Hybrid Phishing Detecting with Recommendation Decision Trees

Duncan Eric O. Ogonji, Cheruiyot Wilson, Waweru Mwangi

***Abstract:** Phishing is performed by trying to trick the victim into accessing any computing information that looks original and then instructing them to send important data to unrestricted/unwanted private resources. For prevention, it is essential to develop a phishing detection system. Recent phishing detection systems are based on data mining and machine learning techniques. Most of the related work literature requires the collection of previous phishing attack logs, analyzing them creating a list of such activities, and blocking traffic from such sources. However, this is a cumbersome task because the data size is very large, continues changing, and is dynamic in nature. [1]. Instead of using a single algorithm approach, it would be better to use a hybrid approach. A hybrid approach would be better at mitigating phishing attacks because the classification of different formats of data is handled; whether the intruder wants to use images or textural input to gain into another user system for phishing. Hybrid recommendation decision trees enhance any of the machine learning and deep learning algorithms' performance. The decision path of the model followed a series of if/else/then statements that connect the predicted class from the root of the tree through the branches of the tree to detect true positives and false negatives of phishing attempts. 10 decision trees were considered and used the features to train the recommendation decision regression model. The developed hybrid recommendation decision tree approach provided an overall true positive rate of the model of 92.28 % and a false negative rate is 7.4%.*

***Keywords:** Phishing, Decision Tree, Detection, Hybrid, Attack*

I. INTRODUCTION

The most challenging risks in online systems today are online scams and phishing attacks. Phishing attacks have become one of the primary hacking methods used against organizations [2]. It is a type of attack in which criminals use fake emails and bogus websites to trap people into giving up delicate information. The end goal for phishing attacks is to gather sensitive and important information such as credit card number or email address and password [3].

In a phishing attack, the attacker points or directs the victim to fake pages using social engineering techniques. Phishing is performed by trying to trick the victim into accessing any computing information that looks original, then instructing them to send important data to unrestricted or any unwanted private resource that i not authorized[4]. Phishing is a method for taking on the appearance of a disclosed substance to hoodwink an injured individual into opening an email, text, or instant message [5]. The beneficiary is then fooled into clicking a noxious connection which can prompt the establishment of malware, the solidifying of the framework as a major aspect of a ransomware attack, or the noteworthy of delicate data [6]. The objective of a phishing attack is to trick receivers into divulging sensitive information such as bank account numbers, passwords, and credit card details [7].

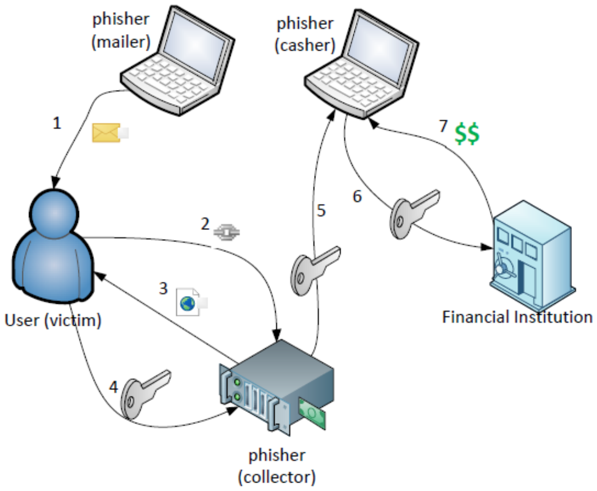

Fig. 1.Phishing Information Flow

Decision Trees are considered to be one of the most popular approaches for representing classifiers [8]. Researchers from various disciplines such as statistics, machine learning, pattern recognition, and data mining have dealt with the issue of growing a decision tree from available data[9]. Recommendation decision trees are considered to be one of the most popular approaches for representing classifiers. Researchers from various disciplines such as statistics, machine learning, pattern recognition, and data mining have dealt with the issue of growing a decision tree from available data [7]. The key point of constructing the decision tree is to determine the best attribute to split the considered node. [10] define a structure of a recommendation decision tree is structured as follows:

Manuscript received on 22 June 2024 | Revised Manuscript received on 28 June 2024 | Manuscript Accepted on 15 July 2024 | Manuscript published on 30 July 2024.

*Correspondence Author(s)

Duncan Eric O. Ogonji*, School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology Nairobi, Kenya. E-mail: ogonjimsc@gmail.com, ORCID ID: 0009-0005-1700-642X

Cheruiyot Wilson, School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology Nairobi, Kenya. E-mail: wilchery68@gmail.com

Prof. Waweru Mwangi, School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology Nairobi, Kenya. E-mail: wawerumwangi@jkuat.ac.ke

i. **Root Node:** The root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
ii. **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
iii. **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
iv. **Branch/Sub Tree:** A tree formed by splitting the tree.
v. **Pruning:** Pruning is the process of removing the unwanted branches from the tree.
vi. **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

II. METHODOLOGY

A. Introduction

The purpose of this section of the thesis implement a hybrid model for detecting phishing attacks using recommendation decision trees.

B. Problem Statement

Development of an effective detection system while minimizing false positives and negatives remains a challenge while detecting phishing attacks[1]. Instead of using a single algorithm approach, it would be better to use a hybrid approach. A hybrid approach would be better at mitigating phishing attacks because the classification of different formats of data is handled; whether the intruder wants to use images or textural input to gain into another user system for phishing [11]. This research developed a hybrid model approach for phishing detection model using recommendation decision trees. Recommendation decision trees enhance any of the machine learning and deep learning algorithms' performance.

C. Research Objective.

The main objective of this research was to implement a hybrid model for detecting phishing attacks using recommendation decision trees.

D. Related work

[12] implemented detection of phishing URLs using Bayes net and Naïve Bayes algorithm and evaluation of risk regarding phishing URLs is done with the help of attributable risk. A training dataset of 1800 URLs (containing 1080 legitimate and 720 phished URLs) was made to train the algorithms. A testing dataset of 720 URLs (containing 288 legitimate and 432 phished URLs) was used for making predictions using the DAG model classifier which is generated after the training of Bayes Net and Naïve Bayes Algorithm.

[2] presented a new move for detecting phishing web pages in real-time as they are visited by a browser. It relies on modelling inherent phisher limits stemming from the constraints they face while building a webpage. Consequently, the implementation of this approach, Off-the-Hook, exhibits several notable properties including high accuracy, brand independence good language-independence, speed of decision, resilience to dynamic phishes, and flexibility to evolution in phishing techniques.
This research by [13][18][19][20] implemented a hybrid anti-phishing approach using some of the well-known phishing detection factors like the MAC address of web pages. [14] Research presented various phishing techniques and their effects on our daily life additionally finding some acceptable and/ or adoptable detection and prevention techniques by which a system automatically detects a phishing web URL using data mining techniques.

E. The Hybrid Recommendation Decision Tree Model

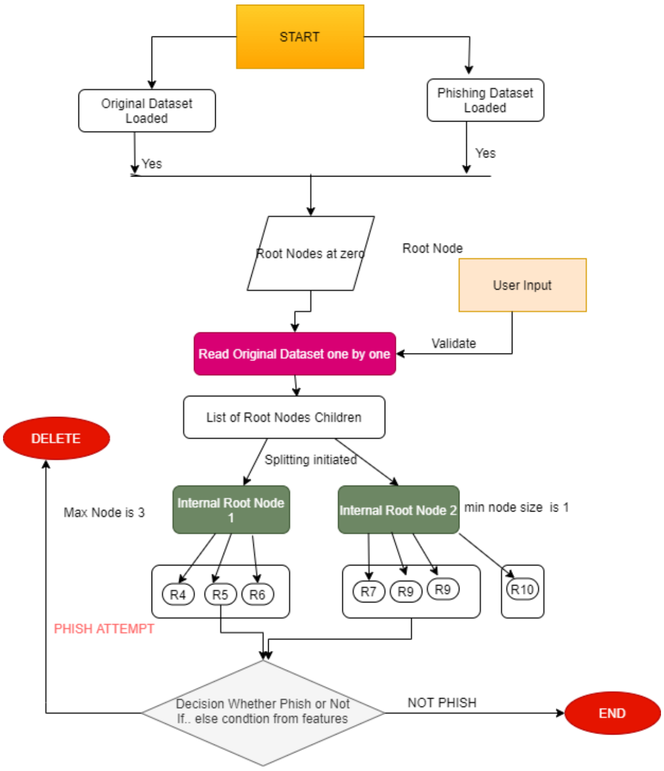

Fig. 2.The Implemented Hybrid Recommendation Decision Tree Model

The parameters for this R-tree from Fig 2 are m==1 and M==3. Internal nodes only hold bounding boxes, while leaves hold the actual points (or, in general, k-dimensional entries) detecting phishing malicious activities and processes by traversing the tree, pruning branches that can't contain a point, or, for NN search, are certainly further away than the current minimum distance. The first was to find the leaf that should host the new point where a list of root node children was created. There are three possible cases: as the root point lies exactly within one of the leaves rectangles internal root nodes, then just add the root point of the tree to internal root nodes and move to the next step.

The root node lies within the overlapping region between two or more leaves' bounding rectangles. For example, referring to Fig 2, it starts overlapping lie in the intersection of R7 and R10. Several points are added to the leaf's rectangle internal nodes R and how many points it contains afterward.

The internal nodes 1 are removed from its parent where a list of root node children is created. R4, R5, and R6 were then added to internal nodes 1.

Then now internal root node 2 has more than internal node 1 children, it is split and repeats this step recursively. To complete the insertion algorithm outlined here, we need to provide a few heuristics to R9 to break ties for overlapping rectangles and to R8 to choose the closest rectangle that could contain phishing attempts. This choice, together with the heuristic R5 R6 for choosing the insertion subtree, deter R7 R8 R2 R9 mines the behavior and shape, not to mention performance) of the R-tree.

F. Dataset Population and Sample

The dataset was from a freemium Canadian Centre for Cyber Security https://www.unb.ca/cic/datasets/andmal2020.html. A purposive sample was used and the final sample was selected based on the knowledge about the study and population. The participants were chosen based on the purpose of the sample, hence the name. From the repository of 149 malware families and 170,134 population samples where purposive sampling from the population was used to get a final sample for the study. The research narrowed down to scareware and riskware malware family variables with a final sample of 4 and 4 variables respectively from the massive dataset. The four variables included hyperlinks to website URLs, Email clicks, user account registration, and user account activation. The basic information of the computer system in which the experiments were running was: windows 7 professional Operating system, Intel Core i3 of 1.80GHZ CPU, RAM of 16 GB. Google Colab was used for its powerful GPU.

III. RESULTS AND DISCUSSION

A. Introduction

This chapter presents the detailed results, analysis, and comparisons from the simulations of the hybrid recommendation decision tree model for detecting phishing attacks.

B. Phishing Detection Results

The most powerful and easy-to-use Python library for developing and evaluating decision tree classifier used was numpy on the dataset which showed the following results: The confusion matrix plot indicates accuracy of 4 correct prediction of phishing attacks. The recommendation decision tree classifier predicts the missing values based on the created confusion matrix. Earlier the recommendation decision tree classifier model that was saved needs to be loaded first after that, it is re-evaluated on the current test dataset. Based on the generated recommendation decision tree classifier model after the evaluation of the trained dataset the testing dataset then a prediction is made and a confusion matrix is created. If the value of the result node which is a parent node is -1 then the website is legitimate and if the value of the result node is 1 then the website is phished.

Table 1: Confusion Matrix Results

Phishing Output Class	Output 1	Output 2	Output 3	Output 4	
Target 1	Yes	0	0	0	
Target 2	0	No	0	0	
Target 3	0	0	Yes	0	
Target 4	0	0	0	Yes	

C. Results Comparison

The developed hybrid recommendation decision tree approach tested phishing datasets from a freemium Canadian Centre for Cyber Security to detect phishing webpages based on hyperlinks information. The overall true positive rate of the model was 92.28 % and the false negative rate was 7.4%. The comparison results in the above table II from the experiments proved that the implemented recommendation decision tree model approach for detecting phishing attacks has the lowest percentage of false positive rate which resulted in a higher accuracy of 88.67%. When the false negative rate is low, it gives the impression of an efficient model approach to detecting phishing attacks. The true Negative (TN) rate measures the rate of correctly detected legitimate sites about all existing legitimate sites [15]. False Negative (FN) rate measures the rate at which phishing websites are incorrectly identified as legitimate about all existing phishing websites. Accuracy (A) measures the rate of phishing and legitimate websites that are identified correctly concerning all the websites [15].

TABLE II: Comparison Results

Technique Used	*True Positive*	*False Negative*	*Accuracy*
Fuzzy Logic Classifiers	82.3%	9.8%	80.5%
AdaBoost with Random Tree	76.9%	8.6%	785%
Implemented Approach	92.8%	7.4%	88.67%

IV. CONCLUSION

This research developed a hybrid model approach for phishing detection model using recommendation decision trees. The developed hybrid recommendation decision tree approach provided an overall true positive rate of the model was 92.28 % and a false negative rate is 7.4%[16][17]. The recommendation decision tree starts with the root node and follows the branch creation logic as entropy to create subsequent nodes that reach the terminal node, which is also considered the leaf node. The entire path from the start of the root node to the leaf node is considered as a rule that detects any true positive variables that may be triggered as phishing attempts. The hybrid decision tree framework is to decide if the best attribute determined for the current set of data elements in the node is also the best according to the whole stream. When the false negative rate is low, it gives the impression of an efficient model approach to detecting phishing attacks. The limitation of this approach is detecting phishing attacks with time series data.

A. Areas of Future Research

Future work should extend to time series dynamic data so that phishers can take advantage of the real-time execution of data. The hybrid recommendation decision tree can be enhanced with other unsupervised machine-learning techniques to detect phishing attacks.

DECLARATION STATEMENT

Funding	No, We did not receive any financial support for this article.
Conflicts of Interest	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material	Not relevant.
Authors Contributions	All authors have equal participation in this article.

REFERENCES

1. P. S. Gayke, S. Kardile, N. Dongare, S. Pathare, and P. Sakat, “Spammer Detection and Fake User Identification in E-Commerce Site,” vol. 9, no. 7, pp. 22–25, 2021. https://doi.org/10.26438/ijcse/v9i7.2225
2. P. Priyadevi and V. Lalithadevi, “An Efficient and Usable Client-Side Phishing Detection Application,” no. 2, 2018.
3. C. Natalino, A. Udalcovs, L. Wosinska, O. Ozolins, and M. Furdek, “Spectrum Anomaly Detection for Optical Network Monitoring Using Deep Unsupervised Learning,” IEEE Commun. Lett., vol. 25, no. 5, pp. 1583–1586, 2021, doi: 10.1109/LCOMM.2021.3055064. https://doi.org/10.1109/LCOMM.2021.3055064
4. V. R. Reddy, C. V. M. Reddy, and M. Ebenezar, “A Study on Anti-Phishing Techniques,” no. 1, pp. 30–36, 2016.
5. H. K. N. G, G. Pooventhiran, and K. R. D, “Landslide Type Prediction using Random Forest Classifier,” no. 2, pp. 7–11, 2020.
6. S. Khatana and A. Jain, “Malware Detection Using the Behavioral Analysis of the Web-based Applications and User,” Int. J. Comput. Sci. Eng., vol. 7, no. 5, pp. 1026–1031, 2019, doi: 10.26438/ijcse/v7i5.10261031. https://doi.org/10.26438/ijcse/v7i5.10261031
7. S. Bansal and A. Singh, “Machine learning in the prediction, determination and further study of different cyber-attacks,” no. 10, 2019. https://doi.org/10.26438/ijcse/v7i10.2736
8. P. Re-identification, H. Xie, Y. Zhou, and Q. Liu, “Deep Learning Feature Representation Applied to Cross Dataset,” no. 2, pp. 1–11, 2018. https://doi.org/10.26438/ijcse/v6i2.111
9. R. V Kotawadekar, A. S. Kamble, and S. A. Surve, “Automatic Detection of Fake Profiles in Online Social Networks,” no. 7, 2019. https://doi.org/10.26438/ijcse/v7i7.4045
10. R. R. Biradar and G. S. Nagaraja, “Anomalous Traffic Detection System for Enterprise using Elastic Stack with Machine Learning,” vol. 9, no. 6, 2021. https://doi.org/10.26438/ijcse/v9i6.1318
11. A. Kulkarni, “Credit Card Fraud Detection Using Random Forest and Local Outlier Factor,” Int. J. Res. Appl. Sci. Eng. Technol., vol. 7, no. 4, pp. 1170–1175, 2019, doi: 10.22214/ijraset.2019.4209. https://doi.org/10.22214/ijraset.2019.4209
12. P. Raj and M. Mittal, “Detection of Phishing URLs using Bayes Net and Naïve Bayes and evaluating the risk assessment using Attributable Risk,” no. 5, 2018. https://doi.org/10.26438/ijcse/v6i5.750755
13. P. Saklecha and J. Raikwar, “Prevention of Phishing Attack using Hybrid Blacklist Recommendation Algorithm,” no. 6, pp. 188–191, 2018. https://doi.org/10.26438/ijcse/v6i6.188191
14. N. S. Reddy and V. K. M, “Review Paper Detection of E-Banking Phishing Websites,” no. 14, pp. 49–52, 2019. https://doi.org/10.26438/ijcse/v9i7.5359
15. H. Agrawal and R. R. Singh, “An Ensemble Approach for Detecting Phishing Attacks,” vol. 9, no. 7, 2021. https://doi.org/10.35940/ijitee.H6540.069820
16. Mabuni, D. (2020). A Novel Impurity Measuring Technique for Decision Tree Learning in Machine Learning. In International Journal of Innovative Technology and Exploring Engineering (Vol. 9, Issue 8, pp. 506–512). https://doi.org/10.35940/ijitee.h6540.069820
17. Panhalkar, A. R., & Doye, D. D. (2020). Improving Decision Tree Forest using Preprocessed Data. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 6, pp. 4457–4460). https://doi.org/10.35940/ijrte.f8136.038620
18. Assegie, T. A. (2021). K-Nearest Neighbor Based URL Identification Model for Phishing Attack Detection. In Indian Journal of Artificial Intelligence and Neural Networking (Vol. 1, Issue 2, pp. 18–21). https://doi.org/10.54105/ijainn.b1019.041221
19. Joshma K J, & Sankar P, V. (2024). Phishing Website Detection. In Indian Journal of Data Mining (Vol. 4, Issue 1, pp. 38–41). https://doi.org/10.54105/ijdm.a1642.04010524
20. Dawood, M., Ibrahim, O. B., & Abu-Ulbeh, W. A. R. A. (2019). Enrich Awareness of Users to Detect Phishing Websites. In International Journal of Engineering and Advanced Technology (Vol. 8, Issue 6s3, pp. 648–650). https://doi.org/10.35940/ijeat.f1119.0986s319

AUTHORS PROFILE

D.E. Ogonji Bsc.IT, MBA, Strategic Management, MSc. Computer Systems (on-going) is a Business focused & result-oriented Technology leader with over 15 years of hands-on experience in leading IT strategy, digital transformation and collaboration with business leaders to deliver transformative digital strategies that improve profitability, drive Growth and efficiency. Skilled in design and implementation of IT services aligned with the business operating model to achieve business outcomes and build technical capability. Proven track record of success in the Financial Services, Public & Private Sectors with demonstrated relevant industry certifications on Cyber-Security, Project Management. Strong background in strategic planning, team development, team leadership, Digital Transformation, vendor management, project management, successful IT implementation and presentation to key stakeholders.

Wilson Cheruiyot has acquired the following degrees: Bachelor of Science in Mathematics and computer science (1994), Masters of Science in Computer Application Technology (2002) and PhD in Computer Science Applications Technology (2012). He is also certified with Microsoft association in the following: Microsoft Certified Database Administrator (MCDBA) and Microsoft Certified Professional (MCP). Prof. Cheruiyot is currently a Senior Lecturer and Dean at TTU. Between 2008 and 2017, Prof. Cheruiyot won international scholarships and research grants amounting to over $ 205,610, from the China Scholarship Council (CSC), to train PhD and M.Sc. JKUAT staff, and other Kenyan students, especially in the areas of Computer Science, Engineering and Business-Related courses.

Prof. Waweru Mwangi holds Bachelor of Science Degree in Mathematics from Kenyatta University (Kenya), a Master of Science Degree in Operation Research from Shanghai University (China) and PhD in Systems and Information Engineering from Hokkaido University (Japan). He is currently an Associate Professor in the Department of Computing, School of Computing and Informatics at JKUAT. His research areas include; systems modelling and development, smart agent computing, simulation and ICT policy formulation

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.