DATA ANALYSIS: IDENTIFYING THE MOST INFLUENTIAL ATTRIBUTES OF HEART DISEASES THAT CAUSES DEATHS

Sk. Husne Tanveer[1], S.Sai Siva Teja[2], V. Arunkanth[3], V.L.N. Aditya[4], Madda. Varalakshmi[5]

[1,2,3,4] Student , Vasireddy Venkatadri Institute of Technology, Nambur, Guntur(Dt), AP, India

[5] Assistant Professor , Vasireddy Venkatadri Institute of Technology, Nambur, Guntur(Dt), AP, India

1. ABSTRACT **2.**

Heart diseases are considered as one of the major causes of death in recent years. They cannot be easily predicted by the medical practitioners as it is a difficult task which requires higher knowledge for identification.. This application identifies the major attributes that cause heart diseases and predicts the occurrence of heart disease. The goal is to get conclusions by applying machine learning techniques on the dataset. The prediction of heart disease requires a large sized data which is difficult to analyse by normal methods. Our research is based on selecting the suitable machine learning technique that can identify responsible attributes that cause heart diseases and predicting heart disease.

keywords : Heart Diseases, classification , K-Nearest Neighbors, Prediction, attribute detection.

INTRODUCTION

Human heart is the most complicated part which is next to the brain. It pumps the blood and supplies to all organs of the whole body. Heart disease is the main reason behind the deaths in various countries of the world and even in India also. In case of USA one person is dying every 36 seconds just because of some kind of heart disease .There are many types of heart disease such as coronary heart disease, cardiovascular disease and cardiomyopathy disease . Coronary disease is which oxygen and blood is not properly is not supplied to the heart because of reduction in the size of coronary arteries .Cardiovascular disease leads to various illnesses in the body like high BP, coronary artery disease, stroke and finally leads to death. Cardiomyopathy is a disease of the heart muscle. The heart muscles become enlarged, thick or rigid when cardiomyopathy affects which makes it harder for the heart to pump blood to the rest of your body. Cardiomyopathy can lead to heart failure[1]. The types of cardiomyopathy include hypertrophic, dilated, restrictive, arrhythmogenic right ventricular dysplasia, unclassified cardiomyopathy.

The purpose of this research is to predict the heart diseases and identify the major factors that cause the heart diseases. If we can find the attributes that majorly influence the heart diseases then we can control and take the necessary precautionary measures to control them. We use machine learning techniques to identify the attributes and for predicting the occurrence of heart diseases.

3. LITERATURE SURVEY

There are many numerous works that have been done related to detection of heart disease based on given data. There are many authors with an accurate diagnosis in medical centers.

The performance of accuracy calculated by K-Nearest Neighbor and Decision Tree algorithms is 98.17% and 98.99% are obtained respectively[1]. The algorithms used in the paper give the better result of the performance of accuracy to characterize the coronary artery disease as per the research done by The Vidya K. Sudarshan. He focused on the Application of higher-order spectra for the characterization of coronary artery disease using electrocardiogram signals.

As per the research analysis of Rajendra Acharya , computer-aided diagnosis of the diabetic subject by heart rate variability signals using discrete wavelet transform method using different classifiers that include Decision Tree (DT), K-Nearest Neighbor (KNN), Naïve Bayes (NB), and Support Vector Machine (SVM). The average accuracy obtained is 92.02% by using DT within ten-fold cross-validation[2]. The computed accuracy is important for prediction, however, it is not enough as needed.

The survey done by Muhammad Saqlainet shows that the identification of Heart Failure by using unstructured Data of Cardiac Patients using Logistic Regression, Neural Network, SVM, Random Forest Decision Tree and Naïve Bayes. The algorithms achieve accuracy of 80%, 84.8%, 83.8%, 86.6%, 86.6% and 87.7% respectively for each individual's algorithm. Naïve Bayes provide the highest accuracy compared to others algorithms[3].

The description of a survey paper on heart disease prediction by M Gandhi , describes the different methodology and the way in which proposed methods are implemented. It also provides some overview of heart disease[4], as well as the role of data mining in healthcare centers and how to apply or use data mining in a healthcare organization, is explained.

