iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.43986

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 VI June 2022- Available at www.ijraset.com

URL Based Phishing Website Detection by Using Gradient and Catboost Algorithms

B. Deekshitha[1], Ch. Aswitha[2], Ch. Shyam Sundar[3], A. Kavya Deepthi[4]
[1, 2, 3, 4]*Computer Science and Engineering Department, Lendi Institute Of Engineering and Technology(Affiliated to JNTUK), Jonnada, Vizianagaram, Andhra Pradesh, India*

Abstract: ***Phishing is one of the most common and most dangerous attacks among cybercrimes. The aim of these attacks is to steal the information used by individuals and organizations to conduct transactions. Phishing websites contain various hints among their contents and web browser-based information. In existing system the Random forest algorithm is used. In our proposed system, we are using different classification algorithm like bagging and boosting algorithms that are Gradient Boosting, Cat boosting to increase accuracy. The features extracted based on the features of websites in UC Irvine Machine Learning Repository. Here, we have performed the performance analysis between the boosting algorithms like Gradient boost, Cat boost and the random forest. From the performance analysis we can determine the best suitable algorithm to detect the phishing website .This study is considered to be an applicable design in automated systems with high performing classification against the phishing activity of websites.***

Keywords: Gradient boosting, Cat boost, Random forest, Machine learning.

I. INTRODUCTION

1) Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without beingexplicitly programmed. It focuses on the development of computer programs that can access data and use it learn for themselves. Machine learning algorithms are often categorized as supervised or unsupervised. Supervised algorithms require a data scientist or data analyst with machine learning skills to provide both input and desired output, in addition to furnishing feedback about the accuracy of predictions during algorithm training. Data scientists determine which variables, or features, the model should analyze and use to develop predictions. Once training is complete, the algorithm will apply what was learned to new data. Machine learning algorithms are often categorized as

2) Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

- Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output Y = f(X). The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.
- Techniques of Supervised Machine Learning algorithms include linear and logistic regression, multi-class classification, Decision Trees and support vector machines. Supervised learning problems can be further grouped into Regression and Classification problems. The difference between these two is the fact that the dependent attribute is numerical for regression and categorical for classification.

3) Regression: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

4) Classification: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. In short Classification either predicts categorical class labels or classifies data (construct a model) based on the training set and the values (class labels) in classifying attributes

and uses it in classifying new data. There are a number of classification models. Classification models include logistic regression, decision tree, random forest, gradient-boosted tree, multilayer perceptron, one-vs-rest, and Naive Bayes.

5) Unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data. Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Unsupervised learning is classified into two categories of algorithms:

6) Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

7) Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

A. Applications of Machine Learning

- Web Search Engine
- Photo tagging Applications
- Spam Detector
- Credit card fraud detection.
- Analysis of the stock market.
- Automated diagnostic procedures.

B. Challenges to Machine Learning

- Time Consuming Implementation
- Affordability
- Over-fitting of Training Data

C. Applications of Machine Learning

- Speech Recognition – Speech to text
- Traffic prediction - Real Time location of the vehicle
- Virtual Personal Assistant Google assistant, Alexa
- Online Fraud Detection- Detection of fake accounts, fake ids

D. Project Deliverables

- Project Information
- Project Documentation
- Proposed System
- Requirements List
- Program

E. Project Scope

- Internet use has become an essential part of our daily activities as a result of rapidly growing technology. Due to this rapid growth of technology and intensive use of digital systems, data security of these systems has gained great importance. The primary objective of maintaining security in information technologies is to ensure that necessary precautions are taken against threats and dangers likely to be faced by users during the use of these technologies. Phishing is defined as imitating reliable websites in order to obtain the proprietary information entered into websites every day for various purposes, such as usernames, passwords and citizenship numbers. Phishing websites contain various hints.
- Among their contents and web browser-based information. Individual(s) committing the fraud sends the fake website or e-mail information to the target address as if it comes from an organization, bank or any other reliable source that performs reliable transactions.

