Intelligent Security for Phishing Online using Adaptive Neuro Fuzzy Systems

G. Fehringer
Computer and Information Sciences
University of Northumbria
Newcastle upon Tyne, NE1 8ST, United Kingdom

P. A. Barraclough
Computer and Information Sciences
University of Northumbria
Newcastle upon Tyne, NE1 8ST, United Kingdom

***Abstract*—Anti-phishing detection solutions employed in industry use blacklist-based approaches to achieve low false-positive rates, but blacklist approaches utilizes website URLs only. This study analyses and combines phishing emails and phishing web-forms in a single framework, which allows feature extraction and feature model construction. The outcome should classify between phishing, suspicious, legitimate and detect emerging phishing attacks accurately. The intelligent phishing security for online approach is based on machine learning techniques, using Adaptive Neuro-Fuzzy Inference System and a combination sources from which features are extracted. An experiment was performed using two-fold cross validation method to measure the system's accuracy. The intelligent phishing security approach achieved a higher accuracy. The finding indicates that the feature model from combined sources can detect phishing websites with a higher accuracy. This paper contributes to phishing field a combined feature which sources in a single framework. The implication is that phishing attacks evolve rapidly; therefore, regular updates and being ahead of phishing strategy is the way forward.**

Keywords*—Phishing websites; fuzzy models; feature model; intelligent detection; neuro fuzzy; fuzzy inference system

I. INTRODUCTION

Phishing attacks are increasing rapidly costing the global economy billions of dollars per year [1]. Although various studies have concentrated on phishing attacks and used a variety of solutions in the recent years to combat phishing [2], [3]-[6] there is still a lack of accuracy in real-time causing vast amount of losses annually [7]. In general, detection techniques are classified in 2 main categories namely, URL blacklist-based and web-page feature-based. URL blacklist use human-verified URL based on server-side that performs URL matching with real-time website URLs to detect phishing websites. This category works on the principle of detecting phishing attacks and provide warning to users to prevent them from taking risky actions that could otherwise result in compromising their sensitive information. Although existing approaches are effective to some extent, effective generalisation to new threats is still a challenge. For instance, it was discovered that zero-hour protection provided by blacklist-based toolbar systems offers true-positive rate between 15% and 40% [8]. These systems can make users vulnerable to phishing threats. Moreover, statistics were posted in March 2009 stating that it takes 10 hours on average to verify submitted URLs [9]. This study proposes a novel Intelligent Phishing Security (IPS) using features and adaptive neuro fuzzy inference systems approach to combat the above limitations.

Feature-based approaches have been studied by researchers [10]-[12] using extracted features from sources such as web-pages and utilizes machine learning algorithms to enhance phishing detection systems, but inaccuracy is still a problem. Different to blacklist-based approaches, feature-based approaches can be generalized to new phishing attacks. However, feature-based approach has a tendency to have high false positives. Inaccuracy has been the main barrier in the existing phishing detection technologies [7].

Fuzzy rules techniques that models the qualitative side of human reasoning process with no precise quantitative analysis was extended to deal with imprecise conditions [13]. Fuzzy rules have been investigated by [13] and have found various practical applications in imprecise reasoning, control and in prediction. However, fundamentals of this method when dealing with data requires consideration. Particularly, there is a lack of effective feature models that cover specific and effective features.

The intelligent phishing security approach (IPS) is based on Adaptive neuro-fuzzy inference system (ANFIS), using features from a combined sources. ANFIS is chosen because it has ability to learn given features. While neural network handles features well, fuzzy logic deals with reasoning on a high level utilizing given features. This enables membership function parameters tuning for classification between phishing, suspicious and legitimate websites. The two combined feature sources 'phishing email and web-form' are used because they are main phishing targets. These enable effective feature extraction for training and testing sets. Phishing emails carry hyperlink that leads users to phishing web-form. The goal of web-form is to collect the main users sensitive information. The phishing e-web-form is not only used for classification, but also used for training and testing fuzzy models and to generate fuzzy rules as demonstrated in experiment section. This study is focused on answering: how can effective feature model be constructed to reduce inaccuracy in online transactions?

A. This research objectives are to:

1) Conduct relevant literature review in the relevant field.
2) Identify sources in order to extract effective features.
3) Perform feature preparation and normalization to a format suitable for the chosen tools.

4) Design and carry out experimental procedure based on Adaptive Neuro-Fuzzy Inference Systems with features.

5) Train and test fuzzy models using training and testing sets.

6) Compare the result with the existing methods in the field to measure the model's effectiveness.

This paper contributes to the field of phishing detection combined sources of features, phishing Email and phishing e-web-form in a single framework. This is novel as the idea has not been used in literature. Features are extracted from these sources to construct a feature model. This feature model is expected to reduce inaccuracy in existing detections.

The remaining sections are structured as: Section II critically reviews relevant literature in the field and describes phishing deception procedures. Section III describes the intelligent phishing security approach including sources and features identification, feature normalization and size, a description of ANFIS and fuzzy rules. Section IV describes the experimental procedures including intelligent phishing security structure and learning procedure, training and parameters, testing and membership functions. Results are presented in Section V. Section VI provides comparison and discussions. Section VII offers evaluation and conclusions including feature work.

II. Relevant Work

Although various solutions are available that models automatic phishing detection, major studies have focused on blacklist-based and feature-based approaches applying machine learning techniques to identify phishing sites.

A. Feature-based and Machine Learning Approaches

Ozarkar and Patwardhan [14] focused on feature-based applying machine learning methods to detect spam email. A set of conventional features model was used and summarized a prediction error rate using Associative Classification (AC) algorithms. Their method achieved 97.5% accuracy. The results indicated that C4.5 outperformed other algorithms with 5.76% average error-rates, which is relatively putting users at risk when faced with phishing attacks.