Mr. R.Rao propose application of knowledge discovering process on prediction of stroke patients based on Artificial Neural Network (ANN) and Support Vector Machine (SVM), which give accuracy of 81.82% and 80.38% for ANN and SVM respectively for training data set and 85.9% and 84.26% for Artificial Neural Network (ANN) and Support Vector Machine (SVM) in test dataset respectively[5] . ANN shows more accurate results than Support Vector Machine (SVM) for proposed work. The accuracy obtained by the paper is not enough in prediction of the stroke patients.

From the above reference works the most effective models to predict Heart Diseases appears to be Naïve Bayes, KNN, Logistic regression, Decision Trees[7] and Neural Network.

4. DATA SET

This dataset consists of 4000 patient records with attributes like Gender, Age, Current Smoker, Cigarettes, Blood Pressure, Prevalent Stroke, Hypertension, Diabetes, Total Cholesterol, Systolic Blood Pressure, Diastolic Blood Pressure, Body Mass Index, Heart Rate, Glucose and Heart Disease. We use modules such as SKLEARN, PANDAS, NUMPY etc to pre-process and handle the data. Sometimes the data may contain missing values which must be handled for the algorithm to work efficiently. After pre-processing is done, we apply classification methods on the data.

5. PROPOSED ALGORITHM

5.1. K-Nearest-Neighbors

The KNN algorithm assumes that similar things will be nearer to one another. In other words, similar things are in close proximity[2][6].

5.1.1 The KNN Algorithm

- The data should be loaded
- Choose number of neighbors and initialize it to K.
- For each example in the data
  - From the data, calculate the distance between the current and the query examples.
  - Add the index and the distance of the example to an ordered collection
- Sort the ordered collection of indices and distances in ascending order by the distance values.
- Get the first K entries from the sorted collection
- From the selected K entries, get the labels
- Return the mean of the K labels, if it is regression
- Return the mode of the K labels, if it is classification

5.1.2 Applying KNN algorithm on our data set

- First we set the number of neighbors to 2
- Next we obtain the test data and train data
  - We train the model with the training data and then we perform testing[8]
  - The algorithm predicts the outcome for test data and returns the result
  - We check the f1 score accuracy, precision and recall for this algorithm.

5.1.3 Advantages

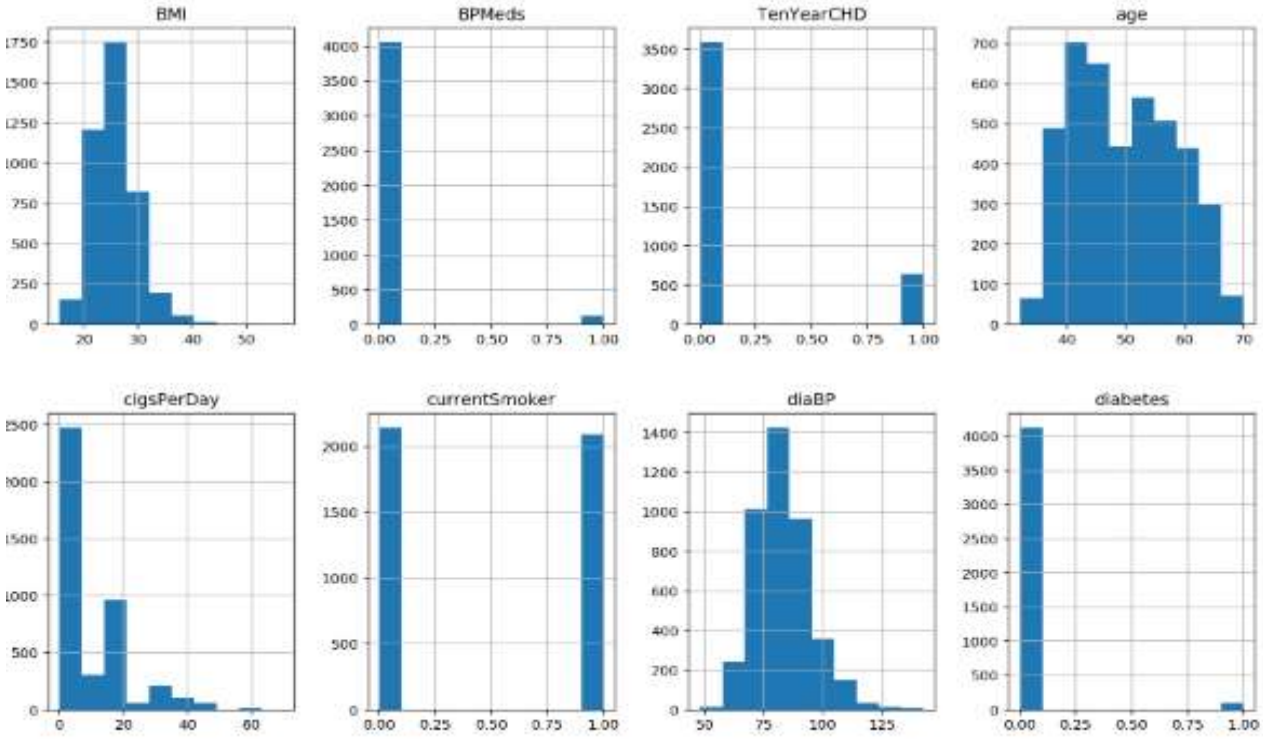
- This algorithm is very simple and easy to implement.
- There is no need of building a model, tuning several parameters, or making additional assumptions.
- The algorithm is versatile, as it can be used for both classification, regression.
- This algorithm is more effective for large training data.

5.1.4 Disadvantages

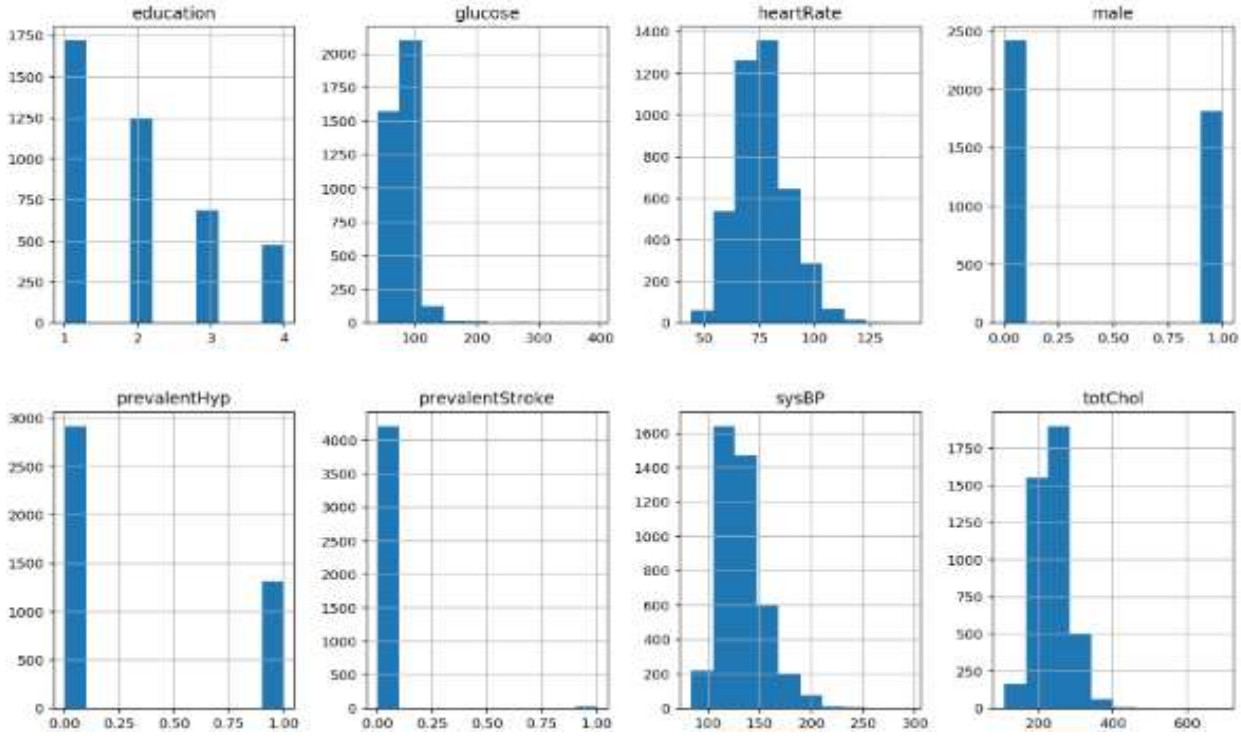
- Always needs to determine the value of K which may be complex some time.
- Because of calculating the distance between the data points for all the training samples, the computation cost is high .

6. HANDLING MISSING VALUES

First, we load the dataset and handle the missing values by replacing them with the median of all the values[10]. Visualizing the data from dataset after handling missing values



The Data on Y axis represents the count of patients and the data on X axis represents the value of the attribute



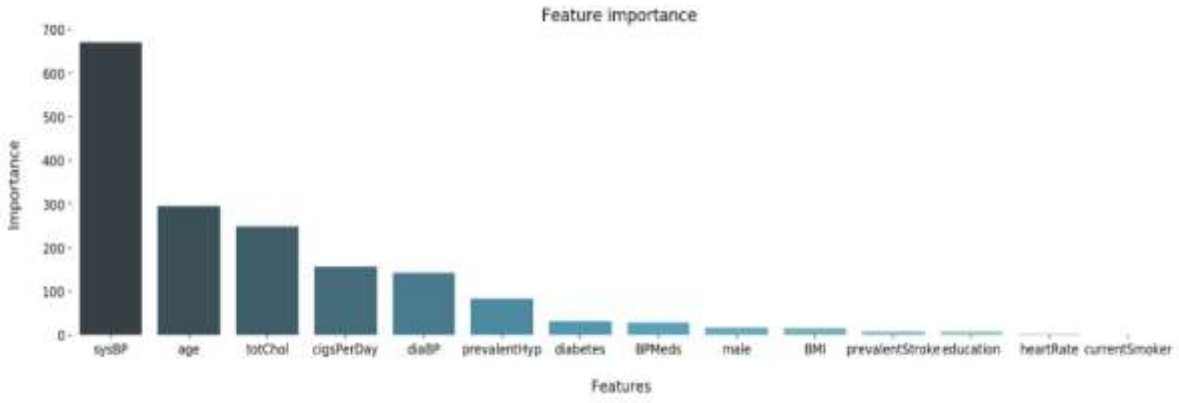
7. IDENTIFYING THE MOST INFLUENTIAL ATTRIBUTES

For this research we have used a classification algorithm called K-Nearest Neighbors. Here we identify the major attributes that are causing heart diseases and apply the algorithm on that attributes[3]

```
[ ]  # Identify the features with the most importance for the outcome variable Heart Disease

     from sklearn.feature_selection import SelectKBest
     from sklearn.feature_selection import chi2

     # separate independent & dependent variables
     X = df.iloc[:,0:14]  #independent columns
     y = df.iloc[:,-1]    #target column i.e price range

     # apply SelectKBest class to extract top 10 best features
     bestfeatures = SelectKBest(score_func=chi2, k=10)
     fit = bestfeatures.fit(X,y)
     dfscores = pd.DataFrame(fit.scores_)
     dfcolumns = pd.DataFrame(X.columns)

     #concat two dataframes for better visualization
     featureScores = pd.concat([dfcolumns,dfscores],axis=1)
     featureScores.columns = ['Specs','Score']  #naming the dataframe columns
     print(featureScores.nlargest(12,'Score'))  #print 10 best features
```

	Specs	Score
10	sysBP	669.506552
1	age	295.507761
9	totChol	249.153078
4	cigsPerDay	156.567318
11	diaBP	142.878574
7	prevalentHyp	82.967184
8	diabetes	31.027987
5	BPMeds	28.153003
0	male	17.120430
12	BMI	15.730717
6	prevalentStroke	8.497823
2	education	7.679797

The result of this KNN algorithm specifies that the attributes like systolic blood pressure, age, cholesterol, cigarettes per day and diastolic blood pressure are the most influential factors that causes the heart deaths.



This graph represents the feature score of each attribute the higher the feature score the more the attribute is tend to cause heart disease[5][6]. In order to apply the algorithm, we split the data as 'Training Data' and 'Test Data'. Training data contains 20% of data from the dataset and the remaining 80% is test data.

8. PERFORMANCE ANALYSIS OF KNN AND LOGISTIC REGRESSION

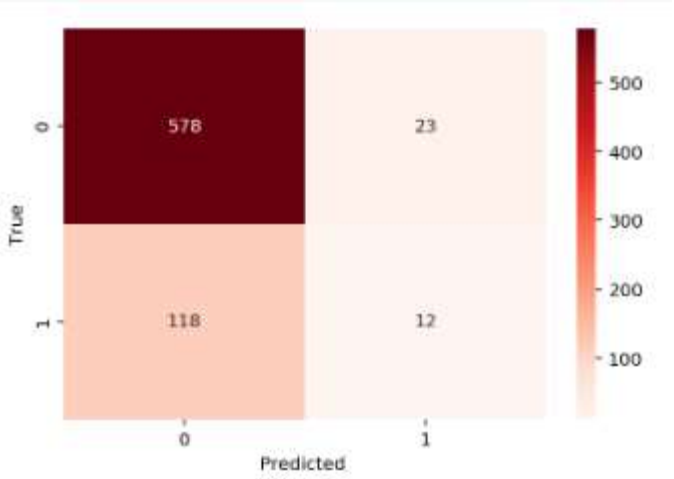
8.1 Using KNN:

After applying the KNN algorithm, the results obtained are as follows

```
[ ] # Check overfit of the KNN model
    # accuracy test and train
    acc_test = knn.score(X_test, y_test)
    print("The accuracy score of the test data is: ",acc_test*100,"%")
    acc_train = knn.score(X_train, y_train)
    print("The accuracy score of the training data is: ",round(acc_train*100,2),"%")

    The accuracy score of the test data is:  80.71135430916553 %
    The accuracy score of the training data is:  88.92 %
```

The Confusion Matrix obtained for KNN is as follows:

```
[ ] # plotting confusion matrix KNN

    cnf_matrix_knn = confusion_matrix(y_test, normalized_df_knn_pred)

    ax= plt.subplot()
    sns.heatmap(pd.DataFrame(cnf_matrix_knn), annot=True,cmap="Reds" , fmt='g')

    ax.set_xlabel('Predicted ');ax.set_ylabel('True');
```



Analysing the confusion matrix of KNN:

- The first box shows that 578 records were observed to be negative and they are actually negative(True Negative)
- The second box in the first row predicts that for 23 patients observation is negative but the actual result is positive(False Positive)
- The first box in the second row shows that for 118 records, the observation is positive but the prediction is negative(False Negative)
- The last box shows that for 12 records the observation is positive and the predicted result is also positive(True positive)
- This confusion matrix plays an important role in determining the performance measure of an algorithm and analysing the results.

8.2 Using Logistic Regression:

After applying the logistic regression algorithm, the results obtained are as follows

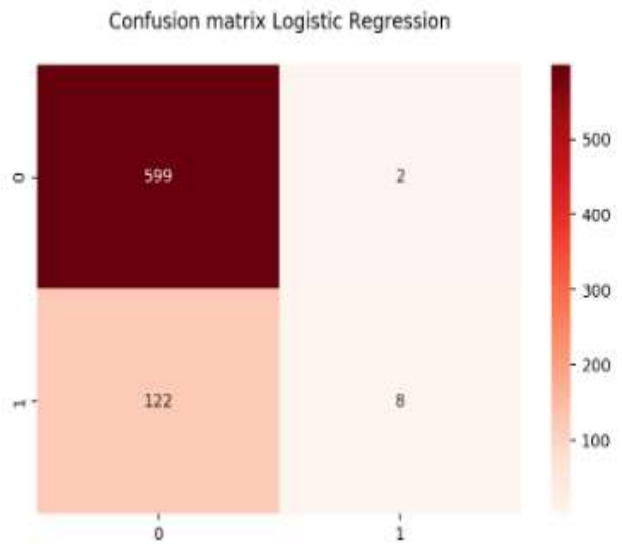
Applying logistic regression

```
[ ] # logistic regression again with the balanced dataset

    normalized_df_reg = LogisticRegression().fit(X_train, y_train)

    normalized_df_reg_pred = normalized_df_reg.predict(X_test)

    # check accuracy: Accuracy: Overall, how often is the classifier co
    acc = accuracy_score(y_test, normalized_df_reg_pred)
    print(f"The accuracy score for LogReg is: {round(acc,3)}")

    # f1 score: The F1 score can be interpreted as a weighted average o
    f1 = f1_score(y_test, normalized_df_reg_pred)
    print(f"The f1 score for LogReg is: {round(f1,3)}")

    # Precision score: When it predicts yes, how often is it correct? P
    precision = precision_score(y_test, normalized_df_reg_pred)
    print(f"The precision score for LogReg is: {round(precision,3)}")

    # recall score: True Positive Rate(Sensitivity or Recall): When it'
    recall = recall_score(y_test, normalized_df_reg_pred)
    print(f"The recall score for LogReg is: {round(recall,3)}")
```

```
The accuracy score for LogReg is: 0.83
The f1 score for LogReg is: 0.114
The precision score for LogReg is: 0.8
The recall score for LogReg is: 0.062
```

The confusion matrix obtained for Logistic regression is as follows:

```
[ ] # plotting confusion matrix for logistic regression

    cnf_matrix_log = confusion_matrix(y_test, normalized_df_reg_pred)

    sns.heatmap(pd.DataFrame(cnf_matrix_log), annot=True,cmap="Reds" , fmt='g')
    ax.xaxis.set_label_position("top")
    plt.tight_layout()
    plt.title('Confusion matrix Logistic Regression\n', y=1.1)
```



Analysing the confusion matrix for Logistic Regression:

- Analysing the confusion matrix for logistic regression:
- The first box shows that 599 records were observed to be negative and they are actually negative(True Negative)
- The second box in the first row predicts that for 2 patients observation is negative but the actual result is positive(False Positive)
- The first box in the second row shows that for 122 records, the observation is positive but the prediction is negative(False Negative)
- The last box shows that for 8 records the observation is positive and the predicted result is also positive(True positive)

The accuracies obtained for both the algorithms on the dataset are as follows:

Algorithm	**Accuracy**
KNN	88%
Logistic regression	83%

9. CONCLUSION

We have worked with KNN and logistic regression out of which KNN was found to be more accurate and precise on the data we choose. It provided more accurate results compared to Logistic regression. From the above study it is clear that the attributes like systolic blood pressure(sysBP), age, cholesterol (totChol), cigarettes per day (cigsPerDay) and diastolic Blood Pressure(diaBP) are mostly influencing the heart diseases. If we can keep these in control, the rate of the heart deaths can be controlled as much as possible.

10. REFERENCES

[1] U. Rajendra Acharya, K. S. Vidya, D. N. Ghista, W. J. E. Lim, F. Molinari, and M. Sankaranarayanan, "Computer-aided diagnosis of diabetic subjects by heart rate variability signals using discrete wavelet transform method," Knowledge-Based Syst., vol. 81, pp. 56–64, 2015.

[2] U. R. Acharya et al., "Application of higher-order spectra for the characterization of Coronary artery disease using electrocardiogram signals," Biomed. Signal Process. Control, vol. 31, pp. 31–43, 2017.

[3] M. Saqlain, W. Hussain, N. A. Saqib, and M. A. Khan, "Identification of Heart Failure by Using Unstructured Data of Cardiac Patients," 2016 45th Int. Conf. Parallel Process. Work., pp. 426–431, 2016.

[4] M. Gandhi, "Predictions in Heart Disease Using Techniques of Data Mining," Int. Conf. Futur. trend Comput. Anal. Knowl. Manag., 2015

[5] R. Rao, "Survey on Prediction of Heart Morbidity Using Data Mining Techniques," Knowl. Manag., vol. 1, no. 3, pp. 14–34, 2011.

[6] T. Karthikeyan, B. Raghavan, and V. A. Kanimozhi, "A Study on Data mining Classification Algorithms in Heart Disease Prediction," Int. J. Adv. Res. Comput. Eng. Technol., vol. 5, no. 4, pp. 1076–1081, 2016.

[7] M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," 2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. iCEEiCT 2016, 2017.

[8] A. DavariDolatabadi, S. E. Z. Khadem, and B. M. Asl, "Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM," Comput. Methods Programs Biomed., vol. 138, pp. 117–126, 2017.

[9] C. Colak, E. Karaman, and M. G. Turtay, "Application of knowledge discovery process on the prediction of stroke," Comput. Methods Programs Biomed., vol. 119, no. 3, pp. 181–185, 2015.

[10] S. Kiruthika Devi, S. Krishnapriya, and D. Kalita, "Prediction of heart disease using data mining techniques," Indian J. Sci. Technol., vol. 9, no. 39, pp. 21–24, 2016.