Comparative study of different algorithms in Disease Prediction

Prof. Shivani

*Department of Computer Science*

*(of Affiliation)*

*K.I.E.T Group Of Institutions*

*(of Affiliation)*

Ghaziabad, India

[shivani.cs@kiet.edu](mailto:shivani.cs@kiet.edu)

Deepak Goswami

*Department of Computer Science*

*(of Affiliation)*

*K.I.E.T Group Of Institutions*

*(of Affiliation)*

Ghaziabad, India

[deepak.1923co1010@kiet.edu](mailto:deepak.1923co1010@kiet.edu)

Mohd Zubair

*Department of Computer Science*

*organization*

(of Affiliation)

*K.I.E.T Group Of Institutions*

*(of Affiliation)*

Ghaziabad, India

[mohd.1923co1027@kiet.edu](mailto:mohd.1923co1027@kiet.edu)

Lovely Mangal

*Department of Computer Science*

*(of Affiliation)*

*K.I.E.T Group Of Institutions*

*(of Affiliation)*

Ghaziabad, India

[lovely.1923co1098@kiet.edu](mailto:Lovely.1923co1098@kiet.edu)

***Abstract*—With the technological advancement in the field of medical health care we need a best possible health care system that can predict the disease only on the basis of symptoms .In this study we applied various machine learning algorithm including KNN , support vector machine , decision tree , Naïve Bayes and logistic regression on various disease dataset to find the most accurate algorithm on particular disease .The objective of this study is to provide fast and efficient machine learning algorithm that will help doctors to choose the best suited algorithm for the particular disease.**

***.Keyword: Logistic Regression ,SVM, Decision Tree and KNN.***

# Ⅰ. Introduction

Nowadays health issues are quite common. We need a model which can predict the disease accurately on the basis of symptoms and help in early treatment and saves vital organs from damaging . In our research we are applying different machine learning algorithm on the disease of liver, lung cancer, heart , diabetes ,brain stroke .The data sets used in this model are taken from an open source and the model is trained in such a manner that it comes up with the best algorithms among different machine learning technique to prognosticate the health state of patient accurately.

*A .Heart Disease*

In a recent survey conducted by world Health organization it has been found that disease related to heart is responsible for approximate 20 million decease every year . These decease can be reduced to half by using a proposed system which can  be used in early detection of heart disease . According to a survey of World Health Organisation(WHO), it is found that India have lost more than $230 billion in last 10 years due to cardiac attacks [2].The dataset for heart disease consists of input such as sex(gender), Constrictive pericarditis(cp), resting blood pressure(trestbps),fasting blood pressure(fbs) etc. which is taken in a model , and the most accurate algorithm is used in prediction of heart disease.

*B*. *Liver Disease*

Liver is a vital organ of our body. It performs many biological functions and it also synthesize protein that is necessary for body growth. It helps our body tissues to keep alive and also provides nutrition and energy to them[2]. Nearly 20 lakh people loose their life due to this disease[3].

*C.Lung cancer*

Cancer can be cured only when it is diagnosed in an early stage . cancer cells grows at a very fast rate and moved to various body parts . Human body is made up of cell they grows and multiply with a process of cell divison. . There are 15 different type of cancer, among these lung cancer is the most deadly having a death ratio of 46.7 per 100,000 persons. Long term tobacco addiction is the main cause of developing lung cancer [4]. Small cell lung cancer is associated with smoking and air pollution evolve more rapidly by becoming large tumors that spread throughout the body[23].

1. *Diabetes*

Diabetes is the imbalance of glucose in our body[1]. It is one of the most common disease nowadays mostly in the developed countries[22]. If the blood glucose is more than the normal glucose level then a person has a high chances of developing diabetes[5].By 2035 nearly 592 million will suffer diabetes[6].According to the survey in 2012 it is leading cause of death in women[7].

*E. Brain Stroke*

A stroke is a medical emergency it is a condition in which blood flow stops to any part of the brain that cause brain cell incapable to function . In an estimation around 70 lakh people in America above the age of 20 year have suffered from a stroke[8].

# Ⅱ. Literature Survey

Nisha Gupta [9] proposed a system based on machine learning techniques in prediction of heart disease. Decision tree and SVM are applied over the dataset of heart patient and accuracy achieved by SVM is highest of 83% followed by decision tree 79%.

|  |  |
| --- | --- |
| ***Algorithm Used*** | ***Accuracy*** |
| SVM | 83% |
| Decision Tree | 79% |

Fig.1. Accuracy comparison[9]

Nikita Ahure [10] presented a model to predict the heart disease and to find the efficiency of algorithm by using two machine learning techniques named as decision tree and random forest in which it is found that random forest is more accurate with the efficiency of 75%.

Asha Rajkumar [11] worked on the prediction of heart disease using machine learning classification techniques. Naïve Bayes classifier and KNN are used to predict the heart disease and to find the accurate algorithm for the disease. Among the two Naïve Bayes showed lower error ratio.

|  |  |
| --- | --- |
| ***Algorithm Used*** | **Accuracy** |
| KNN | 45.6% |
| Naïve Bayes | 52.3% |

Fig.2. Accuracy comparison[11]

Harshit Jindal [12] proposed a prediction model to predict heart disease based on medical attribute like gender age, fasting sugar level etc. These attributes are trained under three algorithm named as Logistic Regression , KNN and Random Forest classifier. KNN is found to be most accurate algorithm with the accuracy of 88.52%.

AKM Sajjadur Rehman [13] have worked to predict the liver disease at a very cost effective method by applying various machine learning algorithm that are Naïve Bayes , SVM ,KNN , Decision Tree and Random Forest over the disease dataset. Among all the other techniques Logistic Regression found to be the most accurate with the accuracy of 75%

|  |  |
| --- | --- |
| ***Algorithm Used*** | ***Accuracy*** |
| KNN | 62% |
| Naïve Bayes | 53% |
| SVM | 64% |
| Decision Tree | 69% |
| Logistic Regression | 75% |
| Random Forest | 74% |

Fig. 3. Accuracy comparison[13]

Dr. S Vijay Rani [2] applied a famous classification algorithm to predict the liver disease. Naïve Bayes and SVM classification techniques are applied and these are compared on the basis of classification accuracy in which SVM is considered as the better classifier to predict the liver disease.

|  |  |  |
| --- | --- | --- |
| ***Algorithm Used*** | ***Accuracy*** | ***Precision*** |
| Naïve Bayes | 61.28% | 0.552 |
| SVM | 79.66% | 0.786 |

Fig.4. Accuracy comparison[2]

## Anju Gulia [14] in their research work have done the prediction of liver disease by implementing a hybrid model.

They improve the accuracy of dataset in three steps. After using various algorithms SVM is found to be highly accurate for the prediction of liver disease without featured selection.

Dakhaz Mustafa Abdullah [15] modelled and designed a system to predict the lung cancer at a very initial stage so that many lives can be saved. They inspect the accuracy ratio of SVM, KNN and Convolutional Neural Network where SVM gives the most accurate result with 95.56% followed by CNN 92.11% and KNN 88.40%.

Muhammad Imran Faisal [16] used several classification algorithms such as decision Tree, Multi Layer perceptron, Neural Network ,SVM ,Gradient Boosted Tree and Naïve Bayes in detection of Lung cancer. UCI repository is used for the dataset where Gradient Boosted tree was found to be most accurate by achieving 90% precision.

Tarig Mohamed Ahmed [17] proposed a system for the classification of diabetes disease. Three algorithms Naïve Bayes, Logistic Regression and J48 were used and by using WEKA application the system was implemented. Experimental result shows the Logistic regression with the best efficiency of 74.8%.

Md. Faisal Farque [18] employed some famous machine learning algorithms in prediction of diabetes disease by applying SVM , KNN, Decision Tree and KNN where Decision Tree attain highest accuracy of 73%.

Olta Llaha [19] have used a data mining and machine learning approach over the collected data of diabetes. Method such as SVM , Decision Tree, Logistic Regression , Naïve Bayes are evaluated to find the best performed method on the dataset. Obtained result shows that Naïve Bayes is the most accurate.

|  |  |
| --- | --- |
| ***Algorithm Used*** | ***Accuracy*** |
| Decision Tree | 63.75% |
| SVM | 64.7% |
| Logistic Regression | 68.02% |
| Naïve Bayes | 68.4% |

Fig.5. Accuracy comparison[19]

