduce the number of features while retaining relative accuracy, it is very difficult to tell which features were found to be more useful than others. Table 1 presents the comparison of proposed work with existing works based on important parameters. This is the

first work which identifies the important biomedical markers using ensemble learning and able to predict the risk of heart disease with an accuracy of 95%, much higher than the baseline methods on the same set of data.

Table 1: Comparison of proposed work (BioLearner) with related works

| **Work** | Neural Networks | Ensemble Learning | Heart Disease Risk Estimation | Identification of Important Biomedical Markers | Reducing features to increase accuracy |
|---|---|---|---|---|---|
| Polat and Güneş [21] | | | | ✓ | ✓ |
| Spencer, Thabtah, Abdelhamid and Thompson [19] | ✓ | | ✓ | | ✓ |
| Khourdifi and Bahaj [20] | ✓ | | ✓ | | ✓ |
| Satyanandam, N. and Satyanarayana, C [16] | | | ✓ | | |
| Tarawneh and Embarak [15] | ✓ | | ✓ | | |
| Fida, Nazir, Naveed and Akram [11] | ✓ | ✓ | ✓ | | |
| Abdeldjouad, Brahami and Matta [18] | ✓ | | ✓ | | ✓ |
| BioLearner (this work) | ✓ | ✓ | ✓ | ✓ | ✓ |

## III. METHODOLOGY

This Section presents the system architecture for determination of key biomedical markers using machine learning and architecture is shown in Figure 1. The objectives are to use and develop machine learning models to get a percentage as a measure of the risk an individual faces of developing heart disease, and to identify the most important biological markers that correlate with heart disease.

### *A. Dataset*

The dataset used for this study is similar to the UCI Heart Disease dataset (UCII). It represents a sample from the anonymized Northern Indian patients received from health authorities. It is similar to UCI dataset in all aspects but is useful to develop the model for a different population sample. This dataset was chosen because a majority of the studies which have already been conducted and all the studies mentioned previously (as discussed in Section 2), have been conducted using same dataset, so comparison of accuracy would be far easier and more useful than if a completely different dataset were chosen, with different features and different variation between the values of each Biomedical marker. This dataset provides valuable biomedical markers which are already known to be significant in the development of heart disease. The dataset is described in Table 2 and the sample of the dataset is given in Table 3. The Data is trained for the 'Target' attribute which denotes whether the patient developed heart disease or not. It can only take two values: 1 represents heart disease, and 0 represents a lack of heart disease. The 'Target' attribute is thus the classification attribute.

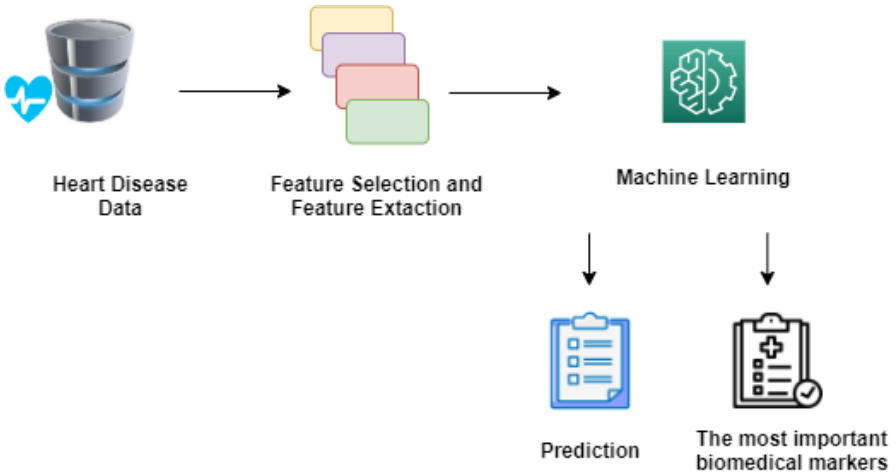



Fig. 1. System Architecture of BioLearner

### *B. System Architecture*

Figure 1 shows the system architecture of BioLearner. Our system inputs heart disease data from a CSV file. This data is then pre-processed to clean it and to make it suitable for feature extraction. The data is analyzed and compared to each other using analytical tools such as Pandas in conjunction with Jupyter Notebook (iPython) to

give a visual representation of the data so that any obvious correlations may be immediately recognized. The features are compared to the 'Target' attribute to see how the 'Target' attribute varies as the other features take their range of values. In this analysis, the features are the *Independent variables* and the 'Target' attribute is the *Dependent variable.*

Table 2: Explanation of Dataset features.

