[1]**Mr. M. Gokulkannan,**

[2]**V. Rohithkumar Reddy,**

[3]**Singu Vinod Kumar,**

[4]**Shaik Abbas**

Phishing Detection System Through Hybrid Machine Learning Based on URL

Journal of Electrical Systems

Abstract - Phishing attacks on the internet by leveraging a comprehensive phishing URL-based dataset. Employing an array of machine learning algorithms, including Decision Tree [4], Linear Regression[4], Random Forest [4], Naive Bayes, Gradient Boosting Classifier, Support Vector Classifier, and a novel hybrid LSD model, the study aims to enhance cyber threat detection. Through meticulous cross-fold validation and Grid Search Hyper parameter Optimization, As an extension we have applied a hybrid model by combining the predictions of multiple individual models like Stacking Classifier, an ensemble technique, to combine predictions from Random Forest [4] Classifier[4] and MLP Classifier as base classifiers. It uses LGBM Classifier as a meta-estimator to make the final prediction, extending the project's capabilities for improved classification performance. Evaluation metrics such as precision, accuracy, recall, and F1-score are employed to assess model effectiveness. The results underscore the efficacy of the hybrid LSD model in mitigating phishing threats, providing a robust defense mechanism against evolving cyber threats. This research contributes to the advancement of cybersecurity measures and demonstrates the potential of machine learning in bolstering online security.

Keywords:- *Phishing attacks, Machine learning algorithms, Cyber threat detection, Hybrid LSD model, Cyber security measures*

I. INTRODUCTION

Phishing is a sneaky online threat where cybercriminals impersonate trustworthy sources, like banks or popular websites, to trick individuals into revealing sensitive information such as passwords,credit card numbers, or personal details. Detecting phishing attempts is crucial because it helps prevent valuable data from falling into the wrong hands and protects against financial losses. Machine Learning, a type of artificial intelligence, is highly effective in the fight against phishing. It works by examining large volumes of data, learning patterns from it, and using this knowledge to identify phishing attempts. One significant advantage is that ML systems can adapt to new and evolving phishing techniques, making them very robust. One way to detect phishing is by analyzing website addresses or URLs. Phishers often make mistakes in URLs, like using misspelled domain names or adding too many subdomains. Machine Learning models excel at spotting these subtle irregularities. nEffective phishing detection systems can be seamlessly integrated into various online tools such as web browsers, email clients, or corporate networks. These integrated systems work in real-time, continuously scanning incoming data for potential phishing threats and providing immediate protection to users.

In this technological era, the Internet has made its way to become an inevitable part of our lives. It leads to many convenient experiences in our lives regarding communication, entertainment, education, shopping and so on. As we progress into online life, criminals view the Internet as an opportunity to transfer their physical crimes into a virtual environment. The Internet not only provides convenience in various aspects but also has its downsides, for example, the anonymity that the Internet provides to its users.[7] With the exponential growth of Internet users, incidents of cybercrimes are also correspondingly expanding in a rapid way. Both people and associations are losing millions worth daily (Hong, 2012, Ragucci and Robila, 2006, University of Portsmouth, 2016). Phishing is one of

[1] Department Of Computer Science And Engineering, Mahendra Engineering College Namakkal,Tamilnadu,India.
singuvinodkumar82@gmail.com

the basic cybercrimes, which is exponentially increasing day by day.[12] With the rise of the internet era, malicious actors have also been increasing in number. Phishing attacks became a trend in the age where websites are part of everyday life. The exploitation of human weaknesses is a major factor in the victimization of users. Phishing websites are set up to look legitimate or similar to other well established websites, causing the victims of the scam to fall prey to it. Since the malicious sites are sometimes indistinguishable from the legitimate source, nonprofessional users of the internet cannot distinguish between the two. This led to the creation of phishing blacklists. Phishing blacklists are software datasets that are kept by professionals. They allow nonprofessional users to become aware of potential phishing websites that they might navigate to.[18]

II. LITERATURE SURVEY

Y. Lin, R. Liu, D. M. Divakaran, J. Y. Ng, Q. Z. Chan, Y. Lu, Y. Si, F. Zhang, and J. S. Dong present "Phishpedia," a pioneering logo-based phishing identification system characterized by exceptional accuracy and minimal runtime impact. This innovative deep learning system excels in precise phishing identification, particularly in logo recognition and matching, surpassing current methods. Its proficiency not only outperforms existing techniques but also uncovers previously unidentified phishing sites, thereby fortifying defense against phishing attacks. Phishpedia stands out as a unique and powerful tool for enhancing cybersecurity. *Cons:* Phishpedia's performance relies on logo availability and quality on webpages. Ongoing updates and maintenance are essential for adapting to evolving phishing tactics.[1]

Shirazi, Haynes, and Raya present a pioneering mobile-friendly phishing detection algorithm leveraging Artificial Neural Networks (ANNs) to scrutinize URL and HTML features. Their approach integrates cutting-edge deep transformers such as BERT, ELECTRA, RoBERTa, and MobileBERT for efficient learning from URL text. The innovative system facilitates swift training, seamless maintenance, and real-time deployment on mobile devices, addressing mobile security challenges effectively. This ensures competitive performance, establishing a robust defense against phishing threats while optimizing resource utilization for enhanced cybersecurity on mobile platforms. *Cons:* Limited to URL detection may miss complex phishing within legitimate pages. Depends on pre-trained transformers, subject to variations in availability and quality. [2]

The thesis by A. Akanchha delves into the realm of SSL certificates within phishing sites, scrutinizing attacker attributes and crafting an auto-detection system reliant on SSL certificate features. Embracing Decision Tree [4] machine learning for its transparency and efficacy, the research presents a pioneering SSL certificate-based phishing detection system, boasting impressive accuracy and a user-friendly Web API. The work underscores the need for future adaptations to combat evolving phishing techniques and ensure ongoing system updates, providing a comprehensive approach to cybersecurity challenges. *Cons:* The system's effectiveness relies on SSL certificate attributes, which could be undermined if attackers develop new methods to mimic genuine certificates. The scalability of the system for managing numerous domains is not extensively discussed.[3]

In the collaborative work of H. Shahriar and S. Nimmagadda, their chapter focuses on Network Intrusion Detection Systems (IDS) leveraging machine learning techniques such as Gaussian Naive Bayes, logistic regression, Decision Tree [4], and neural networks. The study aims to discern normal and anomalous network activities, particularly across TCP/IP layers. Notably, the Decision Tree [4] exhibits commendable performance on public datasets, yet the authors underscore the imperative of real-world testing and scalability assessments for comprehensive validation of its accuracy and efficiency in practical network intrusion detection scenarios. *Cons:* Evaluation may not reflect real-world conditions or evolving attacks. Algorithm choice not exhaustive; different methods may yield different results.[4]

A. K. Dutta's innovative approach utilizes Random Forest [4], a supervised machine learning technique, to construct an advanced system dedicated to identifying phishing websites. The method involves meticulous analysis and selection of pertinent features that distinctly define phishing sites. Implemented as an intelligent browser extension, the system achieves an impressive 98.8% accuracy in detecting phishing sites, strategically addressing human vulnerabilities in online security. While occasionally presenting false alerts, the overarching goal is to significantly enhance online security measures and provide users with a robust defense against potential cyber threats. Cons: Feature quality impacts adaptability to new phishing tactics. Potentialfor false results affects user trust. [5]

III. METHODOLOGY

Modules:

- Importing required Packages
- Exploring the dataset - Phishing URLFeature Data
- Data Processing - Using Pandas Data frame
- Visualization using seaborn & matplotlib
- Label Encoding using Label Encoder
- Feature Selection
- Train & Test Split
- Training and Building the model
- Trained model is used for prediction
- Final outcome is displayed through front-end

A) System Architecture

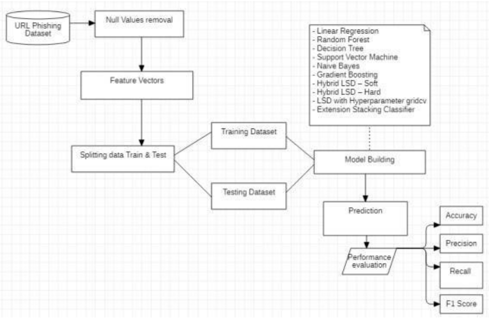


Fig 1: System Architecture

Proposed work

The proposed system employs a cutting-edge hybrid machine learning approach for phishing attack detection based on URL attributes. Leveraging a diverse set of machine learning algorithms, it fortifies defenses against attacks and safeguards users. The integration of cross-fold validation and grid search hyper parameter optimization techniques significantly enhances predictive accuracy. To further bolster its capabilities, Extension of the project introduces a hybrid model through the implementation of aStacking Classifier. This ensemble technique combines predictions from Random Forest [4] Classifier and MLP Classifier as base classifiers, synergistically blending their strengths. The inclusion of LGBM Classifier as a meta-estimator refines the final prediction, elevating the project's classification performance. This comprehensive approach ensures arobust and accurate defense mechanism againstphishing attacks, marking a notable advancement in cybersecurity.

B) Dataset Collection

The "URL-based phishing dataset" is a collection of data designed for the purpose of studying and developing systems to detect and differentiate between phishing and legitimate URLs. It was sourced from Kaggle, a popular platform for datascience competitions and datasets.

Here is a general description of the dataset:*Name*: URL-based Phishing Dataset *Source*: Kaggle

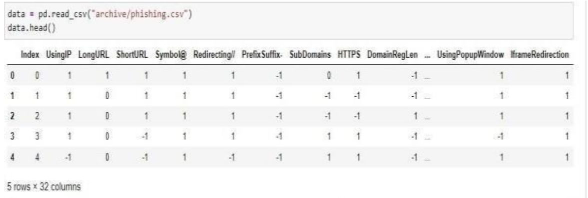
Purpose: To facilitate research and development ofphishing detection systems.

Size: Contains data from over 11,000 websites.

Format: Presented in vector form, implying that eachURL is likely represented as a set of features or attributes.

The dataset is likely structured in a way that each entry or instance corresponds to a URL, and the features (vector form) associated with each URL provide information that machine learning models can use to make predictions about whether a given URL is associated with phishing or is legitimate.

Typical features in a phishing detection dataset might include characteristics such as the length of the URL, the presence of certain keywords, the use of HTTPS, domain age, and other relevant indicators. These features are crucial for training machine learning models to discern patterns that can differentiate between legitimate and phishing URLs.



```
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: In this step, we leverage Pandas, a powerful data manipulation library in Python, to clean, transform, and prepare the dataset. This involves handling missing values, converting data types, and structuring the data for further analysis or modeling.

Visualization with Seaborn & Matplotlib: Utilizing Seaborn and Matplotlib, we create visualizations such as charts and graphs to gain insights into the dataset's characteristics. This step helps us understand patterns, relationships, and distributions within thedata, aiding in informed decision-making for subsequent analysis.

Label Processing: Here, we employ a label encoder,a preprocessing technique, to convert categorical labels into numerical values. This is crucial for machine learning models, as they typically require numerical inputs. Label processing ensures that the models can effectively interpret and learn from the categorical information present in the dataset.

Feature Selection: In this step, we identify and select the most relevant features from the dataset. Feature selection is vital for improving model performance by focusing on the most informative variables and reducing noise. Techniques such as statistical tests, correlation analysis, or machine learning algorithms can be applied to identify the features that contribute significantly to the predictive power of the model.

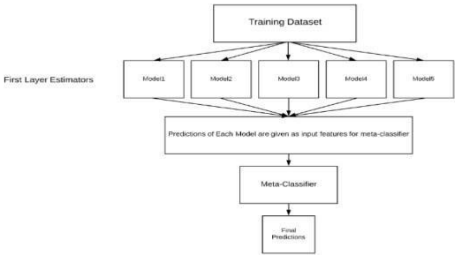
D) Training & Testing

In the initial phase of our project, we implemented the first machine learning model (Model 9) to analyze and interpret the preprocessed dataset. Following this, during the extension phase, we sought to enhance predictive accuracy by creating a hybrid model that amalgamates predictions from multiple models. This innovative approach aims to leverage the strengths of diverse models, fostering improved overall accuracy in our predictions. Simultaneously, we developed a user-friendly Flask-based frontend, fortified with authentication measures, to streamline user interaction with the models. This frontend provides a seamless interface for users to input data and obtain predictions, ensuring a practical and accessible experience. The heart of our project lies in training the aforementioned machine learning models on the preprocessed dataset, allowing them to discern intricate data patterns and relationships. Following the training phase, rigorous evaluations are conducted on a distinct test dataset. Performance metrics such as accuracy, precision, recall, and F1-score are meticulously employed to assess the effectiveness of these models in detecting phishing URLs. This robustevaluation process serves as a crucial quality assurance step, ensuring that the models not only exhibit accuracy but also reliability, affirming their suitability for real-world applications. Through this comprehensive methodology, our project aims to deliver advanced and trustworthy solutions in the realm of phishing URL detection.

E) Algorithms.

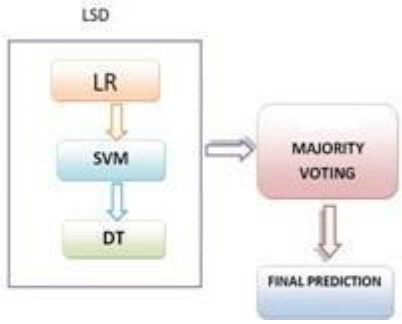
Stacking Classifier:

The project employs a Stacking Classifier, an ensemble technique, to combine predictions from Random Forest [4] Classifier and MLP Classifier as base classifiers. It uses LGBM Classifier as a meta- estimator to make the final prediction, extending the project's capabilities for improved classification performance.



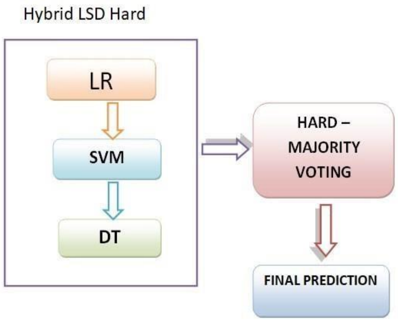
LSD:

The LSD (Logistic Regression, Support Vector Machine, Decision Tree [4]) model with Hyperparameter GridCV is a hybrid classification model that combines the strengths of Logistic Regression, Support Vector Machine, and Decision Tree [4] algorithms, enhancing accuracy and efficiency. GridCV systematically searches through hyperparameter combinations to optimize modelperformance, making it effective in various classification tasks.



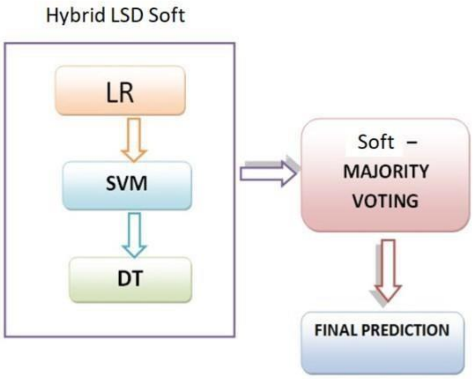
Hybrid LSD (Hard):

The Hybrid LSD (Hard) model combines Logistic Regression, Support Vector Machine, and Decision Tree [4] algorithms with a hard voting technique to make classification decisions. Each component model contributes its prediction, and the final decision is made by majority voting, enhancing accuracy and robustness in various classification tasks.



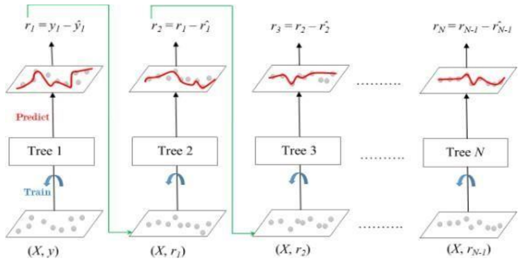
Hybrid LSD (Soft):

The Hybrid LSD (Soft) model combines Logistic Regression, Support Vector Machine, and Decision Tree [4] using soft voting to classify data. It leverages the strengths of each model to make predictions, with the flexibility to handle different types of data and improve accuracy in classification tasks.



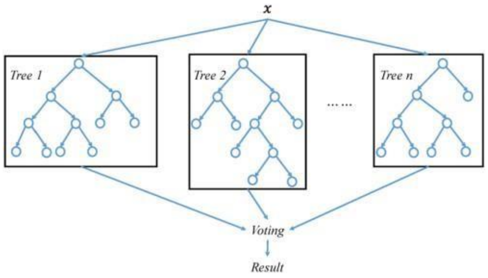
Gradient Boosting:

Gradient Boosting is an ensemble machine learning technique that sequentially builds a predictive model by combining the strengths of multiple weak learners,typically Decision Tree [4]s. It does so by focusing on the errors made by the previous models and adjusting its predictions to reduce those errors, ultimately creating a powerful and accurate predictivemodel that excels in various tasks, including regression and classification.



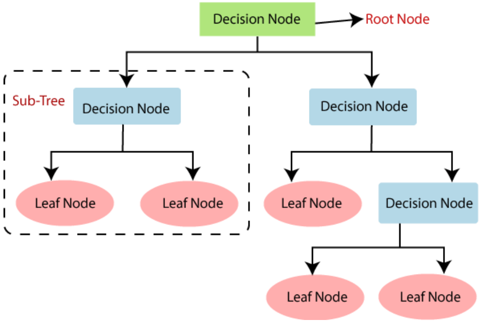
Random Forest:

Random Forest [4] is an ensemble learning method that combines multiple Decision Tree [4]s to make predictions. It works by training a collection of Decision Tree [4]s on random subsets of the data and then averaging their predictions. This ensemble approach enhances accuracy, reduces overfitting, and provides robust performance for both classification and regression tasks.



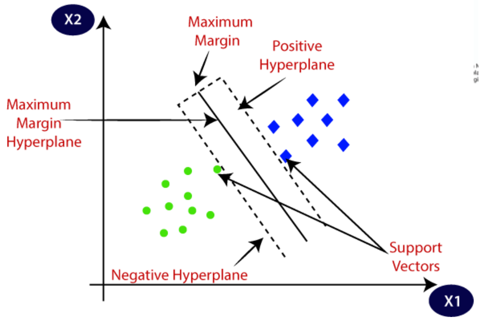
Decision Tree:

A Decision Tree [4] is a machine learning model that makes decisions by recursively splitting data into subsets based on the most significant feature, aiming to classify or predict outcomes. It creates a tree-like structure where each node represents a feature and each branch represents a possible decision, making it interpretable and effective for various tasks.



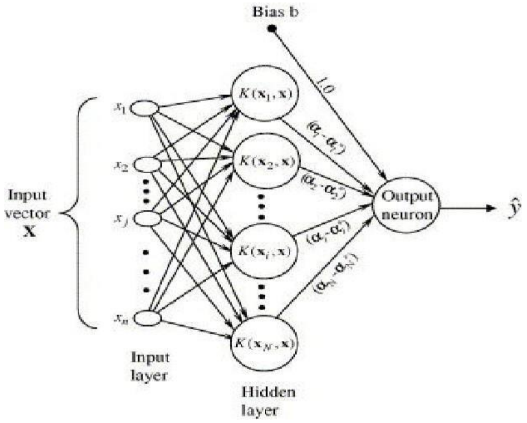
Support Vector Classifier:

A Support Vector Classifier (SVC) is a machine learning model that finds the best possible boundary (hyperplane) to separate different classes of data while maximizing the margin between them. It identifies key support vectors to make accurate classifications, making it effective for both binaryand multi-class classification tasks.



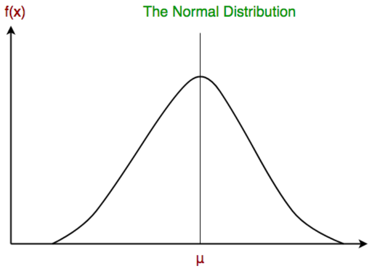
Logistic Regression:

Logistic Regression is a classification algorithm that predicts the probability of an input belonging to a specific category. It employs the sigmoid function to map the input features to a probability score between 0 and 1, and a threshold is applied to classify the input into one of two or more categories based on thisprobability. The model learns coefficients during training to best fit the data and make accurate classifications.


Bias b
1.0
$K(\mathbf{x}_1, \mathbf{x})$
$K(\mathbf{x}_2, \mathbf{x})$
$K(\mathbf{x}_i, \mathbf{x})$
$K(\mathbf{x}_N, \mathbf{x})$
$(\alpha_1 - \alpha_1^*)$
$(\alpha_2 - \alpha_2^*)$
$(\alpha_i - \alpha_i^*)$
$(\alpha_N - \alpha_N^*)$
Output neuron
$\hat{y}$
Input vector
X
x_1
x_2
x_j
x_n
Input layer
Hidden layer

Naive Bayes:

Naive Bayes is a probabilistic classification algorithm that works by applying Bayes' theorem with the "naive" assumption of feature independence. It calculates the probability of a data point belonging to a particular class based on the probabilities of itsindividual features. Naive Bayes is particularly efficient for text classification tasks, spam detection, and other situations where feature independence is a reasonable approximation.



IV. EXPERIMENTAL RESULTS

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

Accuracy: A test's accuracy is defined as its ability to recognize debilitated and solid examples precisely. To quantify a test's exactness, we should register the negligible part of genuine positive and genuine adverse outcomes in completely examined cases. This might be communicated numerically as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

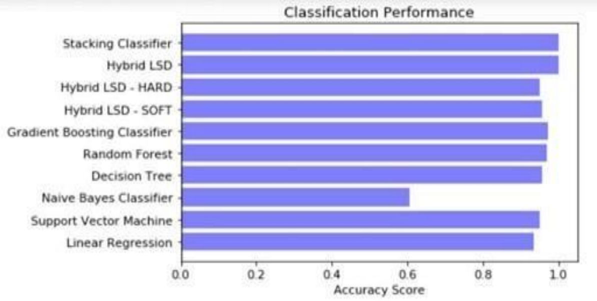


Fig 2: Accuracy Graph

Precision: Precision measures the proportion of properly categorized occurrences or samples among the positives. As a result, the accuracy may be calculated using the following formula:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

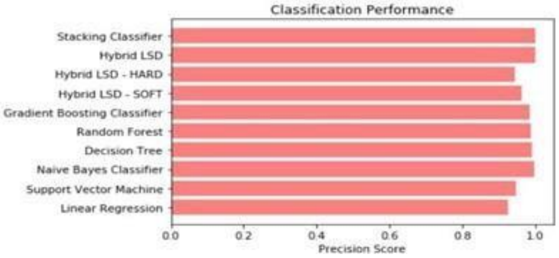


Fig 3: Precision Score Graph

Recall: Recall is a machine learning metric that surveys a model's capacity to recognize all pertinent examples of a particular class. It is the proportion of appropriately anticipated positive perceptions to add up to real up-sides, which gives data about a model's capacity to catch instances of a specific class.

$$Recall = \frac{TP}{TP + FN}$$

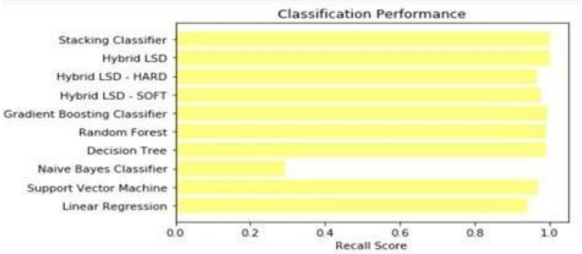
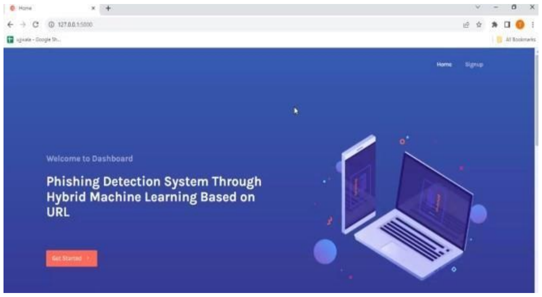


Fig 4: Recall Score Graph

F1-Score: The F1 score is a machine learning evaluation measurement that evaluates the precision of a model. It consolidates a model's precision and review scores. The precision measurement computes how often a model anticipated accurately over the fulldataset.

$$\textbf{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\textbf{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



Welcome to Dashboard

Phishing Detection System Through Hybrid Machine Learning Based on URL

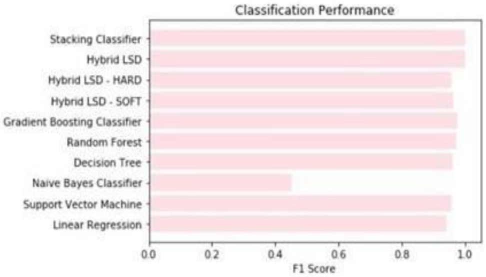
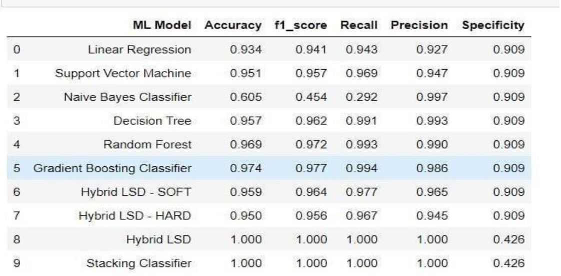


Fig 5: F1 Score Graph

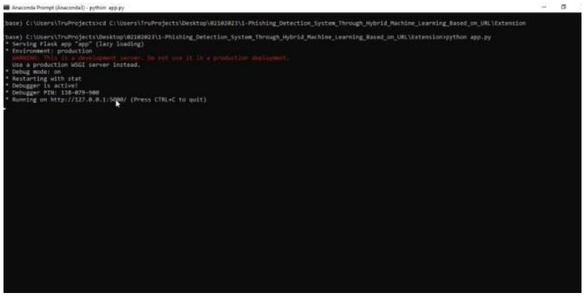
B) Performance Evaluation table.



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

Fig 6: Performance Evaluation Table

C) Frontend



```
(base) C:\Users\TruProjects>cd C:\Users\TruProjects\Desktop\02102023\1-Phishing_Detection_System_Through_Hybrid_Machine_Learning_Based_on_URL\Extension

(base) C:\Users\TruProjects\Desktop\02102023\1-Phishing_Detection_System_Through_Hybrid_Machine_Learning_Based_on_URL\Extension>python app.py
 * Serving Flask app "app" (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: on
 * Restarting with stat
 * Debugger is active!
 * Debugger PIN: 138-079-900
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

Fig 7: Url Link to Web Page

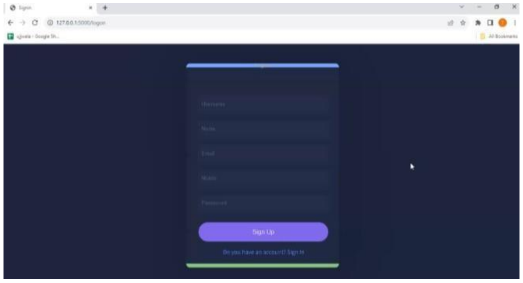


Fig 8: Home page

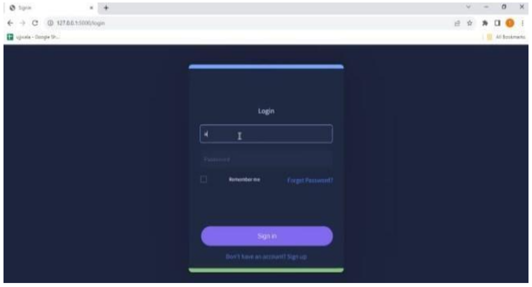


Fig 9: User Signup page

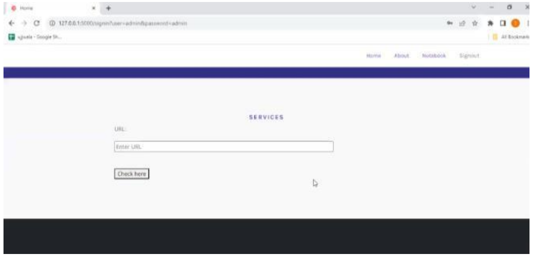
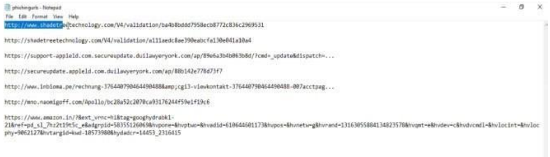
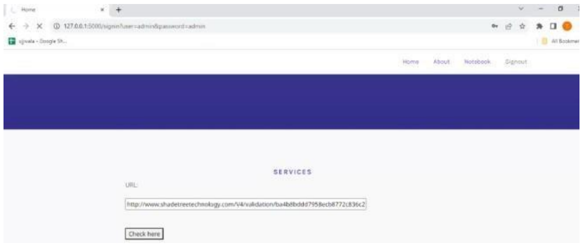


Fig 10: User Sign in Page



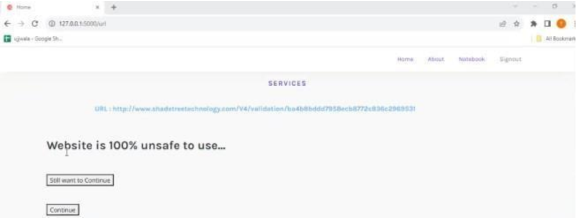
```
http://www.shadetreetechnology.com/V4/validation/ba4b8bddd7958ecb8772c836c2969531

http://shadetreetechnology.com/V4/validation/a111aedc8ae390eabcfa130e041a10a4

https://support-appleld.com.secureupdate.duilawyeryork.com/ap/89e6a3b4b063b8d/?cmd=_update&dispatch=...

http://secureupdate.appleld.com.duilawyeryork.com/ap/88b142e778d73f7

http://www.inbioma.pe/rechnung-376440790464490488&amp;cgi3-viewkontakt-376440790464490488-007acctpag...

http://mno.naomigoff.com/Apollo/bc28a52c2070ca93176244f59e1f19c6

https://www.amazon.in/?&ext_vrnc=hi&tag=googhydrabk1-
21&ref=pd_sl_7hz2t19t5c_e&adgrpid=58355126069&hvpone=&hvptwo=&hvadid=610644601173&hvpos=&hvnetw=g&hvrand=13163055884134823578&hvqmt=e&hvdev=c&hvdvcmdl=&hvlocint=&hvloc
phy=9062127&hvtargid=kwd-10573980&hydadcr=14453_2316415
```

Fig 11: Enter URL



http://www.shadetreetechnology.com/V4/validation/ba4b8bddd7958ecb8772c836c2

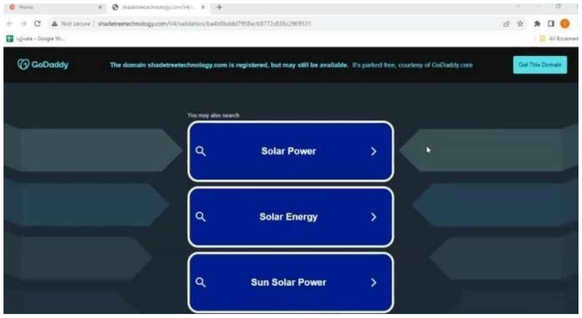
Fig 12: Sample data for testing



URL : http://www.shadetreetechnology.com/V4/validation/ba4b8bddd7958ecb8772c836c2969531

Website is 100% unsafe to use...

Fig 13: Entered Url


Not secure | shadetreetechnology.com/V4/validation/ba4b8bddd7958ecb8772c836c2969531
GoDaddy
Get This Domain
You may also search
Solar Power
Solar Energy
Sun Solar Power

The domain shadetreetechnology.com is registered, but may still be available. It's parked free, courtesy of GoDaddy.com

Fig 14: Url result unsafe 100%

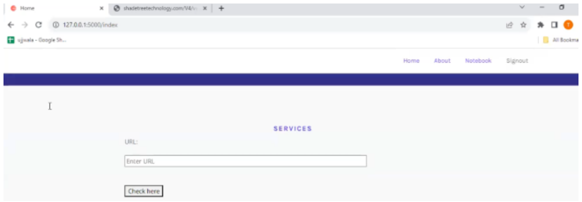
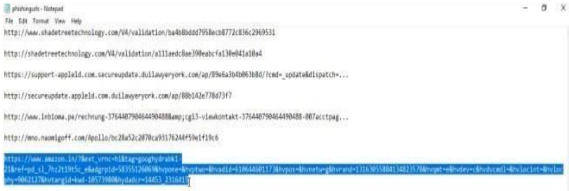
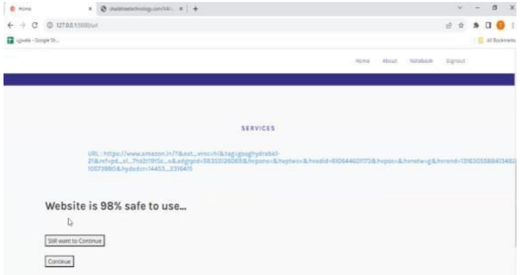


Fig 15: Search Other Urls too



```
http://www.shadetreetechnology.com/V4/validation/ba4b8bddd7958ecb8772c836c2969531

http://shadetreetechnology.com/V4/validation/a111aedc8ae390eabcfa130e041a10a4

https://support-appleld.com.secureupdate.duilawyeryork.com/ap/89e6a3b4b063b8d/?cmd=_update&dispatch=...

http://secureupdate.appleid.com.duilawyeryork.com/ap/88b142e778d73f7

http://www.inbioma.pe/rechnung-3764407904644904888amp;cgi3-viewkontakt-3764407904644904888-007acctpag...

http://mno.naomigoff.com/Apollo/bc28a52c2070ca93176244f59e1f19c6

https://www.amazon.in/?&ext_vrnc=hi&tag=googhydrabk1-
21&ref=pd_sl_7hz2t19t5c_e&adgrpid=58355126069&hvpone=&hvptwo=&hvadid=610644601173&hvpos=&hvnetw=g&hvrand=13163055884134823578&hvqmt=e&hvdev=c&hvdvcmdl=&hvlocint=&hvloc
phy=9062127&hvtargid=kwd-10573980&hydadcr=14453_2316415
```

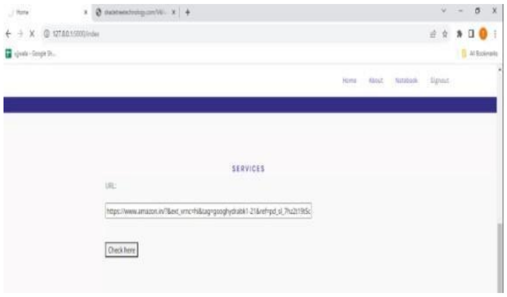
Fig 15: Enter New Url



URL : https://www.amazon.in/?&ext_vrnc=hi&tag=googhydrabk1-21&ref=pd_sl_7hz2t19t5c_e&adgrpid=58355126069&hvpone=&hvptwo=&hvadid=610644601173&hvpos=&hvnetw=g&hvrand=1316305588413482 10573980&hydadcr=14453_2316415

Website is 98% safe to use...

Fig 16: Sample data for testing



SERVICES

URL:

https://www.amazon.in/?&ext_vrnc=hi&tag=googhydrabk1-21&ref=pd_sl_7hz2t19t5c

Check here

Fig 17: Entered New Url

V. CONCLUSION

In conclusion, the project successfully employed a hybrid machine learning approach, prioritizing URL attributes for phishing detection. By leveraging diverse models such as Decision Tree [4]s, Random Forest [4]s, support vector classifiers, and an LSD- based stacking classifier, the system achieved remarkable improvements in accuracy and efficiency.The selection of an extension stacking classifier stoodout due to its exceptional accuracy and F-score, marking a significant enhancement in the overall effectiveness of the phishing detection system. This comprehensive approach addresses a critical cybersecurity challenge by providing robust protection against severe phishing attacks. The integration of multiple machine learning models not only diversified the system's capabilities but also ensured a higher level of adaptability to evolving phishing techniques. The project's success in enhancing accuracy and efficiency underscores its potential impact on bolstering cybersecurity measures, offering a valuable contribution to the ongoing efforts to combat cyber threats. As phishing attacks continue to evolve in sophistication, the developed system stands as a robust defense mechanism, showcasing its potential for real-world applications in safeguarding sensitive information and mitigating the risks associated with cyber threats.

VI. FUTURE SCOPE

The future scope of this project involves continuous refinement and adaptation to emerging phishing tactics. Further research could explore the integration of deep learning techniques, behavioral analysis, and real-time threat intelligence to enhance the system's proactive defense capabilities. Additionally, collaboration with cybersecurity experts and industry stakeholders can contribute to the development of a more comprehensive and resilient solution. Exploringdeployment in cloud environments and IoT devices, along with user-friendly interfaces, will extend the reach of this system. Ongoing updates to the model based on evolving threat landscapes will ensure its sustained effectiveness, making it a cutting-edge solution in the dynamic field of cybersecurity.

REFERENCES

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

[2] H. Shirazia, K. Haynesb, and 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 and S. Nimmagadda, ‘‘Network intrusion detection for TCP/IP packets with machine learning techniques,’’ in Machine Intelligence and 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 and 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, and M. Supramaniam, ‘‘Phishing website detection: An improved accuracy through feature selection and ensemble learning,’’ Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 1, pp. 252–257, 2019.

[8] A. Aggarwal, A. Rajadesingan, and 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, and E. Snekkenes,Computer Security— ESORICS 2017, vol. 10492.Oslo, Norway: Springer, Sep. 2017.

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

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

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

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

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

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

[16] T. Nathezhtha, D. Sangeetha, and 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 and 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 and P. Komisarczuk, ‘‘An analysis of phishing blacklists: Google safe browsing, OpenPhish, and PhishTank,’’ in Proc. Australas. Comput. Sci. Week Multiconf. (ACSW), Melbourne, VIC, Australia. New York, NY, USA: Associationfor Computing Machinery, 2020, pp. 1–11, Art. no. 3,doi: 10.1145/3373017.3373020.

[20] A. K. Jain and 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, and 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 and J. A. Kumar, ‘‘Mobile phishing attacks and defence mechanisms: State of art and open research challenges,’’ Comput. Secur., vol. 73, pp. 519–544, Mar. 2018.

[23] M. Khonji, Y. Iraqi, and 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, and J. Downs, ‘‘Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions,’’ in Proc. SIGCHI Conf. Hum. Factors Comput. Syst., Apr. 2010, pp. 373–382.

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

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

[27] M. Sharifi and 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, and 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 and Public Policy, Jul. 2009.

[29] Y. Zhang, J. I. Hong, and L. F. Cranor, ‘‘Cantina: A content-based approach to 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, and 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