In research work [20] stroke prediction was made by applying many algorithm like KNN, Logistic Regression Decision Tree, Random Forest ,Naïve Bayes and SVM. Naïve Bayes provides the best result with the efficiency of 82.3%

|  |  |
| --- | --- |
| ***Algorithm Used*** | ***Accuracy*** |
| KNN | 80.4% |
| Logistic Regression | 77.6% |
| Naïve Bayes | 82.3% |
| Random Forest | 72.7% |
| SVM | 81.1% |
| Decision Tree | 77.6% |

Fig.6. Accuracy comparison[20]

In research paper [21] model was trained on famous classification algorithm on medical recorded dataset for brain stroke prediction using Neural Network ,Decision Tree and Random Forest. The performance of feed forward Neural Network was the highest with the accuracy of 75.62%

|  |  |
| --- | --- |
| ***Algorithm used*** | ***Accuracy*** |
| Decision Tree | 74.3% |
| Random Forest | 77.6% |
| Neural Network | 75.02% |

Fig.7. Accuracy comparison[21]

# Ⅲ. METHODOLOGY

We compared various machine learning algorithm by applying over various disease data sets and designed a health prediction system. Algorithms we used are Naïve Bayes, Decision Tree , KNN, Logistic Regression.

1. *Selection of data***:** Data is selected from open repository that are Kaggle, UCI etc. The data set contain medical attributes of all the patients.
2. *Data Gathering***:** Data gathering is the process to gather all the data to form a single unit of dataset i.e. to make the data ready for the input.
3. *Preprocessing on data***:** It is the first step to process the data and to make the data ready for the next step.

In this data is filtered and cleaned .This process eliminate the duplicacy in the data that helps to make our model accurate. It also ensures to eliminate the values that are caused due to human error.

1. *Matplotlib***:**It is used in comparison and representation of model.
2. *Numpy* **:** It is used to form the stack of machine learning.
3. *Pandas***:** It Provide support to Numpy and used to clean and analyze the data.

