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.2022.45503

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

Diabetes Prediction Model

Nikita Joshi[1], Abhishek Singh[2], Omkar Bhanushali[3], Rishikesh Datar[4], Deepali Kayande[5]
[1, 2, 3, 4, 5]*Department of Computer Engineering, A.P. Shah Institute of Technology, Thane, Maharashtra, India*

Abstract: Everyone is currently quite aware how dangerous and adverse issues Diabetes causes on a human body. In today's world filled with all sorts of impurities and all other adulterations, even slightest carelessness in maintaining lifestyle can cause serious diseases, disorders and consequences on health. Although with advancement of medical science, we do have treatment cures of Diabetes, but still lacks speed in detection of presence of Diabetes in a human body.
Here, in this study proposed a system that can predict whether a person has diabetes or not with the help of Machine Learning. This project uses Logistic Regression Machine model for the prediction of presence of Diabetes in a person.
Keywords: Diabetes, human body, adulterations, lifestyle, diseases, Machine learning, Logistic Regression, Prediction.

I. INTRODUCTION

Diabetes is a chronic health disorder which affects the body's natural process of converting food into energy. Our body produces natural hormone called Insulin that moves sugar from the blood to the cells for storage or for later use of energy. What Diabetes does is, either not allow enough Insulin production or restrict the effective use of Insulin produced.

Due to all the speeding environment, number of people affected by Diabetes is rising up rapidly. And most among the diabetics, know not much about the risk factors they face prior to detection.

In the past 30 years, of overall developments, we can also evidently see a rise in number of diabetics. People have now slowly begun to realize how deeply Diabetes impacts one's health and his everyday life. When observed, there is a constant inclining trend in the proportion of diabetics in the general population, and the specific growth rate in males is evidently higher than that in females as shown in Fig.1. Globally, China has the largest diabetic population in the world followed by United States and India.

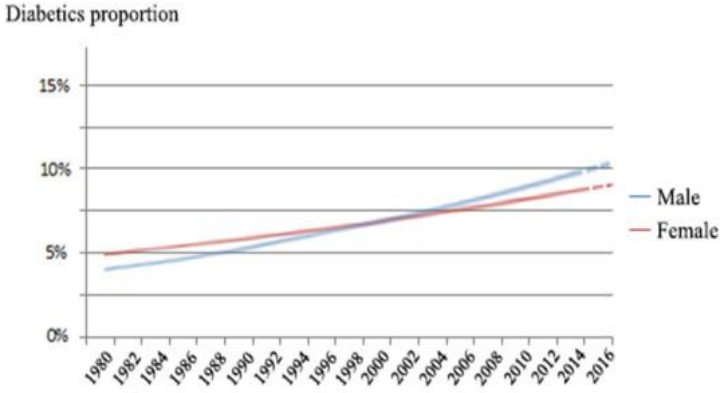


Fig.1 Trend of Diabetes proportion.

To effectively lower the morbidity and reduce the impact of Diabetes, we need to focus on the high-risk age group of people. According to WHO standards, these are the common categories of groups under high risk of Diabetes [4]:

- Age ≥ 45 and infrequent exercising
- BMI ≥ 24 kg/m2
- Family history of DM
- Hypertension or cardiovascular and cerebrovascular disease
- Gestation female whose age ≥30.

As we are steadily learning about these diseases and disorders, we get to know that Diabetes is an incurable disease. But with the help of modern science, Diabetes is still manageable and can be well controlled with regular treatment. However, modern science would show its greatest miracles when detection of diabetes in a person is done at an early stage.

For avoiding and reducing the critical impact, there exists an urgent need to create a system that will detect the presence of diabetes disease with optimal cost and better performance.

Our proposed system has an objective to fulfil this urgency and has used Logistic Regression, a machine learning method to develop the model required to predict the presence of diabetes.

II. LITERATURE REVIEW

Machine Learning, a study of computer algorithms has vast use in medical fields. Many medicine and health researches require the use of various machine learning algorithms and Deep learning algorithms.

