SVM and Random Forest Algorithm are used well to detect phishing attacks for enhanced accuracy.

P.Ganesh [1], S. Kalaiarasi [2*]

[1]Research Scholar, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Science, Saveetha University, Chennai, Tamilnadu. India. Pincode: 602105

[2*]Project Guide, Corresponding Author, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Science, Saveetha University, Chennai, Tamilnadu. India. Pincode: 602105

ABSTRACT

AIM: The desire is to predict phishing attacks to obtain sensitive information from innocent users using SVM compared with Random Forest algorithms to improve accuracy. **Materials and methods:** Two groups such as SVM compared with Random Forest algorithms are applied. The total number of samples examined using this suggested technique is 110 pictures. Here in this sample dataset, 125 photos (70 percent) were used as a training sample and 30% as a testing dataset. Experiments were carried out for N=20 iterations for the SVM and Random forest algorithm. Computation processes were executed and verified for exactness. SPSS was used for predicting significance value of the dataset considering G-Power value as 80%. **Results:** Random forest algorithm demonstrates a high accuracy for Phishing attack detection, and significantly higher recognition rate of 90.54 (p=0.043). **Conclusion:** This article focussed on machine learning methods that are used to detect phishing attacks. The descriptive statistical findings reveal that the Random forest algorithm detects phishing assaults more accurately than the SVM.

Keywords: Machine Learning, SVM, Random Forest Algorithm, URL, Website, Phishing Attack.

INTRODUCTION

The method in which stealing of personal data such as login id, passwords by the attackers is called a phishing attack. Meanwhile, there are some inescapable security issues on the Internet, such as phishing, harmful software, and privacy revelation, that have already gravely affected users. Phishing is not restricted to traditional channels such as email, Text, and pop-ups (Sindhu et al. 2020). While the advent of mobile internet and social media provided consumers with ease, they were also used to spread phishing, such as QR code phishing (Xu 2017). The main advantages of the Novel Random Forest algorithm are it can perform both regression, classification tasks and it can handle large datasets. Google maintains a continually updated blacklist of harmful websites (Spirin 2014). Users can use Google's Private Browsing mode APIs to validate the security of URL links. The reasoning for this is that harmful URLs or site pages have features that set them apart from legal websites. The applications of novel random forest algorithms include banking information and health care (Hou et al. 2017).

The number of exploration papers published in IEEE Xplore is 10, while Google researchers discovered around 1200 publications. Current Conventional Machine Learning Methods for site

Detection extracts statistical characteristics from URLs and Server or extracts relevant characteristics from web pages such as layout, CSS, text and then classifies these characteristics (Tu 2020). Sites generally replicate authentic website URLs by modifying or adding certain characters (Kadam 2021). The Internet also brings unavoidable security vulnerabilities, such as phishing, malware, and privacy disclosures, which have seriously threatened users (Mulligan 1999). Phishing is no longer restricted to traditional means like email, SMS messaging, and popups. While an increase from mobile internet and social media provides facilities for users, it is also used to spread scams (Patil, Rane, and Bhalekar 2017). Google provides blacklists of malicious websites which are consistently updated. Phishing is no longer restricted to traditional means like email, SMS messaging, and popups. This is because malicious URLs or web pages have certain distinguishable characteristics. The best method to protect ourselves from these phishing assaults is to see one of them (Christodorescu et al. 2007). (Bhavikatti et al. 2021; Karobari et al. 2021; Shanmugam et al. 2021; Sawant et al. 2021; Muthukrishnan 2021; Preethi et al. 2021; Karthigadevi et al. 2021; Bhanu Teja et al. 2021; Veerasimman et al. 2021; Baskar et al. 2021)

To overcome the drawbacks of the blacklisted heuristic method based method, many security researchers have now turned to machine learning techniques. Many algorithms in machine learning use historical data to make decisions or predictions about future data. The algorithm uses this method to reliably discover sites, including zero-hour sites, by analyzing multiple banned and valid URLs and their functioning (Furnell and Dowland 2010). The objective is to improve theaccuracy in the recording of phishing attacks detection by using Machine Learning Analytics.

MATERIALS AND METHODS

The suggested work is studied at the Image Processing Laboratory, Department of Computer Science and Engineering, Saveetha School of Engineering. The number of study groups is 2. Random Forest is in group one, and SVM is in 2 groups. The sample size was generated using clinical analysis, with 10 sample sizes estimated per group and a total of 20 samples using the sample size calculator, with a confidence interval of 90.54, an 80 percent pretest power, and a ratio of 1 as the enrollment ratio.

Random Forest Algorithm

In bootstrapping, the attributes and the data sample are randomly selected by substitution to form a single tree. The random forest algorithm selects the correct ones for similar classification and decision tree algorithms. Each tree in the forest predicts a target value, then an algorithm calculates the value for each predicted target. Rather than using an individual tree that contains all of the data and features, this strategy uses a random selection of attributes and training data from the given sets to produce a series of tree models. The output of the random forest is calculated based on the results of the decision trees. Each model gives a classification for the new data.

SVM

The SVM should give a regularization coefficient each time. The SVM technique includes a dependent variable that can be

given in binary values which means output. For example, it can apply when we need to determine the probability of positive or no events. The same technique is used here with the additional sigmoid function. SVM are the most commonly used basic forms of regression. SVM is used while the structured variable is non-stop and the character of the road regression is linear. Regression is a method for predicting the value of a numeric response variable from one or more predicted factors. There are different forms of regression such as linear, multiple, logistic, polynomial, etc. SVM is a statistical analysis technique used in machine learning to predict the value of data based on previous observations in a data set. The better the algorithm should predict the rankings within the data sets. SVM may also help with data preparation by allowing data sets to be saved in a preset repository during the extract, transform, and load process. This model allows us to get a probability that users of a particular website will click on certain ads using machine learning.

Statistical Analysis

SPSS software version 21 The dataset is created using a variety of samples, and the accuracy values acquired from the dataset are used by both methods. The categories are labeled as accurate and are noted in the statistical tool's data and variable displays. The data is an independent variable and accuracy is a dependent variable.

RESULTS

Table 1 shows the comparison of Descriptive Statistical analysis between Random Forest Algorithm and svm with Minimum, maximum, mean, standard Deviation. We discovered that our model predicts the outcome of the final result extremely well after examining the results. Table 2 illustrates the Group Statistical Analysis of the Random Forest Algorithm, which has a mean accuracy of 90.54 and SVM has an accuracy of 86.50. Table 3 displays the Independent sample analyses, which compares mean accuracy and homogeneity. Significant value for homogeneity is 0.043. The findings clearly show that the Random forest algorithm performs better than the svm since it has more accuracy.

Figure 1 shows mean accuracy between Random Forest Algorithm and SVM. The findings clearly show that the Random Forest Algorithm, with the greatest accuracy value of 90.54, outperforms the SVM, which has a low accuracy score of 86.50.

DISCUSSION

This proposed model exceeds the results of the students during 3 semesters. All events benefit from the training and validation suite's optimal outcomes and accuracy. Finally, the model achieved the perfect result and the best precision with the Random Forest algorithm and SVM with an accuracy of 90.54 and 86.50. If all variables are considered, then the precision might be greater.

When compared to earlier efforts, this model allows us to get the greatest results (Cui 2019). Any other research article that does not conflict with the result of our finding(Sonowal 2021). Here we have done comparative study between novel random forest algorithms and SVM (Vrbančič, Fister, and Podgorelec 2020). The novel random forest algorithm accuracy value will be close to 100% for any ideal detection algorithm (Timm and Perez 2010). Our goal is not only to

identify websites but also to improve their performance targeted domains (Silva and da Silva 2017).

This model's shortcoming is that it only works for one curriculum. Tapping on a link in an email can give up the information and arrangement of an association to an attacker. The input and output data will also change.

CONCLUSION

In this research, it shows the accuracy of the Random Forest algorithm has better mean accuracy approximately 90.54 in comparison with SVM which has accuracy of 86.50 which leads to convenient detection of phishing attacks. The review of prior research publications also demonstrates that the Random Forest method outperforms the SVM in terms of accuracy.

DECLARATION

Conflict of Interest

The author does not have any conflict of interest associated with this manuscript.

Authors Contribution

Author EC was involved in Image collection, data analysis and manuscript writing. Author SKA was involved in conceptualization, data validation and critical review of manuscript.

Acknowledgements

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS) (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding: We thank the following organizations for providing financial support that enabled us to complete the study.

1. HKM Info Systems Pvt.Ltd., Chennai
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha school of Engineering

REFERENCES

1. Baskar, M., R. Renuka Devi, J. Ramkumar, P. Kalyanasundaram, M. Suchithra, and B. Amutha. 2021. “Region Centric Minutiae Propagation Measure Orient Forgery Detection with Finger Print Analysis in Health Care Systems.” *Neural Processing Letters*, January. https://doi.org/10.1007/s11063-020-10407-4.
2. Bhanu Teja, N., Yuvarajan Devarajan, Ruby Mishra, S. Sivasaravanan, and D. Thanikaivel Murugan. 2021. “Detailed Analysis on Sterculia Foetida Kernel Oil as Renewable Fuel in Compression Ignition Engine.” *Biomass Conversion and Biorefinery*, February. https://doi.org/10.1007/s13399-021-01328-w.
3. Bhavikatti, Shaeesta Khaleelahmed, Mohmed Isaqali Karobari, Siti Lailatul Akmar Zainuddin, Anand Marya, Sameer J. Nadaf, Vijay J. Sawant, Sandeep B. Patil, Adith Venugopal, Pietro Messina, and Giuseppe Alessandro Scardina. 2021. “Investigating the Antioxidant and Cytocompatibility of Mimusops Elengi Linn Extract over Human Gingival Fibroblast Cells.” *International Journal of Environmental Research and Public Health* 18 (13). https://doi.org/10.3390/ijerph18137162.
4. Christodorescu, Mihai, Somesh Jha, Douglas Maughan, Dawn Song, and

Cliff Wang. 2007. *Malware Detection*. Springer Science & Business Media.
5. Cui, Qian. 2019. *Detection and Analysis of Phishing Attacks*.
6. Furnell, Steven, and Paul Dowland. 2010. *E-Mail Security*. It Governance Pub.
7. Hou, Su, Tianliang Lu, Yanhui Du, and Jing Guo. 2017. "Static Detection of Android Malware Based on Improved Random Forest Algorithm." *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*. https://doi.org/10.1109/isi.2017.8004913.
8. Kadam, Sonali. 2021. "Feature Based Phishing Website Detection Using Random Forest Classifier." *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2021.35400.
9. Karobari, Mohmed Isaqali, Syed Nahid Basheer, Fazlur Rahman Sayed, Sufiyan Shaikh, Muhammad Atif Saleem Agwan, Anand Marya, Pietro Messina, and Giuseppe Alessandro Scardina. 2021. "An In Vitro Stereomicroscopic Evaluation of Bioactivity between Neo MTA Plus, Pro Root MTA, BIODENTINE & Glass Ionomer Cement Using Dye Penetration Method." *Materials* 14 (12). https://doi.org/10.3390/ma14123159.
10. Karthigadevi, Guruviah, Sivasubramanian Manikandan, Natchimuthu Karmegam, Ramasamy Subbaiya, Sivasankaran Chozhavendhan, Balasubramani Ravindran, Soon Woong Chang, and Mukesh Kumar Awasthi. 2021. "Chemico-Nanotreatment Methods for the Removal of Persistent Organic Pollutants and Xenobiotics in Water - A Review." *Bioresource Technology* 324 (March): 124678.
11. Mulligan, Geoff. 1999. *Removing the Spam: Email Processing and Filtering*. Addison Wesley Longman.
12. Muthukrishnan, Lakshmipathy. 2021. "Nanotechnology for Cleaner Leather Production: A Review." *Environmental Chemistry Letters* 19 (3): 2527–49.
13. Patil, Prajakta, Rashmi Rane, and Madhuri Bhalekar. 2017. "Detecting Spam and Phishing Mails Using SVM and Obfuscation URL Detection Algorithm." *2017 International Conference on Inventive Systems and Control (ICISC)*. https://doi.org/10.1109/icisc.2017.8068633.
14. Preethi, K. Auxzilia, K. Auxzilia Preethi, Ganesh Lakshmanan, and Durairaj Sekar. 2021. "Antagomir Technology in the Treatment of Different Types of Cancer." *Epigenomics*. https://doi.org/10.2217/epi-2020-0439.
15. Sawant, Kashmira, Ajinkya M. Pawar, Kulvinder Singh Banga, Ricardo Machado, Mohmed Isaqali Karobari, Anand Marya, Pietro Messina, and Giuseppe Alessandro Scardina. 2021. "Dentinal Microcracks after Root Canal Instrumentation Using Instruments Manufactured with Different NiTi Alloys and the SAF System: A Systematic Review." *NATO Advanced Science Institutes Series E: Applied Sciences* 11 (11): 4984.
16. Shanmugam, Vigneshwaran, Rhoda Afriyie Mensah, Michael Försth, Gabriel Sas, Ágoston Restás, Cyrus

Addy, Qiang Xu, et al. 2021. "Circular Economy in Biocomposite Development: State-of-the-Art, Challenges and Emerging Trends." *Composites Part C: Open Access* 5 (July): 100138.

17. Silva, Jaime A. Teixeira da, and Jaime A. Teixeira da Silva. 2017. "Fake Peer Reviews, Fake Identities, Fake Accounts, Fake Data: Beware!" *AME Medical Journal*. https://doi.org/10.21037/amj.2017.02.10.
18. Sindhu, Smita, Sunil Parameshwar Patil, Arya Sreevalsan, Faiz Rahman, and Ms Saritha A. N. 2020. "Phishing Detection Using Random Forest, SVM and Neural Network with Backpropagation." *2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*. https://doi.org/10.1109/icstcee49637.2020.9277256.
19. Sonowal, Gunikhan. 2021. *Phishing and Communication Channels: A Guide to Identifying and Mitigating Phishing Attacks*. Apress.
20. Spirin, Vladimir. 2014. "Improving Email Security and Management." *Computer Fraud & Security*. https://doi.org/10.1016/s1361-3723(14)70482-8.
21. Timm, Carl, and Richard Perez. 2010. "Phishing Attacks." *Seven Deadliest Social Network Attacks*. https://doi.org/10.1016/b978-1-59749-545-5.00003-3.
22. Tu, Yuanfen. 2020. "Implementation of Improved SVM Algorithm in Tourism Economic Data Analysis." *2020 5th International Conference on Smart Grid and Electrical Automation (ICSGEA)*. https://doi.org/10.1109/icsgea51094.2020.00113.
23. Veerasimman, Arumugaprabu, Vigneshwaran Shanmugam, Sundarakannan Rajendran, Deepak Joel Johnson, Ajith Subbiah, John Koilpichai, and Uthayakumar Marimuthu. 2021. "Thermal Properties of Natural Fiber Sisal Based Hybrid Composites – A Brief Review." *Journal of Natural Fibers*, January, 1–11.
24. Vrbančič, Grega, Iztok Fister Jr, and Vili Podgorelec. 2020. "Datasets for Phishing Websites Detection." *Data in Brief* 33 (December): 106438.
25. Xu, Yulong. 2017. "Research and Implementation of Improved Random Forest Algorithm Based on Spark." *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)(*. https://doi.org/10.1109/icbda.2017.8078683.

TABLES AND FIGURES

Table 1. Comparison, Descriptive Statistical analysis between Random Forest Algorithm and SVM with Minimum, maximum, mean, standard Deviation.

	N	Minimum	Maximum	Mean	Std.Deviation
Groups	20	1	2	1.50	0.513
Accuracy	20	82	95	88.52	3.633

Table 2. Group Statistical analysis of Random Forest Algorithm (mean accuracy of 90.54) and SVM (mean accuracy of 86.50) with Sample size, Mean, Standard deviation, Standard Error Mean.

	Groups	N	Mean	Std.Deviation	Std.Error Mean
Accuracy	RF	10	90.54	3.101	0.981
	SVM	10	86.50	3.028	0.957

Table 3. Independent Samples T test for Random Forest Algorithm and SVM considering variance and statistical significance of 0.043 considering accuracy.

		Levene's Test of Equality		T Test for Equality of means					95% confidence into Differences	
		F	Sig	t	df	Sig (2-tailed)	Mean difference	Std Error Difference	Lower	Upper
Accuracy	Equal Variances assumed	0.004	0.043	2.950	18	0.001	4.043	1.370	1.164	6.922
	Equal Variances not assumed			2.950	17.990	0.001	4.043	1.370	1.164	6.922

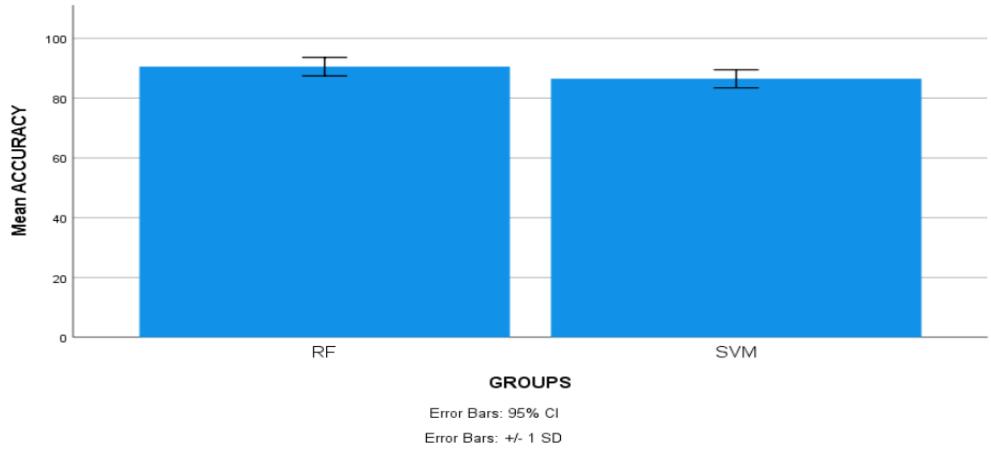

Fig. 1. Bar chart represents the comparison of mean accuracy of svm and Random forest algorithm. Comparison of Random forest algorithm and SVM in terms of mean accuracy. The mean accuracy of the Random Forest algorithm is 90.54 and SVM with 86.50. X-Axis: RF vs Y-Axis: SVM and the Mean accuracy of detection is ±1SD.