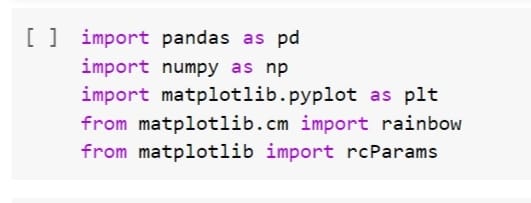


Fig .8. Libraries used

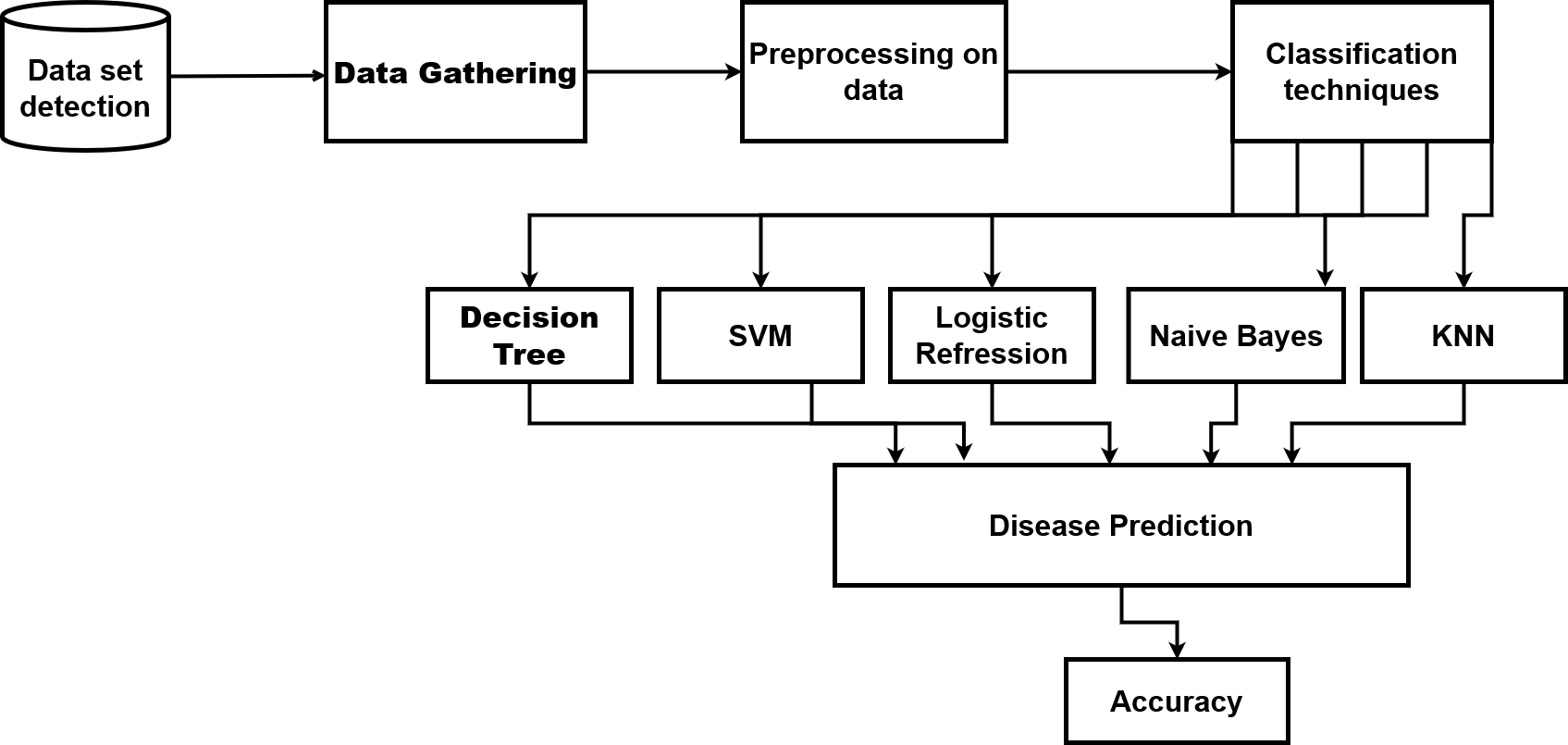


Fig. 9. Proposed Model

*d*)*Sklearn***:** It is the robust library used to support machine learning techniques like SVM , KNN and random forest.

 Fig.10. Import Algorithms

Ⅳ. APPLIED ALGORITHMS

Many algorithms are used in this model to find the best suited algorithm in particular disease prediction.

1. *SVM***:** Support Vector Machine is a model of supervised learning. It is a learning model widely used in the field of medical health care. It use kernel to transform the data into multidimensional space and also normalize the attributes[19].
2. *K- Neigbors*: Classifier: It is a supervised learning algorithm. It stores and allocate data point on the basis of similarity. As the name suggest it is mostly used in the classification of problem. It comes under the category of non-parametric algorithm.
3. *Naïve Bayes*: It is a group of classifier algorithm. It is based on Bayes theorem and is used in the wide variety of classification task.. Health care professional used this learning approach to predict the risk of disease.

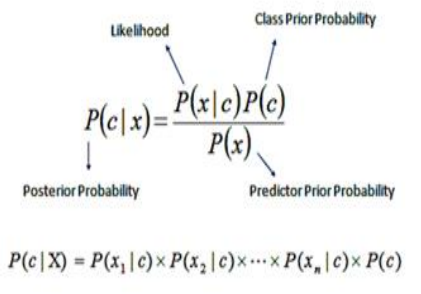


Fig 11: Naive Bayes

1. *Logistic Regression*: It is a supervised learning classification technique. In this target Output should be given with the input data and Output variable is in binary form only either in 0 or 1, true or false, yes or no etc.
2. *Decision Tree*: It is the easiest and popular classifier technique. It belongs to the family of supervised learning method and can be used for regression and classification problem. It falls in the category of non parametric algorithm.

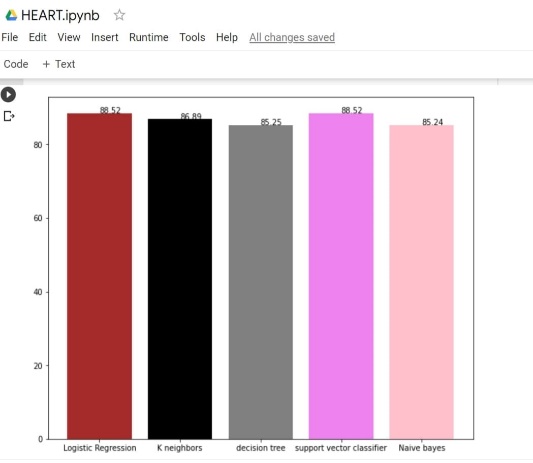


Fig.12. Accuracy of Heart Disease

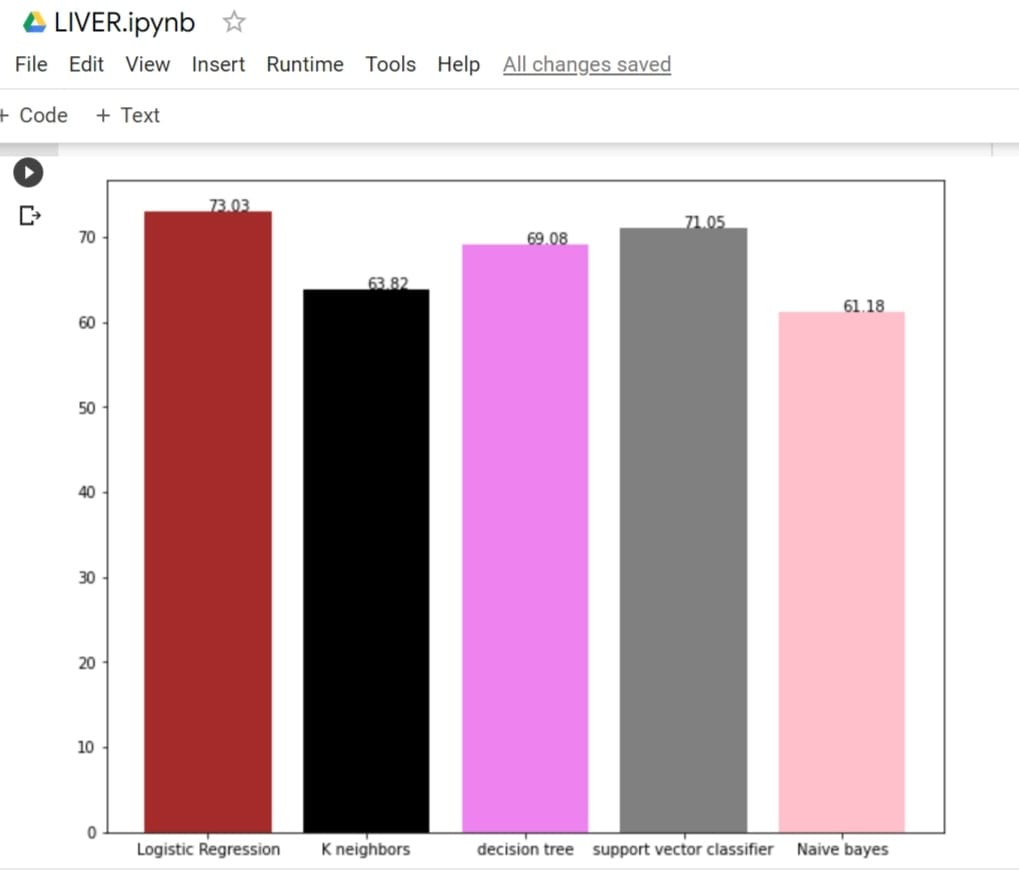


Fig.13 . Accuracy of Liver Disease

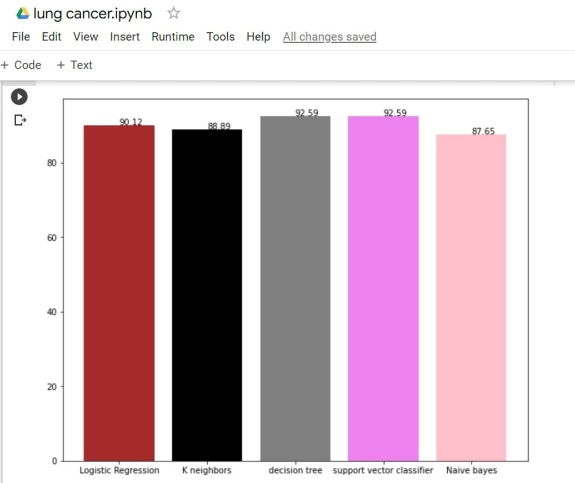


Fig.14. Accuracy of Lung Cancer

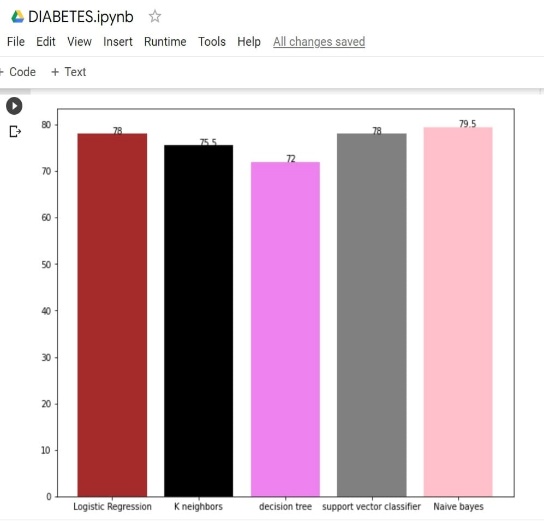


Fig.15. Accuracy of Diabetes

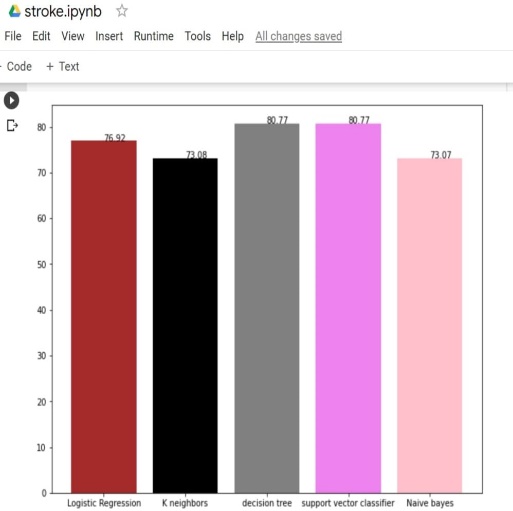


Fig.16 : Accuracy of Brain Stroke

# Ⅴ. RESULT

The result acquired after applying different algorithm in this “ Health Prediction Model” on various disease dataset we arrived at the conclusion that decision tree and SVM are best suited for Lung cancer prediction with accuracy of 92.59% . Logistic Regression and SVM are best suited for Heart disease with accuracy of 88.52%. Decision Tree and SVM are best suited for brain stroke with accuracy of 80.77%. Logistic Regression is best suited for Liver disease prediction with accuracy of 73.03% . Naive Bayes is best suited for Diabetes with accuracy of 79.5%.

For better understanding refer fig 17

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | ***SVM*** | ***Decision Tree*** | ***Logistic Regression*** | ***Naïve Bayes*** | ***K-***  ***NN*** |
| Lung Cancer | 92.59% | 92.59  % | 90.12% | 87.65% | 88.89  % |
| Heart disease | 88.52% | 85.25% | 88.52% | 85.24% | 86.89  % |
| Brain Stroke | 80.77% | 80.77% | 76.92% | 73.07% | 73.08  % |
| Liver | 71.05% | 69.08% | 73.03% | 61.18% | 63.82  % |
| Diabetes | 78% | 72% | 78% | 79.5% | 75.5  % |

Fig.17. Accuracy Table

# Ⅵ. CONCLUSION

Machine learning technology is very useful in the development of medical field .It can make the diagnose of disease at a very early stage that will help patients to avail the facility of disease detection at a very low cost. From this we can conclude that there is a wide use of Machine Learning in the field of medical science. Each above mentioned Machine Learning algorithm give us general idea about the algorithm to be used in particular disease prediction.

# REFERENCES

[1.] Shubham Salunke,  Shubham Rajiwade,  Deepak Yadav,  S.K.Sabnis,   "SMART HEALTH PREDICTION SYSTEM USING MACHINE LEARNING", IJRAR - International Journal of Research and Analytical Reviews (IJRAR), E-ISSN 2348-1269, P- ISSN 2349-5138, Volume.7, Issue 1, Page No pp.483-488, March 2020,

[2.] Mohan, Vijayarani. (2015). Liver Disease Prediction using SVM and Naïve Bayes Algorithms.

[3.] Rahman, A. K. M. & Shamrat, F.M. & Tasnim, Zarrin & Roy, Joy & Hossain, Syed. (2019). A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms. 8. 419-422.

