IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Domain and Content Effectiveness Methodical Analysis Phishing Attacks and Websites Classification using Ensemble Learning

[1] Palagara Uday Kumar, [2] Narayanan.S, [3] Mridula Kannan
[1]Student, [2]Student, [3]Student,
[1]Department of computer science and engineering,
[1]SRMIST, Ramapuram, Chennai, India

Abstract : There are so many risks for inexperienced or negligent users, as well as a variety of tools and techniques used by infamous users to victimize people and access their private information, the internet or public internetwork has become a vulnerable place in modern society. resulting in sometimes smaller penalties But a large number of these victims experience severe losses as a result of falling for traps like phishing malware, tampering with data, , web jacking, Trojan attacks, cracking and salami attacks. As a result, despite online users' and software and application developers' ongoing efforts to build and keep the IT infrastructure safe and secure through the use of various techniques such as encryption, digital signatures, digital certificates, and so on. This research focuses on the topic of detecting and predicting phishing websites. On two distinct datasets, we employ URLs, fundamental new ensemble-based methods, and machine learning classifiers This investigation is done in three steps, once again using a consolidated dataset. They begin with classification using basic classifiers, Cross-validation is used both with and without evaluating ensemble classifiers. Finally, a review of their performance is conducted, and the findings are published to help other researchers use this study in future investigations. Because of the comfort and quick development of web applications, internet users take full advantage of these advantages and use the internet for nearly all of their daily activities, including reading the newspaper, shopping, paying bills, booking tickets, and finding entertainment. This phenomenon forces internet users to stay online for longer periods of time, which increases the likelihood that users will fall victim to phishing, an attack designed by hackers to steal sensitive information by initially luring users with lucrative offers before diverting them to a dubious website (which the user may not suspect) where they can trick the user.

Index Terms **-Ensemble learning, phishing websites.**

I. INTRODUCTION

The internet is becoming more prevalent in our lives, and We depend more and more on the services provided online. Everything from internet banking to smart home solutions, people's working cultures have been impacted, and as a result, the number of risks is increasing at a commensurate rate. There are numerous types of risks on these globally maintained network platforms. Aside from well-known terms such as hacking, cracking, web jacking, and online terrorist organizations, phishing is a common threat. Phishing is a means of perpetrating victims of such assaults are either unaware of them or fail to pay close attention to them. This is true of crimes committed online.

Businesses now experience about 1185 phishing attacks each month, according to Great Horn's report, the landscape for phishing attacks in 2020 is described in the report. The COVID 19 Pandemic, according to about 53% of cyber security experts, saw an increase in these attacks. Resolving a cyberattack requires enterprise security teams to spend one to four days. Phishing assaults were extremely successful during this pandemic, according to the same report's findings, which were supported by around 30% of cyber security professionals. The landscape of phishing attacks in 2020. The number of phishing emails used in their analysis (2020 Phishing Attack Landscape 2020) that target businesses globally was disclosed. Since so many phishing attempts are successful, productivity and profit have decreased as a result of the time needed to fix them. A lot of work has gone into finding phishing websites and stopping this type of fraud. But there is still no definite technique to tell reputable websites from phishing websites. Therefore, efforts are concentrated on techniques that can more accurately identify phishing websites. Different methods are used to identify phishing. Most Modern methods for spotting phishing websites rely on machine learning and intelligent models, like classifying websites based on their attributes. The most prominent elements of a website can be used to determine its validity, which can help to optimize the outcomes of these methods. In another study, two methods—wrapper-based feature selection and correlation-based feature selection—were used to examine the efficacy of discovering features in effectively detecting phishing websites. A significant subset of features was chosen in accordance with the calculated criterion. To choose the most crucial features, wrapper-based feature selection requires supervised algorithms and labels for each instance of the dataset. A selection of features

that produce the most precise categorization or prediction are chosen in the wrapper technique. Finally, in their study, the researchers evaluated the effectiveness of both approaches. In this study, we provide a heuristic method for identifying a website's key properties, which will aid intelligent models in phishing detection. We were able to identify the key characteristics that distinguish trustworthy websites from phishing websites with the aid of knowledge graph representation.

This study aims to propose a workable solution for URL-based phishing website prediction. attributes that may possibly be relevant for further research. This paper is divided into six sections, with the "Introduction" portion including an introduction and the "Motivation" section containing Motivation. The "Literature review" section offers a review of prior research publications and studies undertaken by various researchers around the world. The "Methodology" section contains the Methodology used in this paper. The "Results and discussion" column contains conclusion of all the study's findings and outcomes. Finally, the results are presented in the "Conclusion and Future Scope" section. This journal discusses machine learning techniques, their results, properties.

1.1 Motivation

Phishing is one of the most hazardous and heinous types of harmful crimes committed on the internet However, in today's online workplace, there are a lot of different security risks. Because of our increasing reliance on data, networks, and related technologies, the Internet is often referred to as a workplace. Several analysts and researchers are working hard and may or may not be affiliated with a phishing attack research organization. Phishing is a cyberattack that uses a phoney website that seems like a genuine website to deceive online visitors into disclosing sensitive information. The attackers who have the stolen credentials can access other well-known genuine websites in addition to the one that was the target. Although there are various anti-phishing toolbar extensions and strategies to block phishing sites, phishing attempts continue to be a big worry in the modern digital world. This inspires us However, the goal of all of these individuals and organizations is the same: to lower the danger of phishing, the majority of the time, the activities of infamous Cybercriminals are successful because there is no tested mechanism to provide people with accurately predicted information at the right moment or as and when it is needed.

The researcher's justification includes the forecast about the phishing website. The study by Gupta and colleagues (Gupta et al. 2021) focuses on phishing website prediction. The authors, BB Gupta et al., use nine characteristics and four classifier algorithms: Random Forest, KNN, SVM, and Logistic Regression.

Sahingoz (2019) predicted phishing websites using seven algorithms: Decision Tree, Adaboost, Kstar, KNN, Random forest, SMO, and Nave Bayes.

For phishing website prediction using URL-based features, Jain and Gupta (2018a) used 19 features and 5 classifiers, including Random Forest, SVM, Neural Networks, Logistic Regression, and Nave Bayes.

In order to identify phishing websites, Moghimi et al. (2016) propose a browser add-on plugin that employs both an SVM classifier and a rule-based method for URL characteristics. The research study presented here investigates 12 machine learning classifier methods divided into two categories: ensemble and basis classifiers.

1.2 Literature review

To identify and anticipate phishing websites using various techniques at various periods, multiple studies have been carried out by numerous authors and researchers. For instance, some studies suggested using visual cues, while others suggested using an image-based technique; logos are also seen to provide a basis for detection, and some researchers recommended HTML-based features, as well as domain blacklisting and whitelisting, to be examined for this purpose. The current research presents a URL feature-based approach for detecting and predicting whether these websites are phishing or not. The URL dataset came from the UCI machine learning repository (2020). Now on to the second dataset.

A model called PhiDMA was created by Sonowal and Kuppusamy in 2020. It has five levels, including a whitelist layer, a features layer, lexical signature, string matching, and an accessibility score comparison layer. They also debuted an algorithm for predicting phishing website URLs.A system called AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites was created by Yazan Ahmad Alsariera, Victor Elijah Adeyemo, Abdullateef Oluwagbemiga Balogun, and Ammar Kareem Alazzawi (2020). The purpose of this study is to offer society a fantastic response to the problem that phishing poses today. The drawbacks in this model are Problem of Diminishing Feature, Reuse Sensitive to the specifics of the training data and Low Robustness for Noisy Data.

The case-based reasoning method for phishing detection developed by Abutair et al. (2019) is known as CBR-PDS. They used a genetic algorithm weighting technique, feature extraction, URL blacklisting, and URL blacklisting to foresee the phishing URLs.

A phishing detection model based on feature classification that employs an artificial neural network (ANN) along with Nave Bayes and extreme learning machines (ELM) was proposed by Satapathy et al. (2019).

Phishlimiter, a phishing detection and mitigation approach developed with software defined networking (SDN) and a verified testbed environment, was presented by Chin et al. (2018). They also looked at support vector machines (SVM), J48 trees, Nave-Bayes, and Logistic regression approaches, which are all part of the ANN framework for detecting phishing.

PHISH-SAFE, a feature-based phishing URL detection system based on Machine Learning methods, was proposed and named by Jain and Gupta (2018b). To train their model, they employed support vector machines (SVM) and Nave Bayes classifiers.

Kumar and Gupta (2018) also discussed a method for detecting phishing website URLs based on hyperlink information. To predict the phishing website URLs, they used feature selection and CSS in addition to several machine learning classification methods as SMO, Nave Bayes, Random Forest, support vector machine (SVM), Adaboost, Neural Networks, C4. 5, and Logistic Regression utilising the WEKA tool.

A phishing detection method based on the model's traits and content was given by Abdelhamid and Abdel-jaber in 2017. They used a variety of machine learning classifiers in their experimental work, including eDRI, RIDOR, Bayes Net, C4.5, OneRule, Conjunctive Rule, SVM, and Boosting.

Fresh-Phish, a method for automatically identifying phishing websites, was introduced by Shirazi et al. (2017). In order to test two classifiers from the sci-kit-learn library and four classifiers from the Tfcontrib (2020) package, they utilised Whois data from whoixmlsapi.com (WHOIS API gives access to domain registration information WhoisXML API 2020) (Varoquaux et al. 2015). In addition, a DNN with built-in optimisations like Adadelta, Adagrad, and Gradient Descent was constructed using TensorFlow and TFcontrib.

1.3 Objective

To produce a deep hierarchical representation from a given dataset.
To detect ever-changing phishing websites.
To assign a value to each feature's validity and to choose features.
To offer important information about whether a website is associated with phishing assaults.
Analyzing a website's textual content to determine similarities between it and other websites.

II. PROBLEM STATEMENT

Sovereign bodies (for example) support the development of trustworthy ML systems; nonetheless, any enhancement must be economically justifiable. No system (ML-based or not) is flawless, and ensuring protection against omnipotent attackers is an appealing but unrealistic goal. In our situation, a security system should raise the cost of an attacker achieving their goal. Real attackers have a cost/benefit mindset: they may try to dodge a detector if the benefits outweigh the costs. In reality, worst-case outcomes are the uncommon rather than the rule.

2.1 Purpose

Phishing is a cyber assault that mislead internet users into disclosing personal information by impersonating a reputable website with a bogus website. Attackers utilising stolen credentials may use them to get access not just to the targeted website, but also to other prominent legitimate websites. To counteract phishing sites, there are various anti-phishing methods, toolbars, and extensions available, but phishing efforts remain a major concern in today's digital environment. This motivates us.

III. EXISTING SYSTEM

In order to detect phishing websites, we provide a feature-free method based on the Normalised Compression Distance (NCD), a parameter-free similarity measure that computes the similarity of two websites by compressing them, doing away with the necessity for feature extraction. Additionally, it removes any reliance on a certain set of website

features. By analyzing their HTML and calculating how closely they resemble known phishing websites, this approach classifies websites. We employ the Furthest Point First approach to find samples that are suggestive of a cluster of phishing websites in order to extract phishing prototypes. We suggest employing an incremental learning approach as a foundation for continuous and adaptive detection without extracting new features when there is idea drift. With an AUC score of. %, a high true positive rate (TPR) of about%, and a low false positive rate (FPR) of. %, our proposed strategy significantly outperforms existing methods in identifying phishing websites on a large dataset. Our technique can be used in real systems and employs prototypes to do away with the necessity for long-term data storage in the future. In this study, we provide a feature-free approach for detecting phishing websites based on the Normalised Compression Distance (NCD), which measures website similarity by compressing it, removing the need for manual examination. This method reads webpage HTML source codes and compares them to known phishing websites. We advise The Furthest Point First technique is used to find samples that are indicative of a cluster of phishing webpages in order to extract phishing prototypes. When there is idea drift, the approach may be utilised as the foundation for continuous and adaptive detection without the need for additional feature extraction. With an AUC score of. %, our proposed approach exceeds earlier studies in detecting phishing websites when evaluated on a current large-sized dataset.

3.1 Drawbacks

Create models with a relatively low accuracy but a relatively high false-positive rate.
As the number of hidden layers increases, so does the computational complexity.
Unsatisfactory when dealing with data that contains a high number of features, noisy data, and so forth.
Sensitive to the details of the training data; • Unable to reduce prediction variance; thus, unable to reduce generalisation error.

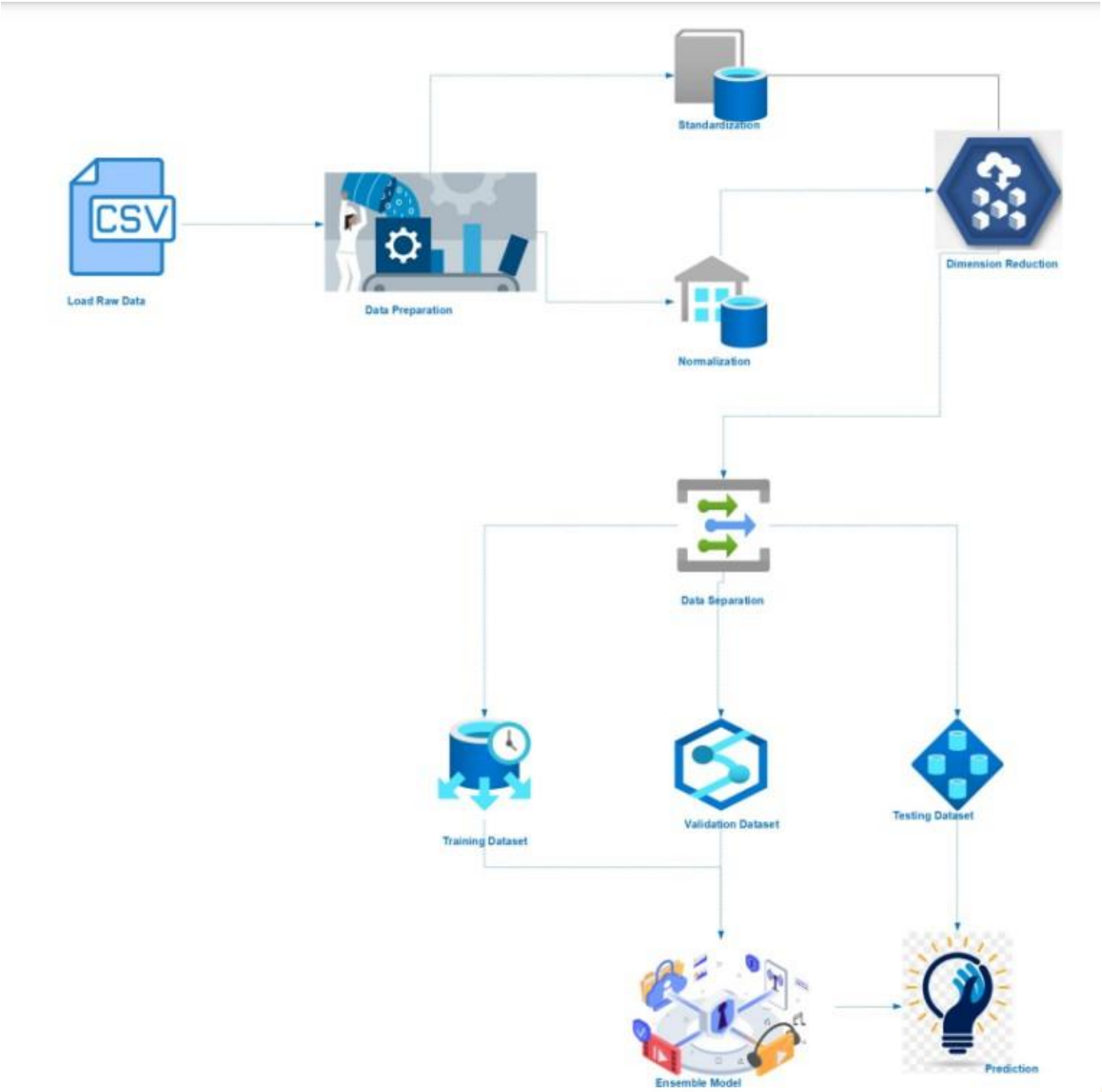


Fig. architecture diagram

IV. PROPOSED SYSTEM

The suggested technique leverages Fuzzy Rough Set(FRS) using theory to select the most useful characteristics from three benchmarked data sets.The parameters gathered are fed into three widelyused phishing detection classifiers. SMO, amultilayer perceptron, and a random forest are allemployed. To put FRS feature selection to the test in building a generalizable phishing detection system,the proposed method trains each classifier on a separate out-of-sample dataset. This training set consists of URLs pulled at random from the internet. All hyper-parameters are also set in accordance withearlier works' settings. We train each classifier on adifferent out-of-sample dataset to evaluate FRS feature selection's ability to build a generalizable phishing detection. The suggested system differ significantly, the retrieved domain name will be flagged as phishing. Because phishing campaigns are only active for a short period of time, this approach basedon domain rank can effectively detect phishing.

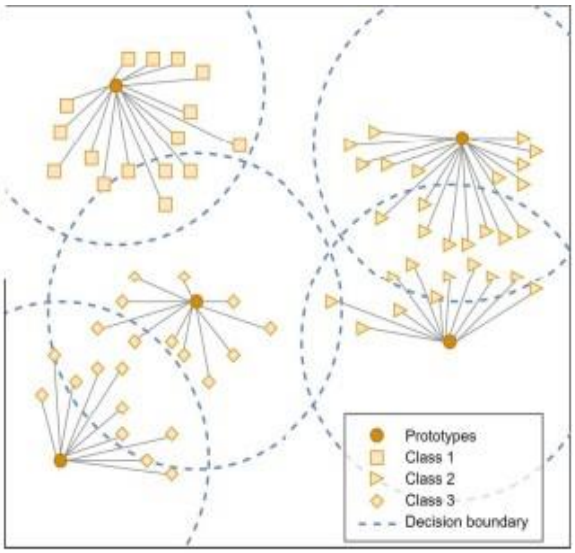
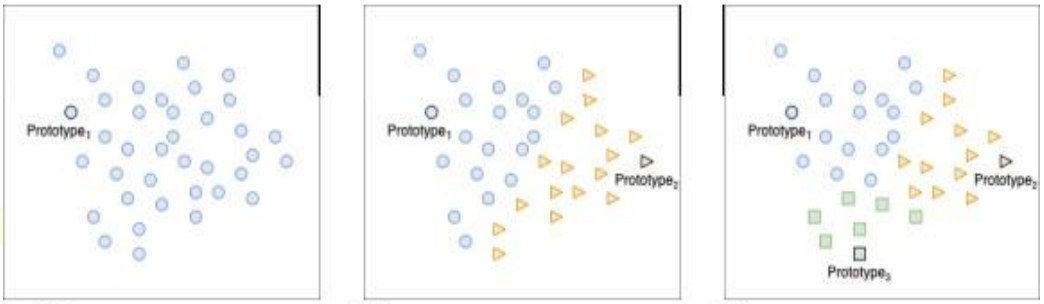


Fig. 1. Classification using Prototypes.



(a) First prototype (b) Second prototype (furthest point from $prototype_1$) (c) Third prototype (furthest point from $prototype_2$)

Fig. 2. Furthest Point First Algorithm.

4.1 System Requirements :

Hardware Requirements
- processor: Minimum i3 Dual Core
- Ethernet connection (LAN) OR a wireless adapter (Wi-Fi)
- Hard Drive: Minimum 100 GB; Recommended 200 GB or more
- Memory (RAM) : Minimum 8 GB; Recommended 32 GB or above

Software requirements
- Python
- Anaconda
- Jupyter Notebook
- TensorFlow

4.2 Advantages :
Detect phishing websites with a high degree of accuracy
while attaining low false positive and negative rates.
- Can be used to improve prediction in industrial applications.
- There is no significant distortion of the forecast outcome. The model learns discriminative characteristics.
- Easy to comprehend and interpret.

V. RESULTS AND DISCUSSION:

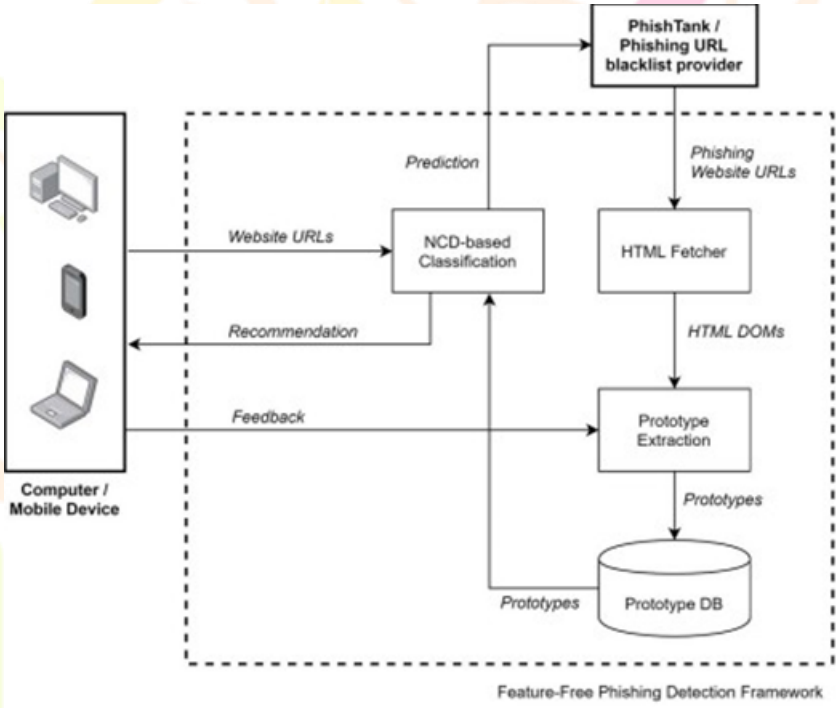
The initial batch of data utilised in this study, 2456
instances of website URL data with 31 different features, was taken from the UCI Machine Learning Repository. The second dataset shares the same 11,055 occurrences and 31 characteristics as the first dataset and comes from the same Kaggle.com source. The last one (1) attribute out of

the 31 total attributes, designated as a result, contains values that reflect 1 as (Phishing website), 1 as (non-phishing website), and 0 as (Suspicious website). Based on URL features, the 30 attributes in both datasets define URL characteristics. The dataset was previously preprocessed, however we nevertheless did data standardisation to ensure accurate and efficient processing while doing classification and prediction on the aforementioned. Because all components are significant, the study considered all 30 criteria. All 30 characteristics were examined since they are all significant. The study looks at how all variables affect how well classifiers work to identify phishing websites. The examples provided in this article illustrate the best results. The Extra Trees and XGBoost classifier algorithms attained the highest accuracy of 99.18% on two datasets selected for these experiments from the UCI Machine Learning Repository: Phishing Websites Data Set (UCI Machine Learning and Repository: Phishing Websites Data Set 2020) and Kaggle. com (Phishing website dataset Kaggle 2020). The datasets are preprocessed, normalised datasets containing 2456 and 11,055 occurrences, respectively, of website URL data in the first dataset and 30 unique URL characteristics plus 1 result attribute with numerical values obtained from the data. These criteria are used to detect whether a website is phishing, non-phishing, or suspicious.

The data acquired using the mentioned metadata was processed using a variety of methods. There are eight columns in a table. The name of the appropriate classification technique is represented by the classifier in the first column. The second column displays the confusion matrix for the appropriate classification algorithm. The phishing or non-phishing classification for the particular algorithm and row is shown in the third column of the confusion matrix. The results produced for each method are summarised in the following columns by precision, recall, f1 score support, and accuracy.

The second approach, linear discriminant analysis, yields precision values of 0.95 for non-phishing and 0.94 for phishing, recall values of 0.96 for non-phishing and 0.93 for phishing, F1 score values of 0.95 for non-phishing and 0.94 for phishing, and overall accuracy of 92.87 percent.

The third technique is K-Nearest Neighbour, which yields accuracy results of 0.97 for non-phishing and 0.94 for phishing, recall results of 0.95 for non-phishing and 0.96 for phishing, and an F1 score of 0.96 for non-phishing and 0.95 for phishing.



Algorithm 1 Prototype Extraction

```
prototypes ← ∅
for all x ∈ data do
    distance[x] ← ∞
    cluster[x] ← ∅
end for
while max(distance) > d_threshold do
    z ← arg max_{x∈data} distance[x]
    for x ∈ data do
        if distance[x] > NCD(x, z) then
            distance[x] ← NCD(x, z)
            cluster[x] ← z
        end if
    end for
    prototypes ← prototypes ∪ {z}
    data ← data \ {z}
end while
```

Algorithm 2 NCD-Based Classification

```
for x ∈ data_test do
    z ← arg min_{p∈prototypes} NCD(x, p)
    if NCD(x, z) < d_threshold then
        classify x as phishing
    else
        classify x as non-phishing
    end if
end for
```

Algorithm 3 Incremental Learning

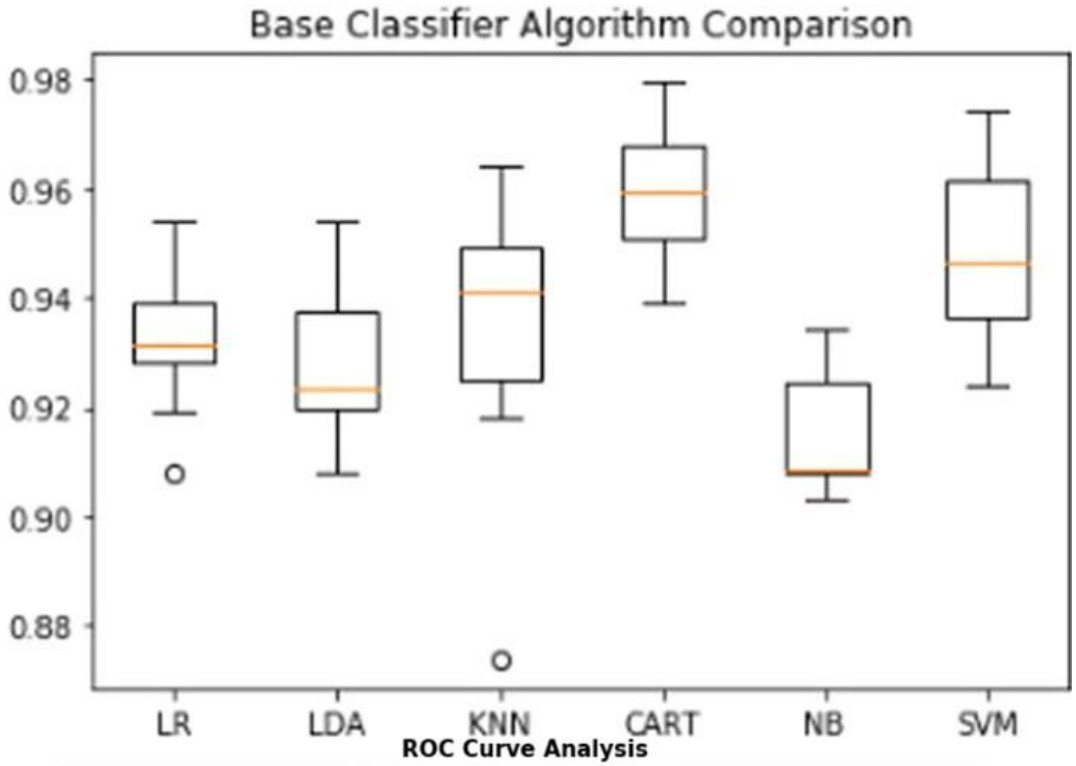
```
for data ← source do
    for x ∈ data do
        classify x using prototypes          ▷ Algorithm 2
    end for
    rejected ← samples in data rejected as phishing
    prototypes_new ← prototypes in rejected  ▷ Algorithm 1
    prototypes ← prototypes ∪ prototypes_new
end for
```

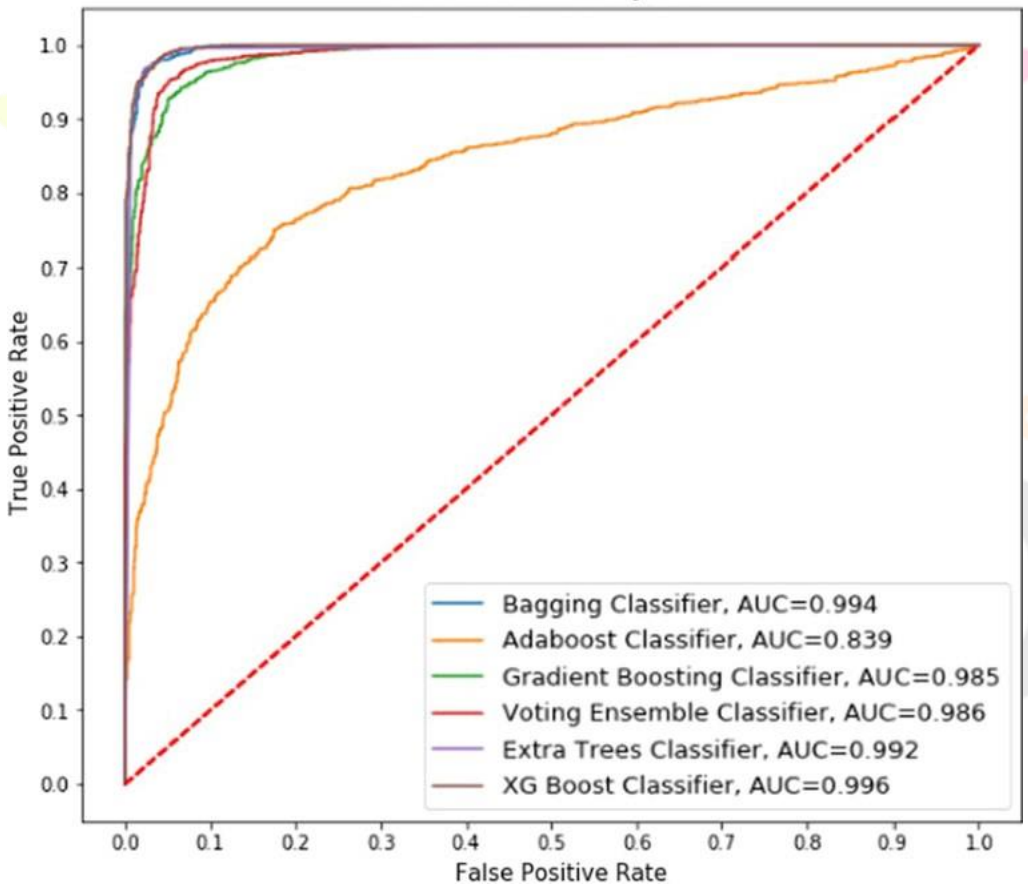
Logistic regression, the first algorithm, yields precision of 0.96 for non-phishing and 0.95 for phishing, recall of 0.96 for non-phishing and 0.95 for phishing, and F1 score of 0.96 for non-phishing and
0.95 for phishing, for a total accuracy of 93.28 percent.

Decision Tree, the fourth strategy, yields accuracy scores of 0.99 for non-phishing and 0.97 for phishing, recall scores of 0.98 and 0.99 for phishing, and F1 scores of
0.98 and 0.98, respectively, for non-phishing and phishing. 95.92 percent of the time, this strategy is
accurate.

The fifth approach, Gaussian Naive Bayes, generates accuracy values for non-phishing of 0.96 and 0.89, recall values for non-phishing of 0.91 and 0.95, and F1 score values for non-phishing of 0.94 and 0.92, respectively.

The sixth algorithm, Support Vector Machines, yields F1 scores of 0.96 for non-phishing and 0.95 for phishing as well as precision values of 0.96 for non-phishing and 0.95 for phishing. The results are presented as a boxplot using Python code.





.

REFERENCES

[1] W. Stats, Internet usage statistics - the internet bigpicture: World internet users and 2020 populationstats, Tech. rep.(2023)
[2] J. Singh, Detection of phishing e-mail, InternationalJournal of Computer Science and Technology 2(3) (2011) 547549.
[3] A.Almomani, B. B. Gupta, S. Atawneh, A. Meulen-berg, E. Almomani, A survey of phishing email ltering als 15 (2013) 20702090. doi:10.1109/SURV.2013.030713.00020.
[4] R. Nidhin A Unnithan, Harikrishnan NB, S. KP, Gualberto et al.: The Answer is in the Text: Multi-Stage Methods for Phishing Detection basedon Feature Engineering Detecting phishing e-mailusing machine learning techniques, in: Proceedings of the 1st Anti-Phishing Shared TaskPilot at 4th ACM IWSPA co-located with 8th ACM Conference on Data and Application Security
[5] Y. Goldberg, G. Hirst, Neural Network Meth- ods for Natural Language Processing, Morgan & doi : 10 . 2200/Claypool Publishers, 2017. S00762ED1V01Y201703HLT037.
[6] P. D. Turney, P. Pantel, From frequency to meaning: Vector space models of semantics, J. Artif. Int. Res. 37(1) (2010) 141188. doi:10.1613/jair. 2934.
[7] M. Verleysen, D. Franois, The curse of dimension- ality indata mining and time series prediction, in: Articial Neural Networks: Computational Intelligence and Bioinspired Systems, IWANN05, Springer-Verlag, 1007/11494669_93.
[8] M. Najafabadi, F. Villanustre, T. Khoshgoftaar, N. Seliya, R. Wald, E. Muharemagic, Deep learning applicationsand challenges in big data Journal of Big Data 2 (122015). analytics, doi:10.1186/s40537-014-0007-7.
[9] D. F. Tenrio, J. P. C. L. da Costa, R. T. Sousa Jr, Great- est eigenvalue time vector approach for blind detectionForensic Computer Science (ICoFCS) (2013).
[10] R. A. Hamid, J. Abawajy, T.-h. Kim, Using fea- ture selection and classication scheme for automatingphishing email detection, Studies in Informatics and Control 22 (Mar 2013).
[11] M. R. Islam, J. Abawajy, A multi-tier phishing detec- tionand ltering approach, Journal of Network and ComputerApplications 36 (2013) 324335. doi: 10.1016/j.jnca.2012.05.009.
[12] M. Zareapoor, K. R. Seeja, Feature extraction or featureselection for text classication: A case study on phishingemail detection, International Journal of Informa- tion Engineering and Electronic Business 7 (Mar 2015).
[13] Ramanathan, H. Wechsler, phishgillnetphishing detection methodology using probabilistic latent se- mantic analysis, adaboost, and co-training, EURASIP Journal on Information Security 2012 (Mar 2012). Gualberto et al.: The Answer is in the Text: Multi-StageMethods for Phishing Detection based on Feature Engineering doi:10.1186/1687-417X-2012-1.
[14] T. Gangavarapu, C. Jaidhar, B. Chanduka, Applicabil- ityof machine learning in spam and phishing email ltering:review and approaches, Articial Intelligence Review (02 2020). doi:10.1007/s10462-020- 09814-9.
[15] J. M. et al., The apache spamassassin public corpus,Tech. rep. [50] G. Mujtaba, L. Shuib, R. Raj, N. Majeed, M.al garadi, Email classication research trends: Review and open ACCESS.2017.2702187.
[16] S. Maldonado, G. LHuillier, Svm-based feature selec- tion and classication for email ltering, in: P. La- torre