www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Robust Detection Strategies for Phishing across Diverse Data Streams and Methodologies

Vadde Varun Tej Reddy, varuntejreddy999@gmail.com

Abstract: Phishing attacks have become an enormous cybersecurity issue, leading to comprehensive observe aimed at identifying the best techniques for classifying and detecting these misleading strategies designed to mislead individuals and organizations into disclosing vital facts. This venture fills a big void in present research through methodically assessing diverse category techniques across various facts situations, ensuring they are not confined to particular datasets or methodologies, thereby offering a comprehensive view of their efficacy in countering phishing attacks. The take a look at evaluated 13 modern-day type methods frequently hired in preliminary phishing research. It uncovered them to 10 varied performance metrics, proceeding to furnish a thorough comprehension in their abilities. This research presents widespread insights into phishing class techniques, improving the existing knowledge base and helping in the formulation of extra effective countermeasures in opposition to phishing threats. The task makes use of the Stacking Classifier, an effective ensemble method, integrating RF, MLP, and LightGBM models to attain a hundred% accuracy in phishing assault category. A Flask-based interface allows truthful consumer trying out and performance assessment. The implementation of person authentication guarantees comfy access, aiding in a thorough assessment of phishing categorization methodologies throughout various data sources and frameworks.

"Index Terms - *Benchmark testing, classification algorithms, performance evaluation, phishing".*

I. INTRODUCTION

Phishing constitutes a significant danger to cybersecurity, as described by way of the national Institute of standards and technology, involving attempts to acquire sensitive data, consisting of bank account numbers, or gain access to extensive computerized systems thru misleading requests thru emails or web sites. The average probability of exposure to this attack throughout distinctive sectors is 11%. [1] "Phishing is a socially engineered attack" that often causes physical or psychological harm to persons and groups. The company sectors embody era, energy or utilities, retail, and financial offerings. These groups exhibit significant susceptibility to phishing attacks. Consequently, measures grounded on cyber security are essential to thwart those attacks [3]. Numerous studies have been conducted on phishing avoidance, particularly specializing in its identification and category.

diverse methodologies are hired in the classification procedure, including Random forest "[4], [5], [6], [7], [8], [9], [10]", "support Vector machine (SVM)" [11], [12], [13], [14], Logistic Regression [15], [16], [17], "Multilayer Perceptron (MLP)" [18], C4.five [19] and [20], and Naïve Bayes [21]. Every demonstrates most appropriate overall performance primarily based on its unique application. The effects of the type approach do now not always practice universally. Therefore, a comparative take a look at need to be conducted to address this deficiency.

Nevertheless, handiest a limited number of studies have performed comparisons of phishing categorization techniques, including [8], [18], [22], [23], and [24]. This contrast observe is frequently segmented into 4 key components: phishing, dataset type, performance assessment, and hired strategies. The data sources utilized by [8], [18], [22], [23], and [24] have been acquired through a phishing internet site and URL, at the same time as [24] hired raw emails retrieved from Apache SpamAssassin and Nazario. The number one performance metrics are "accuracy, precision, and F-measure. Random forest, help vector machines, and Naïve Bayes" are the maximum customary methodologies. This comparative study identifies a gap regarding the effect of existing tactics on diverse public datasets, encompassing each balanced and unbalanced classes.

This research evaluates the performance of the category technique using a particular unbalanced dataset for distinct phishing sorts. This resembles the methodologies hired by using research that did no longer juxtapose those class structures. Vaitkevicius and Marcinkevicius [18] employed balanced datasets and one unbalanced dataset. It was revealed that they performed advanced consequences compared to earlier evaluations. Gana and Abdulhamid [23] exclusively applied unbalanced public datasets, demonstrating that category performance varies primarily based at the subset approach hired. This research is based on multiple investigations that did not demonstrate how

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

performance evaluation affects the methods employed to categorize one of a kind subsets of dataset schemes. Some only articulated the limited influence of this performance on generic designs, such ninety:"10, 80:20, 70:30, and 60:40". furthermore, overall performance evaluation and class methodologies are constrained by means of various metrics, along with “accuracy, F-measure, Precision, true positive rate (TPR), Receiver operating characteristic (ROC), false positive rate (FPR), Precision-recall Curve (p.c), Matthews Correlation Coefficient (MCC), Balanced Detection rate (BDR), and Geometric mean (G-mean)". Studies have shown that each subgroup of the scheme in each balanced and unbalanced data set affects the overall evaluation of the performance of classification technology. This increases dramatically and reduces the performance of positive subgroups.

II. RELATED WORK

Globalization in the 21st century significantly impacts the world because of improvements in technology and communication, allowing usual access to a worldwide market and data interchange through English. Therefore, electronic verbal exchange has integrated into the regular routines of contemporary international professionals across numerous sectors who operate consistently in front in their digital screens. Every so often, these professions may additionally come upon Nigerian 419 rip-off emails while criminals use victims to make in advance payments for non-existent monetary benefits. Those emails expertly combine various persuasive strategies. consequently, the victim vulnerable to the proposition is more willing to react and in the end be enticed into economic loss. This take a look at analyzed a corpus of fifty “Nigerian 419 scam” emails through textual analysis to research the linguistic factors of the persuasion techniques hired by fraudsters to fulfill their conversation goals of lures and deceptions. The examine [2] has identified two number one categories of deceptive strategies employed in conjunction: framing-rhetoric triggers, masquerading as conventional electronic communication genres, and human weakness-exploiting triggers, designed to initiate receivers' emotions. The paper presents pedagogical suggestions for commercial enterprise English educators concerning study room activities, in addition to cautions for both amateur and pro enterprise specialists approximately deciphering the messages of strange emails with vigilance.

Phishing techniques use north -based features based on source code and 0.33 pages to perceive phishing websites. Those strategies own a few drawbacks, considered one of that's their incapability to address force-via downloads. They similarly hire 1/3-party offerings for the identification of phishing URLs, which prolongs the class method. Therefore, on this have a look at [4], we introduce a lightweight utility, CatchPhish, which evaluates the legality of the URL without accessing the web. Counseling technology uses the host name, complete URL [4, 13, 21 and 26], frequency (TF-ID) and phish-indicative phrases From the URL suspected to the classifyer via Random One Classifier. The recommended model has achieved 93.25%accuracy by using full TF -DF functions on our dataset. The TF-IDF mixture and manually designed capabilities have achieved a great accuracy of 94.26% and 98.25% and 97.49% on the benchmark data sets on our data set, exceeding the overall performance of current basic fashion.

In latest years, net phishing assaults have continually developed, ensuing in a decline in client consider in e-commerce and on line services. Diverse techniques and processes utilizing a blacklist of phishing web sites are employed to become aware of phishing web sites [8, 9, 10, 11, and 13]. Unfortunately, the rapid advances of the generation led to the emergence of extra complex techniques for the construction of web pages for customer involvement. As a result, the maximum latest and newly installed phishing websites, such as phishing sites with zero bottom, eliminate detection using blacklist -based methodologies. Numer projects of contemporary studies have used machine learning algorithms to deploy phishing websites, using their use as a system of timely caution to identify these risks. However, widespread websites have been selected on the basis of human use or frequency assessment of the website attributes in most of these methodologies. This article [5] suggests a reasonable detection of phishing websites by weighing the function primarily based on optimizing particle swarm to increase the efficiency of detection. The proposed method advocates by “optimizing particle swarms (PSO)” to successfully assign weights to exceptional websites and therefore increases the accuracy in Find phishing websites. The recommended part of the PSO website has been excellent in terms of their ability to validate fishing places within the fantastic website's habitat categories. Experimental results have confirmed that the proposed weighing features based on PSOs significantly increased accuracy of the category, correct positive and bad citations and false positive and negative devices that will learn about fashion, using fewer websites to detect websites.

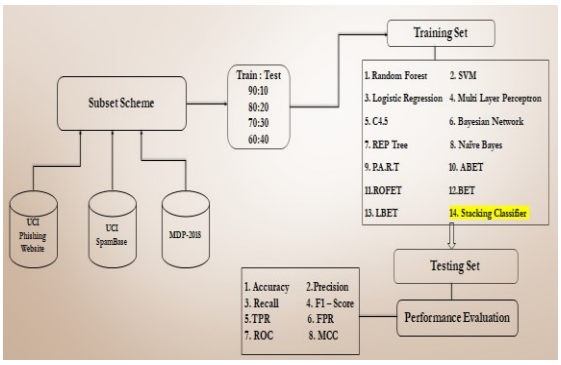
www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Phishing is a cyber-attack that exploits unsuspecting web customers to extract non-public data, including usernames, passwords, social protection numbers, and credit card details. Attackers mislead net customers by way of disguising webpages as credible or legitimate to obtain non-public data. severa anti-phishing solutions, including blacklists, whitelists, heuristic algorithms, and visual similarity strategies, have been offered; but, net consumers continue to fall sufferer to disclosing sensitive information on phishing websites. This studies [6] affords a singular categorization model utilising heuristic capabilities derived from URLs, source code, and 1/3-birthday party offerings to cope with the restrictions of current anti-phishing strategies. The model became assessed using 8 wonderful system gaining knowledge of techniques, among which the Random wooded area (RF) set of rules [4], [5], [6], [7], [8], [9], [10] exhibited the best overall performance, with an accuracy of ninety nine.31%. To determine the optimum classification to identify fishing spots, pain was stopped using multiple orthogonal and indirect random one classifiers. Analysis The most important demonstrated great performance among all "oblique random forests (ORF)" with an accuracy of 99.55%. Each of them tested our model using a 1/3-linked function to assess the effectiveness of 0.33 electoral services in classifying questionable websites. In addition, we created the opposing model using our findings "(Cantina and Cantina+)". Our proposed solution has been tested using current approaches and has accepted zero-day fishing attacks.

This research introduces a unique characteristic selection framework for the phishing detection system based on machine learning, called "HEFS) [7]. The revolutionary technique of "cumulative distribution function (CDF-G)" is applied throughout the HEF segment for generating subset number one, which could be sooner or later to enter a set of statistical failure set to produce subset of secondary functions. The second step generates a group of basic functions from secondary characteristic subsets using a set of functions. Experimental findings suggest that HeFS the famous line of line power in combination with the random forest classifier, precisely distinguishes ninety -four. In a separate test of the basic line function (10) used in a random forest exceeds a comprehensive set of functions (48 V overall) used in SVM classifiers, naive Bayes, C4.Five, Jrip and elements. HEFS demonstrates encouraging results while they are evaluated against a recognized phishing data set from the "University of California Irvin (UCI)" storage. Therefore, HeFS is a very great and pragmatic technique of choosing functions for the structures of phishing learning -based gadget.

III. MATERIALS AND METHODS

This study does an extensive assessment of phishing category methodologies utilising diverse facts sources and frameworks. The analysis encompasses the comparison of 13 unique type methodologies. The research employs both unbalanced and balanced fish datasets, as well as evaluating the influence of various strategies under scenarios that cause data to vary in different proportions. This study sheds light on the adaptability and efficacy of these strategies in the realm of fish development. The stacking classifier, an effective method to the file, was employed to improve the accuracy of the fish attack classification. The integration of "Random Forest (RF), multilayer perceptron (MLP)" and LightGBM models guarantee higher and reliable final prediction in the file and achieve an exceptional 100% accuracy. A person-friendly front give up using the Flask framework is provided to decorate consumer testing and performance evaluation. Person authentication protocols are hooked up to guarantee secure access, facilitating a radical and dependable assessment of phishing categorization methodologies throughout diverse data sources and frameworks.



"Fig.1 Proposed Architecture"

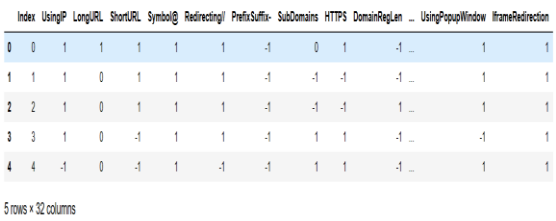
The subset scheme was formulated in such a way that it was in accordance with the actual conditions and analogous findings were achieved from the subsequent test. There was a tenfold approach of cross validation to guarantee the perfection and reliability of the classification model. It is unreasonable to depend only on accuracy such as metrics for performance evaluation [18], [24]. This resulted inside the use of 10 performance assessment metrics: accuracy, F-degree, precision, "true positive rate (TPR), receiver operating characteristic (ROC), false fine rate (FPR), precision-recall curve (p.c), and balanced detection rate (BDR), Matthews

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

correlation coefficient (MCC), and geometric mean (G-mean)". A class technique that accomplished incredibly well in all these tests has been identified, as illustrated in Fig 1.

A) Dataset Collection:

3 public data sets, namely MDP-2018, UCI Phishing and Spam base, were used to evaluate the categorization strategies. UCI Phishing and Spam base data sets have a jerky class distribution, while the MDP-2018 data file is balanced. It consists of 5000 phishing and criminal websites. The MDP-2018 contains 48 functions, while UCI Spam base contains 58 capabilities with a distribution of 2,788 real "e-mails and 1,813 phishing emails". The UCI Phishing data file consists of 31 attributes that include the facts of "6 157 phishing sites and 4,898" real websites.



	Index	UsingIP	LongURL	ShortURL	Symbol@	Redirecting//	PrefixSuffix-	SubDomains	HTTPS	DomainRegLen	...	UsingPopupWindow	IframeRedirection
0	0	1	1	1	1	1	-1	0	1	-1	...	1	1
1	1	1	0	1	1	1	-1	-1	-1	-1	...	1	1
2	2	1	0	1	1	1	-1	-1	-1	1	...	1	1
3	3	1	0	-1	1	1	-1	1	1	-1	...	-1	1
4	4	-1	0	-1	1	-1	-1	1	1	-1	...	1	1

5 rows × 32 columns

"Fig.2 Dataset Collection"

B) Processing:

Data processing changes unrefined data to usable statistics for businesses. Scientists usually interact in the processing of records, include a collection, company, cleaning, verification, evaluation and transformation of statistics into a clear format together with graphs or articles. Data processing can be done in three ways: Vizard, mechanical, and digital. It tries to raise the cost of data and make decisions more effective. This enables businesses to enhance their image and make swift strategic decisions. In this context, data processing skills such as computer programming are significant. This can alter the comprehensive dataset for enormous knowledge of quality and decision-making, particularly large data.

Feature selection: Choosing tasks is the process of creating the most practical, non-oriented, and pleasant functions for model development. As the dataset's importance and diversity grow, systematic minimizing of its dimensions becomes important. The fundamental purpose of selecting elements is to maximize the efficiency of the model that predicts the future while also lowering data expenses for modeling. The choice of tasks, the original engineering factor, entails recognizing the most significant properties of machine learning algorithms. Functional choice strategies are used to limit the number of input variables, as opposed to superfluous or unnecessary functions, resulting in better visibility for those that are most significant to machine learning models. The key benefits of selecting additional attributes rather than allowing machine learning models to set the machine learning model automatically.

C) Algorithms:

Random Forest: "Random forest" is a technique of learning a file that integrates a number of "decision -making trees and provides predictions". It creates a number of decision -making trees and aggregates their predictions to increase accuracy and alleviate excessive amounts. Random forest is resilient, contains excessive-dimensional data, and is gifted in both class and regression applications. Inside the realm of phishing class, it can yield a substantial stage of accuracy [4], [5], [6], [7], [8], [9], [10].

Support Vector Machine (SVM): "support Vector machine (SVM)" It is a technique of learning under supervision that identifies the ideal hyperplane for the distinction between classes and maximizes between them. "Support for vector machines (SVM)" are hired for binary classification tasks and are very talented in the management of complex decision boundaries. It is often utilized in phishing categorization because of its ability to manage non-linear data [11], [12], [13], [14].

The equation of linear hyperplan is expressed as:

"wTx+b=0$wTx+b$=0 (1)"

Where:

- W is a normal vector for hyperplan, which indicates the direction perpendicular to it.
- BB is the term offset or distortion, which shows at the distance of hyperplan from origin along the normal vector WW.

Logistic Regression: Logistic regression is a statistical model that uses a logistical function to symbolize the likelihood of a binary result. Its kilometers of linear class algorithm. Logistic regression is direct, interpreted and is often used as a basic approach to binary classification problems such as phishing detection [15], [16], [17].

The simple linear regression line,

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

"^y=a+bx y ^ = a + b x (2)"

Can be interpreted as follows:

^y represents the anticipated value of y, a denotes the intercept indicating where the regression line intersects the y-axis, and b forecasts the change in y corresponding to each unit change in x.

Multilayer Perceptron (MLP): MLP is a form of "artificial neural network", including several layers of interconnected nodes (neurons) that could explore complex input formulas. "Multilayer perceptron's (MLPs)" are utilized for their capacity to simulate non-linear connections and constitute a crucial detail of "deep learning". They are capable of managing a diverse array of categorization tasks, such as phishing detection [18].

The MLP networks consist of numerous interconnected functions. A network along with three features or levels would be established.

"f(x) = f (3)(f (2)(f (1)(x))). (3)"

Each layer consists of units that execute an affine translation of a linear combination of inputs.

C4.5: C4.5 is an algorithm for classification that uses decision -making trees. Iteratively divides the data file into subset according to the most important characteristics for creating a selection tree. C4.5 is an access to a tree with a selection and its simplicity and interpretability makes IT advantageous for clarifying the decision system in phishing categorization [19, 20].

Upon determining the value of every incidence and its associated possibility, compute the anticipated value of every outcome utilizing the subsequent formula:

"Expected value (EV) = (First possible outcome x Likelihood of outcome) + (Second possible outcome x Likelihood of outcome) - Cost. (4)"

Bayesian Network (Bernoulli NB): Bayes Network is a probability-graphic model that depicts the possible correlation of a set of variables. Childrenuli Bole is a version optimized for binary data. Bayesian Networks can encapsulate dependencies and conditional chances within the statistics, facilitating the modelling of phishing event likelihood depending on observable traits.

REP Tree (Decision Tree): Rep Tree is a categorization version of Timber decision -making. It creates a hierarchical structure using data distribution. BUSHES REP are decision -making trees adapted to positive data sets that can provide exceptional accuracy in class applications, including phishing detection.

Naive Bayes: "Naive Bayes" is a probabilistic method derived from "Bayes' theorem". It classifies via presuming that features are impartial, a "naive" yet frequently powerful assumption. Naive Bayes is an efficient and speedy algorithm for text class, rendering it appropriate for phishing type obligations, mainly whilst handling textual data [21].

In generalized notation we write:

"p (A,B|A) = p(A) * p(B|A) (5)"

The statement reads, "the probability of A and B, given A, is equivalent to the probability of an improved by means of the probability of B, conditioned on A being known or having occurred." This is called conditional opportunity, or more accurately, joint conditional possibility, because the opportunity is classified based on a preceding event or condition.

PART (Passive Aggressive Random Forest decisionTree): Part is a classifier based on the rules that formulate a collection of rules derived from statistics. Passive aggressive techniques are generally hired in the online and sequential educational environment. Part can build rules elucidating the rationale behind unique decisions, that's advantageous for comprehending and alleviating phishing threats.

ABET (AdaBoost ExtraTree): ABET is an ensemble learning technique that integrates greater trees with AdaBoost. Extra timber constitute a kind of Random forest. AdaBoost with extra timber complements type efficacy by using amalgamating the advantages of each methodologies. It is able to be particularly efficacious for coping with imbalanced datasets [29].

ROFET (Random Forest ExtraTree): ROFET integrates Random forest with extra trees, both of which are stochastic decision trees. ROFET integrates the resilience of Random woodland with the variance mitigation of more trees, potentially enhancing overall classification precision.

BET (Bagging ExtraTree): bet integrates Bagging with more timber, utilizing extra trees as the foundational estimator. Bagging can improve the

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

"accuracy and robustness" of extra trees by mitigating overfitting and version.

LBET (Logistic Gradient ExtraTree): LBET is a hybrid model that integrates logistic regression with extra trees. LBET offers a synthesis of the interpretability inherent in logistic regression with the efficacy of extra trees, rendering it advantageous for elucidating and categorizing phishing occurrences.

Stacking Classifier (RF + MLP with LightGBM): Stacking is an ensemble method that integrates numerous base models "(Random forest and MLP)" through a "meta-version (LightGBM)". Stacking utilizes the advantages of various algorithms, potentially improving overall class precision and resilience in phishing detection.

IV. RESULTS AND DISCUSSION

Accuracy: The accuracy of a test is its ability to distinguish between patients and wholesome people. To evaluate test accuracy, compute the ratio of genuine positives and true negatives across all assessed instances. This can be expressed mathematically as:

$$"Accuracy = \frac{TP + TN}{TP + FP + TN + FN}(6)"$$

Precision: Precision quantifies the proportion of positively identified cases or samples that are correctly classified. Precision is determined by the formula:

$$"Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}(7)"$$

Recall: "machine learning" don't forget assesses a model's capability to perceive all pertinent instances of a category. It demonstrates a version's efficacy in encapsulating times of a class by means of comparing accurately predicted positive observations to the overall range of positives.

$$"Recall = \frac{TP}{TP + FN}(8)"$$

F1-Score: The accuracy of a "machine learning" model is assessed using the F1 score. Integrating model precision and recall metrics. The accuracy metric quantifies the frequency of real predictions made through a model throughout the dataset.

$$"F1\ Score = 2 * \frac{Recall\ X\ Precision}{Recall + Precision} * 100(9)"$$

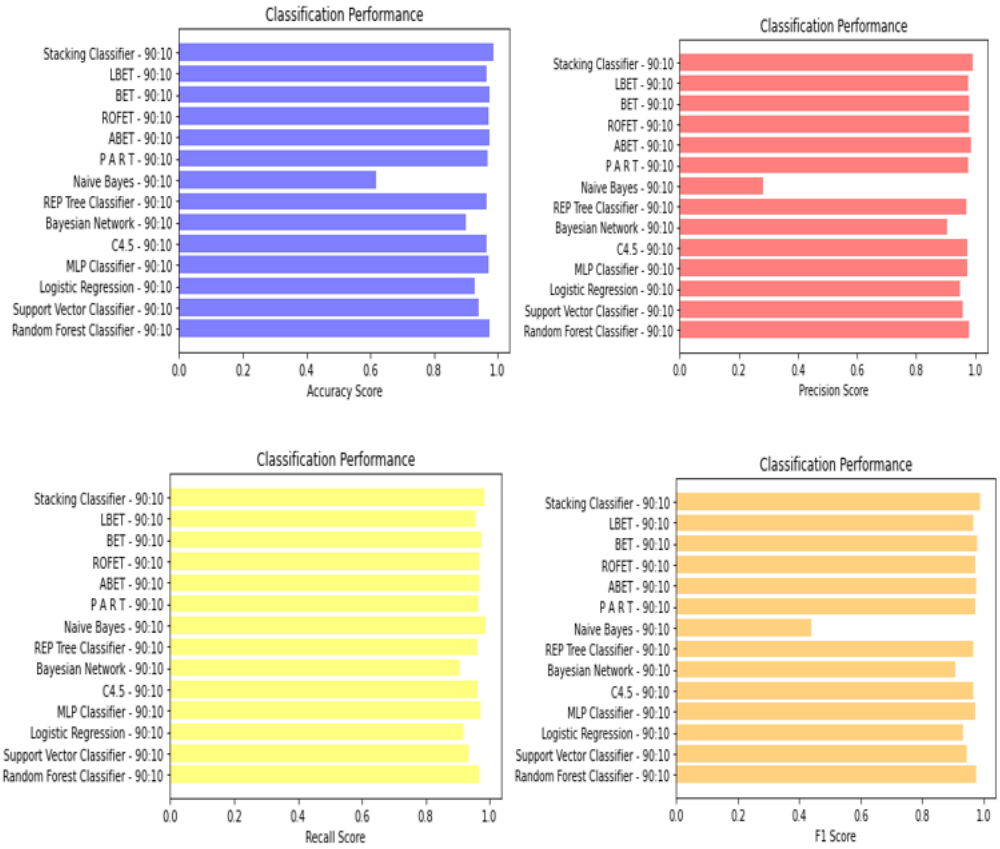
Desk (1) they will assess performance metrics-accuracy, withdrawal, and F1-scóre-for each set of rules. The stacking classifier regularly overcomes all other algorithms. The tables provide a comparative test of metrics for alternative methods.

"Table.1 Performance Evaluation Table-MDP"

ML Model	Accuracy	F1-score	Recall	Precision
Random Forest	0.983	0.983	0.980	0.986
SVM	0.867	0.870	0.829	0.916
Logistic Regression	0.939	0.939	0.917	0.961
MLP	0.958	0.958	0.930	0.988
C4.5	0.970	0.969	0.963	0.975
Bayesian Network	0.931	0.930	0.920	0.940
REP Tree	0.966	0.966	0.952	0.979
Naïve Bayes	0.833	0.808	0.916	0.722
PART	0.991	0.991	0.992	0.990
ABET	0.990	0.990	0.990	0.990
ROFET	0.990	0.990	0.988	0.992
BET	0.988	0.988	0.992	0.984
LBET	0.989	0.989	0.990	0.988
Stacking Classifier	0.999	0.999	0.999	0.988

"Graph.1 Comparison Graphs-MDP"

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86



Accuracy is depicted in blue, precision in red, take into account in yellow, and F1-score in orange in Graph (1). Compared to other models, the stacking classifier demonstrates additional performance across all criteria and reaches the highest values. Graphs above visually represent these findings.

V. CONCLUSION

This project has performed in depth of evaluation of multiple machine learning strategies to detect phishing, along with different data sets and record ratios to ensure careful analysis. The incorporation of ensemble processes, specially the Stacking Classifier, markedly more suitable model accuracy and validated the effectiveness of integrating many models for more predictive overall performance. The undertaking achieved user-friendly interactions and superior user authentication through the smooth integration of Flask with SQLite, creating a safe and user-centric platform for URL access and phishing prediction access. This look at not handiest showcases top notch technological achievements however also affords important An overview of reasonable use of file techniques and web interfaces, drastically promoting our understanding and using cyber security strategies.

Utilizing hyper-parameter adjustment to evaluate performance within subsets of destiny research. Broadening the assessment parameters to comprise additional classification methodologies past the authentic thirteen. Examining an extensive array of overall performance measures for an intensive knowledge of categorization technique efficacy. Investigating numerous records assets, such as authentic phishing datasets and sector-specific facts, to evaluate the efficacy of classification techniques in different situations [18, 23].

REFERENCES

[1] Proofpoint. (Jun. 2022). 2021 State of the Phish. [Online]. Available: https://www.proofpoint.com/sites/default/files/gtd-pfpt-us-tr-state-of-thephish-2020.pdf
[2] C. Naksawat, S. Akkakoson, and C. K. Loi, ‘‘Persuasion strategies: Use of negative forces in scam E-mails,’’ GEMA Online J. Lang. Stud., vol. 16, no. 1, pp. 1–17, 2016.
[3] M. A. Pitchan, S. Z. Omar, and A. H. A. Ghazali, ‘‘Amalan keselamatan siber pengguna internet terhadap buli siber, pornografi, e-mel phishing dan pembelian dalam talian (cyber security practice among internet users towards cyberbullying, pornography, phishing email and online shopping),’’ Jurnal Komunikasi, Malaysian J. Commun., vol. 35, no. 3, pp. 212–227, Sep. 2019.
[4] R. S. Rao, T. Vaishnavi, and A. R. Pais, ‘‘CatchPhish: Detection of phishing websites by inspecting URLs,’’ J. Ambient Intell. Hum. Comput., vol. 11, no. 2, pp. 813–825, Feb. 2020.
[5] W. Ali and S. Malebary, ‘‘Particle swarm optimization-based feature weighting for improving intelligent phishing website detection,’’ IEEE Access, vol. 8, pp. 116766–116780, 2020.

[6] R. S. Rao and A. R. Pais, ''Detection of phishing websites using an efficient feature-based machine learning framework,'' Neural Comput. Appl., vol. 31, no. 8, pp. 3851–3873, Aug. 2019.
[7] K. L. Chiew, C. L. Tan, K. Wong, K. S. C. Yong, and W. K. Tiong, ''A new hybrid ensemble feature selection framework for machine learning-based phishing detection system,'' Inf. Sci., vol. 484, pp. 153–166, May 2019.
[8] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, ''Machine learning based phishing detection from URLs,'' Expert Syst. Appl., vol. 117, pp. 345–357, Mar. 2019.
[9] S. W. Liew, N. F. M. Sani, M. T. Abdullah, R. Yaakob, and M. Y. Sharum, ''An effective security alert mechanism for real-time phishing tweet detection on Twitter,'' Comput. Secur., vol. 83, pp. 201–207, Jun. 2019.
[10] V. Muppavarapu, A. Rajendran, and S. K. Vasudevan, ''Phishing detection using RDF and random forests,'' Int. Arab J. Inf. Technol., vol. 15, no. 5, pp. 817–824, 2018.
[11] A. S. Bozkir and M. Aydos, ''LogoSENSE: A companion HOG based logo detection scheme for phishing web page and E-mail brand recognition,'' Comput. Secur., vol. 95, Aug. 2020, Art. no. 101855.
[12] S. E. Raja and R. Ravi, ''A performance analysis of software defined network based prevention on phishing attack in cyberspace using a deep machine learning with CANTINA approach (DMLCA),'' Comput. Commun., vol. 153, pp. 375–381, Mar. 2020.
[13] M. Sameen, K. Han, and S. O. Hwang, ''PhishHaven—An efficient real-time AI phishing URLs detection system,'' IEEE Access, vol. 8, pp. 83425–83443, 2020.
[14] R. S. Rao, T. Vaishnavi, and A. R. Pais, ''PhishDump: A multi-model ensemble based technique for the detection of phishing sites in mobile devices,'' Pervasive Mobile Comput., vol. 60, Nov. 2019, Art. no. 101084.
[15] Y. Ding, N. Luktarhan, K. Li, and W. Slamu, ''A keyword-based combination approach for detecting phishing webpages,'' Comput. Secur., vol. 84, pp. 256–275, Jul. 2019.
[16] A. K. Jain and B. B. Gupta, ''A machine learning based approach for phishing detection using hyperlinks information,'' J. Ambient Intell. Hum. Comput., vol. 10, no. 5, pp. 2015–2028, May 2019.
[17] A. E. Aassal, S. Baki, A. Das, and R. M. Verma, ''An in-depth benchmarking and evaluation of phishing detection research for security needs,'' IEEE Access, vol. 8, pp. 22170–22192, 2020.
[18] P. Vaitkevicius and V. Marcinkevicius, ''Comparison of classification algorithms for detection of phishing websites,'' Informatica, vol. 31, no. 1, pp. 143–160, Mar. 2020.
[19] Y.-H. Chen and J.-L. Chen, ''AI@ntiPhish—Machine learning mechanisms for cyber-phishing attack,'' IEICE Trans. Inf. Syst., vol. E102.D, no. 5, pp. 878–887, May 2019.
[20] C. L. Tan, K. L. Chiew, K. S. C. Yong, S. N. Sze, J. Abdullah, and Y. Sebastian, ''A graph-theoretic approach for the detection of phishing webpages,'' Comput. Secur., vol. 95, Aug. 2020, Art. no. 101793.
[21] S. Mishra and D. Soni, ''Smishing detector: A security model to detect smishing through SMS content analysis and URL behavior analysis,'' Future Gener. Comput. Syst., vol. 108, pp. 803–815, Jul. 2020.
[22] M. Karabatak and T. Mustafa, ''Performance comparison of classifiers on reduced phishing website dataset,'' in Proc. 6th Int. Symp. Digit. Forensic Secur. (ISDFS), Mar. 2018, pp. 1–5.
[23] N. N. Gana and S. M. Abdulhamid, ''Machine learning classification algorithms for phishing detection: A comparative appraisal and analysis,'' in Proc. 2nd Int. Conf. IEEE Nigeria Comput. Chapter (NigeriaComputConf), Oct. 2019, pp. 1–8.
[24] T. Gangavarapu, C. D. Jaidhar, and B. Chanduka, ''Applicability of machine learning in spam and phishing email filtering: Review and approaches,'' Artif. Intell. Rev., vol. 53, no. 7, pp. 5019–5081, Oct. 2020.
[25] S. Priya, S. Selvakumar, and R. L. Velusamy, ''Evidential theoretic deep radial and probabilistic neural ensemble approach for detecting phishing attacks,'' J. Ambient Intell. Hum. Comput., vol. 14, no. 3, pp. 1951–1975, Jul. 2021.
[26] P. L. Indrasiri, M. N. Halgamuge, and A. Mohammad, ''Robust ensemble machine learning model for filtering phishing URLs: Expandable random gradient stacked voting classifier (ERG-SVC),'' IEEE Access, vol. 9, pp. 150142–150161, 2021.
[27] A. Ozcan, C. Catal, E. Donmez, and B. Senturk, ''A hybrid DNN-LSTM model for detecting phishing URLs,'' Neural Comput. Appl., vol. 35, no. 7, pp. 4957–4973, Aug. 2021.
[28] S.-J. Bu and H.-J. Kim, ''Optimized URL feature selection based on genetic-algorithm-embedded deep learning for phishing website detection,'' Electronics, vol. 11, no. 7, p. 1090, Mar. 2022.
[29] V. Zeng, S. Baki, A. E. Aassal, R. Verma, L. F. T. De Moraes, and A. Das, ''Diverse datasets and a customizable benchmarking framework for phishing,'' in Proc. 6th Int. Workshop Secur. Privacy Anal., Mar. 2020, pp. 35–41.
[30] A. Ihsan and E. Rainarli, ''Optimization of k-nearest neighbour to categorize Indonesian's news articles,'' Asia–Pacific J. Inf. Technol. Multimedia, vol. 10, no. 1, pp. 43–51, Jun. 2021.
[31] Y. A. Alsariera, V. E. Adeyemo, A. O. Balogun, and A. K. Alazzawi, ''AI meta-learners and extra-trees algorithm for the detection of phishing websites,'' IEEE Access, vol. 8, pp. 142532–142542, 2020.
[32] E. Sukawai and N. Omar, ''Corpus development for Malay sentiment analysis using semi supervised approach,'' Asia–Pacific J. Inf. Technol. Multimedia, vol. 9, no. 1, pp. 94–109, Jun. 2020. [33] C. L. Tan, ''Phishing dataset for machine learning: Feature evaluation,'' Mendeley Data, V1, 2018, doi: 10.17632/h3cgnj8hft.1.
[34] X.-Y. Lu, M.-S. Chen, J.-L. Wu, P.-C. Chang, and M.-H. Chen, ''A novel ensemble decision tree based on under-sampling and clonal selection for web spam detection,'' Pattern Anal. Appl., vol. 21, no. 3, pp. 741–754, Aug. 2018.
[35] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, ''A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches,'' IEEE Trans. Syst., Man C, Appl. Rev., vol. 42, no. 4, pp. 463–484, Jul. 2012.
[36] E. S. Gualberto, R. T. De Sousa, T. P. D. B. Vieira, J. P. C. L. Da Costa, and C. G. Duque, ''From feature engineering and topics models to enhanced prediction rates in phishing detection,'' IEEE Access, vol. 8, pp. 76368–76385, 2020.
[37] H. Zhang, G. Liu, T. W. S. Chow, and W. Liu, ''Textual and visual contentbased anti-phishing: A Bayesian approach,'' IEEE Trans. Neural Netw., vol. 22, no. 10, pp. 1532–1546, Oct. 2011.
[38] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Amsterdam, The Netherlands: Elsevier, 2017.
[39] T. Saito and M. Rehmsmeier, ''The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,'' PLoS ONE, vol. 10, no. 3, pp. 1–21, Mar. 2015.
[40] T. Fawcett, ''An introduction to ROC analysis,'' Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, Dec. 2006.
[41] A. E. Aassal, L. Moraes, S. Baki, A. Das, and R. Verma, ''Anti-phishing pilot at ACM IWSPA 2018: Evaluating performance with new metrics for unbalanced datasets,'' in Proc. Anti-Phishing Shared Task Pilot 4th ACM IWSPA, 2018, pp. 2–10.
[42] H. A. Alshalabi, S. Tiun, and N. Omar, ''A comparative study of the ensemble and base classifiers performance in Malay text categorization,'' Asia– Pacific J. Inf. Technol. Multimedia, vol. 6, no. 2, pp. 53–64, Dec. 2017.
[43] N. Japkowicz and M. Shah, Evaluating Learning Algorithms. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[44] R. Gowtham and I. Krishnamurthi, ''PhishTackle—A web services architecture for anti-phishing,'' Cluster Comput., vol. 17, no. 3, pp. 1051–1068, Sep. 2014.