

# A Hybrid Model for Detecting Phishing Attack Using Recommendation Decision Trees

Duncan Eric O. Ogonji, Cheruiyot Wilson, Waweru Mwangi

School of Computing and Information Technology, Jomo Kenyatta University of Agriculture and technology, Nairobi, Kenya  
[eogonji@gmail.com](mailto:eogonji@gmail.com), [wilchery68@gmail.com](mailto:wilchery68@gmail.com), [waweru@gmail.com](mailto:waweru@gmail.com)

**Abstract.** Phishing performs by trying to trick the victim into accessing any computing information which looks original, then instructing them to send important data to unrestricted/unwanted privacy resource. 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 require collection of previous phishing attack logs, analyze them and create a list of such activities and block traffic from such sources. But this is a cumbersome task because the data size is very large, continue changing and dynamic nature. [1]. Instead of using single algorithm approach it would be better to use a hybrid approach. A hybrid approach would be better at mitigating phishing attacks because classification of different format of data is handled; whether the intruder want to use images or textural input to gain into another user system for phishing. Hybrid recommendation decision tress enhances any of 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 positive 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 false negative rate is 7.4%.

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

## I. INTRODUCTION

The most challenging risk in online system today are online scam 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 web sites 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 the phishing attack, the attacker points or directs victim to fake pages using social engineering techniques. Phishing performs by trying to trick the victim into accessing any computing information which looks original, then instructing them to send important data to unrestricted or any unwanted privacy resource that i not authorized[4].

Phishing is a method for taking on the appearance of a disclosed in substance hoodwinks 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 phishing attack is to trick receivers into divulging sensitive information such as bank account numbers, passwords and credit card details [7].

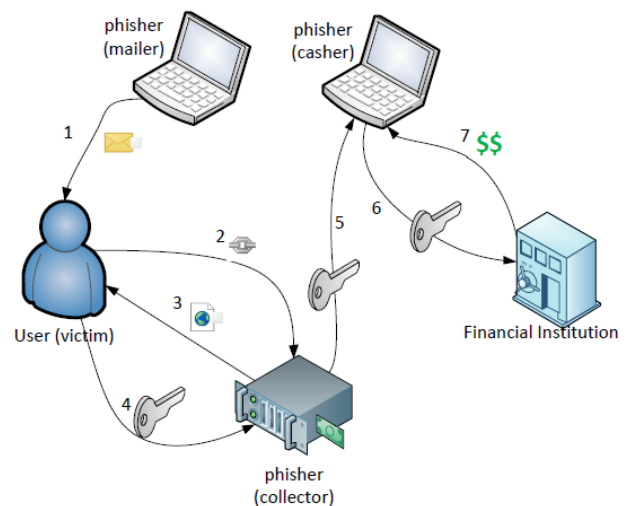


Fig. 1. Phishing Information Flow

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:

- i. **Root Node:** 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 implemented 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 is still a challenge while detecting phishing attacks[1]. Instead of using single algorithm approach it would be better to use a hybrid approach. A hybrid approach would be better at mitigating phishing attacks because classification of different format of data is handled; whether the intruder want 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 enhances any of 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 was 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. 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 Webpages 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 and good language-independence, speed of decision, resilience to dynamic phishes and flexibility to evolution in phishing techniques.

This research by [13] implemented a hybrid anti-phishing approach using some of the well-known phishing detection factors like 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 system automatically detects a phishing web URL uses data mining techniques.

### E. The Hybrid Recommendation Decision Tree Model

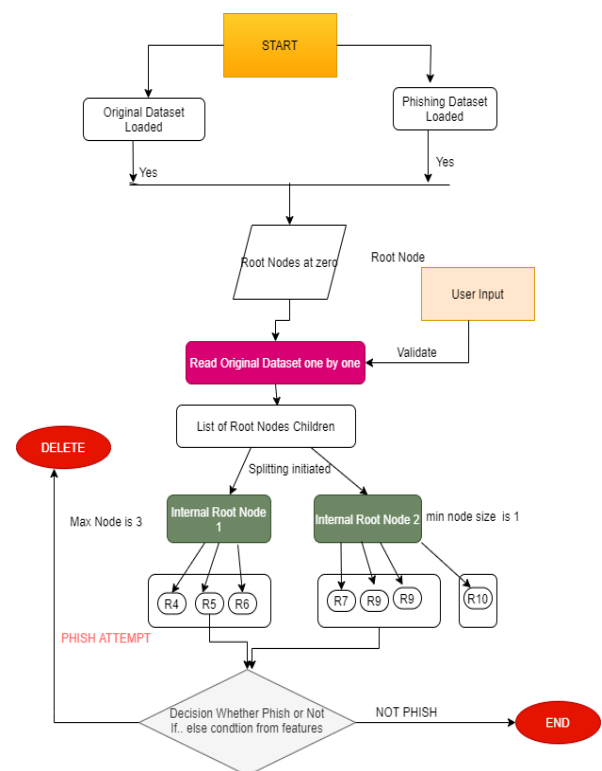


Fig. 2. The Implemented Hybrid Recommendation Decision Tree Model

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. First was to find the leaf that should host the new point where a list of root nodes 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 checks how many points it contains afterward.

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

Then now internal root node 2 has more than internal nodes 1 children, it is split and repeat this step recursively. To complete the insertion algorithm outlined here, we need to provide a few heuristics to R9 break ties for overlapping rectangles and to R8 choose the closest rectangle which 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*

Dataset was from a freemium Canadian Centre for Cyber Security <https://www.unb.ca/cic/datasets/andmal2020.html>. A purposive sample was used where 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 final sample for the study. The research narrowed down to scareware and riskware malware family's variables with a final sample of 4 and 4 variables respectively from the massive dataset. The four variables included hyperlinks of website URLs, Email clicks, user account registration and user account activation.

The basic information of the computer system in which the experiments was running was: windows 7 professional Operating system, Intel Core i3 of 1.80GHZ CPU, RAM memory of 16 GB. Google Colab was used due to it's 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 libraries 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 on the basis of created confusion matrix. Earlier the recommendation decision tree classifier model that was saved which needs to be loaded first after that, it is re-evaluated on the current test dataset.

On the basis of the generated recommendation decision tree classifier model after the evaluation of trained dataset will test the testing dataset and then prediction is made and confusion matrix is created. If the value of result node which is a parent node is -1 then the website is legitimate and if the value of 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 was tested phishing datasets from a freemium Canadian Centre for Cyber Security to detect phishing webpage based on hyperlinks information. The overall true positive rate of the model was 92.28 % and false negative rate is 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 least percentage of false positive rate which resulted in a higher accuracy of 88.67%. When false negative rate is low, it brings an impression of an efficient model approach to detecting phishing attacks. True Negative (TN) rate measures the rate of correctly detected legitimate sites in relation to all existing legitimate sites [15]. False Negative (FN) rate measures the rate of phishing web sites are incorrectly identified as legitimate in relation to all existing phishing websites. Accuracy (A) measures the rate of phishing and legitimate websites which are identified correctly with respect to all the websites [15].

TABLE II: Comparison Results

<i>Technique Used</i>	<i>True Positive</i>	<i>False Negative</i>	<i>Accuracy</i>
<b>Fuzzy Logic Classifiers</b>	82.3%	9.8%	80.5%
<b>AdaBoost with Random Tree</b>	76.9%	8.6%	78.5%
<b>Implemented Approach</b>	92.8%	7.4%	88.67%

#### IV. CONCLUSION

This research developed a hybrid model approach for phishing detection model using recommendation decision tree. The developed hybrid recommendation decision tree approach provided an overall true positive rate of the model was 92.28 % and false negative rate is 7.4%. The recommendation decision tree starts with the root node and follows the branch creation logic as entropy to create subsequent nodes that reaches the terminal node, which is also considered as 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 false negative rate is low, it brings an 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 on time series dynamic data that phishers are taking advantage of real time execution of data. The hybrid recommendation decision tree can be enhanced with other unsupervised machine learning techniques to detect phishing attacks.

#### V. 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.
- [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.
- [4] V. R. Reddy, C. V. M. Reddy, and M. Ebenezer, "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.
- [7] S. Bansal and A. Singh, "Machine learning in the prediction , determination and further study of different cyber-attacks," no. 10, 2019.
- [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.
- [9] R. V Kotawadekar, A. S. Kamble, and S. A. Surve, "Automatic Detection of Fake Profiles in Online Social Networks," no. 7, 2019.
- [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.
- [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.
- [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.
- [13] P. Saklecha and J. Raikwar, "Prevention of Phishing Attack using Hybrid Blacklist Recommendation Algorithm," no. 6, pp. 188–191, 2018.
- [14] N. S. Reddy and V. K. M, "Review Paper Detection of E-Banking Phishing Websites," no. 14, pp. 49–52, 2019.
- [15] H. Agrawal and R. R. Singh, "An Ensemble Approach for Detecting Phishing Attacks," vol. 9, no. 7, 2021.