iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** XII **Month of publication:** December 2021

DOI: https://doi.org/10.22214/ijraset.2021.39559

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

An Efficient Artificial Neural Network for Coronary Heart Disease Prediction

Priyam Vinay Sheta[1], Preet Jayendrakumar Modi[2]
[1, 2]*Dharmsinh Desai University, India*

Abstract: ***Coronary heart disease is rapidly increasing over these days also with a significant number of deaths. A large population around the world is suffering from the disease. When surveys were carried out of the death rate and the number of people suffering from the coronary heart disease, it was understood that how important is the diagnosis of this disease at an early stage. The old way for detecting the disease was not found effective. This paper suggests a different method and technology to detect the disease and the proposed method is more effective than the old traditional methods. In this paper, an artificial neural network that predicts the coronary heart disease is used with 14 features as the input. Feature selection, data preprocessing, and removing irrelevant data was done before training the neural network. The backpropagation algorithm was used for making the neural network learn the features. The output of data was basically binary but the neural network was trained to give the output as a probability between 0 and 1. Two algorithms were proposed for this prediction named Logistic Regression and Artificial Neural Network but the later was selected because of the accuracy of 94%. The accuracy of Logistic Regression was 87%.***

I. BACKGROUND AND MOTIVATION

Heart diseases are increasingly continuously from the last decade. Among them the coronary heart disease is the major one. Our heart gets pure blood and oxygen through arteries and veins. Due to the swelling of these veins, heart does not get enough amounts of blood and oxygen and because of that reason the heart stops working eventually leading to heart attack and death. This swelling of arteries and veins is basically known as coronary heart disease.

Actually, there is no system right now that is working on the values given by the user. There can be a system that works on the images of the heart but images cannot be true every time. Hence a system that worked on numbers will be very helpful to the users. The current system is working on images but it has been found that the prediction on images has many false negatives. It can be very daunting as the user feels that they are not having any disease but finally end up getting the disease. And the user has to have a scan of his heart for using this system and that can be costly every time when they want to check the disease. Hence the system we are going to prepare will overcome all the problems. The use of direct numbers will provide greater accuracy to the user.

In this work, the dataset has 14 features and that are used for the prediction of probability of the disease. There were previous works that used different number of hidden layers in their neural network. As per one research that was published, Chowdhury et. al proposed artificial neural network (ANN) and the overall predictive accuracy acquired was 75%, with one hidden layer. Subbalakshmi et. al proposed Naïve Bayes and the increased accuracy achieved was 82.31%. Nahar et. al proposed techniques of computational intelligence for the prediction of heart disease and got an accuracy of 86.77%. The need of the system was required where the accuracy provided is more efficient as the coronary heart disease prediction is increasing rapidly. Hence in this paper, an artificial neural network having 4 hidden layers is used for getting a high accuracy.

An ANN has the ability to recognize the relationship between the data in the dataset. A neural network is a system of neurons that can adapt according to the input and output.

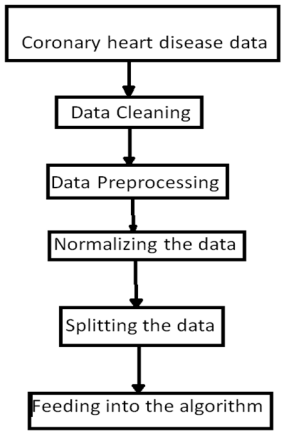
The backpropagation algorithm can set the values of all the neurons so that it can give a correct output. The 14 features used for the prediction are Gender, Age, Smoker or not, Number of cigarette, BP medications, Prevalent Stroke, Prevalent Hypertensive, Diabetes, Cholesterol, Systolic BP, Diastolic BP, BMI, Heart rate and Glucose.