Zuhair [15] investigated different feature-based methods using different sizes based on 58 hybrid futures and four machine learning classifiers, including Naïve Bayes, DT, C4.5, and SVM to classify phishing websites effectively. The author's results revealed that the features were highly significant with a recommendation drawn that there is no golden filtering method that fits all classifications algorithms on the data sets used in the experiments. However, a better way forward should be provided from the finding.

Form's [16] study applied Support Vector Machine classifier to classify emails using a set of 9 structure-based and behaviour-based features. The method achieved 97.25% success rates, however it has a relatively small training dataset (1000 emails with 50% spam and 50% non-spam). In the attempt to improve feature-based approach, 27 features from [17] were used based on multi-class classification-based association rules (MCAR) and Association classification (AC) algorithm to assess the detection system. Their method classified webpages with more than 98.5% accuracy, but it was not clear how many rules were generated by employing the MCAR algorithm.

Moreover, work by [11] concentrated on feature-based approach to explore how rule-based classification data-mining techniques are applied in phishing site detection. They employed 450 legitimate and phishing websites for features extraction. By using JavaScript feature extractor, features related to the address bar were extracted. Rule-based classification algorithms employed C4.5, RIPPER, PRISM and Classification using Associate algorithm and 17 features to identify phishing attacks. Using experiment, C4.5 attained 5.76% error-rates, RIPPER obtained 5.94% errors and PRISM achieved 21.24% error rates. After reducing feature size to 9, classification based on association (CBA) achieved 4.75% error rates which was lower compared to other classifiers used. However, the features are not wide-ranging enough to cover phishing characteristics that users can experience in their daily browsing. Additional features could improve the system.

As seen above, Abdelhamid, Ayash and Tabatah [10] method explored data-mining utilizing Associate Classification algorithms such as Multi-class Associative Classification (MCAC), classification based on association (CBA), missing completely at random (MCAR), Multi-attribute co-cluster (MMAC), rule induction and decision trees algorithms including C4.5, PART, RIPPER with 16 features to distinguish between phishing and legitimate sites. The results attained are as follows: MCAC algorithm was misclassified by 0.8% error rates, C4.5 misclassified 1.24% error rates, RIPPER misclassified 1.86% errors and PART misclassified 4.46% error rates. Although MCAC achieved a high accuracy, their features have been extracted from one source only without considering all other different possible sources. In this case, more sources could improve features to detect emerging phishing attacks.

Another work explored feature-based approach [18] based on a neuro-fuzzy, using features extracted from five different sources (also known as inputs) to detect phishing websites accurately. Their approach achieved 98.5% accuracy, but suffered 1.5% error rates.

In the attempt to improve feature-based approaches, a neuro-fuzzy system was employed to deal with parameters in detecting phishing attacks [19]. Given that machine learning algorithms are parameter driven and parameters are difficult to tune. This applies to this study because in training and testing, parameters are tuned for a desirable outcome. The study used 300 features based on a neuro-fuzzy system to tune different parameters for the discovery of a desirable parameters. The method achieved 98.74% accuracy. The outcome can be used to increase user's confidence on tuning parameters.

Another area of study investigates email-based approaches. A method utilized 9 hybrid features of email header and body extracted from approximately 10000 emails divided equally between genuine and spam emails [4]. Their method applied J48 classification algorithm to classify phishing and legitimate emails. Their approach obtained 98.1% success with 1.9% false positives.

B. Blacklist-based approaches

In the attempt to improve URL blacklist-based approaches, PhishStorm was introduced to identify potential phishing websites [20]. Their approach applied machine learning classification, using lexical features including low-level domain, upper-level domain, path and query based on STORM and Bloom, and big data architectures. Their method attained 94.9% accuracy, but suffered 1.44% false positives. Although the use of data from parts of URL exhibits higher positive rates, features extracted from URLs alone are not comprehensive enough to detect phishing websites accurately due to variations of phishing characteristics and regular evolving techniques. Therefore, extracting data from various sources is a possible means to solve this issue.

Phishing Email classification model was explored to detect phishing [21]. The method applied text stemming and wordNet based on Knowledge discovery, data mining and text processing. Using Random forest algorithm achieved 99.1% accuracy, while 98.4% accuracy was achieved applying J48 algorithm. However, this method used 0.9 training data-set and the remaining 0.1 as testing data-set which is a very small data-size for testing using 10-fold-cross-validation measuring method.

Although blacklist-based approaches are largely utilized in industry due to low false-positives, these approaches cannot handle new phishing attacks [7] due to employing human-verified blacklist which takes longer to update. Also there is an ill-fame group known as "rock phish gang" that utilizes toolkits to make a large number of unique phishing URLs, placing extra pressure on blacklist-based techniques.

In contrast to blacklist-based, feature-based techniques exist to combat this problem. Mohammad, Thabtah and McCluskey [22] used automatically extracted features instead of manually extracted to predict phishing website. The method is said to be effective. However, using URL, IP addresses, adding prefix and suffix to request URL may not be adequate to detect evolving phishing.

In this paper, a feature-based approach is used based on adaptive neuro-fuzzy system with phishing e-web-form features. Specifically, 56 features in total are used. Thirty-four features are taken from the previous study [19] and twenty-two significant novel features are added from phishing e-web-form sources. Features are reduced to decrease redundant records in the training and testing sets and applied the state-of-the-art advanced machine learning algorithm. This should enables the construction of e-web form feature model to solve the errant problem of false positive rates and keep up with emerging phishing attacks.

C. Phishing deception procedures

In order to tackle phishing attacks effectively, it is important to understand how phishing works. Phishing attack is a deception that causes humans to take risky action. As shown in Fig. 1, phishing website attack takes the follow form:

1) **Step 1:** Attacker starts by constructing malicious websites with a form, and e-mail with hyperlink. The malicious sites and email are meant to look exact copies of known organization websites. For example, financial institute sites.

2) **Step 2:** Attacker sends malicious emails with a link pointing a user to malicious website.

3) **Step 3:** A user accesses malicious web form by clicking the link. After the user accessed malicious web form, the user types the required fields with sensitive information as requested on the form and clicks submit button.

The information goes to attacker, while the user believes it is received by a known financial institute.

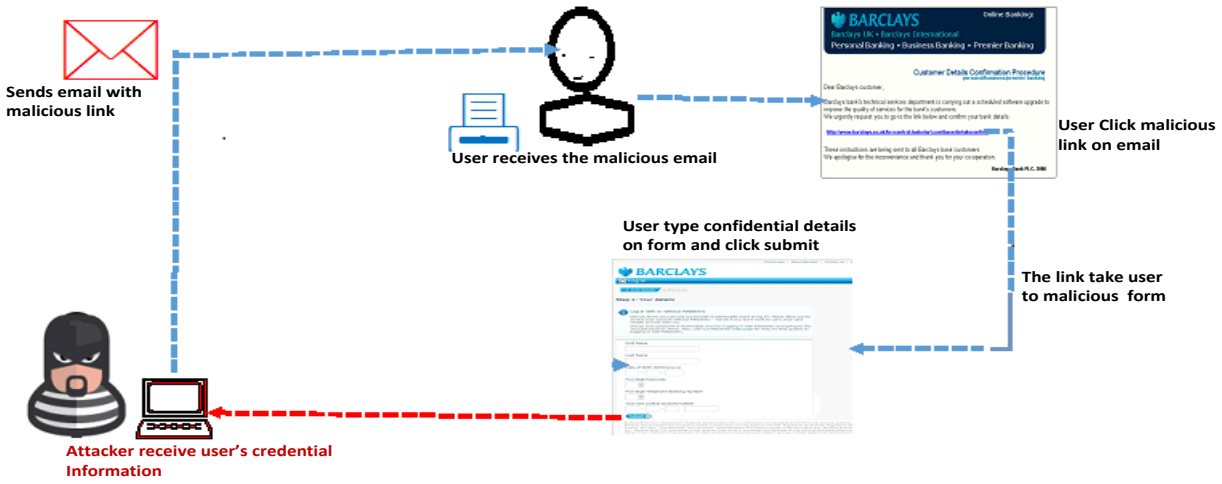

Fig. 1. Phishing deception process.

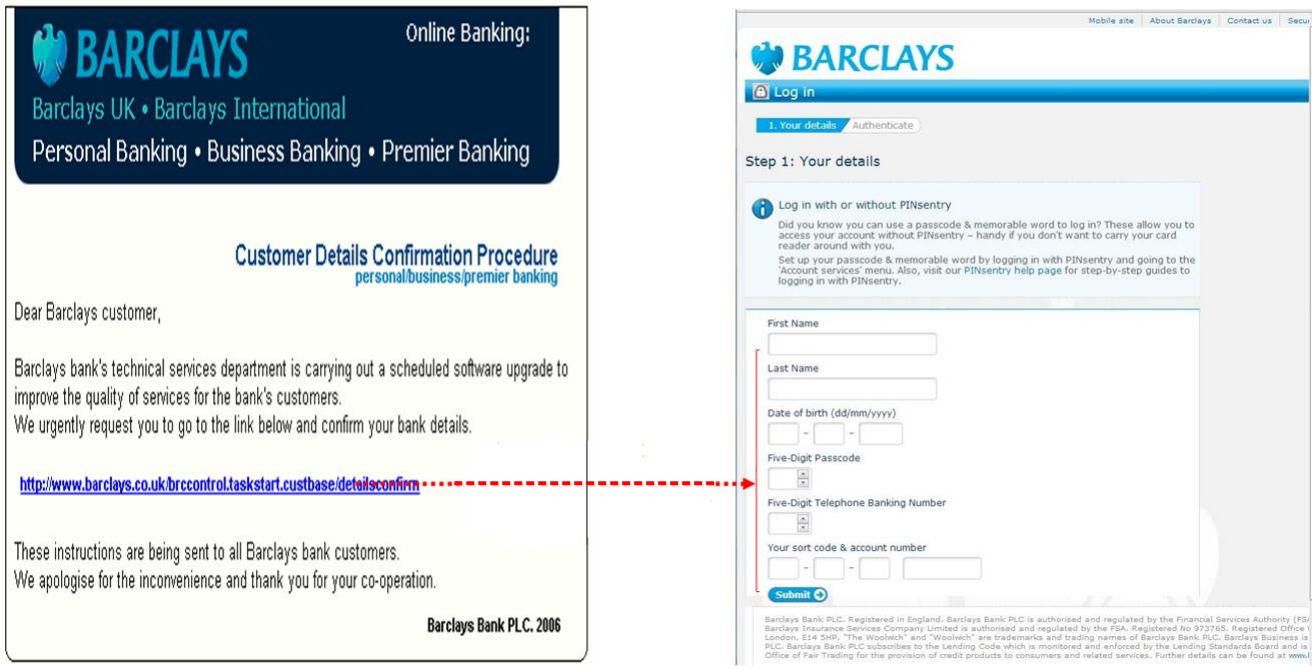

Fig. 2. Phishing attack methodology.

In summary, phishing website attack starts by a phisher creating a malicious website and a form with a hyperlink. The link is sent to users by email, which convince users to click a link that take victims to malicious websites with a form for users to fill in their sensitive details. Examples of phishing email and web-form imitating Barclays bank's genuine email and site are presented in Fig. 2(a) and 2(b).

III. Intelligent Phishing Security

The proposed intelligent phishing security (IPS) take the form of a zero order Sugeno type that consists of four main components, which include: features sources, features, adaptive neuro-fuzzy inference system (ANFIS). ANFIS consists of a rule base and a feature base that makes knowledgebase and a decision-making unit. ANFIS is used because it can learn and validate given features using IF-THEN fuzzy rules. The intelligent phishing security architecture is presented in Fig. 3. The construction of fuzzy rules and the process of the fuzzy inference are provided in detail in sections below. Input features enable fuzzy modelling and Gbell shape membership function to be used for its efficiency. As can be seen in the intelligent phishing architecture, the two sources are phishing emails and phishing web-forms jointly known as phishing e-web-forms

1) The sources are the core of the combined framework from which comprehensive features are extracted that is used to generate fuzzy models and to test the models.

2) Features are phishing website identifiers, which are utilized to classify between phishing, suspicious and legitimate websites. The features can be seen in Fig. 4.

3) A rule-base carries a number of fuzzy IF-Then statement.

4) The Takagi and Sugeno type rules are utilized because it is a zero order [23]. The output of every rule is a combination of linear, input and constant term. Whereas a final output is the weighted average of all rule's output.

5) An adaptive neuro-fuzzy inference system is a neural network with multilayers that performs fuzzy reasoning [13].

6) Features extracted from comprehensive sources are also known as identifiers to distinguish phishing website. The methodology choice is scientifically appropriate to meet the study's objectives.

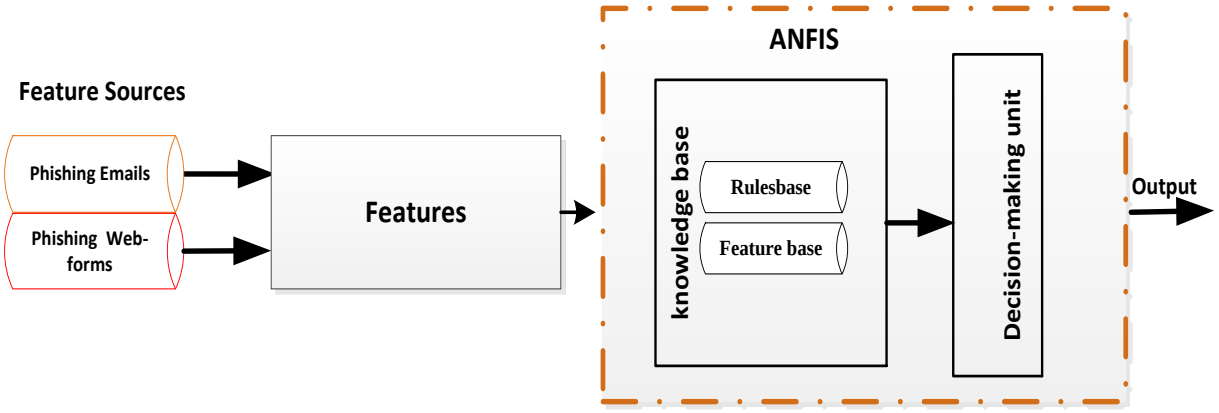

Fig. 3. Intelligent phishing security architecture

A. Sources and feature identification

To reduce phishing attacks, it is important to utilize effective techniques and identify important sources to extract comprehensive features that can detect phishing websites accurately. One hundred phishing emails that contain links to phishing websites and two hundred web-forms used to collect sensitive information from users are gathered to be the representative of sources to extract features. From these sources, 56 features were extracted in which 22 features are novel, while 34 features have been used in the previous work [19] as stated in the above section. These features are presented in Fig. 4. Emails and web-forms are selected from Millersmiles' website archive [24]. Millersmiles is one of the leading anti-phishing web services dedicated to maintain a large archive of phishing websites and emails. Other anti-phishing services that maintains huge archives for phishing websites is PhishTank [9]. The sources are chosen because new phishing attacks are added to them regularly and they supplement each other well.

B. Feature normalization

To comply with fuzzy inference system principles, variable ranges are determined. Intelligent phishing security has four main linguistic variables as detailed in Section D: Legitimate (low), suspicious (medium) and phishing (high). The degree of risk is the most important criteria to determine the accuracy of models. The average percentage accuracy should not exceed the limit acceptable in phishing detection. To determine the degrees of risk, feature's linguistic values (legitimate, suspicious and phishing) ranges are specified, the degree which are normalized to values within the range of (0, 1) by assigning the numerical value. Website are classified between legitimate, suspicious or phishing. Other values like 'very low' are not practical to use. Features are normalized since they are textual and the assigned values are used to define the level of risk of a website. Also, features are of different types. However, some tools for intelligent systems like MATLAB toolbox requires a specific type of data [12].

C. Feature size

Features size can vary depending on the type of measuring tool. However, features should have sufficient size that can be analyzed when using standard measurement method. If for instances, two-fold cross validation method is used, the features size should be split into two pairs with a reasonable amount in each pair. Therefore, 56 feature size are used for experiment in the intelligent phishing security is within the minimal sufficient amount. They are split into 28 training set and 28 testing set. These features are the most frequent phishing features found across all the two hundred phishing web-forms and a hundred phishing emails.

D. Adaptive neuro fuzzy inference system

Adaptive Neuro-Fuzzy Inference System (ANFIS) is a type of adaptive network that is functionally equivalent to Fuzzy inference systems. It combines both fuzzy logic principles and a neural network. It represents Sugeno Tsukamoto fuzzy models that utilize a hybrid learning algorithm. Its outputs depend on input data and the parameters relating to the neurons. ANFIS is used in this study not only because of its advantage of imprecise reasoning, but also for its suitable functionalities when modelling the intelligent phishing security fuzzy models [13].

E. Intelligent phishing security fuzzy rules

In phishing detection structure, there are inputs which are represented in the form of inputs as x , y and one output z. A single input is represented by two-fuzzy sets and the output by a first order polynomial then the rules are presented in the following form in Fig. 5:

Email	Features	Value range
Email	Security alert	
	Details confirmation	
	A transfer was attempted	[0, 0.3]
	Encrypt reference	
	Have your card ready to hand	
Personal Detail	Date of birth	
	Mobile phone number	
	Primary telephone number	
	Home address	[0.2, 0.6]
	Postal code	
	E-mail	
	e-mail password	
	Last name	
	First name	
Card secret Items	Start date	
	Expiration date	
	Credit card number	
	Sort code	
	Security number	
	Account number	
	Password	
	first digits of your password	
	second digits of your password	
	third digits of your password	
	fourth digits of your password	
	fifth digits of your password	
	sixth digits of your password	
	seven digits of your password	
	eighth digits of your password	
	ninth digits of your password	
	tenth digits of your password	
	eleventh digits of your password	[0.5,1]
	twelves digits of your password	
	Confirm your ID	
	Bank account number	
	Your card details	
	Card security code	
	3-digit security code	
	Card verification number	
	Membership number	
	Five-digit passcode	
	Memorable word	
	Mother maiden name	
	Card verification number	
	Personal logon	
	Telephone banking PIN	
	Telephone passcode	
	User code	[1]

Fig. 4. E-web-form Feature Model.

R_1:		Rule 1:	
IF	x is $A1$	IF	feature is security_alert
OR	y is $B1$	OR	feature is details_confirmation
	⋮		⋮
THEN	z is $C1$	THEN	Website is Legitimate
R_2:		Rule 2:	
IF	x is $A2$	IF	feature is credit_card_start_date
OR	y is $B2$	OR	feature is credit card_expiry_date
	⋮		⋮
THEN	z is $C2$	THEN	website is suspicious
R_3:		Rule 3:	
IF	x is $A3$	IF	feature is credit card_number
OR	y is $B3$	OR	feature is sort_code
	⋮		⋮
THEN	z is $C3$	THEN	website is phishing

Fig. 5. Rule representation.

Where x, y and z (feature and website) are linguistic variables, $A1$, $A2$, $A3$ (details_confirmation, credit_card_start-date and card_number) are linguistic values decided by fuzzy-set on the universe of discourse x; $B1$, $B2$ and $B3$ are fuzzy sets on the universe of discourse z (website).

IV. Experimental Procedures

The experiment for the intelligent phishing security (IPS) is based on an ANFIS, using 56 features as training and testing sets. The fuzzy system is used since it is best suited for feature-base fuzzy modeling. 56 comprehensive features are used for training and testing fuzzy models. A two-fold cross-validation method is employed which involves randomly splitting the features into two parts, training set and testing set. First, training is carried out on a training-set only once and testing is done on a test-set only once. Then the roles of training and testing sets are reversed. The results are assembled to get the average errors.

A. Phishing security structure and learning procedure

An Intelligent phishing security structure generated from phishing e-web-form features and ANFIS learning procedure can be seen in Fig. 6. It is equivalent to first-order Sugeno-fuzzy-model and is a multilayer network feedforward where every single neuron performs a specific function. It has six layers including: Layer I – input, Layer II – fuzzification, Layer III – fuzzy rule, Layer IV – normalization, Layer V – defuzzification and Layer VI – crisp output.

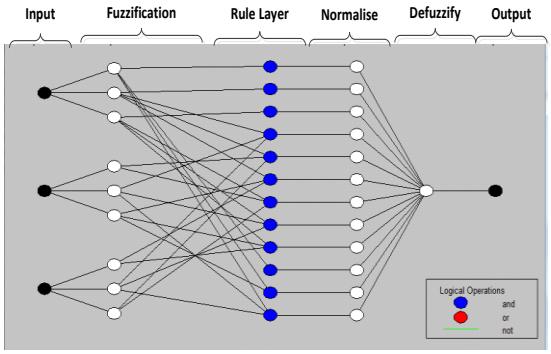

Fig. 6. Fuzzy structure with six layers.

The learning procedures take the following form:

a) Layer 1: This is the input. Neurons in Layer I transmit external inputs to Layer II. This is specified as $y_i^{(1)} = x_i^{(1)}$ (1) where the input is $y_i^{(1)}$ and the output in Layer I is $x_i^{(1)}$.

b) Layer II: This is the fuzzification. Neurons in Layer II perform fuzzification. The Sugneo model fuzzification have a Gbell shape activation function. This is specified as:

$$y_i^{(2)} = \frac{1}{\left(\frac{x_i^{(2)} - ai}{Ci}\right)^{2b_i}} \quad (2)$$

Where, the input is $x_i^{(2)}$ and the output is $y_i^{(2)}$ of neuron I Layer II, and a_i, b_i and c_i are parameters that control the center width and slope of the bell-shape function of neuron.

c) Layer III: This is a rule layer. Every neuron in Layer III matches a single fuzzy rule. A rule neuron receives inputs from fuzzification and calculates the execution strength of the rule it represents. Thus the output in Layer III is expressed as:

$$y_i^{(3)} = \prod_{j=1}^{k} x_{ji}^{(3)} \quad (3)$$

Where, inputs are $x_{ji}^{(3)}$ and $y_i^{(3)}$ are output of rule neuron in Layer III.

d) Layer IV: This is normalization. Every neuron in this layer receives outputs from the rule layer and the normalized execution strength of any given rule is calculated. Therefore, the output of Layer IV is expressed as:

$$y_i^{(4)} = \frac{x_{ii}^{(4)}}{\sum_{j=1}^{n} x_{ji}^{(4)}} = \frac{ui}{\sum_{j=1}^{n} uj} = {}_{ui} \quad (4)$$

where the input $x_{ji}^{(4)}$ from neuron j is allocated in Layer III to Layer IV.

e) Layer V: This is defuzzification. Every neuron in Layer V is linked to the normalisation neuron, as well as receives the initial input x_1 and x_2. A defuzzification neuron calculates the value of weight consequent of any given rules. It is specified as:

$$y_i^{(5)} = x_i^{(5)}\left[k_{i0} + k_{i1}x1 + k_{i2}x2\right] = u_i\left[k_{i0} + k_{i1}x1 + k_{i2}x2\right] \quad (5)$$

Where, the input is $x_i^{(5)}$ and the output in the defuzzification neuron is $y_i^{(5)}$ in Layer V. k_{i0}, k_{i1} and k_{i2} is a set of parameter consequent of rule I [25].

f) Layer VI: This is represented by a single neuron sum. The neuron calculates the total of outputs of every defuzzification and produces the overall output y which is specified as:

$$y = \sum_{i=1}^{n} x_i^{(6)} = \sum_{i=1}^{n} u_i \left[k_{i0} + k_{i1} x1 + k_{i2} x2 \right] \quad (6)$$

Indeed, the fuzzy structure is functionally equal to a first-order Sugeno-fuzzy model.

B. Training and parameters

To facilitate the learning of intelligent phishing security feature model, parameters are assigned. Membership functions assigned to each input is arbitrarily set to 3, Gbell shape was chosen, 30 epochs and learning optimization method known as hybrid. These parameters are identified carefully to optimize model performances. A training-set was presented to the input layer of the network in which the network propagates the inputs from layer to layer till it reaches the output layer. When it finds a different pattern from a desired output, an error is calculated and back-propagated through the network from the output layer to the input layer. When errors are propagated, weighting is modified. After training is complete, the training set is used only once to train the feature model.

C. Testing

Testing begins as the testing feature-set is presented in the input layer of the system. The neurons propagate the inputs from Layer 1 to every layer till it reaches the output layer. Since the measurement method is two-fold cross-validation, testing is completed. After which, roles are reversed and the testing set is used to train, while the training set is used to test models. The measurement at the testing stage is the final average measurement in which the model is evaluated for its merit. After the training and testing processes, outputs are achieved which are presented and discussed in sub-section *D* whereas results are displayed and discussed in Section V.

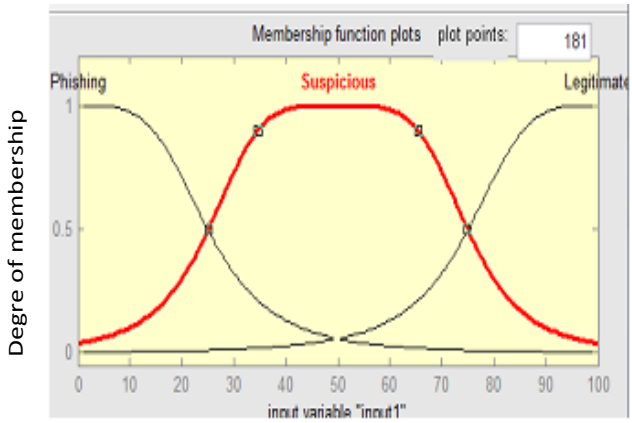

Fig. 7. Membership function using feature-set.

D. Membership functions

Due to efficiency, membership function (MFs) used is Generalized bell (Gbell) shape, while the default MF has triMFs. Membership function of fuzzy sets is a curve that defines how each point in the input space is mapped to a membership value between (0, 1) and the output can vary between (0, 100). The Gbell shape MF is illustrated in Fig. 7.

The intelligent phishing security membership function plot for feature model contains values including Low=Legitimate, Medium=Suspicious and High=Phishing:

Linguistic variables	Numerical value range
Legitimate	(0, 0, 0.3)
Suspicious	(0.2, 0.4, 0.5)
Phishing	(0.4, 0.6, 1)

This means legitimate features has a degrees of risk range between (0, 0.3). Phishing website has degrees of risk range between (0.4, 1). Suspicious has degrees of risk range between (0.2, 0.5). While an output with a risk between (0, 0.3) is classified as low in phishing detection, an output with a risk between (0.2, 0.5) is classified as medium and an output with a risk between (0.4, 1) is classified as high.

V. Results

Using two-fold cross-validation, training was first performed on a training set and tested on a testing set. The roles are reversed and training is performed on a testing set. This is presented in the graph in Fig. 8.

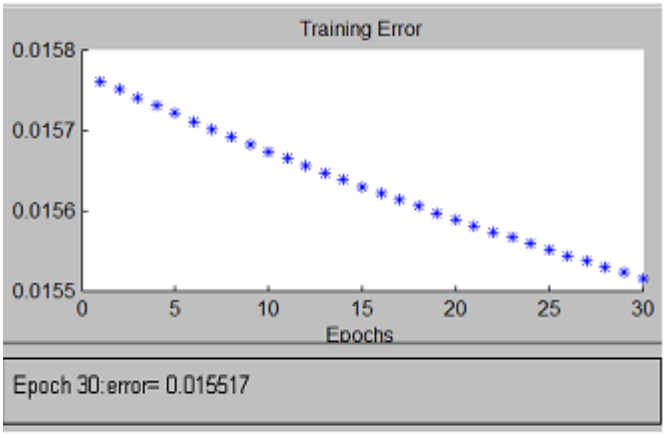

Fig. 8. Training results error rates.

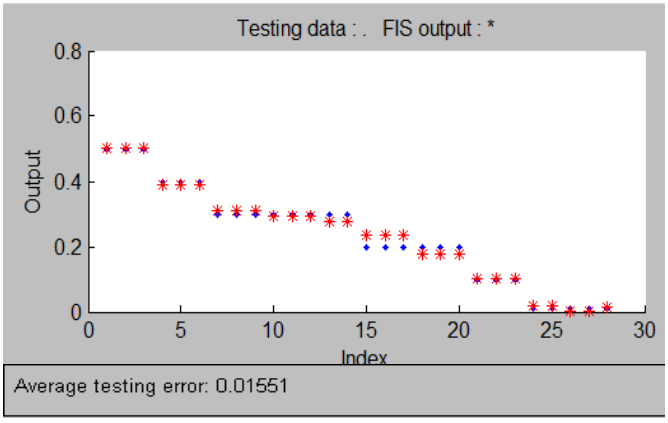

Fig. 9. Test results in error rate.

TABLE. I. A COMPARISON BETWEEN THE EXISTING AND THE IPS RESULTS

Authors	Approach	Feature size	Learning algorithm	Testing results
Form's Approach	Feature-based	9	SVM	97.25%
The IPS Approach	Feature-based	56	ANFIS	98.4%
Barraclough et al.	Feature-base	300	A neuro-fuzzy	98.74%

TABLE. II. RESULTS SUMMARY

Training & Testing	Training error	Testing error	Errors in 2 decimal point	Overall testing average accuracy in percentage
Experiments	0.015517	0.01551	1.6	98.4%
	0.015517	0.01551	1.6	

The estimation for training offered a single crisp output errors of 0.015517, which is lower and is equal to the first run. The overall training measure obtained an average of 1.6% error rate. This is illustrated in Table 1.

Similarly, after the training was completed, testing was performed on a testing set and the estimation for testing gave 0.01551 average testing errors. This is presented in the graph in Fig. 9. The roles are reversed and testing is performed on training set. The output errors produced is 0.01551 which is the same to the first run, an indication of valid results.

The results for each execution are summed-up and divided by two. When converted into a percentage and to two decimal places, the overall average error rates attained 1.6% for both training and testing. Thus, the overall testing average accuracy percentage is 98.4%, which is an excellent result. The summary of results is illustrated in Table 2.

VI. COMPARISON AND DISCUSSIONS

Since the intelligent phishing security (IPS) focuses in enhancing feature-based approaches, the study is similar with Form's study [16] in that both methods used features. The main difference is that IPS method employed 56 features based on ANFIS in comparison to Form's 9 features applying SVM classifier. As it can be seen in summary in Table 1, IPS results is comparable with Barraclough's results [19]. Although features differ in size, IPS method obtained an average accuracy of 98.4%, compared to Barraclough's method, which achieved 98.74% accuracy. Although Barraclough's result is higher by a small margin (0.3 per cent), their method used 300 feature size in comparison to 56 utilized in IPS. This can be explained that phishing evolves regularly and some become redundant. Therefore, a well selected IPS features have dealt with redundant record and are covering the specific sensitive information attackers capture from users. Hence, IPS has offered a promising results in-line with [19] results.

This result is important since it evaluates the effectiveness of phishing e-web-form feature model. This is the first study to demonstrate the effectiveness of phishing e-web-form features. Using two-fold cross-validation in testing allows estimating the validity of phishing e-web-form feature model and its generalization to new attacks. The results indicate that phishing e-web-form feature model can classify between phishing, suspicious and legitimate websites with higher accuracy.

VII. EVALUATION AND CONCLUSIONS

Keeping in mind that phishing strategy evolves regularity; it is a challenge that puts pressure on the existing detection systems. This would apply to the e-web form feature model if they are not updated regularly in future. The limitation in this study is that the evolving phishing characteristics identification is difficult to be automated, hence are manually selected.

This paper analyzed phishing e-web-form sources to identify and extract effective features to classify and detect emerging phishing websites. 56 features were used based on ANFIS algorithms. These features are specific to the main sensitive information that attackers acquire from users. The intelligent phishing security approach obtained promising results which demonstrates effectiveness of phishing e-web-form and feature model to classify and detect phishing websites with a higher accuracy. This is the first study to use phishing e-web-form framework, which is a source for effective features that has demonstrated effectiveness to detecting emerging phishing attacks accurately.

Future work would be to construct fuzzy rules that can be used to identify phishing websites, using human problem-solving methods. Based on the effective phishing e-web-form, an effective phishing detective system will be built to detect real-world phishing websites.

ACKNOWLEDGMENT

Our greatest thanks goes to the head of the department of Computer and Information Sciences and the School of Engineering and Environment in Northumbria University for providing a space and facilities during the process of the research.

REFERENCES

[1] D. Correa, "Fraud costs UK £193bn per year, rise in phishing attacks seen". Online available < https://www.scmagazineuk.com/fraud-costs-uk-193bn-per-year-rise-in-phishing-attacks-seen/article/531293/> Accessed on 22nd May 2017, published in May 2016.

[2] L. Fang, W. Bailing, H. Junheng, S. Yushan, and W. Yuliang, "A Proactive Discovery and Filtering Solution on Phishing Websites", IEEE International Conference on Big Data (Big Data), 2015.

[3] B. Kumar, P. Kumar, A. Mundra, and S. Kabra, "DC Scanner: Detecting Phishing Attack", IEEE Third International Conference on Image Information Processing, 2015.

[4] S. Smadi, N. Aslam, L. Zhang, R. Alasem, and M. A. Hossain, "Detection of Phishing Emails using Data Mining Algorithms", 9th

International Conference on Software, Knowledge, Information Management and Applications (SKIMA), 2015.

[5] Z. Dong, A. Kapadia, J. Blythe, and L. J. Camp, "Beyond the Lock Icon: Real-time Detection of Phishing Websites Using Public Key Certificates", APWG Symposium on Electronic Crime Research (eCrime), 2015.

[6] M. Aburrous and A. Khelifi, "Phishing Detection Plug-In Toolbar Using Intelligent Fuzzy-Classification Mining Techniques, International Journal of Soft Computing and Software Engineering [JSCSE], 2013, Vol. 3, pp. 54-61.

[7] G. Xiang, J. Hong, C. P. Rose, and Cranor, L. "Cantina+: A feature-rich machine learning framework for detecting phishing web

[8] Sites", ACM Transactions on Information and System Security (TISSEC*)*, 2011, 14 (2), pp. 2 - 21.

[9] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, "An empirical analysis of phishing blacklists". In proceedings of the 6th conference on email and Anti-Spam. Mountain view, CA, USA, 2009.

[10] PhishTank. Statistics about phishing activity and phishTank usage. Online available <http://www.phishtank.com/stats.php. 2.2.4, 3.1, 4.4.2> Accessed on 15th November 2016.

[11] N. Abdelhamid, A. Ayash, and F. Tabatah, "Phishing Detection Based Associative Classification Data Mining", Expert System with Application, 2014, 41, pp. 5948-5959.

[12] R. Mohammad, T. L. McCluskey, and F. A. Thabtah, "Intelligent Rule based Phishing Websites Classification". IET Information Security, (2014) 8 (3). pp. 153-160. ISSN 1751-8709.

[13] M. Ajlouni, W. Hadi, and J. Alwedyan, "Detecting phishing websites using associative classification", European Journal of Business and Management, 2013, *5* (15), 36–40.

[14] N. Walia, H. Singh, and A.Sharma, "ANFIS : Adaptive Neuro-Fuzzy Inference System" A Survey," International Journal of Computer Applications (IJCA), 2015, vol. 123, pp. 32-38.

[15] P. Ozarkar and M. Patwardhan, "Efficient Spam Classification by Appropriate Feature Selection", International Journal of Computer Engineering and Technology (IJCET), ISSN 0976 – 6375 Vol. 4, (3), May – June 2013.

[16] H. Zuhair, A. Selamat, and M. Salleh, "Feature selection for phishing detection: a review of research," International Journal of Intelligent Systems Technologies and Applications, 2016, Vol. 15 (2) pp. 147–162.

[17] L. M. Form, K. L. Chiew, S. N. Szeand, and W. K. Tiong, "Phishing Email Detection Technique by using Hybrid Features", IT in Asia (CITA), 9th International Conference, 2015.

[18] M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah "Intelligent phishing detection system for e-banking using fuzzy data mining," Expert Systems with Applications: An International Journal, 2010, pp. 7913-7921.

[19] P. A. Barraclough, M.A. Hossain, M. A. Tahir, G. Sexton, and N. Aslam. "Intelligent phishing detection and protection scheme for online transactions" Expert Systems with Applications, Vol. 40, (11) pp. 4697-4706, 2013,

[20] P. A. Barraclough, M.A. Hossain, M. A. Tahir, G. Sexton, and N. Aslam, "Parameter optimization for intelligent phishing detection using adaptive neuro-fuzzy" International Journal of Advanced Research in Artificial Intelligence (IJARAI), 2014, Vol. 3, (10), 2014

[21] S. Marchal, J. Francois, R. State, T. Engel, "PhishStorm: Detecting Phishing with Streaming Analytics," Network and Service Management, IEEE Transactions, 2014 , vol.11, (4), pp.458-471.

[22] A. F. Yasin, A. Abuhasan, "An Intelligent classification model for phishing email detection" International Journal of Network Security & Its Applications (IJNSA), 2016, Vol.8, No.4.

[23] R. M. Mohammad, F. A. Thabtah, and T. McCluskey, "An Assessment of Features Related to Phishing Websites using an Automated Technique,". In the 7th International Conference for Internet Technology and Secured Transactions (ICITST-2012), London, 2012.

[24] O. Celik and S. Ertugrul," Predictive human operator model to be utilized as a controller using linear", Neuro-fuzzy and fuzzy-ARX modeling techniques, Engineering Applications of Artificial Intelligence, 2010, (4) 23, pp. 595-603.

[25] Millersmiles, "the web's dedicated anti-phishing services". Online available < http://www.millersmiles.co.uk/> accessed on 14th November 2016.

[26] O. E. Dragomir, F. Dragomir, V. Stefan, and E. Minca, "Adaptive neuro-fuzzy inference systems as a strategy for predicting and controlling the energy produced from renewable sources". Energies, 2015, vol *8*, pp. 13047–13061