iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45895

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 10 Issue VII July 2022- Available at www.ijraset.com

Stroke Analysis Using 10 ML Comparison

S K Uma[1], Rakshith S R[2]
[1]*Professor, Department of Computer Science and Engineering, PES College of Engineering, Mandya, Karnataka, India*
[2] *Department of Computer Science and Engineering, PES College of Engineering, Mandya, Karnataka, India*

***Abstract:** An illness known as a stroke damages the brain by rupturing the blood arteries in the brain. It can also happen when the passage of blood and other functioning of the brain is disrupted. The biggest cause of death and disability worldwide, according to the WHO, is stroke. The majority of research has been done on heart attack predictions, but very little research has been done on brain stroke risk. Considering this, numerous ML models are created to estimate the risk of a brain stroke. This work uses Ten ML Models such as Gaussian Naive Bayes, Bernoulli Naive Bayes, Gradient Boosting Classifier, Stochastic Gradient Descent, KNN, SVM, Decision Tree, Random Forest, Logistic Regression and MLP Classifier Classification. To analyze a wide range of physiological parameters and for precise prediction, The KNN algorithm, which had an accuracy of almost 94 percent, accomplished this task the best.*
***Keyword:** Stroke, KNN, MLP Classifier, Deep Leaning, Gradient Boosting Classifier, Stroke Prediction Dataset.*

I. INTRODUCTION

Brain Stroke is the 5th leading cause [1] of death in the USA, according to the Centers for Disease and Prevention (CDC). About 11 percent of all fatalities are identified by non-communicable diseases like stroke. Over seven lakhs ninety-five thousand people in the US suffer from the aftereffects of a stroke on a regular basis [2]. In India, it ranks as the 4th most important cause of death.

Thanks to technological breakthroughs in the field of medicine, ML can now be used to predict the onset of a stroke. The algorithms used in ML are beneficial in providing accurate analysis and producing accurate predictions. The majority of the prior efforts on stroke concern heart attack prediction. Brain stroke research has not been done very much. This study uses ML to forecast the likelihood of a brain stroke. KNN excelled the other ten classification algorithms, attaining more accuracy measure, according to the essential elements of the techniques utilized and results achieved. This model has a drawback because it was trained on textual dataset which is unbalanced. Ten ML classification techniques are implemented in this work. This work can be expanded upon to incorporate all of the existing machine learning techniques.

To continue with this challenge, a dataset from Kaggle [3] is picked that has a variety of physiological features as its attributes.

Based on an examination of the attributes in the Dataset, the final prognosis is made. The dataset is prepared for the ML model's comprehension by being cleaned. Data preprocessing is the name of this stage. The dataset is examined for empty values, and then filled in as necessary. If required, one-hot encoding is done after label encoding to encode string values into numbers.

After preprocessing the Data, the dataset is split into train data and test data. The dataset is then used to create a model using a variety of classification techniques. To find the best-trained prediction model, accuracy for each of these techniques is determined and then compared.

A Flask application is created after building the classifier and evaluating its accuracy. the user enters the parameters for prediction through a web application. The flask application acts as a link between the application and the model. After careful consideration, the study identifies which algorithm is most suited for stroke prediction.

II. RELATED WORK

Using four machine learning algorithms, [4] predicted strokes using the Cardiovascular Health Study (CHS) dataset. To arrive at the optimal solution, the authors combined C4.5, Decision Tree, Support Vector Machine, Artificial Neural Networks, and Support Vector Machine. The Dataset, on the other hand, contains fewer attributes.

In [5] stroke analysis prediction was performed using people's social media posts. The DRFS technique was used by the report's authors to determine the specific symptoms associated with stroke illness. the model's execution time is increased by Text extraction from social media feeds using NLP, which is undesirable.

To predict stroke analysis, the researchers of [6] employed a modified random forest algorithm. It was utilized to evaluate the risk of strokes. It is estimated that this method outperformed the ones already in use, as indicated by the article. This study has only analyzed a few different types of strokes; it cannot be expanded to include other types of stroke in the future.

©IJRASET: All Rights are Reserved

The model for stroke analysis prediction was trained using Decision Tree, Random Forest, and MLP, according to the research report [7]. The three techniques accuracy ratings were very close to one another, with only small differences. The determined efficiency of Decision Tree was 74.31 percent, that of Random Forest was 74.53 percent, and that of MLP was 75.02 percent. According to the study, MLP is more efficient than other two approaches. The only factor used to determine performance was total accuracy, and it did not always have a positive impact.

In Paper [8] demonstrates the development of a ML model to predict cardiac attack. They built the models using several machine learning approaches like Naive Bayes, Decision Tree, and Support Vector Machine and then compare their results. They acquired a higher precision of 60 percent from the methods used, which is quite low.

The researchers of [9] used information retrieval classification methods for calculating the chance of having a stroke. The dataset is collected from the Kingdom of Saudi Arabia's Department of National Guards Health Affairs Hospitals. C4.5, Jrip, and MLP were the three different classification techniques used (MLP). These techniques helped to reach an accuracy of about 95%. Although this Paper claims an efficiency of 95 percent, the training and prediction processes take more time because the authors used a combination of complicated algorithms.

According to study published in [10], the probability of a stroke can be predicted using Naïve Bayes, Neural Networks, and Decision Tree. This study has shown that decision tree now has improved precision (75 percent). However, the results of the confusion matrix showed that this approach could not be applied in real world scenarios.

The authors in [11] used the CHS dataset to predict strokes. They suggested a novel automated feature extraction technique on the conservative means they proposed. They used this strategy with the SVM ML for greater efficiency. However, this led to the creation of several vectors that were reducing the algorithm's efficacy.

In order to forecast thrombotic stroke disease, research [12] suggests artificial neural networks (ANN). The Back-propagation method was used to make predictions.. An accuracy of about 89 percent was achieved by this model. However, because of their complex architecture and larger number of hidden neurons, neural systems require more training time and time to operate.

III. EXISTING SYSTEM

As per the literature review almost all the authors have proposed the Decision Tree, MLP and SVM algorithms are more efficient models for the brain stroke analysis and the max accuracy outcome of the existing papers is 95 percent.

IV. PROPOSED SYSTEM

The data has been examined and is prepared for model building. For the model construction, a precompiled dataset and ML algorithms are required. Among the methods utilized are Gaussian Naïve Bayes, Bernoulli Naïve Bayes, Gradient Boosting Classifier, Stochastic Gradient Descent, KNN, SVM, Decision Tree, Random Forest, Logistic Regression and MLP Classifier Following the development of ten alternative models, they are contrasted using performance metrics as accuracy, precision, recall, and F1 score.

V. METHODOLOGIES

To continue with the development, various Kaggle datasets were explored. An adequate dataset for model training was selected from among all publicly available datasets.

After the dataset was gathered, the new stage is to prepare the dataset so that it is clearer and easily comprehended by the computer. This is known as data preparation. Handling imbalanced data Addressing null values and label encoding for this particular dataset are all part of this stage.

Now the dataset has been preprocessed, it is ready for model construction. A preprocessed dataset, as well as ML methods, are necessary for model construction. Following the development of ten distinct models, they are evaluated using four metrics: Accuracy Score, Precision Score, Recall Score, and F1 Score.

The model comparison produces the optimal model for the testing procedure in terms of accuracy measures. A web application is created to make it simpler for users to enter input data and receive results when a model has been implemented. The model receives the input data via a flask app, which is essentially a Python web framework which connects the model and the web app. The model uses input data to make predictions about the output, which it then sends back to the flask app. This Application then displays the outcome for users to review on the website page.

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

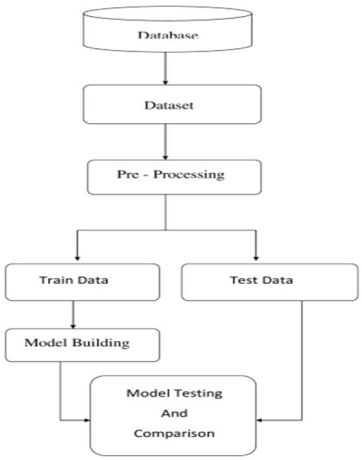


Fig-1: Proposed Flow Diagram

A. Data Collection

Kaggle provided the dataset for stroke prediction in accordance with the publication [3]. 5110 records with 12 characteristics are included in this collection. The columns main characteristics include id, gender, age, hypertension, heart disease, if a person has ever been married, work kind, residence type, average glucose level, body mass index, smoking status, and stroke. One of two values can be found in the output column "stroke": 1 or 0. The number 0 indicates that no stroke risk was found, while the number 1 indicates that a stroke risk was found. This dataset is strongly biased since the likelihood of a 0 in the outcome ("stroke") exceeds the likelihood of a "1" in the same column. The number 1 appears in just 249 records in the stroke field, whereas 0 appears in 4861 records. Data pre-processing is used to balance the data and increase performance.

B. Data Pre-Processing

Data preparation is essential before building a model in order to remove undesirable noise and outliers from the dataset that could stray from the expected training. Everything that hinders the model from operating more effectively is addressed in this step. Following the collection of the suitable dataset, the following step is to clean the data and ensure that it is suitable for model creation. The dataset used comprises 12 features. First, the field 'id' is removed because its presence makes little impact in model creation. The dataset will be examined for missing values, any discrepancy that is found are filled in. In this scenario, the field 'bmi' has missing value that are filled using the field data's mean categorized by the genders. The following step is label encoding, which comes after the dataset has been cleared of any missing values.

Label encoding converts categorical variables in a dataset into numeric variables that the machine can interpret. Because numbers are frequently used in computer training, Strings need to be transformed into integer. The collected dataset includes 5 fields of the data type string. During label encoding, every string is encoded, converting the entire dataset into a sequence of numbers.

The dataset that was utilized to predict stroke analysis is severely unbalanced. The complete dataset comprises 5110 records with 4861 entries indicating no stroke. While developing a machine learning model with such data may produce accuracy, other accuracy metrics like recall and precision are constrained. The results are incorrect and the forecast is ineffective if such skewed data is not handled correctly. The uneven data should therefore be dealt with first in order to create an adequate system.

C. Algorithm

Stroke is the most common illness identified by medical professionals. And its prevalence is increasing every day. This work tested ten commonly used ML approaches for detecting brain stroke recurrent using the publicly accessible, stroke prediction dataset, which are as follows:

1) GaussianNB
2) BernoulliNB
3) Random Forest Classifier
4) Logistic Regression

5) Decision Tree Classifier
6) K-Neighbors Classifier
7) Gradient Boosting Classifier
8) MLP Classifier (Neural Nets)
9) Support Vector Machine
10) Stochastic Gradient Descent

As per our algorithm analysis for the stroke prediction model the above ten algorithm were applied to predict stroke and select the better one as per that we have resulted with the testing and training scores of each algorithm which are listed below in the table both metrics and graphically.

Algorithms	Train Score of Trained Model	Test Score of Trained Model
GaussianNB	17.94	18.17
BernoulliNB	95.72	95.82
Logistic Regression	96.07	96.12
Random Forest	100.0	96.02
SVM	96.07	96.12
Decision Tree	100.0	92.85
KNN	96.12	96.22
Gradient Boosting	96.49	96.02
Stochastic Gradient Descent	96.07	96.12
MLP	99.87	92.35

Table-1: Training and Testing Score of Trained Model

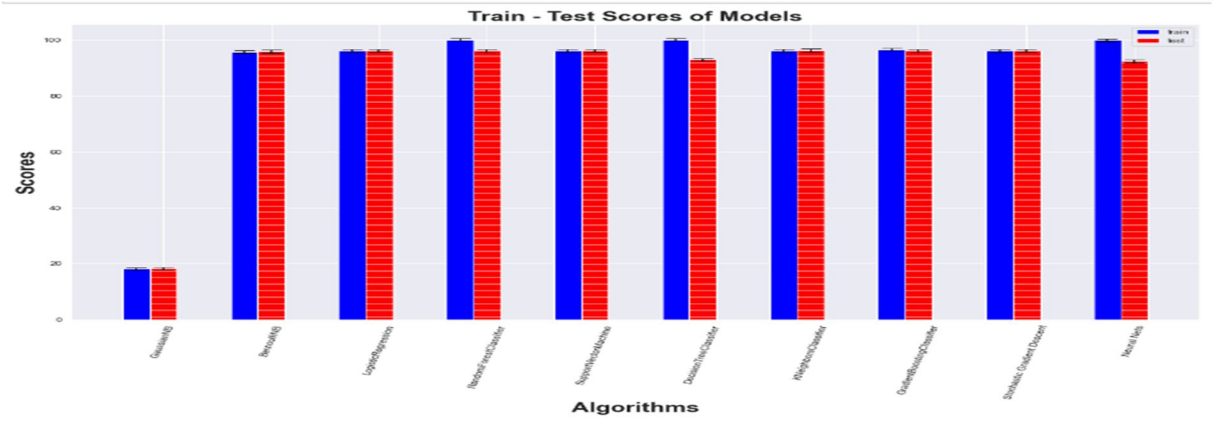


Fig-2: Graphical representation of Training and Testing Score of Trained Model

As per the metrics of Training Score of Trained Model of DecisionTree, RandomForest, GradientBoosting and MLP classifier is more we have chosen the KNN model for prediction as its Testing score of trained models is more and as it is the best algorithm for model prediction.

