

# Prediction of Heart Disease Mortality Rate Using Data Mining

Prasenjit Das<sup>a</sup>, Shaily Jain<sup>b</sup>, Chetan Sharma<sup>c</sup>, Shankar Shambhu<sup>a</sup>, Sakshi<sup>d</sup>

<sup>a</sup>Chitkara University School of Computer Applications, Chitkara University, Himachal Pradesh, India

<sup>b</sup>Chitkara University Institute of Engineering and Technology, Chitkara University, Himachal Pradesh, India

<sup>c</sup>Chitkara University Himachal Pradesh, India

<sup>d</sup>Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

## Abstract

Heart disease is the most acute disease with the highest mortality rate in the world. Prediction and timely treatment of this deadly disease only can reduce its effectiveness. Our paper aims to predict heart disease death using different data mining algorithms with utmost accuracy. In this context, we have used five data mining algorithms, Naive Bayes, LBLInear, Naive tree, Regression and Bayesian network on weka implementing on a dataset from UCI repository. According to the results obtained after execution, all data mining algorithms are predictive with good accuracy. We have evaluated accuracy, f-measure, recall, and precision to compare different data mining algorithms in consideration. However, the Bayes network outperforms all with a maximum accuracy of 79.26%. The values of other parameters are also highest in the Bayes network compared to the other four algorithms.

**Keywords:** Classification, Prediction, Algorithms, Heart Disease, Data Mining, WEKA

## 1. INTRODUCTION

Data Mining is a branch of computer science that is widely used in many fields. Data mining means that mining or digging out knowledge or useful information from a vast amount of data. Through data mining, we can explore small to large datasets to dig out any useful data previously hidden or unknown and detect relationships between different parameters that were not possible with statistical methods. In the health care industry, by applying data mining techniques, we can diagnose and predict the occurrence of disease

and the probability of death. Early prediction and diagnosis of the disease can further decrease the death rate. Cardiovascular disease is the most commonly occurring disease, leading to maximum deaths around the globe [1]. According to WHO, more than 19 million people died from cardiovascular diseases in 2018, and around 4 million of these deaths are of non-senior citizens.

Large amounts of data are available with our health care industry which can be mined to determine hidden information about diseases and be used for effective decision making beforehand [2]. Many researchers have already been motivated by the increasing mortality rate of cardiovascular diseases and started working in the direction of extracting useful information using various data mining techniques [3]. Hence, if we can design a prediction system for different diseases like heart using machine learning or deep learning methods, medical professionals can forego symptoms or problems related to the heart based on the available data about patients and various attributes that contribute to the occurrence of heart disease. One major

---

ACI'21: Workshop on Advances in Computational Intelligence  
at ISIC 2021, February 25-27, 2021, Delhi, India  
EMAIL: prasenjit.das@chitkarauniversity.edu.in (P. Das);  
shaily.jain@chitkarauniversity.edu.in (S. Jain);  
chetan.sharma@chitkarauniversity.edu.in (C. Sharma);  
shankar.shambhu@chitkarauniversity.edu.in (S. Shambhu);  
sakshi@chitkara.edu.in (Sakshi)  
ORCID: 0000-0002-7988-2418 (P. Das); 0000-0001-6078-3607  
(S. Jain); 0000-0001-5401-8503 (C. Sharma);  
0000-0002-2348-1041 (S. Shambhu); 0000-0002-8757-4001  
(Sakshi)

© 2021 Copyright for this paper by its authors. Use  
permitted under Creative Commons License Attribution  
4.0 International (CC BY 4.0).



CEUR Workshop Proceedings  
(CEUR-WS.org)

challenge in assisting doctors in diagnosing the world's most deadly disease needs utmost accuracy[4]. Hence, most of the research is aiming to improve diagnosis accuracy.

This paper used different classification algorithms to evaluate and compare some parameters like accuracy in predicting death rate, f-measure, precision, and recall. Section 2 of this article is about the related work in this domain, and the proposed methodology is discussed in section 3. Section 4 tabulates our experimental setup along with our results and a discussion on them. Finally, the paper is concluded in section 5.

## 2. Background and Motivation

Intensive research has been going on for the past few decades to predict heart disease using data mining techniques. Various data mining algorithms like Naive Bayes, Decision Tree, Neural Network, Support Vector, Logistic Regression, Machine(SVM), k-Nearest Neighbour, Artificial neural network, Random Forest, J48 have already been used by researching in determining different levels of accuracy on multiple datasets around the globe [5].

Guidi et al. and others in [6] designed a clinical decision support system (CDSS) for the heart failure analysis. In their paper, performance comparison of various machine learning classifiers like Artificial neural network (ANN), support vector machine (SVM), CART system with fuzzy rules, and Random forests has been made in which CART model and Random forest outperformed by achieving an accuracy of 87.6%. In [7], the authors proposed a logistic regression classifier after feature selection based upon a decision support system for the classification of heart disease and achieved an accuracy of 77%. Authors in [8] used two approaches, multilayer perceptron (MLP) and support vector machine, to classify heart disease and reached an accuracy of 80.41%. In [9], the authors proposed and evaluated a hybrid classification system of heart disease and achieved an accuracy of 87.4%. They combine the fuzzy and artificial neural network techniques for classification to find the results. Palaniappan et al. In [10] have applied Naive Bayes, ANN, and Decision Tree algorithms to diagnose the existence of heart

disease. According to their results, ANN comes out as the best predictive model with an accuracy of 88.12% compared to Naive Bayes with the accuracy of 86.12% and Decision Tree with only 80.4%. The authors proposed a three-phase model in [11] for heart disease diagnosis. They achieved an accuracy of only 88.89%.

Accuracy is the most important factor in prediction, but this is not only the one. Some researchers have taken some other parameters like precision, recall, f-measure, and R2 values into heart disease prediction. In [12], the authors used the Dimensionality reduction technique to process the raw data of 74 features first and then divide them into three groups. They could achieve the highest accuracy of 99.4% for CH, 100% precision, and 97.1% recall while using CHI-PCA with RF classifier. Shamsollahi in [13] has used combined predictive and descriptive approaches for predicting Coronary Artery Disease. They have selected the k-means method for clustering (descriptive) and various classification methods (predictive), including CHAID, Quest, C5.0, C&RT decision tree, and ANN method. They compared the results on parameters precision, accuracy, specificity, sensitivity, and error rate. As per the results, C&RT comes out as the best method for the entire dataset with only 0.074 errors. In [14], authors applied decision tree classification using J48, random forest, and logistic model trees algorithms on the UCI repository. It is concluded from their results that the J48 tree classification algorithm is the most excellent classifier for heart disease prediction because it achieves the highest accuracy and smallest amount of total time to build. Moreover, effect is pruning is clearly visible. They could achieve an accuracy of only 56.76% and time to build is 0.04 seconds for J48 while logistic model trees reach the only accuracy of 55.77% with a total time to build 0.39 seconds.

Authors have implemented five different classifying algorithms: Naive Bayes, Decision Tree, discriminant, Random Forest, and Support Vector Machine with big datasets and compared their performance in terms of accuracy, precision, specificity, recall, and F-measure [15]. Among all five classifiers, the decision tree ranks first, achieving an accuracy of 99.0%, with random forest stands at the second position with an accuracy of 93.4%.

### 3. Proposed Methodology

The experiment's process flow is explained in Figure 1 and further sections explain the proposed methodology used.

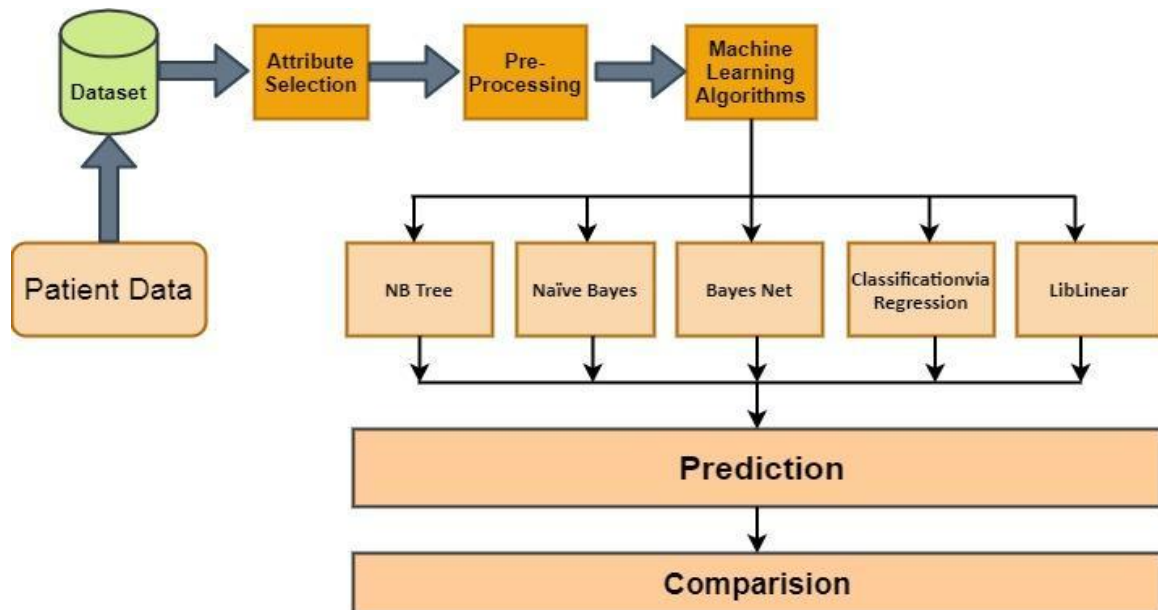


Figure1: Methodology used

#### 3.1 Dataset

We have taken the UCI repository dataset from Kaggle [16] named as Heart Failure prediction. The dataset has in total 13 attributes and 299 records. This patient's data includes the data of 194 men patients and 105

women patients. However, we have used only 12 attributes for this experimentation, as shown in table 1. We have not taken the Time attribute considering it is the consultation duration, and we feel it not so relevant for our study.

Table 1: Dataset Information

Attribute ID	Attribute Used	Attribute Information
A1	Age	Age of Patient. The value ranges from 40 years to 95 years
A2	Sex	Gender of the patient represented in binary form 1 = male. 0 = female
A3	Anemia	Reduction in hemoglobin 1:Yes 0:No
A4	Creatinine Phosphokinase	Level of the creatinine phosphokinase (CPK) enzyme in the blood measured in micrograms per liter
A5	Diabetes	Fasting blood sugar of the patient. If greater than 120 mg/dl the value is 1 (true), otherwise value is 0 (false). 1 = true. 0 = false.
A6	Ejection Fraction	Percentage of blood leaving ranges from 14 to 80
A7	High Blood Pressure	If a patient has high blood pressure (BP>120/80) 1:Yes 0:No
A8	Platelets	Platelet count in the blood and its unit is shown in kplatelets/ml

<b>A9</b>	Serum Creatinine	Creatinine level in blood and its unit of measure is mg/dl
<b>A10</b>	Serum Sodium	Sodium level found in patient blood and its unit is milliequivalents per liter
<b>A11</b>	Smoking	Patient smokes 1: Yes 0: No
<b>A12</b>	Time	This is follow up time with patients
<b>A13</b>	DEATH_EVENT	The occurrence of death due to heart disease 1 = yes. 0 = no

### 3.2 Data Pre Processing

The real-life data consists of redundant values and lots of noise. The data needs to be cleaned, and the missing values need to be filled before the data is fed to generate a model. In the pre-processing process, these issues are taken care of so that the prediction can be made accurately. Once the cleaning of data is done, i.e., the noise is removed, and the missing values are filled, we need to transform it. Many supervised learning algorithms work on nominal or cardinal data. So data transformation is applied to the dataset obtained from UCI in the present work. Reduction of the dataset is applied to convert the complex dataset into a more straightforward form, improving the model's accuracy

### 3.3 Tool Used

WEKA 3.8.4 machine learning tool is used to conduct this study written in Java and developed at the University of Waikato. WEKA tool provides us with different classifiers to examine the performance. WEKA is used to evaluate different data mining tasks like pre-processing, classification, regression, and many more. WEKA accepts .csv and .arff file format and the chosen dataset has already created the required data in the mentioned format.

### 3.4 Classification Algorithms

After going through an intensive literature review, we have selected five classification

algorithms: regression, naive Bayes tree, naive Bayes classification, Bayes network, and Liblinear.

**Regression** [17][18]: Regression is a supervised learning technique used to predict the class of the dataset when the target values are known[19]. The current study includes the regression to generate a model with the parameters, namely, age, gender, etc., and we have predicted the unknown class. The technique of regression works as follows:

The parameters used to make the prediction are continuous variables ( $\theta_1, \theta_2, \dots, \theta_n$ ). Based on these parameters, the model tries to find the best fit to predict Y's target variable and improve upon the accuracy. Using the function F of more predictors ( $x_1, x_2, \dots, x_n$ ) and a factor  $e$  as an error, the formula for calculation Y (value of the target variable) as

$$Y = F(x, \theta) + e \quad (1)$$

The target variable Y is dependent on the predictor variables, which are independent of each other. The model is generated based on the relation between the predictors and the target class. This is done in the training process. The model thus built is now fed with different unknown datasets for which the target value is predicted. The number of correctly predicted classes constitutes the accuracy and establishes the effectiveness of the model.

**Naive Bayes Tree:** It is a hybrid approach in which the model is generated using the naïve Bayes and Decision tree Approach. The naïve Bayes classification assumes that the features are unbiased of each other, and the decision

tree assumes that the features are dependent on each other. So the hybrid approach takes advantage of both approaches. The decision tree is built by considering only one feature, and output is fed to the node. Based on the outcome of each node, other features are selected. In this hybrid approach, the split is done in the same manner by considering only one feature at every node but with Naive-Bayes classifiers at the leaves. In large datasets, data splitting is considered a vital and important task for classification using the features we have implemented the naive Bayes tree classification.

**Naive Bayes Classification** [20]–[22]: This classification technique is based on the Bayes theorem, which works on the assumption that the existence of one feature is independent of the other feature. The advantage of the Naive Bayes classification is that it requires a small amount of data to create/train the model.

Bayes theorem provides a way of calculating posterior probability (conditional probability where we are finding probability under a given condition assumed to be true)  $P(c|x)$  from  $P(c)$ ,  $P(x)$ , and  $P(x|c)$ . The following is the formula to calculate posterior probability:

$$P(c|x) = \frac{P(x|c) \cdot P(c)}{P(x)} \quad (2)$$

Where:

$P(c|x)$  is the conditional probability that occurs when  $x$  has already occurred

$P(c)$  is the known probability of the class.

$P(x|c)$  is the conditional probability of  $x$  condition that  $c$  has occurred.

$P(x)$  the known probability of the class.

**Bayes Network:** The naïve Bayes algorithm assumes the independence of features. This hypothesis hampers the performance of the NB classifier. To improve the performance of the classifier, the Bayes networking algorithm is applied. The network is an acyclic graph that shows the joint probability distribution of the random variables/features. Each node/vertex of the graph represents a feature, and the edge represents the correlation between the features. This, in a way, reduces the effect of the hypothesis that the features are independent of each other. The independence of the features is then evaluated to reduce the number of parameters needed to calculate the probability distribution and compute the posterior probabilities. The acyclic graph is a joint

probability distribution of random variables, say  $U$ . mathematically, we can say that it is an ordered pair  $U = (G, Y)$ . The first component of the ordered pair  $G$  is the acyclic graph. In this graph, the vertices represent the random variable  $X_1, X_2, \dots, X_n$ , and the edges represent the relationship between these variables. The second component,  $Y$ , is the set of features that constitute the network. It contains a feature  $Y_{xi}|x_i = P_B(x_i|x_i)$  for each possible value  $x_i$  of  $X_i$ , and  $\Pi_{xi}$  of  $\Pi_{Xi}$ , where  $\Pi_{Xi}$  denotes the set of parents of  $X_i$  in  $G$ . A Bayesian network  $B$  defines a joint probability distribution (PDF) over  $U$ , and this is a unique PDF.

$$P_B(X_1, X_2, \dots, X_n) = \prod P_B(X_i | \Pi_{Xi}) \quad (3)$$

**LibLinear:** LIBLinear is an open-source library for linear classification. It supports two linear classifications, one logistic regression, and another is the Linear Support vector machine. Given a set of instance-label pairs  $(x_i; y_i); i = 1; \dots; l; x_i \in \mathbb{R}^n$ , both methods solve the following unconstrained optimization problem with different loss functions  $\xi(w; x_i; y_i)$ :

$$\begin{aligned} \min \frac{1}{2} w^T w + C \sum_{i=1}^l \xi(w; x_i, y_i), \\ \min \frac{1}{2} w^T w + C \sum_{i=1}^l \xi(w; x_i, y_i), \end{aligned} \quad \begin{array}{l} C \text{ is a} \\ \text{penalty parameter, and } C > 0 \end{array} \quad (4)$$

### 3.5 Evaluation Matrices

We have considered four parameters for our paper. In the present work, the prediction class is if the person having certain attributes has died because of heart disease or not, so the class  $C$  in the above table is no. of instances belonging to the class. Figure 2 is the confusion matrix.

TP is the number of people who died because of heart disease, and the model also predicted the same. Similarly, TN is the person who didn't die of a heart ailment, and our model also predicted the same. False Positive (FP) is a Type I error because the model predicted that the person died of an ailment, but actually, the patient didn't. False-negative is a type II error. The model predicted that the person didn't die of the ailment, but he/she did.

The accuracy of the model is calculated through the formula given below:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / \text{Total no. of instance} \quad (5)$$

The recall is the measure of correctly predicted classes out of the total positive classes. The formula is as follows:

$$\text{Recall} = (\text{TP}) / (\text{TP} + \text{FN}) \quad (6)$$

Precision is the measure of actual positive classes out of all the correctly predicted positive classes. The formula for the recall is as follows:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (7)$$

Comparing the two models becomes difficult when the precision is low, and the recall value is high. In the case of vice versa, the two parameters are not of much use for comparison of the models. F-score is used to compare the models in such cases. F-score uses the harmonic mean of the two values. This helps to measure the recall and precision at the same time. Instead of the Arithmetic mean, the harmonic mean is used because the Arithmetic mean is sensitive to extreme values.

$$\text{F-score} = (2 * \text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

Actual class \ Predicted class	C	Not in C
C	True Positives (TP)	False Negatives (FN)
Not in C	False Positives (FP)	True Negatives (TN)

Figure2: Confusion Matrix

### 3.6 k-Fold Cross-Validation

Dividing the dataset into k parts of equal size in which k-1 sets are used for training purposes and rest are used for evaluation is termed as k-fold cross-validation [23]. For instance, if we use 10-fold cross-validation, 90 percent of total data is used for training the classifier, and the rest 10 percent is used for evaluation.

## 4. Results and Discussion

The chosen five different classification algorithms were implemented on the heart disease dataset of the UCI repository. The experimental results have been obtained on the framework of WEKA 3.8.4. We used different k as 5, 10, and 20 for cross-validation and evaluated the above mentioned four parameters using five classification algorithms on WEKA. Table 2 tabulates the results obtained when taken 5-fold CV classification with five algorithms to evaluate the accuracy, F-measure, precision, and recall. Similarly, table 3 and table 4 show our experiment's simulation results on weka with 10-fold and 20-fold CV classification. Table5 tabulates the results when we have used 66% data for training the system and the rest 34% data for

evaluating the results. From the results, we can easily predict that Bayesian Network outperforms all with the highest accuracy, precision, f-measure, and recall in each method. Naive Bayes network uses an acyclic graph where each node represents a feature, and the edge represents its relation with other features. In the present work, the features such as age, gender, blood pressure, diabetes, etc., contribute towards heart disease [24]. Hence, the accuracy for this classifier outperforms the other. This establishes our hypothesis that the features such as age, gender, etc., when classified in the form of a graph (where these are dependent on each other), means that the heart-related ailment depends on these factors. So we can use this technique for the prediction of heart disease[25].



**Table 2: Performance Comparison of classifiers (k=5)**

Algorithms	Accuracy (%)	F-Measure (%)	Precision (%)	Recall (%)
LibLINEAR	74.24	73.1	73.1	74.2
Naïve Bayes	77.92	77.6	77.4	77.9
NB Tree	73.91	73.1	72.9	73.9
Bayes Net	78.59	78.7	78.8	78.6
Classification via Regression	71.90	70.1	70.2	71.9

**Table3: Performance Comparison of classifiers (k=10)**

Algorithms	Accuracy (%)	F-Measure (%)	Precision (%)	Recall (%)
LibLINEAR	76.58	75.6	75.7	76.6
Naïve Bayes	77.92	77.8	77.7	77.9
NB Tree	77.59	77.4	77.3	77.6
Bayes Net	79.26	79.5	79.8	79.3
Classification via Regression	75.25	73.4	74.2	75.3

**Table 4: Performance Comparison of classifiers (k=20)**

Algorithms	Accuracy (%)	F-Measure (%)	Precision (%)	Recall (%)
LibLINEAR	72.90	71.9	71.7	72.9
Naïve Bayes	75.25	75.2	75.1	75.3
NB Tree	74.91	74.7	74.5	74.9
Bayes Net	75.91	76.2	76.7	75.9
Classification via Regression	75.91	74.3	75	75.9

**Table5: Performance Comparison of classifiers (percentage split= 66%)**

Algorithms	Accuracy (%)	F-Measure (%)	Precision (%)	Recall (%)
LibLINEAR	74.50	73	74.8	74.5
Naïve Bayes	72.54	71.8	72	72.5
NB Tree	72.54	71.8	72	72.5
Bayes Net	74.50	74.4	74.3	74.5
Classification via Regression	73.52	73.7	74	73.5

## 5. Conclusion and Future Scope

In this paper, five data mining classifiers (LibLinear, Naive Bayes, Naive Bayes tree, Bayes network, and classification via regression) on heart disease data taken from the UCI repository have been implemented. The goal of this experimentation is to detect the accuracy in the prediction of heart disease of patients. We successfully achieved the highest accuracy of 79.28% with the Bayesian network classifier followed by naive Bayes. The reason behind excellent performance by the Bayesian network is the use of graphs in it, as graphs can reflect the relationship better between dependent variables as we have in our dataset like smoking habit, diabetes, high BP,

etc. Hence, we get better accuracy and prove that these factors contribute to heart disease occurrence.

In the future, we could use these results to design an effective prediction system that could help our medical practitioners diagnose and treat heart disease. Also, we could implement these data mining techniques for other diseases like diabetes, etc.

## References

- [1] S. Gupta, D. Kumar, and A. Sharma, "Performance analysis of various data mining classification techniques on healthcare data," *Int. J. Comput. Sci. Inf. Technol.*, vol. 3, no. 4, pp. 155–169, 2011.

- [2] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *Int. J. Comput. Appl.*, vol. 17, no. 8, pp. 43–48, 2011.
- [3] C. S. Dangare and S. S. Apte, "Improved study of heart disease prediction system using data mining classification techniques," *Int. J. Comput. Appl.*, vol. 47, no. 10, pp. 44–48, 2012.
- [4] S. Sa, "Intelligent heart disease prediction system using data mining techniques," *Int. J. Healthc. Biomed. Res.*, vol. 1, pp. 94–101, 2013.
- [5] S. Nazir, S. Shahzad, S. Mahfooz, and M. Nazir, "Fuzzy logic based decision support system for component security evaluation," *Int. Arab J. Inf. Technol.*, vol. 15, no. 2, pp. 224–231, 2018.
- [6] G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, "A machine learning system to improve heart failure patient assistance," *IEEE J. Biomed. Heal. informatics*, vol. 18, no. 6, pp. 1750–1756, 2014.
- [7] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Am. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, 1989.
- [8] M. Gudadhe, K. Wankhade, and S. Dongre, "Decision support system for heart disease based on support vector machine and artificial neural network," in *2010 International Conference on Computer and Communication Technology (ICCCT)*, 2010, pp. 741–745.
- [9] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Syst. Appl.*, vol. 35, no. 1–2, pp. 82–89, 2008.
- [10] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," in *2008 IEEE/ACS international conference on computer systems and applications*, 2008, pp. 108–115.
- [11] E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, "Heart diseases diagnosis using neural networks arbitration," *Int. J. Intell. Syst. Appl.*, vol. 7, no. 12, p. 72, 2015.
- [12] A. K. Garate-Escamilla, A. H. E. L. Hassani, and E. Andres, "Classification models for heart disease prediction using feature selection and PCA," *Informatics Med. Unlocked*, p. 100330, 2020.
- [13] M. Shamsollahi, A. Badiiee, and M. Ghazanfari, "Using combined descriptive and predictive methods of data mining for coronary artery disease prediction: a case study approach," *J. AI Data Min.*, vol. 7, no. 1, pp. 47–58, 2019.
- [14] J. Patel, D. TejalUpadhyay, and S. Patel, "Heart disease prediction using machine learning and data mining technique," *Hear. Dis.*, vol. 7, no. 1, pp. 129–137, 2015.
- [15] I. A. Zriqat, A. M. Altamimi, and M. Azzeh, "A comparative study for predicting heart diseases using data mining classification methods," *arXiv Prepr. arXiv1704.02799*, 2017.
- [16] G. J. Davide Chicco, "Heart Failure Prediction," 2015. <https://www.kaggle.com/andrewmvd/heart-failure-clinical-data> (accessed Nov. 10, 2020).
- [17] F. E. Harrell, "Ordinal logistic regression," in *Regression modeling strategies*, Springer, 2015, pp. 311–325.
- [18] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.
- [19] K. Larsen, J. H. Petersen, E. Budtz-Jørgensen, and L. Endahl, "Interpreting parameters in the logistic regression model with random effects," *Biometrics*, vol. 56, no. 3, pp. 909–914, 2000.
- [20] L. Li, Y. Wu, and M. Ye, "Experimental comparisons of multi-class classifiers," *Informatica*, vol. 39, no. 1, 2015.
- [21] P. Ahmad, S. Qamar, and S. Q. A. Rizvi, "Techniques of data mining in healthcare: a review," *Int. J. Comput. Appl.*, vol. 120, no. 15, 2015.
- [22] S. S. Nikam, "A comparative study of classification techniques in data mining algorithms," *Orient. J. Comput. Sci. Technol.*, vol. 8, no. 1, pp. 13–19, 2015.
- [23] V. Madaan and A. Goyal, "Predicting



- Ayurveda-Based Constituent Balancing in Human Body Using Machine Learning Methods," in IEEE Access, vol. 8, pp. 65060-65070, 2020, doi: 10.1109/ACCESS.2020.2985717.
- [24] Vishu Madaan and Anjali Goyal, "Analysis and Synthesis of a Human Prakriti Identification System Based on Soft Computing Techniques", Recent Patents on Computer Science, 12(1), pp 1-10, 2019. DOI: 10.2174/2213275912666190207144831
- [25] Prateek Agrawal, Vishu Madaan, Vikas Kumar, "Fuzzy Rule Based Medical Expert System to Identify the Disorders of Eyes, ENT and Liver", International Journal of Advanced Intelligence Paradigm (IJAIP), vol 7, issue3-4, pp. 352-367, Inderscience Publications, 2015.