iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** III **Month of publication:** March 2023

DOI: https://doi.org/10.22214/ijraset.2023.49644

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

Multiple Disease Prediction System

Tanmay Ture[1], Amol Sawant[2], Rohan Singh[3], Prof. Chetna Patil[4]
SSJCOE, Dombivli (E)

Abstract: ***There are number of hospitals in the world with advanced diagnostic equipment. But although having this equipment some patients cannot get proper treatments and may suffer to death. Main reason behind this is time, our medical systems lack time and it is not easy for them to manage time. With help of machine learning techniques we created project that identifies patients with major diseases like Heart disease, Kidney disease and Diabetes disease at early stage so that proper treatments can be given to them. We collected three datasets for three models from Kaggle [1], analyzed[2]them, cleaned them and choose best algorithm [3] for each dataset. We achieved 98.52% accuracy on heart disease prediction model [4], 98.73% accuracy on kidney disease prediction model [5], 80.55% accuracy on diabetes disease prediction model [6]. For all three models we used Random Forest algorithm [7] .At the end we created web application using Flask [8] for easy user interaction.***
Keywords: Machine Learning, Data Analysis, Data Science, Data Visualization

I. INTRODUCTION

Many times, we see that patients lose their life because of not getting treatment on time. Healthcare industries lack time and they cannot determine which patient they should treat first. But on the other hand, healthcare industries generate huge amount of data regarding patients health. High level of insights can be drawn from this data. So, by using this data and advanced machine learning techniques we have decided to come up with project 'Multiple Disease Prediction System'. Our project is combination of three machine learning models that are going to identify patients having heart disease, kidney disease and diabetes at early stage, so that the patients having risk to particular disease will get treatment first. For our project we firstly understood problem statement and determined what type of data will be required for our project. We collected three different datasets for our three machine learning models from Kaggle. After collecting data, we analyzed data properly and visualized it for better understanding. Then we cleaned the data by imputing null values, encoding categorical features. Next step was we split dataset intro training and testing set such that we used 80% data for training machine learning model and remaining 20% data we used for testing our machine learning model. After that we tried several classification algorithms like Logistic regression, Random Forest classifier, Support Vector Machine classifier, XGBOOST classifier on all three datasets. Out of these algorithms we found that Random Forest classification algorithm was performing better than others for all three datasets. Using Random Forest algorithm we got 98.52% testing accuracy on heart disease dataset ,98.73% testing accuracy on kidney disease dataset and 80.55% testing accuracy on diabetes dataset. So, we dumped the model using Python's Pickle library and created web application using Python's Flask framework for easy user interaction. Also, in order to improve the project, we created dashboards of visualization for all three datasets using Tableau.

II. LITERATURE SURVEY

1) As heart plays an important role in living organisms. So, the main aim of this paper is diagnosis and prediction of heart related disease should be perfect and correct because it is very crucial which can cause death cases related to heart. They collected dataset from UCI machine learning repository. Used 13 features for prediction. Using Hybrid Random Forest with Linear model algorithm achieved 88% accuracy. In this paper they have use complex algorithm for predicting the heart disease.

2) Chronic Kidney Disease (CKD) implies that the human kidneys are harmed and unable to blood filter in the manner which they should. They employ experiential analysis of ML techniques for classifying the kidney patient dataset as CKD or NOTCKD. They used Composite Hypercube on Iterated Random Projection (CHIRP) algorithm, and it performs well in terms of diminishing error rates and improving accuracy. They achieve 99% accuracy. They have used 24 features which are high in number can be leads to overfitting. And also, they have use complex algorithm which is hard to implement.

3) Diabetes is one of the dangerous diseases in the world. In this paper they have used machine learning techniques to find out diabetes disease. Their aim of this analysis was to invent a system that can help the patient to detect the diabetes disease of the patient with accurate results. Several Machine Learning algorithm they have try to achieve maximum accuracy finally they decide to go with Ensembling classifier AdBoost and XGBoost which give 95% accuracy. They have use PIMA dataset which contain only information about female patient. So, their model will not suitable for the male patient to predict the diabetes disease.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

III. PROBLEM DEFINITION AND OBJECTIVES

A. *Problem Definition*

Modern healthcare systems are fulfilled with latest and effective diagnostic systems then also a lot of patients suffer from death due to lack of treatment on time. One thing that healthcare systems lack is time so they cannot determine which patient should be treated first. As a result, a needy patient could not get treatment on time which may cost his life too.

B. *Objectives*

1) Support healthcare systems
2) Reduce workload of healthcare systems and save their time
3) Identify patients having high risk to particular disease at an early stage
4) Build Relationships
5) Effectively predict if patient have chances of developing particular disease

IV. METHODOLOGY

A. *Building Machine Learning Models*

First, we understand the problem definition of our project to build a require and right machine learning models. Then we collected data from various open data sources like Kaggle, UCI Machine Learning Repository, etc. Quality and quantity of data is very important as it affect to our model. After that we have done the data preprocessing to ensure the that the collected data is in right format or not. Right after we analyze the available data to remove duplicate and deal with the missing values. We also check for the outliers. We visualize the data to find the relationship between the variable. Here we draw some useful insights and handle the skewness. Analyze data divided into training and testing dataset. We use 80% of our data for training and 20% of data for testing. This is most important step in building robust and accurate machine learning model. We have tried several machine learning algorithms to get best result as outcomes. We decide to go with Random Forest Algorithm because it gives us maximum accuracy for all of our machine learning model. And it is easy to implement as well.
Here our all models were ready for the prediction of diseases.

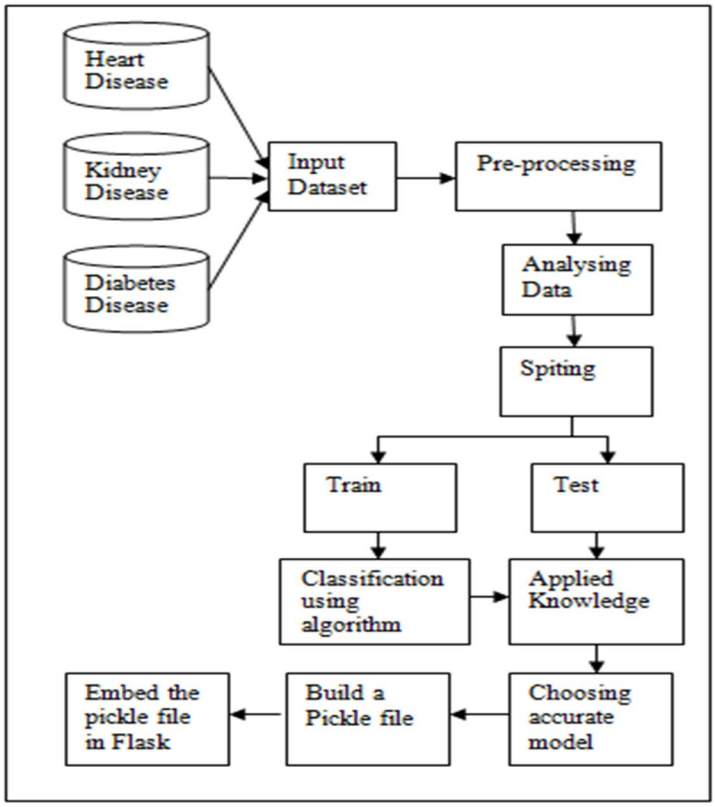


Fig. 1 Block Diagram

B. Deploying ML Models in Flask

After Creating the ML models for diseases, we deploy it in Flask framework of python language. We use Pickle module for that.

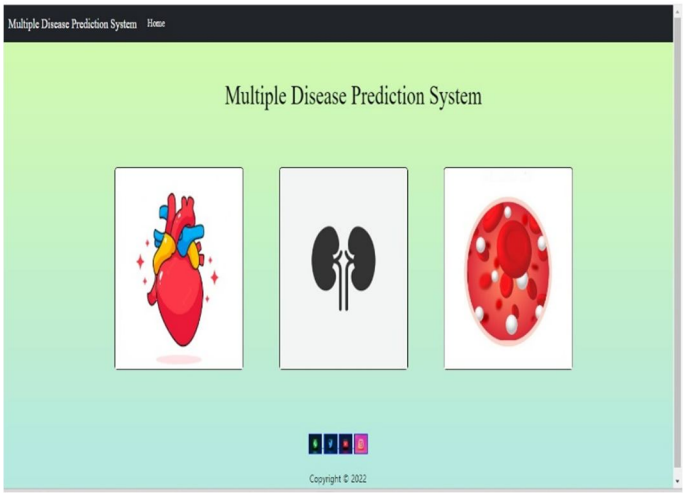


Fig. 2 Graphical User Interface

1) Random Forest

Random Forest working is possible in two phases, first is to create the random forest by merging N decision tree, and second is making prediction for each tree created in the first phase.

The working of the random forest is as follows:

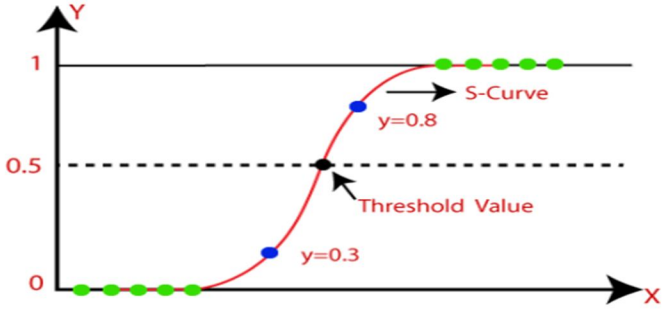
a) Step-1: Firstly, it will select random K data points from the training set.
b) Step-2: After selecting k data points then building the decision trees associated with the selected data points (Subsets).
c) Step-3: Then choosing the number N for decision trees that you want to build.
d) Step-4: Repeating step 1 and 2.
e) Step-5: Finding the predictions of each decision tree, and assigning the new data points to the category that wins the majority votes.

2) Logistic Regression

Logistic Regression is a supervised machine learning algorithm most often used for the classification purpose.

Logistic Regression is used to establish the relation between one dependent and one or more independent variable. It is used to make prediction about categorical values like True or False, 1or0, Yes or No. It gives the probability between the values 1 and 0 rather than showing exact value as 1 or 0.

In logistic Regression instead of fitting a regression line, we fit an "s" shaped logistic function, which predict to maximum values (0 or 1).



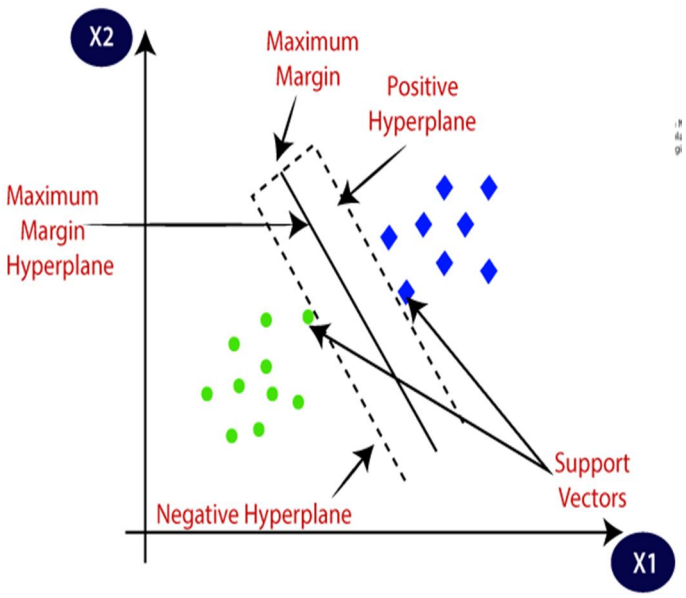
Mathematical steps to get the Logistic regression Equation:

$$\log\left[\frac{y}{1-y}\right] = b0 + b1x1 + b2x2 + \quad \ldots \quad + bnxn$$

3) Support Vector Classifier

Support vector machine is supervised machine learning algorithm used for classification as well as regression. Most often it is well suited for classification purpose.