VI. WORKING OF THE SYSTEM

1) *Step 1:* Upload data by entering the values using the selection and text box given.
2) *Step 2:* Uploaded data will be loaded to the KNN prediction model. Model Checks whether the person is having Stroke or not.
3) *Step 3:* If the person is having stroke message will be displayed to user.

VII. EXPECTED RESULT

We tested the system with both stroke and no-stroke values to ensure that the algorithm utilized in the system is robust. The algorithm predicted the no-stroke and stroke metrics as per the data entered. Below is the classification report for KNN Algorithm

```
KNeighborsClassifier
Train score of trained model: 96.1271102284012
Test score of trained model: 96.22641509433963

Confussion Matrix:
[[968   38]
 [   0    1]]

Accuracy : 0.9622641509433962
Specificity : 0.9622266401590457

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98      1006
           1       0.03      1.00      0.05         1

    accuracy                           0.96      1007
   macro avg       0.51      0.98      0.52      1007
weighted avg       1.00      0.96      0.98      1007
```

Fig-3: Representation of Classification Report for KNN Model

VIII. CONCLUSION

Stroke is a serious medical illness which has to be handled right away. The development of an ML model can help to diagnose stroke early and minimize the severity of its after effects. This work shows how a number of machine learning algorithms perform in accurate prediction of stroke based on various physical measures. KNN Classification performs better than all other techniques, with a 96.22 % accuracy rate.

IX. FUTURE WORK

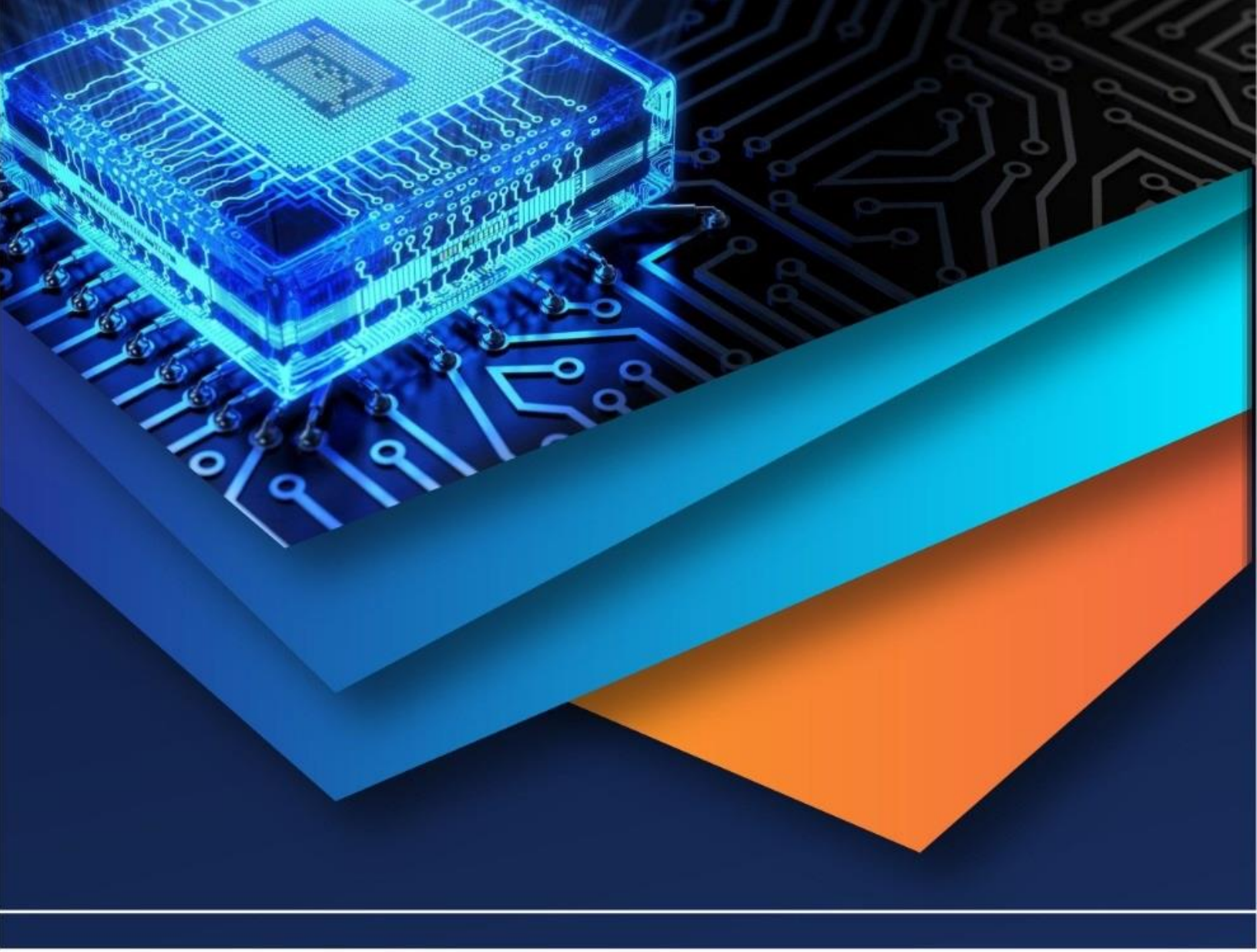
We utilized a jupyter notebook to construct the software, and it was a success. In Python, our project has been successfully tested. We also looked into the project's uses and future scope. Our solution can be linked as API and it can used as mobile app for both Android and IOS.

REFERENCES

[1] Concept of Stroke by Healthline.
[2] Statistics of Stroke by Centers for Disease Control and Prevention.
[3] Dataset named 'Stroke Prediction Dataset' from Kaggle: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.
[4] Singh, M.S., Choudhary, P., Thongam, K.: A comparative analysis for various stroke prediction techniques. In: Springer, Singapore (2020).
[5] Pradeepa, S., Manjula, K. R., Vimal, S., Khan, M. S., Chilamkurti, N., & Luhach, A. K.: DRFS: Detecting Risk Factor of Stroke Disease from Social Media Using Machine Learning Techniques. In Springer (2020).
[6] Vamsi Bandi, Debnath Bhattacharyya, Divya Midhunchakkravarthy: Prediction of Brain Stroke Severity Using Machine Learning. In: International Information and Engineering Technology Association (2020).
[7] Nwosu, C.S., Dev, S., Bhardwaj, P., Veeravalli, B., John, D.: Predicting stroke from electronic health records. In: 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE (2019).
[8] Fahd Saleh Alotaibi: Implementation of Machine Learning Model to Predict Heart Failure Disease. In: International Journal of Advanced Computer Science and Applications (IJACSA) (2019).

[9] Ohoud Almadani, Riyad Alshammari: Prediction of Stroke using Data Mining Classification Techniques. In: International Journal of Advanced Computer Science and Applications (IJACSA) (2018).
[10] Kansadub, T., Thammaboosadee, S., Kiattisin, S., Jalayondeja, C.: Stroke risk prediction model based on demographic data. In: 8th Biomedical Engineering International Conference (BMEiCON) IEEE (2015).
[11] Aditya Khosla, Yu Cao, Cliff Chiung-Yu Lin, Hsu-Kuang Chiu, Junling Hu, Honglak Lee:An Integrated Machine Learning Approach to Stroke Prediction. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (2010).
[12] Shanthi, D., Sahoo, G., Saravanan, N.: Designing an artificial neural network model for the prediction of thrombo-embolic stroke. Int. J. Biometric Bioinform. (IJBB) (2009).
[13] 7 Techniques to Handle Imbalanced Data – Kdnuggets.
[14] Documentation for Logistic Regression from Scikit-learn.org.
[15] Documentation for Decision Tree Classification from Scikit-learn.org.
[16] Documentation for Random Forest Classification from Scikit-learn.org.
[17] Documentation for K-Nearest Neighbor from Scikit-learn.org.
[18] Documentation for Support Vector Machine from Scikit-learn.org.
[19] Documentation for Naïve Bayes Classification Algorithm from Scikitlearn.org

©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)