| Feature Number | Attribute Description | Type of Values |
|---|---|---|
| 1 | *Age:* Age of the person | Multiple values |
| 2 | *Sex:* The gender of the person ['0' means female, '1' means male'] | 0 and 1 |
| 3 | *CP:* The level of chest pain the patient is suffering. | 0, 1, 2, 3 |
| 4 | *RestBP:* The value of the patients resting Blood Pressure | Multiple Values |
| 5 | *Chol:* The cholesterol levels of the person | Multiple Values |
| 6 | *FBS:* The Fasting Blood Sugar level of the person. Two values check if it is greater or lower than 120mg/dl | 0 and 1 |
| 7 | *restECG:* Three different levels to show the waveforms as shown by an ECG machine | 0,1, 2 |
| 8 | *HeartBeat:* The maximum heart rate | Multiple Values |
| 9 | *Exang:* Exercise induced Angina | 0 and 1 |
| 10 | *OldPeak:* Depression included by exercise relative to rest. Between 0 and 6.2 | Multiple Values |
| 11 | *Slope:* The condition of the person during the peak exercise segment. Represents the gradient of the tangent to the slope (increasing, decreasing, flat). | 0, 1, 2 |
| 12 | *CA:* The number of blood vessels colored by fluoroscopy | 0,1,2,3,4 |
| 13 | *Thal:* four different values as results of Thalium tests | 0,1,2,3 |
| 14 | *Target:* The label attribute, signifies whether a person had heart disease or not | 0 and 1 |

Table 3. Sample of the dataset (using pandas and Scikit Learn).

| age | sex | cp | trtbps | chol | fbs | restecg | thalachh | exng | oldpeak | slp | caa | thall | output |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 63 | 1 | 3 | 145 | 233 | 1 | 0 | 150 | 0 | 2.3 | 0 | 0 | 1 | 1 |
| 37 | 1 | 2 | 130 | 250 | 0 | 1 | 187 | 0 | 3.5 | 0 | 0 | 2 | 1 |
| 41 | 0 | 1 | 130 | 204 | 0 | 0 | 172 | 0 | 1.4 | 2 | 0 | 2 | 1 |
| 56 | 1 | 1 | 120 | 236 | 0 | 1 | 178 | 0 | 0.8 | 2 | 0 | 2 | 1 |
| 57 | 0 | 0 | 120 | 354 | 0 | 1 | 163 | 1 | 0.6 | 2 | 0 | 2 | 1 |
| 57 | 1 | 0 | 140 | 192 | 0 | 1 | 148 | 0 | 0.4 | 1 | 0 | 1 | 1 |
| 56 | 0 | 1 | 140 | 294 | 0 | 0 | 153 | 0 | 1.3 | 1 | 0 | 2 | 1 |
| 44 | 1 | 1 | 120 | 263 | 0 | 1 | 173 | 0 | 0 | 2 | 0 | 3 | 1 |
| 52 | 1 | 2 | 172 | 199 | 1 | 1 | 162 | 0 | 0.5 | 2 | 0 | 3 | 1 |

Machine Learning algorithms are then applied on the dataset independently of each other. The results for each are retained. The machine learning methods used in this study are Naïve Bayes, Logistic Regression, SVM (Support Vector Machines), Random Forest, K Nearest Neighbor, and Neural Networks. These are all explained in more detail below. As mentioned before, the 'Target' attribute is the classification or label attribute. This means that this is the label for which all of the different machine learning models in this study are trained. The data is split in a 4:1 ratio for Training and testing data. The classifiers are trained using 80% of the data and 20% of the data is held back. After the classifiers have been trained, the 'Target' attribute is removed, and the 20% of which was held out is tested. The classifiers determine whether or not the records belong to 'Target' category 1 with an estimation that can range from between 0 to 100. An accuracy closer to 0 means that the record is less likely to signify heart disease. An accuracy closer to 100 means that the record is more likely to signify heart disease. K-fold cross validation is also performed as it provides a less biased model than other cross validation methods. Our dataset is also relatively small, so this cross validation is a resampling procedure to give a less biased result.