- Contents of the website includes requests aiming to lure the individuals to enter or update their personal information or to change their passwords as well as links to websites that look like exact copies of the websites of the organizations concerned. Phishing Web sites Features Many articles have been published about how to predict the phishing websites by using artificial intelligence techniques. We examined phishing websites and extracted features of these web sites. We need these features in order to explain phishing attacks characterization.

II. BACKGROUND AND RELATED WORK

A. *Altyeb Taha*

"Intelligent Ensemble Learning Approach for Phishing Website Detection Based on Weighted Soft Voting." Ensemble learning combines the predictions of several separate classifiers to obtain a higher performance than a single classifier. This paper proposes a intelligent ensemble learning approach for phishing website detection based on weighted soft voting to enhance the detection of phishing websites

B. *Mohammad, R.M., Thabtah, F. and McCluskey*

"Predicting Phishing Websites Based on Self-Structuring Neural Network".The Artificial Neural Networks (ANN) are computational models inspired by the structure of the brain and aim to simulate human behaviour, such as learning, association, generalization and abstraction when subjected to training. In this paper, an ANN Multilayer Perceptron (MLP) type was applied for websites classification with phishing characteristics. The results obtained encourage the application of an ANN-MLP in the classification of websites with phishing characteristics.

C. *Doyen Sahoo, Chenghao Liu, Steven C.H. Hoi*

Malicious URL, a.k.a. malicious website, is a common and serious threat to cybersecurity. Malicious URLs host unsolicited content (spam, phishing, drive-by exploits, etc.) and lure unsuspecting users to become victims of scams (monetary loss, theft of private information, and malware installation), and cause losses of billions of dollars every year. It is imperative to detect and act on such threats in a timely manner. Traditionally, this detection is done mostly through the usage of blacklists. However, blacklists cannot be exhaustive, and lack the ability to detect newly generated malicious URLs. To improve the generality of malicious URL detectors, machine learning techniques have been explored with increasing attention in recent years. Further, this article provides a timely and comprehensive survey for a range of different audiences, not only for machine learning researchers and engineers in academia, but also for professionals and practitioners in cybersecurity industry, to help them understand the state of the art and facilitate their own research and practical applications.

D. *Alisha Maini; Navan Kakwani; Ranjitha B; Shreya M K; Bharathi R*

Technology is evolving at an exponential rate, and so are human minds. One of the cybercrimes is phishing attacks. Traditional anti-phishing techniques which use blacklists to iterate and check if the URL is legitimate or phishing is not very useful as the phishers can attack using new URLs. Therefore, Machine learning algorithms can be used to train models to learn the semantic differences between legitimate and phishing URLs. To perform classification of legitimate and phishing URLs, eight ML algorithms which are Random Forest, Decision tree, Naive Bayes, AdaBoost, KNN, XGBoost, Support Vector Machines (SVM) and Logistic Regression are trained and tested. To improve the standard of the classification model, an ensemble model is built using the above-mentioned machine learning algorithms. From the results observed, the machine learning algorithms, XGBoost achieved the highest accuracy and the ensemble model achieved an accuracy higher than all individual machine learning models.

III. METHODS AND FUNCTIONING

A. *Machine Learning Algorithm*

Three machine learning classification model Gradient boost classifier, Cat boost classifier and Random forest has been selected to detect phishing websites.

B. *Random Forest*

It is one of the Supervised Algorithm. It is mainly used to perform the Classification and Regression problems. It mainly build's the Decision trees on different samples and takes majority vote on the classification and average in case of Regression.

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

1) The Random Forest is also an Ensemble Learner. The main theme of the ensemble learner is to combine all the multiple classifiers to solve the complex problem and to improve the performance of the model.
2) It is also an ensemble modeling technique that attempts to build a "Strong classifier" from the "number of weak classifiers". It is done by building a model by using weak models in series. Firstly, a model is built from the training data.
3) Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.

C. Gradient Boosting

It is one of the Boosting Technique. The main theme, of the Boosting is to combine all the weak learners together to form the strong model.