A user friendly website is created where doctors can enter all these details of their patients and can check the probability of the coronary heart disease. The history of all these details of the patients with the date of checkup can also be saved in the system. We will discuss the two algorithms used for the prediction after doing the data preprocessing and data cleaning.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

II. SYSTEM ANALYSIS

Firstly the important libraries required for the data processing and neural network implementation were installed like Numpy, Pandas, Tensorflow, Flask, Pickle, matplotlib, Seaborn, SkLearn, ROC curve, Smote and keras. The data was downloaded from the online platform kaggle and it contained 15 features. The features that are given in the data for the prediction must be a contributing feature towards the heart disease. But, eventually we grasped that out of all the columns, 1 column named “education” was not an important factor leading to a heart disease. All the unnecessary data will be known after the research analysis
but this column was one that was identified by us directly by seeing it.



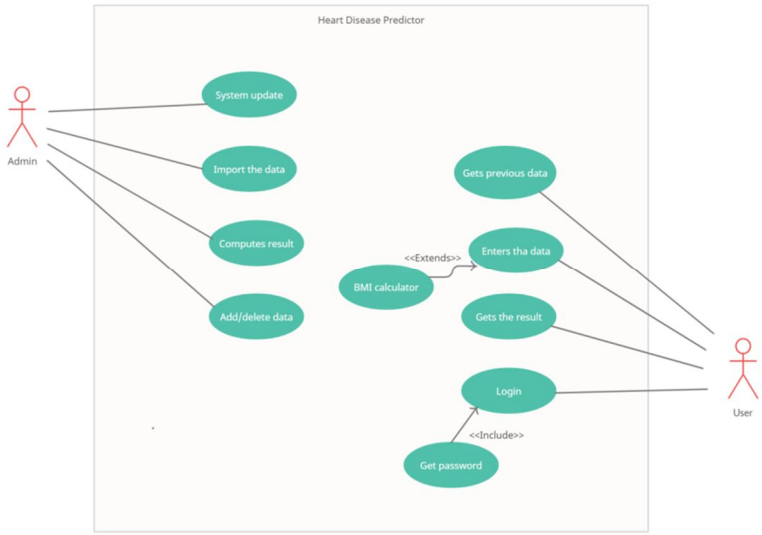
A. Functional Requirements

User can not only create the account on the website by signing up or can also login to the website if account is there but also can check the history of their previous check-up. Also, all the details with the date and time will be displayed in the history section. User can know all the doctors in his city if he is getting a high risk of the heart disease. They can also check the probability of getting the coronary heart disease by entering their latest conditions. In addition, user can know their BMI by entering their weight and height if they are not aware of their current BMI. Moreover, user can freely contact us in case of any problems.

B. Non-Functional Requirements

Users are requiring more and more accurate results and hence the prediction must be done accurately. Similarly, user must be able to use the interface at any place and hence it must be portable. For prediction, always new features will be raised up. Hence the system must be maintained and updated regularly.

C. Use-Case Diagram



III. DATA PREPROCESSING

A. *Removing the NAN data*

The Not a Number (NAN) data cannot be read by the neural network hence we have to process it before feeding it to the algorithm. The processing can be done by 3 ways that is either replacing the NAN value by the value above it, either replace it by the down value or either replace it by the average of all values of that feature. Here the last method is used to remove NAN values. By the help of the SkLearn library as mentioned above we will fill all the NaN values with the average of all the values of that respective feature.

```
In [5]: data.isnull().sum()
Out[5]: male               0
        age                0
        currentSmoker      0
        cigsPerDay        29
        BPMeds            53
        prevalentStroke    0
        prevalentHyp       0
        diabetes           0
        totChol           50
        sysBP              0
        diaBP              0
        BMI               19
        heartRate          1
        glucose          388
        TenYearCHD         0
        dtype: int64
```

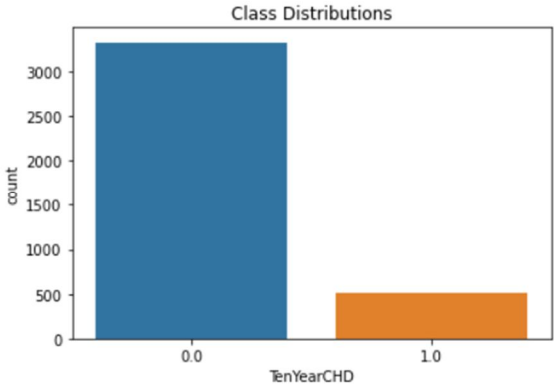
```
In [7]: data_new.isnull().sum()
Out[7]: male               0
        age                0
        currentSmoker      0
        cigsPerDay         0
        BPMeds             0
        prevalentStroke    0
        prevalentHyp       0
        diabetes           0
        totChol            0
        sysBP              0
        diaBP              0
        BMI                0
        heartRate          0
        glucose            0
        TenYearCHD         0
        dtype: int64
```

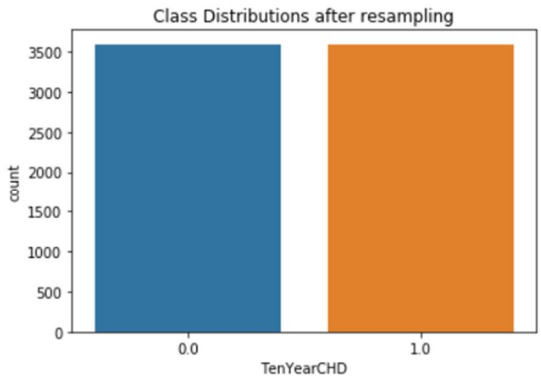
Normalization of data means bringing all the data values in almost same range so the neural network can adjust the weight and bias accordingly. If the range of all values will be different then it will be difficult to get good accuracy.
Value= (Value – minimum of that feature)/(maximum of that feature – minimum of that feature)

IV. DATA NORMALIZATION

A. *Balancing the Distribution*

The problem is detection that whether the user is having a coronary heart disease or not. But if there is a bias data in which the data of user not having disease is more, then it is possible that the final trained model will remain biased towards the data which is more in number. This is called overfitting. This can be seen in the first diagram below. It can be stated that this data is biased. Hence the library named SMOTE as mentioned above is used. SMOTE duplicates the data with the less number and makes it equal to the higher number data. Hence the problem of overfitting can be prevented by using SMOTE. The second figure is the new distribution after using SMOTE.





V. SHUFFLING THE DATA

After the data was equally distributed, it was found that first half of the data contained a large number of negative labels. And the second half of the data contained a large number of positive labels. For a machine learning model to be trained accurately, it is necessary that we do not provide it with same data continuously. Hence shuffling of the data becomes very important here. After shuffling the data has become uniform and the problem of overfitting is avoided.

VI. DIVIDING THE DATA

Usually the Machine Learning model gives a good accuracy on the data it is being trained on. The real accuracy is measured by giving it a new data that is unseen by the model. Hence the dataset is divided in unequal 2 parts so that one of the part is used for training the model and the other part is used to test the accuracy on the unseen data. This can be done by a module named **"train-test-split".** It is used to divide the data into 2 parts.

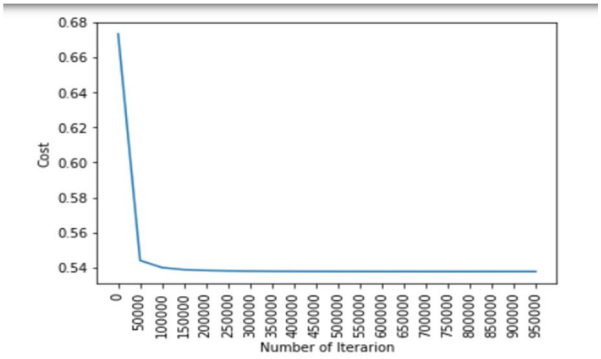
```
X_train,X_test,Y_train,Y_test=train_test_split(X,y,test_size=0.2)
```

We have kept 80% of the data into training part and 20% of the part into testing phase. So final data is:
Training data : 5750 x 15 (14 features + 1 output)
Testing data : 1438 x 15
Hence after all these techniques, the final data is ready and we can now feed it into the algorithm.

VII. ALGORITHMS

A. Logistic Regression

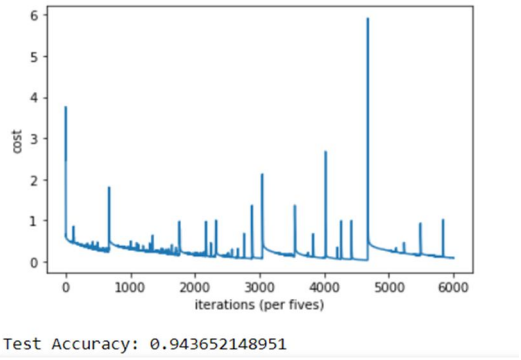
The algorithm is specially used for the classification problems in machine learning. The algorithm was implemented and the accuracy of the prediction on the test data was observed.



The accuracy of the Logistic Regression came out to be near to 88% which was not bad but the same accuracy was achieved by the different researchers mentioned above in the related works section. Hence neural network was implemented to check whether the accuracy is good enough.

B. Artificial Neural Network

ANN algorithm was built to feed the data of 14 features. It had total 6 layers: one input layer of 14 nodes, one output layer of 2 nodes (0 and 1 classification) and 4 hidden layers with 64 nodes each. Different number of nodes was tried and the accuracy was best at 64 nodes. The accuracy observed here was not even higher than the Logistic Regression algorithm but also higher than the previous works mentioned in the related works section.



The accuracy achieved by the artificial neural network around 94% is around and is greater compared to other works and algorithms. Hence the next parts of the system were carried out in consideration to the artificial neural network.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

VIII. THRESHOLD VALUE

Although the neural network is used for the classification problem (0 and 1 problems) but at the output layer there was a sigmoid function used so we can get the probability of getting the heart disease. But what probability can be considered as a low risk and what probability can be considered as a high risk should be determined. Hence we need a value that can be considered as the determining value i.e. under that value, the probability would be considered as **Low** and above that value, it would be considered as **High.** That value is called the Threshold value. Decision tree is used for calculating the threshold value. It is a supervised learning algorithm which can be used on both regression and classification. . A module named ROC_Curve is used to get the "True Positives" and "False Positives" by which we will get a good threshold value. We have here used 4 models to get a very accurate threshold value. AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. These are the 4 models with their respective ROC Scores.

1) Random Forest Classifier

```
RF train roc-auc: 1.0
RF test roc-auc: 0.9955430782895571
```

2) Logistic Regression

```
Logistic train roc-auc: 0.7227370811991269
Logistic test roc-auc: 0.721235068066054
```

3) AdaBoost Classifier

```
Adaboost train roc-auc: 0.8467610296104937
Adaboost test roc-auc: 0.8291052622038537
```

4) KNeighbours classifier

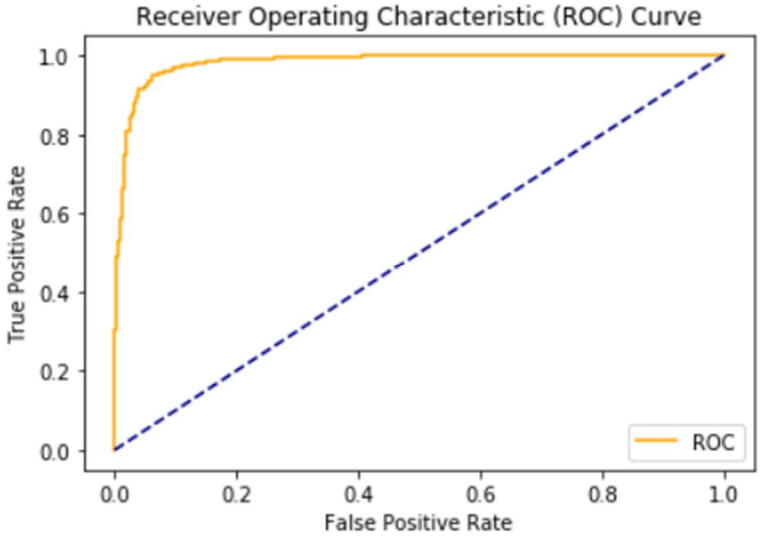
```
train roc-auc: 0.9898356402285621
test roc-auc: 0.9472408776986243
```

Diagnostic Test Result	Disease Status		Total
	Positive	Negative	
Positive	True Positive (TP)	False Positive (FP)	All tests positive (T+)
Negative	False Negative (FN)	True Negative (TN)	All tests negative (T-)
Total	Total with disease (D+)	Total without disease (D-)	Total Sample Size

By combining all these 4 models we will get the values at which the accuracy is highest. By the table given below, we will get the threshold value.

	thresholds	accuracy
204	0.641804	0.947844
230	0.571984	0.947844
225	0.580975	0.947149
226	0.578468	0.947149
229	0.572965	0.947149
...	...	...
4	0.863463	0.317107
3	0.864066	0.316412
2	0.866186	0.309458
1	0.869938	0.308762
0	1.869938	0.308762

Here there are 2 threshold values at the same accuracy and hence we will select the one that is near to 0.5. Hence the threshold value of the proposed system is 0.57. So it can be said that the patients having the probability of coronary heart disease more than 0.57 are at a high risk whereas the patients having the probability less than 0.57 are having a low risk.



IX. MODEL ANALYSIS

During the worktime of this project, we have analyzed many things. We observed the features used in the project and analyzed all that features. We can finally say that all attributes selected after the elimination process show P-values lower than 5% and thereby suggesting significant role in the Heart disease prediction. Men seem to be more susceptible to heart disease than women. Increase in age, number of cigarettes smoked per day and systolic Blood Pressure also show increasing odds of having heart disease. Total cholesterol shows no significant change in the odds of CHD. This could be due to the presence of good cholesterol (HDL) in the total cholesterol reading. Glucose too causes a very negligible change in odds (0.2%). The factors like Blood Pressure, BMI, and Glucose are definitely important in terms of heart disease but in terms of coronary heart disease, they contribute very little. Along with that, observation of the artificial neural network was also done. It was observed that when the number of hidden layers were increased or decreased than 4 then the accuracy used to fall by a good amount. The similar case was seen when the number of nodes were changed from 64. All the previous researchers had used different technologies and techniques for carrying out the system but out of all papers that were studies by us, none had the accuracy greater than 92%. The preprocessing and cleaning of the data played an important role in acquiring the accuracy.

X. TEST CASES

Functional Test Cases	Expected Output Positive / Negative
Verifying that the account is created and updated in the database successfully.	Positive
Verifying the working of the history section.	Positive
Verifying predicted output is more than 85%.	Positive
Verifying the working of the BMI calculator.	Positive
Verifying the working of all the alert boxes.	Positive
Verifying the working of the suggestions of the doctors.	Positive

XI. USER MANUAL

A. *Home Page*

My History Contact us LogIn SignUp

How to know if a heart is healthy?

A healthy heart is the key to good life .The heart is a vital organ of the human body which ensures the effective pumping of blood throughout the circulatory system. Due to our sedentary lives and food habits, the heart is prone to malfunctioning, and heart attack (i.e. coronary artery disease), is one of the primary cause of death [1]. Heart attack is caused by a blockage of the coronary arteries, typically at a site of narrowing (stenosis) caused by atherosclerosis. It is difficult to accurately determine the degree of atherosclerosis in arteries, particularly in the early stages of disease. One method that has been introduced is the intravascular ultrasonic catheter (IVUS), which sends a pulse of sound from a receiver and uses the returned echo to deduce the properties of the arterial tissues.

Heart disease is the leading cause of death for both men and women. Take steps today to lower your risk of heart disease.

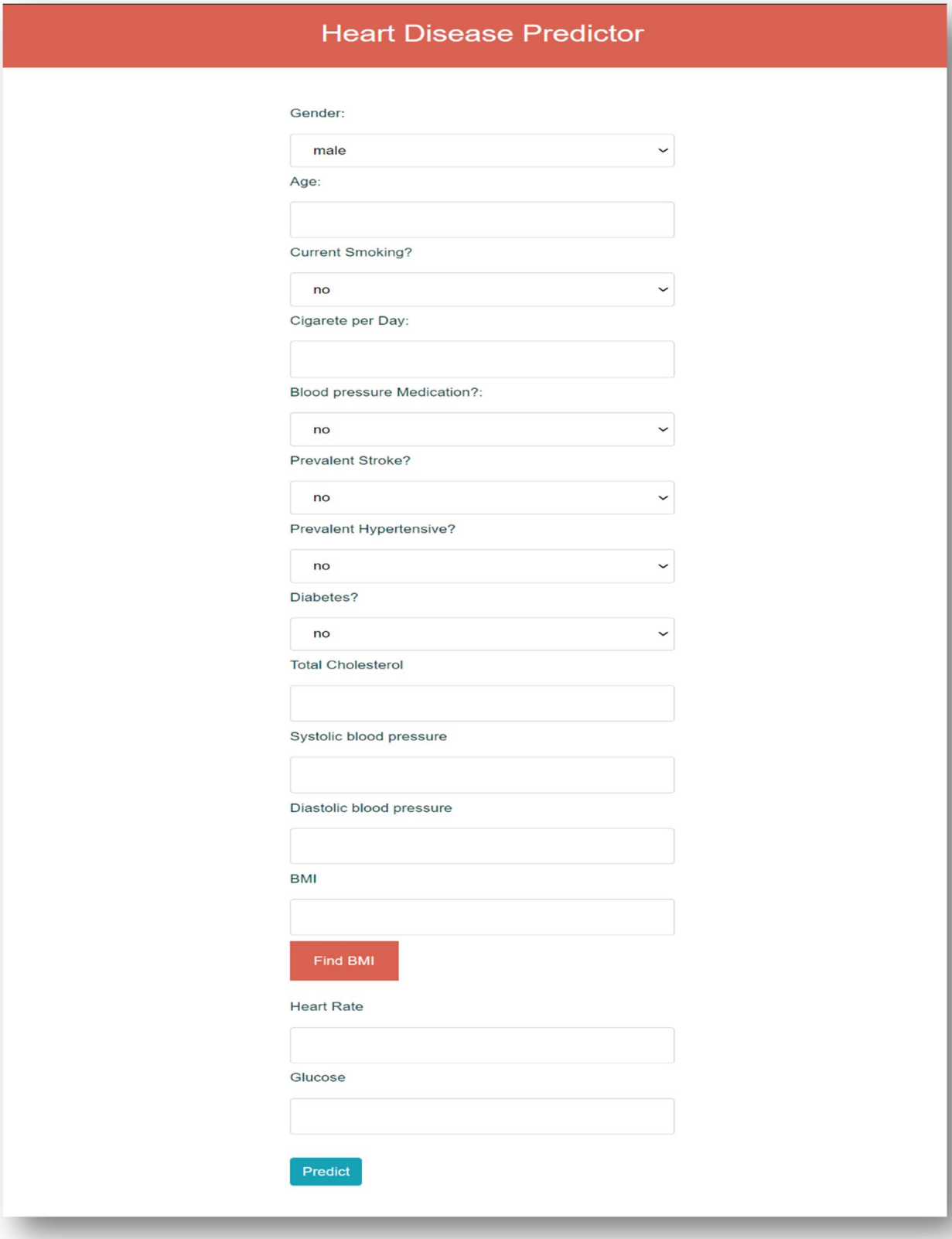
To help prevent heart disease, you can:

- Eat healthy.
- Get active.
- Stay at a healthy weight.
- Quit smoking and stay away from secondhand smoke.
- Control your cholesterol and blood pressure.
- Drink alcohol only in moderation.
- Manage stress.

Calculate the heart risk today!

