URL-BASED HYBRID MACHINE LEARNING PHISHING DETECTION SYSTEM

Neeli Sarvani
Student
Department of CSE
Koneru Lakshmaiah Education Foundation,
Vaddeswaram
AP, India.
neelisarvani@gmail.com

N. Ravinder
Asst. Prof.
Department of CSE
Koneru Lakshmaiah Education Foundation,
Vaddeswaram
AP, India.
ravindernellutla@kluniversity.in

Abstract: By using a large dataset based on the fishing url, they begin attacks on the Internet. This work is trying to detect cyber threats such as different types of machine learning methods, such as Decision Tree, Linear Regression, Random Forest, Naive Bayes, Gradient Boosting Classifier, Support Vector Classifier, and a new hybrid LSD model. As an extension, we have used a hybrid model that combines predictions of many individual models. This is achieved through rigid cross-fold validation and Grid Search Hyper parameter Optimization. Such a model makes classification stacking, which uses a dress technique to combine Random Forest Classifier and MLP Classifier, two base classifiers. As a meta-estimator, it appoints the LGBM classifier to reach the final prediction, which extends the project's ability to perform better classification. The effect of the model is evaluated using matrix including F1-score, recall, accuracy, and precision. The results show that the Hybrid LSD model effectively reduces the risk of fish attacks and provides strong protection against the ever -changing cyber danger. This study contributes to the development of better cyber security measures, and shows how you can improve the safety of the Internet by learning machine.

"Index Terms: *Phishing attacks, Machine learning algorithms, Cyber threat detection, Hybrid LSD model, Cyber security measures".*

I. INTRODUCTION

A smart danger on web is fishing, where thieves stated as legitimate businesses or websites in an attempt towards obtain important information (such as password, credit card details or personal history). In order towards avoid financial losses & towards ensure that sensitive information does not get into wrong hands, it is extremely important towards detect fishing efforts. fight against fishing is equipped among machine learning, a kind of help. It detects fish efforts through analyzing data from large & scale, finding patterns in it & using this knowledge. An important advantage is that ML systems abide extremely flexible, as they can meet new & changing fishing efforts. Checking URL, or site address, is a technique for identifying fishing efforts. incorrect or misspelled domain names or excessive number of underdoman enemies abide

common url tabs. Such a nice irregularity is quite easy towards detect machine learning algorithm. successful fish declaration system can easily endure combined among a wide range of web-based applications, including email clients, corporate networks & browser. These interconnected systems abide always looking for new fishing efforts & protecting users immediately against them.

Internet has evolved into an integrated component of modern life, thanks towards progress of communication & information technology. It facilitates profits of life reform opportunities in communication, entertainment, education, retailers & many other fields. Criminals see Internet as a way towards take their physical crimes online because our online life develops. While there abide many positive benefits towards using Internet, there abide also some negatives, such as Annomination provides it towards users. According towards Research through Partsmouth University (2016), Raguchi & Robila (2006), & Hong (2012), individuals & organizations lose millions of dollars every day. Cybercrime, one of most basic forms of fishing, grows at an exponential speed. [12] Only players have spread among expansion of Internet time. among extensive use of Internet, fishing attacks have increased in popularity. One of main methods is utilized through playing weaknesses. People who suffer from fishing fraud sometimes come for fraud as websites used towards fool them or look like other popular sites. For most parts, unskilled internet users cannot show difference between legitimate & dangerous websites. Because of this, fishing blacklists were developed. Fishing blacklists abide databases among malicious software maintained through experts. They enable ordinary people towards learn about fishing sites that they can travel. [18]

II. LITERATURE SURVEY

Introduction towards "Phishpedia", A Groundbreaking Logo-Based Phishing Identity System That Stands For Its Remarkable Accuracy & Minimal Impact on Runtime. author of system, Y. Lynn, R. Liu, D. M. Diwakaran, J. Y. Ng, Q. Z Chan, Y. Lu, Y. C., F. Zhang & JS abide Dongs. When Compared towards Existing Methods, Our State -of -The -Art Deep Learning System Achieves Better Results in Correctly Identifying Phishing Efforts, Especially When It Comes towards Recognition & Matching. Not Only Does It Perform Better than Current Methods, But It Also Finds Fishing Sites That Were Not Before, Making Defense Much Stronger Against Fishing Efforts. When it comes towards improving cyber security, Phishpedia is in its own league. Negative: effectiveness of Phishpedia depends on presence & quality of people on websites. Sometimes there is regular upgrading, & maintenance must endure ahead of changed phishing strategies. [1]

Using Artificial Nerve Network (ANNs) towards Analyze HTML & URL Properties, Introduction A Groundbreaking Algorithm for Shirazi, Hens & Raya Mobile -Friendly Fishing Detection. Modern deep transformers such as Burt, Electra, Robert & Mobilebert have been included in their method of effective learning from URL text. state -Art -art system effectively administers fast training, smooth maintenance & real -time delaying on mobile devices. It Guarantes Top-Oriented

Performance, Strength's Prevention Against Fishing Attacks, & Maximizes Resource Usage For Better Mobile Cyber Security. Negative: Complicated fishing on actual pages cannot go towards anyone's attention if URL is only way towards detect. availability & quality of pre - trained transformers may vary. [2]

A. Akanka's dissertation examines SSL certificates used through fishing spots, analyzing properties of attackers & developing an auto-detection system that uses these aspects. Research introduces a groundbreaking SSL certificate-based fish-declaration system that uses decision tree [4] machine learning due towards its openness & efficiency. system claims excellent accuracy & has a user-friendly web API. This letter presents a holistic approach towards cyber security problems & highlights need for future adjustment towards develop a fishing strategy & guarantee continuous system updates. Lack of Missing: If malicious actors find ways towards copy real SSL certificates, efficiency of system can endure compromised. scalability of system towards handle many domains is barely affected. third

Logistics regression, decision tree [4], neural networks & Gausian Bole Bayes, H. Shahiriyar & S. Using machine learning methods that Nimgadda focuses on Network Intrusion Detection Systems (IDS) in their joint efforts. Finding specific & unusual network behavior, especially TCP/IP layers, is primary goal of research. Decision Tree [4] works on a publicly available dataset, but authors emphasized need for evaluation & real world testing towards confirm its efficiency & accuracy in Real -world network infiltration scenarios. Lack of deficiencies: results may not endure reflective towards real world or changing dangers. This algorithm choice is not overall; Alternative approaches can produce different results. [4]

In his groundbreaking work, A. of. Dutta creates a refined system for detecting fishing sites using a monitored machine learning technique random forest [4]. This process examines & selects relevant features that carefully separate fishing sites. solution, when distributed as a smart browser plugin, detects phishing sites among 98.8 percent accuracy, towards handle human deficiencies in online security, continuously. main goal is towards improve Internet security & give users a strong security against potential cyber dangers, even though it causes a false alarm sometimes. Resistance: quality of compatibility plants for new phishing strategies is influenced through quality of facilities. Users may lose confidence in results if they lie. [5]

III. METHODOLOGY

A) System Architecture

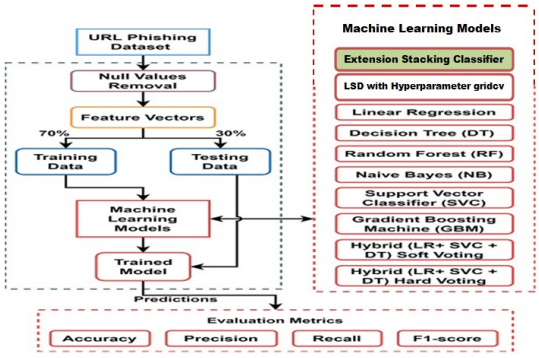


Fig 1: System Architecture

Proposed work: proposed technique uses a state-of-the-art hybrid machine learning method

towards identify fish attacks using URL properties. This strengthens rescue against attacks & defends users through using a wide range of machine learning techniques. Grid Search Hyperparameter optimization & verification of cross -folding, when common, prediction accuracy is greatly improved. expansion of project adds a stacking classifies towards create a hybrid model, which further improves possibilities. In this artist's contingent, two base classifies - Random Forest [4] Classifier & MLP Classifier - towards create a strong general model. classification performance of project is improved through incorporating LGBM classification as a meta-estimator, which refines final prediction. This all -encompassing method represents an important step in cyber security through guaranteeing an effective & reliable defense system against fish attacks.

B) Dataset Collection

"URL-based Phishing dataset" is a set of data created among intention of researching & creating devices towards identify & distinguish between phishing & authentic URLs. It came from Kaggle, a well -known website for data sets & computer science challenges.

This is a broad overview of dataset:

Name: Phishing Dataset based on URLs

Kaggle is source.

goal is towards make phishing detection system research & development easier.

Size: Compiles information from more than 11,000 websites.

Format: Shown in vector form, suggesting that every URL is probably a collection of characteristics.

The dataset may endure carried out so that each entry or example corresponds towards a URL. Machine learning models can use each URL properties (vector form) towards determine whether a certain URL is valid or is associated among fishing.

URL length, use of https, inclusion of specific keywords, domain age & other relevant indicators abide examples of specific properties in a dataset for phishing detection. These features abide necessary towards teach machine learning models that can distinguish between authentic & scams.

```
data = pd.read_csv("archive/phishing.csv")
data.head()
```

	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

C) Pre-processing

Using Pandas Data Frame:: towards clean, replace & prepare dataset at this stage, we use panda, a strong python data manipulation library. This includes addressing missing values, changing data formats & organizing information for further modeling or analysis.

Visualization among Seaborn & Matplotlib: towards learn more about properties of dataset, we use seborn & food plotelib towards create visualizations such as diagrams & graphs. In order towards make well -informed decisions for further analysis, this phase helps us understand distribution, relationships & patterns present in data.

Label Processing: towards translate classified label into numerical values, we use a label codance, a pre

-rosaring method. Given that machine learning model usually requires numerical entrance, it is necessary. dataset is guaranteed through possibility of model for understanding & learning models from classified data.

Feature Selection: most relevant properties from dataset abide found & selected at this stage. through focusing most useful variable & reducing noise, construction choices abide needed towards increase model performance. model can endure found using methods such as statistical testing, correlation analysis or machine learning techniques.

D) **Training & Testing**

Our first machine learning model, Model 9, was used on preprossed data sets during initial phase of project for analysis & interpretation. After that, we wanted towards improve accuracy of our predictions during extension phase, so we created a hybrid model, which combined output from other models. through combining best features of many models, this new strategy expects towards increase accuracy of our predictions. At same time, we created an authentication dream, flask -based fronts, so that users interact among model. A user -friendly & available interface is provided through this friend, so users can enter & easily rebuild predictions. towards train first indicated machine learning models towards detect complex computer patterns & correlations, we will use Predicable Dataset. This will endure basis for our project. A separate test is fully evaluated on a separate test data set after training process. towards evaluate effect of these models in identifying fishing urls, performance matrix is used as recalling, accuracy & F1 score carefully. In order towards guarantee that models abide accurate & reliable, & therefore suitable for use in real world, this is a perfect assessment process is an important quality assurance step. aim of our project is towards provide reliable & advanced phishing url identity using this all-encompassing methodology.

E) Algorithms.

In order towards integrate predictions from **Random Forest** [4] classify & MLP classify, which is base classify, project uses an artist -fee technique called Stacking Classifier. As a meta-estimator, it appoints LGBM classificationer towards reach final prediction, which extends project's ability towards perform better classification.

Decision Tree [4], Logistics Regression [5] & Vector Machine [6] -GridCV create a hybrid classification model through integrating best properties of multiple algorithms among **Hyperpieme LSD (**logistics region, support vector, degeneration wood [4]). When it comes towards classification jobs, GRIDCV actually shines because this model undergoes hyperpieme combinations towards increase performance.

To produce classification decisions include **Hybrid LSD (HARD)** model logistic region, supporting Vektader & a hard voice technique among decision tree [4] algorithms. This method improves accuracy & strength in different classification problems through allowing each component model towards present its prediction & then a majority towards make final decisions.

As a computer classification model, **Hybrid LSD (Soft)** integrates logistic region, supporting voter & decision -making three [4] using a soft poll. among

ability towards handle different types of data & increase accuracy of classification functions, it does predicts through taking advantage of abilities of each model.

A clutch machine learning method, known as **gradient boosting**, uses strength of many weak students - usually decides trees [4] towards create a prediction model. This is achieved through taking full attention towards mistakes made through previous models & through changing predictions accordingly. result is a strong & accurate future saying model that works very well in many tasks, such as classification & regression.

A **ensemble learning technique** that uses many decisions trees [4] is random forest [4] towards generate predictions. towards work for this, several decisions abide three [4] arbitrarily trained on data & average their predictions. Both classification & regression tasks abide performed strongly through this outfit strategy, which improves accuracy & reduces overfitting.

A **decision tree** is a type of machine learning model that tries towards classify or predict results through repeating data in most relevant feature. This method is easy towards understand both successful & simple because it creates a structure similar towards a tree, where each node stands for a function for a possible conclusion & for each branch.

Machine learning models such as **Support Vector Classifier (SVC)** maximize margin between data classes through determining optimal area (hyperplan) towards divide them. Both binary & multi -class classification functions abide effectively controlled through identifying effective support vectors.

As a classification algorithm, **logistic regression** estimates possibility that an entrance is of a certain category. towards classify entrance towards one of two or more categories, it uses an area based on probability point that occurs from Sigmoid feature, which maps input functions up towards 0. Through workouts, model teaches coefficients that allow it towards produce correct classification that fits data properly.

Employing "naive" assumption of feature independence, probabilistic classification method called **Naive Bayes** uses Bayes' theorem. through applying probabilities for each feature towards a data point, it can calculate probability that it will endure in a particular class. Text classification, spam filtering, & other uses where feature independence is an acceptable approximation abide where Naive Bayes excels.

IV. EXPERIMENTAL RESULTS

A) Comparison Graphs → Accuracy, Precision, Recall, f1 score

Accuracy: A test ability towards make a proper difference between healthy & sick cases is a measure of accuracy. We can determine accuracy of a test through calculating proportion of cases undergoing proper positivity & genuine negative. It is possible towards express this mathematically:

$$\text{“}Accuracy = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}(1)\text{"}$$

Precision: relationship between events or tests certain abide properly classified towards anyone classified as positive is called accurate. Therefore, there is a formula considering determining accuracy:

$$"Precision = \frac{\text{True Positive}}{\text{True Positive } + \text{False Positive}} (2)"$$

Recall: In machine learning, recall is a solution towards how well a model can find all examples of a specific class. ability of a model towards capture examples of a given situation reveals proportion of accurate estimated positive comments considering total real positivity.

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

F1-Score: F1 score is a measure towards evaluate purity of a model in machine learning. It takes memory & accuracy of a model & mixes them. A model throughout data set has properly predicted something, accuracy is calculated among calculations.

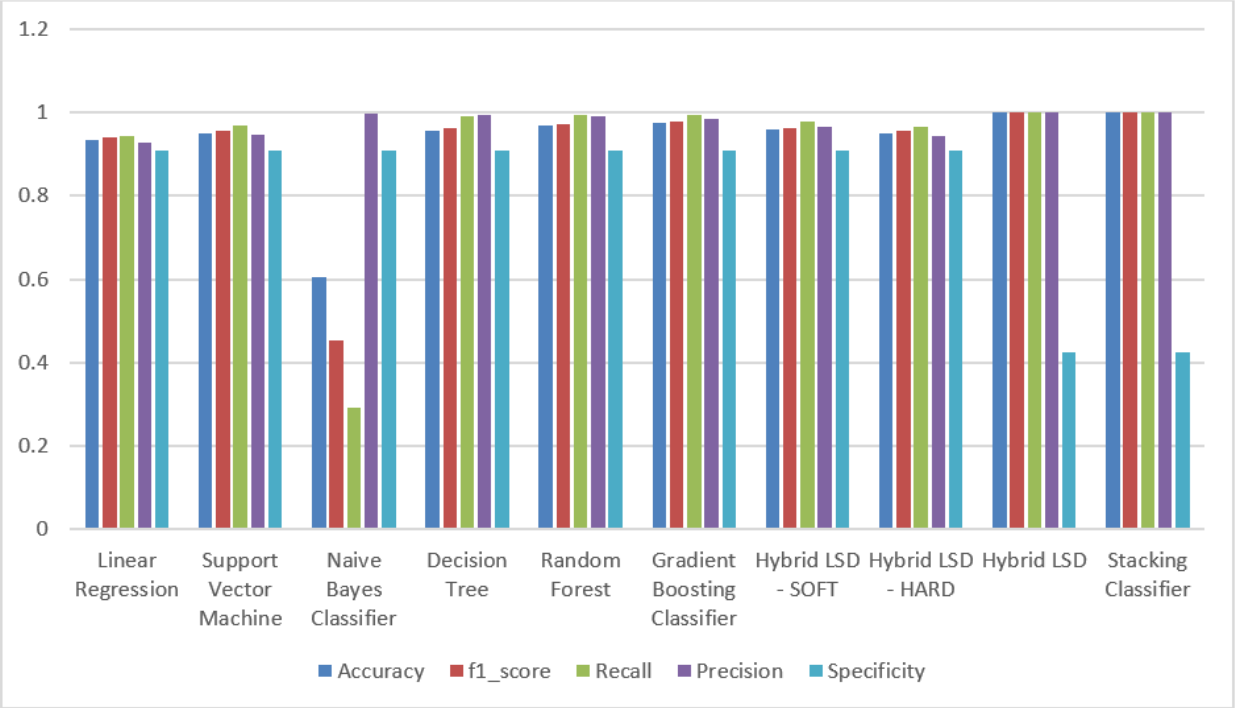
$$"F1\ Score = 2 * \frac{Recall \text{ X } Precision}{Recall + Precision} * 100(1)"$$

Table (1) compares algorithms' performance based on four metrics: accuracy, precision, recall, & F1 - Score. Stacking Classifier routinely beats all competing algorithms on every single metric. You may also see a comparison of other algorithms' metrics in tables.

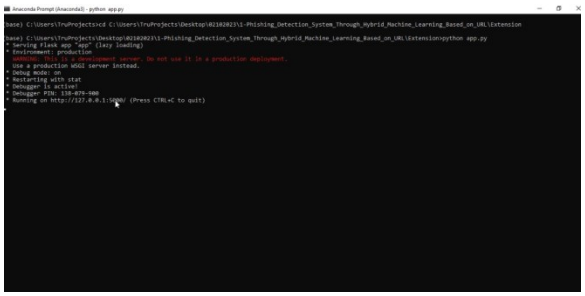
Table.1 Performance Evaluation Table

ML Model	Accuracy	f1_score	Recall	Precision	Specificity
Linear Regression	0.934	0.941	0.943	0.927	0.909
Support Vector Machine	0.951	0.957	0.969	0.947	0.909
Naive Bayes Classifier	0.605	0.454	0.292	0.997	0.909
Decision Tree	0.957	0.962	0.991	0.993	0.909
Random Forest	0.969	0.972	0.993	0.990	0.909
Gradient Boosting Classifier	0.974	0.977	0.994	0.986	0.909
Hybrid LSD - SOFT	0.959	0.964	0.977	0.965	0.909
Hybrid LSD - HARD	0.950	0.956	0.967	0.945	0.909
Hybrid LSD	1.000	1.000	1.000	1.000	0.426
Stacking Classifier	**1.000**	**1.000**	**1.000**	**1.000**	**0.426**

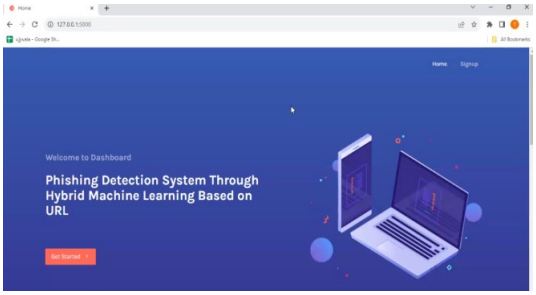
Graph.1 Comparison Graph



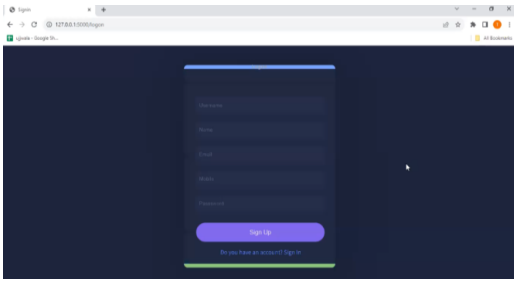
The graph shows that accuracy is blue, recall is green, precision is purple, & specificity is sky blue. F1 score is red. All measures suggest that Stacking Classifier outperforms other models, among highest values achieved. These results abide graphically shown in graphs up above.



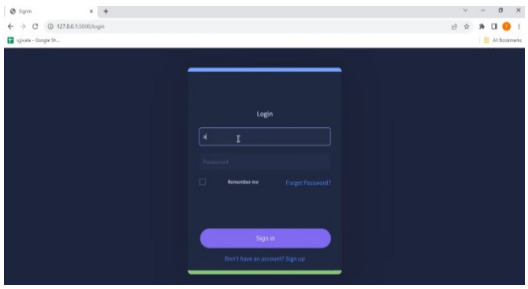
"Fig 2: URL Link towards Web Page"



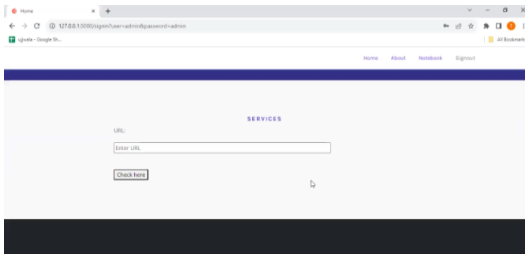
"Fig 3: Home page"



"Fig 4: User Signup page"

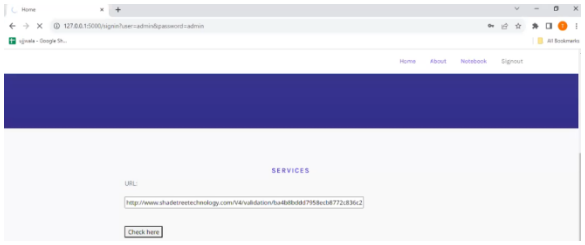


"Fig 5: User Sign in Page"

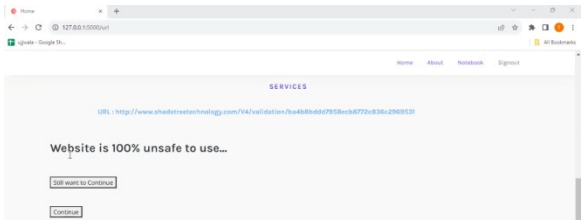


"Fig 6: Enter URL"

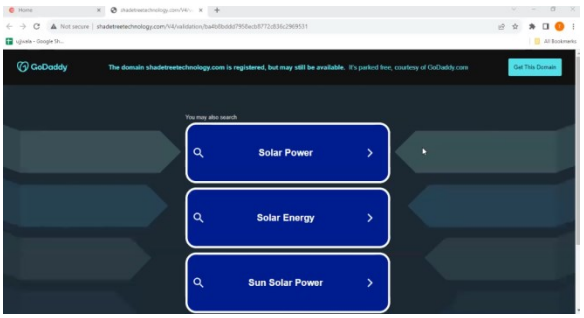
"Fig 7: Sample data for testing"



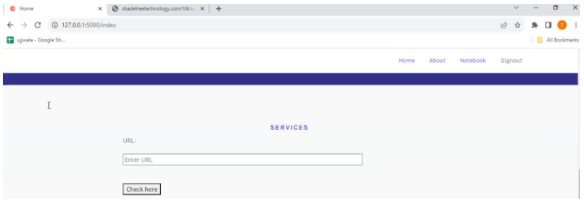
"Fig 8: Entered URL"



"Fig 9: URL result unsafe 100%"

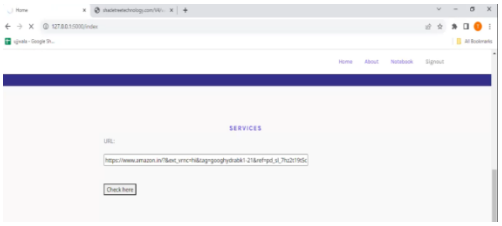


"Fig 10: Search Other Urls too"

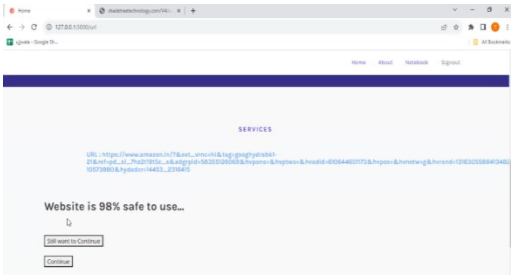


"Fig 11: Enter New URL"

"Fig 12: Sample data for testing"



"Fig 13: Entered New URL"



"Fig 14: URL result page (safe/ unsafe)"

V. CONCLUSION

With help of a hybrid machine learning strategy, project was able towards prefer URL properties for fish decisions, & it was a success. algorithm improved its accuracy & efficiency using different types of models, including decision trees, random forests, vector classifies & an LSD-based stacking classifies. through using an expansion stacking classifies, fish declaration system was greatly improved, which was responsible for its high accuracy & F-point. This all-dedicated method provides a strong defense against refined fishing efforts, & solves a big question in cyber security. A better degree of adaptability towards develop fishing techniques was ensured through integrating different machine learning models, which brought diversity towards system's abilities. As a result of its positive effects on accuracy & efficiency, project has opportunity towards strengthen cyber security measures & make a significant contribution in fight against cyber threats. developed system shows its ability towards become real -world applications towards protect sensitive information & reduce risk associated among cyber threats. It stands as a strong defense mechanism against quickly sophisticated fishing work.

Constant improvement & adjustments in new fishing techniques abide in future plans for this project. In order towards improve active safety opportunities of system, future studies can use deep learning, behavioral analysis & real -time panties intelligence. Working among professionals for cyber security & other industry players can also help create a strong solution. Access towards this system can endure expanded through checking blame environment & distribution on Internet of Things devices, as well as through developing a user -friendly interface. A top modern solution in ever -existing scope of cyber security, model undergoes constant changes towards change hazard landscape, guarantees its constant effect.

REFERENCES

[1] Y. Lin, R. Liu, D. M. Divakaran, J. Y. Ng, Q. Z. Chan, Y. Lu, Y. Si, F. Zhang, & J. S. Dong, ‘‘Phishpedia: A hybrid deep learning based approach towards visually identify phishing webpages,’’ in Proc. 30th USENIX Secur. Symp. (USENIX Security), 2021, pp. 3793–3810.

[2] H. Shirazia, K. Haynesb, & I. Raya, ‘‘Towards performance of NLP transformers on URL-based phishing detection for mobile devices,’’ Int. Assoc. Sharing Knowl. Sustainability (IASKS), Tech. Rep., 2022.

[3] A. Akanchha, ‘‘Exploring a robust machine learning classifier for detecting phishing domains using SSL certificates,’’ Fac. Comput. Sci., Dalhousie Univ., Halifax, NS, Canada, Tech. Rep. 10222/78875, 2020.

[4] H. Shahriar & S. Nimmagadda, ‘‘Network intrusion detection for TCP/IP packets among machine learning techniques,’’ in Machine Intelligence & Big Data Analytics for Cybersecurity Applications. Cham, Switzerland: Springer, 2020, pp. 231–247.

[5] A. K. Dutta, ‘‘Detecting phishing websites using machine learning technique,’’ PLoS ONE, vol. 16, no. 10, Oct. 2021, Art. no. e0258361.

[6] A. K. Murthy & Suresha, ‘‘XML URL classification based on their semantic structure orientation for web mining applications,’’ Proc. Comput. Sci., vol. 46, pp. 143–150, Jan. 2015.

[7] A. A. Ubing, S. Kamilia, A. Abdullah, N. Jhanjhi, & M. Supramaniam, ‘‘Phishing website detection: An improved accuracy through feature selection & ensemble learning,’’ Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, pp. 252–257, 2019.

[8] A. Aggarwal, A. Rajadesingan, & P. Kumaraguru, ‘‘PhishAri: Automatic realtime phishing detection on Twitter,’’ in Proc. eCrime Res. Summit, Oct. 2012, pp. 1–12.

[9] S. N. Foley, D. Gollmann, & E. Snekkenes, Computer Security— ESORICS 2017, vol. 10492. Oslo, Norway: Springer, Sep. 2017.

[10] P. George & P. Vinod, ‘‘Composite email features for spam identification,’’ in Cyber Security. Singapore: Springer, 2018, pp. 281–289.

[11] H. S. Hota, A. K. Shrivas, & R. Hota, ‘‘An ensemble model for detecting phishing attack among proposed remove-replace feature selection technique,’’ Proc. Comput. Sci., vol. 132, pp. 900–907, Jan. 2018.

[12] G. Sonowal & K. S. Kuppusamy, ‘‘PhiDMA—A phishing detection model among multi-filter approach,’’ J. King Saud Univ., Comput. Inf. Sci., vol. 32, no. 1, pp. 99–112, Jan. 2020.

[13] M. Zouina & B. Outtaj, ‘‘A novel lightweight URL phishing detection system using SVM & similarity index,’’ Hum.-Centric Comput. Inf. Sci., vol. 7, no. 1, p. 17, Jun. 2017.

[14] R. Ø. Skotnes, ‘‘Management commitment & awareness creation—ICT safety & security in electric power supply network companies,’’ Inf. Comput. Secur., vol. 23, no. 3, pp. 302–316, Jul. 2015.

[15] R. Prasad & V. Rohokale, ‘‘Cyber threats & attack overview,’’ in Cyber Security: Lifeline of Information & Communication Technology. Cham, Switzerland: Springer, 2020, pp. 15–31.

[16] T. Nathezhtha, D. Sangeetha, & V. Vaidehi, ‘‘WC-PAD: Web crawling based phishing attack detection,’’ in Proc. Int. Carnahan Conf. Secur. Technol. (ICCST), Oct. 2019, pp. 1–6.

[17] R. Jenni & S. Shankar, ‘‘Review of various methods for phishing detection,’’ EAI Endorsed Trans. Energy Web, vol. 5, no. 20, Sep. 2018, Art. no. 155746.

[18] (2020). Accessed: Jan. 2020. [Online]. Available: https://catches-of-themonth-phishing-scams-for-january-2020

[19] S. Bell & P. Komisarczuk, ‘‘An analysis of phishing blacklists: Google safe browsing, OpenPhish, & PhishTank,’’ in Proc. Australas. Comput. Sci. Week Multiconf. (ACSW), Melbourne, VIC, Australia. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–11, Art. no. 3, doi: 10.1145/3373017.3373020.

[20] A. K. Jain & B. Gupta, ‘‘PHISH-SAFE: URL features-based phishing detection system using machine learning,’’ in Cyber Security. Switzerland: Springer, 2018, pp. 467–474.

[21] Y. Cao, W. Han, & Y. Le, ‘‘Anti-phishing based on automated individual white-list,’’ in Proc. 4th

ACM Workshop Digit. Identity Manage., Oct. 2008, pp. 51–60.

[22] G. Diksha & J. A. Kumar, ‘‘Mobile phishing attacks & defence mechanisms: State of art & open research challenges,’’ Comput. Secur., vol. 73, pp. 519–544, Mar. 2018.

[23] M. Khonji, Y. Iraqi, & A. Jones, ‘‘Phishing detection: A literature survey,’’ IEEE Commun. Surveys Tuts., vol. 15, no. 4, pp. 2091–2121, 4th Quart, 2013.

[24] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, & J. Downs, ‘‘Who falls for phish? A demographic analysis of phishing susceptibility & effectiveness of interventions,’’ in Proc. SIGCHI Conf. Hum. Factors Comput. Syst., Apr. 2010, pp. 373–382.

[25] P. Prakash, M. Kumar, R. R. Kompella, & M. Gupta, ‘‘PhishNet: Predictive blacklisting towards detect phishing attacks,’’ in Proc. IEEE INFOCOM, Mar. 2010, pp. 1–5.

[26] P. K. Sandhu & S. Singla, ‘‘Google safe browsing-web security,’’ in Proc. IJCSET, vol. 5, 2015, pp. 283–287.

[27] M. Sharifi & S. H. Siadati, ‘‘A phishing sites blacklist generator,’’ in Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl., Mar. 2008, pp. 840–843.

[28] S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, & C. Zhang, ‘‘An empirical analysis of phishing blacklists,’’ in Proc. 6th Conf. Email Anti-Spam (CEAS), Mountain View, CA, USA. Pittsburgh, PA, USA: Carnegie Mellon Univ., Engineering & Public Policy, Jul. 2009.

[29] Y. Zhang, J. I. Hong, & L. F. Cranor, ‘‘Cantina: A content-based approach towards detecting phishing web sites,’’ in Proc. 16th Int. Conf. World Wide Web, May 2007, pp. 639–648.

[30] G. Xiang, J. Hong, C. P. Rose, & L. Cranor, ‘‘CANTINA+: A featurerich machine learning framework for detecting phishing web sites,’’ ACM Trans. Inf. Syst. Secur., vol. 14, no. 2, pp. 1–28, Sep. 2011