IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HEART DISEASE PREDICTION USING MACHINE LEARNING AND DEEP LEARNING

[1]Aniket Tewari, [2]Atul Yadav, [3]Shweta Mayor Sabharwal

[1]Student, [2]Student, [3]Professor

[1]Bachelor of Technology (B. Tech – CSE),

[1]Galgotias University, Greater Noida, Uttar Pradesh, India

Abstract: **In this era, diseases are very common and heart disease is the most common one. Machine learning made a great impact on almost every sector. Health sector is also much helped by the machine learning. By the help of machine learning, we can make predictions of the heart disease. The correct and accurate prediction of heart disease on various parameters can save the life of many people. There are various factors on which the heart disease can be judged like blood pressure, cholesterol level, and many more. With this model, the differentiation of healthy and non – healthy person can e easily done.**

So, on considering all the parameters the model predicts the heart disease. The model is trained on a large dataset which contains various attributes. The dataset is split into two main parts namely training and testing dataset. The model is trained on the dataset named training dataset and the testing part is carried out using the test dataset. After successful training, the model predicts the disease. The various machine algorithms are used namely Support Vector Machine, K-NN Algorithm, Naïve Baye's Classifier Algorithm, Random Forest, Decision Tree, etc.

The accuracy table is made for various algorithms in order to judge the model upon various scenarios. The confusion matrix is also prepared which also gives the better view of the accuracy. Various graphs are also made using the appropriate libraries which gives better visualization of the dataset.

So, in this way the model predicts the heart disease and gives accurate results and is also useful for the people.

Index Terms **- MACHINE LEARNING, DEEP LEARNING, HEART DISEASE, SVM, KNN, CNN.**

I. INTRODUCTION

In this busy life, person do not get time to look after their health. This as a result causes various health problems like heart disease. Heart disease includes a wide range of symptoms that affects the heart. Today, heart disease is a major cause of the death in worldwide. According to the WHO Report, approximately 18 million deaths occur annually due to cardiovascular disease [1]. There are various factors which affect the condition of the heart of a person. The unhealthy factors which are responsible for the heart disease are blood pressure, high cholesterol level, hypertension, high level of triglycerides, and many more.

American Heart Association has conducted a research and enlisted the factors which make difficult to predict the heart disease. The symptoms are like irregular heartbeat, swollen legs, weight loss, hypertension, tiredness, and many more [2]. These factors if ignored causes difficulty in future.

The model is firstly trained on a large dataset which contains various attributes. The dataset is mainly split into two parts namely training dataset and test dataset. The training dataset is used to train the model while the testing dataset is used for the testing purpose. The dataset is split into 70-30 ratio. The 70 part comprises of the training and 30 part comprises of the testing dataset. The model can be trained to predict the output and the predicted output can be analysed deeper for better visualisation [3].

The figure below demonstrates the split of the dataset into two parts in the ration 70:30.

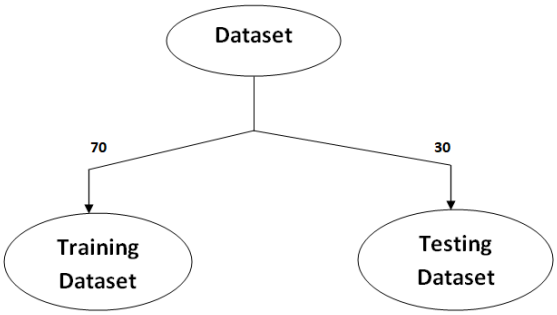

Fig.1- Train – Test Dataset Split

The accuracy table is prepared which shows the accuracy of the model for various splits like 70-30, 80-20, 70-30, 60-40, 50-50, 10 fold and so on.
Based on the table, we get the accuracy of the model. The confusion matrix is basically used to determine the performance and accuracy of the prediction model. The confusion matrix resembles the following. It easily represents the accuracy of given model.

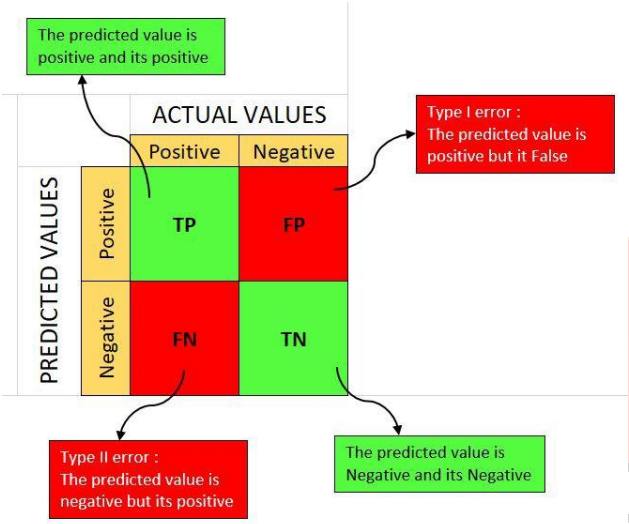

Fig.2- Confusion Matrix

The various algorithms are used in making this model. The algorithms used in the model are like Support Vector Machine (SVM), Naive Bayes, Random Forest Algorithm, Decision Tree, K - NN Classifier, CNN, etc.
Support Vector Machine (SVM) is the supervised machine learning algorithm which is basically used to solve the classification and regression problems. SVM aims to find the best line or we can say the decision boundary that segregates the n-dimensional space into various classes and we can easily put the new coming data point in the appropriate class.
The below figure demonstrates the SVM algorithm on how the plane is divided into classes and any new data point can be put easily in a appropriate class.

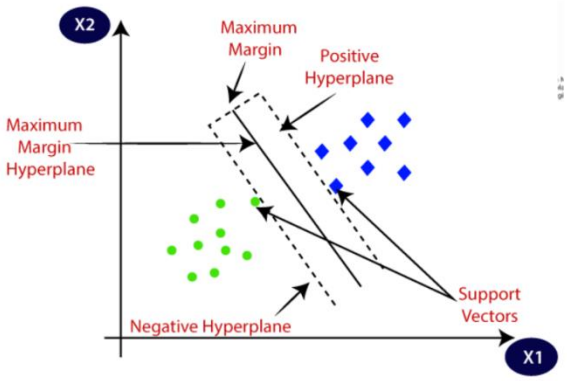

Fig.3- Support Vector Machine Algorithm

K-Nearest Neighbour (KNN) is another Machine Learning Algorithm. It also uses supervised machine learning algorithm. KNN assumes similarity between the already present classes and whenever any new data point comes the algorithm puts this into an appropriate class.

II. LITERATURE REVIEW

Heart Disease Prediction is not any new task. Many other approaches have also used in order to predict the heart disease. They all used different algorithms which gives different accuracy in terms of percentage.

1. According to the report prepared by Gárate-Escamila et al, it predicted that heart diseases are more common in men than women. In this ANN and the DNN algorithms were used with the statistical model [4].
2. According to the Harvard Medical School report, the large dataset were used in order to predict the heart diseases namely Hungarian – Cleveland Dataset. They applied different machine learning algorithms in order to easily predict the heart disease. The dimensionality reduction and the feature selection technique is also used in order to make the dataset appropriate. The various techniques of feature selection are used to extract the most important features. The techniques used are like Filter method, Embedded method and Wrapper method. The algorithm used is KNN classifier where model is trained on existing classes and whenever any new datapoint comes, it can easily be put in an appropriate class [5].
3. The paper in Computational Intelligence and Neuroscience, predicts the heart disease as well using the machine learning technique. It used the concept of machine learning algorithms and deep learning as well. It compared the model with various algorithms like support vector machines, CART algorithms, KNN algorithms, random forest, decision tree, and several others.