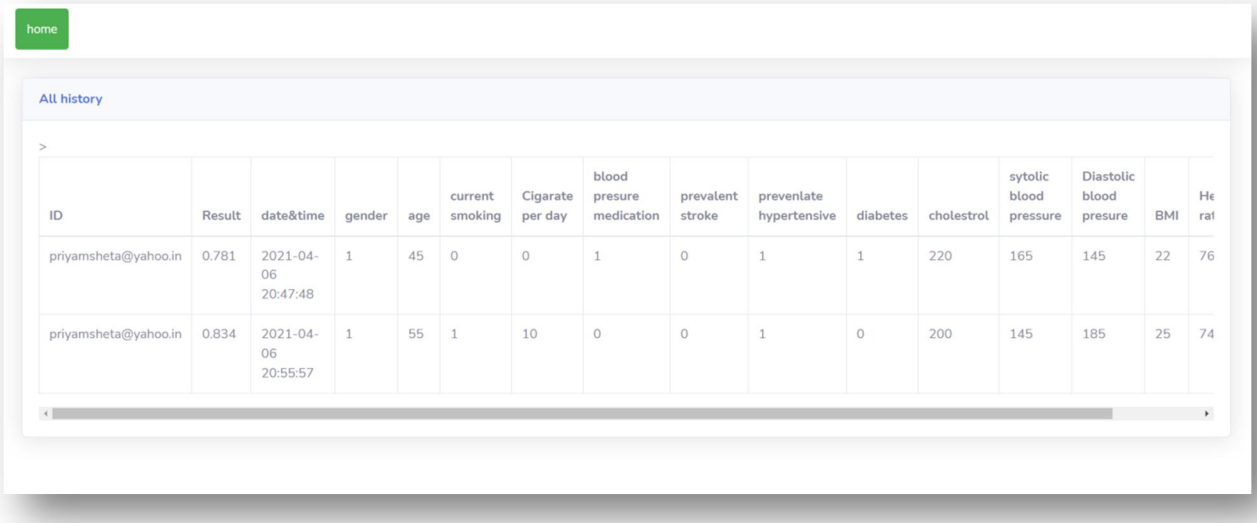
Click here!

B. Prediction Page



C. Predicting the Values

D. History Section



XII. CONCLUSION

The user can use the website to enter the details required and that values will be feed as an input to the neural network for predicting the probability of the disease and the amount of risk will also be displayed. There was a need for a system that can detect the coronary heart disease very accurately as the disease is increasing rapidly since last decade. There were old traditional systems that were working well but had many false negatives and that lead to more number of deaths. Hence it was necessary to improve the system. Many researchers have proposed system with accuracy of 85%, 87% and also one had an accuracy of 92%. But the proposed system mentioned in this paper had an accuracy of 94% that is obtained by implementing a neural network that is having 4 hidden layers. The features that are contributing more to the disease are also observed and analyzed after the system was implemented.

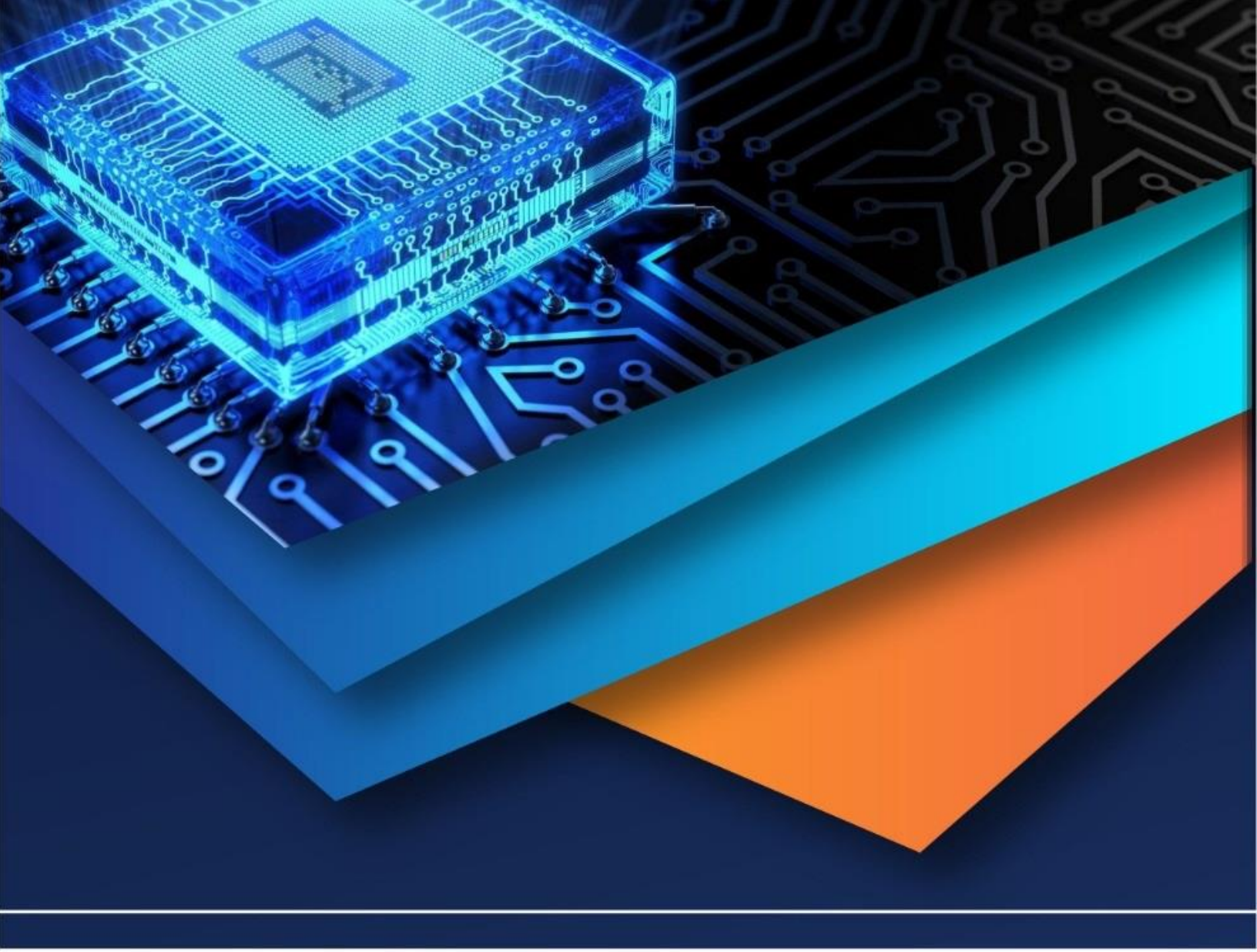
REFERENCES

Conference Papers
[1] https://www.researchgate.net/publication/326733163_Prediction_of_Heart_Disease_Using_Machine_Learning_Algorithms
[2] Book: Baldi, P. and Brunak, S. (2002). Bioinformatics: A Machine Learning Approach. Cambridge, MA: MIT Press.
[3] W. H. Organization and others, “World health statistics 2017: monitoring health for the SDGs, sustainable development goals,” 2017.
[4] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, “Computational intelligence for heart disease diagnosis: A medical knowledge driven approach,” Expert Syst. Appl., vol. 40
[5] https://comengapp.unsri.ac.id/index.php/comengapp/article/view/288
[6] https://www.sciencedirect.com/science/article/abs/pii/S0957417420302323
[7] F. Amato, A. López, E. M. Peña-Méndez, P. Va\vnhara, A. Hampl, and J. Havel, “Artificial neural networks in medical diagnosis.” Elsevier, 2013.

[8] N. A. Sundar, P. P. Latha, and M. R. Chandra, "Performance analysis of classification data mining techniques over heart disease database," IJESAT] Int. J. Eng. Sci. Adv. Technol. ISSN, pp. 2250–3676, 2012.
[9] A. Shinde, "Heart Disease Prediction System using Multilayered Feed Forward Neural Network and Back Propagation Neural Network," 2017.
[10] S. Nurmaini, A. Gani, and others, "Cardiac Arrhythmias Classification Using Deep Neural Networks and Principle Component Analysis Algorithm.," Int. J. Adv. Soft Comput. Its Appl., vol. 10.
[11] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," Inf. Sci. (Ny)., vol. 340.
[12] J.-S. Wang, W.-C. Chiang, Y.-L. Hsu, and Y.-T. C. Yang, "ECG arrhythmia classification using a probabilistic neural network with a feature reduction method," Neurocomputing, vol. 116.
[13] Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network by S. Radhimeenakshi.
[14] Jae Kwon Kim and Sanggil Kumar , "Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis", Journal of healthcare engineering.

Book name: - Hands-On Machine Learning-Aurelian Gerona

Websites
[1] researchgate.com
[2] https://keras.io/
[3] towardsdatascience.com
[4] https://www.w3schools.com/html/
[5] https://www.w3schools.com/html/html_styles.asp
[6] https://www.w3schools.com/html/html_css.asp



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY