SN Computer Science (2022) 3:488
https://doi.org/10.1007/s42979-022-01387-4

ORIGINAL RESEARCH

Prediction of Phishing Websites Using Stacked Ensemble Method and Hybrid Features Selection Method

Mithilesh Kumar Pandey[1] · Munindra Kumar Singh[1] · Saurabh Pal[1] · B. B. Tiwari[2]

Received: 31 January 2022 / Accepted: 25 August 2022 / Published online: 25 September 2022
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd 2022

Abstract

Phishing is considered a big concern in this age of data and digital technologies because of its significant influence on the banking and online retailing industries. Cybercriminals target all economic activity on the Internet; thus, it is critical to take security precautions to safeguard assets. One of the first steps in constructing a safe cyberspace is to prevent phishing attacks before they happen. The detection mechanisms for these assaults were created using machine learning and other methods. However, there is still room for improvement in terms of detection accuracy. This paper proposes the optimization of an ensemble classification algorithm for phishing website (PW) detection. The suggested technique was optimised using a hybrid features selection method (Chi-square, extra tree, and heatmap) by modifying numerous machine learning (ML) method parameters, including random forest, naive Bayes, J48, and KNN. These were achieved by rating the optimal classifiers and selecting the top classifiers to serve as the foundation for the suggested technique. The obtained results by all experiments show that assigned optimized stacking ensemble approach outperforms previous ML-based detection methods. The level of precision attained was 99.7%.

Keywords Phishing websites · Random forest · Naïve Bayes · KNN · J48 · Stacked ensemble method and features selection methods: Chi-square, extra tree, and heatmap

Introduction

The Internet, covering a broad area of our daily lives, is an indispensable element. Many individuals use it for a variety of purposes, including shopping, bill payment, banking, and communication. Users suffer security issues as a result of increased usage, as well as in identifying theft, hacking phishing, and other cybercrimes. The most prevalent cybercrime assault is phishing. It is characterised as a social engineering technique used to trick customers into visiting phoney websites to steal sensitive details of customers such as bank details. People often fall for the information included in phishing emails and websites due to a lack of awareness, which is utilised by the attacker as a way of penetrating the user's privacy and obtaining critical information. This occurs when an attacker creates a phishing website that is so similar to legal websites that it is impossible for certain users to tell the difference. Sending an email with links to bogus websites that are identical to actual websites is one of the most prevalent strategies employed by fraudsters. They appear to be legitimate pages when they are opened, regarding details of bank account or check account regarding details [1].

This article is part of the topical collection "Advances in Computational Approaches for Artificial Intelligence, Image Processing, IoT and Cloud Applications" guest edited by Bhanu Prakash K N and M Shivakumar.

✉ Saurabh Pal
drsaurabhpal@yahoo.co.in

Mithilesh Kumar Pandey
mithileshkumarmca@gmail.com

Munindra Kumar Singh
munindra09_vbspu@yahoo.in

B. B. Tiwari
bbtiwari62@gmail.com

[1] Department of Computer Applications, VBS Purvanchal University, Jaunpur, Uttar Pradesh 222001, India

[2] Department of Electronics and Communication, VBS Purvanchal University, Jaunpur, Jaunpur, Uttar Pradesh 222001, India

In recent years, cybercrime has become a worry for many companies and academics. Phishing is a sort of cybercrime that is often regarded as one of the most dangerous. In phishing, the attackers steal the user's credentials and information by impersonating legitimate emails or websites. Because it impacts a large number of Internet users and companies, this form of assault has become a problem. In phishing, an attacker impersonates a certain organization's LW and distributes it to victims via fraudulent means. Bogus websites are opened by clicking on links in the emails [2].

Phishers are constantly refining and improving their methods for creating false websites that appear to be authentic. The problem is figuring out which bogus websites are being utilised in phishing scams. To avoid these assaults and safeguard user privacy and security, new and updated solutions must be developed. Phishing is a type of cybercrime that uses social engineering and technology tactics to defraud people. Its goal is to hurt consumers by stealing personal information, passwords, and bank account information. Phishing attempts often use social engineering to trick the target into clicking on a spoofed link that takes them to a false web page that looks just like the real one. As a result, instead of being forwarded to the specified website, motivated for more secure. A firewall and anti-virus software alone will not protect from an online phishing assault. Users lose millions of dollars each year as a result of this type of attack. From APWG's most recent data on phishing events, the number of complaints of mishaps reached 138,328 in the fourth quarter of 2018. Then, in the first half of 2020, it climbed by 15% to 162,155, indicating a 15% rise over the previous year. Furthermore, in 2019, phishing assaults were the most common web attacks. According to intelligence study, these assaults are predicted to continue to rise. Several machine learning algorithms have been presented to detect phishing websites automatically [3].

Related Work

Abdelfettah and Hassan suggested increasing the performance in identifying web pages and predicting phishing websites by adopting the GA approach. Although the performance of the GA-based URL detector was improved, the prediction time was quite long when dealing with a large number of URLs [4].

Rao and Pais used page elements such as logo, favicon, scripts, and styles. The technique used a server to update the page characteristics, which slowed down the detection system's speed [5].

Aljofey et al. proposed a CNN-based detection algorithm for detecting the phishing page. To discover URLs, a sequential pattern is employed. According to previous studies, CNN performs better when fetching pictures rather than text [6].

AlEroud and Karabatis developed a generative adversarial network to get around a discovery system. A KNN organized system can detect the impression of an unfavourable network [7].

Geo et al., for example, introduced a detective ensemble model that blended two models to build a new classifier that outperformed each model alone. Furthermore, a combined feature selection technique of phishing website detection that focussed on improving phishing site features was introduced to increase detection [8].

Jain and Gupta discussed about phishing website and suggested about malicious URLs, identified using both NB and SVM algorithms. Both SVM and NB are sluggish learners, who do not remember their past outcomes. As a result, the URL detector's efficiency may be diminished [9].

Purbay and Kumar observed multiple machine learning approaches to categorise URLs. They matched the execution of several types of machine learning algorithms. However, there were no comments concerning the algorithms' retrieval capabilities [10].

Gandotra and Gupta discussed multiple categorisation techniques to detect dangerous URLs. The results of the studies showed that the system performed better than other machine learning approaches. However, it has limitations when it comes to managing massive amounts of data [11].

Basit et al. proposed a unique ML ensemble technique for detecting phishing assaults. This model was created using the results of three different classifier combinations to improve detection. However, because phishing attacks are extremely risky and have devastating consequences for people and companies, the detection rate of phishing websites must be improved. As a result, implementing accurate, effective, and up-to-date phishing detection tools to protect against the phisher's adaptive approaches is critical. Other techniques were used to improve the accuracy of PW detection systems. In comparison to other recent methodologies, the results of these research showed that the recommended approaches improved significantly. However, there is still room for improvement in terms of detection accuracy [12].

Hung Le et al. organised a URL detector, based on deep learning. The authors claimed that the approach can extract information from URLs. Deep learning algorithms take longer to achieve results. It also parses the URL and compares it to the library to provide an output [13].

Hong et al. extracted URLs crawler from data repositories. Phishing websites were identified using a lexical characteristics technique. The crawler-based dataset was used to assess performance. As a result, there is no guarantee that the URL detector will work with real-time URLs [14].

Kumar et al. organised ML-based URL for shortlisted dataset. A lexical feature technique was also used to distinguish between dangerous and authentic websites. The

authors used an older dataset, which might lower the detector's performance when using real-time URLs [15].

The current work proposes an optimal ensemble classification approach for detecting PWs. Training, feature optimisation, and testing are the three primary steps in this process. The classifiers (RF, J48, KNN, and naive Bayes) were first trained, and no optimisation strategy was used in this stage. In the second stage, a hybrid features selection approach is utilised to optimise these classifiers that may be used to improve the classifiers' overall accuracy. Following that, depending on their ranking, optimised classifiers were used as the stacking ensemble technique basis classifiers. Finally, a test dataset is created by gathering new websites, which is then utilised to predict the websites' eventual class designation. The following is the study's structure. "Materials and Methods" provides associated literature. In "Conclusion", the approach and materials were not employed. In the next section, the results of the current study's experiments are reported. The findings are described and compared with related literature in the same section. The results and recommendations are reported in the final section, which summarises the current study's conclusion.

Materials and Methods

Data Preparation

The dataset from UCI repository for phishing websites was utilised to perform experiments and assess the efficacy of the suggested strategy in this study. We used the publicly available informative index for the execution and testing of our machine learning computation, which provides the following assets that may be used as contributions for model structure: a collection of site URLs for (11,430, 89) locations. Each example contains 89 site limits and a class name that indicates if the site is phishing or not (1 or − 1). There are 5715 phishing sites in the sample Fig. 1.

Methods

Random Forest

Random forest uses a supervised learning method to detect phishing websites. In this experiment, random forest predicted an ensemble classifier which integrated different weak decision trees. This algorithm is organised to enhance the average accuracy in various experiments. It depends on the majority of voting for strong prediction in various predictions [16].

Naïve Bayes

The Bayes' uses naive Bayes classifiers to detect phishing websites. Naïve Bayes algorithms always support an object likelihood and do not predict different ideas for each pair as dependent on others. Naïve Bayes well predicted by rapid machine learning algorithms [17].

J48

Phishing data overfitting is a problem encounter, in which decision trees provide balance to solve the problem. In this situation, the ID3 algorithm experiences data overfitting. The difficulty with decision trees is that they separate the data into pure sets. Pruning is directed to detect data overfitting in J48's extension. J48 is an ID 3-based machine learning decision tree classification technique. It is quite useful for categorising and continually examining data. J48 always generates new patteren in each experiment with correct observation [18].

KNN

There is no such thing as the best classifier; it all relies on the situation and the type of data or problem at hand. Because it does not generalise across data in advance, kNN is sluggish when there are a number of observations. Instead, it reads the historical database each time a prediction is needed, as mentioned. The qualities of KNN are that it is a non-parametric technique and a sluggish learning algorithm. Because it simply saves the stage, the algorithm requires practically no time to consider. After that, the saved data is utilised as a new observation point. There is no such thing as the optimal classifier; it is always dependent on the situation and the type of data or problem at hand. Because it does not generalise across data in advance, kNN is sluggish when there are a lot of observations. Instead, it reads the historical database each time a prediction is needed, as mentioned [19].

Proposed Method

The proposed approach was as follows: a cross fold was applied to the dataset with ten folds ranging from 1 to 10. Heat matrix feature optimisation is achieved as a result of the resultant subsets. The phishing websites were then classified using four different classifiers: J48, KNN, NB and RF, naive Bayes, J48, and KNN. The enhanced accurate rate, F-measure and precision were used to determine the best classifier for detecting phishing websites in Fig. 2.

Following the identification of key traits, the ML systems could be trained to detect whether or not the site was real. As one rises, the other rises with it, and the closer one gets to 1, the stronger does the bond become. Overlaying a

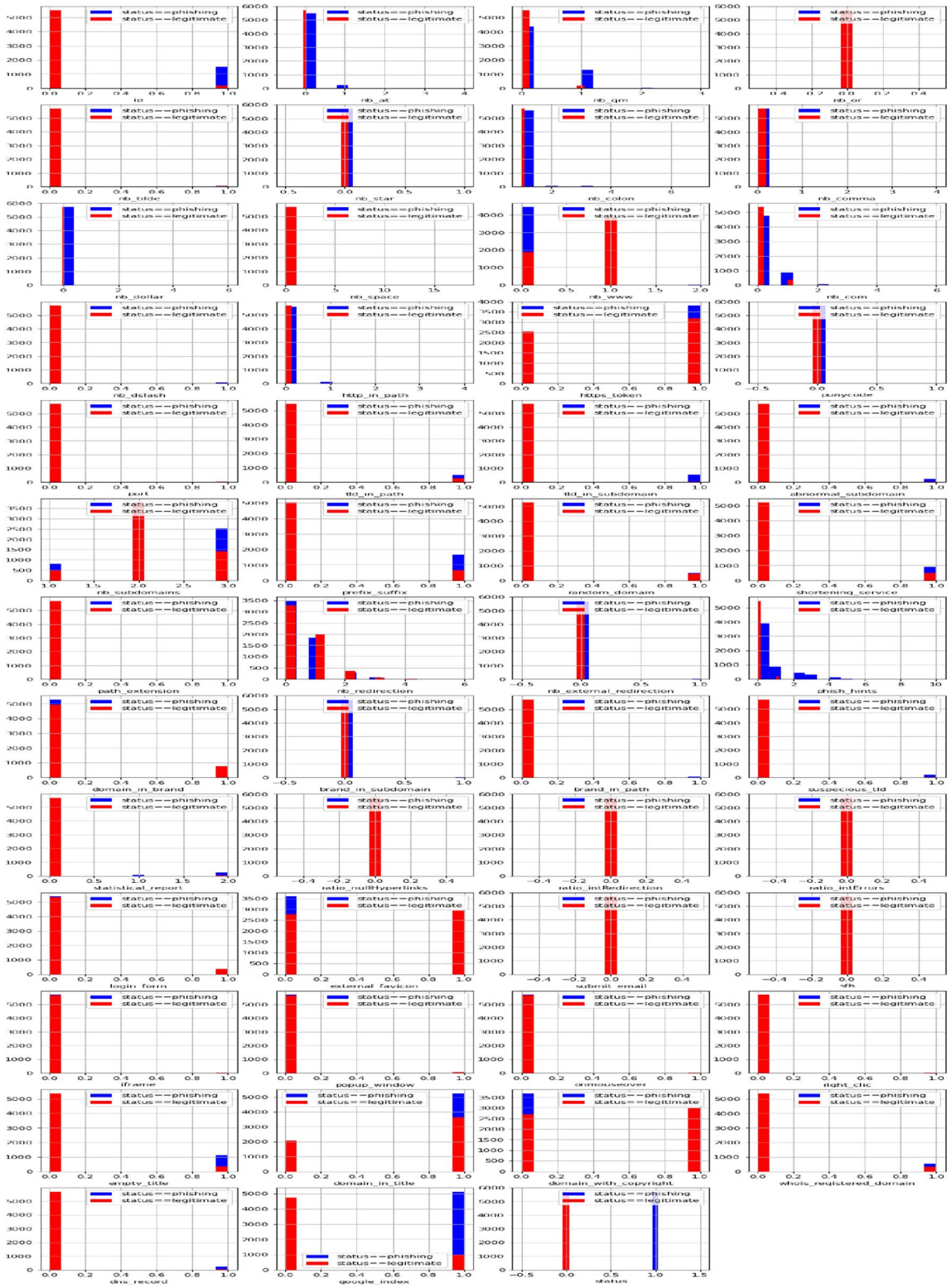

Fig. 1 Graphical representation of phishing websites attributes by histogram

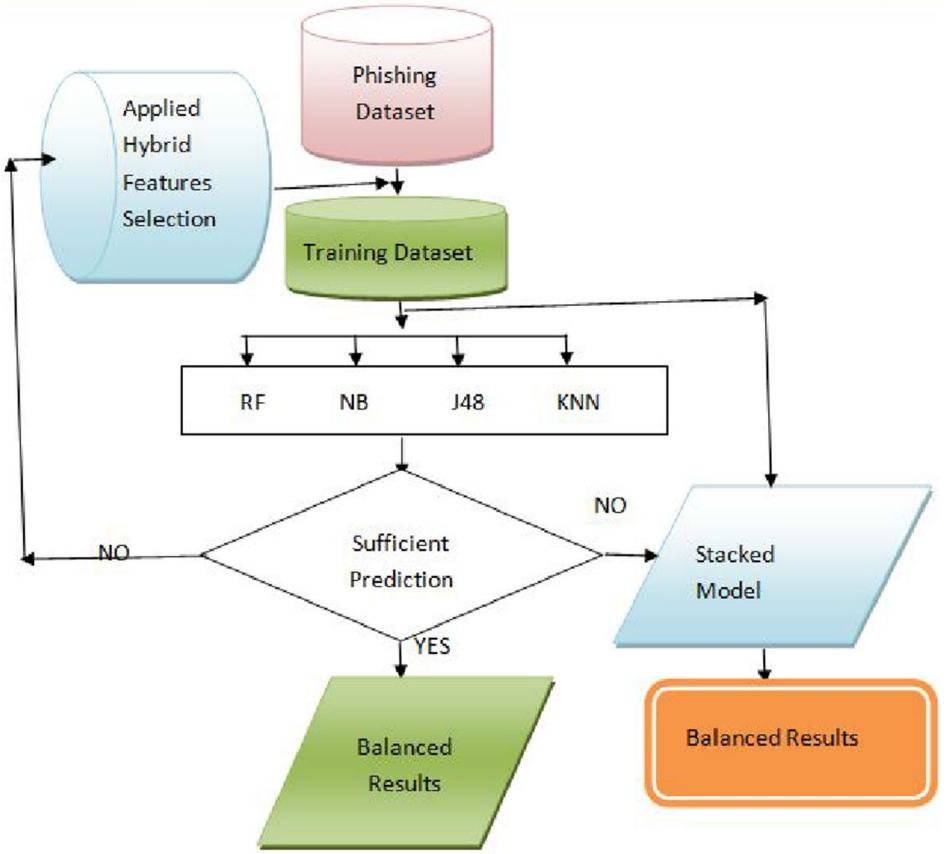

Fig. 2 Graphical representation of the proposed model for phishing website attributes

data-driven "paint by numbers" canvas on top of a picture creates a heat map in Fig. 3. The dataset was subjected to a set of classifiers [20].

The proposed stacking ML ensemble model for improved phishing website detection is shown and explained in this part.

The experiments are divided into three stages: training, rating, and testing. The next sub-sections go through these processes in further detail. The classifiers (RF, NB, J48, and KNN) are trained without optimisation in the training stage. After that, the goal is to first gain a general idea of the classification performance before optimising it, and then to figure out which PW properties are the most useful. To improve the above-mentioned classifiers, the Chi-square, extra tree, and heatmap are utilised. The stacking approach was used to assemble the optimised classifiers and form an ensemble classifier in the ranking step, and the stacking ensemble method was used to boost the overall accuracy of the recommended model by selecting the ideal values of model parameters. The efficacy of each strategy was determined by comparing the outcomes of each method separately. We employed random forest, naive Bayes, J48, and KNN. The strategy that was the most effective in terms of improving and developing the identification of phishing sites was established. For each combination, the accuracy, precision, and F-measure values were determined.

Result and Discussion

In the studies, the tenfold cross-validation approach was employed to validate the models and minimise prediction uncertainty. Using this method, the training dataset was divided into ten subgroups. Each of these subgroups has to be assessed in each of the remaining nine subsets. Each evaluation subgroup was utilised once in each of the ten repetitions. Figure 4 depicts the performance of the four classifiers (RF, NB, KNN, and J48) as well as the derived average accuracy score of 97.024%, 93.171%, 94.446%, and 96.726%, respectively.

In this experiment, we consider the genuine value near to accuracy. The precision is restricted and always assigned the same value in the same experiment for other processes. In the studies, the tenfold cross-validation approach was employed to validate the models and minimise prediction uncertainty. Figure 5 depicts the performance of the four classifiers (RF, NB, KNN, and J48) as well as the estimated average accuracy score (96.582, 94.695, 94.519 and 94.985).

In this experiment, we examine the true positive values. The recall always shows goodness of that model and search

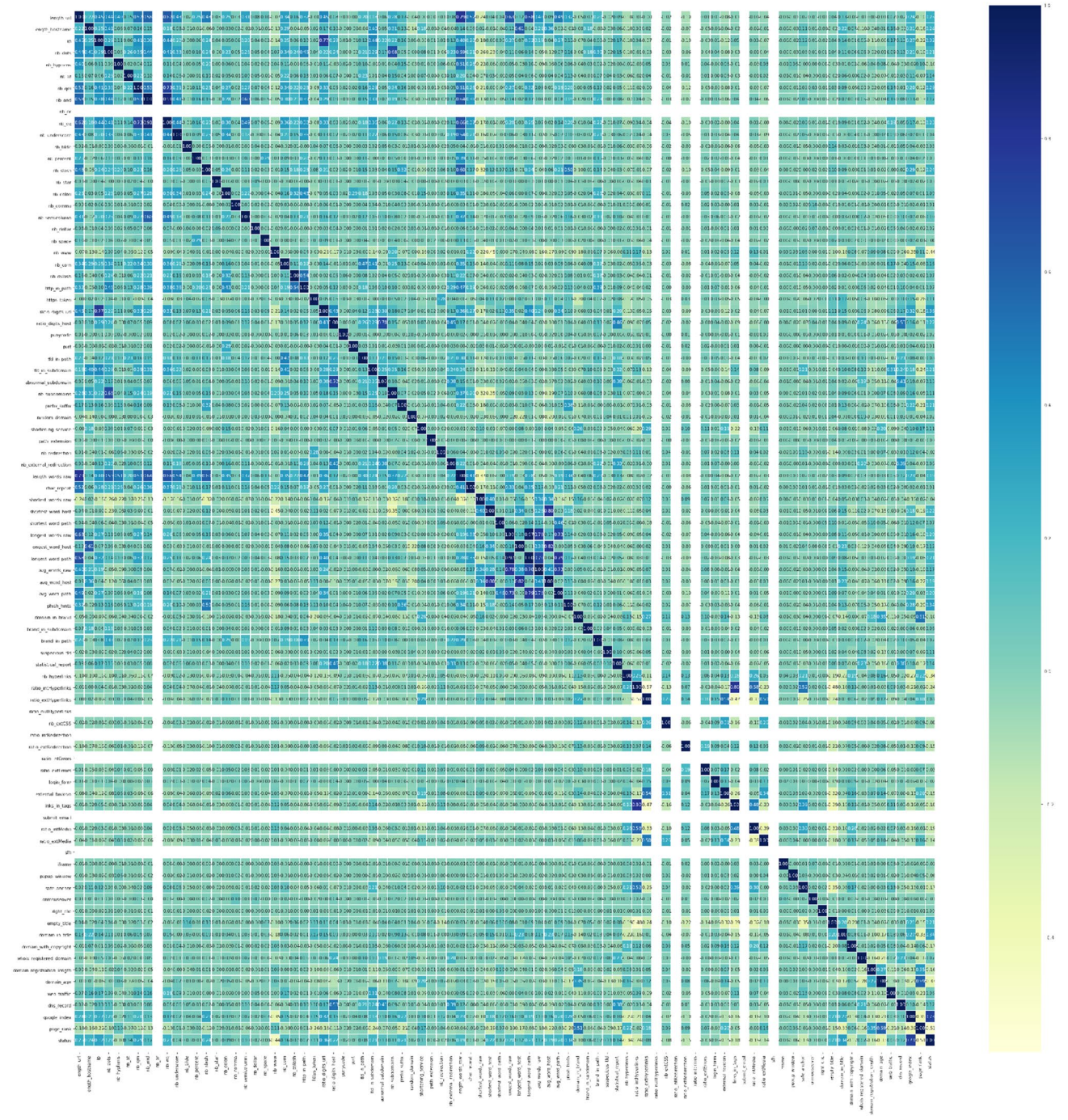

Fig. 3 Graphical representation of phishing websites 89 features by heatmap

and identify the genuine position values. In the studies, the tenfold cross-validation approach was employed to validate the models and minimise prediction uncertainty. Figure 6 depicts the performance of the four classifiers (RF, NB, KNN, and J48) as well as the estimated average recall score (98.084, 96.601, 96.393 and 96.730).

Optimisation by Hybrid Features Selection Methods

In the second phase, the dataset with the selected features were obtained by Chi-square, extra tree and heatmap methods, then, on each of the subsets, four classifiers (Fig. 7).

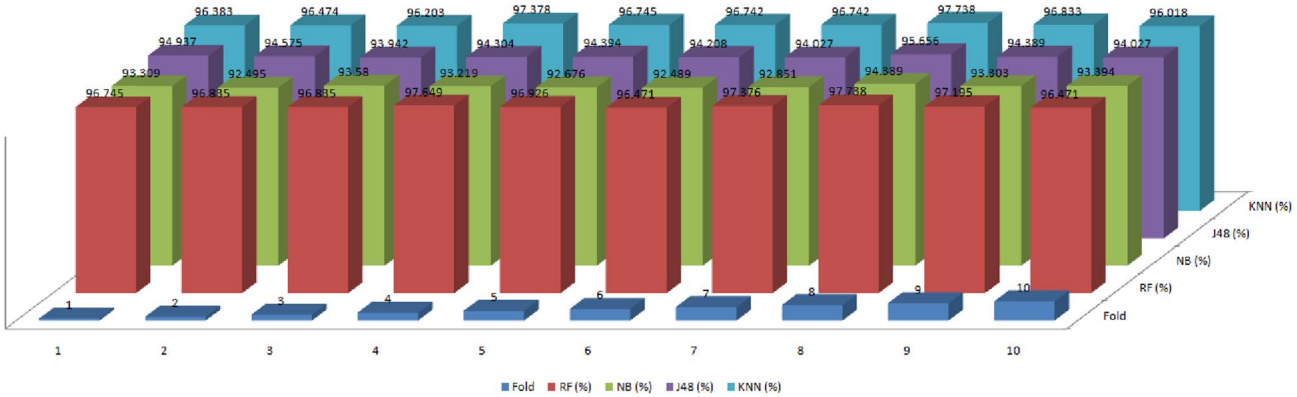

Fig. 4 Representation accuracy of RF, NB, KNN, and J48 classifiers for phishing websites attributes

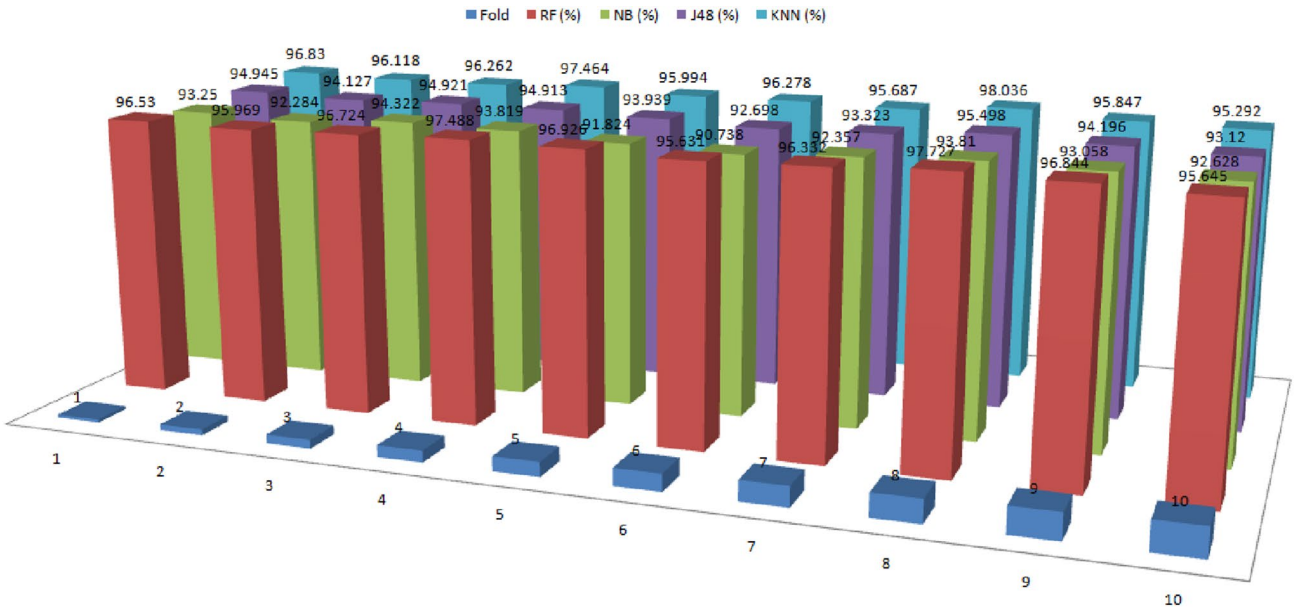

Fig. 5 Representation precision of RF, NB, KNN, and J48 classifiers for phishing websites attributes

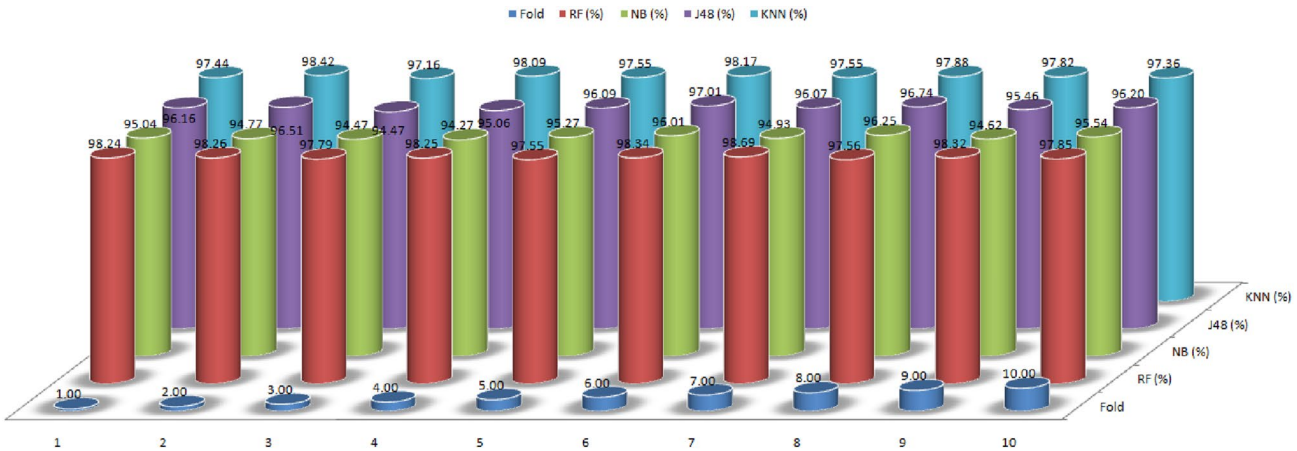

Fig. 6 Representation recall of RF, NB, KNN, and J48 classifiers for phishing websites attributes

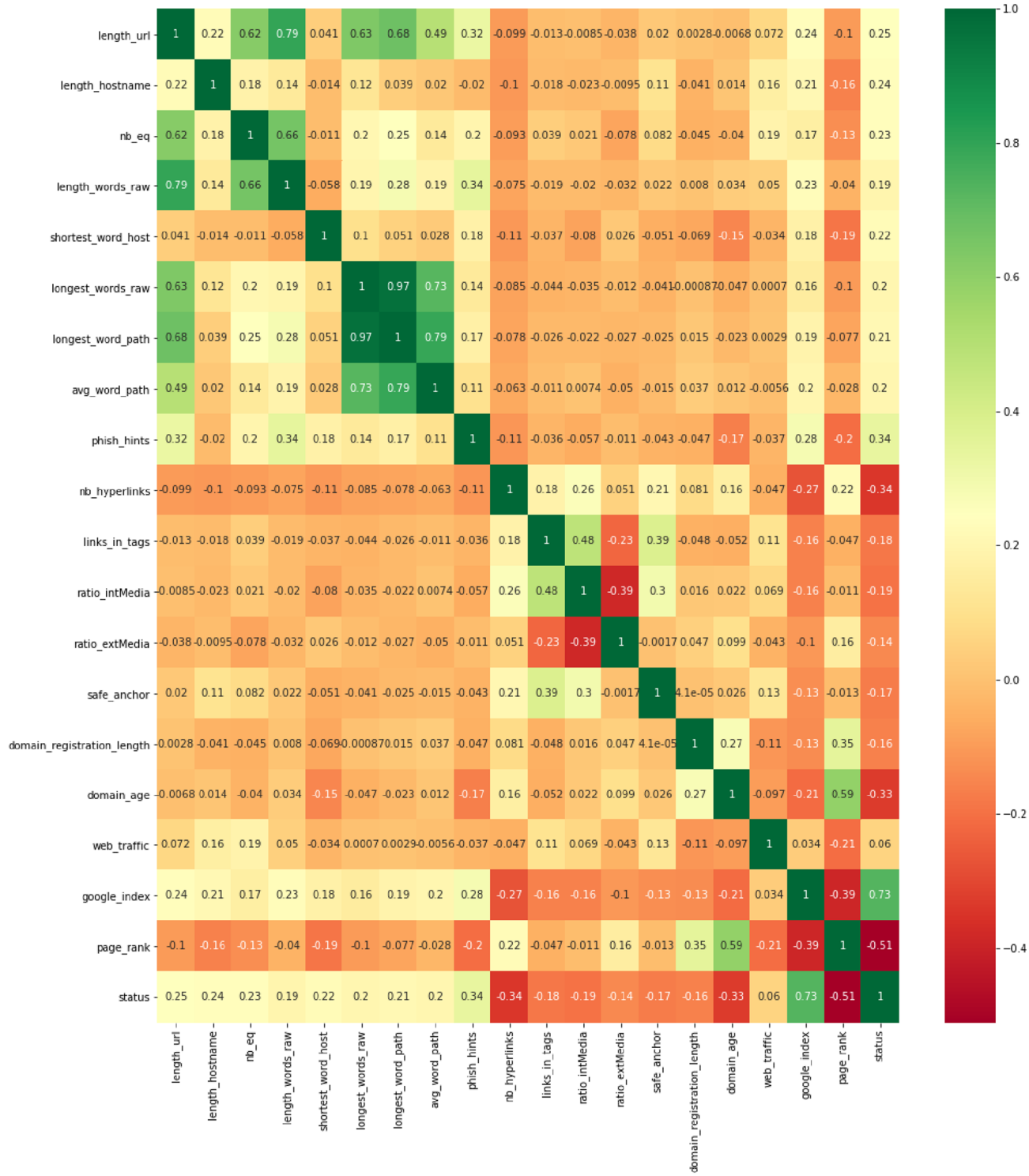

Fig. 7 Representation heatmap for 20 selected important features for phishing websites attributes

Correlate heat_map generates boundary for good relationship of two variables. The boundary limit assigned by 1, − 1 and 0 always have very weak or no relationship. The attributes' boundary variables go to a positive direction, indicating strength association between attributes [20].

Table 1 Representation of Chi-square for phishing websites attributes

S. No	Specs	Score
0	length_url	35,328.04
1	length_hostname	3574.91
2	nb_eq	2116.26
3	length_words_raw	2099.19
4	shortest_word_host	1760.36
5	longest_words_raw	14,504.34
6	longest_word_path	26,076.34
7	avg_word_path	4460.65
8	phish_hints	2785.08
9	nb_hyperlinks	427,921.20
10	links_in_tags	12,891.23
11	ratio_intMedia	21,314.94
12	ratio_extMedia	14,287.97
13	safe_anchor	14,154.79
14	domain_registration_length	402,875.40
15	domain_age	2,992,713.00
16	web_traffic	193,730,400.00
17	google_index	2847.87
18	page_rank	6032.52
19	status	5715.00

Chi-square

Chi-square creates a relation between the observed and predicted frequencies of collected events or variables. Chi-square is a valuable tool for assessing category differences, especially ones that are nominal in character [21]. Table 1 shows 20 important features of a good score.

Extra Tree Features Selection Method

[0.00417966 0.00464072 0.00366814 0.00175792 0.00289009 0.0046448 0.00391138 0.00215644 0.008263 0.01136815 0.01407321 0.00750137 0.0043768 0.01343768 0.00148478 0.01369747 0.00198299 0.21401142 0.04441774 0.63753626].

In this experiment, we applied extra tree classifier as ensemble learning technique. Extra tree does well in generating decision tree by de-correlated decision tree as aggregrates of trees outcomes [22]. In Fig. 8, the important phishing website attributes are selected.

The previous experiment employed a tenfold cross-validation approach to assess the accuracy, precision, and recall; however in this phase, we used Chi-square, additional tree, and heatmap to calculate the feature intensity and significance. The findings of hybrid performance of feature selection approaches lower the uncertainty of prediction. Each evaluation subgroup was utilised once in each of the ten repetitions. Figure 9 displays the performance of the four classifiers (RF, NB, KNN, and J48) as well as the estimated average accuracy score (96.744%, 93.633%, 97.014% and 96.897%).

During the first and second phases of the experiment, the values for accuracy, precision, and recall were calculated, but no viable classifier that provided continuous results was found. We employed these classifiers in a single unit as a stacking ensemble approach (RF + NB + KNN + J48) to forecast phishing websites, since some algorithms produce high values in phase-1 and then compute low values after features selection.

For the same dataset, the accuracy results were compared to suggested hybrid feature selection approaches and stacking ensemble methods. In phase three, the ensemble stacking approach was proven to be more accurate than the other four classifiers in Table 2. Finally, the acquired accuracy was 99.7% with a tenfold increase, which outperformed the other classifier techniques in the previous phase.

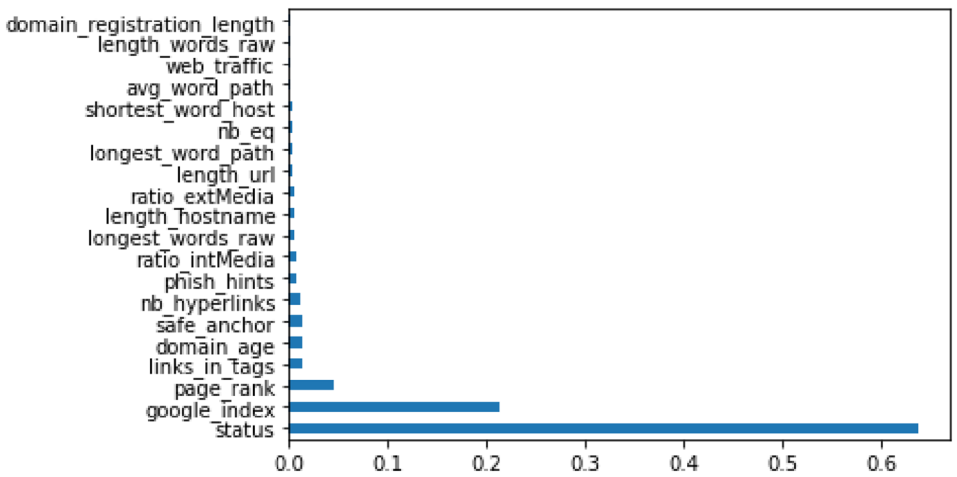

Fig. 8 Representation of extra tree features selection for important features for phishing website attributes

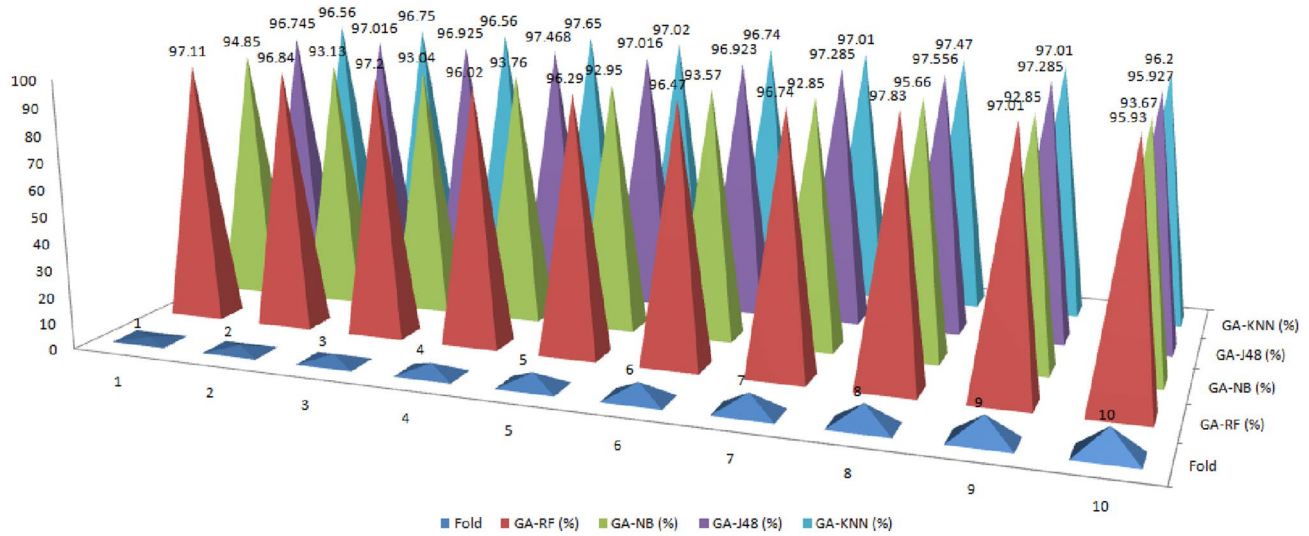

Fig. 9 Representation classification accuracy after using hybrid features selection methods

Conclusion

An efficient stacking ensemble model is proposed in this work to detect phishing sites. Using Chi-square, additional tree, and heatmap, the optimisation approach was utilised to determine the optimal parameter values of multiple classifier learning algorithms. The suggested model is made up of three steps. Several classifier learning methods, including RF, NB, J48, and KNN, were learned in the phase-1, training stage, without employing hybrid feature selection methods. Phase 2 is used to optimise these classifiers by picking the hybrid feature selection method values and calculating the accuracy, precision, and recall score; however if there are any imbalances, the following phase is employed. Certain classifiers were utilised as foundation classifiers for the stacking ensemble approach in phase 3. The best ensemble approaches were these classifiers (RF, NB, J48 and KNN). Finally, all methods were tested using the same dataset as the class's test dataset (legitimate or phishing). With the suggested optimised stacking ensemble approach, phase 3 obtains sufficient findings and a superior performance compared to previous ML-based detection methods. The level of precision attained was 99.7%. A statistical study was undertaken to establish that the gained improvements were statistically significant. Furthermore, the findings revealed that the proposed approaches were more accurate than previous research that employed the same phishing dataset. More light detection methods will be more accurate with IoT surroundings, as a guideline for future investigations. It is also a good idea to use deep learning algorithms to analyse and increase the detection rate of PWs, as well as to use more phishing datasets.

Table 2 The accuracy of the optimised stacking ensemble method

Evaluation	Random forest	Naïve Bayes	J48	KNN	Ensemble (stacking)
Accuracy	96.744	93.633	96.897	97.014	99.7
Precision	0.964	0.975	0.970	0.951	0.957
Recall	0.941	0.958	0.957	0.942	0.981

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Buber E, Demir Ö, Sahingoz OK. Feature selections for the machine learning based detection of phishing websites. In: 2017 International artificial intelligence and data processing symposium (IDAP). IEEE; 2017. pp. 1–5. https://doi.org/10.1109/IDAP.2017.8090317.
2. Vijayalakshmi M, Mercy Shalinie S, Yang MH, Raja Meenakshi U. Web phishing detection techniques: a survey on the state-of-the-art, taxonomy and future directions. IET Networks. 2020;9(5):235–46.
3. Jain AK, Gupta BB. A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP J Inf Secur. 2016;2016:9.
4. Jain AK, Gupta BB. "PHISH-SAFE: URL features-based phishing detection system using machine learning", Cyber Security. Adv Intell Syst Comput. 2018. https://doi.org/10.1007/978-981-10-8536-9_44.
5. Purbay M, Kumar D. Split behavior of supervised machine learning algorithms for phishing URL detection. In: Lecture Notes in

Electrical Engineering, vol. 683, 2021; https://doi.org/10.1007/978-981-15-6840-4_40.
6. Gandotra E, Gupta D. An efficient approach for phishing detection using machine learning. In: Algorithms for Intelligent Systems. Singapore: Springer; 2021. https://doi.org/10.1007/978-981-15-8711-5_12.
7. Basit A, Zafar M, Javed AR, Jalil Z. A novel ensemble machine learning method to detect phishing attack. In: 2020 IEEE 23rd international multitopic conference (INMIC). IEEE; 2020. pp. 1–5. https://doi.org/10.1109/INMIC50486.2020.9318210.
8. Le H, Pham Q, Sahoo D, and Hoi SCH. URLNet: Learning a URL representation with deep learning for malicious URL detection. Conference'17, Washington, DC, USA, arXiv:1802.03162, 2017.
9. Hong J, Kim T, Liu J, Park N, Kim SW. "Phishing URL detection with lexical features and blacklisted domains", Autonomous Secure Cyber Systems. Springer, https://doi.org/10.1007/978-3-030-33432-1_12.
10. Kumar J, Santhanavijayan A, Janet B, Rajendran B and Bindhumadhava BS. Phishing website classification and detection using machine learning. In: 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1–6, https://doi.org/10.1109/ICCCI48352.2020.9104161.
11. Gao X, Shan C, Hu C, Niu Z, Liu Z. An adaptive ensemble machine learning model for intrusion detection. IEEE Access. 2019;7:82512–21.
12. Hassan YA, Abdelfettah B. Using case- based reasoning for phishing detection. Procedia Comput Sci. 2017;109:281–8.
13. Rao RS, Pais AR. Jail-Phish: an improved search engine based phishing detection system. Comput Secur. 2019;1(83):246–67.
14. Aljofey A, Jiang Q, Qu Q, Huang M, Niyigena JP. An effective phishing detection model based on character level convolutional neural network from URL. Electronics. 2020;9(9):1514.
15. AlEroud A, Karabatis G. Bypassing detection of URL-based phishing attacks using generative adversarial deep neural networks. In: Proceedings of the sixth international workshop on security and privacy analytics. 2020. pp. 53–60. https://doi.org/10.1145/3375708.3380315.
16. Wen Y, Wu R, Zhou Z, Zhang S, Yang S, Wallington TJ, et al. A data-driven method of traffic emissions mapping with land use random forest models. Appl Energy. 2022;305: 117916.
17. Anand R, Sakkari DS. Classification of fake news on Twitter by using Naïve Bayes classifier. In: Ranganathan G, Fernando X, Shi F, El Allioui Y, editors. Soft computing for security applications. Singapore: Springer; 2022. pp. 399–408. https://doi.org/10.1007/978-981-16-5301-8_30.
18. Tanvir Fayaz S, Tejanmayi GS, Kanaka Ruthvi Y, Vijaya Shetty S, Shenoy SU, Bhat G. Prediction of liver patients using machine learning algorithms. In: Shetty NR, Patnaik LM, Nagaraj HC, Hamsavath PN, Nalini N, editors. Emerging research in computing, information, communication and applications. Singapore: Springer; 2022. p. 135–45. https://doi.org/10.1007/978-981-16-1338-8_12.
19. Wang Y, Pan Z, Dong J. A new two-layer nearest neighbor selection method for kNN classifier. Knowl-Based Syst. 2022;235: 107604.
20. Lin CW, Hong S, Lin M, Huang X, Liu J. Bird posture recognition based on target keypoints estimation in dual-task convolutional neural networks. Ecol Ind. 2022;135: 108506.
21 Sumant AS, Patil D. Ensemble Feature Subset Selection: Integration of Symmetric Uncertainty and Chi-Square techniques with RReliefF. J Inst Eng (India). 2022. https://doi.org/10.1007/s40031-021-00684-5.
22. Kharwar AR, Thakor DV. An ensemble approach for feature selection and classification in intrusion detection using extra-tree algorithm. Int J Inf Secur Privacy (IJISP). 2022;16(1):1–21.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.