iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54415

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 VI Jun 2023- Available at www.ijraset.com

Diabetes Prediction Using ML

Shubham Choubey[1], Subham Agrahari[2], Anisha Shaw[3], Samprit Dhar[4], Rahul Raj Sarma[5], Shailesh Kumar Singh[6], Pallabi Das[7], Bidyutmala Saha[8]
[1, 2, 3, 4, 5, 6]*Student, CSE, Guru Nanak Institute of Technology*
[7, 8]*Asst. Professor, CSE, Guru Nanak Institute of Technology*

Abstract: ***The goal of this research is to create a machine learning algorithm-based system that is effective in detecting diabetes with high accuracy. Machine learning approaches have the potential to develop into trustworthy tools for diabetes diagnosis by utilising data analytics and pattern identification. Utilising feature selection techniques, the most pertinent elements that significantly influence diabetes prediction are found. Implemented and assessed using performance metrics including accuracy, recall, precision, and F1 Score are various machine learning algorithms, such as K-Nearest Neighbour, Logistic Regression, Random Forest, Support Vector Machine (SVM), and Decision Tree. The suggested technique works better than conventional methods, providing a more automated and effective method of diabetes detection. It could transform diabetes diagnosis, enhance patient outcomes, and enable individualised treatment plans.***
Keywords: Feature selection, Machine learning, Performance evaluation, Support Vector Machines.

I. INTRODUCTION

Diabetes is a chronic disease brought on by the body's inability to produce enough insulin or to utilise it effectively. A hormone called insulin controls blood sugar levels. High blood sugar levels caused by uncontrolled diabetes can harm nerves and blood vessels, among other physiological systems. Insulin injections are necessary for people with type 1 diabetes, also known as insulin-dependent diabetes mellitus (IDDM), because their bodies are unable to manufacture enough insulin on their own. When the body cells do not correctly utilise insulin, type 2 diabetes, also known as Non-Insulin Dependent Diabetes Mellitus (NIDDM), develops. Pregnant women who have type 3 gestational diabetes, which is characterised by elevated blood sugar levels, are affected. Diabetes increases the likelihood of developing certain problems and poses long-term health hazards. In India, a sizable portion of adults have pre-diabetic or have type 2 diabetes. The body's capacity to absorb glucose from food, which is necessary for energy, is impacted by diabetes. In order to use glucose as fuel, insulin is essential. Diabetes develops when the pancreas is unable to make enough insulin or when the body has trouble using it. The complications of diabetes can include heart attacks, strokes, neuropathy, foot ulcers, amputation, blindness, and kidney failure, and the disease is frequently misdiagnosed.

II. LITERATURE SURVEY

A variety of healthcare databases have been used to do substantial research on the common and chronic condition of diabetes. To accurately anticipate diabetes, researchers have created prediction systems employing machine learning algorithms. Many studies have concentrated on improving accuracy and contrasting various algorithms. For instance, the SVM method was applied to a diabetic dataset in a study by Omolara Adekoya, Walker Scott, and Stewart Robinson, which produced high accuracy. S Saru and S Subashree employed resampling methods in conjunction with decision tree, KNN, and Naive Bayes algorithms to achieve a 94.4% accuracy.

Decision tree, ANN, Naive Bayes, and SVM algorithms were employed by Priyanka Sonar and K. JayaMalini to achieve precision rates between 77% and 85%. Decision tree, SVM, and Naive Bayes algorithms were utilised by Deepti Sisodia and Dilip Singh Sisodia, with Naive Bayes showing the greatest performance at 76.30% accuracy. There are many uses for machine learning in many industries, including healthcare.

It can be divided into three categories: reinforcement learning, unsupervised learning, and supervised learning. While unsupervised learning includes building models on unlabeled data to find patterns and structures, supervised learning involves training models using labelled data. Models are trained through reinforcement learning to make choices based on rewards and consequences. Specifically, classification is a subset of supervised learning that applies to the current project.

Overall, machine learning techniques have shown to be excellent in understanding and predicting diabetes, which has helped to advance medical diagnosis and treatment. Our work comes under Supervised learning(Classification).

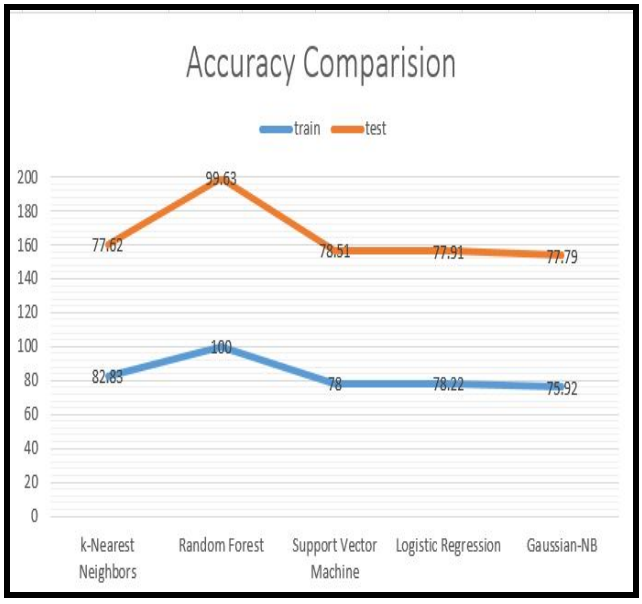
©IJRASET: All Rights are Reserved

III.ALGORITHMS USED

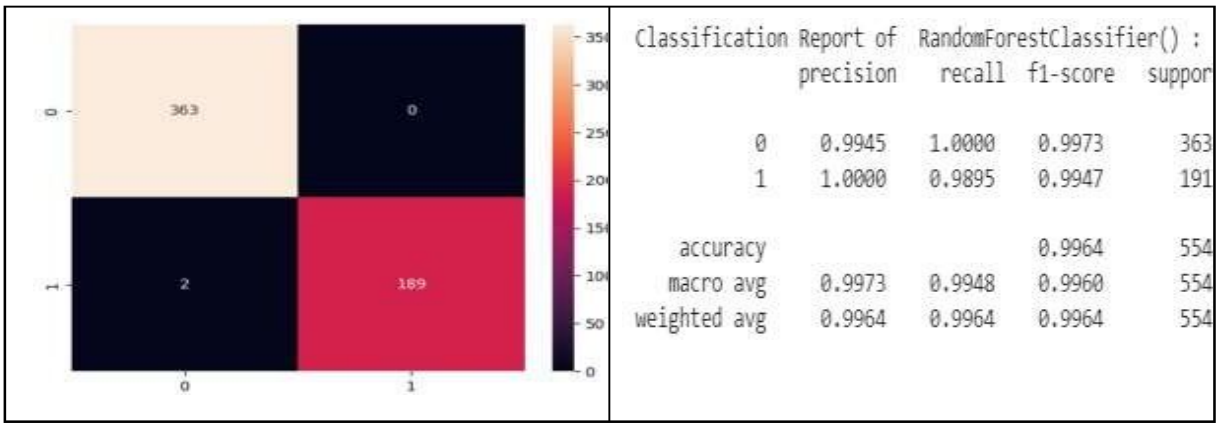
1) *The k-Nearest Neighbors are:* Possibly the simplest machine learning algorithm is the k-NN algorithm. The only step in creating the model is saving the training data set. The technique locates a new data point's "nearest neighbors," or the data points that are closest to it in the training data set. Let's first see whether we can verify the relationship between model complexity and accuracy.
2) *Rough Forest:* The idea of decision trees is advanced by this classifier. Each tree in the resulting forest is made up of a random selection of features drawn from the entire set of features.
3) *Support Vector Machine:* Support Vector Machine, sometimes known as SVM, is one of the most well-liked algorithms for supervised machine learning. It may be used to solve both classification and regression issues. However, It is typically employed for classification issues. The SVM algorithm's objective is to establish the best line or decision boundary that can divide an n-dimensional space into two classes, making it simple to classify new data points in the future. A hyperplane defines this optimum decision boundary.
4) *Logistic Regression:* One of the most often used Machine Learning algorithms, within the category of Supervised Learning, is logistic regression. Using a predetermined set of independent factors, it is used to predict the categorical dependent variable. In a categorical dependent variable, the output is predicted via logistic regression. As a result, the result must be a discrete or categorical value. However, it can only be either True or False, 0 or 1, or Yes or No. nonetheless, it provides probabilistic values that fall between 0 and 1 rather than the precise values, which are 0 and 1.
5) *Gaussian-NB:* Based on the Bayes theorem, the probabilistic machine learning technique known as Naive Bayes is employed for numerous classification applications. Gaussian The extension of naive Bayes is nave Bayes. It is mostly employed in text categorization with a large training set. One of the simplest and most efficient classification algorithms is the Naive Bayes Classifier, which aids in creating machine learning models that can be built quickly and anticipate outcomes.

IV. RESULT & ANALYSIS

A dataset created by Jonda Silva and the PIMA Indian diabetes dataset were used for the test's analysis.One dependent attribute and eight independent attributes are present in it. It was necessary to do feature engineering to make most out of dataset. We have done undersampling to make the dataset balanced. Very few had null values,so we took care of deleting them in the former process. Features were contributing unequally,so standardization was necessary here. Different models use different standardization functions(like StandardScaler & MinMaxScaler). For the test, Google Colab and Python 3 were used.To predict diabetes in women, five machine learning algorithms, including Logistic Regression, KNN, Random Forest, SVM, and GNB, were utilized during training & testing. Below image shows a comparative study of the accuracies we achieved.



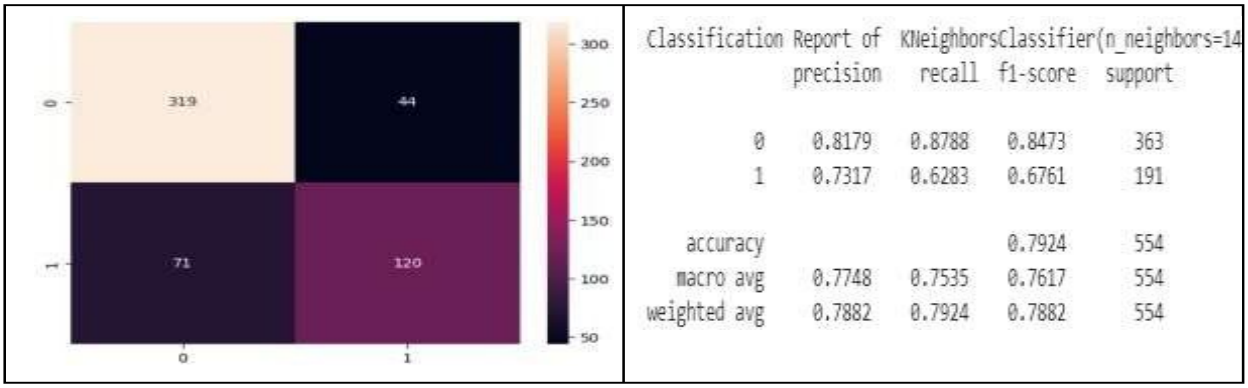
On the divided dataset(80:20 for training & testing respectively), we have applied a variety of machine learning methods in the model one by one, and Random Forest produces the classification results with the highest accuracy 99.63%.



Classification Report of RandomForestClassifier() :

	precision	recall	f1-score	suppor
0	0.9945	1.0000	0.9973	363
1	1.0000	0.9895	0.9947	191
accuracy			0.9964	554
macro avg	0.9973	0.9948	0.9960	554
weighted avg	0.9964	0.9964	0.9964	554

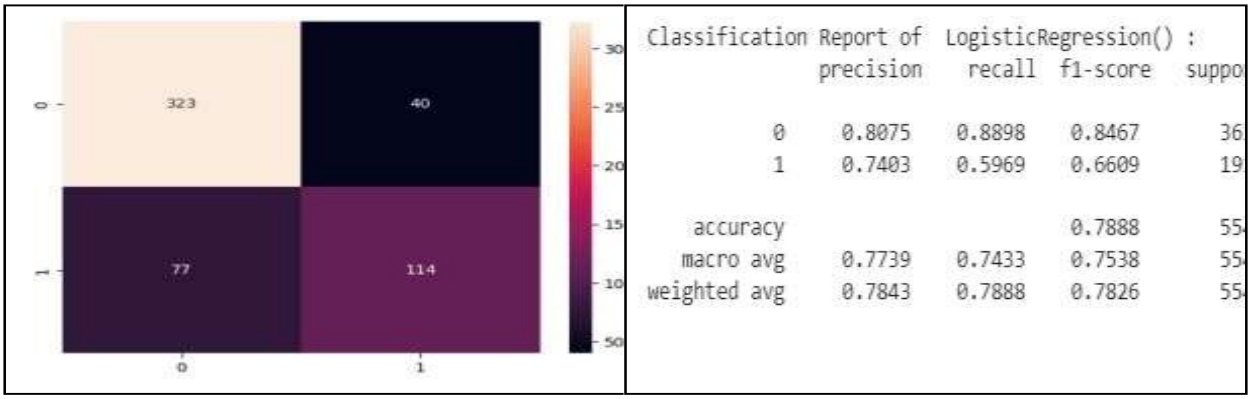
Fig: Classification report of Random forest



Classification Report of KNeighborsClassifier(n_neighbors=14

	precision	recall	f1-score	support
0	0.8179	0.8788	0.8473	363
1	0.7317	0.6283	0.6761	191
accuracy			0.7924	554
macro avg	0.7748	0.7535	0.7617	554
weighted avg	0.7882	0.7924	0.7882	554

Fig: Classification report of K-Nearest Neighbour

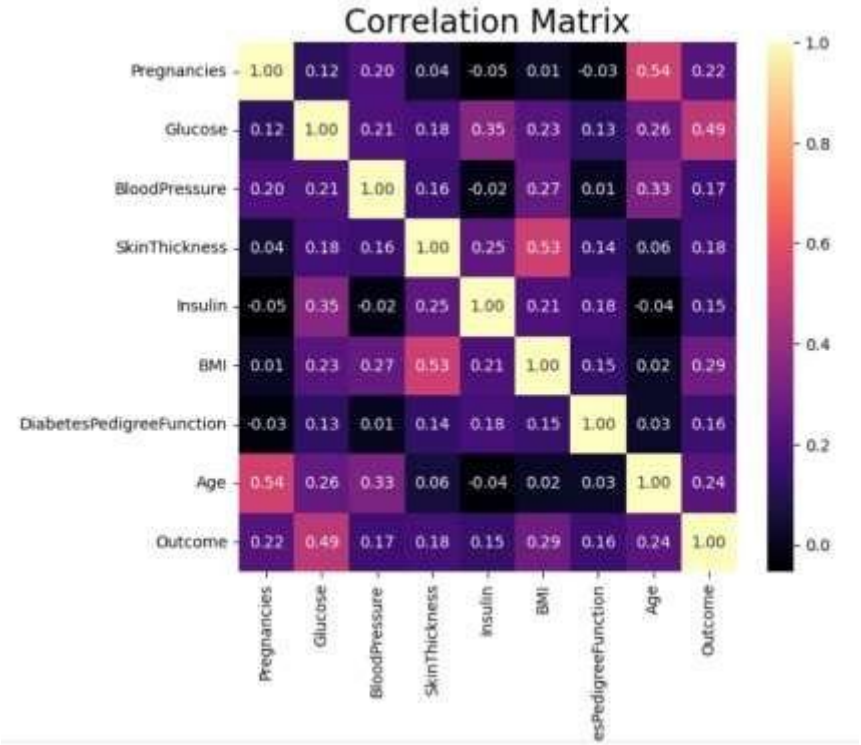


Classification Report of LogisticRegression() :

	precision	recall	f1-score	suppo
0	0.8075	0.8898	0.8467	36
1	0.7403	0.5969	0.6609	19
accuracy			0.7888	55
macro avg	0.7739	0.7433	0.7538	55
weighted avg	0.7843	0.7888	0.7826	55

Fig: Classification report of Logistic Regression

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

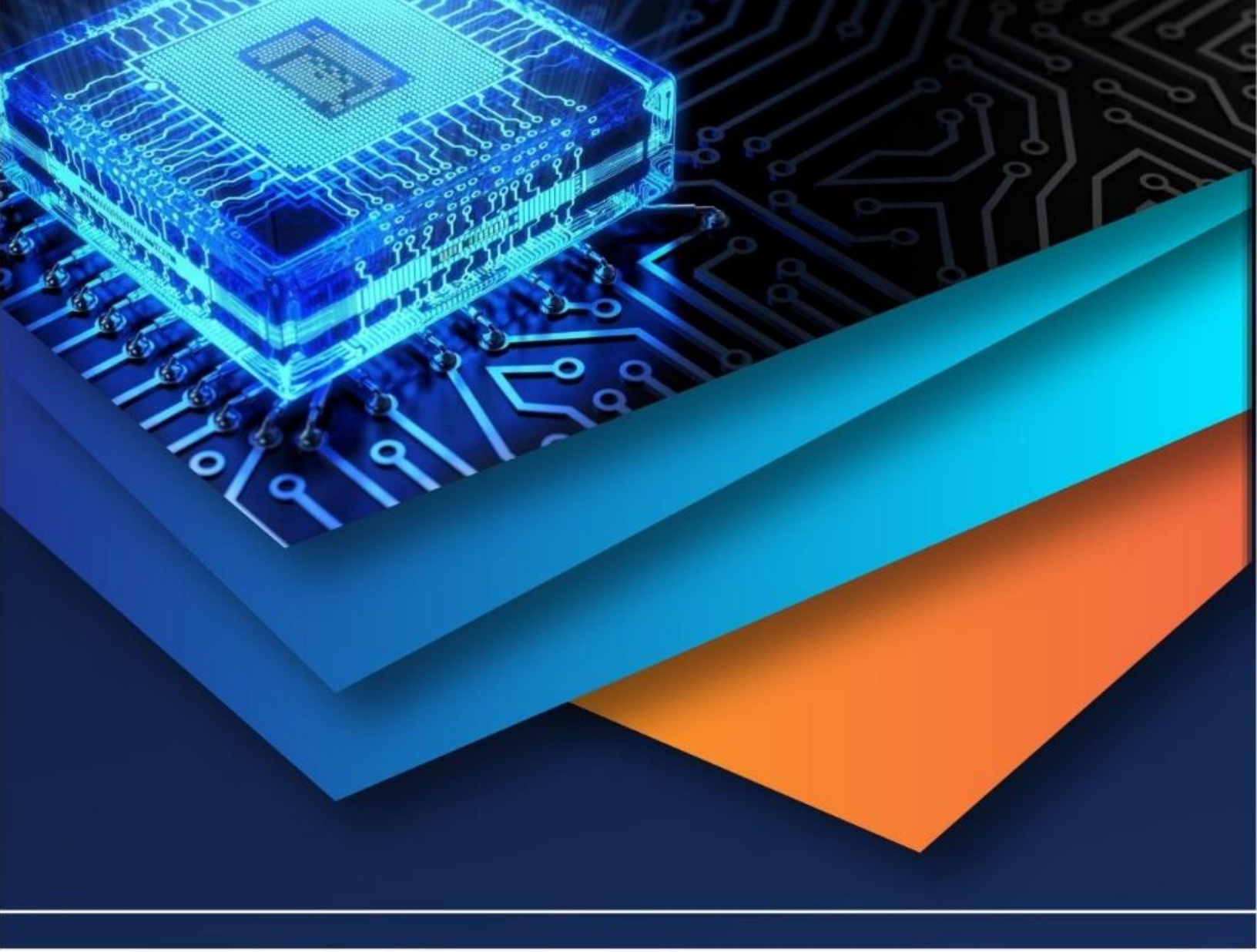


Four other algorithms were also used in the test, including Logistic Regression (which had an accuracy of 77.91%), Support Vector Machine (78.51%), K-Nearest Neighbor (which had an accuracy of 77.62%), and Gaussian Naive Bayes (which had an accuracy of 77.79%). Additionally, it has been found that half of all women with gestational diabetes go on to develop type 2 diabetes, meaning that women who have previously given birth have a higher risk of having the disease than women who have never given birth.

REFERENCES

[1] Reddy, Mrs Y. Anitha, et al. "Diabetes Disease Prediction Using Machine Learning Algorithms." Journal Of Engineering Sciences 14.04 (2023).
[2] Sisodia, Deepti, And Dilip Singh Sisodia. "Prediction Of Diabetes Using Classification Algorithms." Procedia Computer Science 132 (2018): 1578-1585.
[3] Kalyankar, Gauri D., Shivananda R. Poojara, And Nagaraj V. Dharwadkar. "Predictive Analysis Of Diabetic Patient Data Using Machine Learning And Hadoop." 2017 International Conference On I-Smac (Iot In Social, Mobile, Analytics And Cloud)(I-Smac). Ieee, 2017.
[4] Rani, Km Jyoti. "Diabetes Prediction Using Machine Learning." International Journal Of Scientific Research In Computer Science Engineering And Information Technology 6 (2020): 294-305.
[5] Walker, Scott, And Robinson Stewart. "Machine Learning Approach For Accurate Prediction Of Diabetes Mellitus."
[6] Sonar, Priyanka, And K. Jayamalini. "Diabetes Prediction Using Different Machine Learning Approaches." 2019 3rd International Conference On Computing Methodologies And Communication (Iccmc). Ieee, 2019.
[7] Saru, S., And S. Subashree. "Analysis And Prediction Of Diabetes Using Machine Learning." International Journal Of Emerging Technology And Innovative Engineering 5.4 (2019).

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)