ISSN 2278-3091
Volume 11, No.3, May - June 2022
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse101132022.pdf
https://doi.org/10.30534/ijatcse/2022/101132022

IoT and Machine Learning approach for Early Heart disease Prediction & Diagnosis

Iqbal Ahmed[1], Swapnil Diganta Thakur[2]
[1]Dept. of Computer Science & Engineering, University of Chittagong, Bangladesh, iqbal.ahmed@cu.ac.bd
[21]Dept. of Computer Science & Engineering, University of Chittagong, Bangladesh, s.digu79@gmail.com

Received Date : April 18, 2022 Accepted Date : May 21, 2022 Published Date : June 06, 2022

ABSTRACT

Heart disease is one of the most prominent causes of deaths globally. Every year almost 17.9 million people lose their life due to heart disease which account for 31% of total deaths worldwide. Most of the time patients know about their heart disease after reaching a severe heart condition from where total recovery is impossible. However, if the Cardiovascular state is monitored regularly, heart disease can be detected at an early stage and early detection can prevent the severity of most heart diseases. In most cases patients do not feel any kind of pain when the cardiovascular diseases grow slowly. By the time someone feels uneasiness and pain, their heart condition gets seriously bad. Moreover, it is also not feasible for everyone to check up on their heart condition periodically by visiting a heart specialist. Our proposed system will work for the early detection of heart diseases using Machine learning classifiers and IoT technologies. Our system has two subsystems. First one is our trained machine learning model which will be implemented as a WebApi. Second one is our IoT setup with heartbeat sensors. Sensors will collect data from the user's body and send those to the machine learning model. Then, the model will predict the result about the user's heart condition and send it back to the IoT device. Model will classify the user's heart condition either as "Normal" or "Abnormal". Based on the result, the user should go to a cardiologist for a checkup. We have used the Heart Disease Dataset from UCI Machine Learning Repository. In addition, we trained seven machine learning algorithms after preprocessing the dataset. Further we will also build an IoT setup with sensors to communicate with the WebApi and complete our proposed system of predicting heart diseases.

Key words : Heart Disease, IoT, Machine Learning , UCI Heart Disease Dataset.

1. INTRODUCTION

Cardiovascular system is one of the important organs for human being, which is formed by heart arteries, vein, capillaries. Heart is the muscular organ which pumps blood through veins and arteries. The heart resides between lungs and the middle chamber of the chest. Heart is divided into two parts by muscular walls. The right portion of the heart accumulates blood, lacking from the body through veins and pumps that to the lungs. The left portion accumulates blood, full of oxygen from lungs and pumps it through arteries to the body. The pumping rhythm is maintained by the heart's electrical system consisting of pacemaker cells in the sinoatrial node [1]. Any circumstances affecting the heart's pumping rhythm is considered as heart disease. Generally, it is caused by blood-vessel troubles, hereditary defects, infections. According to WHO from 2000 till 2019 the number of deaths due to Ischaemic heart disease (Ischemia) rose from 2 million to 8.9 million [2]. According to their statistics 16% of the world's total mortality is caused by Ischaemic heart disease. Ischaemic heart disease has another name, known as coronary artery disease, is the result of narrowed heart arteries. Heart muscle suffers from less blood and oxygen due to the blockage in arteries which eventually lead to heart attack. Patients sometimes feel pain and discomfort due to Ischemia. But Ischemia without pain known as Silent Ischemia is more common nowadays. Patients may have a heart attack with no warning in such cases. Therefore, we have to go to heart-specialist for check-ups periodically to avoid Ischaemic heart disease. Besides Ischemia, Heart Arrhythmias, Pericardial Disease, Cardiomyopathy are also some cardiovascular diseases which can be detected by a cardiologist during check-up [3]. However, visiting a heart-specialist for a check-up is costly and visiting multiple times is not affordable for most of the people in developing countries. Therefore, our target is to notify people exactly when it is necessary to go for a check-up. We are building a system utilizing machine learning and IoT.

The progression in the field of Cloud and distributed computing huge computational tasks can be done at low cost [4]. Additionally, Machine Learning Algorithms need more computational power, it can utilize the processing power of Cloud Computing. Cloud also provides an ample amount of storage. Therefore, machine learning is getting popular to solve problems related to different sectors in recent days. Most Machine learning (ML) solve classification and regression problems. Internet of Things (IoT) is another popular technology nowadays. IoT technology achieves data from the environment through sensors, can process the data as most IoT devices include microprocessors and can send data to and receive from the internet [5, 6]. In recent years a good amount

of research has been published related to detection of heart disease utilizing machine learning and IoT [5, 6]. As a huge number of people have heart disease like heart ischemia, arrhythmia, dilated cardiomyopathy, strokes etc. datasets related to heart disease patients are generated for various research purposes for different sectors. Data Scientists and ML engineers process the raw data to make usable for research purposes because there might be many outliers, missing values due to human error [7, 8, 9]. With the processed dataset and various machine learning algorithms available in Python and R and other tools like WEKA, research is now easily accessible [7].

Our proposed system will work for the early detection of heart diseases using Machine learning classifiers and IoT technologies. Our system has two subsystems. First one is our trained machine learning model which will be implemented as a WebApi. Second one is our IoT setup with heartbeat sensors. Sensors will collect data from the user's body and send those to the machine learning model. Then, the model will predict the result about the user's heart condition and send it back to the IoT device. Model will classify the user's heart condition either as "Normal" or "Abnormal". Based on the result, the user should go to a cardiologist for a checkup.

2. LITERATURE REVIEW

In recent years a good amount of research has been published related to detection of heart disease utilizing machine learning algorithms and IoT based sensors.

Amin et al. and Ajit et al. [10, 11] developed a machine learning classifier with different predictive models for heart disease detection utilizing feature selection algorithms such as Relief, Lasso, mRMR. They used UCI Cleveland dataset and compared Logistic Regression, K-th nearest neighbor, Artificial Neural Network, Support Vector Machine, Naive Bayes, Decision Tree and Random Forest algorithm.

Senthilkumar et al. [12] extracted features from UCI dataset using Less Error Classifier and proposed a hybrid method combining Logistic Regression, Multivariate Adaptive Regression Splines (MARS) and Artificial Neural Networks (ANN). They achieved maximum accuracy with Hybrid Random Forest with a linear model (HRFLM).

Dr. A. A. Gurjar et al. [13] worked on a heart attack detection system by heartbeat sensing using IoT. Their system includes a pressure sensor, heartbeat sensor and ATMega microprocessor with a WiFi module. The IoT collected data were sent to cardiologists for monitoring.

C. Beaulah et al. [14] used ensemble techniques thus bagging and boosting for effective improvement of prediction accuracy. They combined Naive Bayes, C4.5(extension of ID3 algorithm), Bayes Net, Projective Adaptive Resonance Theory (PART), MultiLayer Perceptron (MLP) to build their ensemble classifier. The Cleveland heart dataset from UCI was used for training and testing by them.

Gowrishankar et al. [15] Developed an IoT based heart attack detecting system which measures human body temperature using LM35 sensor, pulse rate and use Arduino uno with wifi module to upload data on the web. This system sends this data on the web from where a specialist can make decisions about heart diseases.

3. METHODOLOGY

3.1 Brief View of the Proposed System

We will first build our machine learning model after training with publicly available datasets for heart disease detection. We are solving a classification problem which will detect whether a person is prone to heart disease or not based on their sex, age, blood pressure, heart rate and other medical experiment results. After finalizing the model, we will make a REST Api which will contain our machine learning model. Then we will build an IoT setup with NodeMCU/Arduino and sensors which can measure different metrics from a human body. Measured data will be of such type that matches with the attributes in the datasets we are using. We may not be able to get all the attributes via sensors which are present in our dataset. So, we will rank those attributes which are obtainable via available sensors compatible with our Microcontroller. In case, if all the important attributes are non-obtainable via sensor, we can just build a system with available IoT sensors to keep record of the heart condition of a person. In the next Figure 1 elaborates the overview of our proposed system.

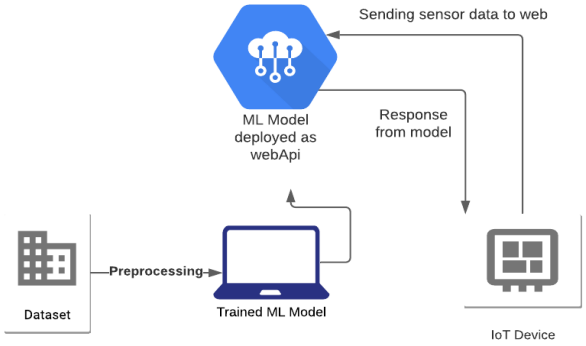

Figure 1: High level view of the proposed system

The following figure 2 represents the Heart rate and SPO2 monitor system using MAX30100 and NodeMCU ESP8266.