[4.] Abdullah, Dakhaz & Mohsin Abdulazeez, Adnan & Sallow, Amira. (2021). Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques. Qubahan Academic Journal. 1. 141-149. 10.48161/qaj.v1n2a58.

[5.] Kishan Patel, Manu Nair, Subham Phansekar, “Diabetes prediction using Machine Learning”, 2021 IJSER, Volume 12, issue 3, March 2021.

[6.] Rani, KM. (2020). Diabetes Prediction Using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 294-305. 10.32628/CSEIT206463

[7.] Mahboob Alam, Talha & Iqbal, Muahammad & Ali, Yasir & Wahab, Abdul & Ijaz, Safdar & Baig, Talha & Hussain, Ayaz & Malik, Muhammad & Mehdi, Muhammad & Ibrar, Salman & Abbas, Zunish. (2019). A model for early prediction of diabetes. Informatics in Medicine Unlocked. 16. 100204. 10.1016/j.imu.2019.100204.

[8.] Alanazi EM, Abdou A, Luo J. Predicting Risk of Stroke From Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models. JMIR Form Res. 2021 Dec 2;5(12):e23440. doi: 10.2196/23440. PMID: 34860663; PMCID: PMC8686476.

[9.] Gupta, Nisha & Dharmale, Gulbakshee & Parmar, Darshana. (2021). HEART DISEASE PREDICTION USING MACHINE LEARNING. 10.13140/RG.2.2.16604.92800.

[10.] Rindhe, Baban & Ahire, Nikita & Patil, Rupali & Gagare, Shweta & Darade, Manisha. (2021). Heart Disease Prediction Using Machine Learning. International Journal of Advanced Research in Science, Communication and Technology. 267-276. 10.48175/IJARSCT-1131.

[11.] Rajkumar, Asha & Reena, Mrs. (2009). Diagnosis Of Heart Disease Using Datamining Algorithm. Global J Comput Sci Technol. 10.

[12.] Jindal, Harshit & Agrawal, Sarthak & Khera, Rishabh & Jain, Rachna & Nagrath, Preeti. (2021). Heart disease prediction using machine learning algorithms. IOP Conference Series: Materials Science and Engineering. 1022. 012072. 10.1088/1757-899X/1022/1/012072.

[13.] Rahman, A. K. M. & Shamrat, F.M. & Tasnim, Zarrin & Roy, Joy & Hossain, Syed. (2019). A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms. 8. 419-422.

[14.] Gulia, Anju, Rajan Vohra and P. Ithaya Rani. “Liver Patient Classification Using Intelligent Techniques.” (2014).

[15.] Abdullah, Dakhaz & Mohsin Abdulazeez, Adnan & Sallow, Amira. (2021). Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques. Qubahan Academic Journal. 1. 141-149. 10.48161/qaj.v1n2a58.

[16.] Faisal, Muhammad & Bashir, Saba & Khan, Zain & Khan, Farhan. (2018). An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer. 10.1109/ICEEST.2018.8643311.

[17.] Ahmed, Tarig Mohamed. “USING DATA MINING TO DEVELOP MODEL FOR CLASSIFYING DIABETIC PATIENT CONTROL LEVEL BASED ON HISTORICAL MEDICAL RECORDS.” (2016).

[18.] Faruque, Md & , Asaduzzaman & Sarker, Iqbal. (2019). Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus. 1-4. 10.1109/ECACE.2019.8679365.

[19.] Llaha, Olta & Rista, Amarildo. (2021). Prediction and Detection of Diabetes using Machine Learning.

[20.] Sailasya, Gangavarapu & Kumari, Gorli. (2021). Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. International Journal of Advanced Computer Science and Applications. 12. 10.14569/IJACSA.2021.0120662.

[21.] Nwosu, Chidozie & Dev, Soumyabrata & Bhardwaj, Peru & Veeravalli, Bharadwaj & John, Deepu. (2019). Predicting Stroke from Electronic Health Records. Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference. 2019. 5704-5707. 10.1109/EMBC.2019.8857234.

[22.] Juneja, Abhinav, et al. "Predicting diabetes mellitus with machine learning techniques using multi-criteria decision making." International Journal of Information Retrieval Research (IJIRR) 11.2 (2021): 38-52.

[23.] Juneja, Sapna, et al. "An approach for thoracic syndrome classification with convolutional neural networks." Computational and Mathematical Methods in Medicine 2021 (2021).