- Gradient boosting is a highly robust technique for developing predictive models. It applies to several risk functions and optimizes the accuracy of the model's prediction. It also resolves multicollinearity problems where the correlations among the predictor variables are high.
- Gradient Boosting is an ensemble machine learning algorithm and typically used for solving classification and regression problems. It is easy to use and works well with heterogeneous data and even relatively small data. It essentially creates a strong learner from an ensemble of many weak learners.

D. Cat Boost or Categorical Boosting

It is an open-source boosting library developed by Yandex. In addition to regression and classification, Cat Boost can be used in ranking, recommendation systems, forecasting and even personal assistants.

- Cat Boost is based on gradient boosted decision trees. During training, a set of decision trees is built consecutively. Each successive tree is built with reduced loss compared to the previous trees.

IV. IMPLEMENTATION AND RESULTS

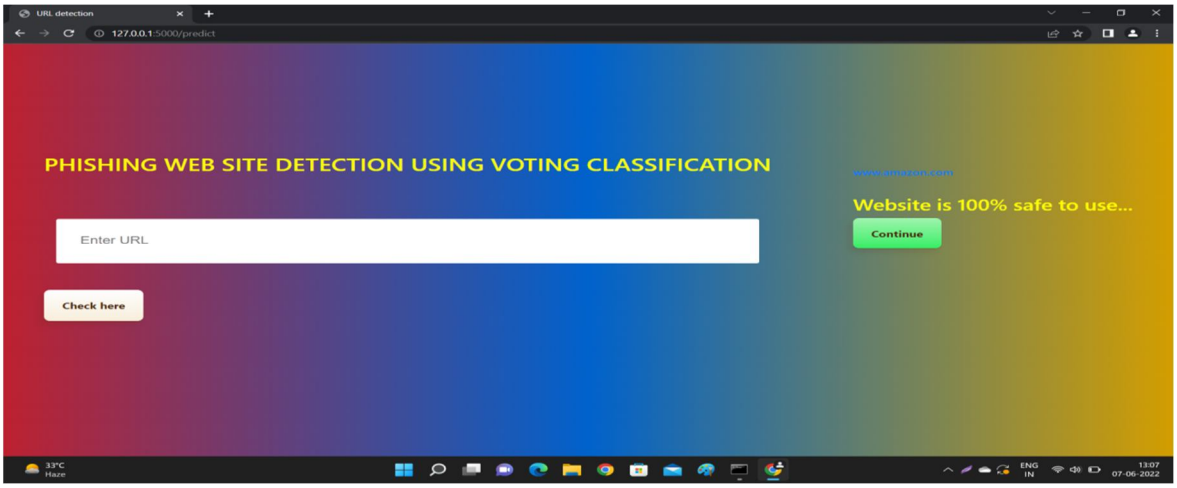
Scikit-learn tool has been used to import Machine learning algorithms. Dataset is divided into training set and testing set in 80:20 ratios respectively. Each classifier is trained using training set and testing set is used to evaluate performance of classifiers. Performance of classifiers has been evaluated by calculating classifier's accuracy score, false negative rate and false positive rate

ML Model	Accuracy	F1_score	Recall	Precision
Gradient boost classifier	0.974	0.977	0.994	0.986
Cat boost classifier	0.972	0.975	0.994	0.989
Random forest	0.976	0.970	0.995	0.988

