A Dynamic and Combined Phishing Detection Technique

D. Deva Hema, C. Akshaya, Talluri Venkata Sai Sumanth, Diganta Saha

***Abstract:** With the quick advancement of web applications, internet users are spending more and more time with these applications .They utilize the benefits of the internet in doing all the day-to-day chores from reading newspaper to grocery shopping .This makes them prone to various kinds of cyber-attacks such as phishing , password attack , malwares etc...Phishing is one of the most common cyber-attack which is made by the attackers to take the users' delicate data . In phishing attack the users are first tempted with attractive offers and are then redirected to illegitimate (phishing) websites which ask for their credentials .In spite of the alert and awareness spread against these types of cyber-attacks , people continue to fall prey and get affected .The attackers have evolved with time and craft the attacks in such a way that the phishing websites and emails may seem real .Many systems and algorithms have been developed to predict phishing attacks .However ,the achievement rate of phishing attacks stays high and it's detection is prone towards high true negative and false positive ratio. Therefore ,to deal with this conundrum we are putting forward a generalized algorithm for phishing detection with improved accuracy.*

***Keywords:**Phishing ,cyber-attack ,websites ,email ,false-positive ,true-negative*

I. INTRODUCTION

The Phishing is the illicit use of electronic communication to trap user and take advantage of them. It's a cyber attack in which the victims are approached by electronic mail, telephone or text message by an attacker masquerading as a legal organization or person to entrap people into giving delicate information such as individually identifiable information, net-banking and credit card data, and passphrases. These confidential data like PIN are used to access and cyber jack delicate accounts of the victim thereby causing personality theft and monetary loss .The cyber attackers use social engineering to convince the victims to click on malicious links or attachments or make them reveal sensitive information .This is one of the oldest technique of cyber-attack dating back to early 90s but is still the most effective and widespread technique to lure victims.

TYPES OF PHISHING

There are several different types of phishing attacks which the attackers use to target specific group of users. These are:

- Email phishing: In this phishing attack the attacker impersonates as a legitimate and well known organization and send fake (phishing) emails to thousands of internet users .These mails usually contain links which on clicking directs the user to a fake websites .These websites closely resemble the original ones thus tricking the users to give away their credentials. In general ,these mails always indicate that they need immediate and urgent responses.
- Search Engine Phishing: Search engine phishing is where a fake webpage is created and specific keywords that would masquerade as keywords of a genuine webpage are targeted while waiting for the searcher to land on the fake webpage.
- Clone Phishing: The clone phishing attack is a phishing technique where the hacker makes use of the legitimate messages that the victim may have already received to create or clone a malicious version of it. This technique creates a virtual replica of the genuine message and sends the fake message from an email address that resembles the genuine mail id.
- HTTP Phishing: In this method the hackers send an email which just has a legal-looking link in the email body.There is no other content except for the link itself.
- Spear Phishing :This is a targeted attack .Unlike the general phishing attack in which large number of people are sent generalized spam emails ,in this attack individuals from specific organization are targeted .The attackers use social engineering to study their interest and send mails to trap them.
- Watering Hole Phishing:It is a lesser known form of phishing attack.These attacks target businesses by identifying the websites the company or employees visit the most and infect them.
- URL phishing: In URL phishing attacks, attackers use the phishing page's URL to infect the target.

These different types of phishing attack are a big threat to the users' privacy and sensitive information. Thus, machine learning and deep learning based prediction algorithms have been developed to limit the success rate of these attacks.
Machine learning is a fast developing and one of the most promising field of AI , that is evolving as a crucial and powerful technology for the future. Machine Learning uses AI to enable systems to learn by itself and develop an ability to read and analyze data to provide outputs.

Revised Manuscript Received on March 30, 2020.
* Correspondence Author

D. Deva Hema*, Assistant Professor, Department of Computer Science and Engineering in SRMIST, Chennai, Tamil Nadu, India. Email: devahema2010@gmail.com

C. Akshaya, Department of Computer Science and Engineering,SRM IST, Chennai, Tamil Nadu, India. Email: akshaya1998@gmail.com

Talluri Venkata Sai Sumanth, Department of Computer Science and Engineering, SRMIST, Chennai, Tamil Nadu, India. Email:tallurisaisumanth77@gmail.com

Diganta Saha, Department of Computer Science and Engineering,SRM IST, Chennai, Tamil Nadu, India. Email: digantasaha99@gmail.com

Retrieval Number: E2819039520/2020©BEIESP
DOI: 10.35940/ijitee.E2819.039520
Journal Website: www.ijitee.org

The process of ML is striking similar to the way the humans learn and develop the ability to think logicaly.The logical thinking ability is then automatically refined and sharpened by the Human brain based on the experience, to apply it to various inputs / situations , analyse them to arrive at conclusions. Machine learning methods are used to spontaneously find the underlying sequences within composite data that we would have difficulty to find out. The hidden sequences and information related to a problem can be used to determine further events and perform various complex decision making.

Machine learning uses several techniques automatically to recognize a particular pattern or trend in a complex and voluminous set of data .Then by using the analytical ability they can predict an event (say the probability of rain or on time arrival of train or probability of getting a confirmed reservation etc..) which can be usedby humans to take better decisions .They can even provide decisions where the complexity limits human decision making ability.

Machine learning process: The following terminologies are key to ML.

Dataset: A collection of information samples, that contain attributes necessary to solving the problem.

Features: Necessary sets of data that facilitate in understanding a problem. These are supplied in to a Machine Learning algorithm to help it learn.

Model: The demonstration of a characteristic that a Machine Learning algorithm has learnt. It understands this from the information it is shown during training. The model is the output you get after training the algorithm.

Data Collection: Collection of the information that the algorithm will learn from.

Data Preparation: Arrangement and architect the information into the ideal configuration, extricating significant highlights and performing dimensionality decrease.

Training: This is where the Machine Learning algorithm really learns by analyzing the information that has been gathered and prepared.

Evaluation: It is the stage when the model is fed with a set of test data to see how well it performs.In this case this output is compared with the actual /expected output

Tuning: Based on the accuracy of the test output the model is fine tuned to maximise the performance.

Methods :There are several approaches to train a model:

- Supervised Learning :In this approach you provide the machine with a set of prepared data with inputs and outputs.The machine learns the logical association among the input and output and uses the learning to provide outputs for test data or problem data. Supervised learning comes with the possibility of bias built in. This is similar to class room teaching where a teacher guides the student to problem solving
- Unsupervised Learning:In this approach the model is only given sets of data and allowed to learn by itself. Though this is a more difficult approach it eliminates the trainer bias and provides better analytical results.This approach is mainly used for clustering
- Semi - Supervised Learning:This approach is a combination among unsupervised and supervised methodologies. The training method isn't closely supervised with sample outputs for every input, but we also don't let the algorithm perform its own methods and provide no feedback.

Reinforcement Learning This approach encourages the machine to learn by rewarding .This is like behavior molding in humans when the expected behavior is rewarded thereby reinforcing the behavioural pattern.

II. LITERATURE SURVEY

The increased use of internet and its applications has increased the opportunities for the cyber attackers to steal users' data and misuse them .Phishing attack has become the most prevalent and successful technique adopted by the cyber criminals to entrap the users .Many predicting mechanisms and algorithms using machine learning have been developed for predicting various types of possible phishing attacks have been proposed by researchers .

In this paper [1], an effective mechanism to predict email phishing using deep learning algorithm called THEMIS has been proposed. In this technique, email body and email headers are modeled at character and word level and RCNN is used to predict the phishing email. Authors[2], proposed a methodology called SAFE-PC to predict newer kinds of email phishing attacks, where in, email headers and email body are extracted and RUSBoost classifier is trained for prediction of phishing emails. The paper authored by [3] suggests, a mechanism to detect phishing websites using the Google's PageRank and eight other features .The features used are aging of name of website, known images, distrustful URL, distrustful links, IP address, dots in URL and forms. In this paper [4],a "case based reasoning –phishing detection system"(CBR-PDS) has been introduced .This system is checked against simple URL characteristics and intra-URL relatedness features to predict genuineness of URL. This paper[5], describes in detail a competent methodology for phishing detection based on PNNs and K-medoids clusters. In the paper [6],a system in which a "Multi-Agent System"(MAS) , an adaptive intelligence technique is superimposed above the distributed "Case-Based Reasoning Phishing Detection Systems"(CBR-PDSs). The addition of multi-agent system to CBR-PDS provides a different efficient method to analyze phishing attacks in a greater scale. A mixed model to detect phishing has been put forward by performing classification using RF, SMO, J48, BN, NB and IBK models and selecting the best three based on performance and accuracy to build the hybrid model. [7]. Feature selection procedures like Information Gain, Information Gain Ratio, ChiSquare, and Correlation-Based Feature Selection have been studied and it has been concluded by the authors [8],that though they reduces the accuracy of classification algorithms ,they also reduce the computational time of these algorithms.Authors [9] proposed a method for classifying phishing web pages with reduced error rate containing speed and time - accurate efficiency. This is achieved by blending Polynomial neural network with genetic algorithm and optimizing the predicting techniques. Authors[10] have proposed a method to use random forest classifier for phishing detection. Classifying accuracy, region under receiver operating characteristic(ROC) curves and F-measure is used to evaluate the performance of the technique. In this paper[11] ,phishing URL differentiation in distributed cloud environment has been done using MapReduce framework.

Multi - Modal images and textual features extraction and classification is done by employing MapReduce framework in cloud environment.

The accuracy of prediction mechanism depends on the selection of the most effective and essential features that are extracted from the website, URL and email.

The effectiveness of selection will substantially reduce false positive and true negative ratios .Many feature selection methods have been studied and deployed by the researchers in their prediction model. A procedure of feature selection by mixing the adaptive F-Ratio principles with enhanced genetic algorithm has been proposed by authors[12], which selects the best subset of attributes using the method. Feature selection methods like Information Gain, Information Gain Ratio, ChiSquare, and Correlation-Based Feature Selection have been studied and it has been concluded by the authors [8],that though they reduce the accuracy of classification algorithms ,they also reduce the computational time of these algorithms.In this paper [13],CFS subset selection and Consistency Subset feature selection methods were used. These feature selection methods were applied to classification methods like Naïve bayes and SMO to study the accuracy of the methods with different classifiers. Mathematical intersection principle based innovation method using genetic algorithm (GA) for feature selection is proposed[14].Also the performance of this approach with classifiers like naïve bayes and J48 has been compared with other commonly used feature selection methods like CFS, IG and CAE. In this paper [15] ,role of classifiers, feature selection methods with dimensionality reduction in phishing detection have been studied. The paper makes a comparison of the performance of five classifying algorithms, three well known feature selection methods and one dimensionality reduction algorithm on the publicly available website phishing information.

III. PROPOSED SYSTEM

This proposed system applies “Fuzzy Rough Set”(FRS) method as a principle to choose the most prominent characteristics from the various features extracted from the phishing website and email dataset .The commonly used features for phishing website prediction can be classified into four categories. These are:

1. URL-Based : Uniform Resource Locater**(**URL) is used to examine a website to predict if it is genuine or not.URL of phishing domain have distinct features. Some of them are count of digits in URL ,complete length of URL, checking for typosquatting of URL etc

2. Domain-Based : Use passive queries to identify certain key features like whether the domain is blacklisted, if the domain is newly created or the registrant is hidden etc

3. Page-Based: Uses information from the reputed page ranking services eg Google’s page rank. Page features like number of visits on monthly, weekly, daily basis, average visit timing ,internet traffic share per country, popularity among social networking sites among others

4. Content-Based : Analyses the contents like titles of pages, meta data, hidden text, Images etc,,,Analyzing this features are used to extract information about web site category and audience profile ,which can guide to detect if the domain is used for phishing are not.

“Fuzzy rough set theory”(FRS) is the amalgamation of rough set theory and fuzzy set theory. Fuzzy set theory is used to represent the degree of membership x in A. It is used when a situation cannot be expressed through crisp set of values. Rough set theory was introduced to deal with the problem of incomplete information. The problem of incomplete data shows that it would not be possible to predict the concept based on the equivalence relation . Fuzzy set theory and rough set theory being complements of each other help us to model vague as well as incomplete information.

The features selected using FRS theory are supplied into the classifiers for phishing detection .These classifiers are multilayer perceptron and random forest. The system trains each classifier on a collection of data to analyze the potential of FRS feature selection in creating a generalization in phishing detection. The training set contains a varied collection websites from an online database.

This system uses the most efficient classifiers which are multi-layer perceptron and random forest.

Random forest is a classifier developed from decision tree classifier. It is a collection of multiple decision trees. When a new instance needs to be classified, each decision tree supplies a differentiation for input data out of which the random forest selects the most voted prediction as the result.