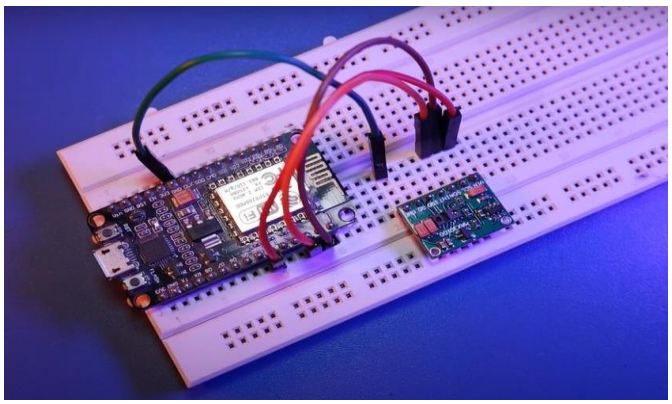

Figure 2: Heart rate and SPO2 monitoring system

3.2 Dataset Description

We primarily used the heart disease dataset from UCI. This dataset was made in 1998 and consists of 4 different datasets from Cleveland, Hungary, Switzerland and Long Beach V. Originally it has a total of 76 attributes including the attribute to be predicted. But a subset of only 14 attributes is used in all the published experiments. Again, due to missing values in more than one column, data from Switzerland, Hungary and Long Beach V dataset are not generally used. Only data from the Cleveland dataset is used by Machine Learning Researchers. This dataset has 303 rows in total. Next table 1 and table 2 is the short description of each column containing information of individuals submission.

Table 1: Dataset short description

Column Heading	Description	Value
age	Age in year. It's numerical data.	It Ranges from 29 to 77
Sex	Gender.	It's categorical data.
cp	Chest pain type.	It's categorical data.
trestbps	Resting blood pressure (in mmHg). It's numerical data.	It ranges from 94 to 200.
chol	Serum cholesterol in mg/dl. It's numerical data.	It ranges form 126-564.
fbs	Fasting blood sugar > 120 mg/dl.	It's categorical data.
restecg	Resting electrocardiographic value.	It's categorical data.
thalach	Maximum heart rate achieved. It's numerical data.	It ranges from 71 to 202.
exang	Exercise induced angina.	It's categorical data.
oldpeak	St depression induced by exercise relative to rest. It's numerical data.	It ranges from 0.0 to 6.2.
slope	Slope of the peak exercise ST segment.	It's categorical data.
ca	Number of major vessels colored by fluoroscopy.	It's categorical data.
thal	Thallium stress test.	It's categorical data.
target	Presence of heart disease.	It's categorical data.

Table 2: Possible values for different categorical attributes

Attributes	Possible Values
sex	male (1), female (0)
cp	typical angina (0), atypical angina (1), non-anginal pain (2), asymptomatic (3)
fbs	fasting blood sugar > 120 mg/dl (1), fasting blood sugar < 120 mg/dl (0)
restecg	normal (0), having ST-T wave abnormality (1), definite left ventricular hypertrophy by Estes' criteria (2)
Attributes	**Possible Values**
thalach	Maximum heart rate achieved. It's numerical data.
exang	yes (1), no (0)
slope	upsloping (0), flat (1), downsloping (2)
ca	Number of major vessels colored by fluoroscopy.
thal	normal (0), fixed defect (1), reversible defect (2)
target	no (0), yes (1)

3.3 Data Preprocessing

The accuracy of any machine learning model depends on the dataset it was trained by. The quality of data is crucial for an accurate prediction. Therefore, we have to do some essential operations on our dataset to make it perfect for training a model.

1) Removing Missing Values

ML algorithm can't get any insight from a missing value. If there are one or two missing values in a row, one possible solution is removing the row for better training. If attributes of other columns play an important role in decision making, we need to impute the missing values with appropriate placeholder as mean. But if the row contains missing values for a significant number of columns, we must remove the whole row.

2) Removing Outliers

Outliers cause faulty decisions by ML model. Outliers are uncommon patterns in a dataset which are the sign of rare cases in a statistical survey. They don't represent the normal trend for a population. Generally, values which are three standard deviations away from mean are considered as outliers. If the dataset has enough instances for training, then rows having outliers should be removed.

3) Encoding Categorical Data

Categorical data implies those attributes which can take a few numbers of possible values and generally are string values.

Machine learning modules utilize statistical formulas deep down and those formulas work best on numerical data. That is the reason we convert categorical values to numerical values for better performance. For further improvement we use a technique named "One Hot Encoding" which makes a new column for every unique value in categorical columns and each of the new columns contains binary values. One Hot Encoding is applied on columns with high cardinality.

4) Normalizing Data

In dataset different columns hold different kinds of data. Some of them can have larger ranges, some of them might have much smaller ranges. Some columns might have very large values compared with other columns. Machine Learning algorithms perform much better in a dataset where this variation among column values is minimum. To achieve this ML engineer suggests normalizing the data which will turn every value of each column a real number between 0 and 1.

The UCI Cleveland dataset which is being used by us doesn't contain any missing values. It has some outliers in around 15 rows but due to the small number of total instances we are not going to remove those now. There are seven categorical columns other than target column which are ***sex, cp, fbs, restecg, exang, ca, thal***. All of them already contain categories in numerical format. We applied One Hot Encoding on columns named ***cp, thal and slope***, which shows in next figure 3.

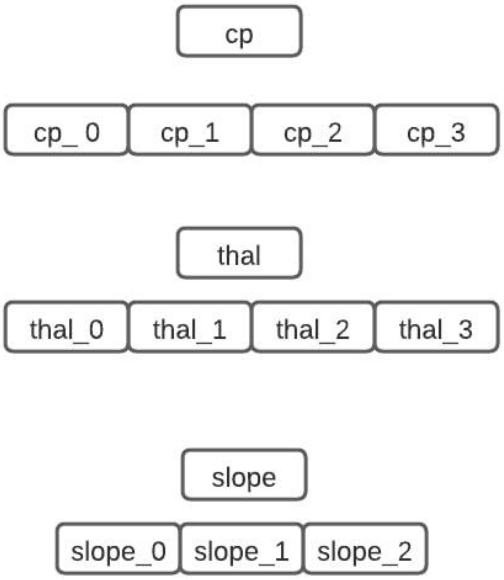

Figure 3: Encoding of Categorical Data

3.4 Dividing the dataset into feature and target

For this research we want to predict the possibility of having heart disease. The column named 'target' in our dataset is our target column whose value we will predict by our machine learning models. We will separate our target column from the rest of the other columns. The other columns are features which will cause the target values. In a single row of dataset, all the feature attributes will have effect of various degrees over the target attribute. In our coding we keep feature columns in a variable named X and our target columns in a variable name Y. Then we split both of X and Y for training and testing. Y_train contains the targets corresponding to X_train and Y_test contains the targets corresponding to X_test. In our experiment we used 75% data for training and 25% data for testing.

3.5 Building Classifier Model

We ran eight machine learning algorithms on the preprocessed dataset and got eight models. Our used machine learning algorithms are, *Naïve Bayes, Decision tree, Logictic Regression, KNN, Random Forest, SVM, XGBoost and CatBoost.*

3.6 Applying Expandable AI

A trained machine learning model is in some sense like a black-box model. End users and in some cases even developers of a model do not have an idea about exactly which attributes of the dataset contributed most to the decision-making process of the model. Explainable Artificial Intelligence (XAI) describes the behavior of a model by showing the impacts of different attributes of the dataset and giving stats like fairness of the model. XAI helps us to both understand and interpret the predictions made by a machine learning model. We have used a game theory based XAI named SHAP (SHapely Additive exPlanations) to visualize the behavior of the model which gives the best accuracy score. It gives us the most important attributes based on which we can design our IoT setup and select which sensors are needed most for our proposed system.

4. EXPERIMENT AND RESULTS

We used eight machine learning algorithms thus built eight classification models. To check the performance of our models we will use each classifier's confusion matrix.

4.1 Confusion Matrix

Our models are binary classifiers. Therefore, each of our models (total 8 models) will have confusion matrices with 2 rows and 2 columns.

	Predicted as True	**Predicted as False**
Labeled as True	True Positive(TP)	False Negative(FN)
Labeled as False	False Positive(FP)	True Negative(TN)

Precision Score: The model's precision score reflects its ability to correctly forecast the positives out of all the positive predictions it has made.

Mathematically**, Precision Score = TP / (FP + TP)** (1)

Recall Score: The model's ability to correctly forecast positives out of real positives are represented by the recall score.

Mathematically**, Recall Score = TP / (FN + TP)** (2)

Accuracy Score: The model's accuracy is defined by the true positives and true negatives as a percentage of all positive and negative observations.

Mathematically,

Accuracy Score = (TP + TN)/ (TP + FN + TN + FP) (3)

F1 Score: It is represented as the harmonic mean of precision and recall score.

Mathematically, **F1 Score = (2* Precision Score* Recall Score)/ (Precision Score + Recall Score)** (4)

4.2 Training Parameters

The following table 3 represents our training parameters for each model and in our experiment, we used 75% data for training and 25% data for testing.

Table 3: Training parameters

Number of instances/rows	303
Number of instances with missing values	0
Total number of columns	14
Number of instances in training dataset	227
Number of instances in test dataset	76

4.3 Discussion

The following table 4 and figure 4 and 5 elaborately represents the complete picture of our experiments with Heart Disease Dataset from UCI Machine Learning Repository. All our eight model's accuracy, precision, recall and f1 score have been calculated using the equation (1) to equation (4).

Table 4: All Scores of the models

Model Name	*Accuracy Score*	*Precision Score*	*Recall Score*	*F1 Score*
Naïve Bayes	61.84%	83.33%	36.58%	50.85%
Decision Tree	76.32%	75.56%	82.93%	79.07%
XGBoost	85.53%	85.71%	87.80%	86.75%
KNN	85.53%	91.30%	86.42%	85.37%
Random Forest	85.53%	84.09%	90.24%	87.06%
Logistic Regression	86.84%	89.74%	85.37%	87.50%
SVM	86.84%	89.74%	85.37%	87.50%
CatBoost	92.11%	92.68%	92.68%	92.68%

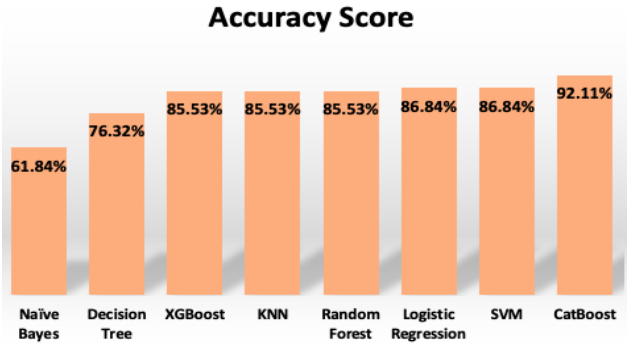

Figure 4: Comparison of accuracies of all the models

From the above figure, it is clearly shows that our **CatBoost Model** outperformed all other eight models with the accuracy of 92.11% on UCI dataset for heart disease predication and diagnosis.

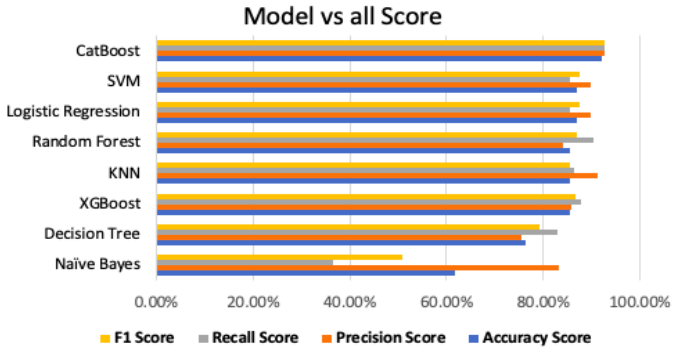

Figure 5: Comparison of all score vs model

The next figure 6 also finds out generated feature rank of CatBoost model. After applying SHAP on our CatBoost model we can see that ***Chest Pain Type(cp), Thallium Stress Test result (thal), ST Depression result (oldpeak), Age of a patient*** are the major detectors of his/her heart disease. The top three attributes are not measurable via available IoT sensors.

