Malicious URL Detection Using Machine Learning and Deep Learning

Prasanna Kumar M.[1], Dhanraj S.[2], Dr.Bhavani Shankar[3]
[1]*Asst.Professor, Dept. of CSE, R.N.S Institute of Technology, Bengaluru*
[2] *Asst. Professor, Dept. of CSE, EWIT, Bengaluru*
[3]*Associate Professor, Dept. of CSE, R.N.S Institute of Technology, Bengaluru*

Abstract—The quantity and magnitude of network information security risks have continually rising. Hackers today mostly employ techniques that target technology from beginning to conclusion and take advantage of human weakness. These methods include pharming, phishing, and social engineering, among others. These attacks include a number of phases, one of which is to trick users through malicious Uniform Resource Locators (URLs). In light of this, malicious URL detecting is a hot subject right now. A variety of academic research have demonstrated several ways to identify malicious URLs using machine learning and deep learning technologies. Based on our hypothesized URL behaviours and characteristics, we provide a machine learning-based solution for detecting malicious URLs in this work. Furthermore, big data technology is applied to enhance the ability to appreciate fraudulent URLs based on aberrant activity. A novel collection of URL traits and behaviours, a machine learning algorithm, and big data technologies make up the suggested detection method, to summarize it. The experimental findings indicate that the specified URL features and behaviour can increase total the capacity to identify dangerous URLs. This suggests that the proposed methodology may be seen as a successful and consumer method of identifying dangerous URLs.

Keywords: phishing, machine learning, malicious URL detection.

I. INTRODUCTION

Internet sources are identified to by respective Uniform Resource Locator (URL). In [1], introduced the traits and two simple factors of the URL: protocol identifier, which suggests what protocol to use, and aid name, which specifies the IP tackle or the area identify the place the aid is located. It can be considered that every URL has a unique shape and format. Attackers regularly attempt to alternate one or extra elements of the URL's shape to deceive customers into spreading their malicious URLs. Malicious URLs are recognized as hyperlinks that adversely have an effect on users. Customers will be referred to resources or pages via certain URLs just so cybercriminals can run code on users systems, and redirect customers to undesirable sites, malicious websites, every malware or another fake website downloads. Additionally, malicious URLs can be hid in downloads that are considered as safe, and they may propagate rapidly via file as well as communication exchange in open networks. utilising malicious URLs in methods of attack includes. Phishing, spam, drive-by downloads, and defacement.

Malicious URLs are a hazardous danger to cyber security, these kinds of assaults can lead to scams, the place humans they lose money, accounts, and personal details. It is essential to be in a position to discover and act towards the Imost common technique for addressing these threats is the use for blacklists, but this strategy has many problems when used in resistance to new URLs, so we concentrate more and more on computer learning algorithms. and that is exactly the focal point of this project. The hazard of community facts insecurity is growing hastily in range and stage of danger. The techniques broadly speaking used by means of hackers these days are to assault end-to-end technological know-how and take advantage of human vulnerabilities.

These tactics comprise pharming, phishing, social engineering, and more. Using malicious Uniform Resource Locators to trick clients is one stage in carrying out these assaults (URLs). Because of this, finding malware URLs is a popular past time today. Numerous scientific studies have found a variety of methods for spotting fraudulent URLs using deep artificial intelligence and computer pedagogical strategies.URL sharing is a core appeal of current social media structures like Twitter and Facebook. Recent research discover that round 25% of all popularity messages in these structures comprise URLs, amounting to hundreds of thousands daily share rate of URLs. However, this potential also brings risks from unscrupulous users .

trying to spread malware, phishing, and other low-quality stuff. Multiple recent efforts acknowledge Spam URL issues eventually result in a decrease in the quality of data made available by these systems. The development of internet businesses like social networking, e-banking, and e-commerce has been significantly influenced by the Covid 19. Tragically, advances in technology come with a multitude of abuse techniques for users. These attacks frequently involve fake websites that can be hacked and steal a variety of confidential information utilize.

Internet resources are characterized with their URLs (Uniform Resource Locators). The attributes of URLs and two basic bits were introduced in [1]The name of the resource indicating the protocol identity, which denotes the protocol to use, and the Port number or regional, which identifies the existence of the resource, are also obligatory fields. You can presume that every URL has a distinct format and shape. Attackers frequently attempt to change at least one of her URLs. form sections in an effort to deceive the users and circulate harmful URLs. Links that hurt users are recognised as malicious URLs. These URLs reroute users to resource or pages that provide attackers accessibility to the user's computer to run code, send users to undesirable, harmful, or other phishing websites, or let people download malware. To do On shared networks, malicious URLs might well be quickly distributed through the exchange of information and messages by masking yourself in seemingly secure downloads. Malicious URLs may be employed in drive-by downloads, phishing, spam, and many other attack methods.

Malicious URLs represent a serious threat to online security as such attacks can result in schemes where individuals lose funds, accounts, and personal data. It is essential to be able to recognise these dangers and respond to them. The most prevalent strategy has been to use blacklists, but this method presents numerous challenges. in opposition to new URLs, so we are increasingly more targeted on laptop getting to know precisely that is the project's main focus: algorithms. The scope and level of peril posed by a lack of trust in public information is rapidly expanding. The techniques broadly speaking use through means of attackers these days are to assault end-to-end technological know-how and take advantage of human vulnerabilities.

Pharming, phishing, social engineering, and other strategies are some of these. One step in carrying out these attacks is to fool the people into browsing a dangerous URL (Uniform Resource Locator).Because of this, detecting harmful URLs is a popular past time today. Multiple malicious URL detection equipment, a lot of which are based on a knowledge of laptops and an in mastery methodologies, have been validated by numerous scientific inquiry. A major perk of contemporary social media platforms like Twitter and Facebook is sharing URLs. According to recent research, URLs are present in around 25% of all popular stories in these frameworks. This translates to daily URL exchange of hundreds of thousands. This potential, meanwhile, is exacerbated by dishonest users who do want to install malware, phishing, and other uninformed. Numerous recent initiatives have recognized the problem of spam URLs eventually. The data which is available in these systems is of questionable quality. The evolution of tech startups, including social networking, e-banking, and e-commerce, has been greatly affected by the Covid 19 Unfortunately, technology advancements include a number of exploitation techniques of individuals. Such attacks frequently use rogue websites that collect a variety of individual information that a hacker may use.

II. LITERATURE SURVEY

Our contributions and contributions to this work are as follows: (1) This work suggests a DCNN-based model for identifying malicious URLs. The remarkable multilayer convolution structure receives a using the dynamic convolution technique, a novel folding layer. The pooling layer is replaced with the k-max-pooling layer. The vector entrance dimension determines the breadth of the The core layer of the dynamic convolution process has an unique mapping. Additionally, In order to obtain more in-depth points all over a wide swath, the depth of the current convolution layer and the size of the URL inputted both affect how the pooling layer values are altered. (2) The components are gathered from the URL sequence during the function extraction and representation process. We found variables that the convolutional neural network right away processes and combines into a vector. in order to evaluate the classification model. This approach combines the strengths of personality embedding and phrase embedding in addition to simplifying the function extraction process and doing rid with the requirements for manual point extraction. Word embedding can be used to collect phrase sequence data. But not through personality embedding. Character embedding may process strange words and characters in the URL. Additionally, the dictionary and vector dimensions are no longer too large. The aggregate can further correctly specify the URL and store a memory space, which will aid in information extraction from the URL. (3) We carried out an extensive number of compared trials to establish the viability of the mannequin provided in this study. In order to prove that phrase More accuracy can be achieved by embedding based on personality than through phrase plus persona embedding. We conducted three separate tests. Further, we do three comparison experiments that demonstrate how using the community shape made up of a DCNN and specific URL parameters may have a greater impact.

Currently, there are two broad categories of approaches for detecting malicious URLs: standard methods that only rely on blacklisting and methods that heavily rely on machine learning. The primary list detection technique is first described in literature. Although this strategy as simple and effective, it is restricted and isn't able to realise freshly produced dangerous URLs. The literature demonstrates that. Currently, there still are two broad categories of techniques for identifying malicious URLs: standard methods that just rely on blacklisting and methods that heavily rely on machine learning. The primary list detection technique is first described in literature. Although this strategy as simple and effective, it is restricted and isn't able to recognize freshly produced dangerous URLs. The literature emphasizes that point.

Currently, there still are two broad groups of techniques for identifying malicious URLs: standard methods that just rely on blacklisting and systems that heavily rely on machine learning. The primary list detection technique is first discussed in the literature. Although this methodology as simple and effective, it is restricted and isn't able to materialize freshly produced malware URLs. The literature makes that point. Numerous scientists have suggested evaluate a URL's legitimacy is harmful based just on the strings included within the URL. These strategies can regularly retrieve reliable information from the URL. For instance, literature uses the personality stage of the cyclic neural community mannequin to categorise URLs given by DGA. The research suggests using a heavy laptop workload to find dangerous URLs. Literature uses character-level semantic factors in tandem with the n-gram model with deep learning to determine whether or not DGA creates the URL. The literature lists a variety of deep learning techniques for detecting bad URLs. It comprises the deep convolution structure, the bidirectional Autoencoder, the blended CNN and LSTM designs, and the single continuous temporary memory (LSTM) styles.

III. METHODOLOGY

MACHINE LEARNING & DEEP LEARNING:

Due to its ability to handle enormous quantities of data, machine learning and deep learning have become increasingly powerful tools in recent years.

SOFTWARE REQUIREMENTS:

- Accurate classification: The system should be able to classify URLs into Benign and Malicious accurately.
- Accessibility: The User should be able to access the application to enter URLs for classification
- User friendly: The software should be user-friendly for users who are not well versed in computer science.

CNNs and RNNs have been included into neural network structures to fulfill this purpose. The diagram appears as follows:

RNNs and LSTMs are examples of sequence generator architectures that can begin by translating a feature vector with a set length from either a picture. To prepare a list of labels or terms for your image, follow this procedure. The encoder used for this project is ResNet50. Using pre-trained algorithms, the millions of images in the ImageNet dataset were separated into 1000 teams. Replace the top layer, which contains 1000 neurons and is only used for ImageNet classification, with a linear layer containing a double the neurons. as add to use this network. This is because the network's weights are adjusted to identify numerous features common to nature. Several neurons create through LSTM. Long Short-Term Memory (LSTM) cells compose an RNN, which is used to generate captions continuously from raw images. To remember things from earlier steps, these cells employ the repetition and gate ideas. For some further detail, you may read or watch this. The outputs the output layer, typically predicts the subsequent word based based on the visuals and present flow, obtains the combined output from the encoder and decoder.

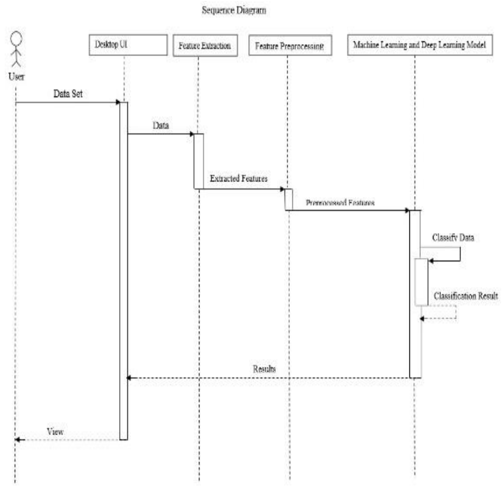


Fig1: Sequence Diagram

The proposed system is:

- The graphical user interface is the first alternative (GUI). At this phase, the user alters the system.
- The user must sign in or register if it is their first first-time visitor.
- At that point, the user can upload an image and obtain a description.
- Once we use a Mri to extract features from the imagery and transform them into fixed-length feature vectors following the user enters a link or provides text.
- Change the preprocessed image's size, orientation, color, brightness, and viewpoint. The captions' noise is greatly reduced as a result of this approach. B. Punctuation. The RNN receives feature vector input and use it to recursively construct captions for the photographs.

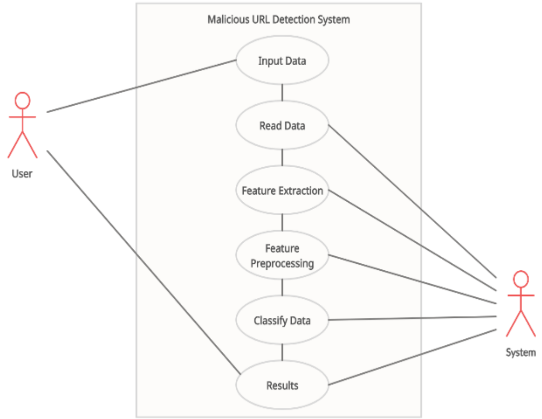


Fig2: Use Case Diagram

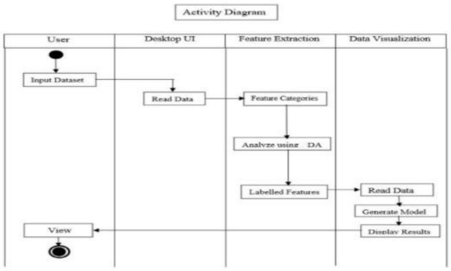


Fig3: Activity Diagram

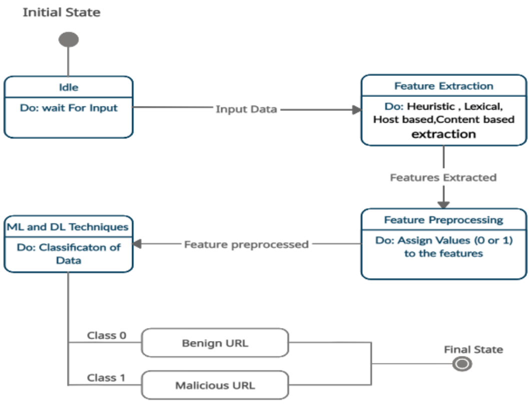


Fig4: State Diagram

CREATING THE MODELS:

- This is a supervised machine learning and deep learning job, which is evident from the dataset. Classification and regression are two of the most common supervised learning issue categories.
- The input URL is categorised as either malicious (1) or real (2), which presents a classification issue for the data collection that was used. (0). The following types of supervised learning models (classifications)

have been taken into account when training the dataset:

- Decision Tree
- Random Forest
- Multilayer Perceptrons
- XGBoost
- Support Vector Machines

PROJECT DESIGN/MODEL:

- Importing the Dataset: 11000 URLs comprise the experimental dataset for the malicious URL identification model, of which 6000 have been determined to be malicious and 5000 to be secure. The Virus Total tool looked at each of these URLs to verify the markings that are on each URL. CSV format has been employed to keep the complete dataset. Some of the sources of data are URLhaus, Phishtank, Alexa, Malicious_n_Non-Malicious URL .
- Training the models: There are two subsets of the dataset, which includes malicious and secure URLs. 20% of the information is used for testing, and the remaining 80% is used for training. For each algorithm, the trial is conducted numerous times.
- Generating the Predictions: After the training stage, the models are tested using test data. The testing is done for all the algorithms used and their performance is recorded by comparing their predictions with actual values and also percentage of model's correct predictions.
- Evaluating and comparing the model: Both the Training and Testing performance are evaluated for every algorithms.Accuracy score is used to measure the accuracy of the algorithms.

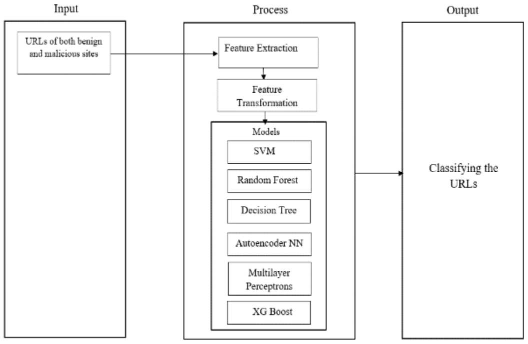


Fig5: Work Flow Diagram

EVALUATION METRICS:

The accuracy rate is the proportion of right choices made across all test sets.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where:

- TP- The amount of malicious URLs that were accurately labelled is a true positive.
- FN - The quantity of malicious URLs mistakenly labelled as secure is known as false negative.
- TN- True negative is the quantity of properly labelled safe URLs.
- FP - False positives are instances where secure URLs are mistakenly labelled as malicious.

After the evaluation, the performance of each model is compared and then the best performing model is chosen for web framework implementation.

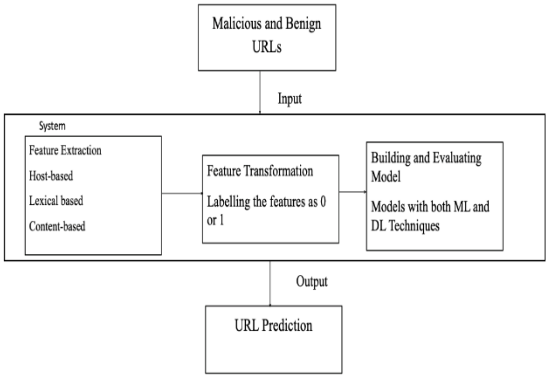

Malicious and Benign URLs
Input
System
Feature Extraction
Host-based
Lexical based
Content-based
Feature Transformation
Labelling the features as 0 or 1
Building and Evaluating Model
Models with both ML and DL Techniques
Output
URL Prediction

Fig 6: Component Design

SOFTWARE IMPLEMENTATION:

Step 1: Importing the Dataset
Step 2: Feature Extraction
Step 3: Feature Transformation
Step 4: Creating the Model
Step 5: Training the Model
Step 6: Generating the Predictions
Step 7: Evaluating and comparing the model

V. RESULTS AND DISCUSSION

Models are rated based on their accuracy value. It shows the proportion of accurate forecasts that were made. compared to all input samples. Compare all models based on training and assessment accuracy.

A Malicious URL detecting Machine Learning and Deep Learning Model for accurately detecting and classifies the Benign and Malicious URLs.

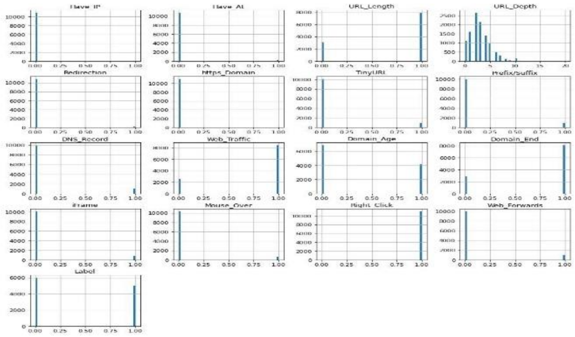


Fig7: Separate Data Set Graphs

The accuracy determines the percentage of all accurate classifications. The precision calculates the proportion of correct classifications to incorrect classifications that are penalised. The recall determines the proportion of accurate classifications that are penalized for missing entries. The sensitivity or true positive rate are other names for memory. The F1-score functions as a derived efficacy measurement by estimating the harmonic mean of precision and recall. Plotting the receiver operating characteristic (ROC) curve with TP R on the Y-axis and F P R on the X-axis indicates the efficacy of the classifier. When classes are equally distributed, the ROC curve is usually employed; when classes are unbalanced, the is the precision-recall curve. We calculated the trade-off between accuracy and recall across the varying threshold in the range of [0, 1] in order to produce a precision-recall curve.

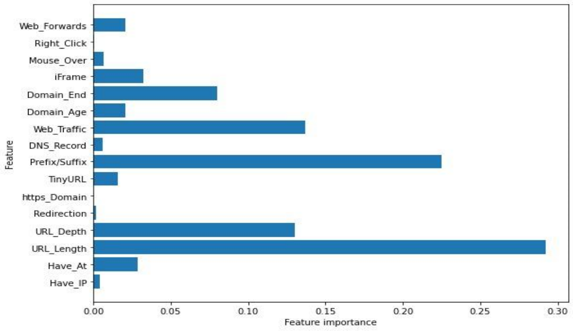


Fig8: Top 16 Sub Domains Grouped By Suffix

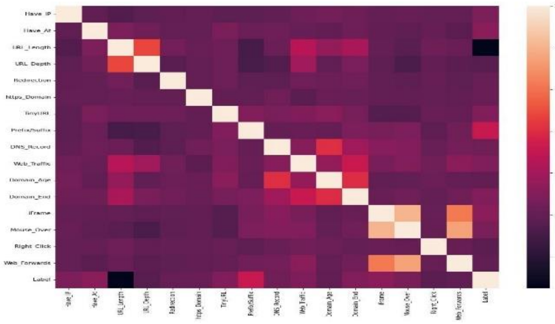
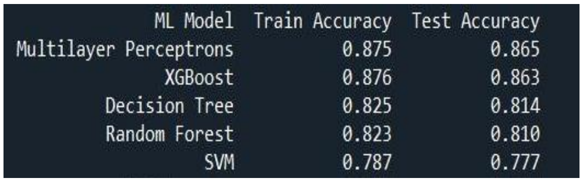


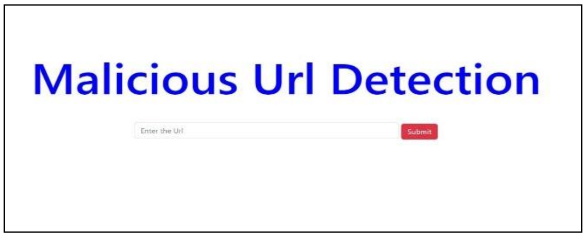
Fig9: Heat Graph



```
                ML Model  Train Accuracy  Test Accuracy
   Multilayer Perceptrons           0.875          0.865
                  XGBoost           0.876          0.863
            Decision Tree           0.825          0.814
            Random Forest           0.823          0.810
                      SVM           0.787          0.777
```

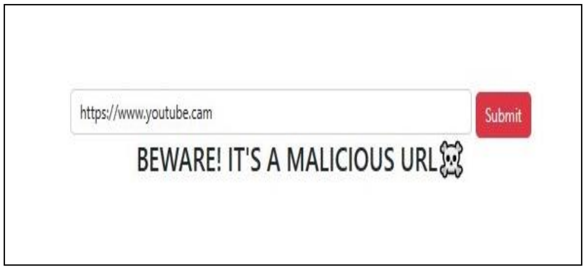
Fig10: Accuracy of the Models

Because of the precision and reliability of the Voting methodology, it is possible to determine whether a URL is secure or malicious. AUC of all models work with only a slight variation. This may help detect malicious URLs with greater frequency. This still represents one of the key directions for future growth. They were all text representations that used character level Keras encoding. technique and outperformed the DNN-based three-gram technique in terms of performance. In comparison to Deep learning combined with hacking-based malicious URL identification can be a reliable alternative to conventional machine learning-based approaches that rely on handcrafted features. This is a result of malicious authors' capability to use subject knowledge to learn the features that were manually created in an effort to avoid detection.



Malicious Url Detection

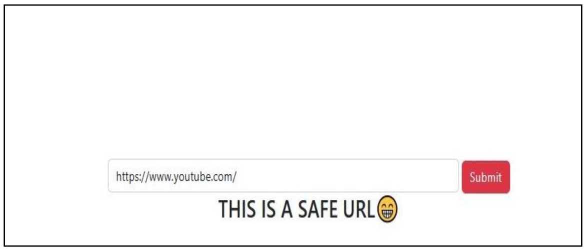
Fig11: Home Page



https://www.youtube.cam

BEWARE! IT'S A MALICIOUS URL☠

Fig 12: Detecting malicious URL



https://www.youtube.com/

THIS IS A SAFE URL😁

Fig 13: Detecting safe URL

VI . CONCLUSION

Its use of malicious websites that appear to be legitimate web pages and URLs is a typical social engineering tactic. In order to predict harmful websites, The goal of this project is to apply the available data sets to train deep neural networks and machine learning models. The targeted URLs and the website's content-based functionality are collected from a dataset containing both the harmful and benign URLs of the website. Measure each model's performance level and make comparisons. In order to more precisely anticipate dangerous URLs, In this research, machine learning and deep learning techniques are combined.

REFERENCE

[1] D. J. Lemay, R. B. Basnet, and T. Doleck, “Examining the relationship between threat and coping appraisal in phishing detection among college students,” Journal of Internet Services and Information Security, vol. 10, no. 1, pp. 38–49, 2020.

[2] H. Kim, “5G core network security issues and attack classification from a network protocol perspective,” Journal of Internet Services and Information Security, vol. 10, no. 2, pp. 1–15, 2020.

[3] K. Aram and J. O. SoK, “A systematic review of insider threat detection,” Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), vol. 10, no. 4, pp. 46–67, 2019.

[4] R. B. Basnet and R. Shash, “Towards detecting and classifying network intrusion traffic using deep learning frameworks,” Journal of Internet Services and Information Security, vol. 9, no. 4, pp. 1–17, 2019.

[5] F. Valenza and M. Cheminod, “An optimized firewall anomaly resolution,” Journal of Internet Services and Information Security, vol. 10, pp. 22–37, 2020.

[6] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, “An empirical analysis of phishing blacklists,” in Proceedings of Sixth Conference on Email and Anti-Spam (CEAS), 2009.

[7] C. Seifert, I. Welch, and P. Komisarczuk, “Identification of malicious web pages with static heuristics,” in Conference on Telecom Networks and Applications, 2008. ATNAC 2008. Australasian. IEEE, 2008, pp. 91–96.

[8] S. Sinha, M. Bailey, and F. Jahanian, “Shades of grey: On the effectiveness of reputation-based “blacklists”,” in Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on. IEEE, 2008, pp. 57–64.

[9] Leo Breiman.: Random Forests. Machine Learning 45 (1), pp. 5- 32, (2001).

[10] D. Sahoo, C. Liu, S.C.H. Hoi, “Malicious URL Detection using Machine Learning: A Survey”. CoRR, abs/1701.07179, 2017.

[11]R.J.Vidmar.(August1992).Ontheuseofatmosphericplasmasas electromagnetic reflectors. *IEEETrans. PlasmaSci.*[Online].21(3).pp. 876-880. Available: http://www.halcyon.com/pub/journals/21ps03-vidmar