Likewise, we could have used various algorithms like Support Vector Machine, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm, Logistic Regression, K-NN, Gaussian Naïve Bayes, Bagging algorithm and Gradient Boost Classifier. But after our study, working with and testing these algorithms with our dataset, we received the highest accuracy of 96% by the Logistic Regression algorithm. The model will be built by using Logistic Regression and then trained and tested with the PIMA Indian dataset, which consists 768 female patient's data who are all more than the age of 21. The dataset consists of 9 features:

1) *Pregnancies:* Number of times the patient was pregnant.
2) *Glucose:* Plasma glucose concentration over two hours in an oral glucose tolerance test.
3) *Blood Pressure:* Diastolic blood pressure (mm Hg).
4) *Skin Thickness:* Triceps skin fold thickness (mm).
5) *Insulin:* Two-Hour serum insulin (mu U/ml).
6) *BMI:* Body mass index (weight in kg/(height in m)^2).
7) *Diabetes Pedigree Function/DPF:* A function that scores the likelihood of diabetes based on family history.
8) *Age:* In years.
9) *Outcome:* Class variable (0 if non-diabetic, 1 if diabetic). This is the target variable.

We have designed a flow chart Fig.2 of the prediction model that shows the implementation and execution that will be achieved.

III. EXPERIMENTAL SETUP

A. *Software Requirements:*

1) *Python:* General Programming Language used to code the model, includes several libraries.
2) *Python libraries:* NumPy, matplotlib, PIL (Python Imaging Library), pandas.
3) *Flask:* Python web framework used for creating web applications in python.
4) *PyCharm:* Python IDE used for production of python web development.

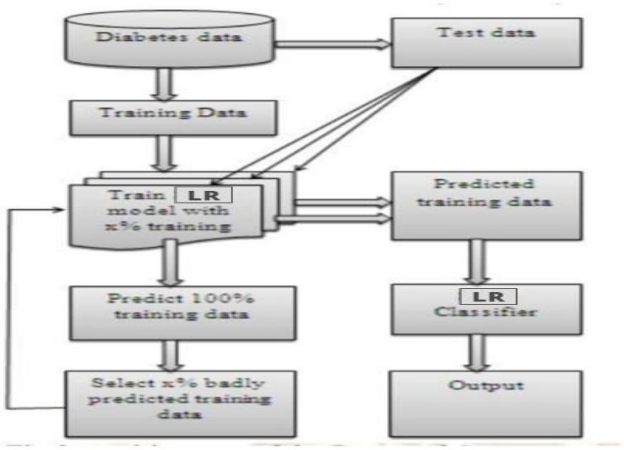


Fig.2 Flow chart of the model

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

IV. METHODS AND PROCESS

Logistic Regression is a classification machine learning algorithm, used to predict a binary outcome of a provided dataset of independent variables or features.

A. Data Pre-processing

The zero or null values in the features of dataset need to be located. The predictor model cannot have zero values in any other feature than the Pregnancies feature in the dataset.

Such zero values are replaced by the mean values of the feature column. This step is a major requirement for the growth in the accuracy as the incorrect values increase the chance of faulty prediction.

B. Training and Testing Model

The Dataset used is split into a ration of 80:20 for training and testing the model. The 80% of the data will be used for training and 20% of it will used for testing the prediction model.

The use of Logistic Regression is for the presence of diabetes prediction and to check the accuracy of the outcome predictions.

In the section of predictions, there are 4 important parts:

True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

Here, TP and TN are the cases when the results and the real outcomes match. Whereas, FP and FN are the cases obtained when there is a mismatch in the results and real outcomes.

A classified report is then generated which consists of Precision Metric, Recall, F1 score and support.

Recall is the percent of positives accurately identified. The Precision metric is the percentage of accurate predictions, F1 score is the percent of positive predictions that are accurate and Support is the actual occurrence of the class in the dataset.