   After comparing the results from all these algorithms an accuracy chart is prepared which depicts the performance and accuracy with different algorithms. The dimensionality reduction techniques are also used in making the prediction. Dimensional reduction techniques is the process of minimising the number of random variables upon consideration. There are various techniques under dimensionality reduction like PCA, LDA, and GDA. PCA works on the principle that till the datapoint in the higher dimensional space gets mapped with the datapoint in the lower dimensional space, the variance measure in the lower dimensional space is maximum [6].
   Dimensionality reduction basically reduces the computational time and removes the redundant features as well. So, in this way this model predicts the heart disease.
4. According to the Zhang et al Journal, a large audience is having the heart disease. The health of the heart depends upon various factors like blood pressure, cholesterol level, and many others. The prediction can be made easily through the support vector machine algorithm as its accuracy is good as compared to other models according to the journal. SVM is used to classify the disease and the data is matched with the data and record of the New York Hearts Association. After this, next all shortcomings are fed to another researchers. The SVM algorithm performed with 93% of f-measure for the model. The paper also showed various graphical models of the dataset which helps in easily visualisation of the dataset [7].
5. The journal published in Wettschereck et al also focuses on heart disease prediction. The main algorithm used in the model is K-NN algorithm. The reason behind using the KNN algorithm is that it is the derivation of lazy learning algorithm. It also uses the feature selection technique. The various techniques of feature selection are used like filter methods and wrapper methods. The KNN uses the already present classes and the model is trained on the basis of these the new data point is put in the appropriate class [8].
6. According to the Imani and Ghassemian research, heart disease prediction is not an easy task. One has to look for various other symptoms like blood pressure, cholesterol level, age of a person, blood sugar level, hypertension, and so on. According to the approach suggested by Imani, sometimes the data is not enough and appropriate, so some additional steps have to be performed. SO, he approached a weighted training sample method in order to achieve the same. This method included feature extraction techniques, in order to achieve the spatial dimension of the images for generating the results which as a result increases the accuracy of the model [9].
7. In the model proposed by Srinivas et. al. the prediction of heart disease was the prime consideration. It used the prediction in coal mines as the prime consideration. It used the algorithms like decision tree, naive bayes, logistic regression and the neural networks for making the prediction of the heart disease. They firstly extracted the important features of the dataset and then they trained the model upon various attributes and finally its accuracy was tested and then it made the predictions [10].

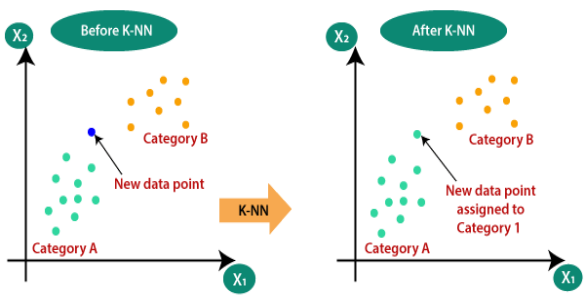

Fig.4- K-NN Algorithm

III. METHODOLOGY

3.1. Dataset—

The dataset used in the model is very large. It consistes of various attributes like age, sex, cholesterol level, trestbps, cp, fbs, restcg, thalach, exang, oldpeak, target, and many more.
The model is trained on this dataset and then it is used to predict the output.

3.2.—
Classifiers Used in this Research—

The main algorithms used in order to make predictions in the model are Support Vector Machine (SVM), K-Nearest Neighour Classifier, Random Forest algorithm, Decision Tree, and CNN.

(i) Support Vector Machine—

Support Vector Machine (SVM) is the supervised machine learning algorithm where the model is trained on a labelled dataset. The dataset used is well – formed. It is used for both classification and the regression problems. The main aim of SVM is to find the best boundary line or decision boundary that segregates the n - dimensional space into several classes, so that whenever any new data point comes it can be easily put into the appropriate class.

(ii) K-Nearest Neighbour—

K - Nearest Neighbour (KNN) is also a supervised machine learning algorithm. The algorithm trains the model upon the already available test cases and whenever any new data point comes it finds the similarity between the already available test cases and puts in an appropriate class. It is also a lazy learner algorithm. It simply classifies the new data according to the similarity between already available test cases.
The letter 'K' in KNN algorithm stands for a numeric value which tells the number of clusters present.

(iii) Random Forest—

Random Forest is also a supervised machine learning (ml) algorithm that predicts the output. It is used for both the classification and the regression problems. It is based on the concept of the ensemble learning. Random Forest contains various decision trees on various subsets and takes the average of all these in order to predict the output.

It takes less time as compared to the others algorithms. It also predicts the output with great accuracy.
It selects random 'k' data points and then builds the decision tree on these data points. Then for any new data point, it assigns to that one which has majority of votes.

(iv) Decision Tree—

Decision Tree is also a supervised machine learning algorithm that is used for both the classification and the regression problems. It is similar to a tree like structured classifier, where representation is like these-
Internal Nodes – Represents Features of the dataset
Branches – Represents Decision rules
Leaf – Represents Outcome
In simple words, decision tree asks question, and if the answer is YES / NO then, its split into various sub – trees.

(v) Logistic Regression –

Logistic regression is the supervised machine learning algorithm and is very popular. It basically predicts the output of the categorical variable, and is mainly used for the classification problems.
The output of the algorithm is either Yes / No, 0 / 1, True / False. But instead of giving the exact value, it gives the probabilistic value which lies within the range of 0 to 1.
In this, instead of finding the regression line, we find the S – shaped curve or S – shaped logistic function which predicts the values ranging between the maximum value as one and the minimum value which is zero.
The logistic function satisfying the algorithm is of the form-

$$\log (y/(1-y)) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots\ldots\ldots + b_nx_n$$

The dependent variables involved in the algorithm must be categorical in nature.
It is of several types like binomial, multinomial, and the ordinal.

Accuracy –

Accuracy is the ratio of the correctly labelled data to the whole pool of the data.
It basically answers the question which is how many disease did we correctly label pout of all the disease.

Mathematically, Accuracy = (TP+TN) / (TP+TN+FP+FN)

Precision –

Precision is basically the ratio of the correctly positive labelled data by the program to all the positive labelled data.
It basically answers the question which is how many of the labelled persons actually have the heart disease.

Mathematically, Precision = TP / (TP+FP)

Sensitivity—

The sensitivity is basically the ratio of the correctly positive labelled data to all the actually positive data.
In our model, the sensitivity is the ratio of the positive persons having the heart disease to all the persons who have heart disease in reality.

Mathematically, Sensitivity = TP / (TP+FN)

Specificity –

Specificity is the correctly negative labelled data in the dataset to all those who are free from it.
In our model, it is the ratio of the negative labelled data to those who in the reality are healthy.

Mathematically, Specificity = (TN) / (TN+FP)

F1 - Score—

F1 – score basically considers both the values of precision and recall before generating the results. In simple words, it is the harmonic mean (HM) of the precision and the recall.

Mathematically, F1 – Score = (Recall * Precision) / (Recall + Precision).

IV. RESULT AND INTERPRETATIONS

The model is successfully made and trained on various splits of the dataset like 80-20, 70-30, 60-40, and so on and also on various algorithms like Logistic Regression, Naive Bayes, Random Forest, Decision Tree, and so on.

First of all, the dataset is imported in the environment. So, we have used the dataset heart_disease_prediction_dataset.csv and then its top five entries are printed as below.
This table contains the top five entries of the dataset with all the attributes of it and with its values.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig. 5- Dataset

Now the info of the dataset is printed to get the type of variables, which shows what are the various attributes in the dataset and belong to which category. All the column names with their data type is shown in the table.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trestbps  303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalach   303 non-null    int64
 8   exang     303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Table -1- Dataset Info

The density graph is prepared which is shown below.
This shows the variation of the density with the given attribute. There is a peak in the density in between the range of values.
This table basically shows which attributes are of which type like integer, string and so on. This gives the full description of the dataset. It also states whether the variable or the attribute is null or not – null.
There are several columns like age, sex, sp, trestbps, chol, fbs, and so on which hold some values and on the basis of them the model is trained and is made to predict the heart disease.

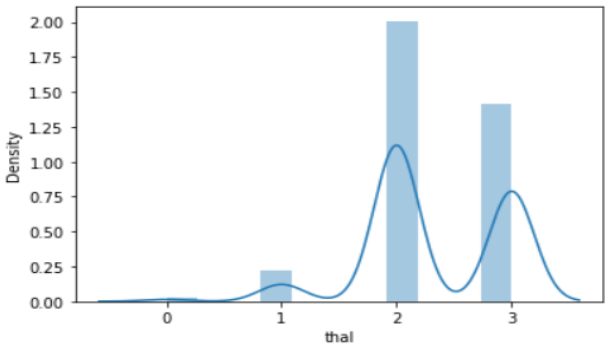

Fig. 6– Density Graph

The bar graph for the target v/s count is depicted below.

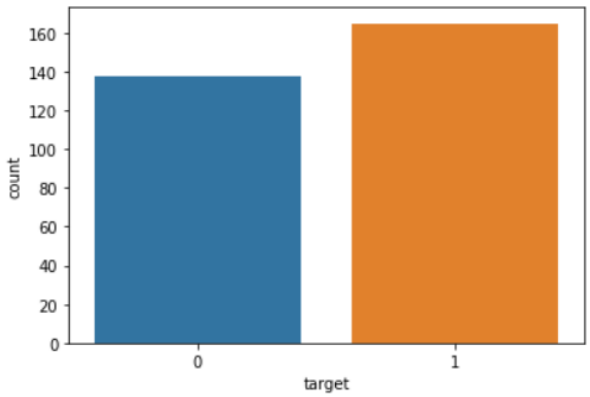

Fig. 7– Bar Graph

Now the various algorithms are applied on different splits.

a) 80-20 split--

In 80-20 split the dataset is split in a way that 80% comprises of the training dataset and the 20% comprises of the testing dataset.

The confusion matrix is prepared and using this, the accuracy score, sensitivity, specificity, precision and the f1-score is calculated. The model has higher accuracy for the Random Forest Algorithm.

The confusion matrix for the given split is plotted as follows.

	Predicted Negative	Predicted Positive
Actual Negative	22	5
Actual Positive	4	30

Fig. 8– Confusion Matrix

The various algorithms are used in order to check the accuracy of the model. Firstly, the model is trained on the

Logistic regression, then on naive bayes, support vector, KNN, decision tree, and on neural networks.
The comparison of these algorithms against accuracy is shown in the given figure.
The model has higher accuracy for the random forest algorithm as it is around 85%.

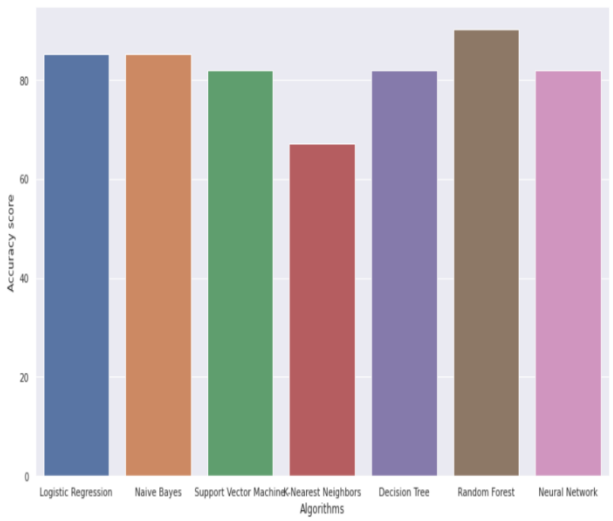

Fig. 9- Comparison in 80-20 split

b) 70-30 split—

Now the splitting of the dataset is done in such a way that 70% of the dataset comprises of the training dataset and the 30% comprises of the testing dataset.

The confusion matrix is prepared which contains data of True Positives, True Negatives, False Positives, and the False Negatives. The accuracy, sensitivity, specificity, precision, and the f1-score are calculated for the various algorithms. The graph comparison of them is depicted below.

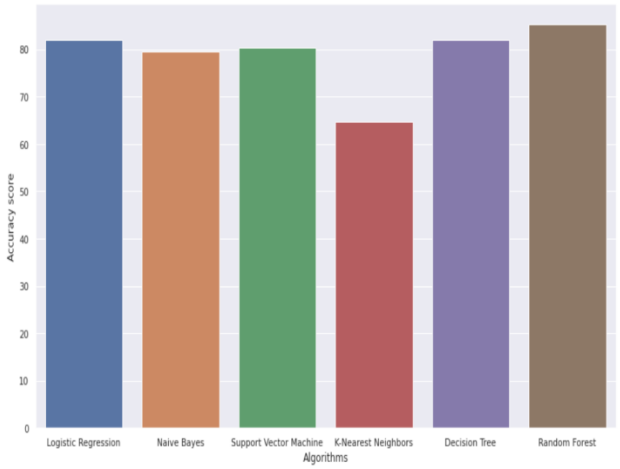

Fig. 10– Comparison in 70-30 split

So, in the 70-30 split also, the accuracy score for the random forest algorithm is maximum as it predicts correctly.

Now, the table showing the comparison of all algorithms is prepared which contains the sensitivity, specificity, precision and the f1 – score for the various algorithms.

Classifier	Partition Scheme	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)
Logistic Regression	80-20	88	81	81	85
	70-30	91	75	75	82
	60-40	87	76	76	81
Naive Bayes	80-20	91	78	78	84
	70-30	87	73	73	79
	60-40	86	73	73	79
SVM	80-20	88	74	74	81
	70-30	89	73	73	80
	60-40	87	73	73	79
KNN	80-20	68	67	67	67
	70-30	74	64	64	69
	60-40	73	56	56	63
Decision Tree	80-20	82	81	81	82
	70-30	77	75	75	76
	60-40	79	85	85	82
Random Forest	80-20	94	85	85	89
	70-30	87	83	83	85
	60-40	87	83	83	85

Table-2 – Comparison of Sensitivity, Precision, Specificity and F1 – Score of Logistic Regression, Naïve Bayes, SVM, KNN, DT and RF classifiers.

V. CONCLUSION

So, we have seen how this model works. The model is firstly trained on a very large dataset, and then is trained using various algorithms, and finally it predicts the output. There are several algorithms used like support vector machine (SVM), K- Nearest neighbor, CNN, Random Forest, and Decision Trees. The accuracy table and the confusion matrix is drawn for the various splits of the training and testing dataset.
The model gives the accurate prediction for the person's heart disease with an accuracy of 90%. So, this way, the model is very useful for all people.

VI. REFERECES

[1] World Health Organization, *Cardiovascular Diseases*, WHO, Geneva, Switzerland, 2020, https://www.who.int/health-topics/cardiovascular-diseases/

[2] American Heart Association, *Classes of Heart Failure*, American Heart Association, Chicago, IL, USA, 2020, https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/classes-of-heart-failure

[3] S. Marsland, "Machine learning," *An Algorithmic Perspective*, CRC Press, Boca Raton, FL, USA, 2020.

[4] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informatics in Medicine Unlocked*, vol. 19, Article ID 100330, 2020.

[5] Harvard Medical School, "Throughout life, heart attacks are twice as common in men than women," 2020, https://www.health.harvard.edu/heart-health/throughout-life-heart-attacks-are-twice-as-common-in-men-than-women

[6] Bharti, Rohit, et al. "Prediction of heart disease using a combination of machine learning and deep learning." *Computational intelligence and neuroscience* 2021 (2021).

[7] R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, and S. Speedie, "Automatic methods to extract New York heart association classification from clinical notes," in *Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1296–1299, IEEE, Kansas City, MO, USA, November 2017.

[8] D. Wettschereck, D. W. Aha, and T. Mohri, "A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms," *Lazy Learning*, vol. 11, no. 1/5, pp. 273–314, 1997.

[9] M. Imani and H. Ghassemian, "Feature extraction using weighted training samples," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 7, pp. 1387–1391, 2015.

[10] K. Srinivas, G. Raghavendra Rao, and A. Govardhan, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," in *Proceedings of 2010 5th International Conference on Computer Science & Education*, pp. 1344–1349, IEEE, Hefei, China, August 2010.

[11] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informatics in Medicine Unlocked*, vol. 19, Article ID 100330, 2020.

[12] P. Ramprakash, R. Sarumathi, R. Mowriya, and S. Nithyavishnupriya, "Heart disease prediction using deep neural network," in *Proceedings of the 2020 International Conference on Inventive Computation Technologies (ICICT)*, pp. 666–670, IEEE, Coimbatore, India, February 2020.

[13] Kanksha, B. Aman, P. Sagar, M. Rahul, and K. Aditya, "An intelligent unsupervised technique for fraud detection in health care systems," *Intelligent Decision Technologies*, vol. 15, no. 1, pp. 127–139, 2021.

[14] K. Divya, A. Sirohi, S. Pande, and R. Malik, "An IoMT assisted heart disease diagnostic system using machine learning techniques," in *Cognitive Internet of Medical Things for Smart Healthcare*, A. E. Hassanien, A. Khamparia, D. Gupta, K. Shankar, and A. Slowik, Eds., vol. 311, pp. 145–161, Springer, Cham, Switzerland, 2021.