Expert Systems With Applications 209 (2022) 118305

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Multi-scale semantic deep fusion models for phishing website detection

Dong-Jie Liu [a], Guang-Gang Geng [a,*], Xin-Chang Zhang [b]

[a] College of Cyber Security, Jinan University, Guangzhou 510632, China
[b] Qilu University of Technology, Jinan 250353, China

ARTICLE INFO

Keywords:
Deep learning
Semantic deep fusion
Phishing detection

ABSTRACT

In view of semantic counterfeiting characteristics of phishing websites and their multi-scale composition, this paper fully considers the semantic information of different scales, and proposes three semantic-based phishing detection models at different depths using various deep learning methods. The proposed three models are Multi-scale Data-layer Fusion (MDF) model, Multi-scale Feature-layer Fusion (MFF) model and Multi-scale In-depth Fusion(MIF) model. Experimental results on a constructed complex dataset show that the three models all have good recognition capabilities and the MIF model achieves the best performance on a complex dataset, with an F1-Measure of 0.9830, AUC value of 0.9993 and a false positive rate of 0.0047. Then with further comparison with both visual and text methods and an active discovery experiment lasting for 6 months with 3016 phishing websites detected in the real network environment, it is found that the proposed model is both competitive and practical for real detection scenarios.

1. Introduction

Phishing is an attempt to lure users to access counterfeit websites to steal confidential information. This strategy involves various channels such as malicious emails, instant messages and social media. With the rapid development of web applications, especially the popularization of electronic payment and online banking, phishing has become one of the most serious cyber Security threats.

Data collected by the Federal Bureau of Investigation (FBI) Internet Crime Complaint Center (IC3) shows that the total economic loss caused by cybercrime in 2018 was approximately US$2.7 billion, which is almost twice the figure in 2017. Among these losses, phishing scams accounted for almost half of the total losses reported in 2018, reaching nearly US$1.3 billion (Gorham, 2019). The 2nd Quarter 2020 APWG Phishing Activity Trends Report says that the average wire transfer loss caused by commercial email compromise (BEC) attacks is increasing: the average wire transfer attempt fee in the second quarter of 2020 was $80,183, which was significantly higher than the $54,000 in the first quarter. The average loss caused by a Russian BEC phishing against the company was US$ 1.27 million (APWG, 2020).

Although there are already a variety of anti-phishing tools and techniques used to combat phishing attacks, the number of phishing continues to grow rapidly. According to Volume 24 of the "Microsoft Security Intelligence Report", from January to December 2018, phishing attacks increased by 250%, and ransomware attacks are on the decline. Phishing website attacks remain a preferred attack method (Microsoft, 2019). According to Cohen, during the 2020 COVID-19 pandemic, phishing attacks increased by 350% year-on-year (Cohen, 2020). According to a report released by the International Anti-Phishing Working Group (APWG), the number of phishing websites reached a record 316,747 in December 2021, a year-on-year increase of more than 59% (APWG, 2021). Given the surge in phishing and its severe economic and privacy toll, how to effectively detect and deal with phishing has become an urgent network security issue.

With the popularity of Facebook, Twitter, TikTok and many other social media, the spread of phishing has become more diversified, which has caused the traditional phishing detection methods starting with emails to be unable to effectively prevent phishing. Since most phishing attacks lead victims to access phishing websites, identifying phishing websites as an entry point has become the most practical means of phishing defense.

Phishing websites often have semantic similarities with brand websites, including similar URLs, similar titles, and similar content. Their purpose is to achieve the effect of counterfeiting, making the victim appear to be browsing the same website, and then obtain high profits. At present, many phishing detection methods have noticed this, but they mainly extract features from a heuristic point of view, and often extract statistical features from the whole webpage. Related research ignores that a webpage is an organic combination of multiple functional modules. From the perspective of webpage display, this paper mines multi-scale semantic information of different modules of webpages, and

* Corresponding author.
E-mail addresses: djliu@jnu.edu.cn (D.-J. Liu), gggeng@jnu.edu.cn (G.-G. Geng), zhangxc@sdas.org (X.-C. Zhang).

https://doi.org/10.1016/j.eswa.2022.118305
Received 23 May 2022; Received in revised form 18 July 2022; Accepted 25 July 2022
Available online 1 August 2022

carries out semantic fusion at different depths to build semantic-based phishing detection models.

The motivation of this paper to conduct multi-scale semantic analysis is that semantic counterfeiting is one of the essential features of phishing websites and mining multi-scale semantic information in different semantic modules is lacking but well worth exploring. What is more, the design is under the premise of not relying on any third-party services and complex image features such as logos, webpage screenshots that consume bandwidth and storage so that it can be applicable in the real large-scale network detection environment.

The main contributions of this paper are as follows:

- This paper considers multi-scale semantic information of different modules of webpages and extracts multi-scale semantic information of URL, Title, Body Text and Invisible text (HTML Tags) from both URL and HTML, and performs their fusion of different depths, which is more efficient compared with methods relying on limited text information, third-party services or heuristic artificial features. On this basis, it further designs three end-to-end deep semantic information fusion models with different depths, and fully mines the complementary phishing identification information of semantic data of different scales.
- This paper proposes three semantic-based phishing detection models at different depths, including Multi-scale Data-layer Fusion (MDF) model, Multi-scale Feature-layer Fusion (MFF) model and Multi-scale In-depth Fusion (MIF) model. These models all use deep learning architecture and do not require any artificial features, or depend on any third-party services or image features that cost bandwidth and storage, and are very suitable for real large-scale network environment compared to methods using these features.
- To effectively verify the effectiveness of the proposed models, a complex phishing dataset is constructed and various comparative experiments with recent related research are carried out. The results prove the effectiveness and competitiveness of the three models and MIF model shows the best performance with an F1-Measure of 0.9830, AUC value of 0.9993 and a false positive rate of 0.0047.
- Not limited to a fixed dataset, an active phishing discovery experiment in the real network environment lasting for six months is also conducted, with 3016 phishing websites detected and thus further confirms the practicality of the proposed MIF model in reality.

The rest of sections are organized as follows. Section 2 presents a literature review. Section 3 introduces the extracted multi-scale semantic information. Section 4 proposes multi-scale semantic analysis based phishing detection models. Section 5 introduces the dataset and conducts detailed experimental evaluation. Section 6 conducts further discussion on different parameters and model robustness. Section 7 presents the active phishing discovery experiment in real network environment. Lastly, Section 8 draws the conclusion and discusses possible future work on anti-phishing.

2. Related work

Machine learning based phishing detection methods are currently a hot topic in anti-phishing research. Machine learning (including deep learning) mines phishing classification information from different angles, including visual information such as logo and favicon (Bozkir & Aydos, 2020; Chiew et al., 2018) and text information such as HTML and URL (Adebowale et al., 2020; Fang et al., 2019; Hiransha et al., 2018; Huang et al., 2019; Li et al., 2019; Peng et al., 2018; Ra et al., 2018; Yerima & Alzaylaee, 2020; Zhang et al., 2017) etc. The visual analysis method needs to identify and extract the corresponding visual elements, which requires a long time to crawl images, as well as high storage and computing resources and is not well applicable to large-scale scenarios. Considering that this paper focuses on semantic analysis, this section will review the anti-phishing research related to text analysis. Since some semantic detection methods for phishing emails can provide useful ideas for phishing website detection, this section includes both semantic-based phishing website and phishing email detection.

Most recent research adopts different models to analyze the URL for phishing detection. Sahingoz et al. extracted some string matching related NLP features from the URL and used 7 different machine learning based classification algorithms and Random Forest with only NLP based features performs the best on the constructed dataset (Sahingoz et al., 2019). Huang et al. used a Convolutional Neural Network (CNN) to extract character-level spatial feature representations of URLs; meanwhile, they employed an attention-based hierarchical Recurrent Neural Network(RNN) to extract word-level temporal feature representations of URLs. They then fused these feature representations via a three-layer CNN to build accurate feature representations of URLs, on which a phishing URL classifier was trained (Huang et al., 2019). Wei et al. treated each character of the URL as a semantic unit, mapping each character to a fixed-length embedding vector, and then used CNN for classification (Wei et al., 2020). Aljofey et al. used a character-level CNN still based on URL but their CNN contained 7 convolutional layers (Aljofey et al., 2020). Zhang et al. proposed a phishing detection method based on CNN and Bi-LSTM and they added Bi-LSTM on the basis of CNN to obtain URL long-distance dependent features (Zhang et al., 2021). Vazhayil et al. compared CNN and CNN-LSTM and found CNN-LSTM performed better (Vazhayil et al., 2018). In paper (Wei et al., 2019), the input of the model is a character-level URL, and the model structure is five parallel convolution layers. The lengths of the convolution kernels at the base layer of each volume are different. After their outputs are further linearly fused, the classification results are generated through three fully connected layers. Dilhara compares seven deep learning methods relying on URLs, including four hybrid deep learning models, namely GRU-LSTM, LSTM-LSTM, BI(GRU)-LSTM, and BI(LSTM)-LSTM . The experimental results in this paper show that the model with BI (GRU)-LSTM structure achieves the best results (Dilhara, 2021). These methods have reference significance for feature extraction and model construction; but they only rely on the URL, not including HTML, which reduces its practicability in the actual network environment, for phishers can evade URL detection easily at a very small price such as by changing one single letter or number.

Some research takes web content into consideration. Yerima and Alzaylaee (2020) utilized a CNN for classification. This paper only used one single scale of web content data, and mainly discussed the effect of the different parameter settings of 1-CNN on the model performance. But its conclusions have some reference significance for the construction of the single-scale semantic model of this paper. Ariyadasa et al. (2020)used a combined model of LSTM and CNN with the use of both URLs and HTML but they abstracted 15-dimensional HTML artificial features, that is, they did not fully use automatic features. Li et al. (2019) proposed a stacking model to detect phishing webpages using heuristic URL and HTML features. They introduced mean HTML string embedding on the whole HTML, which does not differentiate different parts of the webpage. The way they used Word2vec is similar to Zhang et al.'s work (Zhang et al., 2017), in which they also extracted semantic features through Word2vec on the whole HTML text to describe the features of phishing sites, and further fused them with other multiple statistical features to construct a more robust phishing detection model. In paper (Al-Ahmadi, 2020), a phishing detection system called phishing detection based on a multilayer perceptron (PDMLP) was proposed and 41 statistical features extracted from URLs and web content are used as model input. Opara et al. (2020) used CNN to analyze the content of the HTML document of a web page. It adopted three models: HTMLPhish-Character, HTMLPhish-Word, and HTMLPhish-Full. They

are based on character embeddings, word embeddings and concatenation of the two embeddings. The model structure of HTMLPhish-Full is that the two embeddings are cascaded and then enter the volume base layer. The experimental results of this paper show that the effect of HTMLPhish-Full is relatively optimal.

Adebowale et al. proposed a deep phishing detection model based on mixed features, in which semantically related features only included URL strings, anchor text information and image alterative text information. URLs were used to train the LSTM algorithm, without considering the semantic information of web content. Even on a balanced data set with equal numbers of positive and negative samples, the experimental results of the method in this paper are not satisfactory with a best F1-Measure value of 0.9329. Considering the imbalance of positive and negative samples in the real network environment, the practicality of the method needs to be further improved (Adebowale et al., 2020).

Some research on phishing email detection can also be used for reference. Verma et al. presented a general semantic feature selection method for text problems based on the statistical t-test and WordNet, and they show the effectiveness on phishing email detection by designing classifiers that combine semantics and statistics in analyzing the text in the emails (Verma & Hossain, 2013).Peng et al. presented a natural language processing technique approach which is used to analyze text and detect inappropriate statement that are indicative of phishing attacks (Peng et al., 2018).

DeepAnti-PhishNet applied deep neural network to detect phishing emails, and the experimental results prove that word embedding with deep learning, specifically LSTM is appropriate for the email anti-phishing task (Ra et al., 2018). Hiransha et al. used Keras Word Embedding and CNN to build their model to distinguish phishing emails from legitimate emails (Hiransha et al., 2018). Fang et al. proposed a phishing email detection model named THEMIS, which is used to model emails by analyzing the email header and the email body respectively (Fang et al., 2019). In this work, they used a recurrent convolutional neural network (RCNN) model with multilevel vectors and attention mechanism. Although the research object of this paper is email rather than webpage, it proves to a certain extent the necessity of analyzing research objects from different scales.

3. Multi-scale semantic analysis

Both visual phishing and semantic phishing are important means for phishing websites to deceive users. Visual analysis methods need to identify and extract corresponding visual elements, and crawling images is not only time-consuming but also requires large storage and computing resources; while semantic analysis only needs to crawl web content. Under the background that an ordinary server can crawl millions of webpages every day, the semantic analysis method is suitable for real large-scale phishing website detection scenarios. This paper explores models that uses only web content to achieve high-accuracy phishing detection.

The existing anti-phishing research based on semantic analysis mainly extracts features from a heuristic perspective, and extracts statistical features from HTML as a whole. Related research ignores the fact that webpages are an organic combination of multiple functional modules. Regarding the detection of phishing websites based on semantic analysis, if the different semantic modules in the webpage content are not distinguished, the webpage semantic information is likely to be ignored, confused or weakened. If semantic analysis is performed according to different scales and semantic fusion of different depths is performed, it is expected to more comprehensively improve the ability to identify phishing websites based on semantic analysis.

This paper intends to mine multi-scale semantic information of different modules of webpages, perform semantic fusion of different depths, and build a semantic-based phishing detection model. Multiple semantic functional blocks of webpages, which this paper calls multi-scale semantic information carriers, including “URL”, “webpage title”, “visible body text” and “invisible text” (HTML tag, etc.), etc. Among them, “webpage title”, “visible body text” and “invisible text” constitute different functional modules of webpages, and they are actually separated from HTML. The multi-scale semantic modules contain semantic information as follows:

URL is an acronym for Uniform Resource Locator and is a reference to a resource on the Internet. A URL has two main components: protocol identifier and resource name. For instance, the URL https://www.paypal.com/signin, the protocol identifier is https and the resource name is www.paypal.com/signin The resource name is the complete address to the resource, which contains one or more of the following components: host name, port number, path and query string. URL, especially the host name, is part of a brand; and the URL of a phishing website often inserts brand-related strings to read like the benign one. In addition, a phishing URL often contains some special characters such as “@”, “-”, numbers, or more dots to distinguish it from the address of a real website.

Title is a high-level summary of a webpage. Generally speaking, the title of a homepage is the official name of the website, and the title of the topic page is usually the topic name. When web users visit the webpage, the title is the area attracting the highest attention. In addition, the title of a login page of some brand websites that are frequently phished often contains sensitive words and are semantically similar. For example, the title of https://www.paypal.com/signin is: “Log in to your PayPal account”, which contains sensitive words such as the brand name “PayPal”, “log in” and “account”, which phishing sites targeting at PayPal also have.

Body Text refers to the text in the body of a webpage, which is the content that users can mainly see and usually reads very similar. It contains much sensitive information such as the brand name, the login box and so on. In HTML, the <body> tag is used to indicate the main body of a webpage. This paper takes *Body Text* as one input information, hoping to utilize its semantic difference to identify phishing.

Invisible Text in this paper refers to the tags used for syntax and layout in HTML, that is, the parts in the webpage source codes that play an indispensable role but are not visible to users. Phishing websites are often similar to brand websites in structure such as color tones and webpage layouts, and these similarities are often controlled by Invisible Text. In addition, the Invisible Text of phishing websites also contains sensitive words such as “Form” and “Password” used to steal users’ private information. This hidden phishing identification information is expected to contribute to phishing identification.

The following Fig. 1 is a sample of a phishing website targeted at Facebook, which indicates the multi-scale semantic information of different modules more clearly.

4. Multi-scale semantic fusion models of different depth

This paper proposes three phishing detection models based on different depth of multi-scale semantic information analysis, including: Multi-scale Data-layer Fusion (MDF) model, Multi-scale Feature-layer Fusion (MFF) model and Multi-scale In-depth Fusion model(MIF) model. Deep neural network algorithms such as word embedding, 1-CNN, LSTM, etc. have developed rapidly in text analysis tasks, which provides richer means for the semantic analysis in this paper. The proposed models are described separately below.The model architectures are presented in Figs. 2–4.

4.1. Multi-Scale Data-layer Fusion model

In Multi-Scale Data-layer Fusion (MDF) model, data layer fusion refers to cascading unprocessed multi-scale data to form unified fused data. On the basis of these data, phishing detection models are constructed for semantic pattern analysis. The multi-scale data to be cascaded in this paper includes: URL, Title, Body Text, Invisible Text,

Fig. 1. Semantic information of a phishing website.

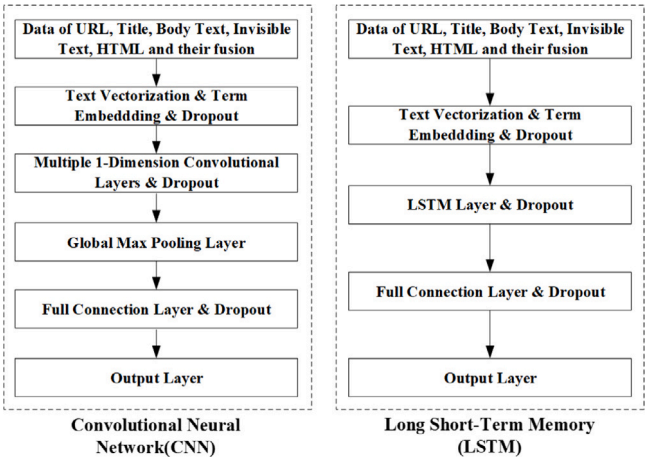

Fig. 2. Multi-scale data-layer fusion models (CNN & LSTM).

and HTML. For URL data, it is necessary to replace punctuation marks other than "-" and "_" with delimiters to facilitate subsequent word embedding analysis. The experiment will compare and analyze the recognition capabilities provided by single-scale data and fusion data of different scales, such as URL, URL+Title, URL+Title+Body Text, URL+Title+Body Text+Invisible Text, URL+HTML, etc. This paper will use CNN and LSTM models to mine the phishing website identification information on fusion data.

The fully connected layers of the CNN and LSTM models use the ReLU activation function, and the output layers both use the Sigmoid activation function.

$$ReLU(x) = \begin{cases} 0, & x < 0, \\ x, & x \geq 0. \end{cases} \tag{1}$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

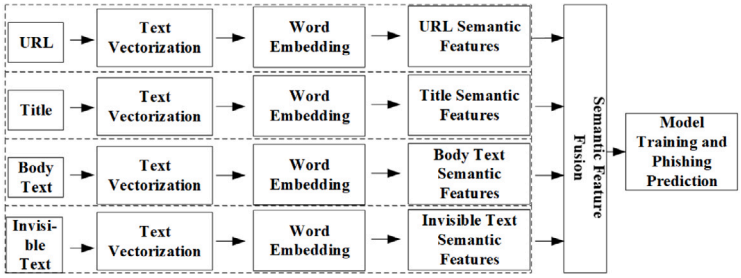

Fig. 3. Multi-scale feature-layer fusion model.

CNN and LSTM models use binary cross-entropy loss and adopt Adam optimization algorithm. Adam optimization is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments.

4.2. Multi-Scale Feature-layer Fusion model

Different from the first proposed MDF model which cascades unprocessed multi-scale data, the Multi-Scale Feature-layer Fusion (MFF) model first learns word embedding representation on the data of different scales. Then the word embedding of the vocabulary contained in the data of a certain scale is synthesized to obtain the embedding vector of the scale. Then multi-scale embedding vectors are fused to form a joint feature vector, and the classifier is trained on the joint vector space to obtain a detection model. The architecture of MFF model is as follows:

The multi-scale semantic vectors extracted in this paper include: URL semantic vector, Title semantic vector, Body Text semantic vector, Invisible Text semantic vector and HTML semantic vector. Zhang et al.'s work (Zhang et al., 2017), as well as the work of Li et al. (2019), only considered HTML semantic vector, but we compare more vectors' effectiveness through experiments. Data at each scale consists of multiple words, and the embedding of one single word cannot represent one certain scale of data. Therefore, it primarily needs to realize the mapping from word embedding to embedding representation of a certain scale of data.

Take HTML as an example. In Zhang et al.'s work (Zhang et al., 2017), the arithmetic average and TFIDF weighted average of all vocabulary embedding vectors in HTML are used as the embedding representation of HTML. The experimental results show that the effect of arithmetic average mapping is slightly better than TFIDF mapping. Therefore, this paper uses arithmetic mean mapping. The calculation process is as follows:

$$SF_i^H = \frac{1}{n_i}\sum_{j=1}^{n_i} W_{ij} \tag{3}$$

In the formula, SF_i^H represents the Semantic Feature representation of the ith HTML text. n_i is the number of words in the ith HTML text. w_{ij} is the word embedding of the jth word in the ith HTML text.

Similarly, the semantic feature representation of URL, Title, Body Text and Invisible Text are SF^U, SF^T, SF^B and SF^I respectively. If the data set is not large enough for word embedding training, some short text scales can also be merged, such as merging URL and Title, and the corresponding semantic feature is SF^{UT}.

Phishing detection based on fusion of semantic features can use a variety of machine learning algorithms, such as: AdaBoost, Random Forest, SMO, neural network, etc. This paper will compare the effectiveness of multiple algorithms.

4.3. Multi-scale In-depth Fusion model

Multi-scale In-depth Fusion (MIF) model designs multiple deep neural network structures for the proposed scales of semantic data including URL, Title, Body Text, and Invisible Text. Different from the MFF models, all modules perform in-depth semantic fusion after the maximum pooling and before the fully connected layer, and the fusion model sets a unified optimization function to optimize the loss end-to-end.

The MIF model minimizes the loss of classification by the optimization algorithm, and is expected to fully mine the integrated phishing website identification ability of multi-scale semantic deep features.

The number of 1-dimensional convolutional layers in each module can vary according to the length of data at different scales. In other words, the number of convolutional layers of Body and Invisible Text modules is larger than that of URL and Title modules. The activation functions of the convolutional layers all use ReLU. The cascade layer is to linearly cascade the deep features of each scale to form in-depth fusion features.

$$\Phi(ID) = \phi(ID^U) \oplus \phi(ID^T) \oplus \phi(ID^B) \oplus \phi(ID^I)\phi(ID^X) \tag{4}$$

In the formula, $\phi(ID^U)$, $\phi(ID^T)$, $\phi(ID^B)$ and $\phi(ID^I)$respectively represent in-depth URL features, in-depth Title features, in-depth Body Text features, and in-depth Invisible Text features, and $\phi(ID^X)$ represent scalable in-depth features at other scales, such as HTML.

The deep cascade feature is used as the input of the fully connected layer. From the cascaded layer to the output, it can be considered as a multi-layer perceptron. The fully connected layer uses the ReLU activation function, followed by a Dropout layer. The output layer predicts whether the website to be detected is a phishing website. As a typical binary classification problem, the output layer uses the Sigmoid function. The loss function of the deep multi-scale phishing detection model adopts binary cross entropy, and the optimization function adopts Adam algorithm.

5. Experiments

In order to verify the effectiveness of the proposed multi-scale semantic detection models of different depths, both phishing and legitimate samples from multiple sources are collected and a complex phishing dataset is constructed. Comparative experiments are then carried out based on the dataset.

5.1. The dataset

To fully verify the phishing detection performance of the models proposed in this paper, the following principles are first proposed to better construct the dataset:

1. There should be as many types of phishing samples as possible in the dataset, reflecting the complexity of the real Web environment and the diversity of phishing targeted brands to support the robustness and generalization of the proposed models.
2. The dataset should contain multiple confusing samples so that it can better verify whether the models are applicable in the complex real network environment.

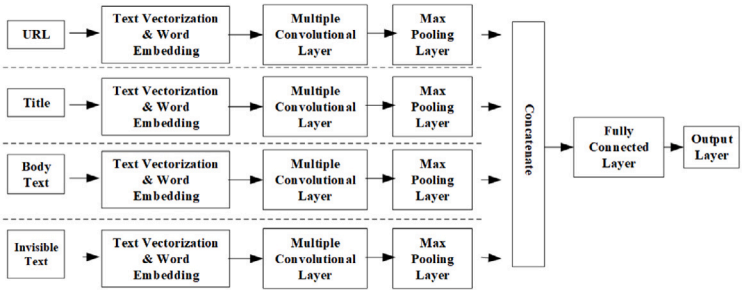

Fig. 4. Multi-scale in-depth fusion model.

Table 1
Confusion matrix.

		True	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

3. The ratio of positive and negative samples needs to reflect the imbalance of phishing detection in reality, that is, the number of phishing websites on the Internet is much smaller than that of legitimate websites.

Based on the above principles, this paper collects 1,564 phishing samples from APWG, APAC, PhishTank, etc., covering nearly 100 brand phishing websites, including PayPal, Yahoo, AOL, HSBC, CHASE, eBay, Bank of America, Facebook, Alibaba and so on. It also collects 3,829 non-phishing websites from DNS logs, DMOZ and search engines. The data obtained from search engines is a list of results returned by retrieving brand names. There are more than ten languages used in the sample webpages, such as English, French, Spanish, Portuguese, Japanese, Korean, Chinese, German, Italian, Dutch, Greek, Hungarian and so on. The domain names and content of multiple samples in the dataset contain brand-similar strings. Some confusing samples require even anti-phishing experts' careful observation. In addition, the ratio of phishing websites to non-phishing websites is 1:2.45, which reflects the imbalance of phishing detection to a certain extent.

To be noted, the complexity of the dataset is embodied in:(1) It contains a large number of legitimate samples that are difficult to identify. The dataset contains multiple confusing legitimate samples with similar brand domain names, legitimate websites containing brand titles and brand elements, etc., which are easily misjudged as phishing. (2) The dataset covers nearly 100 phishing targeted brands, which increases the difficulty of identification to achieve the diversity of phishing targeted brands to support the robustness and generalization of the proposed models. (3) There are more than ten languages used in the sample webpages. The fact that the samples cover multiple languages further increases the complexity of the dataset.

5.2. Evaluation metrics

The detection metrics including Precision, Recall, F1-Measure, Accuracy, True Positive Rate, False Positive Rate, and Area Under Curve are used in the experiments. Table 1 describes the basic terms for binary classification.

Based on the above, the metrics are explained as follows:

- Accuracy (ACC) is the ratio of the correct predictions to all the predictions, which is calculated as $(TP+TN)/(TP+FP+FN+TN)$;
- Precision (P) is the ratio of correct positive predictions to the total positive predictions, which is calculated as $TP/(TP+FP)$;
- Recall (R), also known as True Positive Rate (TPR), is the ratio of correctly predicted positive items to all the positive items, which is calculated as $TP/(TP+FN)$;
- False Positive Rate (FPR) is the ratio of wrongly predicted negative samples to all the true negative items, which is calculated as $FP/(FP+TN)$;
- F1-Measure (F_1) is the weighted harmonic mean of the precision and recall, which takes both Precision and Recall into account and is calculated as $2\times R\times P/(R+P)$;
- Area Under Curve (AUC) shows the performance of a detection model at all detection thresholds. AUC measures the predicted rankings rather than their absolute values.

5.3. Evaluation and results

In the experiments, the dataset is split into training set, validation set, and test set for five times, and for each time their ratio keeps as 3:1:1. All the results are the average of the experimental results on the 5 splits, which are expected to be a more accurate estimate of the performance of the proposed semantic-based phishing detection models on the dataset.

(1) ***Evaluation of MDF Model***

To verify the performance of the proposed semantic based Multi-scale Data-layer Fusion (MDF) model, two deep learning algorithms, CNN and LSTM, which have been proved effective in many fields are used. The settings of the relevant parameters mainly refer to Johnson and Zhang (2017), Wei et al. (2019) and Wielgosz et al. (2017). Different parameters are tried and the best ones are chosen. Further discussion on different parameters is in Section 6.1

The structure of CNN is described as follows:

- One text vectorization layer with a sequence length of 500;
- one embedding layer with the embedding dimension as 256;
- one dropout layer with a dropout rate of 50%;
- two convolutional layers, with the output dimension set as 128, the kernel size set as 7, the stride length of the convolution set as 3, and with ReLU activation function;
- one pulling layer, which conducts global max pooling operation for 1D temporal data;
- two dense layers with ReLU activation function with the output dimension of the first dense layer set as 128, and the second one as 64;
- a dropout layer, with the input units dropout rate of 50%;
- the output layer using a sigmoid activation function.

The LSTM structure includes:

- One text vectorization layer with a sequence length of 500;
- one embedding layer with the embedding dimension as 256;
- one dropout layer with a dropout rate of 50%;
- one LSTM layer, with the output dimension set as 128;
- one dropout layer, with the input units dropout rate of 50%;

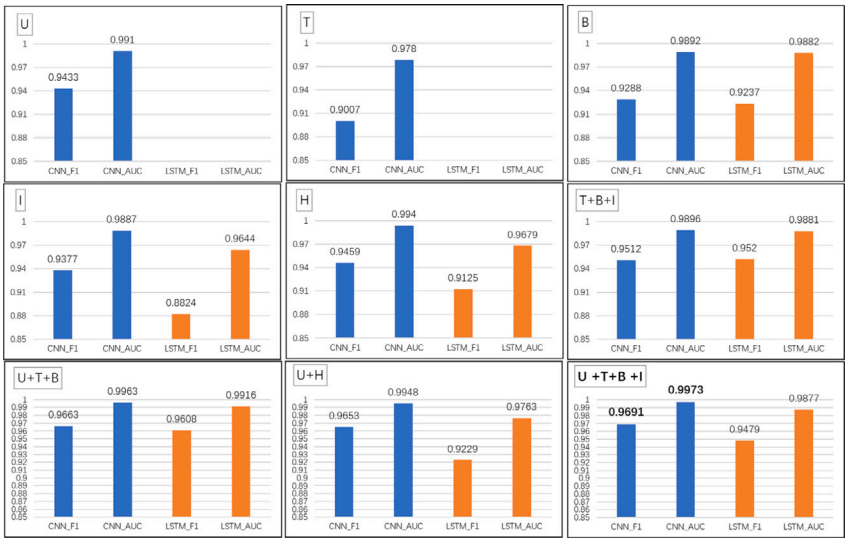

Fig. 5. Performance of MDF model.

- one dense layer with ReLU activation function with the output dimension set as 128;
- one dropout layer, with the input units dropout rate of 50%;
- the output layer using sigmoid activation function.

The multi-scale data used in the experiment include: URL(U), Title(T), Body Text(B), Invisible Text(I) and HTML(H), as well as their various fusions. Fig. 5 shows the comparison of the experimental results of MDF model on the constructed dataset. As mentioned before, since F1-Measure is the weighted harmonic mean of the precision and recall and AUC shows the performance of a detection model at all detection thresholds, to present the results more clearly, F1-Measure and AUC are adopted.

From Fig. 5, it is seen that with both CNN and LSTM, the multi-scale fusion data model shows better performance than the single-scale data model, indicating that the data of multiple scales have complementarity in phishing recognition. For instance, URL or Title using LSTM as the classifier barely shows its classification capability, which is why they do not even appear in the figure. What is more, although the three scales—Title, Body, and Invisible Text are all from HTML, the model based on the fusion data of Title, Body, and Invisible Text shows better phishing recognition ability than the model trained directly using HTML. Additionally, on the same scale of data, the performance of CNN is better than that of LSTM. Comprehensive use of URL, Title, Body Text, and Invisible Text achieves the highest F1-Measure(0.9691) and AUC(0.9973) with CNN as the classifier, better than the best results obtained by LSTM with URL, Title, and Body Text combined.

(2) ***Evaluation of MFF model***

This section experimentally analyzes the effectiveness of the proposed Multi-scale Feature-layer Fusion (MFF) semantic phishing detection model. The semantic features on each scale are set to 128 dimensions, and the embedded value of each scale is the average of word embedding values, including semantic feature representation of URL(F^U), Title(F^T), URL and Title Merged(F^{UT}), Body text(F^B), Invisible text(F^I) and HTML(F^H). Skip-gram algorithm is used to learn word embedding, where maximum distance between the current and predicted word within a sentence is 5. All word embedding values are trained on the corresponding scale data, that is, the embedding values of a word on different scales are different. This design is to mine contextual semantic information within the scale data. The fusion of multi-scale semantic features adopts linear fusion. Four classifiers, Neural Network (NN), SMO, AdaBoost, and Random Forest are adopted. The neural network used is a multilayer perceptron (MLP), which is a fully connected neural network, which consists of an input layer, a hidden layer, and an output layer. Momentum applied to the weight updates is set as 0.2. The learning rate for weight updates is set as 0.3. For SMO, polynomial kernel is used; the calibration method to use is multinomial logistic regression model with a ridge estimator; the epsilon for round-off error is set as 1.0E−12; and the complexity parameter C is set as 1.0. For AdaBoost, decision Stump is used as the base classifier and the number of iterations to be performed is set as 150. Regarding Random Forest, the number of iterations is set as 300 and the maximum depth of the tree is set as 5. Fig. 6 shows its performance trained with different scales of semantic features. For each scale, the values of the best performance are presented.

As shown in Fig. 6, the four classifiers, Neural Network (NN), SMO, AdaBoost, and Random Forest trained on multi-scale semantic fusion features all achieve good performance. Among them, the performance of the AdaBoost algorithm on semantic feature representation of URL and Title Merged, Body text and Invisible text is the best, with the F1-Measure as 0.9780 and AUC as 0.9987. The three-scale semantic feature fusion $F^{UT} + F^B + F^I$ is slightly better than the four-scale semantic feature fusion $F^U + F^T + F^B + F^I$ possibly because: First, the dataset is not large enough to support URL and Title to fully learn word embedding separately; second, there are semantic similarities between URL and Title, and unified training of word embedding is more sufficient.

What is more, we further integrate semantic representation of HTML and the results in Fig. 6 show a decrease in model performance. This indicates that the three scales of Title, Body and Invisible Text have already represented the page well, and adding HTML will increase noise or lead to overfitting. Besides, the F1-Measures of only using semantic feature representation of HTML in (Zhang et al., 2017) are all below 0.94, which are much worse than the other three methods using semantic feature fusion, and thus proves again the complementarity of different scales.

Additionally, by further comparing the results of MDF model and MFF model in Figs. 5 and 6, it is seen that the performance of MFF model, the feature-layer fusion model is much better than MDF model, the data-layer fusion model, especially on the value of F1-Measure, 0.9780 and 0.9691 respectively.

(3) ***Evaluation of MIF model***

The in-depth features of each scale come from the corresponding scale of semantic data trained on the CNN Model. The front-end structure of each CNN includes one text vectorization layer, one embedding layer, one dropout layer, two one-dimensional convolutional layers and one pooling layer, followed by the in-depth feature layer. Multiple in-depth features adopt linear fusion. Taking the in-depth fusion of URL,

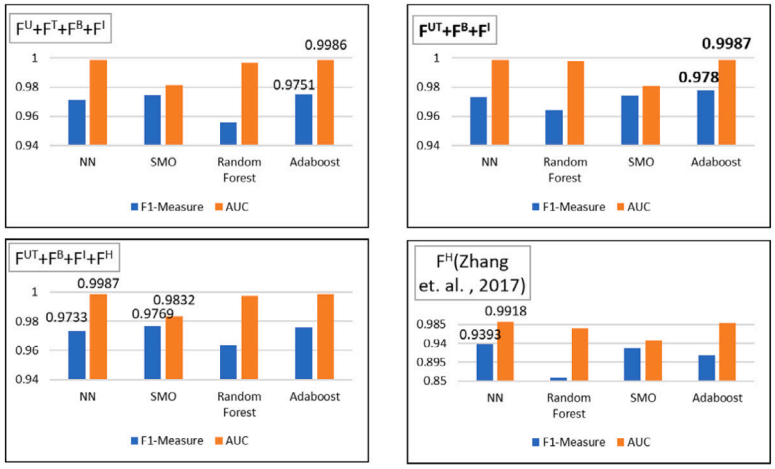

Fig. 6. Performance of MFF model.

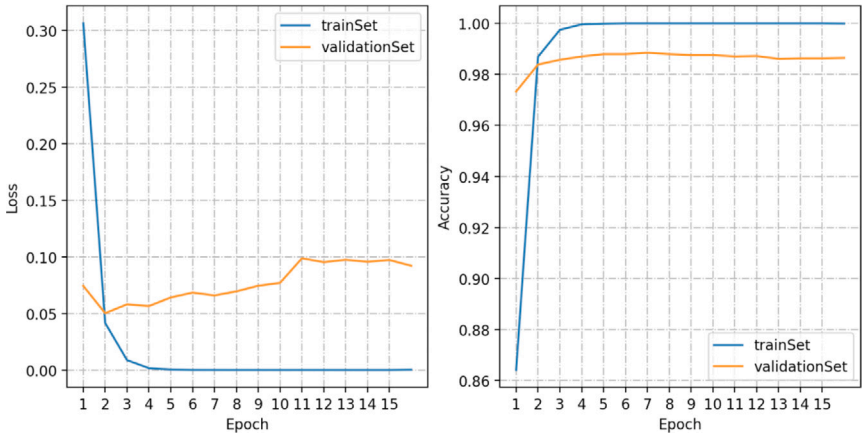

Fig. 7. Performance of MIF model with $\phi(ID^{UT}) \oplus \phi(ID^B) \oplus \phi(ID^I)$ on training and validation set.

Table 2
Performance of MIF model.

Deep Features	FPR	Precision	Recall(TPR)	F1-Measure	AUC
$\phi(ID^U) \oplus \phi(ID^H)$	0.0151	0.9631	0.9667	0.9649	0.9942
$\phi(ID^U) \oplus \phi(ID^T) \oplus \phi(ID^B)$	0.0060	0.9852	0.9738	0.9794	0.9983
$\phi(ID^{UT}) \oplus \phi(ID^B)$	0.0165	0.9606	0.9712	0.9657	0.9972
$\phi(ID^{UT}) \oplus \phi(ID^B) \oplus \phi(ID^I)$	**0.0047**	**0.9884**	**0.9776**	**0.9830**	**0.9993**
$\phi(ID^U) \oplus \phi(ID^T) \oplus \phi(ID^B) \oplus \phi(ID^I)$	0.0050	0.9878	0.9722	0.9799	0.9989
$\phi(ID^{UT}) \oplus \phi(ID^B) \oplus \phi(ID^I) \oplus \phi(ID^H)$	0.0063	0.9847	0.9802	0.9824	0.9993

$\phi(ID^U)$: In-depth URL features, $\phi(ID^{UT})$: In-depth URL and Title merged features, $\phi(ID^T)$: In-depth Title features, $\phi(ID^B)$: In-depth Body Text features, $\phi(ID^I)$: In-depth Invisible Text features, $\phi(ID^H)$: In-depth HTML features.

Title, and Body Text as an example, the dimension of deep fusion features is 384, followed by two fully connected layers, whose output dimensions are 384 and 192 respectively. The ReLU activation function is used; before the output layer it is a dropout layer, and the dropout ratio is set to 0.5; the output layer uses the sigmoid activation function. The loss function uses binary cross entropy and the optimizer of the deep fusion model selects Adam.

Comparing with MDF model and MFF model, the performance of MIF model is the best, achieving the best F1-Measure and AUC values. This also shows that the multi-scale semantic representations based on deep neural networks have good complementarity. What is more, the fusion of URL and Title data as one scale is more effective than using them separately. Further fusion of HTML in-depth feature will also bring noise to the model, which is similar to MFF model.

MIF model with the three scales (URL+Title, Body Text, Invisible Text) that achieves the best results is further discussed in detail. The changes of Loss and Accuracy values with the increase of epochs on the training and validation sets in the model learning stage are presented in Fig. 7.

It shows that as the epoch increases, the Loss on the training set continues to decrease, and the accuracy rate tends to 100%; the results on the validation set show that when the epoch is greater than 10, the model is overfitted, and the accuracy on the validation set begins to decrease.

Fig. 8 shows the model performance on the test set as the epoch increases. It can be seen from Fig. 8 that when the epoch is greater than 10, the Precision becomes larger at the cost of a sharp decline in recall, and the corresponding comprehensive metric F1-Measure decreases, and the accuracy metric also decreases. Generally speaking, when the epoch is 8, F1-Measure and AUC both achieve optimal values. The optimal values of MIF model ($\phi(ID_{UT}) \oplus \phi(ID_B) \oplus \phi(ID_I)$) in Table 2 are thus obtained when epoch is 8.

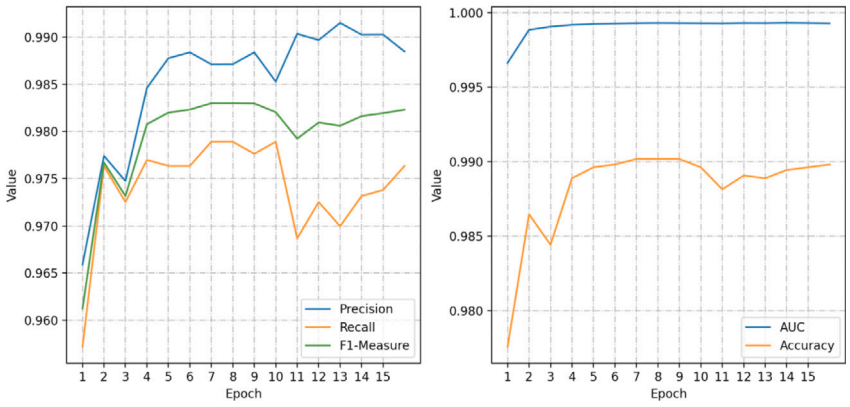

Fig. 8. Performance of MIF model with $\phi(ID^{UT}) \oplus \phi(ID^{B}) \oplus \phi(ID^{I})$ on test set.

Table 3
Comparison of proposed method with visual based analysis methods.

Perspective	Model	FPR	Precision	Recall	F1-Measure	AUC
Visual	CNN (favicon)	0.0175	0.9504	0.8002	0.8684	0.9272
	CNN (webpage screenshot)	0.0288	0.9281	0.9103	0.9191	0.9842
	CNN (favicon + webpage screenshot)	0.0248	0.9373	0.9102	0.9236	0.9849
Semantic	MDF model (CNN on $U + T + B + I$)	0.0100	0.9754	0.9630	0.9691	0.9973
	MFF model (AdaBoost on $F^{UT} + F^{B} + F^{I}$)	0.0094	0.9771	0.9789	0.9780	0.9987
	MIF model (CNN on $\phi(ID_{UT}) \oplus \phi(ID_{B}) \oplus \phi(ID_{I})$)	**0.0047**	**0.9884**	**0.9776**	**0.9830**	**0.9993**

In summary, the multi-scale semantic based phishing detection models at different depths are proposed and all show good phishing recognition abilities. In the real Web environment, these models are easy to be built and upgraded, showing good competitiveness, especially the MIF model with the three scales (URL+Title, Body Text, Invisible Text) which achieves the best phishing recognition performance.

(4) ***Comparison with visual methods***

The phishing detection models proposed in this paper are based on the semantic analysis of texts using machine learning algorithms (including deep learning) at different depths, taking into account the different components of the webpage text. In addition to semantic similarity of phishing websites, visual similarity is also an important feature. Existing phishing detection methods based on visual element analysis such as favicon (Chiew et al., 2018) and screenshot of webpages (Liu & Lee, 2020) have been proven effective. So it is meaningful to compare the proposed semantic models and the visual methods.

Therefore, a visual analysis model experiment on the same dataset was conducted on the same training, validation, and test sets. Experiments were carried out on favicon, screenshot of webpage, and their fusion using CNN. The parameters refer to Liu and Lee (2020) and the convolution kernel size is 3×3, the pool size is (2,2) and the activation function is ReLU. The experimental results are shown in Table 3.

It is obvious from Table 3 that the models proposed in this paper, especially the MIF model, have better phishing detection performance than the three visual models with a much higher F1-Measure, which indicates that the proposed model has obvious advantages in detection accuracy and comprehensiveness. The underlying reasons why the detection effect of the proposed multi-scale deep semantic fusion models are better than that of visual methods are that compared with visual similarity, all phishing websites have semantic similarity in order to deceive users; and the proposed models can effectively use the complementary identification information of multiple scale semantic data, and can effectively mine phishing website patterns.

More importantly, the proposed semantic deep models have more efficiency advantages than visual models. Take MIF with CNN on $\phi(ID_{UT}) \oplus \phi(ID_{B}) \oplus \phi(ID_{I})$ and CNN on screenshot of webpage as an example. On a PC with a CPU of 2.3 GHz, when the number of iterations is set as 65, the model learning time is 1704.88s and 4502.95s respectively. The detection time for a single page is 0.0069s and 0.0130s respectively. The above data is obtained under the premise that the input screenshot is resized as $200 \times 200 \times 3$. If a larger screenshot is used, the efficiency of visual analysis will be further reduced. Besides, visual analysis requires much more computing and storage space than semantic analysis. So the models proposed in this paper are more effective and efficient than the visual analysis models.

In addition, the proposed models are easy to be built and upgraded, and can effectively cope with the evolution of phishing websites, and facilitate the coverage of new phishing targeted brands, for the only thing to do manually is to collect new samples. So the proposed semantic deep models are suitable for actual large-scale complex phishing detection scenarios. Nevertheless, with sufficient computing and storage resources, for application scenarios that do not require high detection efficiency, on the basis of the proposed multi-scale semantic information in this paper, further consideration of fusion of visual information will be expected to further improve the capability of phishing detection.

(5) ***Comparison with text methods***

In Related Work, this paper describes the research on phishing website identification based on text analysis in recent years. This section further analyzes and compares these related methods. To be noted, the implementation of Wei et al. (2020) adopts the three convolutional layers recommended in this paper. For Dilhara (2021), the BI (GRU)-LSTM that the author thinks is optimal among the 7 deep model structures is implemented. The results are presented in Table 4.

From Table 4, it is obvious that the proposed MIF with CNN on $\phi(ID_{UT}) \oplus \phi(ID_{B}) \oplus \phi(ID_{I})$ performs the best. The possible reason is that URL-based analysis does not require the retrieval of target website

Table 4
Comparison of the proposed method with text based analysis methods.

Deep Semantic analysis Method	FPR	Precision	Recall	F1-Measure	AUC
Wei et al. (2020)	0.0279	0.9281	0.8823	0.9046	0.9751
Aljofey et al. (2020)	0.0290	0.9261	0.8875	0.9061	0.9772
Wei et al. (2019)	0.0222	0.9434	0.9041	0.9232	0.9870
Dilhara (2021)	0.0279	0.9291	0.8964	0.9124	0.9810
Opara et al. (2020) HTMLPhish-Character	0.0282	0.9295	0.8990	0.9137	0.9814
Opara et al. (2020) HTMLPhish-Full	0.0283	0.9297	0.9158	0.9227	0.9831
Vazhayil et al. (2018)	0.0298	0.9258	0.9047	0.9149	0.9800
Zhang et al. (2021)	0.0266	0.9344	0.9201	0.9269	0.9844
MIF model (CNN on $\phi(ID_{UT}) \oplus \phi(ID_B) \oplus \phi(ID_I)$)	**0.0047**	**0.9884**	**0.9776**	**0.9830**	**0.9993**

content or the use of any third-party services. It is easy to implement, however in practice there are a large number of confusing URLs on the Internet. The dataset in this paper is constructed with this in mind, and contains a large number of samples with URL similarity, such as brand-similar hosts found from DNS logs and URLs obtained by retrieving brand names from search engines, which leads to the result that the models in the papers such as Aljofey et al. (2020), Dilhara (2021), Wei et al. (2019) and Wei et al. (2020) are not effective for analysis relying only on URLs.

In more detail, the performance of Wei et al. (2019) (F1-Measure is 0.9232) is significantly better than Aljofey et al. (2020) and Wei et al. (2020). The potential important reason is that Wei et al. (2019) uses 5 parallel convolutional layers, and their convolution kernel sizes are: 2, 3, 4, 5, and 10. This multi-scale convolutional network structure fully exploits the semantic information in the URL, which also provides an idea for the subsequent multi-scale convolution kernel optimization of the multi-scale semantic fusion model proposed in this paper. LSTM (Dilhara, 2021) and the serial dual-model structure CNN-LSTM (Vazhayil et al., 2018) and CNN-BLSTM (Zhang et al., 2021) are slightly better than single model structure in Aljofey et al. (2020) and Wei et al. (2020), which shows that different structural models have a certain complementarity in phishing identification.

HTMLPhish-Character (Opara et al., 2020) is a method for phishing identification using HTML character embedding as CNN input. The experimental results show that the F1-Measure is only 0.9137, which is different from the model for HTML word embedding as CNN input in the MDF Model shown in Figure 5 of this article. Compared with the F1-Measure of 0.9459, the HTML character embedding phishing recognition ability is insufficient, which can be further reflected in the results of HTMLPhish-Full. The fusion of character embedding and word embedding features leads to interference between features, resulting in poor performance of the model.

In short, the above-mentioned methods based on text analysis are inferior to the model proposed in this paper. The underlying reason is that the above methods only involve part of the scales of the multi-scale semantic information analyzed in this paper, and do not comprehensively use the complementary recognition ability of multi-scale semantic information. What is more, the MIF deep fusion model proposed in this paper fully mines the above-mentioned multi-scale semantic information for phishing recognition and further strengthens the phishing recognition ability.

6. Further discussion

6.1. Further discussion on different parameters

In order to discuss the influence of the selection of different parameters on the performance of the model, this paper further conducts several sets of comparative experiments with different parameters on the MIF model with CNN on $\phi(ID_{UT}) \oplus \phi(ID_B) \oplus \phi(ID_I)$ that achieves the best results, and the MDF model with CNN on U+T+B+I and LSTM on U+T+B that perform the best for each algorithm (CNN on U+T+B+I performs better than LSTM). The important parameters of the comparative analysis include: kernel size, stride, dropout ratio and number of filters. The parameters are set as $kernalsize = 7$, $stride = 3$, $dropoutratio = 0.5$, and $filters = 128$ in the experiment. The experimental results of the following different parameters are the F1-Measures of the model when the other three parameters are fixed and the other parameter takes different values. The effects of different parameters are presented as follows (see Fig. 9).

Table 5
Comparison of different parameters on MDF model with CNN on U+T+B+I.

Kernel size	F1-Measure	Stride	F1-Measure
3	0.9587	1	0.9682
5	0.9633	2	0.9675
7	0.9691	3	0.9691
9	0.9678	4	0.9633
11	0.9673	5	0.9619
Dropout rate	**F1-Measure**	**Number of filters**	**F1-Measure**
0.1	0.9675	32	0.9638
0.3	0.9661	64	0.9673
0.5	0.9691	128	0.9691
0.7	0.9646	256	0.9665
0.9	0.9675	512	0.9666

Table 6
Comparison of different parameters on MDF model with LSTM on U+T+B.

Dropout rate	F1-Measure	Number of filters	F1-Measure
0.1	0.9542	32	0.9501
0.3	0.9549	64	0.9583
0.5	0.9608	128	0.9608
0.7	0.9605	256	0.9580
0.9	0.9537	512	0.9577

Regarding the size of the convolution kernel, if the size of the convolution kernel is too small, the complete word representation cannot be extracted; if the size is too large, noise will be introduced. The recommended convolution kernel size for the proposed model is 7 or 9. In regard to stride, it is the distance that the window slides each time a convolution is performed. If the stride is set too small, there will be more repeated calculations, the amount of calculation will be large, and the training efficiency will be reduced. If the stride setting is too large, information will be missed and the features behind the data cannot be effectively extracted. The data in the table above shows that when stride is set to 3 or 4, the model works best.With respect to dropout, some nodes in the network are randomly discarded according to a certain probability to prevent overfitting of the model. In this experiment, the best effect is when the dropout value is 0.5. Number of filters is the number of output filters in the convolution (i.e. the dimensionality of the output space). The larger the value of filters, the more complex the model and the longer the training time. The comparative experimental results show that 32, 64, and 128 filters are all good choices.

Under different parameter settings, the performance comparison of the MDF model of CNN on U+T+B+I and the model of LSTM on U+T+B is shown in Tables 5 and 6 .

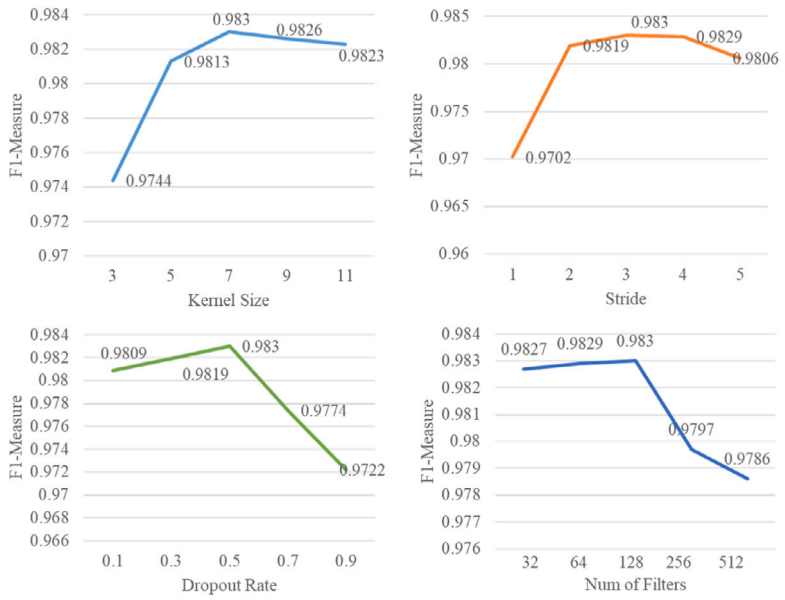

Fig. 9. Comparison of different parameters on MIF model with CNN on $\phi(ID_{UT}) \oplus \phi(ID_B) \oplus \phi(ID_I)$.

As shown in Table 5, in the case of different values of the four parameters, the performance variation of MDF model with CNN on U+T+B+I are similar to MIF model with CNN on $\phi(ID_{UT}) \oplus \phi(ID_B) \oplus \phi(ID_I)$. Table 6 shows the main parameters and F1-Measure values of the MDF model with LSTM on U+T+B. It is seen that when the dropout rate is set to 0.9, the performance is degraded due to too much discarded content. In the actual experiment, when the dropout rate becomes larger, the epochs required for model training increase significantly. In practice, it is recommended that the value should not exceed 0.7.

In summary, through the comparative experimental analysis of multiple groups of different parameters, the parameters set in Section 5.3 of this paper can achieve good results. In practical application, due to the difference of data sets, the relevant parameters need to be adjusted accordingly.

6.2. Further discussion on model robustness

There are many phishing and legitimate websites on the Internet, and some complex samples may not be included in the dataset. The adaptability and robustness of the model are important aspects to evaluate the effectiveness of the model. In addition, driven by huge interests, in order to avoid supervision, criminals often change the form of phishing websites, such as changing templates. These changes often lead to the degradation of the effectiveness of the phishing detection model. It has been 2 years since the data set construction time in Section 5.1, which provides conditions for verifying the robustness of the model by collecting new samples, especially the degradation of the model over time. We again collected multiple new complex samples following the dataset construction principles in Section 5.1. Confusing legitimate samples are collected from DNS recursive resolution log and Google search engine with brand-related string in URL or HTML; phishing samples are some of the latest phishing websites, which contain new templates, pictures, etc. A total of 971 complex test cases were collected, including 284 legitimate samples and 689 phishing samples. In particular, these new phishing sites contain 17 picture-in-picture phishing (picture-in-picture is a cheating method for turning text into pictures). The newly collected complex samples are all used for testing, that is, the training set and the validation set still use the original data, which can better verify the robustness and generalization of the proposed model over time. The experimental method of this part is similar to the previous experiment, that is, the average of the results of 5 random division experiments is still used, the difference is that 971 difficult samples are added to the test set.

The experiment to test the effectiveness of the model on an increased complex test sample is conducted. The experiments compare the performance of the best performing visual method CNN (favicon + webpage screenshot) and the best multi-scale semantic method MIF with CNN on $\phi(ID_{UT}) \oplus \phi(ID_B) \oplus \phi(ID_I)$ in the proposed method.

From Table 7, the test results with new complex samples added show that the visual method has poor adaptability to the continuous evolution of phishing websites, and the comprehensive index F1-Measure has dropped from 0.9236 to 0.8301; while for the MIF model with CNN on $\phi(ID_{UT}) \oplus \phi(ID_B) \oplus \phi(ID_I)$ proposed in this paper, although the F1 value has dropped by 0.0273, it is still as high as 0.9557. It is worth mentioning that 16 of the 17 picture-in-picture phishing sites were correctly identified because the webpage URL, invisible text, and title still provide recognizable information despite the text being pictured. The only unrecognized picture-in-picture phishing is because its URL, HTML Tag, and title do not contain semantic phishing-related information, but such phishing also reduces its ability to deceive users. In conclusion, the experimental results reflect that the proposed semantic model has good robustness, and the degradation degree is slower over time, which further indicates that semantic counterfeiting is a more essential feature of most phishing websites than visual counterfeiting.

7. Active phishing discovery experiment in real network environment

The analysis object of the multi-scale semantic deep fusion models proposed in this paper is the HTML file of the webpage, which has natural advantages in phishing detection in a large-scale network environment, as shown in the following aspects:

- Compared with the methods based on visual analysis (logo Bozkir & Aydos, 2020, favicon Chiew et al., 2018, webpage screenshots (Liu et al., 2019), etc.), the models proposed in this paper do not need to download pictures that consume a lot of bandwidth, and neither do they need to consume a lot of picture storage space, which is more efficient than visual analysis;
- The models proposed in this paper do not rely on any third-party features, and is more practical for large-scale scenarios than methods using third-party features (Liu et al., 2021; Ozcan et al.,

Table 7
Test results after adding new complex samples.

Method	FPR	Precision	Recall	F1-Measure	AUC
CNN (favicon + webpage screenshot)	0.0137	0.9288	0.8296	0.8301	0.9923
MIF model (CNN on $\phi(ID_{UT}) \oplus \phi(ID_B) \oplus \phi(ID_I)$)	0.0082	0.9544	0.9582	0.9557	0.9956

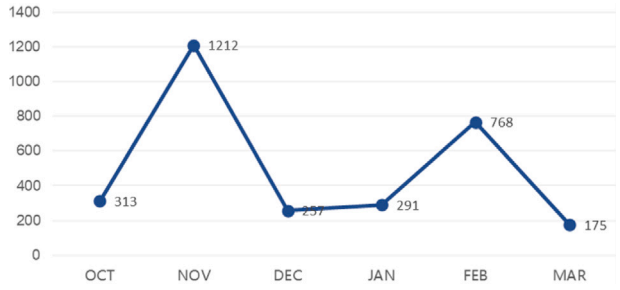

Fig. 10. Number of phishing websites discovered per month.

Table 8
Information about the top 10 newly discovered phishing sites.

Brand name	Number	Type
Agricultural Bank of China	633	Online bank
JD.COM	405	e-commerce
Taobao and Alipay	381	e-commerce
China UnionPay	379	Online bank
China Mobile	315	Mobile communication
Bank of China	276	Online bank
ICBC	227	Online bank
China Merchants Securities	173	Stock investment
Facebook	131	Social media
Apple	96	Mobile communication

2021; Saha et al., 2020; Wang et al., 2019; Yerima & Alzaylaee, 2020).

To verify the effectiveness of the proposed models in a real network environment, we conducted a half-year (2020.10.01–2021.03.31) phishing discovery experiment using MIF model which has the best performance in the dataset experiments. The daily detection data source includes two parts: one is the host data from the local recursive DNS service log, and the other is the Zone file data of COM and NET domain names. Considering the fact that the vast majority of phishing websites use newly registered domain names (Li et al., 2016), the detection object of this experiment is set to new hosts every day. The service object of the DNS recursive server used in this experiment is local users. The daily resolution of the service is about 1.5 billion, the number of hosts after deduplication is about 4 million, and the actual number of hosts detected every day is about 90,000. The Zone File of the .COM and .NET domain names used in this experiment is the difference between the data of two consecutive days, and the number of newly detected hosts is about 140,000 and 8,000 per day.

In order to verify the correctness of the actual detection results of phishing websites, all phishing websites discovered actively are reported to the China Anti-Phishing Website Alliance. The alliance has a professional audit team, which identifies the reported data every day, and handles phishing websites by stopping website domain name resolution and notifying browser manufacturers to block phishing websites. The alliance only deals with the brands of alliance members, mainly local companies, but also includes a few international brands such as Facebook and Apple, so the brands detected are phishing websites around these brands.

In this experiment, only 2 personal computers are used to crawl all webpages and perform detection. A total of 3,016 phishing websites were detected, covering 57 brands, and all websites were reported to APAC in a timely manner on the same day. All of these sites were identified as phishing sites, according to the alliance's audit. The number of phishing sites detected each month is shown in Fig. 10.

The month with the highest number of phishing sites shown in the figure is November, with the potential reason of Double Eleven Shopping Festival. The month with the second largest number is February, the underlying reason being the Chinese New Year, which is also the peak consumption season. Among the 57 phishing brands detected, the top 10 brands are shown in Table 8.

The average life cycle of phishing websites found by active phishing detection in this paper is less than 2.3 days, while the average life cycle of phishing websites reported by the China Anti-Phishing Website Alliance is more than 18 days. Considering that every more second phishing website survives, more people will be deceived, this work greatly improves the efficiency of anti-phishing. One of the data sources for this work is the local DNS recursive service logs, which are only available to millions of users. It is conceivable that if service data such as Google public DNS (8.8.8.8) and DNSPod's Public DNS+ (119.29.29.29) available to hundreds of millions and billions of users are used, the number of phishing websites discovered will be greatly increased. The above discovery results using only two servers for half a year prove the practicability of the proposed model in the actual complex network environment.

8. Conclusion and future work

Phishing has become one of the biggest cyber security threats and how to effectively detect phishing websites through technical means is an important countermeasure. The current research on phishing website detection based on machine learning is mainly relying on the limited text input such as URL only or relying on third-party services or heuristic artificial features. However, the widespread, global and cross-language nature of phishing and its rapid evolution under the constant combat of phishing and anti-phishing have led to a result that it is very difficult to detect phishing fast and accurately based on those features. This paper fully analyzes the semantic information in functional blocks of the webpage and extracts multi-scale semantic information from both URL and HTML, including URL, Title, Body Text and Invisible Text (HTML Tags). Bases on the extracted information and their different fusions of different depths, three end-to-end semantic information fusion models with different depths are designed to fully mine the complementary recognition ability of the multi-scale semantic information. This paper proposes multi-scale data-layer fusion (MDF) model, multi-scale feature-layer fusion (MFF) model and multi-scale in-depth fusion(MIF) model using Word Embedding, LSTM, CNN, etc. to carry out multi-scale semantic deep analysis. There is no need to extract any artificial features or retrieve information from third-party services, and the classification of phishing and non-phishing websites is effectively made through automated deep semantic analysis.

The phishing identification ability of the proposed models are first verified on the constructed unbalanced complex dataset. The experimental results show that: (1) the automatically extracted semantic information has effective phishing website identification ability and is complementary in phishing recognition; (2) the proposed models on the multi-scale semantic information all have good performance and the performance of MIF model is the best; under the premise

of FPR of 0.0047, the F1-Measure and AUC values reach 0.9830 and 0.9993 respectively; (3)through comparative experiments with current visual and text analysis methods, the proposed methods show good efficiency and robustness; (4) an active discovery experiment lasting for half a year was conducted and 3016 phishing websites were found, which further verifies the practicality of the model in real network environment.

In the future, the following work could be carried out: (1) to build a larger-scale phishing dataset, which will not only have more positive and negative samples, but also cover more phished brands and more language samples; (2) to construct a deep phishing detection model that fuses semantic and visual information to more effectively explore the counterfeit characteristic of phishing websites; (3) to recognize phishing websites in the actual Web scenarios, and optimize the model in practice.

CRediT authorship contribution statement

Dong-Jie Liu: Data curation, Writing – original draft. **Guang-Gang Geng:** Methodology, Writing – review & editing. **Xin-Chang Zhang:** Investigation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under grant No. 92067108.

References

Adebowale, M. A., Lwin, K. T., & Hossain, M. A. (2020). Intelligent phishing detection scheme using deep learning algorithms. *Journal of Enterprise Information Management*.

Al-Ahmadi, S. (2020). PDMLP: phishing detection using multilayer perceptron. *International Journal of Network Security & Its Applications (IJNSA) Vol*, *12*.

Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J.-P. (2020). An effective phishing detection model based on character level convolutional neural network from URL. *Electronics*, *9*(9), 1514.

APWG (2020). Phishing activity trends report, 2nd quarter 2020. https://docs.apwg.org/reports/apwg_trends_report_q2_2020.pdf.

APWG (2021). APWG phishing attack trends reports list. http://www.apwg.org/resources/apwg-reports/.

Ariyadasa, S., Fernando, S., & Fernando, S. (2020). Detecting phishing attacks using a combined model of LSTM and CNN. *International Journal of Advanced and Applied Sciences*, *7*(7), 56–67.

Bozkir, A. S., & Aydos, M. (2020). LogoSENSE: A companion HOG based logo detection scheme for phishing web page and E-mail brand recognition. *Computers & Security*, *95*, Article 101855.

Chiew, K. L., Choo, J. S.-F., Sze, S. N., & Yong, K. S. (2018). Leverage website favicon to detect phishing websites. *Security and Communication Networks*, *2018*.

Cohen, J. (2020). Phishing attacks increase 350 percent amid covid-19 quarantine. https://www.pcmag.com/news/phishing-attacks-increase-350-percent-amid-covid-19-quarantine.

Dilhara, B. (2021). Phishing URL detection: A novel hybrid approach using long short-term memory and gated recurrent units. *International Journal of Computer Applications*, *975*, 8887.

Fang, Y., Zhang, C., Huang, C., Liu, L., & Yang, Y. (2019). Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism. *IEEE Access*, *7*, 56329–56340.

Gorham, M. (2019). 2018 Internet crime report. https://pdf.ic3.gov/2018_IC3Report.pdf.

Hiransha, M., Unnithan, N. A., Vinayakumar, R., Soman, K., & Verma, A. (2018). Deep learning based phishing e-mail detection. In *Proc. 1st antiphishing shared pilot 4th acm int. workshop secur. privacy anal.*. Tempe, AZ, USA.

Huang, Y., Yang, Q., Qin, J., & Wen, W. (2019). Phishing URL detection via CNN and attention-based hierarchical RNN. In *2019 18th IEEE international conference on trust, security and privacy in computing and communications/13th IEEE international conference on big data science and engineering* (pp. 112–119). IEEE.

Johnson, R., & Zhang, T. (2017). Deep pyramid convolutional neural networks for text categorization. https://aclanthology.org/P17-1052.pdf.

Li, X., Geng, G., Yan, Z., Chen, Y., & Lee, X. (2016). Phishing detection based on newly registered domains. In *2016 IEEE international conference on big data* (pp. 3685–3692). IEEE.

Li, Y., Yang, Z., Chen, X., Yuan, H., & Liu, W. (2019). A stacking model using URL and HTML features for phishing webpage detection. *Future Generation Computer Systems*, *94*, 27–39.

Liu, D.-J., Geng, G.-G., Jin, X.-B., & Wang, W. (2021). An efficient multistage phishing website detection model based on the CASE feature framework: Aiming at the real web environment. *Computers & Security*, *110*, Article 102421.

Liu, D., & Lee, J.-H. (2020). CNN based malicious website detection by invalidating multiple web spams. *IEEE Access*, *8*, 97258–97266.

Liu, D., Lee, J.-H., Wang, W., & Wang, Y. (2019). Malicious websites detection via cnn based screenshot recognition. In *2019 international conference on intelligent computing and its emerging applications* (pp. 115–119). IEEE.

Microsoft (2019). Microsoft security intelligence report volume 24. https://info.microsoft.com/%20%20ww-landing-M365-SIR-v24-Report-eBook.html.

Opara, C., Wei, B., & Chen, Y. (2020). HTMLPhish: enabling phishing web page detection by applying deep learning techniques on HTML analysis. In *2020 international joint conference on neural networks* (pp. 1–8). IEEE.

Ozcan, A., Catal, C., Donmez, E., & Senturk, B. (2021). A hybrid DNN–LSTM model for detecting phishing URLs. *Neural Computing and Applications*, 1–17.

Peng, T., Harris, I., & Sawa, Y. (2018). Detecting phishing attacks using natural language processing and machine learning. In *2018 Ieee 12th international conference on semantic computing* (pp. 300–301). IEEE.

Ra, V., HBa, B. G., Ma, A. K., KPa, S., Poornachandran, P., & Verma, A. (2018). DeepAnti-PhishNet: Applying deep neural networks for phishing email detection. In *Proc. 1st antiphishing shared pilot 4th acm int. workshop secur. privacy anal.* (pp. 1–11). Tempe, AZ, USA.

Saha, I., Sarma, D., Chakma, R. J., Alam, M. N., Sultana, A., & Hossain, S. (2020). Phishing attacks detection using deep learning approach. In *2020 third international conference on smart systems and inventive technology* (pp. 1180–1185). IEEE.

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345–357.

Vazhayil, A., Vinayakumar, R., & Soman, K. (2018). Comparative study of the detection of malicious URLs using shallow and deep networks. In *2018 9th international conference on computing, communication and networking technologies* (pp. 1–6). IEEE.

Verma, R., & Hossain, N. (2013). Semantic feature selection for text with application to phishing email detection. In *International conference on information security and cryptology* (pp. 455–468). Springer.

Wang, W., Zhang, F., Luo, X., & Zhang, S. (2019). PDRCNN: precise phishing detection with recurrent convolutional neural networks. *Security and Communication Networks*, *2019*.

Wei, B., Hamad, R. A., Yang, L., He, X., Wang, H., Gao, B., & Woo, W. L. (2019). A deep-learning-driven light-weight phishing detection sensor. *Sensors*, *19*(19), 4258.

Wei, W., Ke, Q., Nowak, J., Korytkowski, M., Scherer, R., & Woźniak, M. (2020). Accurate and fast URL phishing detector: a convolutional neural network approach. *Computer Networks*, *178*, Article 107275.

Wielgosz, M., Skoczeń, A., & Mertik, M. (2017). Using LSTM recurrent neural networks for monitoring the LHC superconducting magnets. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *867*, 40–50.

Yerima, S. Y., & Alzaylaee, M. K. (2020). High accuracy phishing detection based on convolutional neural networks. In *2020 3rd international conference on computer applications & information security* (pp. 1–6). IEEE.

Zhang, Q., Bu, Y., Chen, B., Zhang, S., & Lu, X. (2021). Research on phishing webpage detection technology based on CNN-BiLSTM algorithm. *Journal of Physics: Conference Series*, *1738*(1), Article 012131.

Zhang, X., Zeng, Y., Jin, X.-B., Yan, Z.-W., & Geng, G.-G. (2017). Boosting the phishing detection performance by semantic analysis. In *2017 Ieee international conference on big data* (pp. 1063–1070). IEEE.