The objective of SVM is to find the hyperplane in N-dimensional Space that distinctly classifies the data points. Dimension of the hyperplane is depending upon number of features.



We choose the hyperplane whose distance from it to nearest data point on each side is maximized. This hyperplane is known as maximum margin hyperplane.

Hyperplane are decision boundaries that help in classifying that help in classifying the data points. Support vectors are datapoints that are closer to hyperplane.

4) Decision Tree Classifier

Decision tree algorithm is supervised machine learning algorithm used for regression as well as classification. Here we used it for classification purpose.

a) Step 1 – Start with root node to form a tree T containing all the datapoints in dataset.

b) Step 2 – If the node purity is below the purity threshold or not all the records of s belong to class C, then use the purity information attribute to split the node. This create sub – tree.

c) Step 3 – Repeat step 2 until

- All the leaf node satisfy minimum purity threshold
- The tree cannot be further split
- Any other stopping criteria is reached (such as maximum tree depth desire).

5) XGBoost

XGBoost is supervised machine learning algorithm falls under the boosting technique. Boosting is an ensemble technique designed to combine several weak classifiers into strong classifier.

a) Step 1 - Initialize the dataset and assign equal weight to each of the data point.

b) Step 2 - Provide this as input to the model and identify the wrongly classified data points.

c) Step 3 - Increase the weight of the wrongly classified data points and decrease the weights of correctly classified data points. And then normalize the weights of all data points.

d) Step 4 - if (got required results) Go to step 5 else Go to step 2

e) Step 5 – End

6) Accuracy And F-1 Score

Heart Disease

Algorithm	Accuracy	F-1 Score
Logistic Regression	78.04	0.80
Random Forest Classifier	98.29	0.99
Support Vector Classifier	68.29	0.70
Decision Tree Classifier	98.53	0.99
XGBoost	98.53	0.99

Kidney Disease

Algorithm	Accuracy	F-1 Score
Logistic Regression	94.93	0.92
Random Forest Classifier	98.73	0.98
Decision Tree Classifier	98.73	0.98
XGBoost	98.73	0.98

Diabetes Disease

Algorithm	Accuracy	F-1 Score
Logistic Regression	72.22	0.71
Random Forest Classifier	80.55	0.80
Support Vector Classifier	72.22	0.70
Decision Tree Classifier	75.92	074
XGBoost	80.55	0.79

We choose Random Forest algorithm for building our all three Models.
Below are some points why should we use Random Forest Algorithm:

- It takes less training time as compare to other algorithm.
- It predicts output with high accuracy, even for large data set it run efficiently.
- It can also maintain accuracy when the large proportion of data is missing.

V. REQUIREMENT ANALYSIS

A. Hardware Requirements

1) Processor: - intel core2 duo and above
2) Ram: - 2GB and above
3) Rom: - 100GB and above

B. Software Requirements

1) Operating System: - Windows XP and above
2) Programming Language: - Python3
3) IDE: - Jupyter Notebook & VSCODE
4) Libraries: - Numpy, Pandas, Matplotlib, Seaborn, Sci-kit learn, Pickle, Flask.

VI. ADVANTAGES AND DISADVANTAGES OF SYSTEM

A. *Advantages*

1) It is very useful in Hospitals.
2) Reduce the workload of Hospital staff.
3) Easy to use. Anyone can use it.
4) It predicts robust and accurate results.
5) It takes less time to identify the patient have a particular disease or not.
6) Needy patient can get early treatment by identifying early-stage disease.
7) Cost saving.

B. *Disadvantages*

1) There is little bit number of chances of miss prediction in some cases.

VII. CONCLUSION

We successfully build a system that predict more than one disease with high accuracy. It is easy to use as well.
We achieve accuracy on project as follows:

1) Our project identifies heart disease with an accuracy of 98%.
2) Our project predicts diabetes disease with an accuracy of 80%.
3) Our project predicts kidney disease with an accuracy of 98%.

VIII. FUTURE SCOPE

1) It will be very useful to healthcare industries like hospitals to identify diseases at early stage.
2) In the future we can add more diseases in the existing system.
3) We can try to improve the accuracy of prediction in order to decrease the mortality rate.
4) Try to make the system more user-friendly.

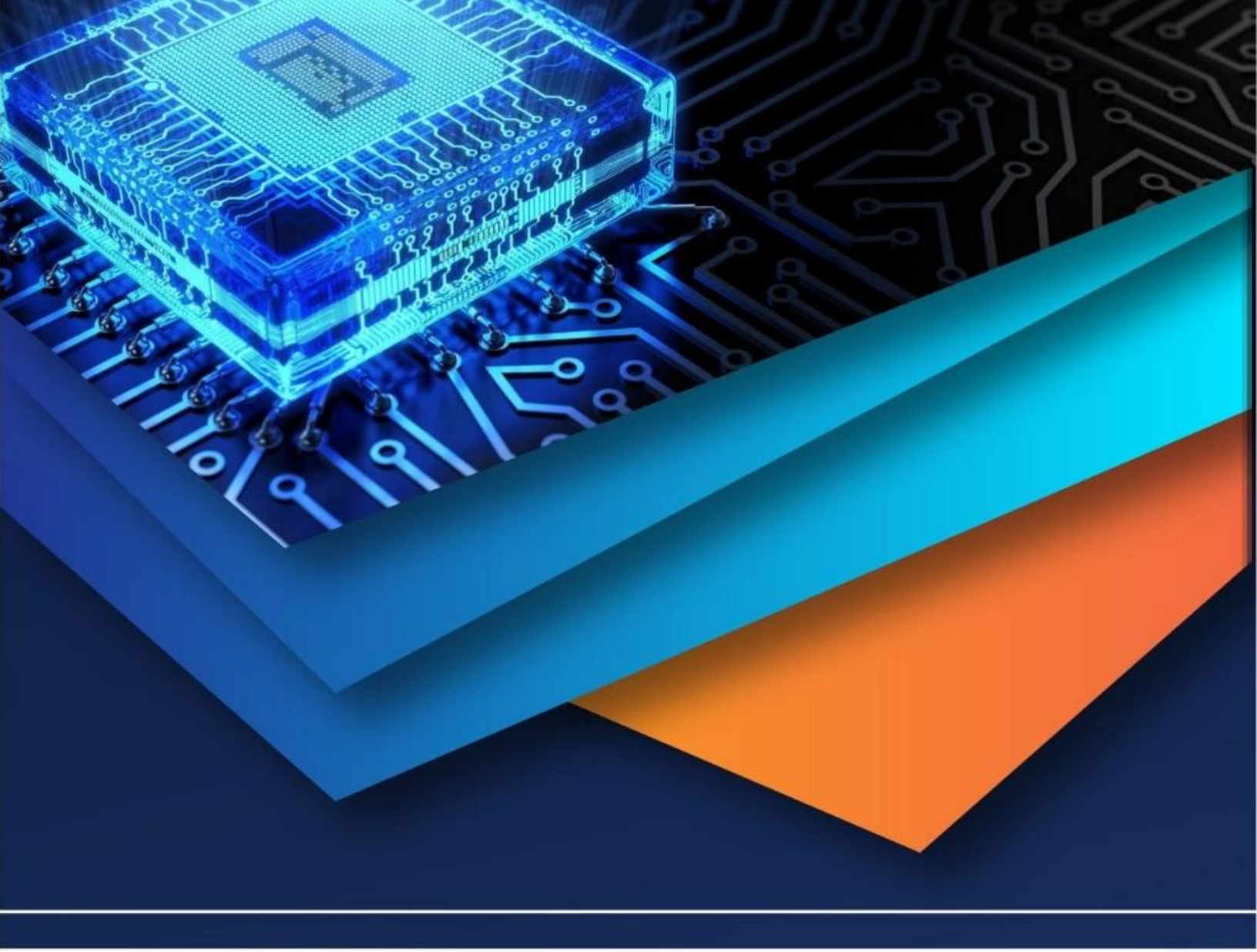
IX. ACKNOWLEDGEMENT

We would like to thank our college principal Dr. Pramod Rodge for providing lab facilities and permitting us to go on with our project. We can also express our deepest thanks to our H.O.D Dr. Uttara Gogate who's benevolent helps us making available the computer facilities to us for our project in our laboratory and making it true success. Without their kind and keen co-operation our project would have been stifled to stand still. We would like to thank our project coordinator Prof. Reena Deshmukh for all the support we need from they for our project. Lastly, we sincerely wish to thank our project guide Prof. Chetna Patil for their encouraging and inspiring guidance helped us to make our project success. Our project guide provided quality guidance to us. We would also like to thank our colleagues who helped us directly or indirectly during our project.

REFERENCES

[1] Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques Author - Senthilkumar Mohan, Chandrasegar Thirumalai, And Gautam Srivastava https://ieeexplore.ieee.org/document/8740989
[2] An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy Author - Bilal Khan, Rashid Naseem, Fazal Muhammad, Ghulam Abbas, And Sunghwan Kim, https://ieeexplore.ieee.org/document/9040562
[3] Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers Author - MD. Kamrul Hasan, MD. Ashraful Alam, Dola Das, Eklas Hossain, (Senior Member, IEEE), and Mahmudul Hasan https://ieeexplore.ieee.org/document/9076634
[4] Kaggle Website(https://www.kaggle.com/)
[5] A. S. Abdullah and R. R. Rajalaxmi, ''A data mining model for predicting the coronary heart disease using random forest classifier,'' in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22–25
[6] M. S. Dr Vijayarani1, ''Liver disease prediction using SVM and Naïve Bayes algorithms,'' Int. J. Sci., Eng. Technol. Res., vol. 4, no. 4, pp. 816–820, 2015
[7] L. Breiman, ''Random forests,'' Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001
[8] I. Jenhani, N. B. Amor, and Z. Elouedi, ''Decision trees as possibilistic classifiers,'' Int. J. Approx. Reasoning, vol. 48, no. 3, pp. 784–807, Aug. 2008.
[9] S. Vijayarani, S. Dhayanand, and M. Phil, ''Kidney disease prediction using SVM and ANN algorithms,'' Int. J. Comput. Bus. Res. (IJCBR), vol. 6, no. 2, pp. 2229–6166, 2015.
[10] N. Al-milli, ''Backpropogation neural network for prediction of heart disease,'' J. Theor. Appl.Inf. Technol., vol. 56, no. 1, pp. 131–135, 2013.
[11] T. Karayılan and Ö. Kılıç, ''Prediction of heart disease using neural network,'' in Proc. Int. Conf. Comput. Sci. Eng. (UBMK), Antalya, Turkey, Oct. 2017, pp. 719–723.

[12] P. Cunningham and S. J. Delany, ''k-Nearest neighbour classifiers,'' Multiple Classifier Syst., vol. 34, pp. 1–17, Mar. 2007
[13] J. Pahareeya, R. Vohra, J. Makhijani, and S. Patsariya, ''Liver patient classification using intelligence techniques,'' Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 4, no. 2, pp. 295–299, 2014.
[14] S. Vijayarani, S. Dhayanand, and M. Phil, ''Kidney disease prediction using SVM and ANN algorithms,'' Int. J. Comput. Bus. Res. (IJCBR), vol. 6, no. 2, pp. 2229–6166, 2015.
[15] J. P. Kelwade and S. S. Salankar, ''Radial basis function neural network for prediction of cardiac arrhythmias based on heart rate time series,'' in Proc. IEEE 1st Int. Conf. Control, Meas. Instrum. (CMI), Jan. 2016, pp. 454–458.
[16] H. A. Esfahani and M. Ghazanfari, ''Cardiovascular disease detection using a new ensemble classifier,'' in Proc. IEEE 4th Int. Conf. Knowl.- Based Eng. Innov. (KBEI), Dec. 2017, pp. 1011–1014.
[17] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, ''Hybrid intelligent modeling schemes for heart disease classification,'' Appl. Soft Comput. J., vol. 14, pp. 47–52, Jan. 2014. doi: 10.1016/j.asoc.2013.09.020.
[18] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, ''Performance analysis of data mining classification techniques to predict diabetes,'' Procedia Comput. Sci., vol. 82, pp. 115–121, 2016.
[19] K. Uyar and A. Ilhan, ''Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks,'' Procedia Comput. Sci., vol. 120, pp. 588–593, 2017.
[20] T. Vivekanandan and N. C. S. N. Iyengar, ''Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease,'' Comput. Biol. Med., vol. 90, pp. 125–136, Nov. 2017.
[21] T. Karayılan and Ö. Kılıç, ''Prediction of heart disease using neural network,'' in Proc. Int. Conf. Comput. Sci. Eng. (UBMK), Antalya, Turkey, Oct. 2017, pp. 719–723.
[22] A. Al-Anazi and I. D. Gates, ''A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs,'' Eng. Geol., vol. 114, nos. 3–4, pp. 267–277, Aug. 2010, doi: 10.1016/j.enggeo.2010.05.005.
[23] S. Taheri and M. Mammadov, ''Learning the naive Bayes classifier with optimization models,'' Int. J. Appl. Math. Comput. Sci., vol. 23, no. 4, pp. 787–795, Dec. 2013.
[24] H. Li, K. Ota, and M. Dong, ''Learning IoT in edge: Deep learning for the Internet of Things with edge computing,'' IEEE Netw., vol. 32, no. 1, pp. 96–101, Jan./Feb. 2018.
[25] P. Cunningham and S. J. Delany, ''k-Nearest neighbour classifiers,'' Multiple Classifier Syst., vol. 34, pp. 1–17, Mar. 07
[26] K. S. Dar and S. M. U. Azmeen, ''Dengue fever prediction: A data mining problem,'' J. Data Mining Genomics Proteomics, vol. 6, no. 3, pp. 1–5, 2015.
[27] B. Khan, R. Naseem, M. Ali, M. Arshad, and N. Jan, ''Machine learning approaches for liver disease diagnosing,'' Int. J. Data Sci. Adv. Anal., vol. 1, no. 1, pp. 27–31, 2019.
[28] Y. Meidan, M. Bohadana, A. Shabtai, J. D. Guarnizo, M. Ochoa, N. O. Tippenhauer, and Y. Elovici, ''ProfilIoT: A machine learning approach for IoT device identification based on network traffic analysis,'' in Proc. Symp. Appl. Comput., Apr. 2017, pp. 506–509.
[29] S. Radhimeenakshi, ''Classification and prediction of heart disease risk using data mining techniques of support vector machine and artificial neural network,'' in Proc. 3rd Int. Conf. Comput. Sustain. Global Develop. (INDIACom), New Delhi, India, Mar. 2016, pp. 3107–3111.
[30] M. Pradhan and G. R. Bamnote, ''Design of classifier for detection of diabetes mellitus using genetic programming,'' in Proc. 3rd Int. Conf. Frontiers Intell. Comput., Theory Appl., Nov. 2015, pp. 763–770



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY