iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** IV **Month of publication:** April 2020

DOI:

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: 7.429
Volume 8 Issue IV Apr 2020- Available at www.ijraset.com

A Comparative Study of Machine Learning Classifiers for Medical Diagnosis

Bhavnath Thakur[1], Harshit Ruhela[2], Kanishk Gupta[3], Chhaya Sharma[4]
[1, 2, 3]*Student,* [4]*Asst. Prof., Department of Computer Science & Engineering, Raj Kumar Goel Institute of Technology, Ghaziabad, India*

Abstract: ***Machine learning offers a principled approach for developing sophisticated, automatic, and objective algorithms for biomedical data. In our review we will be describing our system that takes symptoms from a patient and returns probable disease and then for the predicted disease we do an analysis using same classification algorithms. For classification we are using six classifiers which are - KNN, Decision Tree, Random Forest Naïve Bayes, Support Vector Machine (SVM) and Logistic Regression. For the training and testing we are using structured data and we did a 30% split for testing and 70% for training of model. In this review we will be describing our algorithms used and also methods for improving their accuracy.***
Keywords: Medical Diagnosis, KNN (K-Nearest Neighbor), Decision Tree, Naïve Bayes, Random Forest Tree, Support Vector Machine, Logistic Regression

I. INTRODUCTION

Machine Learning is an application of Artificial Intelligence that provides system the ability to automatically learn and improve from experience without being explicitly programmed. Disease prediction using patient treatment history and health data by applying data mining and machine learning techniques is ongoing struggle for the past decades. Machine learning in medical field has recently made huge progress. It has been able to help identify cancerous tumors. Machine learning lends itself to some processes better than others. Algorithms can provide immediate benefit to disciplines with processes that are reproducible or standardized. Machine learning can offer an objective opinion to improve efficiency, reliability, and accuracy. In this report, we use a proprietary platform to analyze data, and loop it back in real time to physicians to aid in clinical decision making. At the same time a physician sees a patient and enters symptoms, data, and test results into the EMR, there's machine learning behind the scenes looking at everything about that patient, and prompting the doctor with useful information for making a diagnosis, ordering a test, or suggesting a preventive screening. Long term, the capabilities will reach into all aspects of medicine as we get more useable, better integrated data. We'll be able to incorporate bigger sets of data that can be analyzed and compared in real time to provide all kinds of information to the provider and patient. Such a system will prove to be a helping hand to the doctor as he will be able to get an idea about the patient's condition before he begins his initial diagnosis. This will improve the accuracy of diagnosis and well help to reduce False Negative and False Positive diagnosis. There are number of algorithms to deal with both structured and unstructured data and provide possible diagnosis. In this system we will be focusing on structured data and six classification techniques which are KNN, Decision Tree, Random Forest Naïve Bayes, Support Vector Machine (SVM) and Logistic Regression.

II. EXISTING SYSTEM

Prediction using traditional disease risk model usually involves a machine learning and supervised learning algorithm which uses training data with the labels for the training of the models. High-risk and Low-risk patient classification is done in groups test sets. But these models are only valuable in clinical situations and are widely studied. Multiple web and mobile based applications are available for disease pre- diction that generally uses single classification algorithm. The information of patient, test result and statistics is recorded in Electronic Health Record (EHR) which enables potential solutions that reduce the case study of medical. The interaction between human and system is done by traditional approach.

III. PROPOSED SYSTEM

In this paper, the diagnosis process is divided in two parts symptoms prediction and report analysis. In our system on the basis of symptoms most probable disease is predicted and then for the predicted disease corresponding test will be recommended. In report analysis on the basis of parameters of test report, confirmation of disease could be checked. For prediction of disease multiple classifiers are used, namely – KNN, Decision Trees, Random Forests, Support Vector Machine, Logistic Regression and Naïve Bayes classification techniques and the output is the majority of prediction given by the mentioned classifiers. For accuracy

improvement we are using elbow method in KNN for selecting the best K value with least possible error and CART algorithm in Decision Tree. Also, in our system user can input symptoms both via voice and text hence interaction is easy as compared to existing systems. Since the output of system is majority of predictions by multiple classifiers the probable diagnosis become more accurate and thus our system will prove to be a helping hand for the doctors to get an idea of patient health before actual diagnosis.

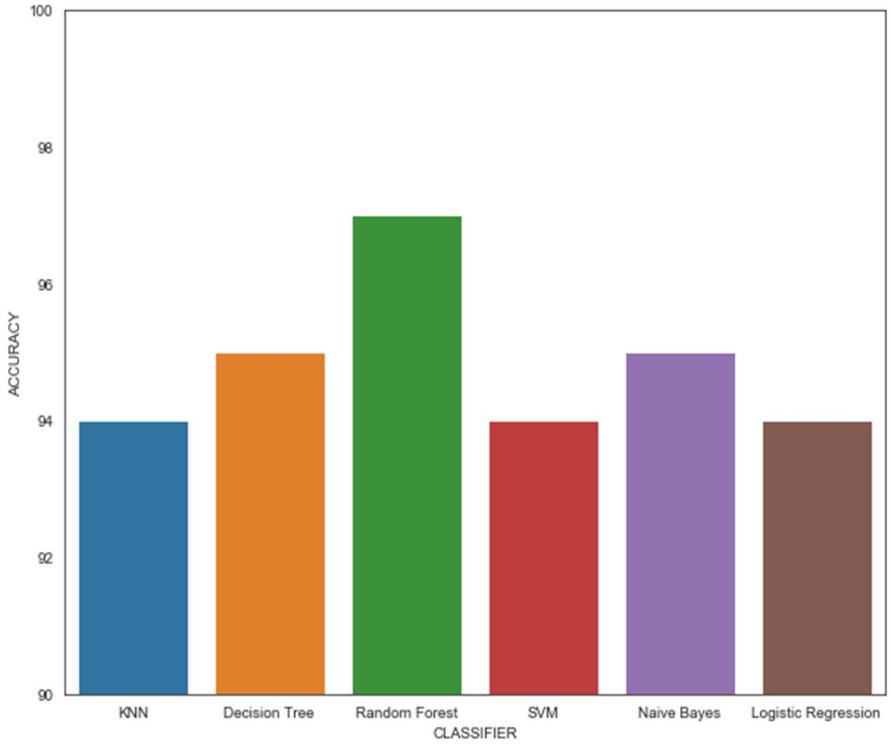


Fig.1.Classifiers with their respective Accuracy scores

Our model will predict majority of predictions given by 6 classifiers. If all six classifiers predict same disease then output will be that disease only. If more than half predict same disease and rest are predicting different disease then final output will be considered of majority classifiers. In case three are predicting some disease and rest three other than we would output the prediction of classifiers with more accuracy and if all are predicting different diseases then we would consider the prediction of random forest as our output since it is having best predictive accuracy for the given dataset. After we predict the most probable disease, we would do an analysis for that disease. Then our system would recommend tests for the predicted disease then user would input all the necessary parameters required for the test of the disease and then using same classifiers as above our system would provide confirmation of the disease.

A. KNN (K-Nearest Neighbor)

KNN or K Nearest Neighbors is a classification algorithm in machine learning. Since KNN is based on feature similarity we can do classification using KNN classifier. In KNN for a new data point we calculate its Euclidean distance from every available data point and then we select K nearest Neighbors based on the distance calculated and then the label or category assigned to the majority of those neighbors is assigned to the data point. Now the most common method to choose k is to take square root of all number of data points but in our system for purpose of improving accuracy we are using Elbow method.

In our system we trained and tested model for values of K ranging from 1 to 40 and calculated error rate. Error rate is calculated by comparing the prediction made by model to actual correct values. Then the final K value is determined by choosing that value where error rate decreased abruptly. For checking the value of K, we plotted K versus error rate with help of seaborn library and then selected appropriate K. Also, in our system we performed feature scaling using standard scalar which is also known as Z-score normalization which helps to scale down data based on standard normal distribution.

Accuracy of our model came out to be 94% with precision and recall of 91% and 87% respectfully.

B. Decision Tree

Decision tree is used here to create a training model that can use to predict the class or value of the target variable by learning decision rules obtained from training data. We start from the root of the tree, compare the values of the root attribute with the record attribute (training data) and based on the comparison, we follow the branch corresponding to that value and proceed to the next node. Since the decision of making strategic splits heavily affects a tree's accuracy, the decision criteria are different for classification and regression trees. Among various decision tree algorithms, we'll be using CART algorithm to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. Ergo, we can say that the purity of the node increases with respect to the target variable. The decision tree splits the nodes on all available variables and then select the splits which results in homogenous sub-nodes.

C. Random Forest

Random Forest is the most common ensemble method, it consists of a collection of decision trees. The idea behind Random Forest is that we repeatedly select data from the data set and build a decision tree with each new sample and then the most predicted label becomes the class for that data point. The intention of using this classifier is to get a more accurate diagnosis.

The reason that random forest works so well is that a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

The low correlation between models is the key as uncorrelated models predict more accurate results than any of the individual predictions. While some trees may be wrong many others will be right so as a group the trees are able to move in the correct direction.

D. Naïve Bayes

Naïve Bayes algorithms is a probabilistic machine learning model that's used for classification task. This algo is based on applying Bayes' theorem with a strong assumption that all the predictors are independent to each other. For example, a mobile may be considered as smartphone if it is having touch screen, internet facility, good camera etc. Here all these features contribute independently to the probability of that the mobile is a smartphone.

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

P(A/B) = (P(B/A) P(A)) / P(B)

Here, (A | B) is the posterior probability of class.

(A) is the prior probability of class.

(B|A) is the likelihood which is the probability of predictor given class.

(B) is the prior probability of predictor.

Naïve Bayes classification are easy to implement and fast. The requirement of training data is less. It is highly scalable in nature, or they scale linearly with the number of predictors and data points. It can make

probabilistic predictions and can handle continuous as well as discrete data. Naïve Bayes classification algorithm can be used for both binary as well as multi-class classification problems.

E. Support Vector Machine

Support Vector Machine or SVM are powerful yet flexible supervised machine learning algorithms. Generally, it is considered to be a classification approach, but it can be used both for classification and regression process. SVM is able to handle multiple continuous and categorical variables. SVM constructs a decision boundary or hyperplane that divide the datasets into classes to find a Maximum Marginal Hyperplane (MMH) where we can easily put the new data point in the correct category in the future.

The closest point to hyperplane is referred as Support Vector. Each dataset has a support vector point. The gap between dataset is known as margin. Greater margin will affect better computation result.

F. Logistic Regression

Logistic Regression allows us to solve classification problems where we try to predict discrete values. Logistic Regression makes use of sigmoid function which takes solution of linear regression and output value between 0 and 1. Now in Logistic Regression we set a probability mark usually .5 below which the point belongs to class 0 and above which it belongs to class 1 in a binary classification situation. Sigmoid Function = $(1/1+e^{-x})$

IV. SYSTEM ARCHITECTURE

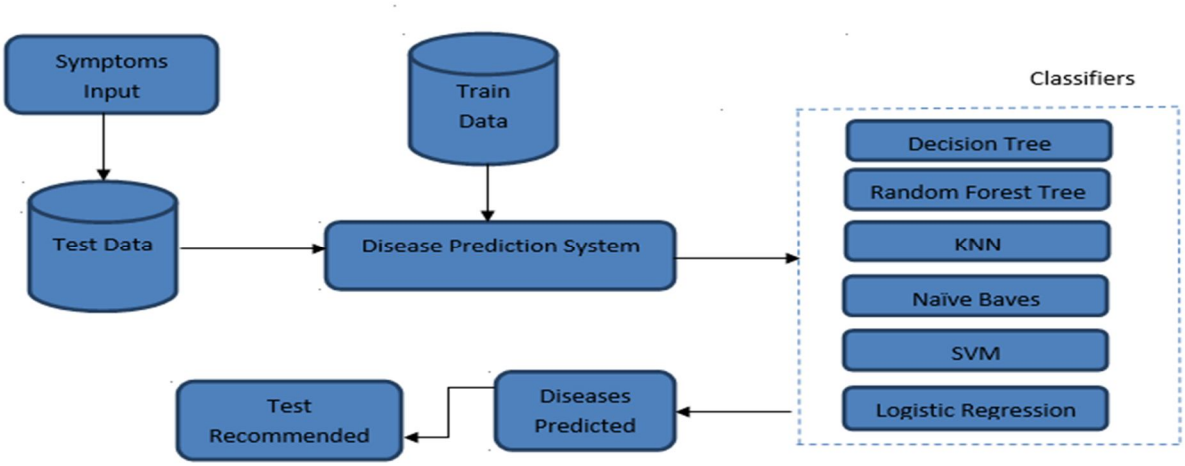


Fig.2.Disease Prediction from symptoms

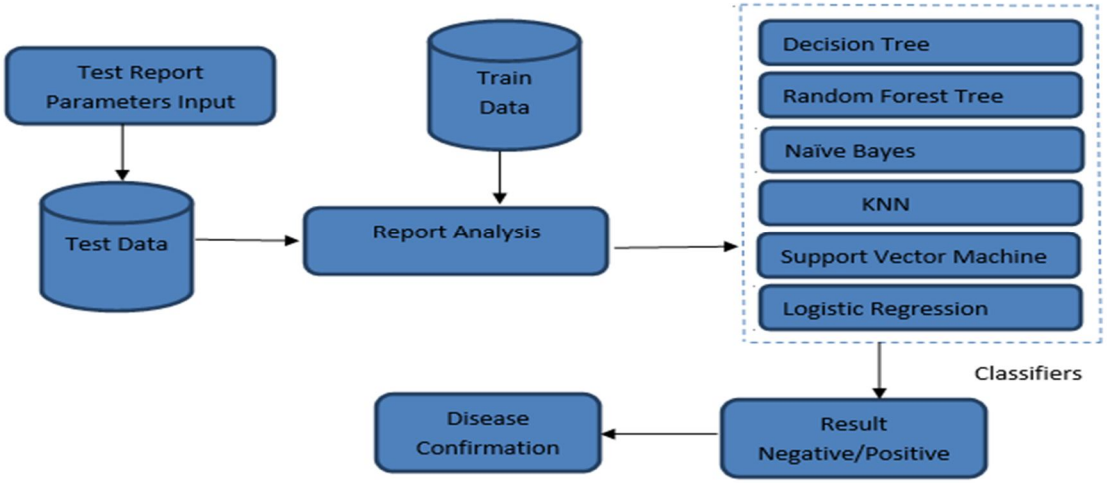


Fig.3.Analysis of recommended Test Reports of predicted diseases and confirmation of disease

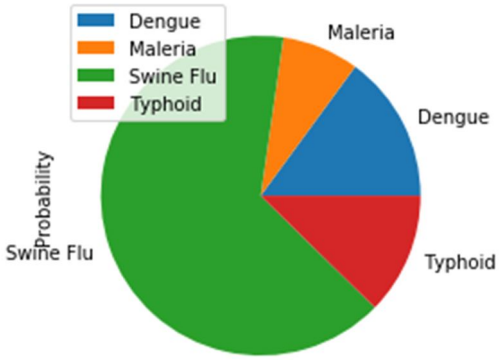


Fig.4.Probabilty of predicted diseases

Symptoms	Predicted Diseases	Test Recommended	Report Analysis	Final Disease
Fever Headache Muscle Pain Vomiting Rashes Fatigue	Swine Flu	Influenza	Positive	Swine Flu

Table.1. Process of disease prediction and confirmation

V. CONCLUSION

Machine learning has emerged as a field critical for providing tools and methodologies for analyzing the data generated by the biomedical sciences. This review has provided a condensed snapshot of applications of machine learning to detection and diagnosis of disease. Fusion of disparate multimodal and multiscale biomedical data continues to be a challenge. For example, current methods have difficulty integrating structural and functional imagery, with genomic, proteomic, and ancillary data to present a more comprehensive picture of disease. Hence, the above described system provides the initial diagnosis of the patient which would help doctors to improve accuracy of their diagnosis.

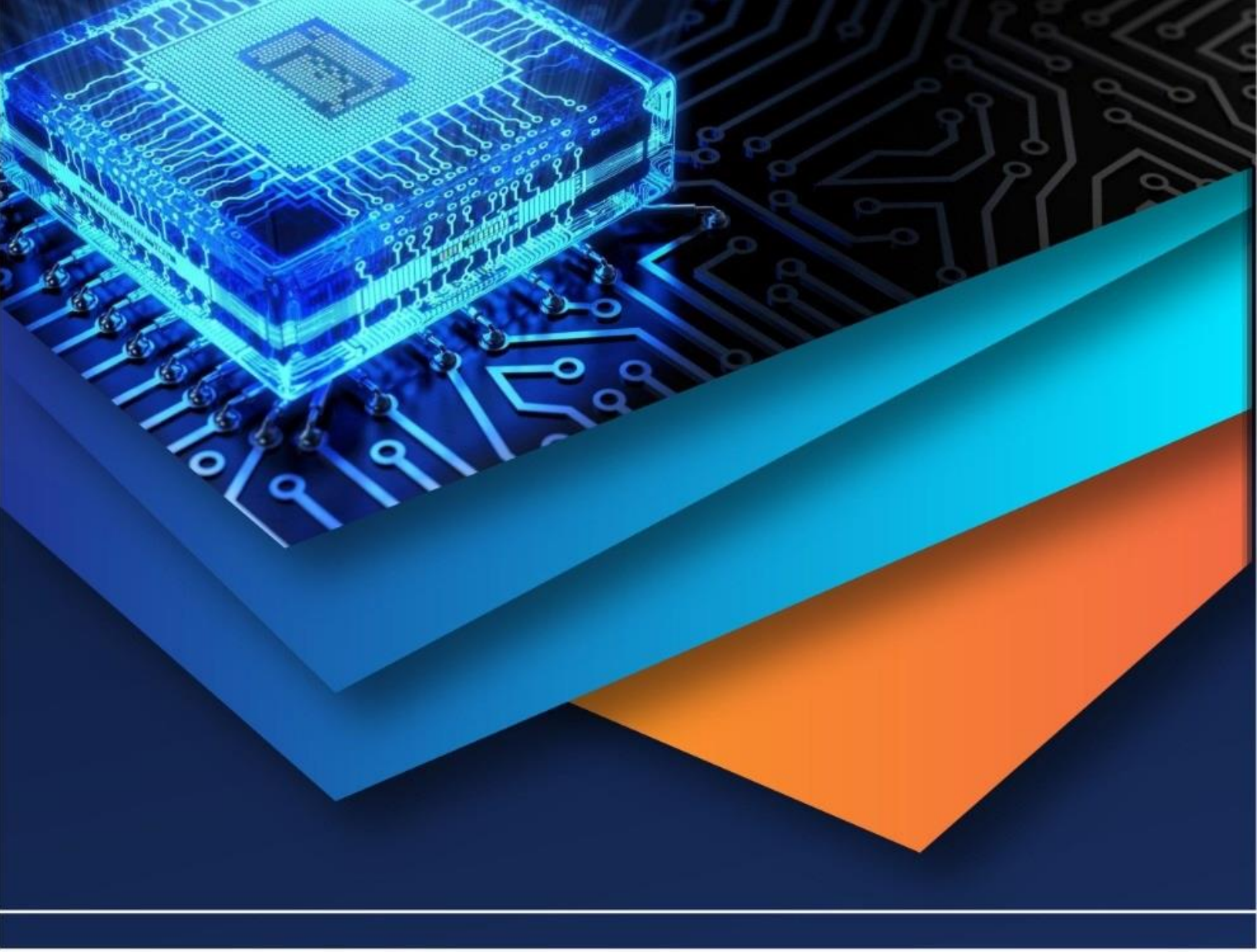
VI. ACKNOWLEDGEMENT

We would like to express our deepest appreciation to all those who provided us the possibility to complete this report. A special gratitude we give to our final year project guide, Ms. Chhaya Sharma, whose contribution in stimulating suggestions and encouragement, helped us to coordinate our project especially in writing this report.

Furthermore, we would also like to acknowledge with much appreciation the crucial role of Mr. Gaurav Agarwal. He helped in providing us an in-depth analysis of the algorithms/classification techniques. Last but not least, many thanks go to the Head of the Department, Dr. Sachi Gupta who invested her full effort in guiding the team in achieving the goal. We have to appreciate the guidance given by other supervisor as well as the panels especially in our project presentation that has improved our presentation skills thanks to their comment and advices.

REFERENCES

[1] Medium.com/sciforce/top-ai-algorithms-for-healthcare-aa5007ffa330, 2019
[2] Machine Learning, 1st Edition by Pearson (by Saikat Dutt, Subramanian Chandramouli, Amit Kumar Das.
[3] Mandal, I., Sairam, N. Accurate telemonitoring of Parkinson's disease diagnosis using robust inference system (2013) International Journal of Medical Informatics, 82 (5), pp. 359-377. DOI: 10.1016/j.ijmedinf.2012.10.006
[4] Torrado, N., Wiper, M.P., Lillo, R.E. Software reliability modeling with software metrics data via gaussian processes (2013) IEEE Transactions on Software Engineering, 39 (8), art. no. 6392172, pp. 1179-1186. DOI: 10.1109/TSE.2012.87
[5] Mandal, I., and Sairam, N. Accurate Prediction of Coronary Artery Disease Using Reliable Diagnosis System Journal of Medical Systems, 2012, Volume 36, Number 5, Pages 3353-3373. DOI: 10.1007/s10916-012-9828-0
[6] Mandal, I., Sairam, N. Enhanced classification performance using computational intelligence (2011) Communications in Computer and Information Science, 204 CCIS, pp. 384-391. DOI: 10.1007/978-3-642-24043-0_39
[7] Mandal, I., Sairam, N. New machine-learning algorithms for prediction of Parkinson's disease (2014) International Journal of Systems Science, 45 (3), pp. 647-666. DOI: 10.1080/00207721.2012.724114
[8] novatiosolutions.com/10-common-applications-artificial-intelligencehealthcare, 2018



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY