ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

Development of Anti-Phishing Model for Classification of Phishing E-mail

Niharika Vaishnaw[1], S R Tandan[2]

M.Tech Scholar, CSE, Dr. C.V. Raman University, Bilaspur (C.G), India [1]

Assistant Professor, CSE, Dr. C.V. Raman University, Bilaspur (C.G), India [2]

Abstract*:* Due to colossal financial losses in recent years, phishing has drawn attention of most of the individuals and organizations in the world of internet. Need for protection against phishing activities through fraudulent emails has increased remarkably. In this paper we propose a hybrid model to classify phishing emails using machine learning algorithms with the aspiration of developing an ensemble model for email classification with improved accuracy. We have used the content of emails and extracted 47 features from it. The processed emails are provided as input to various machine learning classifiers. Going through experiments, it is observed and inferred that Bayesian net classification model when ensemble with CART gives highest test accuracy of 99.32%.

Keywords**:** Phishing, Machine learning, Email classification, Hybrid model

I. INTRODUCTION

One of the major security issues associated with internet users these days is "phishing". Phishing is a fallacious action performed in order to acquire financial and personal information like usernames, passwords, credit card numbers, social security numbers, date of birth etc. It is an email spoofing in which a legitimate-looking email is sent to some target users. These emails appear to come from familiar and authentic websites. It usually includes exciting or bothersome statements and suspicious redirecting hyperlinks towards fake website spoofing innocent internet users.

A diagrammatic explanation of phishing activity is given in fig. 1. The phisher installs phishing website and mass mailer to the victim server. The server unknowingly broadcast these phishing emails to the target users. Users get forged by clicking hyperlinks embedded with the email.

A. Phishing Types

1) Spear phishing: It is one of the most successful techniques accounting 91% of attacks. It is accomplished by using personal information of the victim to earn trust thus increasing probability of success [20].

2) Clone phishing: A type of phishing in which a legitimate email is cloned completely replacing the attachment/link with the spurious version.

3) Whaling: It primarily targets high profile and senior executives. The content of email is often written as a legal subpoena, customer complaint, or executive issue. It involves some kind of falsified company-wide concern [21].

4) Rogue WiFi (MitM) : Attackers compromise free Wifi access-points, and configure them to run man-in-the-middle (MitM) attacks [22].

The Kaspersky Lab study 'Financial Cyberthreats in 2014 reports that 28.8% of phishing attacks in 2014 were intended to steal financial data from users. While carrying out their scams, cybercriminals have shifted their focus from bank brands to payment systems and online shopping sites.

In the Payment Systems category, cybercriminals mostly targeted data belonging to users of Visa cards (31.02% of detections in the Payment Systems category), PayPal (30.03% of detections) and American Express (24.6%). Amazon remains the most commonly-attacked brand in the Online Shopping category – 31.7% of attacks in this category used phishing pages mentioning Amazon. However, this is 29.41 percentage points less than in the previous year [2].

The existing defense system (its designs and technology) against such malicious attacks needs to be greatly improved. Behdad et al. [1] pointed out that improving the defense system is not enough to stop fraudsters as some of them could still penetrate; the system should also be able to identify fraudulent activities and prevent them from occurring.

To ensure cyber security and combat cybercrime, development and implementation of emphatic phishing detection techniques is highly essential. Anti-phishing techniques based on machine learning methodology have already substantiated to be utterly effective due to advances in data mining and learning algorithms.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

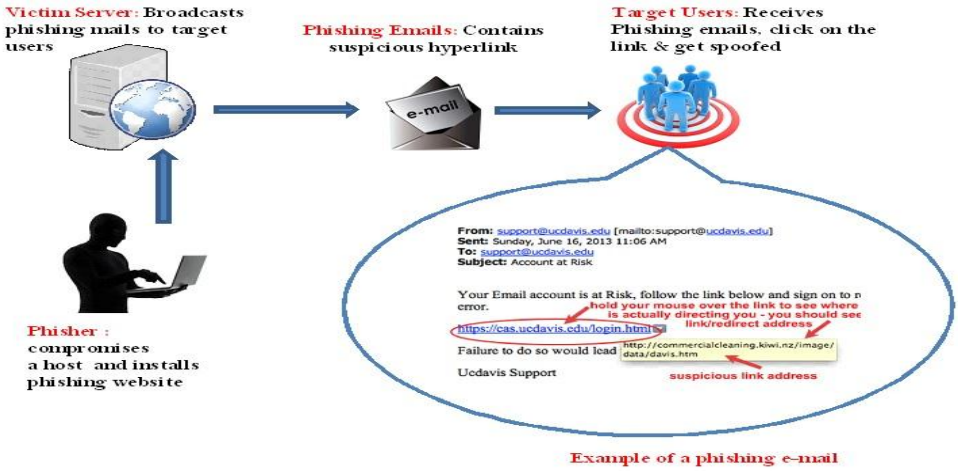

Fig. 1: Phishing Scenario

II. RELATED WORK

Isredza Rahmi A Hamid et.al [3] proposed a hybrid feature selection approach based on combination of content based and behavior-based. The study presented that hybrid features selections are able to achieve 93% accuracy rate as compared to other approaches. F. Toolan et.al [4] introduced an approach to classifying emails into Phishing / non-Phishing categories using the C5.0 algorithm and an ensemble of other classifiers that achieve high recall. The F-Score of the R-Boost method was 99.31% by far the highest of the techniques that have been examined. Gansterer et al. [5] made comparisons between binary (spam vs. not spam) and ternary classification approaches (ham, spam, and phishing). The accuracy reached up to 97% by adapting a support vector machine (SVM).

Semantic ontology concept with adaptive Naive Bayes algorithm was proposed by Bazarganigilani [6] as a new algorithm for text classification of phishing emails using a heuristic way to detect the phishing emails. The accuracy reached up to 94.87%. FRALEC is a hybrid system proposed by Castillo et al. [7] to classify e-mails into two classes, ham email and phishing email. This system consists of three filters:

Naive bays classifier, rule-based classifier, emulator-based classifier. The precision in the best result was 96%. N. Zhang et.al [8] proposed multilayer feedforward neural networks for phishing email detection and evaluated the effectiveness of this approach. NN gives the highest recall while still maintaining a >95% precision, suggesting that NNs are excellent at detecting phishing emails.

III. DATASET

In this research work, we have used publicly available sources of phishing and legitimate emails namely: [23] and [24]. Our phishing dataset is composed of 8266 emails from the following files available at [23]: phishing0.mbox, 20051114.mbox, phishing2.mbox, and phishing3.mbox. The legitimate dataset is gathered from the following files as available at [24]: 20030228_easy_ham.tar.bz2, 20030228_hard_ham.tar.bz2, 20030228_easy_ham_2.tar.bz2. Thus data set contains 8266 instances, 47 features and 1 class having phishing and ham. There is no missing value in this data set.

IV. METHODOLOGY

A. Bayesian net is a statistical processing based on bayes decision theory and is a fundamental technique for pattern recognition and classification. It assumes that pattern possesses random characteristics and they are generated in a random way by some natural phenomena and process. It is a graphical model that encodes probabilistic relationships among variable of interest. The natural choice for dealing with random and uncertain pattern is to use statistical technique based on probabilistic characteristics of data. The Bayesian method is based on the assumption that the classification of patterns is expressed in probabilistic terms. It assumes that the statistical characteristics of random patterns are expressed as known probability values describing the random nature of pattern and their features. These probabilistic characteristics mostly concern a priori probability and conditional probability density of pattern of class [9].

B. CART (Classification and Regression Technique) is one of the popular methods of building decision tree in the machine learning community. It builds a binary decision tree by splitting the records to each node, according to a function of a single attribute. CART uses the Gini index for determining the best split. The initial split produces two nodes, each of which attempts to split in the same manner as the root node. Once again, all the input fields are examined to find the candidate splitters. If no split can be found that significantly decreases the diversity of a given node, labelled as leaf node. At the end of tree growing process, every record of the training set is assigned to some leaf of the full decision tree. Each leaf is assigned a class and an error rate. Error rate of a leaf node is the percentage of incorrect classification at that node [9].

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

C. CHAID (Chi-Squared Automation Interaction Detection) is a derivative of AID (Automatic Interaction Detection). It attempts to stop growing the tree before over fitting occurs. CHAID avoids the pruning phase. In the standard manner, the decision tree is constructed by partition the data set into two or more data subsets, based on the values of one of the non-class attributes. After the data set is partitioned according to the chosen attributes, each subset is considered for further partitioning using the same algorithm. Each subset is partitioned without regard to any other subset. The process is repeated for each subset until some stopping criteria is met. In CHAID, the number of subsets in a partition can range from two up to the number of distinct values of the splitting attribute [9].

D. Artificial Neural Network (ANN**)** is composed of a set of elementary computational units, called neurons, connected together through weighted connections. These units are organized in layers so that every neuron in a layer is exclusively connected to the neurons of the preceding layer and the subsequent layer. Every neuron, also called a node, represents an autonomous computational unit and receives inputs as a series of signals that dictate its activation. Following activation, every neuron produces an output signal. All the input signals reach the neuron simultaneously, so the neuron receives more than one input signal, but it produces only one output signal. Every input signal is associated with a connection weight. The weight determines the relative importance the input signal can have in producing the final impulse transmitted by the neuron. The connections can be exciting, inhibiting or null according to whether the corresponding weights are respectively positive, negative or null. A threshold value, called bias, is similar to an intercept in a regression model [10].The term neural network has moved round a large class of models and learning methods. The main idea is to extract linear combinations of the inputs and derived features from input and then model the target as a nonlinear function of these features. ANN is a large class of algorithms that has the capability of classification, regression and density estimation [11].

E. Support vector machine (SVM) design a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. A SVM is a promising new method for classification of both linear and nonlinear data. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships [12]. Support vector machine algorithms divide the n dimensional space representation of the data into two regions using a hyperplane. This hyperplane always maximizes the margin between the two regions or classes. The margin is defined by the longest distance between the examples of the two classes and is computed based on the distance between the closest instances of both classes to the margin, which are called supporting vectors [13]. The support vector machine is very popular as a high-performance classifier in several domains in classification. The basic idea is to construct a hyper plane as the decision surface such that the margin of separation between positive and negative examples is maximized. Here the error rate of a learning machine is considered to be bounded by the sum of the training error rate and a term depending on the Vapnik Chervonenk is (VC) 1 dimension. Given a labeled set of N training samples (X_i, Y_i), where $X_i \in R_n$ and $Y_i \in \{-1, 1\}$, the discriminate hyper plane is defined as:

$$f(X_q)= \Sigma Y_i \propto i \; K(X_q, X_i)+b$$

Here K( ) is a kernel function and the sign of $f(X_q)$ determines the membership of query sample X_q .Constructing an optimal hyper plane is equivalent to determining all nonzero $\propto$ is which corresponds to the support vectors, and the bias b. The expected loss of making decision is the minimum.

F. Decision tree induction is the learning of decision trees from class labeled training tuples. A decision tree is a flow chart like tree structure, where each internal node denote a test on an attribute, each branch represent an outcome of the test, and each leaf node hold a class label. The topmost node in a tree is the root node. Decision tree can handle high dimensional data. Their representation of acquired knowledge in tree from is intuitive and generally easy to assimilate to human.

The learning and classification steps of decision tree induction are simple and fast. Decision tree algorithm is simple and fast. These tree classifiers have good accuracy. Decision tree induction algorithms have been used for classification in many application areas such as medicine, manufacturing, and production, Financial Analysis, astronomy, and molecular Biology. Decision tree are the basic of Several Commercial rule induction System. Decision tree are built, many of the branches may reflect noise or outliers in the training data. In this research work we will use various data mining based decision tree algorithm like CART, QUEST, CHAID, ID3, C5.0 etc to development of decision support system [14].

