ISSN: 2278-4632
Vol-14, Issue-1, January: 2024

CYBER THREAT PREDICTIVE ANALYSIS FOR IMPROVING SECURITY

Mogilipuram Sravani
Master of Computer Application
Chaitanya Bharathi Institute of Technology (A)
Hyderabad, Telangana, India
sravani.mogilipuram@gmail.com

Srinivasa S. P. Kumar
Master of Computer Application
Chaitanya Bharathi Institute of Technology (A)
Hyderabad, Telangana, India

ABSTRACT

As technology increases day by day, there is an increase in cyberattacks around us. This results in various threats that affect the confidentiality of data putting a risk at the sensitivity of personal information. Some of the threats are phishing, ransomware, spyware, etc. have affected many individuals and organizations. In this study, we used Cyber Threat Intelligence (CTI) data for ML prediction for overall security improvement. Using the malware prediction dataset, determine whether the system is vulnerable to malware attack or not. Phishing attacks are used to steal end-user's information. The intruder can even send malicious codes, such as malware, into the system in different ways, so the system can be damaged. Thus, using different ML techniques, we can predict threats based on different data collected as input to ML algorithms such as SVM, DT, DF, Adaboast, and XG Boost which are used to predict cyber threats. The best accuracy score is 98% which is gained using classification algorithms such as Random Forest and Decision Tree for detection of phishing attacks. Resulting in identifying the threats and taking countermeasures accordingly to improve security of the organization.

KEYWORDS:
Cyber threats, Cyber Threat Intelligence (CTI), Cyber Supply Chain (CSC), phishing, malware.

I. INTRODCUTION

Securing the data from the attackers has always been challenging. There are different domains of attack, such as software, hardware, communication, supply chains, social engineering, etc. [1] The recent data branch attack happened on the "MOVEIts"app. The NCSC report stated that the stolen information relates to employees at a number of organizations, which include British Airways, the BBC, Ofcom, etc. Later, ransom was demanded by the hackers. [2] Similarly, many zero-day exploits occur by groups of people where organizations are demanded with huge ransoms at the cost of their employee's personal data. Phishing attacks are one of the common techniques where, the users think the URL, email, text messages are from trusted sources. When a user clicks on the link, malicious code is automatically downloaded, which causes harm to the system. XSS and SQL injections are web security vulnerabilities that can access user's data. This paper aims to improve the security of threats such as malware, phishing, XSS, and SQL injection using different ML algorithms to predict the type of attack by which the system is affected. We applied different ML algorithms to the different datasets for the attack prediction and found that the XG Boost has the highest accuracy among all the algorithms.

II. LITERATURE SURVEY

The existing work was used to enhance Cyber Supply Chain (CSC) security with consideration of different aspects such as Cyber Threat Intelligence (CTI), tactics, techniques, and procedures, cyber threat analysis, and much more. The approach covered both inbound and outbound supply chain factors, which helped in identifying the vulnerabilities easily. It was classified into four phases: determining strategy, threat analysis, threat prediction, and the control phase.

A. CYBER THREAT INTELLIGENCE (CTI)

To control the increase in cyber-attacks, companies in recent years have begun investigating Cyber Threat Intelligence (CTI). Mainly, the organizations have traditionally collected and analyzed data from systems such as Security Information and Event Management Systems (SIEMS). [3] The organization uses Open-Source Intelligence (OSINT) [4] and different public sources for alerting CTI to protect against unaware threats. It uses different online tools to check whether the emails and URLs are safe to use. For example, we use the "haveIPwned" website to check whether the email is hacked or not.

Copyright @ 2024 Author

ISSN: 2278-4632
Vol-14, Issue-1, January: 2024

B. TACTICS, TECHNIQUES, AND PROCEDURES

Organizations frequently do security audits and assessments to check risks and vulnerabilities. The company uses the National Institute of Standards and Technology (NIST) [5], standard ISO 27001. The open web application security project (OWASP) [6] and others should follow the guidelines for security. Organizations even use risk mitigation methods where an external society audits the organizations to identify vulnerabilities and loopholes where they should provide more security.

C. CYBER THREAT ANALYSIS

In the cyber supply chain (CSC), mostly the analysis of threats is done by using various steps that are
(i) Information Gathering: gathering information about the information, such as IP addresses, searches the organization's websites to find the vulnerable spots, where it contains two phases: active information gathering and passive information gathering.
(ii) Experimentation or scanning: the attacker uses penetration testing and vulnerability assessment measures to explore vulnerable spots. For instance, Nmap uses Nmap to check different open ports and vulnerabilities to gain access to a specific port.
(iii) Exploit: When the vulnerabilities are found, he tries to gain access to the system. Later, manipulate the user's or organization's data for a specific motive.
(iv) Command and Control: Attackers may change the user credentials and try to inject malicious codes. Most organizations are attacked accordingly, using different methods and procedures to gain access to most of the weak and repeated systems in a similar pattern. [7]

D. USAGE OF AI IN CYBER THREATS

Artificial Intelligence (AI) can be utilized by actions to enhance cyber-attacks, boost scanning for vulnerabilities, etc. AI can help in pattern recognition of different attacks, which helps in building Advanced Persistent Threats (APIs). It can use large datasets to develop different techniques that can break into systems easily. It can also be used to analyses huge amounts of data for different phishing and social engineering attacks. Many of them can be used for creating adaptive malware that can change its behavior to enter the system.

AI can be used for improving incident response; algorithms can help in threat detection and network security. AI can improve Intention Detection Systems (IDS) by learning about network traffic and identifying security breaches. It helps to detect and respond known patterns of attacks that are already recognized and fed as data for the algorithm.

E. MACHINE LEARNING IN CSC SECURITY

Machine learning has been used in many security concepts, such as Intrusion Detection systems (IDS), spam filters, and antivirus, to predict threats. There are many ML methods and different data mining techniques in cybersecurity applications [8],[9]. Most of the dataset is in the form of a sequential learning method so that we can use specified label values as input characteristics and understand more correlation among them accordingly. There are many neural networks, decision trees, and SVM algorithms implemented so we get better accuracy.

III. METHODOLOGY

The proposed approach aims to identify the vulnerabilities using CTI and machine learning. This study explains to predict the threats from reference to existing dataset. Later the features are selected by hyper parameter optimization. Optimization technique is used for selecting best hyperparameter for the ML algorithm.

They are various optimization techniques in this study we used Random Search and Bayesian optimization. Bayesian optimization has given the best results for the malware prediction. Later the model is trained with the best features selection and analysis of threats takes place.

A. PHISHING DETECTION

There are various phishing methods, such as whitelisting, blacklisting, context, visual similarity, and URL-based [10]. The present model is context-URL based. In the context-URL based where the details of the webpage are extracted.
Step 1: The various URLs are collected from the dataset which are labelled as good /bad.
Step 2: With the help of web scraping, we can get the contents of the URLs and then perform the feature engineering process. Web scraping is the process to retrieve the data from the webpage which is in the form of Html, CSS, JavaScript into user required form such as csv, file, excel.

Step 3: Different HTML tags are extracted from the URLs using the bs4 module. Some important features like, has_link, has_password, has_hidden_element, and has_input are taken into consideration when the dataset is created.
Step 4: Different machine learning algorithms, such as AdaBoost, Random Forest, Support Vector Machine (SVM), and KNeighbors, are applied to the dataset.
Step 5: Finally, the output is predicted to check if the URL is phishing or legitimate, meaning that it is safe to use. Figure 2 describes the detailed step by step process for the phishing detection.

Copyright @ 2024 Author

ISSN: 2278-4632
Vol-14, Issue-1, January: 2024

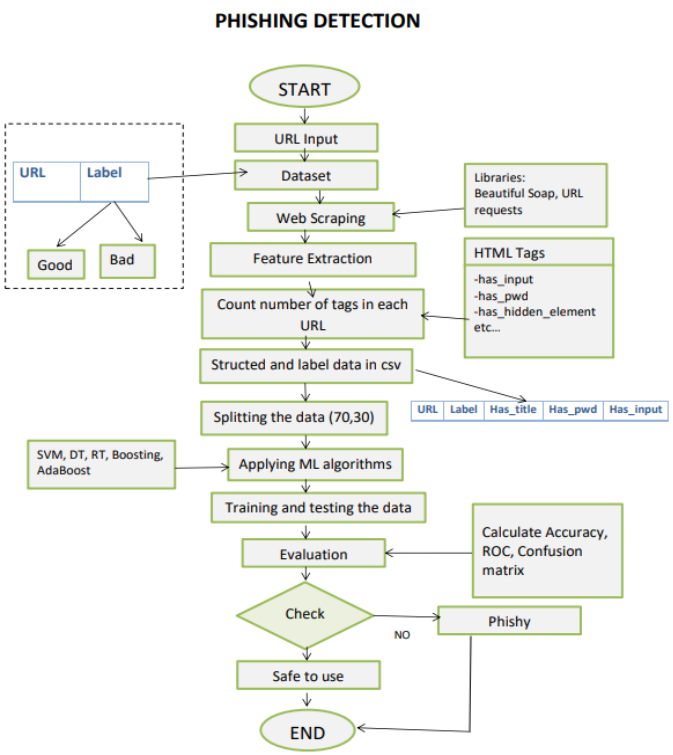

Figure 2. Step by step process for phishing detection.

B. MALWARE PREDICTION

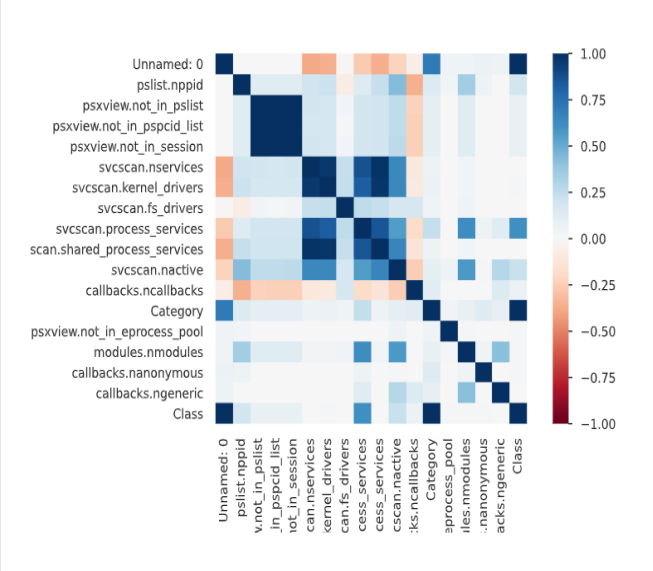

Figure 3. Correlations among various data for malware prediction.

Using the malware dataset [11] to predict the type of malware to which the system is vulnerable, such as spyware, ransomware, and Trojan horses

Step 1: Data analysis is done using the profile report from the Pandas module. The report examines various aspects such as correlations, variable types, and dataset statistics. Figure 3 shows the correlations among various data fields.
Step 2: A data cleaning method is used to remove unwanted data and missing data from the dataset. Then Random Search and Bayesian optimization techniques are applied using pipeline. Thus the result where as follows:

- Random Search: classifier_learning_rate=0.89, max_depth=5, n_estimators=400.
- Bayesian Optimization: classifer_learning_rate_0.93, classifier_n_estimators=222, max_depth=7.

Step 3: Feature selection is done, where the important features are chosen. As a result, 15 columns are selected, and unwanted columns are discarded. Features such as pslist.nproc, pslist.avg_threads, pslist.nprocs64bit, dlllist.ndlls are taken into consideration.
Step 4: A 5-fold cross-validation is applied to the dataset. Later, various machine learning algorithms are used, including gradient boosting, random forest, and XG boost are applied.
Step 5: Training and testing the dataset with the input values, the type of malware is predicted. As Figure 4 describes the process of malware detection.

Copyright @ 2024 Author

ISSN: 2278-4632
Vol-14, Issue-1, January: 2024

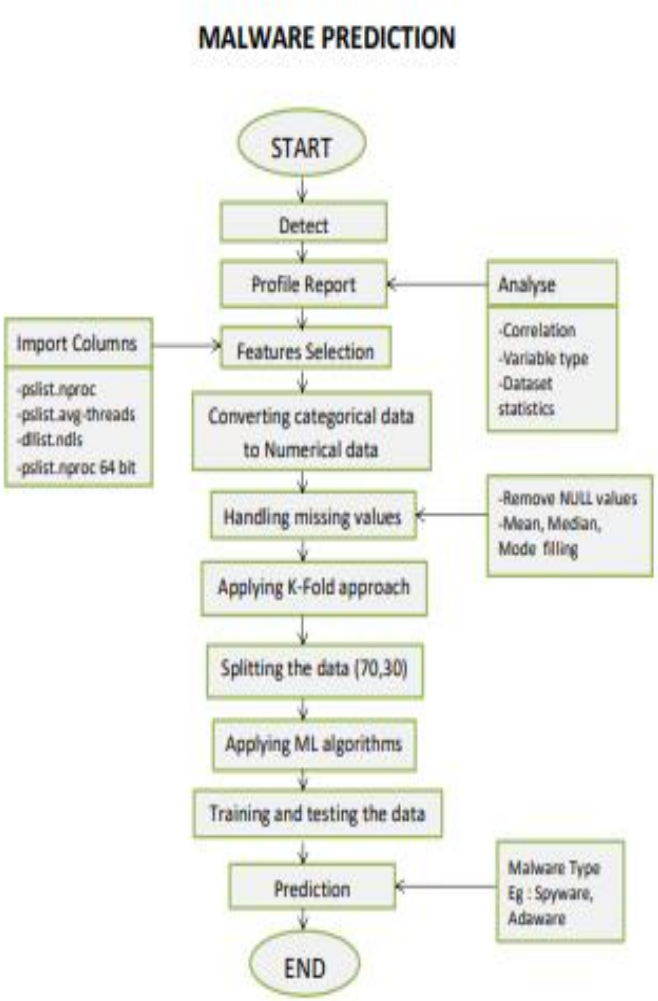

Figure 4. Step by step process for malware detection.

IV. APPLIED MACHINE LEARNING (ML) METHODS

Different machine learning algorithms are applied for phishing detection and malware detection, so we can achieve a better prediction of value to check whether malware or phishing is detected.

The algorithms that are used to train the dataset are

- **A decision tree (DT)** is part of concept learning. It divides the root node into sub-nodes until the values give one specific output, which depends on the features. The features can be spit out using the information gained by calculating the weights of each column.

- **A random forest (RF)** is the combination of two or more decision trees. It is classified in two ways: regression and classification. The regression method uses individual trees that are averaged, and the output is given accordingly. Whereas in classification, the majority vote will be output.

- **Support Vector Machine (SVM)** is a supervised learning algorithm. It creates a hyperplane that is used in the classification of different classes.

- **Naive Bayes** is also known as a probabilistic classifier since it is based on Bayes' Theorem. It is classified into three types, which are

  a. **Gaussian Nave Bayes (GaussianNB)**: Gaussian distributions—normal distributions, continuous variables
  b. **Multinomial Nave Bayes (MultinomialNB)**: multinomial distributions of discrete data, such as frequency counts.
  c. **Bernoulli Nave Bayes (BernoulliNB)**: Boolean variables—such as True and False or 1 and 0.

In the above paper, we use the Bernoulli nave bayes.

Ensemble learning is divided into three subsections: boosting, bagging, and stacking. Boosting and bagging are homogenous weak learners, whereas stacking is a heterogeneous solid learner.

Bagging is used to split the dataset into smaller, individual parts. Later, the algorithms are applied to each small dataset. Then the average of all the outputs is taken. Thus, the final result is obtained.

Boosting is used to correct the errors. The errors of the dataset after training are again sent as input with the subset of the sequential dataset. It is mainly classified into three types of boosting.

- **Ada Boost** is used for solving dynamic allocation problems. It helps a weak classifier perform better than random guessing of outputs. It gives higher accuracy by increasing the weights of incorrect samples, so the next iteration can focus more on these weighted samples. [14]

- **Gradient Boosting** is used by joining several weak learners with strong learners. The optimisation is done by differentiable loss functions. Different loss functions are mean squared error and cross-entropy. It is providing better accuracy. Thus, it is used for both regression and classification. It is used in different fields, like web search and ecology.

- **XGBoost** is an updated version of the gradient boosting algorithm. It is used because it is flexible, portable, and efficient. It uses decision trees as base learners and parallel tree boosting for increased speed and accuracy. Thus, it is used to solve many problems; for example, the Kubernetes XGBoost operator is designed for scheduling and monitoring.
  - Approximate Greedy Algorithm: instead of assigning every total weight split, this algorithm employs weighted quantiles to find the node that can split to get better performance.

Copyright @ 2024 Author

- Cash-Aware Access: It uses cache memory for accessing of data.
- Sparsity: It calculates Gain value with the help of observations that contains missing values with respect to left node where there is change in data,

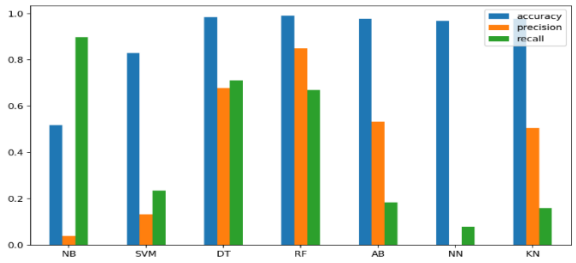

Figure 5. Comparison of various machine learning algorithms.

V. RESULT ANALYSIS

Confusion matrix is used for measurement performance. It helps in calculating 4 different types of predictions. That is true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). To calculate confusion matrix, we use actual and predicted values. For example, for a binary problem for phishing emails let us consider URL is spam to be 1 and URL not spam shall be 0. We can calculate the confusion matrix accordingly:

- True Positives (TP): The instances where the actual and predicted values are 1 (11).
- False Positives (FP): The instances where the actual value is 0 and predicted value is 1 (01).
- True Negatives (TN): The instances where the actual value and predicted value is 0 (00).
- False Negatives (FN): The instances where the actual value is 1 but the predicted value is 0 (10).

The figure 6 describes the confusion matrix for the phishing detection. The figure contains total 3011 sample data from the dataset. It helps in calculating different matrix accordingly.

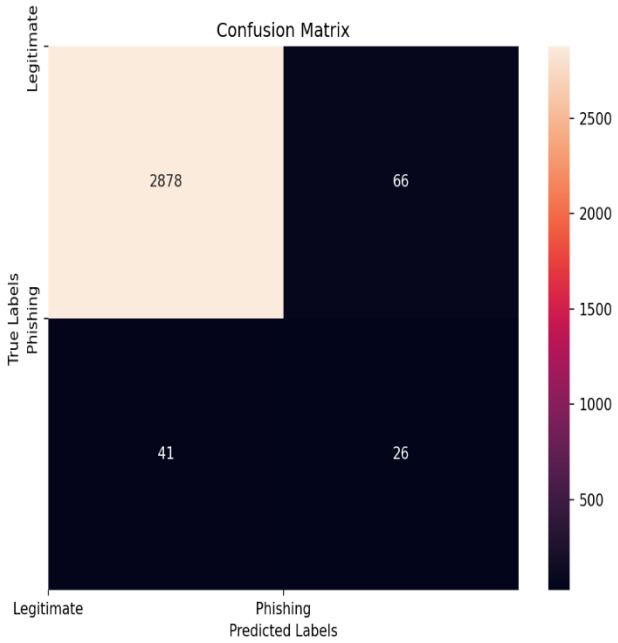

Figure 6. Confusion matrix for phishing detection.

Different metrics such as accuracy, precision, recall is used to calculate the best ml algorithm for the dataset. To calculate it we use 4 values that is True positive (TP), true negative (TN), false positive (FP), and false negative (FN)
Formulae for accuracy, precision, recall are

Accuracy is the proportion of accurately predicted data points among all the data points is known as accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (1)$$

Precision is ratio of accurately categorised positive samples (True Positive) to all mistakenly or correctly classified positive samples.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall is the proportion of Positive samples that were properly identified as Positive to all of the Positive samples

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

This section discusses the results of the threat prediction. Where it takes into consideration many variables and categorical data. The results often focus on phishing and malware detection.
Phishing attack's accuracy, score, and recall of various machine learning algorithms are given below in the table. From the table, we can see that the random forest and AdaBoost give better accuracy and precision than other machine learning algorithms. Thus, when the

Copyright @ 2024 Author

ISSN: 2278-4632
Vol-14, Issue-1, January: 2024

URL is passed to predict, it determines whether the URL is phished or legitimate.
Malware attack's accuracy rate is 95% for detecting whether it is malware or benign. Whereas for the category detection its accuracy rate is 75%. It is divided into 16 different malware such as ransomware, spyware, trojan horse etc.

In the future, the model can be enhanced by including more details about cyber threats and additional information about the attacks to improve accuracy. There can be further studies about cyber threat

Algorithm	Accuracy	Precision	Recall
Random Forest	98.95%	81.39%	66.38%
Decision Tree	98.37%	68.71%	67.75%
AdaBoost	97.89%	93.75%	68.44%
Support Vector Machine	91.40%	92.50%	69.38%
K-Neighbors Classifier	95.36%	93.33%	75.71%

Table 1: The performance evaluation metrics for the ML models in phishing detection.

intelligence (CTI) using AI. That is the automatic detection of threats in the organization when any suspicious activity has occurred. We can implement different big data approaches for a better understanding of the dataset. There are huge amount of data that can be easily analyzed using various methods.

Different ML algorithms can be used for the prediction of better scores and accuracy. Use of the latest technologies such as block chain technology, which helps in better analysis and prediction purposes. For user credentials, we can use quantum technology, which helps increase protection from different password cracking methods.

IV.CONCLUSION AND FUTURE SCOPE

Cyber threats predictive analytics helps in improving security by detecting the threats that can harm the security methods of the systems. The aim of this paper is to identify common threats that can harm the systems with different algorithms and doing a comparison on various ML algorithms to check the best accuracy rate to help predict the better algorithm. As the detection of the threats takes much more time in manual mode compared to automation, The above paper helps us automate a few sections to identify the previous attack patterns as we train the algorithms with datasets.

With the predictions, we can easily take countermeasures depending on the level of danger posed by the threats. We can also use emerging technologies such as block chain and quantum mechanics. Thus, identifying the threats is much easier when automated, and appropriate precautions should be taken, such as regular updates of the software improve security and vulnerability risks.

In the future, the model can be enhanced by including more details about cyber threats and additional information about the attacks to improve accuracy. There can be further studies about cyber threat intelligence (CTI) using AI. That is the automatic detection of threats in the organization when any suspicious activity has occurred. We can implement different big data approaches for a better understanding of the dataset. There are huge amount of data that can be easily analyzed using various methods.

Different ML algorithms can be used for the prediction of better scores and accuracy. Use of the latest technologies such as block chain technology, which helps in better analysis and prediction purposes. For user credentials, we can use quantum technology, which helps increase protection from different password cracking methods.

REFERENCES

[1] CAPEC - CAPEC-3000: Domains of Attack (Version 3.9). (mitre.org)

[2] MOVE it vulnerability and data extortion incident. (NCSC.GOV.UK)

[3] Security Information and Event Management (SIEM): Analysis, Trends, and Usage in Critical Infrastructures. https://doi.org/10.3390/s21144759

[4] OSINT Framework. https://osintframework.com/

[5] NIST Cybersecurity. https://www.nist.gov/cybersecurity

[6] OWASP - Attacks. https://owasp.org/www-community/attacks.

[7] Cyber Threat Predictive Analytics for Improving Cyber Supply Chain Security.

[8] E. G. V. Villano, "Classification of logs using machine learning," M.S. thesis, Dept. Inf. Secur. Commun. Technol., Norwegian Univ. Sci. Technol., Trondheim, Norway, 2018.

[9] R. C. B. Hink, J. M. Beaver, M. A. Buckner, T. Morris, U. Adhikari, and S. Pan, "Machine learning for power system disturbance and cyber-attack discrimination," in Proc. 7th Int. Symp. Resilient Control Syst. (ISRCS), Denver, CO, USA, Aug. 2014, pp. 1–8, doi: 10.1109/ISRCS.2014.6900095.

[10] A Survey of URL-based Phishing Detection.

[11] Kaggle - Starter Cyber Attacks 2010-2018. https://www.kaggle.com/code/kerneler/starter-cyber-attacks-2010-2018-072b39ce-f/notebook

[12] Nmap - Free Security Scanner for Network Exploration & Security Audits. https://nmap.org/

[13] BeautifulSoup4 - PyPI. https://pypi.org/project/beautifulsoup4

Copyright @ 2024 Author

ISSN: 2278-4632
Vol-14, Issue-1, January: 2024

[14] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997. [Online]. Available: https://www.ee.columbia.edu/~sfchang/course/svia/papers/freund95decisiontheoretic-adaboost.pdf

Copyright @ 2024 Author