The results from this machine learning process is then given as the prediction. The features are updated with a running dictionary using the Algorithm 1. Algorithm 1 works by assigning each feature a score. The initial score of each feature is set to zero. Each feature is then sequentially dropped from the dataset (as it is acted upon by all the machine learning algorithm). If the accuracy of the algorithm increases form dropping a feature, that feature is assigned a negative score, and if dropping the feature results in a decrease in accuracy, that feature is assigned a positive score. If dropping the feature has no effect, that feature is assigned a score of 0. These scores are added to each other and the features given in descending order by score. The features correspond to the biomedical markers in the initial data, and their position and score signifies their relative importance in determining whether the patient has risk of developing heart disease. An overview of the machine learning algorithms is given as follows:

- The Naïve Bayes technique is based upon the Bayes Theorem, which supposes that the components are independent of each other. Naïve Bayes assumes that the features are completely independent from each other and are unrelated. It is used for supervised classification learning models. As has been mentioned before, Naïve Byes is a commonly used algorithm for problems such as these.
- Logistic Regression is a classification technique used for predicting outcomes in a scenario where only two outcomes are possible. It does so by performing regression analysis. In our case, the two outcomes correspond to the 'Target' attribute, which can take value of 0 or 1.
- Random Forests are an ensemble learning method for classification [22]. It can also be used for other forms of estimation, but as we have a binary classification problem, we used the RandomForestClassifier library from Sci-kit Learn. Random Forests choose an N (random) number of records and build Decision Trees for those, thus making a Forest. Each tree in the forest gives a prediction of the category to which an unseen record belongs. The mode of these is the class label assigned to the new and unseen record. The Random Forest method do not allow for overfitting like in Decision Trees, and reduce bias. However, they are complex and take more time and resources.

**ALGORITHM 1: ALGORITHM FOR EXTRACTING BEST FEATURES**

```
BEGİN
Calculate Initial Accuracy Of All Algorithms;
    for all algorithms do
        for all features do
            Drop feature;
            Make prediction without feature;
            if new prediction is Greater Than original prediction then
                feature score decreases;
            else
            end
            if new prediction is Less Than original prediction then
                feature score increases;
            else
            end
        end
    end
Features Presented by Score;
END
```

- SVM or Support Vector Machine is a classification algorithm that needs a labelled dataset. This algorithm creates hyperplanes and then attempts to separate the records into different classes. We are dealing with two well defined classes in our study. No heart disease ('Target' = 0) and positive heart disease ('Target' = 1). Thus, we use a linear kernel for our SVM.
- K-Nearest Neighbour (KNN) is a supervised machine learning algorithm widely used for classification problems like the one we have. KNN works by finding the distance between a new record and the records already in the dataset. It then choses from between the K examples closest to the new record and votes for the most frequent label. The number K is defined by the user. Determination of the ideal value for K is done by repeating the algorithm several times with different values of K. In our case, the value of K that yielded the highest accuracy was K = 8.

- Neural Networks are a very powerful machine learning algorithm that assign weights to different edges and try to determine the correct outcome. The amount of forward propagation depends on the weights, which can be changed during back propagation. Forward and backwards propagation are done iteratively on a training dataset to train the Neural Network. The larger the training dataset, the better the accuracy for any predictions should be. There can be a number of layers between the input and output layer. These are called hidden layers.

The classifiers used in this study are sourced from the Scikit Learn and Keras libraries, but their implementation has been tailored to ensure maximum accuracy. For example, the Neural Network contains a variable number of hidden layers depending on the features provided to the model. The Random Forest Classifier also calculates the mode of a variable number of Decision Trees and is implemented to give the highest accuracy possible. The Logistic Regression and Naïve Bayes classifiers however are implemented as given in the code libraries due to their relative simplicity as compared to the others.

## IV. Performance Evaluation

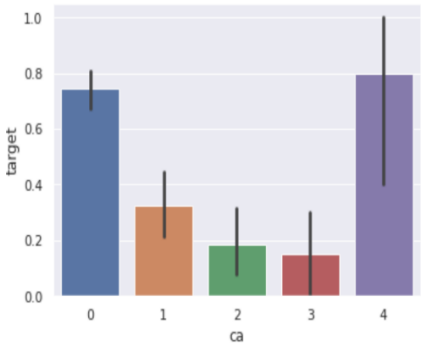
The details of software and hardware tools used for performance evaluation are given in this section. We used a dataset similar to the Cleveland heart disease dataset distributed by the University of California Irvine (UCI) representative of the Northern Indian population [12]. We used various software tools such as Keras, Python 3, Scikit Learn, Jupyter Notebook, Numpy, Pandas, Matplotlib and Seaborne to conduct this study. The machine learning was performed using Google Colab as well as using Windows 10 running on a Core i7 8550U with 32GB RAM, and 1 TB Solid State Drive.

***a) Data Analysis:*** Figure 2 shows the data analysis of different variables and features. Our analysis of the 'Target' variable shows that there are 165 records with the value of 'Target' being 1 and 138 records with the value of 'Target' being 0. To analyze the features that we have, we create bar plots with the features as the independent variable, and the proportion of 'Target' attributes as the dependent variable. Our analysis of the 'sex' attribute shows that the data we have, heart disease is more common in women than in men. This seems to contradictory to established research which is widely agreed on the position that heart disease is much more common in men than women, sometimes by a ratio of two to one [23]. This shows the dataset is not very representative with regard to the genders of the patients. The limited size of the dataset is made apparent by this as well. Analysis of the 'restecg' attribute shows that the majority of patients were suffering from mild pain (restecg = 1). These and the patients with restecg = 0 form a larger proportion of the patients who have heart disease. Patients with restecg = 2 normally do not have heart disease. Analysis of the attribute 'ca' shows that patients with 'ca' = 0 and 'ca' = 4 were much more likely to have a positive diagnosis of heart disease. Analysis of the 'cp' attribute shows that the patients with non-typical anginal pain are much more probable to have heart disease than those suffering from typical anginal pain ('cp' = 0). Analysis of the 'exang' attribute suggests that the patients who do not test positive for exercise induced angina 'exang' = 0) are more likely to have heart disease. Analysis of the 'fsb' attribute shows that those with fasting blood sugar greater than 120mg/dl ('fsb' = 0) are slightly more likely to have heart disease, but not by much. Analysis of the 'slope' attribute reveals that the patients with 'slope' = 2 have a much higher probability of having heart disease. Analysis of the 'thal' attribute shows that patients with positive thalium test type 'thal' = 2 have the highest probability of heart disease.

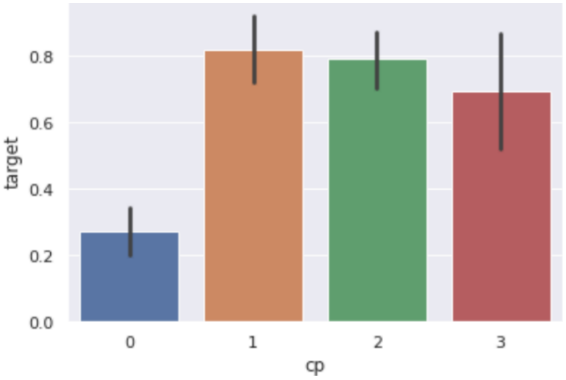
***b) Machine Learning based Performance Analysis:*** A comparison of the accuracy of predicting heart disease after adjusting the features for the most important biomedical markers is given in Figure 3. The following finding have been observed:

- Dropping the attribute 'Age' reduces the accuracies of K-NN, Random Forests, and Neural Networks, while not producing a change in the accuracies of the other algorithms. Age is thought to be a critical factor in the occurrence of heart disease as mentioned previously, and this result was expected.
- Dropping the attribute 'sex' attribute reduces the accuracy for the Naïve Bayes and Logistic Regression algorithms. It improves the accuracy of the K-NN algorithm. It has no effect in the other algorithms. Heart disease is far more common in males, but rates for diagnosis in females is low even though they may suffer from it. Thus, this outcome was expected.
- Dropping the 'cp' algorithm reduces the accuracy in all algorithms substantially. That chest pain is important in determining heart disease comes as no surprise. Different types of heart disease effect all cause chest pain. A reduced accuracy was expected.
- Dropping the 'trestbps' attribute reduces the accuracy of the Naïve Bayes, K-NN, and Neural Network. The resting blood pressure seems to be an important marker in the set of features.

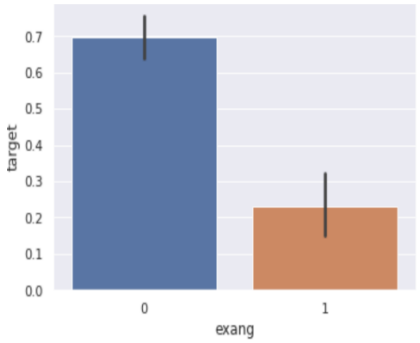
- Dropping the ‘chol’ attribute decreases the accuracy of the neural network, while increasing the accuracy of the K-NN algorithm. The rest of the algorithms give the same accuracy. According to this, the cholesterol levels do not appear to be significant in the occurrence of heart disease. This seems to against conventional wisdom but is in agreement with another study conducted in the USA and Japan in 2016 [24].



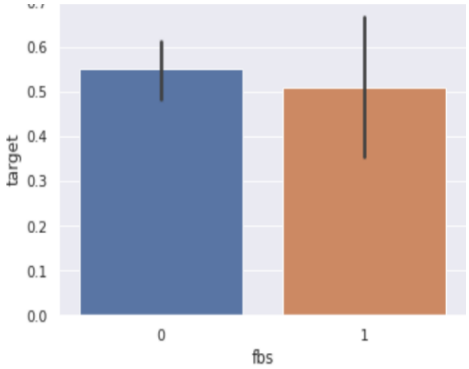

a) Analysis of ‘ca’



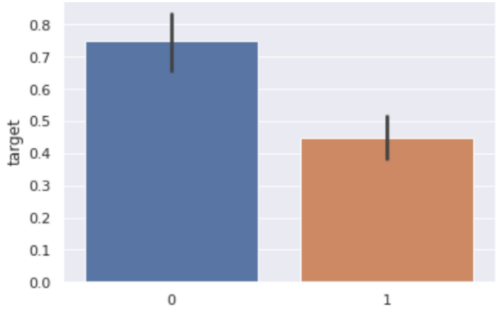

b) Analysis of ‘cp’



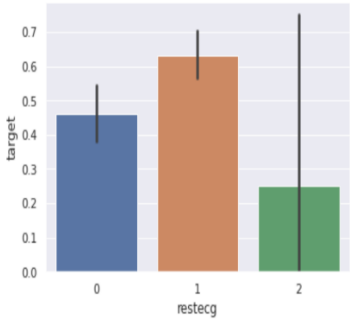

c) Analysis of ‘exang’



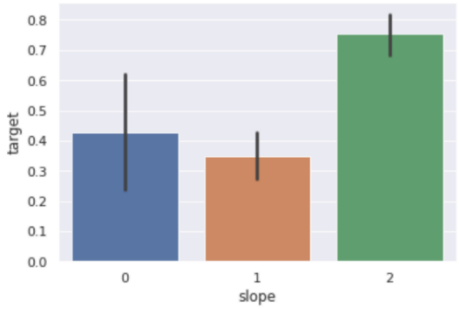

d) Analysis of ‘fbs’



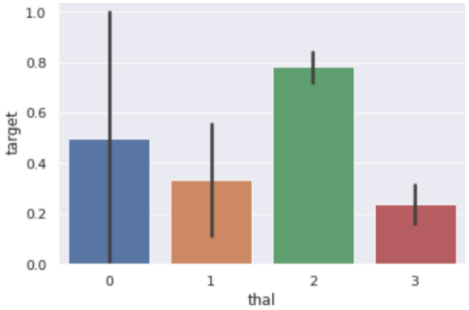

e) Analysis of ‘sex’




f) Analysis of ‘restecg’




g) Analysis of ‘slope’




h) Analysis of ‘thal’

Figure 2: Data Analysis of Different Variables and Features

- Dropping the ‘FSB’ attribute resulted in an increase in accuracy with the SVM, but decreases in accuracy with Naïve Bayes, K-NN, and the neural network. So fasting blood sugar seems to be an important biomedical marker. It is known that diabetes and heart disease are co-morbid to some degree.
- Dropping the ‘RestECG’ attribute causes the accuracy of the SVM and the neural network to decrease, while the rest stay the same. ECG tells about damage to the heart and it seems this information is useful prediction of heart disease.
- Dropping the ‘thalach’ (peak heart rate) attribute causes the accuracy of the SVM to increase while causing the accuracy of the K-NN and neural network to decrease. The rest remained constant. So this feature seems to be somewhat important. The peak heart rate would also normally be assumed to give an indication of cardiac health.
- Dropping the ‘exang’ attribute (exercise induced angina) causes the accuracy of logistic regression to increase while the rest remains constant. This would seem to indicate that pain due to exercise is not related to heart disease, but may involve some other factors.
- Dropping the ‘oldpeak’ attribute (depression induced by exercise relative to rest) causes the accuracy of the logistic regression and the SVM to increase, while causing the accuracy of the Naïve Bayes algorithm, the K-NN, and the neural network to decrease.
- Dropping the ‘slope’ (condition during peak exercise) attribute causes the accuracies of the Naïve Bayes algorithm to decrease. It also causes a significant decrease in the accuracy of the K-NN algorithm. The rest remains the same. Both this and ‘oldpeak’ would seem to indicate that though exercise induced chest pain is not an indicator of heart disease, the depression and gradient of a person’s condition does play a role.
- The ‘CA’ attribute (number of coloured vessels in fluoroscopy) causes the accuracies of the Logistic Regression and the SVM to increase when it is dropped. This does not have any effect on the other algorithms. This indicates that the determination of heart disease based upon the identification of different coloured vessels in this non-invasive diagnosis method is not very useful.
- And finally, dropping the ‘thal’ attribute (thalium test) causes the accuracies of the all the algorithms to decrease. This shows that it is a very important biomedical marker in the determination of heart disease. The Thalium test involves high powered nuclear imaging and is very thorough. It is no surprise that it play a large part in the determination of heart disease.

The most important biomedical markers identified by the algorithm for predicting heart disease are in ascending order: Chest Pain, Thalium Tests, Sex, Age, Resting Blood Pressure, Fasting Blood Sugar, Resting ECG, condition during peak exercise, peak heart rate, depression induced by exercise relative to rest, cholesterol, exercise induced angina, and the number of colored vessels during fluoroscopy. Figure 3 shows the performance comparison of machine learning algorithms. The best result was given by the Random Forest Classifier, which identified heart disease with an accuracy of 95%. This is higher than the study conducted by Haq et al. [25] on the same dataset. It is also higher than this study [16], also conducted on the same dataset.

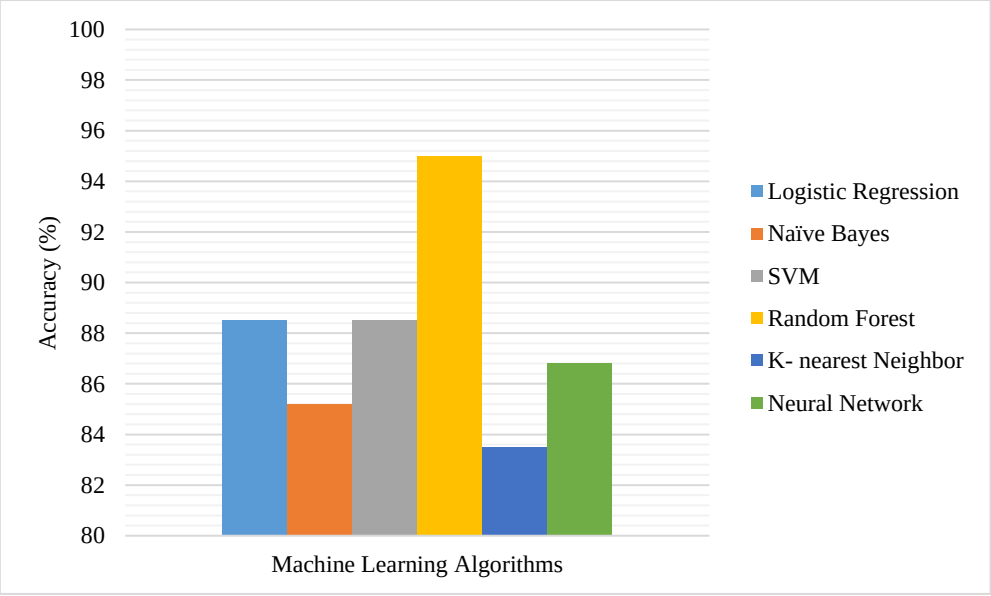



Figure 3: Comparison of Machine Learning Algorithms based on accuracy

## V. CONLCLUSIONS AND FUTURE WORK

This paper proposed a machine learning-based smart heart disease prediction system for risk determination using biomedical markers called BioLearner. This study manages to improve further the task of being able to detect heart disease using the UCII dataset. The results are marginally better than those from other studies. The Random Forest algorithm shows particular improvement. This specific algorithm was not used in most previous studies and accounts for bias and overfitting. Perhaps the Random Forest was not used by earlier studies due to its long time to execute. The Logistic Regression and SVM algorithms give the same accuracy, even though SVM tries to find the best line to divide the categories, while Logistic Regression uses different criteria for the weights of the features. Perhaps some of the rounding off built into the algorithms has obscured the differences in accuracy. But by and large, the results of the algorithms are in line with previous studies. However, the neural network shows a marked improvement in accuracy compared to other studies done using Neural Networks or multilayer perceptrons. The number of hidden layers and the weights assigned is specific to every study, and the combination used in this study has produced better results. The study also ranks the biomedical markers in the dataset by their importance in heart disease. The ranking of these factors confirms some apparent preconceptions such as 'chest pain'. A point to particularly note is that the cholesterol level has been ranked as a minor factor by the algorithm used in this study. That study has faced a lot of criticism over it, concluding that cholesterol is not essential to heart disease. However, no anomalies were found in its conduct. Or it may be that the Cleveland dataset is simply too small to account for all the factors that affect heart disease precisely. The more data we have, the more we can correctly train our machine learning models, and our predictions will be more accurate.

In future, BioLearner can be extended in many ways. The principal problem in detecting heart disease is that the biomedical markers used as features in datasets used to train machine learning models do not have the same real-life impact on the development of heart disease [27]. This research can help in the assigning of more accurate weights. This may be done by weighing more heavily on the factors identified by this study. The number of features can also be increased to include those a category for those who smoke. It has previously been mentioned that smoking has a high impact on heart disease. Exactly how much in relation to other factors is a study that can be pursued by collecting the above data from people smoke and from those who do not smoke. Further, the relationship between non-smoking and smoking people can be identified based on the some important factors such as frequency of smoking. The same approach can be taken by having data about diet and exercise to increase the number of features and be able to arrive at a correct conclusion. The accurate weighting of these factors is vital to correctly identifying heart disease in a patient [28]. Biomedical markers identified by this study can be further studies to see how they vary from typical values and how they can be improved to ensure fewer people suffer from and are at risk of heart disease [29]. Finally, the machine learning models proposed in this study can be run in an ensemble fashion in real Serverless computing environments using frameworks such as iFaaSBus [31].

## SOFTWARE AVAILABILITY

BioLearner has been released as open-source software. The implementation code with experiment scripts and results can be found at the GitHub repository: https://github.com/iamssgill/BioLearner. Any further information can be availed from the corresponding author on reasonable request.

## ACKNOWLEDGMENTS

We would like to thank Muhammed Golec (QMUL, UK) and Shreshth Tuli (Imperial College London) for their useful suggestions and discussion to improve the quality of the paper.

## REFERENCES

[1] Roberts, M., 2013. *Unhealthy Britain: Five Big Killers*. [online] BBC News. Available at: <https://www.bbc.com/news/health-21667065> [Accessed 6 August 2020].

[2] Amir Masoud Rahmani, Zahra Babaei, and Alireza Souri. "Event-driven IoT architecture for data analysis of reliable healthcare application using complex event processing." Cluster Computing 24, no. 2 (2021): 1347-1360.

[3] Pratik Goswami, Amrit Mukherjee, Bishal Sarkar, and Lixia Yang. "Multi-agent-based smart power management for remote health monitoring." Neural Computing and Applications (2021): 1-10.

[4] Mehdi Hosseinzadeh, Jalil Koohpayehzadeh, Ahmed Omar Bali, Parvaneh Asghari, Alireza Souri, Ali Mazaherinezhad, Mahdi Bohlouli, and Reza Rawassizadeh. "A diagnostic prediction model for chronic kidney disease in internet of things platform." Multimedia Tools and Applications 80, no. 11 (2021): 16933-16950.

[5] Ardhendu Sekhar, Soumen Biswas, Ranjay Hazra, Arun Kumar Sunaniya, Amrit Mukherjee, and Lixia Yang. "Brain tumor classification using fine-tuned GoogLeNet features and machine learning algorithms: IoMT enabled CAD system." IEEE Journal of Biomedical and Health Informatics (2021).

[6] Mehdi Hosseinzadeh, Jalil Koohpayehzadeh, Marwan Yassin Ghafour, Aram Mahmood Ahmed, Parvaneh Asghari, Alireza Souri, Hamid Pourasghari, and Aziz Rezapour. "An elderly health monitoring system based on biological and behavioral indicators in internet of things." Journal of Ambient Intelligence and Humanized Computing (2020): 1-11.

[7] Tuli, Shreshth, Shikhar Tuli, Gurleen Wander, Praneet Wander, Sukhpal Singh Gill, Schahram Dustdar, Rizos Sakellariou, and Omer Rana. "Next generation technologies for smart healthcare: Challenges, vision, model, trends and future directions." Internet Technology Letters 3, no. 2 (2020): e145.

[8] Tuli, Shreshth, Nipam Basumatary, Sukhpal Singh Gill, Mohsen Kahani, Rajesh Chand Arya, Gurpreet Singh Wander, and Rajkumar Buyya. "Healthfog: An ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated iot and fog computing environments." Future Generation Computer Systems 104 (2020): 187-200.

[9] Gill, S.S., Arya, R.C., Wander, G.S. and Buyya, R., 2018, August. Fog-based smart healthcare as a big data and cloud service for heart patients using IoT. In International Conference on Intelligent Data Communication Technologies and Internet of Things (pp. 1376-1383). Springer, Cham.

[10] Gagliano, S., Ravji, R., Barnes, M., Weale, M. and Knight, J., 2015. Smoking Gun or Circumstantial Evidence? Comparison of Statistical Learning Methods using Functional Annotations for Prioritizing Risk Variants. *Scientific Reports*, 5(1).

[11] Health Europa. 2019. *Avoidable Deficiencies In Heart Failure Cost NHS £21M*. [online] Available at: <https://www.healtheuropa.eu/avoidable-deficiencies-in-heart-failure-cost-nhs-21m/98696/> [Accessed 6 August 2020].

[12] Archive.ics.uci.edu. 1988. *UCI Machine Learning Repository: Heart Disease Data Set*. [online] Available at: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease> [Accessed 9 August 2020].

[13] Sharma, P., Choudhary, K., Gupta, K., Chawla, R., Gupta, D. and Sharma, A., 2020. Artificial plant optimization algorithm to detect heart rate & presence of heart disease using machine learning. *Artificial Intelligence in Medicine*, 102, p.101752.

[14] Gonsalves, A., Thabtah, F., Mohammad, R. and Singh, G., 2019. Prediction of Coronary Heart Disease using Machine Learning. *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies - ICDLT 2019*,.

[15] Tarawneh, M. and Embarak, O., 2019. Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques. *Advances in Internet, Data and Web Technologies*, pp.447-454.

[16] Satyanandam, N. and Satyanarayana, C., 2019. Heart Disease Detection Using Predictive Optimization Techniques. *International Journal of Image, Graphics and Signal Processing*, 11(9), pp.18-24.

[17] Nikram, S., Shukla, P. and Shah, M., 2020. *Cardiovascular Disease Prediction Using Genetic Algorithm And Neuro-Fuzzy System*. [online] Available at: <https://www.ijltet.org/journal/149094578316.1568.pdf> [Accessed 7 August 2020].

[18] Abdeldjouad, F., Brahami, M. and Matta, N., 2020. A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques. *Lecture Notes in Computer Science*, pp.299-306.

[19] Spencer, R., Thabtah, F., Abdelhamid, N. and Thompson, M., 2020. Exploring feature selection and classification methods for predicting heart disease. *DIGITAL HEALTH*, 6, p.205520762091477.

[20] Khourdifi, Y. and 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), pp.242-252.

[21] Polat, K. and Güneş, S., 2009. A new feature selection method on classification of medical datasets: Kernel F-score feature selection. *Expert Systems with Applications*, 36(7), pp.10367-10373.

[22] Ho, Tin Kam. "Random decision forests." In Proceedings of 3rd international conference on document analysis and recognition, vol. 1, pp. 278-282. IEEE, 1995.

[23] Publishing, H., 2016. *Throughout Life, Heart Attacks Are Twice As Common In Men Than Women - Harvard Health*. [online] Harvard Health. Available at: <https://www.health.harvard.edu/heart-health/throughout-life-heart-attacks-are-twice-as-common-in-men-than-women#:~:text=Researchers%20found%20that%20throughout%20life,mass%20index%2C%20and%20physical%20activity.> [Accessed 10 August 2020].

[24] nhs.uk. 2016. *Study Says There's No Link Between Cholesterol And Heart Disease*. [online] Available at: <https://www.nhs.uk/news/heart-and-lungs/study-says-theres-no-link-between-cholesterol-and-heart-disease/> [Accessed 10 August 2020].

[25] Haq, A., Li, J., Memon, M., Nazir, S. and Sun, R., 2018. A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. *Mobile Information Systems*, 2018, pp.1-21.

[26] Muhammed Golec et al.. "BioSec: A Biometric Authentication Framework for Secure and Private Communication among Edge Devices in IoT and Industry 4.0." IEEE Consumer Electronics Magazine (2020).

[27] Yong Deng, D. Frank Hsu, Zhonghai Wu, and Chao-Hsien Chu. "Combining multiple sensor features for stress detection using combinatorial fusion." Journal of Interconnection Networks 13, no. 03n04 (2012): 1250008.

[28] Alejandro Juan, Richard W. Pazzi, and Azzedine Boukerche. "Using Accuracy-Based Learning Classifier Systems for Adaptable Strategy Generation in Games and Interactive Virtual Simulations." Journal of Interconnection Networks 10, no. 04 (2009): 365-390.

[29] Arjan Durresi, Mimoza Durresi, and Leonard Barolli. "Priority Based Wireless Communications for Health Monitoring on Highways." Journal of Interconnection Networks 9, no. 04 (2008): 337-349.

[30] Uyar, K. and İlhan, A., 2017. Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. Procedia computer science, 120, pp.588-593.

[31] Muhammed Golec at al., iFaaSBus: A Security and Privacy based Lightweight Framework for Serverless Computing using IoT and Machine Learning, IEEE Transactions on Industrial Informatics, 2021