G. C5.0 is one of the more recent in a family of learning algorithms referred to as decision tree algorithms. This algorithm is an improvement of the C4.5 algorithm also developed by Quinlan. The improvements are merely in efficiency, the algorithm remains the same [16].

The algorithm is based on the concepts of entropy, the measure of disorder in the collection, and the information gain of each attribute. Information gain is a measure of the effectiveness of an attribute in reducing the amount of entropy in the collection. The C5.0 algorithm builds a decision tree for the data in question. This can be thought of as a sequence of if then rules that allow new instances to be classified. It begins by calculating the entropy of a collection (S) as shown in Equation 1. In this, c represents the number of classes in the system (2 in the phishing detection problem) and pi represents the proportion of instances that belong to class i.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

$$E(S) = \sum_{i=1} - p_i \log_2 p_i \quad \text{.....(1)}$$

The next step is to calculate the information gain for each attribute. This is the expected reduction in entropy by partitioning the dataset on the given attribute. The information gain for attribute, A, in collection, S is shown in Equation 2 where E(S) is the entropy of the collection as a whole, S_v is the set of instances that have value v for attribute A.

$$G(S,A) = E(S) - \sum_{v=values(A)} (|Sv|/|S|) * E(S_V) \quad \text{....(2)}$$

From these information gain values the best attributes for partitioning the dataset are chosen and the decision tree is built [15].

H. QUEST uses a sequence of rules, based on significance tests, to evaluate the predictor variables at a node. For selection purposes, as little as a single test may need to be performed on each predictor at a node. Unlike C&RT, all splits are not examined, and unlike C&RT and CHAID, category combinations are not tested when evaluating a predictor for selection. Splits are determined by running quadratic discriminate analysis using the selected predictor on groups formed by the target categories. This method again results in a speed improvement over exhaustive search (C&RT) to determine the optimal split [17, 18].

I.Hybrid: Two or more modes combined to form a new model is called an hybrid model. A hybrid model is a combination of two or more models to avoid the drawbacks of individual models and to achieve high accuracy. Bagging and boosting [14] are two techniques that use a combination of models. Each combines a series of k learned models (classifiers), M1, M2,…..Mk, with the aim of creating an improved composite model, M. Both bagging and boosting can be used for classification.

V. EXPERIMENT SETUP

Fig. 2 shows the flowchart of the methodology we pursued to perform our experiments. Emails are collected from the sources mentioned in [23], [24]. 47features are extracted from emails using literature surveys of previous works on phishing. We wrote a series of short PEARL scripts to generate files in specific input formats required by classifiers. XML & HTML parsing is done using PERL libraries. Thus a dataset is prepared and further classified as train data and test data. In phase I, both train and test data are provided to various classifiers like CHAID, CHART, SVM etc. to evaluate their accuracy. In phase II hybrid model is prepared and tested with the dataset. The best hybrid model classifies the data into ham and phishing data with highest accuracy.

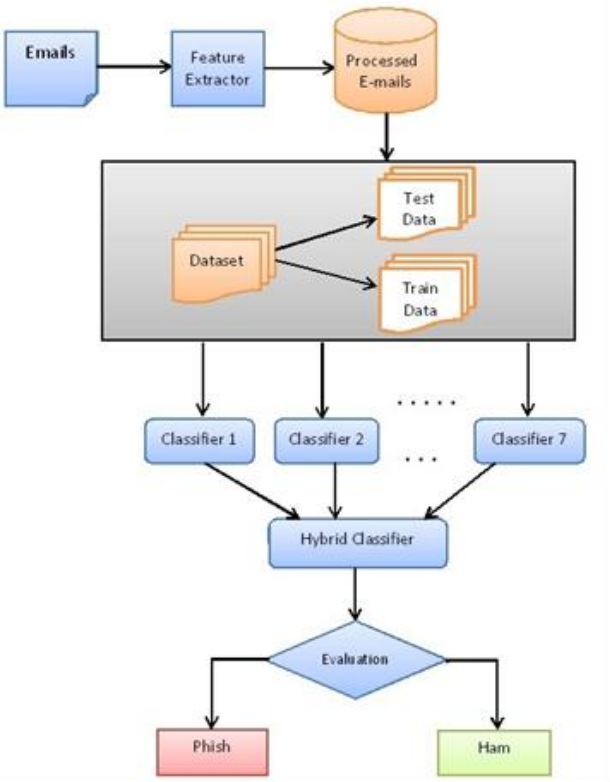

Fig. 2: Model Architecture

VI. EVALUATION MATRIX

A. Model Evaluation

The classification models [19] can be evaluated using Specificity, Sensitivity and Accuracy. This can be accomplished using partition of data set, confusion matrix and other statistical Methods.

1) Confusion Matrix

One of the methods to evaluate the performance of a classifier is using confusion matrix. The number of correctly classified instances is the sum of diagonals in the matrix; all others are incorrectly classified. The following terminology is often used while referring to the counts tabulated in a confusion matrix. A confusion matrix for binary classification is shown in Table I.

The True Positive (TP): corresponds to the number of Positive examples correctly predicted by the classification model.

The False Negative (FN): corresponds to the number of positive examples wrongly predicted as negative by the classification model.

The False Positive (FP): corresponds to the number of negative examples wrongly predicted as positive by the classification model.

DOI 10.17148/IJARCCE.2015.4610

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

The True Negative (TN): corresponds to the number of negative examples correctly predicted by the classification model.

TABLE I
STRUCTURE OF CONFUSION MATRIX FOR BINARY CLASS PROBLEMS

Actual Vs Predicted	**Positive (P)**	**Negative(N)**
Positive(P)	True Positive(TP)	False Negative(FN)
Negative(N)	False Positive(FP)	True Negative(TN)

2) Sensitivity, Specificity, Accuracy

Sensitivity and Specificity are statistical measures of the performance of a binary classification test.

Sensitivity: measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition).

Sensitivity =TP/(TP+FN) …Equation1

Specificity: measures the proportion of negatives which are correctly identified (e.g. the percentage of healthy people who are correctly identified as not having the condition).

Specificity = TN/(TN+FP) …Equation2

Classification Accuracy: Classification accuracy of the classifier is the proportion of instances which are correctly classified.

Classification accuracy = (TP+TN)/N

where N is total number of samples i.e (TP+TN+FP+FN)

VII. RESULT ANALYSIS

This experiment has simulated using Clementine 12.0 for classification of phishing e-mail. Clementine 12.0 (IBM SPSS Modeler) is a data mining and text analytics software application built by IBM .The main motive of this research work is to develop robust classifier and achieve high classification accuracy which can classify ham and phishing mail. The analysis of model is categorized into two phase: first we have analysed individual models and achieved classification accuracy, secondly we have analysed ensemble model.

A. Phase I: Analysis of various models for phishing e-mail classification

In first phase of our experiment, we have used various classification models to classify the data as phishing and ham email. Various classification techniques like CART, CHAID, QUEST, C5.0, ANN and Bayesian Net are trained and tested using 75-25% training testing data set, which means 75% data set is used in training the model and 25% data is used for testing the model. Both the training and testing accuracy is calculated with the help of equation 1.Table II shows accuracy of various individual models. However, three models like C5.0, ANN and SVM achieved above 99% accuracy which shows the robustness of models used, whereas ANN achieved highest classification accuracy 99.61% as a testing accuracy. Fig. 3 shows that training and testing accuracy of the models we have used. Here x-axis represents various models and y axis represents accuracy corresponding to models. Finally we derived that ANN is the best model for classification of phishing emails. The confusion matrix of best model is shown in table III. Other performance measures of best models like sensitivity and specificity are calculated using confusion matrix with the help of equation 1 and 2. Table IV shows performance measures of best model. Here sensitivity of model is 99.41 and 99.81 as training and testing datasets respectively. Similarly specificity of model is 99.35 and 99.40 as training and testing respectively. Figure 4 shows various performance measures of best individual model i.e of ANN.