A multilayer perceptron is better known as an artificial neural network. A perceptron is a single neuron model and building block of ANN. Neurons which are simple computational units use an activation function to produce an output signal ,from weighted input signals.A multilayer perceptron has an input layer which is the visible layer and many hidden layers to process the data and an output layer which is the last hidden layer which outputs the computed value .Number of hidden layers increase with the increase in complexity in the data set. The model makes use of the page rank to classify a domain name as phishing or not. The Page Rank algorithm generates a distribution which represents the probability that a user randomly clicking on links will arrive at any particular website by using various features of the web page **.**The victim URLs page rank is compared with that of legitimate ones .If there is an observable difference between the page ranks then the URL is classified as phishing URL.This approach aims at reducing false positive ratio and can effectively detect phishing.

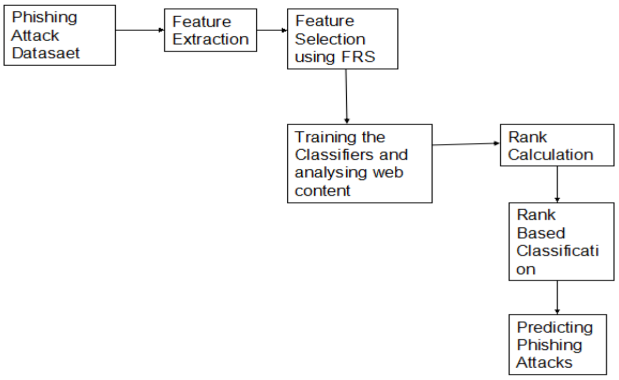

Phishing Attack Datasaet
Feature Extraction
Feature Selection using FRS
Training the Classifiers and analysing web content
Rank Calculation
Rank Based Classification
Predicting Phishing Attacks

Fig 1. Architecture Diagram

IV. IMPLEMENTATION PROCESS

1. Feature Extraction:

Preprocessing: The comparison between two web pages is performed based on the matching contents on the web page. The attributes like URL based, content based, page based and domain based features are used for effective detection of pages from the phishy ones.
This attributes are then used as features for checking their prominence in phishing detection using "fuzzy rough sets".

2. Feature Selection:

The theory of "Rough Sets"(RS) is a method used to conclude how to distinguish a data set, between a decision boundary and a non distinct relation . With features containing continuous values, Rough Set(RS) expanded by "Fuzzy Rough Set"(FRS) theory. FRS makes sure that there exists a subset of attributes denoting the notations of a set member while other attributes are non-label. Heuristic analysis takes four attributes URL based attributes, domain based attributes, content based attributes and page based attributes. The "fuzzy rough sets" determines which of the features will be most prominent in predicting whether the given site is phishing or not.

3. Detection:

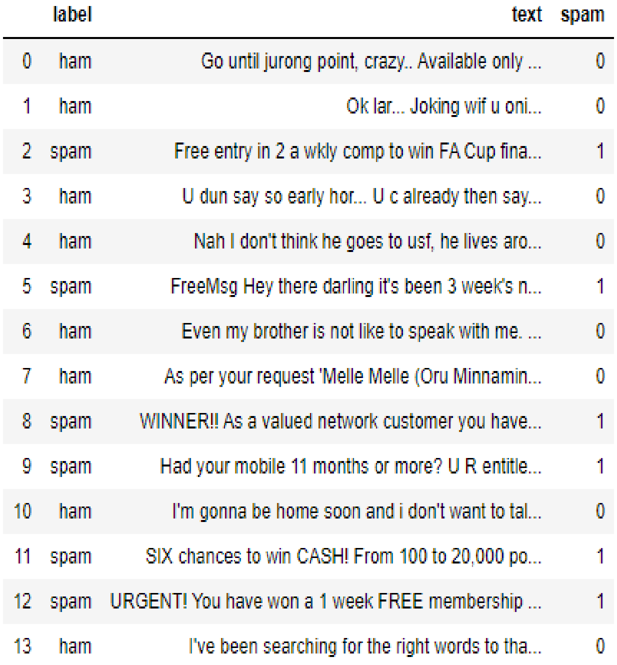
The prediction of whether the dataset is a phishing set is done by two classifiers which are random forest and multilayer perceptron.
Random Forest: The random forest includes a set of trees that are defined by the features extracted in the previous module and are then set as data points to the nodes of the trees in the set by which the respective output is taken of a single tree. The corresponding outputs of all the trees are taken and which value or feature is most derived is considered as the parameter for predicting whether the site is phishing or not.
Multilayer Perceptron: This is a set of feed forward artificial neural networks(ANN) which consists of multiple perceptrons that are binary classifiers which considers the vector representation of the numbers which belongs to a specific class and are used for linear prediction using feature vectors. The multilayer perceptron has three layers which are output layer, hidden layer and input layer. The nodes are mostly neurons except the initial input layer. The neurons uses the non linear activation function. The multiple layers present in the network help to distinguish the linearly inseparable data . The features are then obtained as output from the output nodes which gives the most prominent features for rank calculation.
The features hence obtained are used for rank calculation and prediction.

4.Rank Calculation and Prediction

The features which being most prominent are selected and hence are used for rank calculation. The calculation is done by the rank predicting algorithm which calculates the rank based on the properties or features selected in the previous processs. The calculated rank is then compared with the ranks of the genuine web pages to determine whether or not the web page is a phishing site. Popular ranking sites like Google's page ranking and Alexa's page ranks are also used for reference in case of conflicting results. The combined usage of the two methods results in the output which determines that in the given dataset a phishing web site is present or not.

V. RESULTS

The dataset (Fig 2) supplied to the proposed system contains various messages and their attributes which are then analysed and the frequency of the most frequently used spam words (Fig 4) in the phishing websites or emails is detected. The system provides a graphical representation (Fig 3) of the words which is then used to identify phishing websites. Based on the word frequency in the particular website and the rank of the website we can find if the website is genuine or not. The FRS theory extracts the most prominent features and is very effective in feature extraction.Some features like URL length, domain age and use of special characters gave very promising results. The classifiers trained also showed an almost an accuracy of 97% by using the random forest classifier and a 92% accuracy with the multilayer perceptron.



	label	text	spam
0	ham	Go until jurong point, crazy.. Available only ...	0
1	ham	Ok lar... Joking wif u oni...	0
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	1
3	ham	U dun say so early hor... U c already then say...	0
4	ham	Nah I don't think he goes to usf, he lives aro...	0
5	spam	FreeMsg Hey there darling it's been 3 week's n...	1
6	ham	Even my brother is not like to speak with me. ...	0
7	ham	As per your request 'Melle Melle (Oru Minnamin...	0
8	spam	WINNER!! As a valued network customer you have...	1
9	spam	Had your mobile 11 months or more? U R entitle...	1
10	ham	I'm gonna be home soon and i don't want to tal...	0
11	spam	SIX chances to win CASH! From 100 to 20,000 po...	1
12	spam	URGENT! You have won a 1 week FREE membership ...	1
13	ham	I've been searching for the right words to tha...	0

Fig 2. Tabular Representation of a portion of the dataset

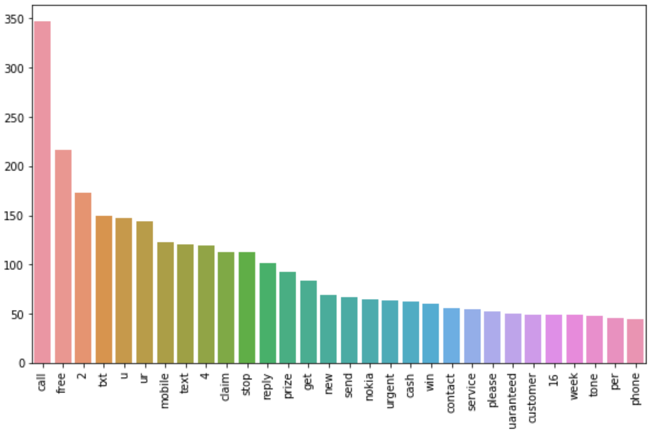


Fig 3. Graphical Representation of the data

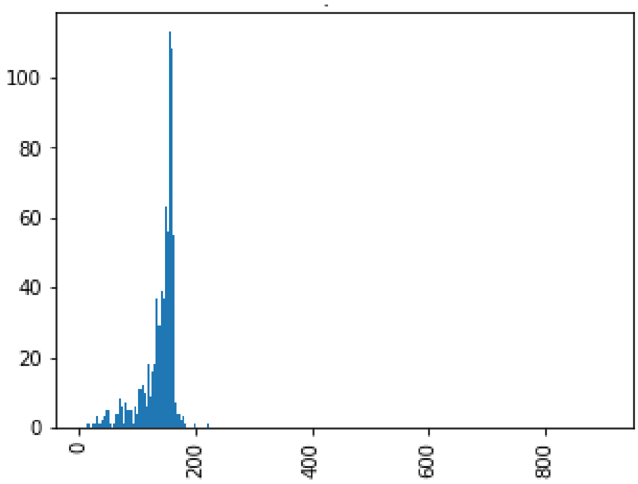


Fig 4. Detection of the Frequency of spam words

VI. CONCLUSION

The proposed system takes a collection of sample websites which are stored in a dataset for checking whether they are genuine or not. It uses "Fuzzy Rough Set"(FRS) theory for feature extraction and classifiers for feature selection like random forest and multilayer perceptron which are trained on the dataset to detect if the site is genuine. We can further improve this system by including the ability to analyse the images present in a website and include captcha recognition methods to increase the accuracy of the system.

REFERENCES

1. Yong Fang ,Cheng Zhang ,Cheng Huang ,Liang Liu ,Yue Yang in "Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism" - 2019 IEEE Access Volume 7
2. "Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks" by Christopher Guiterrez, Taegyu Kim,Rafaelle Della Corte, Jeffery Avery,Dan Goldwasser - IEEE Transactions on Dependable and Secure Computing 2018
3. "A PageRank Based Detection Technique for Phishing Web Sites" by A. Naga Venkata Sunil, Anjali Sardana - 2012 IEEE Symposium on Computer and Informatics
4. "Using Case-Based Reasoning for Phishing Detection" by Hasan Abutair , Abdelfettah Belghith - 2017 Procedia Computer Science(109:281-288)
5. "Detection of Phishing Websites Based on Probabilistic Neural Networks and K-Medoids Clustering" by El-Sayed, M El-Alfy - 2017 The Computer Journal(Volume 60,Issue 12)
6. "A Multi-Agent Case-Based Reasoning Architecture for Phishing Detection" by Hasan Abutair , Abdelfettah Belghith - 2017 Procedia Computer Science(110:492-497)
7. "A Hybrid Model to Detect Phishing-Sites using Supervised Learning Algorithms" by M. Amaad Ul Haq Tahir, Sohail Asghar, Ayesha Zafar. Saira Gillani - 2016 International Conference on Computational Science and Computational Intelligence
8. "The Best Features Selection Method and Relevance Variable for Web Phishing Classification" by Sumarni Adi, Yoga Pristyanto, Andi Sunyoto - 2019 International Conference on Information and Communications Technology
9. "Phishing Websites Classifier using Polynomial Neural Networks in Genetic Algorithm" by S. Gayathri - 2017 Fourth International Conference on Signal Processing, Communication and Networking
10. "Intelligent Phishing Website Detection using Random Forest Classifier" by Abdulhamit Subasi, Esraah Molah, Fatin Almkallawi, Touseef J Chaudhery - 2017 International Conference on Electrical and Computing Technologies and Applications
11. "High-performance Classification of Phishing URLs Using a Multi-modal Approach with MapReduce" by Niju Shreshtha, Rajan Kumar Kharel, Jason Britt, Ragib Hasan - 2015 IEEE World Congress on Services
12. "Application of Genetic Algorithm Based on F-Ratio Rule in Signal Feature Selection" by Ting An - 2017 10th International Symposium on Computational Intelligence and Design
13. "Various Feature Extraction and Classification" by Dalvir Kaur, Sukesha Sharma - 2017 Second International Conference on Microelectronics,Computing and Communication Systems
14. "Genetic Algorithm based Feature Selection Approach for Effective Intrusion Detection System" by Ketan Sanjay Desale, Roshni Ade - 2015 International Conference on Computer Communication and Informatics
15. "Investigating the Effect of Feature Selection and Dimensionality Reduction On Phishing Website Classification Problem" by Pradeep Singh, Niti Jain Ambar Maini - 2015 1st International Conference on Next Generation Computing Technologies

AUTHORS PROFILE

D. Deva Hema, received the M.E Degree in Anna University in 2007 and pursuing Ph.D in Satyabhama Institue of Science and Technoloy. She is currently Assistant Professor in SRM Institute Of Science and Technology in Ramapuram, Chennai, Tamil Nadu, India. Her research interests include the area of Artificial Intelligence and Machine Learning including crash prediction and occupant protection during vehicular crashes.

C. Akshaya, is currently pursuing bachelors of technology in computer science and engineering from SRMIST,Chennai,Tamil Nadu, India

Talluri Venkata Sai Sumanth, is currently pursuing bachelors of technology in computer science and engineering from SRMIST,Chennai,Tamil Nadu, India

Diganta Saha, is currently pursuing bachelors of technology in computer science and engineering from SRMIST,Chennai,Tamil Nadu, India