IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING PREDICTIVE SYSTEM TO IDENTIFY HEART DISEASE AT AN EARLY STAGE

[1]Utsab Ray, [2]Karabi Ganguly, [3]Kinshuk Ganguly

[1]BTech, Department of Biomedical Engineering - JISCE, [2]Associate Professor, Department of Biomedical Engineering - JISCE, [3]BTech, Department of Computer Science and Engineering – IEM Kolkata

[1]Department of Biomedical Engineering,
[1]JIS College of Engineering, Kalyani, Nadia, India

Abstract: According to a new WHO report, cardiac problems are becoming more prevalent. Due to this, 17.9 million individuals pass away annually. It becomes more challenging to diagnose and begin treatment at an early stage as the population grows. Heart is an important organ of all the organisms. The diagnosis and prognosis of heart-related diseases require greater accuracy, perfection, and correctness because even a minor error can result in exhaustion or even death. Deaths due to cardiac diseases are common, and the numbers of these deaths are increasing day by day. In this study, we use the Irvine repositories dataset for training and testing to assess the efficacy of machine learning techniques for predicting heart disease; KNN, SVM and other related supervised algorithms have been employed. The accuracy of the algorithms have been verified and finalized to develop a predictive system. As a decision support system, this predictive model can therefore be used by medical professionals as analytical, diagnostic and prognostic tool in cardio-pathology domain.

***Keywords* - Heart- related diseases, machine learning, dataset, supervised algorithms and predictive model.**

I. Introduction

The maintenance of the heart is crucial due to it being among the largest and most important organs in the human body. The majority of diseases are heart-related, making it necessary to predict heart diseases, and for this reason comparative studies are required in this field. Currently, the majority of patients pass away because their diseases were discovered too late due to instrument inaccuracy, necessitating the knowledge of more effective algorithms for disease prediction.

The primary focus of mankind is on healthcare. According to WHO recommendations, everyone has a fundamental right to good health. It is believed that suitable healthcare facilities should be accessible for routine health checks. Heart-related illnesses account for over 31% of all fatalities worldwide. Because of the absence of diagnostic facilities, qualified physicians, and other assets that impact the precise prognosis of heart disease, early identification [1] and treatment of various cardiac illnesses is very complicated, especially in poor countries.In light of this worry, medical aid software is currently being created using computer applications and methods of machine learning as a support system for the early diagnosis of cardiac disease. Early illness prediction once an individual suffers is the issue the healthcare sector is currently facing. The size of the medical history records and data makes them potentially incomplete and inconsistent in the real world. In the past, it may not have been possible for all patients to receive early-stage treatment and accurate disease prediction [2].The risk of death can be decreased by detecting any heart-related illnesses in their early stages. In order to comprehend the patterns in the data and derive predictions from them, many ML techniques are applied in the field of medicine. In general, healthcare data have enormous volumes and complex structures. Big data may be handled by ML algorithms, which can then be mined for useful information.

One effective testing tool is machine learning, which is centered in training and testing. Machine learning is a particular subset of Artificial Intelligence (AI), a large field of learning in which machines imitate human abilities. A growing area of data science is machine learning, a subfield of AI study [3]. The algorithms used in machine learning are built to handle a wide range of tasks, including prediction, classification, and decision-making. On the other hand, machine learning systems are taught how to process and use data; as a result, the fusion of the two fields of technology is also known as machine intelligence.
Training data is needed in order to educate the ML algorithms. A model is created following the learning phase and is regarded as the product of the ML algorithm. In accordance with the concept of machine learning, which states that it learns from natural phenomena and things, this project uses biological parameters as testing data, such as cholesterol, blood pressure, sex, age, and others, and on the basic principle of these, a comparison is made in terms of algorithm accuracy.

Many authors have already made significant efforts to predict heart disease using machine learning algorithms [4-6], but this is an additional effort to conduct an experiment on benchmarking the Ucberkeley myocardial infarction forecast data - set while going to compare the four widely used ML techniques to determine which ML technique is the most accurate.

II. RELATED WORK

Any algorithm's performance is dependent on the dataset's bias and variance [7]. According to a research [7] on machine learning for heart disease prediction, naïve bayes performed better with minor differences and high biasness than high standard deviation and low biasness, which is KNN. Low biasness & high variance cause KNN to suffer from the issue of over fitting, which is why KNN performance degrades.

One nonparametric machine learning approach is the decision tree, however as we all know, overfitting is a problem that can be resolved using other overfitting removal methods. Support vector machines, which have an algebraic and statics foundation, build linear separable n-dimensional hyperplanes to classify datasets.

Heart disease severity is categorised using a variety of techniques, including KNN, decision trees, generic algorithms, and naive bayes [8]. According to Mohan et al.[8], combining two distinct procedures can result in a hybrid approach, which has the highest accuracy of all others at 88.4%.Data mining has been used by some researchers. In a particular research, [9] researchers explained how the intriguing pattern and understanding are gleaned from the sizable dataset. They compare the accuracy of different data mining, machine learning, and other techniques to decide which is the best one, and the results are in favour of svm.

SVM was shown to be the best among the machine learning and data mining algorithms developed in a particular work [10]; other algorithms included naive bayes, knn, and decision tree. These algorithms were trained using the UCI machine learning dataset, which contains 303 samples and 14 input features.

Using CAD technology, [11] researchers recently developed a multi-layer perceptron algorithm to predict the occurrence of human cardiac illnesses and the accuracy of the method. If more people use prediction systems to diagnose their illnesses, then more people will be aware of the ailments, which will lower the death rate for cardiac patients.

In another work, [12] researchers have shown that decision trees are more accurate than the naive bayes classification algorithm.

Many researchers have worked on this, including a definite multitude [13], where logistic regression has been employed to predict heart disease, support vector machines to predict diabetes, and Adaboost classifiers to predict breast cancer. They found that logistic regression had an exactness of 87.1%, back propagation machines had an accuracy of 85.71%, and Adaboost classifiers had an accuracy of up to 80%.

A survey report on the prediction of cardiac illnesses has demonstrated that hybridization performs well and provides better prediction accuracy than the older machine learning algorithms [14].

By using the logistic regression approach on this dataset, the authors of [15] were able to attain a prediction accuracy of 77%. By comparing different global evolutionary computation algorithms in this study, authors [16] improved their work and saw increased prediction accuracy.

In another article, researchers [17] proposed a study on the diagnosis of diabetic disease using ML techniques. This illness was thought to be an immensely important component of ML. According to a survey carried out by the World Diabetes Federation, 285 million people worldwide have diabetes (IDF).

The significance of ML techniques in numerous fields has been proved by a number of applications in a discrete research [18]. The strategy made advantage of specific machine learning techniques.

A prior piece on analytics and data mining applications was suggested in 2017 [19]. These processes were employed in the business world for a variety of reasons. They have examined 10 supervised learning algorithms and 8 unsupervised learning algorithms here [19]. They demonstrated an application for the tractor trailer type learning algorithms in their research.

.

III. MACHINE LEARNING ALGORITHMS

For the development of the cardiac disease prognostic model, we have selected six well-known ML approaches. These strategies' specifics are as follows:

Linear Regression - The supervised learning method is what it is. It is based on how independent and dependent variables relate to one another.

Support Vector Machine - Support In order to evaluate data and find patterns for classification and regression analysis, machine learning's Vector Machine [20] classification technique is utilised. SVM is frequently considered when the data is categorised as a two-class problem. Finding the appropriate hyper plane that isolates every data point from one class to the other is how this technique characterises data.

Decision Tree - Machine learning's Decision Tree method [21] is used to create Classification models. The structure of a tree is the foundation of this categorization approach. This falls within the supervised learning category because the desired outcome is already known. The decision tree approach can be applied to both categorical and numerical data.

Naïve Bayes - Based on the Bayes' Theorem [22], which assumes that features are statistically independent of one another; this supervised machine-learning technique was developed. High dimensional input data are employed with the Nave Bayes Classifier [23].The naive Bayes approach has many applications in computer vision.

Random Forest - A group of unprized classification-based trees makes up Random Forest [24]. Given that it is insensitive to dataset noise and has a very low risk of over fitting, it exhibits exceptional performance in terms of a variety of real-world issues.

It operates more quickly compared to several tree-based algorithms and typically increases accuracy for testing and validation data.

K Nearest Neighbour - It classifies different types of data with each other based on the distance between the locations where the data are located. The user determines the number of neighbours for each other's data sets, which is a very important factor in the assessment of the dataset.

IV. Methodology

Figure 1 depicts the entire methodology architecture.

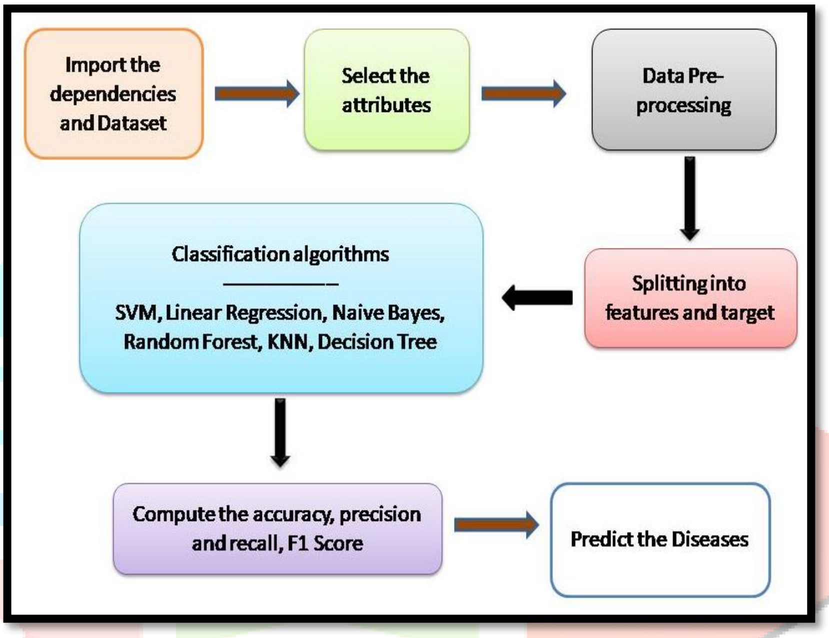

Figure 1: Methodology

Data collection - We used the Cleveland Cardiovascular Disease Dataset, which is available online at the UCI Repository [25]. The 14 qualities taken into account are as follows in Table 1:

Label	Explanation
Age	Depicts the age of the patient normally varying between 29 to 70
Sex	Gender [male -0, female -1]
Cp	Categorization of chest pain
Trestbps	Resting bp [blood pressure]
Chol	Cholesterol value
Fbs	Blood sugar value in fasting
Resting	Electro-cardio graphical result in resting state
Thali	Heart rate in a maximum stage
Exang	Exercise
Oldpeak	ST Slope in depression
Slope	Slope of ST Segment
Ca	Vessel Count
Thal	3 – normal
Targets	1 or 0

Table 1: Dataset classification

Data Pre-processing - Pre-processing is required for the machine learning algorithms to produce prestigious results. For instance, the Random Forest technique does not allow datasets with null values, so we must manage null values in the original raw data. For our work, we must use some categorised values into dummy values in the format of "0" and "1"

Data Balancing - Since the data balancing graph shows both of the target class are equal, data balancing is crucial for accurate results. The target classes are shown in Fig. 2 with "0" denoting patients with heart disease and "1" denoting patients without heart disease.

Figure 2: Target class

Accuracy and other computational factors - Four values—true positive (TP), false positive (FP), true negative (TN), and false negative—determine how accurate the algorithms are (FN).

V. Results and Discussion

```
[ ] import numpy as np
    import pandas as pd
    from sklearn.model_selection import train_test_split
    from sklearn.linear_model import LogisticRegression
    from sklearn.metrics import accuracy_score
```

Figure 3: Code to import Libraries

Data Collection and Processing

```
[ ] # loading the csv data to a Pandas DataFrame
    heart_data = pd.read_csv('/content/data.csv')
```

```
[ ] # print first 5 rows of the dataset
    heart_data.head()
```

	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

```
# print last 5 rows of the dataset
heart_data.tail()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

Figure 4: Code of data collection and processing

```
[ ] # number of rows and columns in the dataset
    heart_data.shape

    (303, 14)
```

```
[ ] # getting some info about the data
    heart_data.info()

    <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
```

```
[ ] # checking for missing values
    heart_data.isnull().sum()

    age         0
    sex         0
    cp          0
    trestbps    0
    chol        0
    fbs         0
    restecg     0
    thalach     0
    exang       0
    oldpeak     0
```

Figure 5: Code of data exploration

```
[ ] # statistical measures about the data
    heart_data.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Figure 6: Code of statistical measures of dataset

1 --> Defective Heart

0 --> Healthy Heart

Splitting the Features and Target

```
[ ] X = heart_data.drop(columns='target', axis=1)
    Y = heart_data['target']
```

```
[ ] print(X)

         age  sex  cp  trestbps  chol  ...  exang  oldpeak  slope  ca  thal
    0     63    1   3       145   233  ...      0      2.3      0   0     1
    1     37    1   2       130   250  ...      0      3.5      0   0     2
    2     41    0   1       130   204  ...      0      1.4      2   0     2
    3     56    1   1       120   236  ...      0      0.8      2   0     2
    4     57    0   0       120   354  ...      1      0.6      2   0     2
    ..   ...  ...  ..       ...   ...  ...    ...      ...    ...  ..   ...
    298   57    0   0       140   241  ...      1      0.2      1   0     3
    299   45    1   3       110   264  ...      0      1.2      1   0     3
    300   68    1   0       144   193  ...      0      3.4      1   2     3
    301   57    1   0       130   131  ...      1      1.2      1   1     3
    302   57    0   1       130   236  ...      0      0.0      1   1     2

    [303 rows x 13 columns]
```

```
[ ] print(Y)

    0      1
    1      1
    2      1
    3      1
    4      1
          ..
    298    0
    299    0
    300    0
    301    0
    302    0
    Name: target, Length: 303, dtype: int64
```

Figure 7: Feature and Target specification

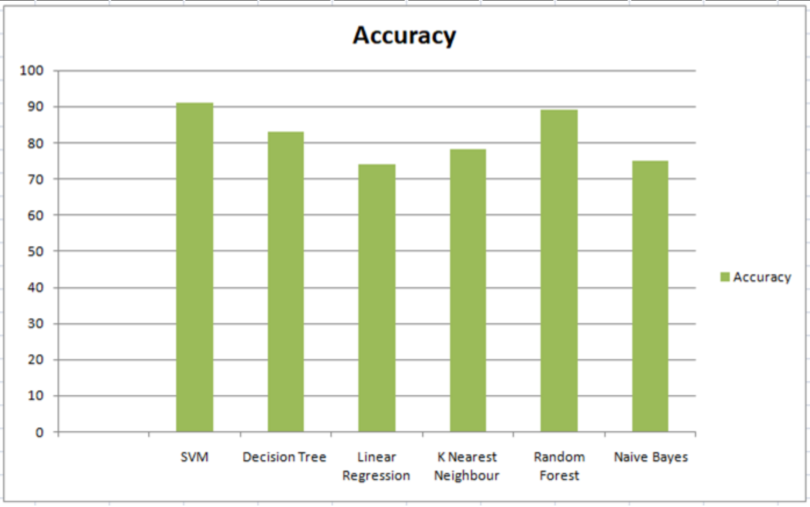

Figure 8: Accuracy of algorithms

Figure 8 shows a specific formula and a contrast of the accuracy outcomes of algorithms tested using the machine learning architecture against one another.

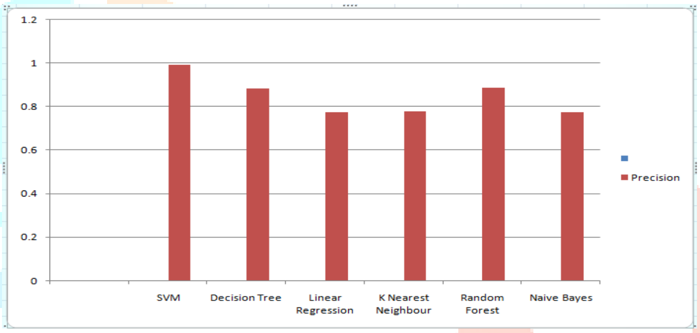

Figure 9: Precision of algorithms

Figure 9 shows a summary statistics and analysis of the precision outcomes of algorithms performed utilizing machine learning model against one another.

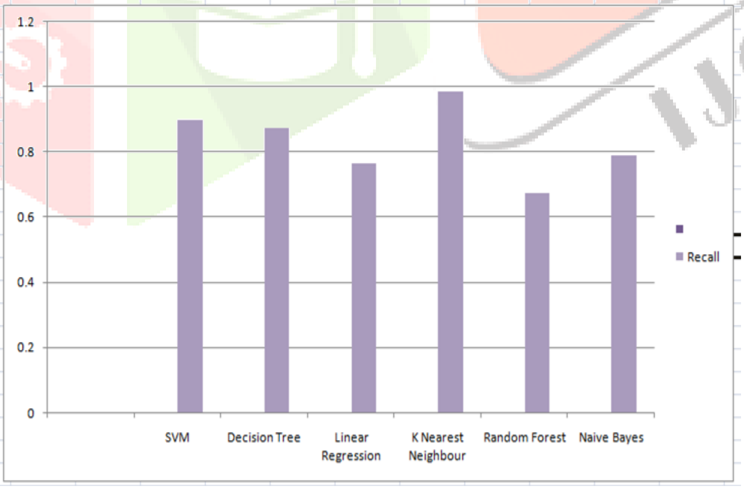

Figure 10: Recall of algorithms

Figure 9 shows a summary statistics and analysis of the recall outcomes of algorithms performed utilizing machine learning model against one another.

Building a Predictive System

```
[ ] input_data = (62,0,0,140,268,0,0,160,0,3.6,0,2,2)

    # change the input data to a numpy array
    input_data_as_numpy_array= np.asarray(input_data)

    # reshape the numpy array as we are predicting for only on instance
    input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

    prediction = model.predict(input_data_reshaped)
    print(prediction)

    if (prediction[0]== 0):
      print('The Person does not have a Heart Disease')
    else:
      print('The Person has Heart Disease')

    [0]
    The Person does not have a Heart Disease
```

Figure 11: Predictive System

Figure 11 portrays a final predictive system which is developed with the desired algorithm that yielded highest accuracy and that model is depicting the final state of heart disease as the related datasets are fed to the system.

VI. CONCLUSION

In this research, we have aimed to assess the numerous machine learning strategies and predict whether or not a specific individual will develop cardiac illness given various individual traits and indications. Our report's main focus was on examining the accuracy and examining the causes of the variations among various algorithms.

Since the human heart constitutes one of the body's most significant organs and heart disease prediction is a major human concern, algorithm accuracy constitutes one of the factors considered when evaluating an algorithm's performance.
The dataset utilised for both training and testing purposes affects how accurate machine learning algorithms are.SVM is the best method when we compare them based on the dataset and other considerations. Other algorithms may function more effectively for various situations and datasets, but in our case, we have found this result. Also, if we increase the amount of training data, we might be able to obtain results that are more accurate, but processing time would be longer, and the system would be slower than it is currently since it would have to deal with more data and be more complex. We made this decision since it is easier for anyone to work with after taking these potential factors into account.

In order to reduce the rate of mortality cases through increased disease awareness, more machine learning techniques will be deployed in the future to analyse cardiac problems more effectively and detect illnesses early.

Acknowledgment

We would like to thank Department of Biomedical Engineering, JIS College of Engineering for their support and help.

REFERENCES

[1] https://www.who.int/hrh/links/en/.
[2] AshishChhabbi, LakhanAhuja, SahilAhir and Y. K.Sharma, "Heart Disease Prediction Using Data Mining Techniques", © IJRAT Special Issue National Conference “NCPC-2016”, pp. 104-106, 19 March 2016.
[3] https://en.wikipedia.org/wiki/Machine_learning.
[4] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 204-207, doi: 10.1109/ISCC.2017.8024530.
[5] S. Dhar, K. Roy, T. Dey, P. Datta and A. Biswas, "A Hybrid Machine Learning Approach for Prediction of Heart Diseases," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-6, doi: 10.1109/CCAA.2018.8777531.
[6] C. Raju, E. Philipsy, S. Chacko, L. Padma Suresh and S. DeepaRajan, "A Survey on Predicting Heart Disease using Data Mining Techniques,"2018 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, 2018, pp. 253-255, doi: 10.1109/ICEDSS.2018.8544333.
[7] Himanshu Sharma and M A Rizvi, “Prediction of Heart Disease using Machine Learning Algorithms: A Survey” International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8 , IJRITCC August 2017.
[8] Senthilkumarmohan, chandrasegarthirumalai and GautamSrivastva, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” IEEE Access 2019.
[9] AmandeepKaur and JyotiArora,“Heart Diseases Prediction using Data Mining Techniques: A survey” International Journal of Advanced Research in Computer Science , IJARCS 2015-2019.

[10] M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools" International Journal of Scientific Research in Computer Science, Engineering and Information Technology ,IJSRCSEIT 2019.

[11] Santhana Krishnan J and Geetha S, "Prediction of Heart Disease using Machine Learning Algorithms" ICIICT, 2019.

[12]AditiGavhane, GouthamiKokkula, IshaPanday, Prof. KailashDevadkar, "Prediction of Heart Disease using Machine Learning", Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology(ICECA), 2018.

[13] Pahulpreet Singh Kohli and ShriyaArora, "Application of Machine Learning in Diseases Prediction", 4th International Conference on Computing Communication And Automation(ICCCA), 2018.

[14] M. Akhil, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia Technol., vol. 10, pp. 85–94, 2013.

[15] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J.Schmid,S.Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, "Internationalapplication of a new probability algorithm for the diagnosis of coronaryartery disease," The American journal of cardiology, vol. 64, no. 5, pp.304–310, 1989.

[16] B. Edmonds, "Using localised 'gossip' to structure distributed learning,"2005.

[17] FsdfsdfBayuAdhi Tama,1 Afriyan Firdaus,2 Rodiyatul FS, "Detectionof Type 2 Diabetes Mellitus with Data Mining Approach Using SupportVector Machine", Vol. 11, issue 3, pp. 12-23, 2008.

[18] Yu-Xuan Wang, QiHui Sun, Ting-Ying Chien, Po-Chun Huang, "UsingData Mining and Machine Learning Techniques for System DesignSpace Exploration and Automatized Optimization", Proceedings of the2017 IEEE International Conference on Applied System Innovation, vol.15, pp. 1079-1082, 2017.

[19] ZhiqiangGe, Zhihuan Song, Steven X. Ding, Biao Huang, "Data Miningand Analytics in the Process Industry: The Role of Machine Learning",2017 IEEE Translations and contentmining are permitted for academicresearch only, vol. 5, pp. 20590-20616, 2017.

[20] https://en.wikipedia.org/wiki/Support_vector_machine

[21] https://en.wikipedia.org/wiki/Decision_tree_learning

[22] https://en.wikipedia.org/wiki/Bayes27_theorem

[23] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[24] https://towardsdatascience.com/understanding-random-forest58381e0602d2

[25] https://archive.ics.uci.edu/ml/datasets/heart+disease