TABLE II
CLASSIFICATION ACCURACY OF MODELS

Techniques	**Accuracy of models**	
	Training	**Testing**
CART	98.29	98.84
CHAID	98.00	98.65
QUEST	97.66	97.97
C5.0	99.22	99.13
ANN	99.39	99.61
SVM	99.14	99.37
Bayesian Net	98.37	98.84

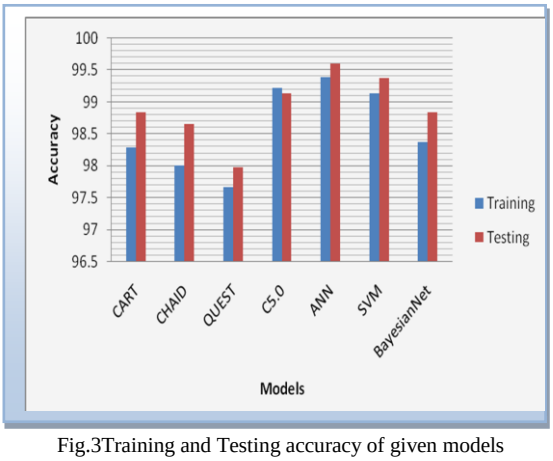

Fig.3Training and Testing accuracy of given models

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

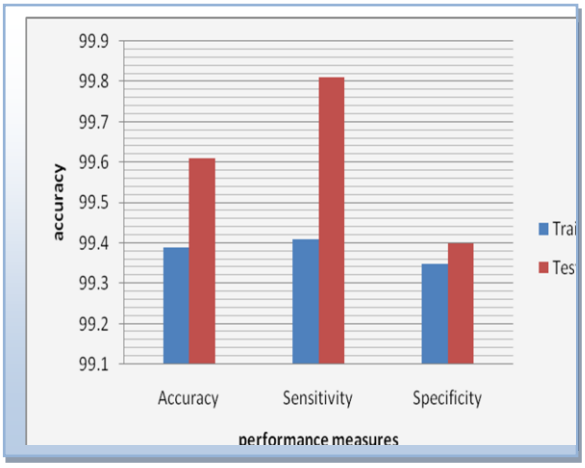

Fig.4 Performance measures of best individual model (ANN)

TABLE III
CONFUSION MATRIX OF BEST MODEL (TRAINING/ TESTING SAMPLES)

Actual Vs Predicted	Training		Testing	
	Ham	Phishing	Ham	Phishing
Ham	3066	18	1064	2
Phishing	20	3088	6	1002

TABLE IV PERFORMANCE MEASURES OF ANN

Performance Measures	ANN	
	Training	Testing
Accuracy	99.39	99.61
Sensitivity	99.41	99.81
Specificity	99.35	99.40

B. Phase II: Development of ensemble model for classification of phishing attacks

In this phase of experiment, we have extended our previous work performed in phase I. We have used various individual classification techniques and ensemble these techniques to develop a hybrid robust model which can classify the phishing emails. An ensemble model is a technique to combine two or more models to achieve high classification accuracy as compared to each individual model. In this phase, first we have trained and tested models like CART, CHAID, QUEST and Bayesian net model and evaluated their accuracies. After that we have ensemble Bayesian net model with rest of the models which gives enhanced accuracy as compared to individual models.

Table V shows training and testing accuracy of individuals and ensemble models.

In this phase, first we have trained and tested models like CART, CHAID, QUEST and Bayesian net model and evaluated their accuracies. After that we have ensemble Bayesian net model with rest of the models which gives enhanced accuracy as compared to individual models. Table V shows training and testing accuracy of individuals and ensemble models. The ensemble models CART+Bayesian Net, CHAID+Bayesian Net and QUEST+Bayesian have achieved high classification accuracy as compared to each individual model. Table V shows that CART+Bayesian Net gives highest training and testing accuracy as best model. Fig. 5 shows performance measures of various individuals and ensemble models. The confusion matrix of best model is shown in table VI. Other performance measures of best models like sensitivity and specificity are calculated using confusion matrix with the help of equation 2 and 3.Table

Techniques	Accuracy of models	
	Training	Testing
CART	98.29	98.84
CHAID	98.00	98.65
QUEST	97.66	97.97
Bayesian Net	98.37	98.84
CART+Bayesian Net	**99%**	**99.32**
CHAID+Bayesian Net	**98.66**	**99.04**
QUEST+Bayesian Net	**98.79**	**99.08**

TABLE VI
CONFUSION MATRIX OF BEST ENSEMBLE MODEL (TRAINING/TESTING SAMPLES)

VII shows that performance measures of best model, where sensitivity of model is 98.86 and 99.15 as training and testing respectively. Similarly specificity of model is 99.13 and 99.50 as training and testing respectively. Fig. 6 show various performance measures of best ensemble model.

Actual Vs Predicted	Training		Testing	
	Ham	Phishing	Ham	Phishing
Ham	3049	35	1057	9
Phishing	27	3081	5	1003

DOI 10.17148/IJARCCE.2015.4610

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 6, June 2015

TABLE VII
PERFORMANCE MEASURES OF BEST ENSEMBLE MODEL

Performance Measures	CART+Bayesian Net	
	Training	Testing
Accuracy	99	99.32
Sensitivity	98.86	99.15
Specificity	99.13	99.50

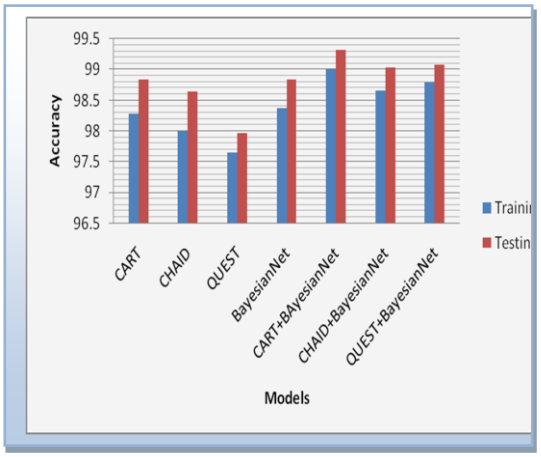

Fig. 5 Performance measures of various individuals & ensemble models

VIII. CONLUSION

With the emergence of phishing as a global security issue, detection and filtration of phishing emails from legitimate ones has become one of the challenging aspects. In this paper, we have extended previous model for some classification techniques like CART, CHAID and QUEST model and ensemble each model to the Bayesian net classification model. We concluded that ensemble of the Bayesian net classification model with these three models individually, gives noticeable increase in classification accuracy for each case. We have achieved highest 99.32% testing accuracy in case of ensemble of CART and Bayesian Net model. The results motivate future work to build an automatic filter detecting phishing emails with the implementation of our hybrid model. We intend to include feature selection mechanism to reduce number of features with elimination of trivial ones.

REFERENCES

[1] M. Behdad, L. Barone,M. Bennamoun, and T. French, "Natureinspired techniques in the context of fraud detection," IEEE Transactions on Systems, Man, and Cybernetics C: Applications and Reviews, vol. 42, no. 6, pp. 1273–1290, 2012.
[2] http://www.kaspersky.com/about/news/virus/2015/Over-a-quarter-of-phishing-attacks-in-2014-targeted-users-financial-data.
[3] Isredza Rahmi A Hamid and Jemal Abawajy," Phishing Email Feature Selection Approach", International Joint Conference of IEEE TrustCom-11, pp. 916-921, 2011.
[4] F. Toolan, J. Carthy.: Phishing Detection using Classifier Ensemble. In E-Crime Researchers Summit,2009.
[5] Wilfried N. Gansterer David P\, et al., "E-Mail Classification for Phishing Defense,", Springer-Verlag, presented at the Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval, Toulouse, France,PP. 449-460, 2009.
[6] M. Bazarganigilani, "Phishing E-Mail Detection Using Ontology Concept and Nave Bayes Algorithm," International Journal of Research and Reviews in Computer Science, vol. 2,no.2, 2011.
[7] del Castillo, M.Iglesias, Ángel Serrano, J., "An Integrated Approach to Filtering Phishing Emails Computer Aided Systems Theory – EUROCAST 2007." vol. 4739, R. Moreno Diaz, et al., Eds., ed: Springer Berlin / Heidelberg, pp. 321-328, 2007.
[8] N. Zhang and Y. Yuan, "Phishing detection using neural network," http://cs229.stanford.edu/proj2012/ZhangYuan-PhishingDetectionUsingNeuralNetwork.pdf.
[9] Arun K. Pujari, Data mining techniques, 4th edition, Universities Press (India) Private Limited,2001.
[10] Paolo Giudici, Silvia Figini, "Applied Data Mining for Business and Industry" ,John Wiley & Sons Ltd,. United Kingdom, 2009.
[11] Alessio Pascucci," Toward a PhD Thesis on Pattern Recognition" , 2006.
[12] V. N. Vapnik, "Statistical Learning Theory" , New York: John Wiley and Sons, 1998.
[13] V. Vapnik, "The Nature of Statistical Learning Theory" ,Springer; 2 edition , 1998.
[14] Han, J.,& Micheline, K., "Data mining: Concepts and Techniques", Morgan Kaufmann ,Publisher, 2006.
[15] Quinlan, J. R. C4.5: Programs for Machine Learning, Morgan-Kaufmann Publishers 1993.
[16] Quinlan, J. R. Is See5/C5.0 Better Than C4.5?, available at http://www.rulequest.com/see5-comparison.html. Last accessed June 2009.
[17] R. D. King, C. Feng, and A. Sutherland(1995), STATLOG-Comparison of classification algorithms on large real-world problems, Applied Artificial Intelligence, Vol.9(3), pp 289-333.
[18] T.S. Lim, W.Y.Loh, Y.S.Shih (2000), A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms, Journal of Machine Learning,Vol 40, pp 203–228.
[19] D. Shanmuga Priya , B. Kavitha, R. Naveen Kumar and K. Banuroopa,"Improving BayesNet classifier using various feature reduction method for spam classification", IJCST, Vol. 1 , Issue 2,2010.
[20] Stephenson, Debbie. "Spear Phishing: Who's Getting Caught?". Firmex. Retrieved 27 July 2014.
[21] "What Is 'Whaling'? Is Whaling Like 'Spear Phishing'?". About Tech. Archived from the original on 2015-03-28. Retrieved March 28, 2015.
[22] "Black Hat DC 2009". May 15, 2011.
[23] Jose Nazario. Phishing corpus. http://monkey.org/~jose/wiki/doku.php?id=phishingcorpus.
[24] SpamAssassin. Public corpus. http://spamassassin.apache.org/publiccorpus.

Fig. 6 Performance measures of best ensemble

BIOGRAPHIES

Niharika Vaishnaw is an M.Tech scholar in the Department of Computer Science and Engineering, Dr. C.V. Raman University, Bilaspur (C.G), India. She received B.E in 2012 from Guru Ghasidas Vishwavidyalaya Bilaspur (C.G).

S.R. Tandan is currently pursuing Ph.D and Assistant Professor in the Department of Computer Science and Engineering, Dr. C.V. Raman University, Bilaspur, India. His research interests include application of Artificial intelligence in robotics, Soft Computing, Dynamic path planning and Mobile robot navigation in cluttered environments, Mobile Computing and Cyber security.