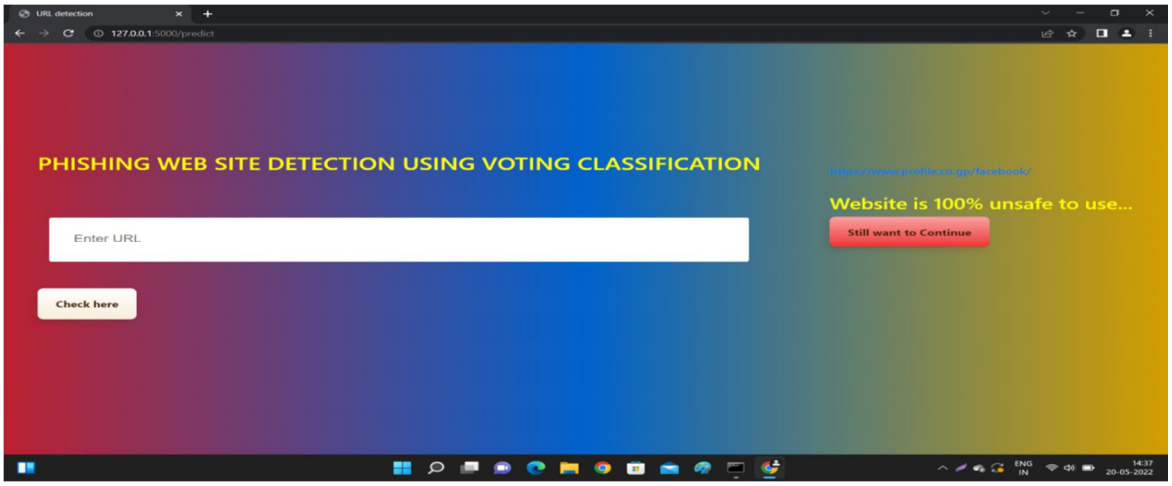
Results shows that Gradient boost classifier gives better detection accuracy which is 97.4 and Cat boost classifier gives detection accuracy which is 97.2% with lowest false negative rate than decision tree and support vector machine algorithms. Result also shows that detection accuracy of phishing websites increases as more dataset used as training dataset. All classifiers perform well when 90% of data used as training dataset.

V. SHOWING HOW MUCH PERCENT A WEBSITE IS SAFE TO USE

This screen presents the results derived from experimental evaluation. These are derived by using the Algorithms used in proposed systems to achieve highest accuracy.



Detection of legitimate website



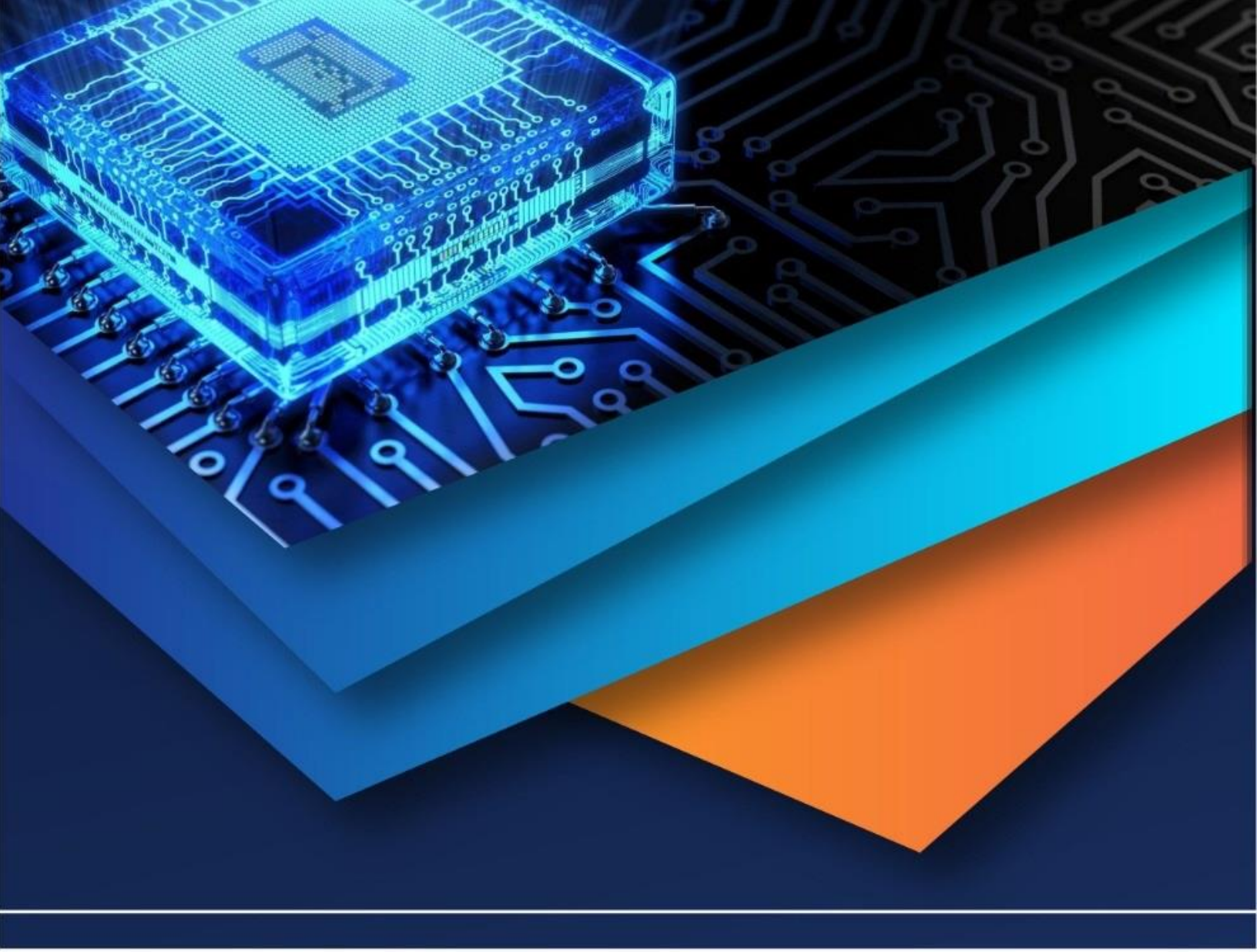
Detection of phishing website

VI. CONCLUSION

Nowadays, phishing websites are increasing rapidly and causing more damage to the users and organizations. It is becoming a biggest threat to people's daily life and the networking environment. In these attacks, the intruder puts on an act as if it is trusted organization with an intention to purloin liable and essential information. Phishing website is a mock website that looks similar in appearance but different in destination. The unsuspected users post their data thinking that these websites come from trusted financial institutions. Hence, there is a need for efficient mechanism for the detection of phishing website. In our project, we developed a model that can be mainly used in determining the website's as either phishing or legitimate by using the features extraction techniques from the URL. These features are compared with the features present in the features extraction dataset and validated accordingly. Here, in our project we applied the algorithms like Gradient Boost, Cat Boost and Random Forest on the model that has been developed. During testing, it has been observed that the system has performed well and as expected. This paper aims to enhance detection method to detect phishing websites using machine learning technology. We achieved 97.4% detection accuracy using Gradient boost classifier and 97.2% using Cat boost classifier with lowest false positive rate. As classifiers give better performance when we used more data as training data. In future hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used.

REFERENCES

[1] "Intelligent Ensemble Learning Approach for Phishing Website Detection Based on Weighted Soft Voting" by Altyeb Taha November 2021.
[2] Mohammad, R.M., Thabtah, F. & McCluskey, L. "Predicting phishing websites based on self-structuring neural network". Neural Comput & Applic 25, 443–458 (2014).
[3] Malicious URL Detection using Machine Learning: A Survey Doyen Sahoo, Chenghao Liu, Steven C.H. Hoi [Submitted on 25 Jan 2017 (v1), last revised 21 Aug 2019 (this version, v3)].
[4] A. Maini, N. Kakwani, R. B, S. M K and B. R, "Improving the Performance of Semantic-Based Phishing Detection System Through Ensemble Learning Method," 2021 IEEE Mysore Sub Section International Conference (MysuruCon), 2021, pp. 463-469.
[5] CatBoost : gradient boosting with categorical features support Anna Veronika Dorogush, Vasily Ershov , Andrey Gulin [v1] Wed, 24 Oct 2018.
[6] Bentéjac, C.Csörgő, A. & Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 54, 1937–1967 (2021).
[7] Singh and Meenu, "Phishing Website Detection Based on Machine Learning: A Survey," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 398-404.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY