Multidimensional Features Driven Phishing Detection based on Deep Learning

P Vigneshwaran[1], A Soumith Roy[2], M Lekha Chowdary[3], D Md Nasirulla[4], B Sunal Sathvik[5]

[1]Professor, Dept. of Computer Science Engineering, Jain University, Bangalore, India
[2-5]UG Scholar, Dept. of Computer Science Engineering, Jain University, Bangalore, India

---

Abstract - *As a criminal crime involving the use of electronic means to steal confidential information from users, phishing is actually an important threat to the network, and phishing losses are increasingly rising. Feature engineering is vital in phishing website detection solutions, but the accuracy of detection is critically dependent on prior knowledge of the features. Moreover, although features extracted from different dimensions are more comprehensive, the drawback is that the extraction of these features requires a considerable amount of time. In order to address these limitations, we propose a multidimensional feature phishing detection approach supported by a quick detection method using machine learning. In the first step, the character sequence features of the given URL are extracted and used for quick classification by algorithms, and this step does not require assistance from third party vendors or any specific knowledge of phishing. In the second stage, we combine statistical URL features, webpage code features, webpage text features, and therefore easily classify the results of deep learning into multidimensional features. The method could reduce the detection time for setting a threshold. Testing on a dataset containing millions of phishing URLs and legitimate URLs, the accuracy reaches maximum, and the false positive rate is minimum. By reasonably adjusting the edge, the experimental results show that the detection efficiency is improved.*

Key Words*: *Security, Phishing, Feature Extraction

1. INTRODUCTION

The Internet has become an important infrastructure that brings great convenience to human society. However, the web is additionally characterized by some inevitable security problems, like phishing, malicious software, and privacy disclosure, which have already brought serious threats to the economy of users. The APWG (Anti-Phishing Working Group) defines phishing as a criminal mechanism employing both social engineering and technical subterfuge to steal identity data and financial account credentials of consumers. Phishing may be a very fashionable method utilized in network attacks and results in privacy leaks, fraud and property damage. consistent with statistics from the Kaspersky Lab, in 2017, 29.4% of user computers were subjected to a minimum of one Malware-class web attack over the year and 199 455 606 unique URLs were recognized as malicious by web antivirus components. Additionally, the share of monetary phishing increased from 47.5% to almost 54% of all phishing detections in 2017. Phishing has become one among the most important security threats within the Internet. The spread of phishing is not any longer limited to traditional modalities like e-mail, SMS, and pop-ups. Though the prosperity of the mobile Internet and social networks have brought convenience to users, they need also been employed to spread phishing, like QR code phishing, spear phishing and spoof mobile applications, etc. additionally, many cunning phishing attacks are hosted on websites that have HTTPS and SSL certificates because many users think that HTPPS websites are likely legitimate. Phishing presents a diversified development trend, which poses new detection challenges. While phishers are pernicious and conceal, security experts and researchers have dedicated many efforts in terms of phishing website detection.

The character sequence of the URL is natural, automatically generated feature that avoids the subjectivity of artificially selected features. In addition, it does not require third-party assistance and any prior knowledge about phishing. However, in the process of character sequencing, the difficulty is to effectively extract association and semantic information.

2. RELATED WORKS

Phishers usually distorts the hostname part and the path part from the URL of the target web page to produce the phishing URL, and thus features can be extracted based on the URL statistical rules or simply based on the URL strings. Studies have proposed many interesting features of various styles of phishing websites from multiple perspectives.

2.1 Anti-Phishing Solutions

Due to rise in phishing attacks, various solutions are endeavored that can provide a solution to this problem. So as to outline a framework that can give assurance from phishing attacks, there exist various ways. Numerous methodologies are being used in the frameworks proposed before. First let us know about these approaches. Anti-Phishing solutions can be classified into following categories:-

i. ***Heuristic Based Approach-*** *This technique makes use of heuristics to classify URLs. Heuristics are the features that are considered to check a website. Here they heuristics like IP address in domain part, '@' symbol in URL, right click disabled, pop-up windows for passwords etc. to derive rules on these heuristics and decide a threshold for it.*

ii. ***Content Based Approach-*** *The comparison of two web pages is done based on the similar contents on the web page. This technique makes use of Term Frequency/Inverse document Frequency (TF-IDF). TF-IDF compares the terms in the original website to the phishy one. Another approach is to capture the screen shots of a website and then process it to compare. This information retrieved from screen shots after processing can be given to a search engine to acquire its page rank and check the legitimacy of the website by comparing the content on it. Website logo can be used is this method to analyze the webpage using the Google image database.*

iii. ***Blacklist Based Approach****- A blacklist includes list of websites that are declared as spam. Such blacklists are maintained by organizations like Google. Spam URLs are added to this blacklist. The disadvantage of this approach is that a newly created phishing URL may not be present in the blacklist. Thus, such URLs will be left undetected. URLs present in the blacklist have denied access. This means a user cannot surf this webpage.*

iv. ***Machine Learning Approach-*** *In this technique, features are extracted and they are classified using the machine learning techniques. The classification accuracy depends on the algorithm chosen. We can see that more than one ML methods are experimented on the same dataset to find the best suitable one. Such comparisons of algorithms can help to give better accuracy in experimentation.*

v. ***Hybrid Approach-*** *In this technique different techniquesare combined to detect if a website is fake or real. For example heuristics and blacklisting of URL can be combined to form a better system. Another hybrid model is combination of Machine Learning algorithm. In such model, the dataset is trained using first algorithm and then the result is again passed to the second algorithm for training.*

Certain tests like comparisons of DNS of actual and visual links were performed by the algorithms and also checks dotted decimal of IP address, checks coded links and pattern matching. The system contains some of the drawbacks that it produces false positive results if any legitimate and trust worthy site has IP address instead of domain name and if the user does not visit the legitimate (original) site then it considers some suspicious sites as normal. Hence the assessment results in false negative conclusions.

The Statistics of suspicious URL‟s have been analyzed by many researchers. The garera uses to classify phishing website URL‟s the work by garera uses logistic regression over hand selected features. It includes features like flag keywords in URL, Google page rank-based features web page quality guidelines. A direct comparison with our approach is difficult without access to same URL‟s and features. The analysis features include IP address who is in the records containing time and date information about geographic, registration and also lexical features of URL such as length, width, distribution of characters and predefined popular names. Joby James, Sandhya L, Ciza Thomas proposed that phishers use spoofed e-mail, phishing software to steal personal information and financial account details such as usernames and passwords. Suggested methods for detecting phishing websites by analyzing various features of benign and phishing URLs by Machine learning techniques. They found methods used for detection of phishing websites based on lexical features, host properties and page importance properties. They considered various data mining algorithms for evaluation of the features in order to get a better understanding of the structure of URLs that spread phishing. The fine-tuned parameters are useful in selecting the apt machine learning algorithm for separating the phishing sites from benign sites.Ram B. Basnet, Andrew H. Sung, and Quingzhong Liu said phishing a hotbed of multibillion dollar underground economy has become an important cybersecurity problem. Therefore, learning machine based approaches have been implemented for phishing detection. Many existing techniques in phishing website detection seem to include as many features as can be conceived, while identifying a relevant and representative subset of features to construct an accurate classifier remains an interesting issue in this particular application of machine learning. Evaluating using correlation-based and wrapper feature selection techniques using real world phishing data sets with 177 initial features.Miss Sneha Mande, Prof D.S.Thosar proposed phishing site detection is truly an unpredictable and element issue including numerous components and criteria that are not stable. On account of the last and in addition ambiguities in arranging sites because of the intelligent procedures programmers are utilizing, some keen proactive strategies can be helpful and powerful tools can be utilized, for example, fuzzy, , neural system and data mining methods can be a successful mechanism in distinguishing phishing sites.Weiping Wang, Feng Zhang, Xi Luo and Shigeng Zhang proposed how CNN used to automatically judge which characters play key roles in phishing detection, capture the key components of the

URL, and compress the extracted features into a fixed length vector space. By combining the two types of networks, CNN achieves better performance than just using either one of them. We built a dataset containing nearly 500,000 URLs which are obtained through Alexa and PhishTank. Experimental results show that PDRCNN achieves a good detection accuracy.

3. PROPOSED SYSTEM

The problem with existing system is machine learning methods are widely used in phishing website detection. The reason is that malicious URLs or phishing web pages have some characteristics that can be distinguished from legitimate websites, and machine learning can be effective in this regard for processing. Current mainstream machine learning methods of phishing website detection extract statistical features from the URL and the host or extract relevant features of the webpage, such as the layout, CSS, text, and then classify these features. The proposed methodology imports data-set of phishing and legitimate URLs from the database and then the imported data is pre-processed. Identifying phishing website is performed based on following category of URL features: domain, address, abnormal and HTML, JavaScript features. The URL features are extracted with processed data and values for each URL attribute. The analysis of URL is performed using machine learning technique which computes the range value and the threshold value for URL attributes. Then it is differentiated as phishing and legitimate URL. The attribute values are computed using feature extraction of phishing websites and it is used to obtain the range value and threshold value. The value for each phishing attribute is ranging from {-1, 0, 1} these values are defined as low, medium and high. The classification of phishing and legitimate website is based on the values of attributes extracted using different types of phishing categories and a machine learning approach. URL Feature Extraction- The attribute value of the URL is computed using corresponding set of attribute values {-1, 0, 1}. The attributes URL_of_Anchor tag and Prefix_Suffix also have inter linked value and that needs to be computed for finding range and threshold value. Finding Attribute Values -The phishing features are extracted for each URL to find whether the website is phishing or legitimate. The URL_of_Anchor tag attribute is selected to identify the overlap values. The overlap value is the summation of selected attribute value which is combined with other attributes. Classification- The units in this layer have connection to every other unit in the succeeding layer.

That's why this layer is called as fully-connected layer. It contains sigmoid non-linear activation function, which gives values 0 for legitimate and 1 for phishing. The prediction loss for both the sub tasks is estimated using binary cross entropy, as given below

$$loss(p,e) = -\frac{1}{N}\sum_{i=1}^{N}[e_i \cdot log(p_i) + (1-e_i) \cdot log(1-p_i)]$$

We propose a multidimensional feature phishing detection approach supported a quick detection method by using deep learning. We build a real dataset by crawling a total of 1021758 phishing URLs as positive samples from phishtank.com, and a total of 989021 legitimate URLs as negative samples from dmoztools.net.

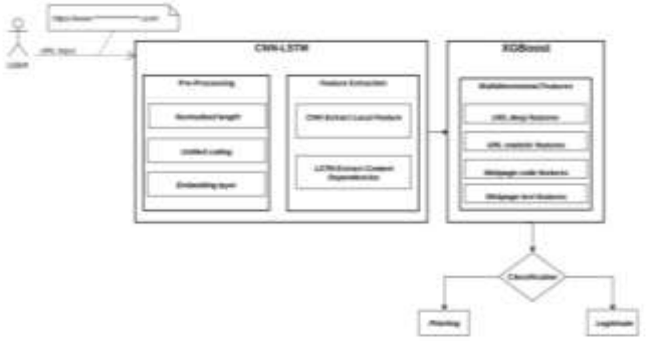

Fig 3.1: System architecture

- The process of phishing website detection using MFPD is explained, and an extensive experiment on the dataset we built is conducted. The results show that our proposed approach exhibits good performance in terms of accuracy, false positive rate, and speed.
- Within the first step, the given URL character sequence features are extracted and used for quick classification by machine learning algorithm, and this step doesn't require third party assistance or any prior knowledge about phishing.Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data.
- Within the second step, we combine multidimensional features like URL statistical features, webpage code features, webpage text features, and therefore the quick classification results of deep learning into multidimensional features. We use Logistic regression to train the text vector and generate the probability that the text belongs from the phishing website, then the probability is employed to represent the webpage text features. After extracting features from different aspects, these features should be fused. In this project, the output of CNN algorithm is used as the deep URL features, and it is combined with the URL statistical features, webpage code

features and webpage text features to make up multidimensional features, which are classified by a machine learning approach.

- Though the multidimensional feature algorithm has greater accuracy than the CNN-LSTM, the acquisition of WHOIS information and Alexa ranking from the URL, and the extraction of webpage code features and webpage text features take a certain amount of time, which cannot meet the needs of real-time detection. We improve the classification output of the softmax layer in the CNN-LSTM algorithm.
- An XGboost algorithm is proposed. By revising the output judgment conditions of the softmax classifier in the deep learning process and setting a threshold, the detection time can be reduced.

4. RESULTS

The following two tables give results of unit testing.

TABLE-4.1: TEST CASE-1

Test Case#	UTC01
Test Name	File import format
Test Description	To test whether an excel file with comma delimited format is accepted or not
Input	A comma separated excel file with valid dataset
Expected Output	The file should be read by the program and few lines of the file should be displayed
Actual Output	The file is read and contents are displayed accordingly
Test Result	Success

TABLE-4.2: TEST CASE-2

Test Case#	UTC02
Test Name	File import format
Test Description	To test whether an excel file with comma delimited format is accepted or not
Input	A text file with valid dataset
Expected Output	It Should show the alert Message Select Only CSV file
Actual Output	Shown alert message
Test Result	Success

The following give the simulation results of the model:

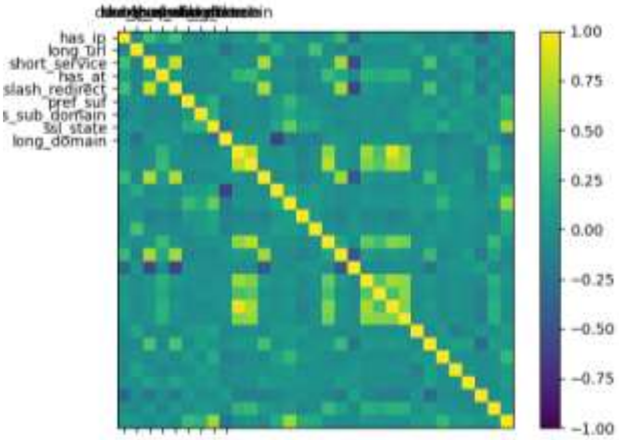

Fig 4.1: Correlation Matrix

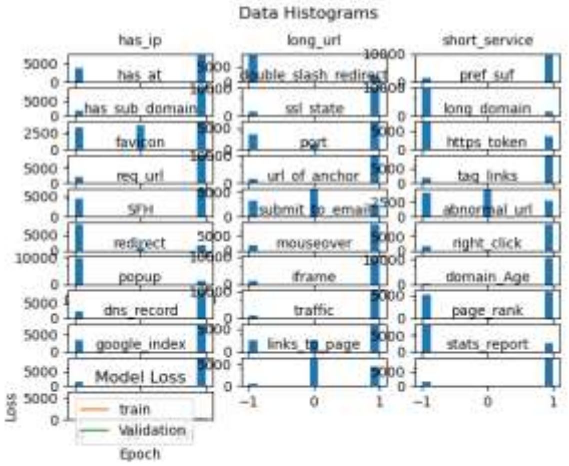

Fig 4.2: CNN Loss

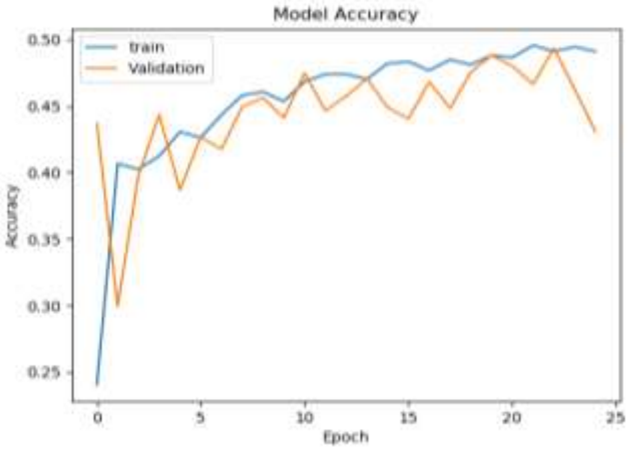

Fig 4.3: CNN Accuracy

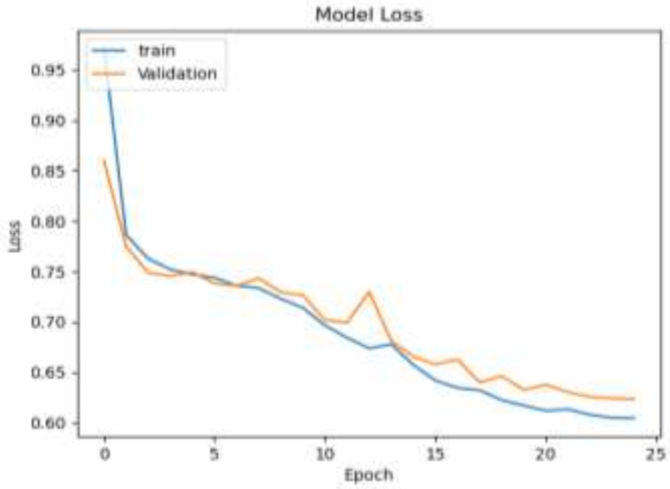

Fig 4.4: CNN-LSTM Loss

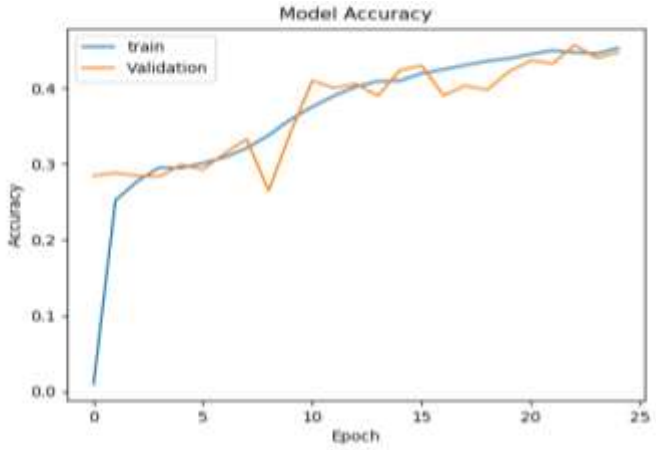

Fig 4.5: CNN-LSTM Accuracy

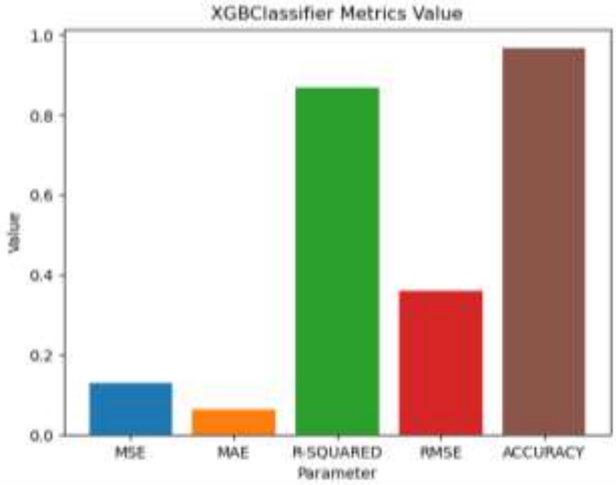

Fig 4.6: XGBoost Classifier Metrics values

5. CONCLUSIONS AND FUTURE WORK

It is well known that a good phishing website detection approach should have good real-time performance while ensuring good accuracy and a low false positive rate. Our proposed MFPD approach is consistent with this idea. Under the control of a dynamic category decision algorithm, the URL character sequence without phishing prior knowledge ensures the detection speed, and the multidimensional feature detection ensures the detection accuracy. We conduct a series of experiments on a dataset containing millions of phishing and legitimate URLs. From the results, we found that the MFPD approach is effective with high accuracy, low false positive rate and high detection speed.

A future development of our approach will consider applying deep learning to feature extraction of webpage code and web page text.

REFERENCES

[1] Yongjie Huang, Qiping Yang, Jinghui Qin, WushaoWen "Phishing URL Detection via CNN and Attention-Based Hierarchical RNN" 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing AndCommunications/13th IEEE International Conference On Big Data Science And Engineering 2324-9013/19/$31.00 ©2019 IEEE.

[2] Amani Alswailem, BashayrAlabdullah, Norah Alrumayh, Dr.AramAlsedrani"Detecting Phishing Websites Using Machine"978-1-7281-0108-8/19/$31.00 2019 IEEE.

[3] G. JaspherWillsieKathrine,Paradise Mercy Praise,A. Amrutha Rose "Variants of phishing attacks and their detection techniques" Proceedings of the Third International Conference on Trends in Electronics and

Informatics (ICOEI 2019) IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8.

[4] Mohammad Mehdi Yadollahi, FarzanehShoeleh, ElhamSerkani, AfsanehMadani, Hossein Gharaee "An Adaptive Machine Learning Based Approach for Phishing Detection Using Hybrid Features"2019 5th International Conference on Web Research (ICWR)978-1-7281-1431-6/19/$31.00 ©2019 IEEE.

[5] bubekir Buber, ÖnderDemir,OzgurKoraySahingoz "Feature Selections for the Machine Learning based Detection of Phishing Websites" 978-1-5386-1880-6/17/ ©2017 IEEE

[6] Karl R. Weiss,Taghi M. Khoshgoftaar" Detection of Phishing Webpages using Heterogeneous Transfer Learning" 2017 IEEE 3rd International Conference on Collaboration and Internet Computing 0-7695-6303-1/17/31.00 ©2017 IEEE DOI 10.1109/CIC.2017.00034.

[7] Priyanka Singh, Yogendra P.S. Maravi, Sanjeev Sharma Phishing "Websites Detection through Supervised Learning Networks" 978-1-4799-7623-2/15/$31.00_c 2015 IEEE.

[8] PurviPujara, M. B.Chaudhari"Phishing Website Detection using Machine Learning" International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2018 IJSRCSEIT | Volume 3 | Issue 7 | ISSN : 2456-3307.

[9] Ram B. Basnet, Andrew H. Sung, and QuingzhongLiu "Feature Selection for Improved Phishing Detection" IEA/AIE 2012, LNAI 7345, pp. 252–261, 2012.© Springer-Verlag Berlin Heidelberg 2012.

[10] Waleed Al "Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 9, 2017.

[11] Miss Sneha Mande, Prof.D.S.Thosar "Detection of Phishing Web Sites Based On Extreme Machine Learning"Vol-4 Issue-6 2018 IJARIIE-ISSN(O)-2395-4396.

[12] Mazharul Islam, Nihad Karim Chowdhury "Phishing Websites Detection Using Machine Learning Based Classification Techniques".

[13] Arun Kulkarni, Leonard L. Brown "Phishing Websites Detection using Machine Learning" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019.

[14] HemaliSampat, Manisha Saharkar, Ajay Pandey ,Hezal Lopes " Detection of Phishing Website Using Machine Learning International Research Journal of Engineering and Technology" (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 03 Mar-2018.

[15] Moitrayee Chatterjee and Akbar SiamiNamin"Deep Reinforcement Learning for Detecting Malicious Websites" arXiv:1905.09207v1 [cs.LG] 22 May 2019.

[16] Sandeep Kumar Satapathy, Shruti Mishra, Pradeep Kumar "Classification of Features for detecting Phishing Web Sites based on Machine Learning Techniques"International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8S2, June 2019.

[17] Alyssa Anne Ubing, SyukrinaKamiliaBinti ,Azween"Phishing Website Detection: An Improved Accuracy through Feature Selection and Ensemble Learning"(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 1, 2019.

[18] Weiping Wang, Feng Zhang, Xi Luo and Shigeng Zhang "Precise Phishing Detection with Recurrent Convolutional Neural Networks" HindawiSecurity and Communication Networks Volume 2019, Article ID 2595794.

[19] AltyebAltaher"Phishing Websites Classification using Hybrid SVM and KNN Approach"(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 6, 2017.

[20] Santhana Lakshmi V, VijayaMSb"Efficient prediction of phishing websites using supervised learning algorithms" © 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of ICCTSD 2011.