Precision = TP/ (TP + FP)

Recall = TP/(TP+FN)

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

C. Calculating the Accuracy

With help of confusion matrix, a tabular form of actual vs predicted values, accuracy of the model is found. Calculation of model accuracy is done by:

$$\frac{\text{True Positive + True Negatives}}{\text{True Positive + True Negatives + False Positives + False Negatives}}$$

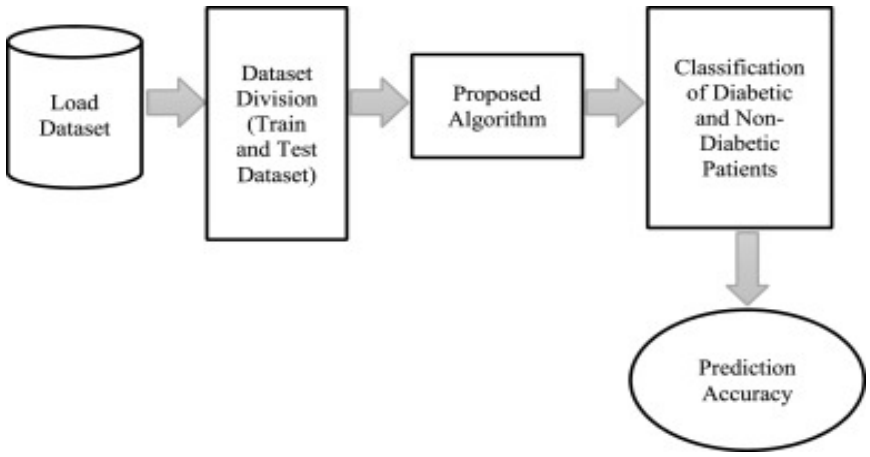


Fig.3 Flow of Modules.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

V. RESULTS

On processing through the dataset, we inferred that 34.9% of the patients are Diabetic and the rest of the 65.1% of them do not have diabetes. We also have the feature wise comparison results.

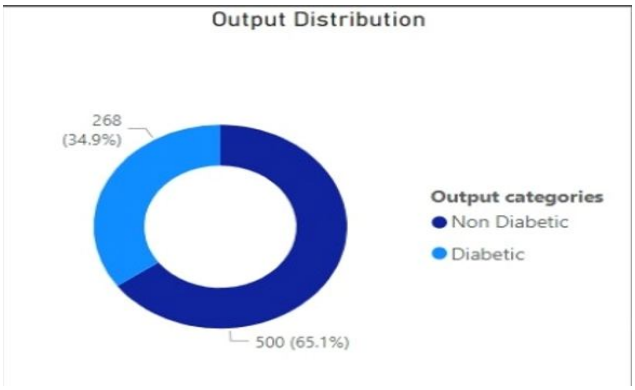


Fig.4 Output Distribution of Diabetic and Non-Diabetes patients.

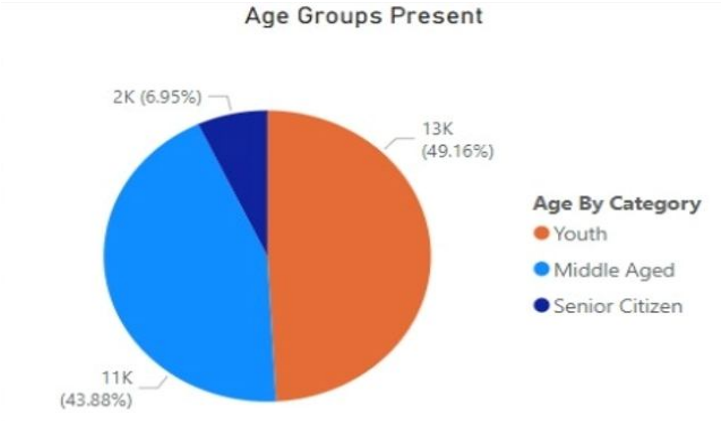


Fig.5 Age Groups present in the dataset.