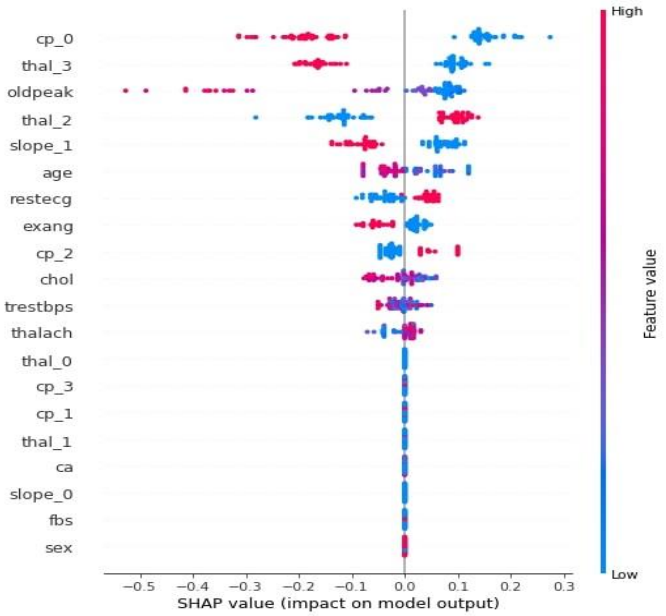

Figure 6: SHAP generated feature-rank for CatBoost

5. CONCLUSION AND FUTURE PLAN

In this research we have compared eight Machine Learning Models trained with UCI Cleveland heart disease dataset. We have found that amongst these eight algorithms, **CatBoost** performs the best with 92.11% accuracy, when using 75% data as train data and 25% data as test data. We believe, our work for the early detection of heart diseases using Machine learning classifiers and IoT technologies help the people for finding out initial signs and symptoms of their heart problems. However, after finding the most important columns of dataset, we can see that these are not currently measurable with IoT sensors. In future, we need a more suitable dataset where column values can be measured with sensors. In addition, we have a plan to improve our machine learning model by applying Ensemble Methods.

ACKNOWLEDGEMENT

We thanks to those researchers who lead us into this field and thanks to our Department of Computer Science and Engineering, University of Chittagong for helping and encouraging to carry out this research.

REFERENCES

1. A. H. Kashou, H. Basit, and L. Chhabra. ***Physiology, Sinoatrial Node***, Book Chapter, *StatPearls Publishing*, USA, October 2021.
2. M. Almatrooshi, M. Alzaabi, R. S. Al Darmaki, and S. Lootah. ***Global Epidemiology of Ischemic Heart Disease: Results from the Global Burden of Disease Study.*** Cureus, Vo.12(7),e9349, 2020.
3. M. H. Khandaker, R. E. Espinosa, R. A. Nishimura, L. J. Sinak, S. N. Hayes, R. M. Melduni, and J. K. Oh. ***Pericardial disease: diagnosis and management***. Mayo Clinic proceedings, 85(6), 572–593, 2010.
4. A. B. Sururah, O. O. Lukumon, O. A. Olugbenga, B. Muhammad, D. D. Juan Manuel, A. A. Lukman, O. Anuoluwapo Ajayi, and A. O. Hakeem. ***Cloud computing in construction industry: Use cases, benefits and challenges***, Automation in Construction, Volume 122, 2021.
5. B. Sana Abdelaziz, and I. A. Johnson. ***A Review of Identity Methods of Internet of Things (IOT)***. Advances in Internet of Things, Vol.11, 2021.
6. S. Ashif, Md. R. Islam, and M. H. Imam. ***Heart Disease Detection by UsingMachine Learning Algorithms and aReal-Time Cardiovascular Health Moni-toring System.*** World Journal of Engineering and Technology, 6, pp.854-873, 2018.
7. M. K. Gawali, and C. Rambabu. ***IoT Model for Heart Disease Detection Using Machine Learning (ML) Techniques.***Techno-Societal 2021. Springer, Cham.
8. I. H. Sarker. ***Machine Learning: Algorithms, Real-World Applications and Research Directions***. SN Computer Science, Vol. 2, 2021.
9. G. Nguyen, S. Dlugolinsky, M. Bobák *et al.* ***Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey.*** Artificial Intelligence Review 52, pp. 77–124, 2019.
10. A. U. Haq, , P. Li Jian, H. M. Muhammad, N. Shah, and S. Ruinan. ***A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms.*** Mobile Information Systems, 2018.
11. G. D. Santhosh, Dr. K. Jayanthi, Dr. A. P. Ajith, *et al*. ***A Comprehensive survey on Heart Disease Prediction using Machine Intelligence***, Research Square, 06 July 2021.
12. S. Mohan, T. Chandrasegar, and S. Gautam. ***Effective heart disease prediction using hybrid machine learning techniques.*** IEEE access 7 (2019): 81542-81554.
13. A. A. Gurjar, and A. S. Neha. ***Heart attack detection by heartbeat sensing using Internet of Things: IoT***. Heart 5, no. 03, 2018.
14. C. Latha, C. Beulah, and S. J. Arolin. ***Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques.*** Informatics in Medicine Unlocked 16 (2019): 100203.
15. S. Gowrishankar, M. Y. Prachita, and A. Prakash. ***IoT based heart attack detection, heart rate and temperature monitor***. International Journal of Computer Applications, Vol.170, No. 5, pp. 26-30, 2017.