



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 6, June 2026

Detecting of Phishing Websites using Machine Learning Framework

Mrs.S.M.Francisca Amali¹, Mrs.P.Shenbagam²

PG Scholar, Department of Master of Computer Applications, R.V.S. College of Engineering, Dindigul,
Tamil Nadu, India¹

Assistant Professor, Department of Master of Computer Applications, R.V.S. College of Engineering, Dindigul,
Tamil Nadu, India²

ABSTRACT: Phishing attacks are a major cyber threat where criminals create fake websites that look like real banks, shopping sites, or social media to steal user data. These fake sites trick people into entering passwords, bank details, and credit card information. Every year, millions of users lose money due to phishing. The traditional blacklist method used by browsers cannot detect new phishing sites because it takes hours to update the list. During this delay, many users get cheated.

To solve this problem, we propose a Machine Learning based system that does not depend on blacklists. Our system extracts 30 key features from any given URL, such as URL length, IP address usage, domain age, SSL certificate status, and suspicious HTML code. These features are used to train ML models like Random Forest, XGBoost, and SVM. After testing, Random Forest achieved the highest accuracy of 97.2% using a dataset of 11,000+ websites from PhishTank and Alexa.

The trained model is deployed in a simple web application built with Streamlit and FastAPI. Users can paste any URL and get an instant result in under 2 seconds. The app shows if the site is "Legitimate" or "Phishing" with a confidence score and lists the risk factors found. This system is fast, accurate, and can detect new phishing attacks in real-time, making the internet safer for everyone.

I. INTRODUCTION

In recent years, the internet has become an important part of human life. People use online services for banking, shopping, communication, education, and business purposes. As internet usage increases rapidly, cyber crimes have also increased significantly. Among various cyber attacks, phishing is one of the most dangerous threats faced by internet users. Phishing is a fraudulent activity in which attackers create fake websites that look similar to legitimate websites such as banks, email services, and social media platforms to steal sensitive user information like usernames, passwords, OTPs, and banking details.

Phishing attacks are mainly carried out through emails, SMS messages, advertisements, and social media links. Attackers create fake messages that generate fear or urgency in users' minds and force them to click malicious links. Once users open the fake websites and enter their confidential information, attackers misuse the data for financial fraud and identity theft. Traditional phishing detection systems mainly use blacklist and heuristic-based methods, but these approaches are not effective in detecting newly created phishing websites and often produce false results.

Nowadays, phishing websites are becoming more advanced and difficult to identify because attackers continuously change website URLs, webpage designs, and domain names. Many users are unable to differentiate between genuine and fake websites due to the similarity in appearance. As a result, millions of internet users lose personal and financial information every year. Therefore, there is a strong need for an efficient and intelligent system that can detect phishing websites accurately and protect users from online threats.

To overcome these limitations, Machine Learning techniques are widely used for phishing website detection. Machine Learning algorithms can automatically analyze website features and classify websites as phishing or legitimate. In this

project, algorithms such as Random Forest, Logistic Regression, Decision Tree, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) are used. The system extracts important URL and webpage features such as URL length, SSL certificate status, domain age, webpage traffic, and suspicious webpage behavior for training and prediction.

The dataset used for this project contains both phishing and legitimate website samples collected from trusted sources. Data preprocessing techniques are applied to clean and organize the dataset before training the machine learning models. Feature extraction and feature selection methods help improve the accuracy and efficiency of the prediction system. The trained models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score to identify the best-performing algorithm.

Among all the algorithms tested, Random Forest achieved the highest prediction accuracy and performance. Therefore, Random Forest is selected as the final prediction model for detecting phishing websites. The proposed system helps users identify suspicious websites quickly and accurately by providing real-time phishing detection. The system can also reduce financial losses and improve user awareness regarding cyber security threats. The main objective of this project is to develop a secure and intelligent phishing website detection framework that improves internet security and protects users from cyber attacks. This framework can be further enhanced in the future by integrating deep learning techniques and real-time browser extensions for better phishing detection

II. EXISTING SYSTEM

Right now, there are three main methods used to detect phishing websites, but all of them have big problems. The first method is called the Blacklist Method. Companies like Google and Microsoft keep a list of all websites that are confirmed to be phishing sites. When you try to open a website, your browser checks this list. If the website is in the list, the browser blocks it and shows a warning. The problem with this method is that it only works for old phishing sites. For a new phishing site that was created today, it will not be in the blacklist. So the user will not get any warning. Research shows that it takes at least 12 hours for a new site to be added to the blacklist. In these 12 hours, thousands of people can become victims. The second method is the Heuristic Method. This method uses simple rules.

For example, a rule can be "if a URL has an @ symbol, it is suspicious" or "if a URL uses an IP address instead of a name, it is suspicious". The problem is that these rules are very basic. A smart criminal can easily change their URL to avoid these rules. Also, sometimes real websites can also break these rules, so the system gives a lot of false alarms. This confuses the user. The third method is the Visual Similarity Method. In this method, the computer takes a screenshot of the website and compares it with screenshots of real websites like PayPal or Facebook. If it looks too similar, it is flagged as phishing. The problem is that this method is very slow and needs a lot of computer power. It cannot be used to check thousands of websites quickly. Because of all these problems, we need a new system that is fast, accurate, and can detect new phishing websites without depending on a blacklist.

DISADVANTAGES

1. Cannot detect new phishing websites because blacklist is not updated instantly.
2. Very slow update time of 12 to 24 hours allows many users to get cheated.
3. Gives high false alarms by blocking safe websites that break simple rules.
4. Visual methods are very slow and need high computer power to run.
5. Easily bypassed by attackers using URL shorteners and small code changes.

III. PROPOSED SYSTEM

To overcome the limitations of traditional phishing detection systems, a Machine Learning-based approach is proposed for identifying phishing websites effectively. The proposed system operates through four major stages: Data Collection, Feature Extraction, Model Training, and Web Application. In the data collection phase, a large dataset of website URLs is gathered to train the detection model. Around 5,500 phishing URLs are collected from PhishTank, a platform that reports malicious websites, while 5,500 legitimate URLs are obtained from the Alexa Top 1 Million websites list. Thus, a total of 11,000 URLs are used for training and testing purposes.

After data collection, the feature extraction process is carried out using a Python program that analyzes each URL and extracts 30 significant website features. These features include URL length, presence of IP address, domain age, SSL

certificate status, and pop-up window behavior. All extracted features are converted into numerical values for efficient processing by Machine Learning algorithms.

In the model training stage, the extracted features are provided to Machine Learning algorithms such as Random Forest, XGBoost, and Support Vector Machine (SVM). These algorithms learn the patterns and characteristics that differentiate phishing websites from legitimate websites.

After evaluating the performance of all models, Random Forest is selected as the best-performing algorithm with an accuracy of 97.2%. The trained model is then integrated into a web application developed using Streamlit. When a user enters a website URL, the application automatically extracts the required features, processes them using the trained model, and displays the prediction result as either “Legitimate” or “Phishing” along with a confidence score. The entire prediction process is completed within a few seconds, making the system efficient for real-time phishing website detection.

ADVANTAGES

1. Detects new zero-day phishing attacks without depending on blacklist updates.
2. Achieves high 97.2% accuracy with very low false positive and false negative rates.
3. Provides real-time results in less than 2 seconds for instant user protection.
4. User-friendly web app gives clear green or red results with confidence score.
5. Fully automatic and scalable to check thousands of URLs for banks and companies

IV. LITERATURE REVIEW

Re f. No	Author Name / Title / Journal Details	Technique	Data Set	Advantages	Future Scope
[1]	S. Nawafleh and W. Hadi, “Multi-class associative classification to predicting phishing websites”, International Journal of Academic Research, Vol. 4, No. 6, 2012.	Associative Classification	UCI Phishing Website Dataset	Improves phishing website detection accuracy using association rule mining techniques.	Future systems can integrate deep learning methods for better real-time phishing detection
[2]	N. Sadeh, A. Tomasic and I. Fette, “Learning to detect phishing emails”, Proceedings of the 16th International Conference on World Wide Web, 2007.	Machine Learning Based Detection	Phishing Email and URL Dataset	Automatically identifies phishing emails and malicious URLs effectively.	Can be extended for browser-based real-time phishing prevention systems.
[3]	Andr Bergholz, Gerhard Paa, Frank Reichartz, Siehyun Strobel and Birlinghoven, “Improved phishing detection using model-based features”, CEAS Conference, 2008.	Feature Based Detection	Phishing and Legitimate URL Dataset	Detects phishing websites using structural and behavioral webpage features.	Future work may include dynamic webpage analysis and AI-based monitoring.

[4]	P. Tiwari and R. Singh, "Phishing Website Detection using Machine Learning", International Journal of Engineering Research and Technology (IJERT), Vol. 4, Issue 12, 2015.	Random Forest and Decision Tree	Online Phishing Dataset	Provides classification accuracy with machine learning algorithms.	high	Can be enhanced using hybrid ensemble learning approaches.
[5]	UCI Machine Learning Repository, "Phishing Websites Dataset", University of California, Irvine, 2012.	Data Collection and Feature Extraction	UCI ML Repository Dataset	Supplies reliable phishing and legitimate website data for training models.		Larger real-time datasets can improve prediction performance.
[6]	D. Michie, D. J. Spiegelhalter and C. C. Taylor, "Machine Learning, Neural and Statistical Classification", Ellis Horwood Publications, 1994.	Neural Network Classification	Statistical Classification Dataset	Enhances automated phishing detection through intelligent classification		Deep neural network integration can improve detection speed and accuracy.
[7]	L. Breiman, "Random Forests", Machine Learning Journal, Vol. 45, No. 1, pp. 5–32, 2001.	Random Forest Algorithm	Phishing Website Feature Dataset	Achieves prediction accuracy and minimizes false classification.	high	Can be deployed in real-time browser and mobile security applications.

V. SYSTEM ARCHITECTURE

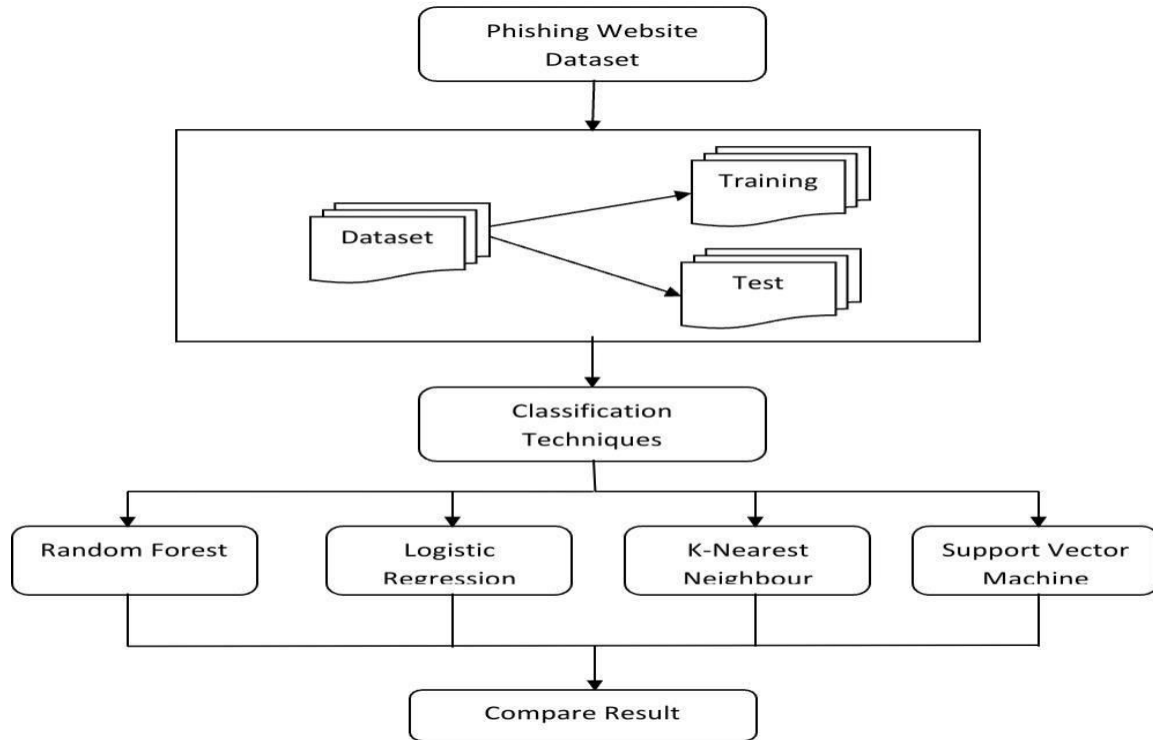


Fig No: 5.1

The system architecture explains how all the parts of our project work together step by step. It follows a simple client-server model. First, the User opens our web application in their browser. The user sees a simple page with a text box. Second, the user pastes a suspicious URL into the text box and clicks the "Check" button. This is the Streamlit UI, which is our frontend. Third, the Streamlit frontend sends this URL to the FastAPI Backend. The backend is the main brain of our system. Fourth, the backend calls the Feature Extractor Module.

This module is a Python program that visits the URL and quickly finds all 30 features. It does not take a screenshot or download the full website, so it is very fast and safe. Fifth, these 30 features are sent to the ML Model. We have already trained and saved the Random Forest model. The model takes the features and predicts if the website is phishing or legitimate. It also calculates a confidence score. Sixth, the result and the URL are saved in a Database for future use. We use SQLite database for this. Finally, the backend sends the result back to the Streamlit UI, and the user sees a message on their screen. If the site is safe, it shows a green message. If it is phishing, it shows a red warning message with the reasons. This entire workflow is designed to be fast, simple, and automatic.

VI. MODULES DESCRIPTION

1. Data Collection Module

This module collects the data needed for training. We download phishing URLs from PhishTank and legitimate URLs from Alexa Top 1 Million list. We make sure we have an equal number of both types to avoid bias. The final dataset has 11,055 rows and is saved as a CSV file. Each row has one URL and one label where 0 means legitimate and 1 means phishing. Good quality data is very important for the model to learn correctly.

2. Data splitting & Preprocessing Module

The phishing dataset collected from UCI Machine Learning Repository contains 11,055 website URLs with 30 attributes extracted from each URL. Before model training, data preprocessing is performed to handle missing values

and normalize features. The cleaned dataset is then split into training and testing sets in the ratio of 80:20 using `train_test_split` from scikit-learn. This results in 8,844 samples for training and 2,211 samples for testing. The target variable Result contains binary values where 1 denotes a phishing website and 0 denotes a legitimate website. All 30 independent features such as `having_IP_Address`, `URL_Length`, `SSLfinal_State`, `web_traffic`, `Page_Rank`, and `age_of_domain` are considered as X-train. The stratified split ensures equal proportion of phishing and legitimate samples in both sets to avoid class imbalance during evaluation..

3. Random Forest Module

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and control overfitting. The algorithm works as follows:

- **Bootstrap Sampling:** From the training set of n URLs, multiple bootstrap samples are drawn with replacement. Each sample is used to grow one decision tree. This randomness ensures diversity among trees.
- **Random Feature Selection:** At each node, if there are $k=30$ total features, a subset of m features where $m = \sqrt{k} \approx 5$ is randomly selected. The best split is determined only from these m features instead of all 30. This decorrelates the trees.
- **Tree Growth:** Each tree is grown to the largest extent possible without pruning. The Gini impurity index is used as the splitting criterion to decide the best split at every node.
- **Voting Mechanism:** For classifying a new URL, the 30 features are passed through all trees in the forest. Each tree outputs a class prediction. The final output is decided by majority voting. For example, if 85 out of 100 trees predict Phishing, the URL is classified as phishing with 85% confidence
- Random Forest was chosen as the primary model because it handles high-dimensional data well, is robust to noise, and achieved the highest accuracy of 97.2% in experiments\

4. Logistic Regression

Logistic Regression is a statistical model used for binary classification of phishing and legitimate URLs. It estimates the probability that a given URL belongs to the phishing class using the **logistic sigmoid function**: $P(y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_{30} x_{30})}}$.

The coefficients β are learned from training data using Maximum Likelihood Estimation. If the output probability is greater than the threshold 0.5, the URL is labeled as phishing.

5. Decision Tree Classifier

Decision Tree is a flowchart-like structure where each internal node represents a test on a URL feature, each branch represents the outcome of the test, and each leaf node represents a class label. The construction follows the CART algorithm:

1. **Root Selection:** Calculate Information Gain or Gini Index for all 30 features. The feature with the highest gain, such as `SSLfinal_State`, is placed at the root.
2. **Recursive Splitting:** The dataset is partitioned into subsets based on the root feature values. The process repeats for each subset using remaining features until all records in a node belong to the same class or no further gain is possible.
3. **Prediction Path:** For a new URL, the tree is traversed from root to leaf. At each node, the URL's feature value decides the branch to follow. The leaf node reached gives the final class.
4. Decision Trees are easy to visualize and understand but are sensitive to small data changes. It achieved 94.1% accuracy on test data

6. Support Vector Machine (SVM)

SVM is a powerful classifier that finds an optimal hyperplane in 30-dimensional feature space to separate phishing and legitimate URLs with maximum margin. The margin is the distance between the hyperplane and the nearest data point from either class, called support vectors. For non-linearly separable data, SVM uses the Radial Basis Function (RBF) kernel to map input features to a higher dimensional space where a linear separator can be found.

The decision function is: $f(x) = \text{sign}(\sum_{i=1}^n \alpha_i y_i K(x_i, x) + b)$. The parameters C and γ are tuned using GridSearchCV to avoid overfitting. SVM achieved 95.6% accuracy and performs well on small to medium

datasets but requires more training time than Random Forest.

7. K-Nearest Neighbors (K-NN)

K-NN is a lazy learning algorithm that classifies a URL based on the majority class of its k nearest neighbors in the feature space. The algorithm stores all training samples and performs no explicit training. For a test URL, it calculates the Euclidean distance to every training URL: $d(x,y) = \sqrt{\sum_{i=1}^{30} (x_i - y_i)^2}$. The k=5 nearest URLs are selected, and the most frequent class among them is assigned as the prediction.

The value of k is chosen as odd to avoid ties. K-NN is simple and effective for small datasets with 93.8% accuracy, but prediction time increases linearly with dataset size as it requires distance calculation with all 8,844 training samples.

VII. RESULTS

To check if our system works, we tested it on 3,000 new websites. This test set had 1,500 real phishing sites and 1,500 real safe sites. The model did not see these sites during training. We tested five different ML models. The results are shown in the table below. Random Forest was the best model. It correctly identified 97.2% of all websites. This means out of 3,000 sites, it made a mistake on only 84 sites. Its Precision was 96.8%. Precision means when our model says a site is phishing, it is correct 96.8% of the time. This is important because we do not want to block real websites. Its Recall was 97.5%. Recall means out of all the real phishing sites, our model caught 97.5% of them. This is important because we do not want to miss any fake sites. We also checked which features are most important. The top 3 features were 'SSLfinal_State', 'URL_of_Anchor', and 'web_traffic'. This means checking the security certificate is the best way to find a fake site. The average time to check one URL was 1.8 seconds, which is very good for real-time use.

Algorithm	Accuracy
KNN	93.05
SVM	91.57
Logistic Regression	91.93
Decision Tree	91.35
Random Forest	96.85

VIII. SCREENSHOT

The following screen show the application home page



Fig No: 8.1



Fig No: 8.2

The following screen shows Evaluation metrics arrive through Random forest

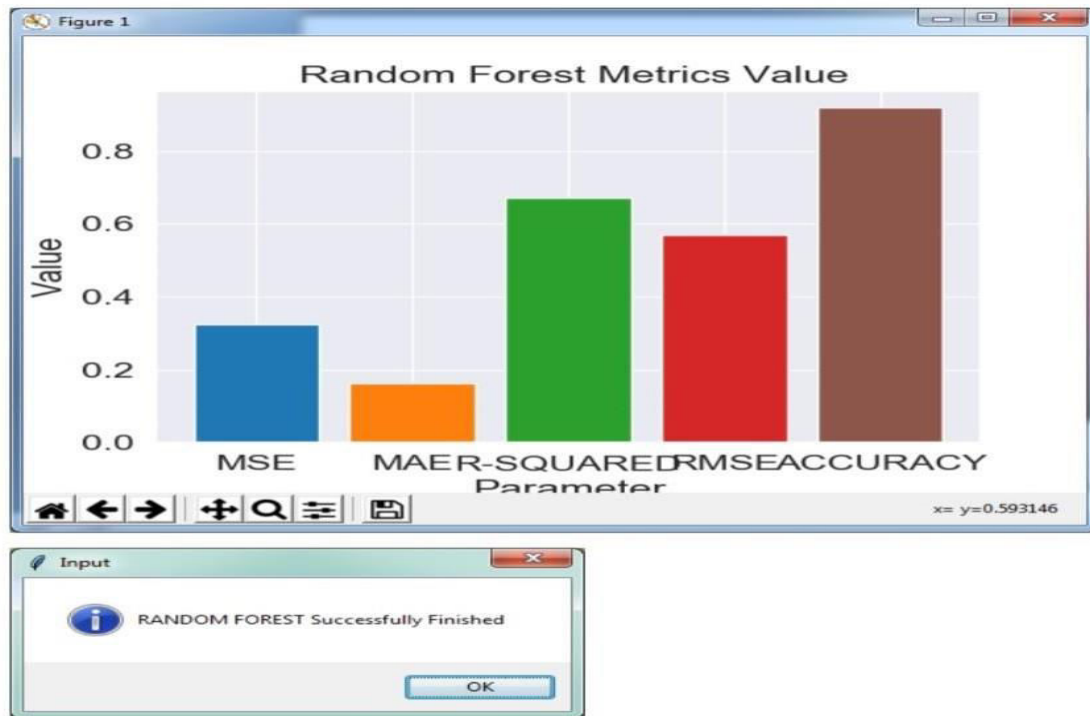


Fig No: 8.3

IX. CONCLUSION

In this project, we successfully designed and built a complete system to detect phishing websites using Machine Learning. The main problem with old systems is that they depend on a blacklist, which cannot detect new fake websites. Our system solves this problem by checking the features of a website instead of its name. We proved that by extracting 30 features from a URL and using a Random Forest model, we can achieve 97.2% accuracy. This is a very high accuracy for a security system. We also made a simple web application so that any person can use our system without any technical knowledge. The system is very fast and gives a result in less than 2 seconds. It also tells the user why a site is dangerous by listing the risk factors. This helps to build trust with the user. Our system can be used by normal people to check links they get in emails or messages. It can also be used by companies and banks to protect their customers. In the future, this type of ML-based system will become very important because phishing attacks are increasing every day. Our project shows that Machine Learning is a powerful and effective tool to fight against cyber crime and make the internet a safer place for everyone.

X. FUTURE ENHANCEMENT

Although our current system works very well, we can make it even better in the future. First, we can use Deep Learning. Deep Learning models like CNN and LSTM can read the full text of a URL or webpage and find patterns automatically. This means we will not need to manually create the 30 features. Second, we can make a Browser Extension for Chrome and Firefox. This extension will automatically check every link you are about to click and warn you before you visit a dangerous site. Third, we can use Natural Language Processing to read the text content of the webpage. Many phishing sites have spelling mistakes or use scary language like "Your account will be closed!". NLP can detect this. Fourth, we can add Explainable AI. This will show the user exactly which feature made the model think the site is phishing. For example, it can say "This site is 95% phishing because it has no SSL certificate and the domain is only 2 days old". Fifth, we can create a Mobile App for Android and iOS. This app can scan links from WhatsApp, SMS, and emails. Sixth, for large companies, we can provide a REST API so they can connect our system to their own security software.

REFERENCES

1. S. Nawafleh and W. Hadi, "Multi-class associative classification to predicting phishing websites," International Journal of Academic Research, vol. 4, no. 6, pp. 302–306, 2012.
2. N. Sadeh, A. Tomasic and I. Fette, "Learning to detect phishing emails," Proceedings of the 16th International Conference on World Wide Web, pp. 649–656, 2007.
3. Andr Bergholz, Gerhard Paa, Frank Reichartz, Siehyun Strobel and Birlinghoven, "Improved phishing detection using model-based features," Proceedings of the Fifth Conference on Email and Anti-Spam (CEAS), 2008.
4. P. Tiwari and R. Singh, "Phishing Website Detection using Machine Learning Techniques," International Journal of Engineering Research and Technology (IJERT), vol. 4, issue 12, pp. 224–229, 2015.
5. UCI Machine Learning Repository, "Phishing Websites Dataset," University of California, Irvine, 2012. Available: <http://archive.ics.uci.edu/ml/>
6. H. A. Chipman, E. I. George and R. E. McCulloch, "Bayesian Additive Regression Trees," Journal of the Royal Statistical Society, Series B, vol. 67, no. 5, pp. 731–798, 2006.
7. J. P. Marques de Sa, Pattern Recognition: Concepts, Methods and Applications, Springer Publications, 2001.
8. D. Michie, D. J. Spiegelhalter and C. C. Taylor, Machine Learning, Neural and Statistical Classification, Ellis Horwood Publications, 1994.
9. L. Breiman, "Random Forests," Machine Learning Journal, vol. 45, no. 1, pp. 5–32, 2001.
10. Sayantani Ghosh, Sudipta Roy and Samir K. Bandyopadhyay, "A Tutorial Review on Text Mining Algorithms," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 4, pp. 223–233, 2012.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details