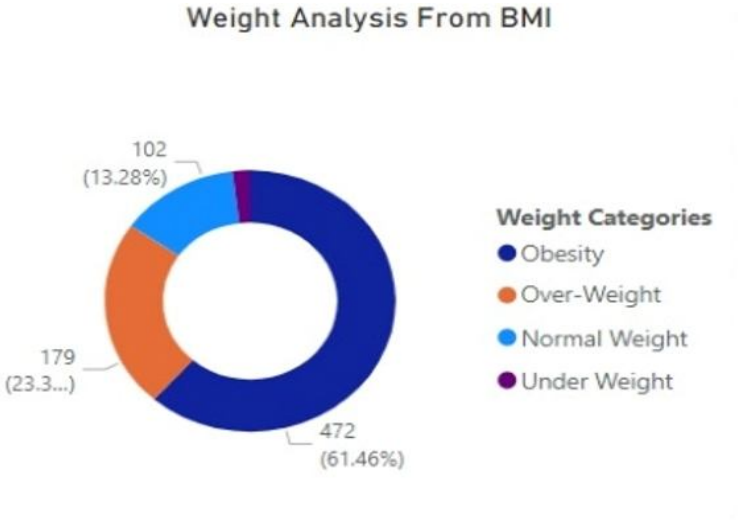


Fig.6 Weight Analysis from BMI given in the Dataset

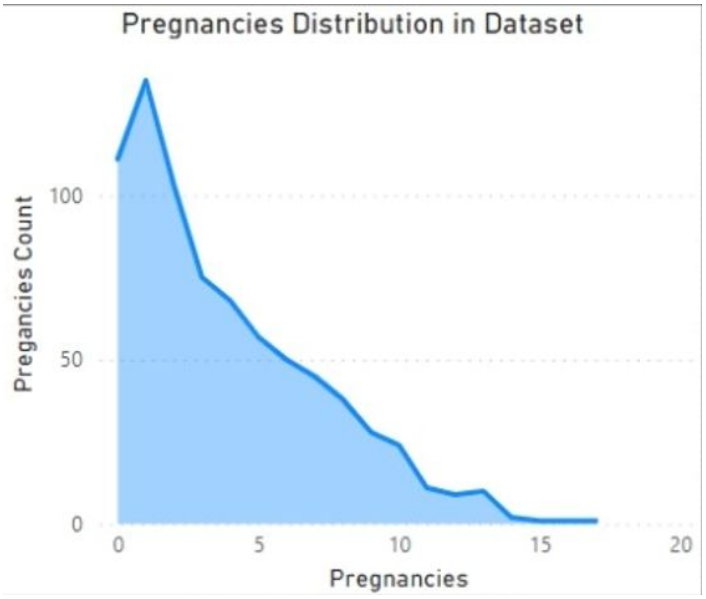


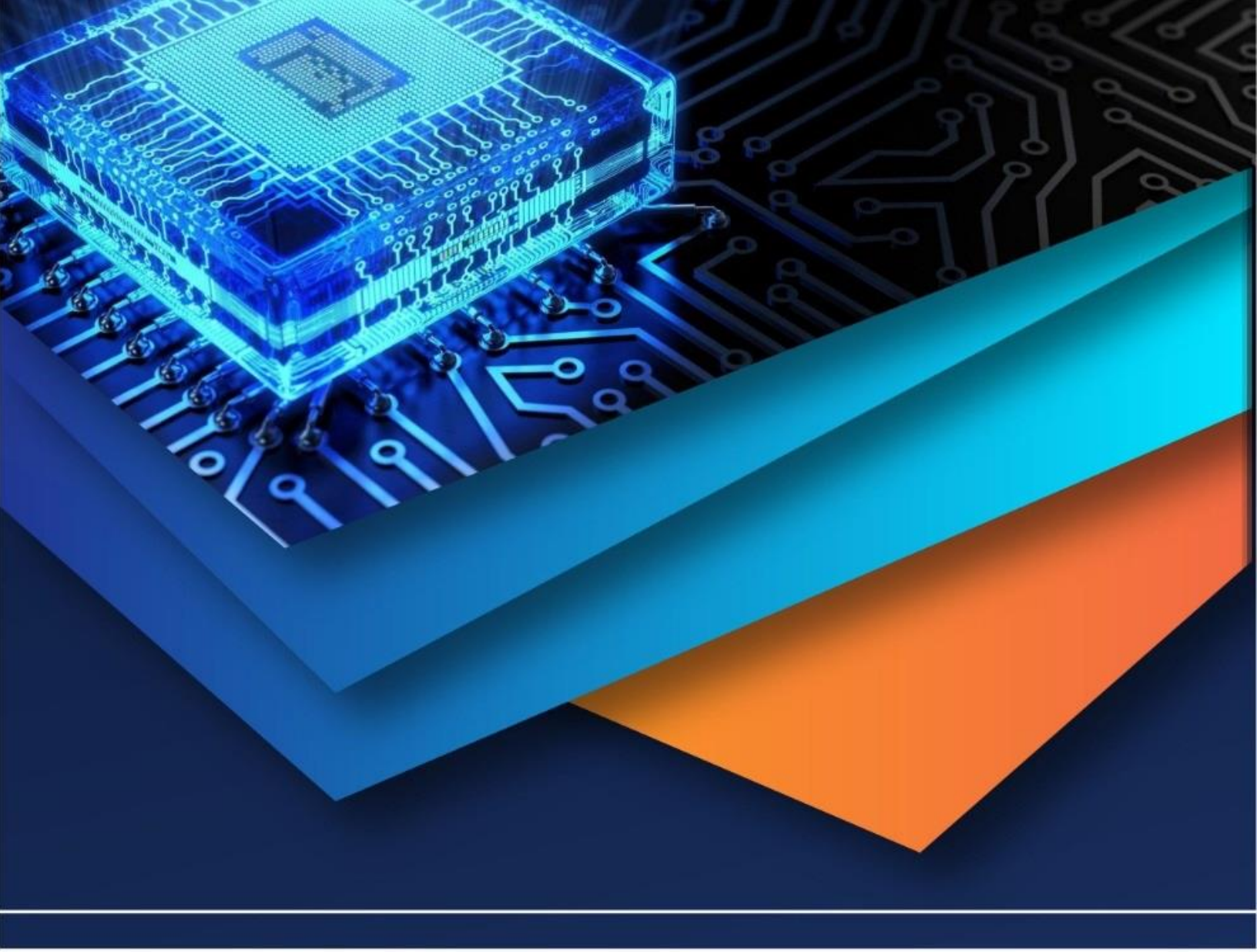
Fig.7 Pregnancies distribution seen in Dataset

VI. CONCLUSION

Logistic Regression has evidently shown to be one of the most effective algorithms to build the predictive model for diabetes.
Study also conveys that choice of algorithm is not only what is required for higher accuracy, but also other factors too. Here other factors include, Data pre-processing, removal and replacement of null values, training and testing the data and many more.
Detecting diseases at earlier stages can help to be treated more easily and effectively [3]. This proposed system serves the exact need of the time in many developing regions. System has successfully solved one of the crucial problems by giving efficient, quicker and higher accuracy of the prediction model than other algorithms and made the model ready to process and adapt new datasets. System is now a platform for intelligence and knowledge prediction in real time handling of larger volume of data [3].
The goal of this research deals with the study of diabetic treatment which may give in healthcare industry by analyzing the data. This system can mainly focus on the patients in the rural areas [3]. Patients there can be treated at a low cost as the prediction will be done in less time compared to the current system.
This system can further be developed to find how likely are the non-diabetic patients to become diabetic in coming years.

REFERENCES

[1] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC, 978-1-5090-3243-3, 2017.
[2] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7, 2017.
[3] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing, 2015.
[4] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang,Type 2 diabetes mellitus prediction model based on data mining, Informatics in Medicine Unlocked, Volume 10, 2018, Pages 100-107, ISSN 2352-9148
[5] Changsheng Zhu, Christian Uwa Idemudia, Wenfang Feng,Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques, Informatics in Medicine Unlocked, Volume 17, 2019, 100179,ISSN 2352-9148
[6] Temurtas, H., Yumusak, N., Temurtas, F., "A comparative study on diabetes disease diagnosis using neural networks", Expert Syst, Vol. 36, pp. 8610–15, 2009.
[7] Chavey, A., Kioon, M., Bailbé, D., "Programming Of Beta-Cell Disorders And Intergenerational Risk Of Type 2 Diabetes Diabetes", Maternal Diabetes, Vol.40, No.5, pp. 323-30, 2014.
[8] Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus, Omar Kassem Khalil Aissa Boudjella, 2016 Sixth International Conference on Developments in eSystems Engineering.
[9] Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY