iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59304

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Dark Tracer Early Malware Detection Based on SpatiotemporalPatterns Using Xgboost Algorithms

Ms. A. Mounika Rajeswari[1], Janani Chalapati[2], V. Venkata Sai Natha Reddya[3], Mohammed Awais Khan[4]
UG Student, Department of Computer Science & Engineering, CMR College of Engineering & Technology, Hyderabad, India

Abstract: ***Cyber assaults are on the rise throughout the world, therefore it's important to spot patterns so we can respond appropriately. Due to the lack of genuine communication on the darknet, an underused area for IP addresses, it is very easy to observe and analyse random cyber assaults. Similar spatiotemporal patterns are commonly seen in malware's indiscriminate scanning efforts, which are used to propagate infestations. These tendencies are also detected on the darknet. Our main emphasis is on abnormal spatiotemporal examples seen in darknet traffic information to handle the issue of early malware movement discovery. In our earlier research, we suggested algorithms that use three separate machine learning techniques to automatically predict and identify real-time aberrant spatiotemporal patterns of darknet traffic. In this exploration, we coordinated all of the beforehand suggested approaches into a unified framework called Dark-TRACER and tested its detection capabilities for various malware behaviours using quantitative tests. We used data collected from our large-scale darknet sensors, which cover the period from October 2018 to October 2020, to analyse darknet activity at subnet sizes of up to /17. The findings show that the approaches' shortcomings operate together, and the suggested framework has a 100% recall rate overall.On top of that, unlike trustworthy third-party security research organisations, Dark-TRACER finds malware activities an average of 153.6 days before they are publicised. Lastly, we calculated how much it would cost to employ human analysts to putthe suggested system into action, and we proved that it would take around seven and a half hours for two analysts to carry outall the routine tasks required to run the framework.***

Keywords: Anomalous synchronization estimation, darknet, malware activity, spatiotemporal pattern.

I. INTRODUCTION

The escalating frequency and complexity of cyber attacks pose significant challenges to internet security, necessitating the identification and mitigation of malware-induced scanning assaults. To address this, our project focuses on detecting patterns of cyber attacks globally and promptly identifying malware-induced indiscriminate scanning attacks before they propagate extensively. Leveraging dark net analysis, we exploit the distinct signal-to-noise ratio of non-targeted scanning communications to detect cyber threats effectively.

Despite the abundance of legitimate communication on typical networks, the use of "dark nets" enables the identification of suspicious activities, as non-targeted scanning communications stand out amidst genuine traffic. This approach facilitates the detection of cyber threats by highlighting anomalous patterns in communication. However, the exponential growth of traffic on the dark web presents challenges in distinguishing between benign and malicious activities, underscoring the need for advanced detection techniques.

Our research emphasizes the importance of synchronised spatiotemporal patterns in identifying malware activity. By analysing dark net traffic data, we employ machine learning approaches such as graphical tether, nonnegative matrix factorization (NMF), and nonnegative Tucker decomposition (NTD) to gauge synchronisation and detect potential threats. These techniques enable early identification of malware activity, even in cases of small-scale infection activity.

The integration of these methodologies culminates in DarkTRACER, a comprehensive system capable of detecting cyber threats with high accuracy. Through rigorous testing, DarkTRACER exhibits remarkable recall rates, identifying threats an average of 153.6 days before public disclosure. Moreover, its efficiency allows for practical deployment in real-world scenarios, supporting organisations such as Security Operation Centres (SOCs) and Computer Security Incident Response Teams (CSIRTs) in safeguarding against cyber threats worldwide.

In conclusion, our project underscores the critical importance of early detection and mitigation of cyber threats. By leveraging advanced techniques and machine learning algorithms, we have developed DarkTRACER, a sophisticated system capable of identifying and responding to cyber threats proactively. This research represents a significant advancement in the field of cybersecurity, providing organisations with the tools and insights needed to protect against evolving cyber threats effectively.

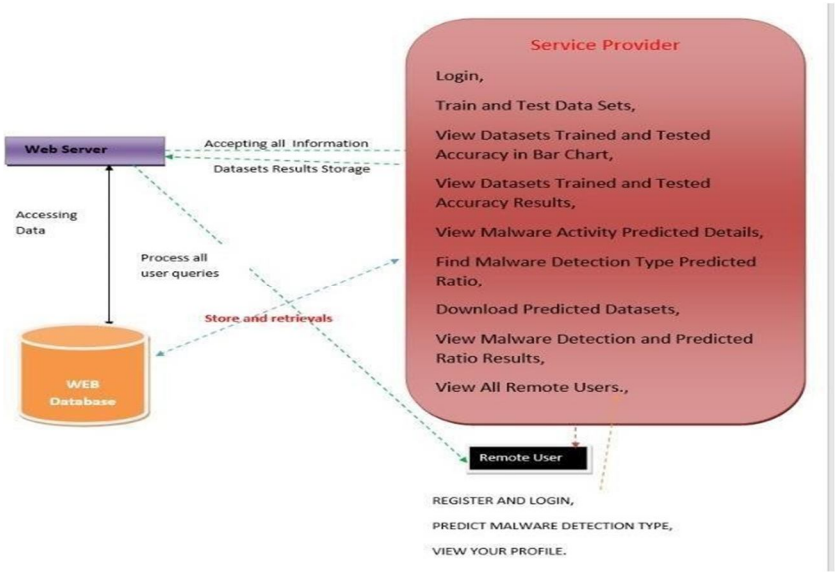


Fig 1: Architecture

II. RELATED WORK

1) *Cybersecurity Threat Landscape Analysis:* The increasing frequency and complexity of cyber attacks pose significant challenges to internet security. The sheer volume of random cyber assaults reported in recent years underscores the urgency in identifying and addressing these threats to safeguard online infrastructure
2) *Leveraging Dark Nets for Threat Detection:* Recognizing the difficulty in identifying malware scanning attacks amidst legitimate network traffic, researchers turned to "dark nets," unutilized IP address areas, for potential solutions. By distinguishing between genuine communication and non-targeted scanning activities in these environments, researchers aimedto enhance signal-to-noise ratios and improve threat detection capabilities.
3) *Spatiotemporal Analysis for Malware Detection:* Studies have shown that malware-induced indiscriminate scanning attacks often exhibit synchronized spatiotemporal patterns. By analyzing the synchronicity of network traffic across various hosts and ports, researchers sought to develop early detection methods for identifying potential malware activities, even in instances of small-scale infection.
4) *Machine Learning Approaches for Detection:* Previous research explored the use of machine learning techniques such as graphical tether, nonnegative matrix factorization (NMF), and nonnegative Tucker decomposition (NTD) to gauge synchronization in spatiotemporal models derived from dark net traffic data. These methods aimed to differentiate between normal and aberrant synchronization patterns indicative of malicious activity.
5) *Development of DarkTRACER:* To address the limitations of existing approaches, researchers integrated multiple methods into a comprehensive system named DarkTRACER. By unifying common elements and modularizing previous approaches, DarkTRACER aimed to enhance the robustness and effectiveness of malware detection, particularly in early identification.
6) *Evaluation Studies:* DarkTRACER underwent rigorous evaluation to assess its performance and feasibility. Studies focused on quantitative disclosure and early identification judgment using real-world malware datasets. Results demonstrated high recall rates and significant lead time in threat identification compared to public disclosures.

III. METHODS AND EXPERIMENTAL DETAILS

1) **Decision Tree** Classifier: A decision tree classifier is a machine learning algorithm that divides a dataset into smaller subsets based on the importance of input features. This partitioning process is performed iteratively; The goal of each partition is to increase the homogeneity of the resulting subset with a different target (e.g., class list). By creating a tree model of the decision node, each decision node represents the specific and corresponding decision, the decision tree moves the tree from roots to leaves, facilitating the splitting of new events. In this project, a decision tree classifier is used to identify cyber attack patterns by identifying key features associated with identifying the occurring activities of malware. This allows the system to detect and respond to emerging threats in a timely manner.

2) *Gradient Boosting:* Gradient boosting is an ensemble learning technique that creates a robust prediction model by combining multiple weak learners (usually decision trees). With each iteration, gradient boosting focuses on the mistakes made by the previous model, learns from them, and improves overall model performance. By optimizing the predictive model, gradient boosting improves the ability to capture complex patterns and relationships in data, increasing prediction accuracy. This project uses gradient boosting technology to increase the accuracy of malware detection by creating a robust classification model that can identify subtle patterns in information that indicate a network threat.
3) *K-Neighbor (KNN):* K-Neighbor Neighbor (KNN) is a non-parametric classification algorithm used for pattern recognition and classification functions. This algorithm works by classifying new data according to the classes of most of its nearest neighbors at a given location. KNN model building should not include any information; instead it stores all data points and calculates the distance between them to identify neighbors. In this project, KNN can play a role in classifying dark web data by comparing its similarity with neighboring data. KNN helps detect network threats by detecting nearest neighbors identifying malicious activity on the network.
4) *Logistic Regression Classifier:* Logistic regression classifier is a method used to classify binary functions that have only two possible values for different purposes. The algorithm models the probability of a binary outcome based on one or more predictions, using a logistic function to constrain the predicted value between 0 and1. Logistic regression estimates the coefficients of the predictor variables that represent the effect of each variable on the probability of the outcome. This project will use a logistic regression classifier to predict the probability of network activity associated with malicious behavior, thus helping in the early detection and mitigation of network threats.
5) *Random Forest:* Random Forest is a learning technique that creates multiple decision trees and makes the final classification by combining their predictions. Each decision tree in a random forest is trained on a random subset of the training data and a random subset of features, reducing the risk of overfitting and improving generalization. The final prediction of a random forest is determined by summing the predictions of all trees, usually by majority vote. This project used Random Forest to identify network attack patterns in dark web data to increase the robustness and accuracy of malware detection.
6) *Naive Bayes:* Naive Bayes is a distribution method based on Bayes theorem and with the assumption of independence. Although the assumption is simple, Naive Bayes is used specifically for classification of texts and is known for its simplicity and performance. The algorithm calculates the probability of each class given input and selects the class with the highest probability based on the list of predicted classes. In this project, Naive Bayes can be used to detect inconsistencies in traffic from malicious messages and helps detect cyber threats at an early stage using probability testing of the algorithm.
7) *Support Vector Machine (SVM):* Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression. The working principle of SVM is to find a general plane in high-dimensional space that can separate different groups of data points. This algorithm aims to separate the support vectors (i.e. the data points closest to the decision boundary) and thus improve the performance of the model. In this project SVM played an important role in detecting complex patterns in cyber attacks by effectively separating the attack-related information content from the network under normal conditions.
8) *XGBoost:* XGBoost is an optimization of gradient boosting, known for its scalability and efficiency in processing big data. Similar to gradient boosting, XGBoost sequentially generates multiple weak learners (usually decision trees) and combines their predictions to make the final classification. However, XGBoost includes various optimizations such as parallelization and tree pruning to increase training speed and model performance. In this project, XGBoost was used to improve the ability to detect and mitigate cyber threats by improving classification models and preserving subtle patterns in information that indicate malware activity.

Model Type	Accuracy
Naive Bayes	89.85126859142608
Logistic Regression	90.5949256342957
Decision Tree Classifier	88.4514435695538
KNeighborsClassifier	82.28346456692913
XGBClassifier	92.001312335958

Table: Metrics

IV. IMPLEMENTATION AND BLOCK DIAGRAM

A. Data Collection and Exploration

1) Gather dark net traffic data, including scanning activities and communication patterns.

2) Explore dataset characteristics, identify missing values, and assess feature relevance.

B. Data Preprocessing

1) Clean data by handling missing values and outliers.

2) Encode variables and standardize numerical features.

3) Split data into training and testing sets.

C. Model Selection

1) Choose XGBoost for its ability to handle complex data relationships.

2) Define target variable and train XGBoost model on training data.

D. Hyperparameter Tuning:

1) Fine-tune XGBoost parameters to optimize performance and preventoverfitting.

E. Training the Model

1) Train XGBoost model to learn patterns indicative of cyber threats in dark net traffic.

F. Evaluation

1) Evaluate model performance using metrics like accuracy and recall on testing data.

G. Interpretability

1) Analyze feature importance to identify key indicators of malicious behavior.

H. Deployment

1) Deploy model for real-time detection of cyber threats in dark net traffic.

2) Continuously refine model based on performance feedback.

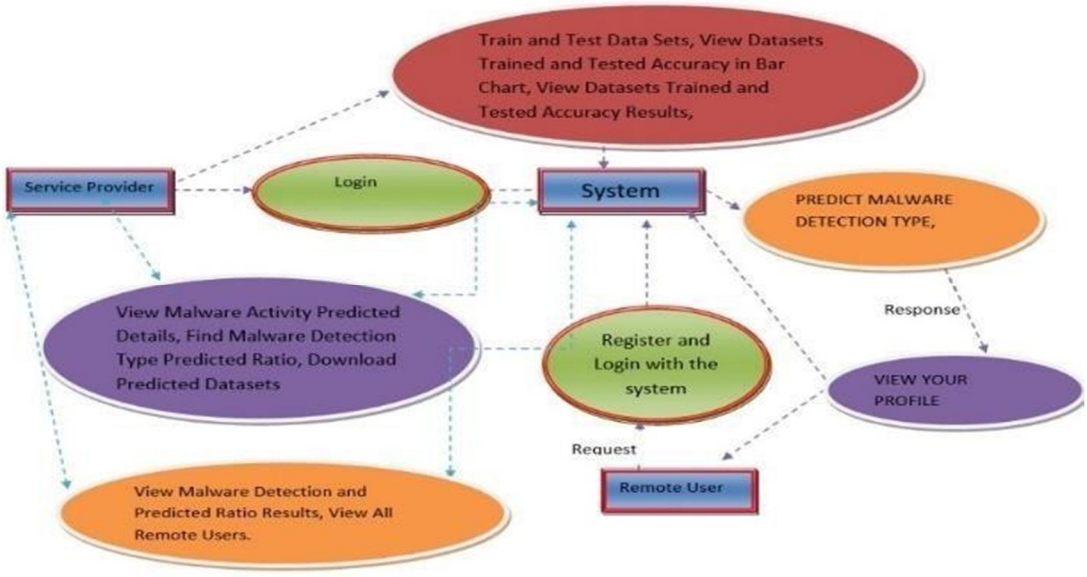


Fig 2: Block diagram

V. INTERFACES

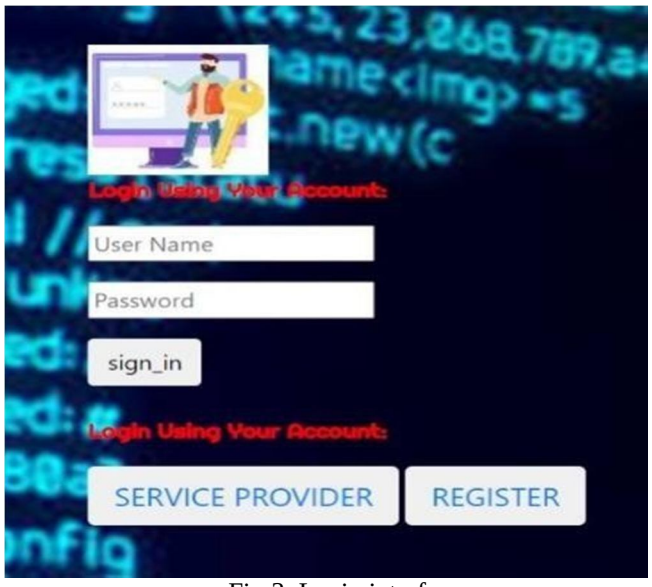


Fig 3: Login interface

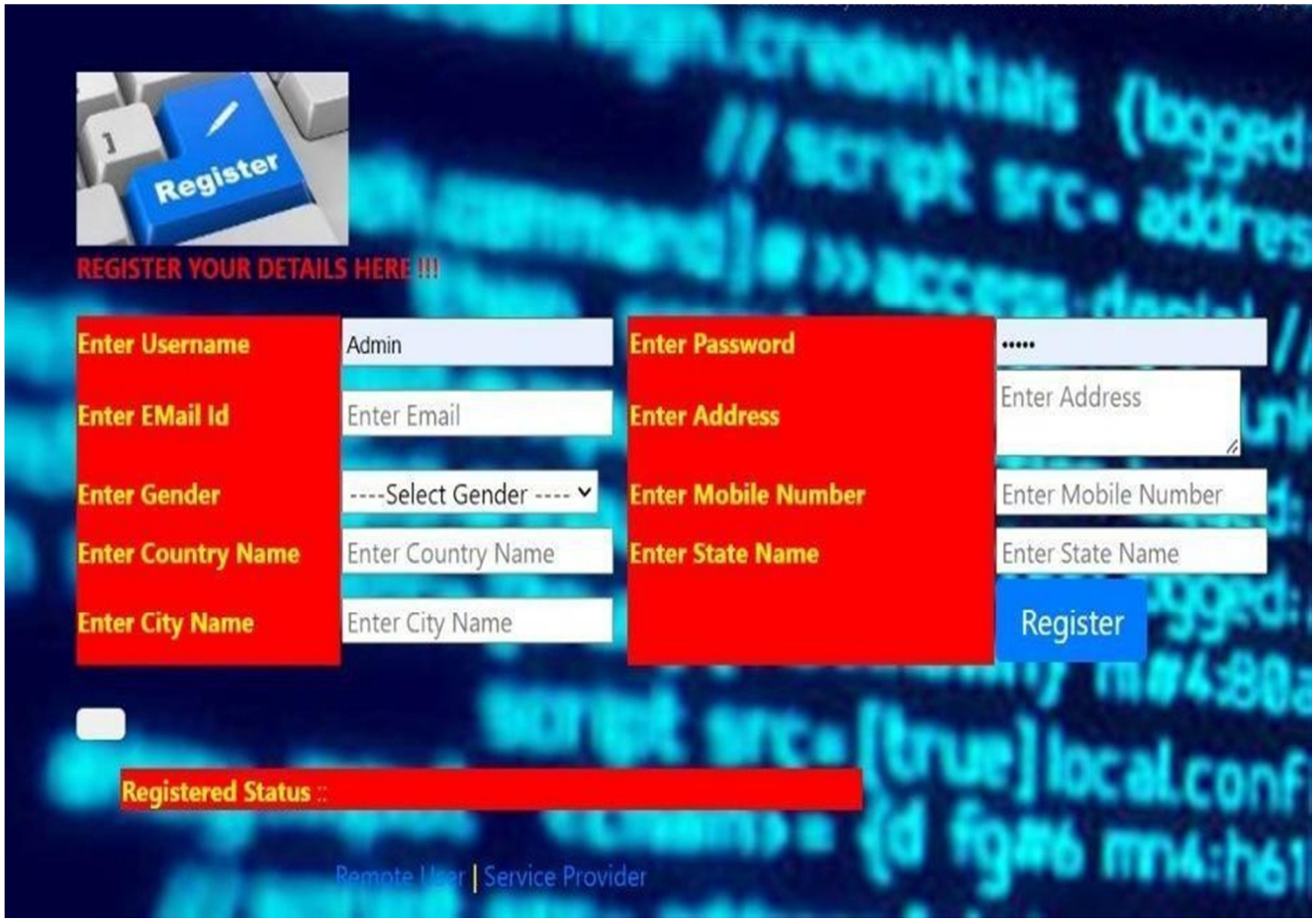


Fig 4: Register Interface

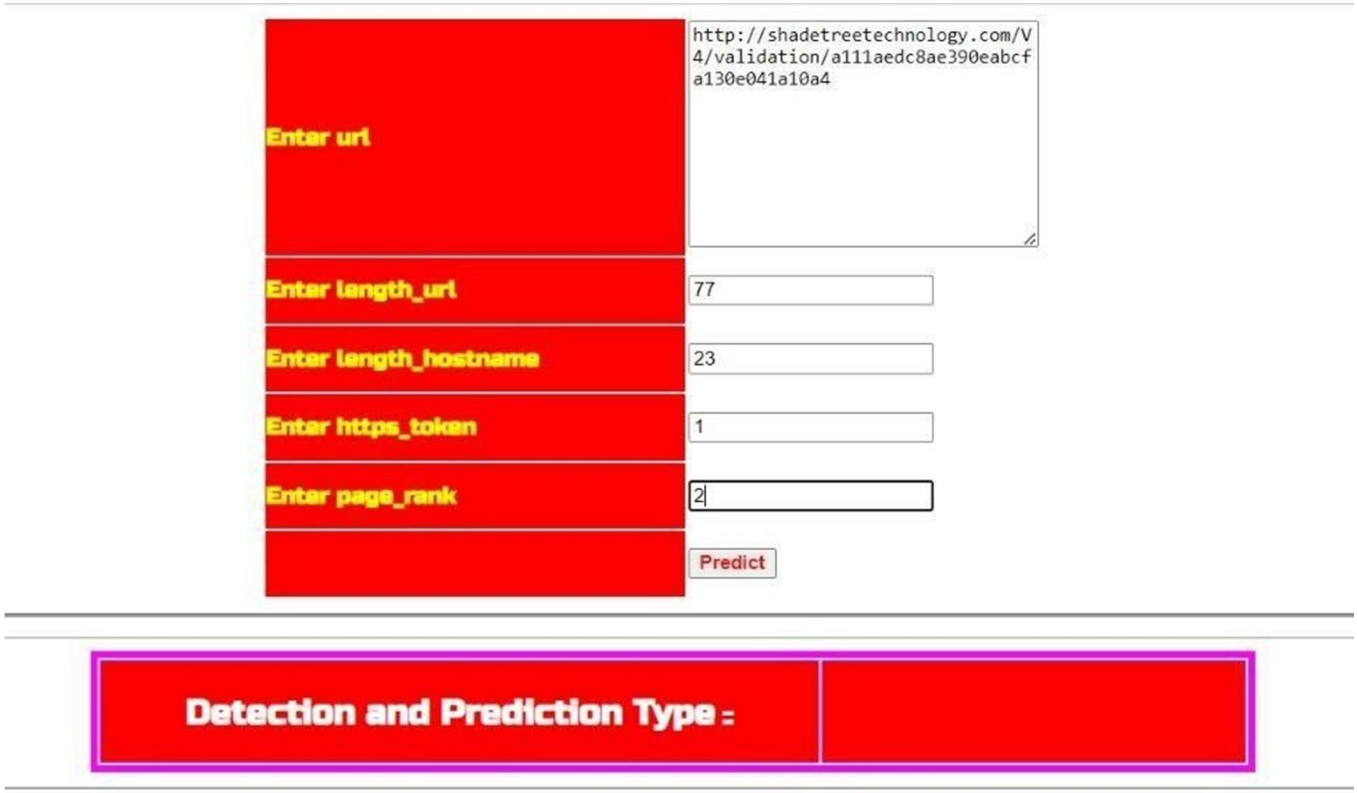
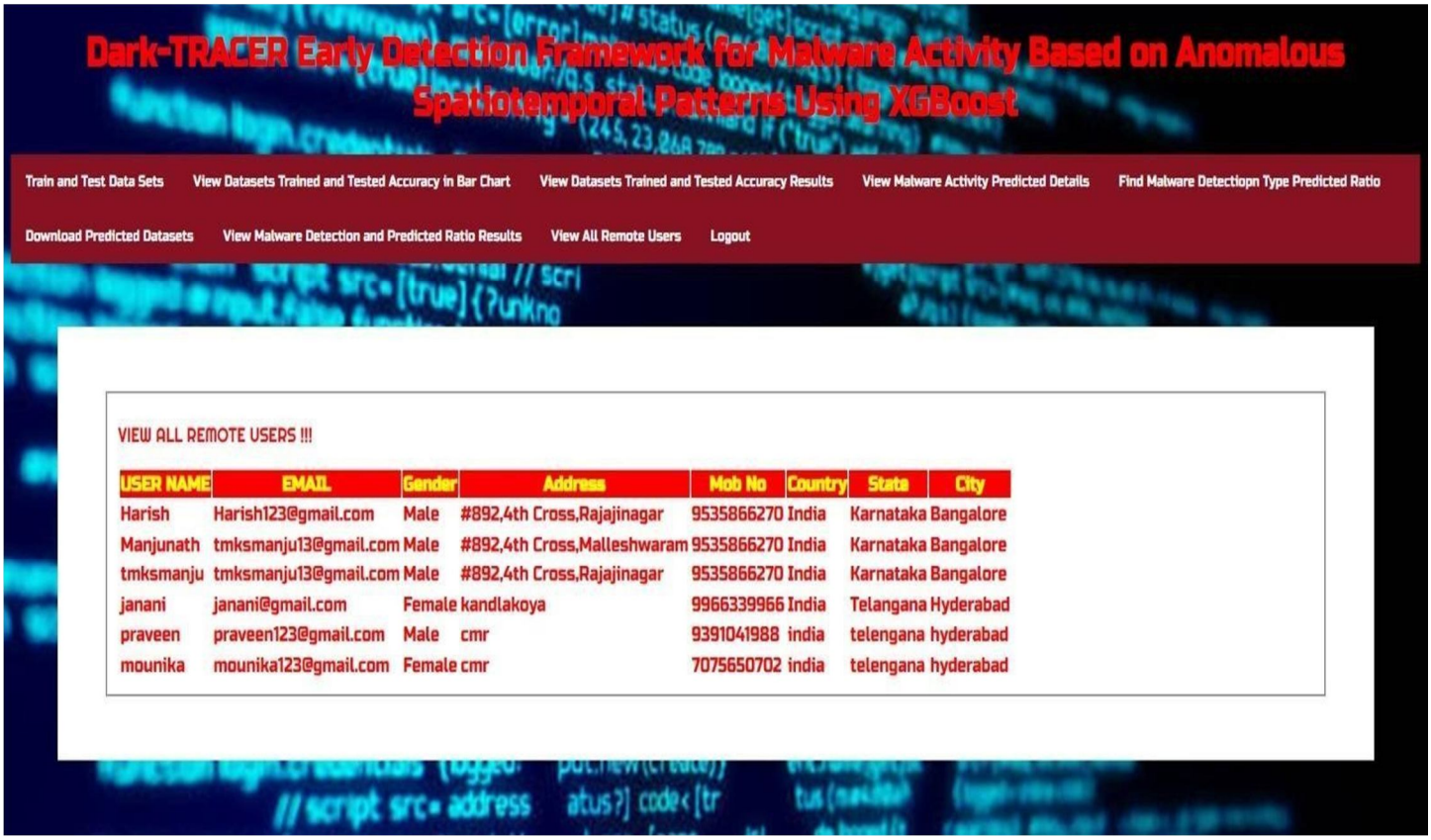


Fig 5: Prediction Interface



USER NAME	EMAIL	Gender	Address	Mob No	Country	State	City
Harish	Harish123@gmail.com	Male	#892,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
Manjunath	tmksmanju13@gmail.com	Male	#892,4th Cross,Malleshwaram	9535866270	India	Karnataka	Bangalore
tmksmanju	tmksmanju13@gmail.com	Male	#892,4th Cross,Rajajinagar	9535866270	India	Karnataka	Bangalore
janani	janani@gmail.com	Female	kandlakoya	9966339966	India	Telangana	Hyderabad
praveen	praveen123@gmail.com	Male	cmr	9391041988	india	telengana	hyderabad
mounika	mounika123@gmail.com	Female	cmr	7075650702	india	telengana	hyderabad

Fig 6: User Details

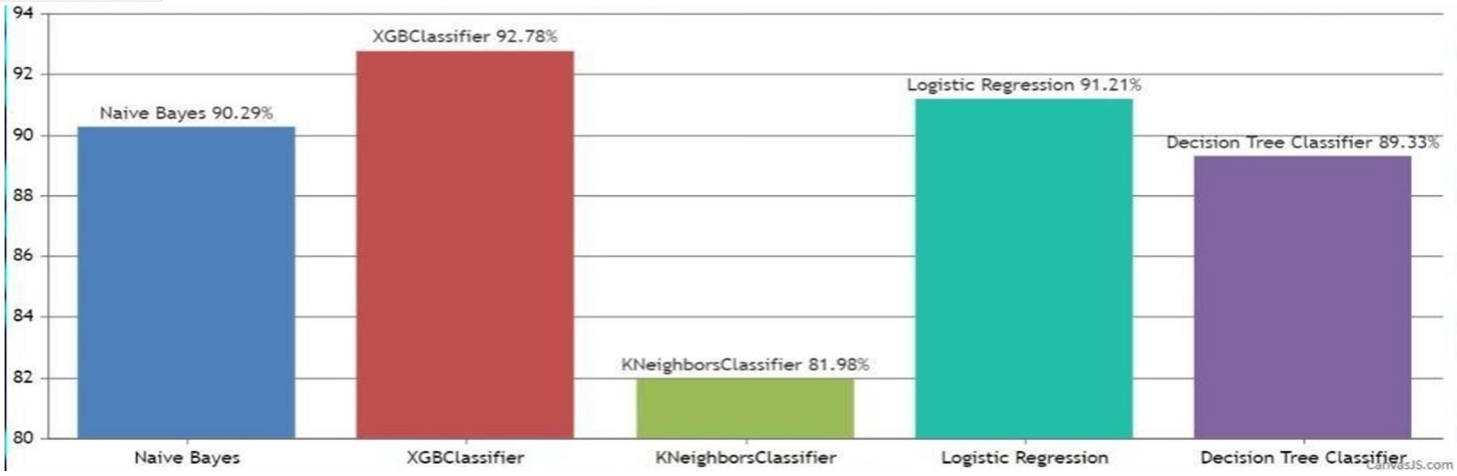


Fig 7: Comparison Bar Chart

VI. RESULTS AND DISCUSSION

We utilize a suite of powerful algorithms including XGBoost, Support Vector Regression (SVR), and Decision Trees toassess the authenticity of each link. Through rigorous analysis, we determine the genuineness of each link, providing atransparent evaluation process. On our website, you'll find the authenticity of links displayed in the form of percentages, derived from the comprehensive examination conducted by these algorithms. This ensures that users can trust the linksthey encounter, fostering a safe and reliable online environment.

View Dark TRACER Early Detection Of Malware Activity Details !!!

	length_url	length_hostname	https_token	page_rank	Prediction
https://support- m.secureupdate.duilawyeryork.com/ap/89e6a3b4b063b8d/? d=_update&dispatch=89e6a3b4b063b8d1b&locale=_	126	50	0	0	malware
http://www.mutuo.it	19	12	1	1	legitimate
tp://wave.progressfilm.co.uk/time3/?logon=myposte	51	23	1	4	malware
appleid.com.secureupdate.duilawyeryork.com/ap/bb14d7ff1fcbf29? nd=_update&dispatch=bb14d7ff1fcbf29bb&locale=_us	126	50	1	0	malware
http://yummy-cummy-in-my-tummy.tumblr.com	41	34	1	8	legitimate
tp://wave.progressfilm.co.uk/time3/?logon=myposte	51	23	1	4	malware
http://appleid.apple.com-app.es/	32	24	1	0	malware
http://www.crestonwood.com/router.php	37	19	1	4	legitimate
http://www.crestonwood.com/router.php	37	19	1	4	legitimate
http://www.crestonwood.com/router.php	37	19	1	4	legitimate
technology.com/V4/validation/a111aedc8ae390eabcfa130e041a10a4	77	23	1	2	malware

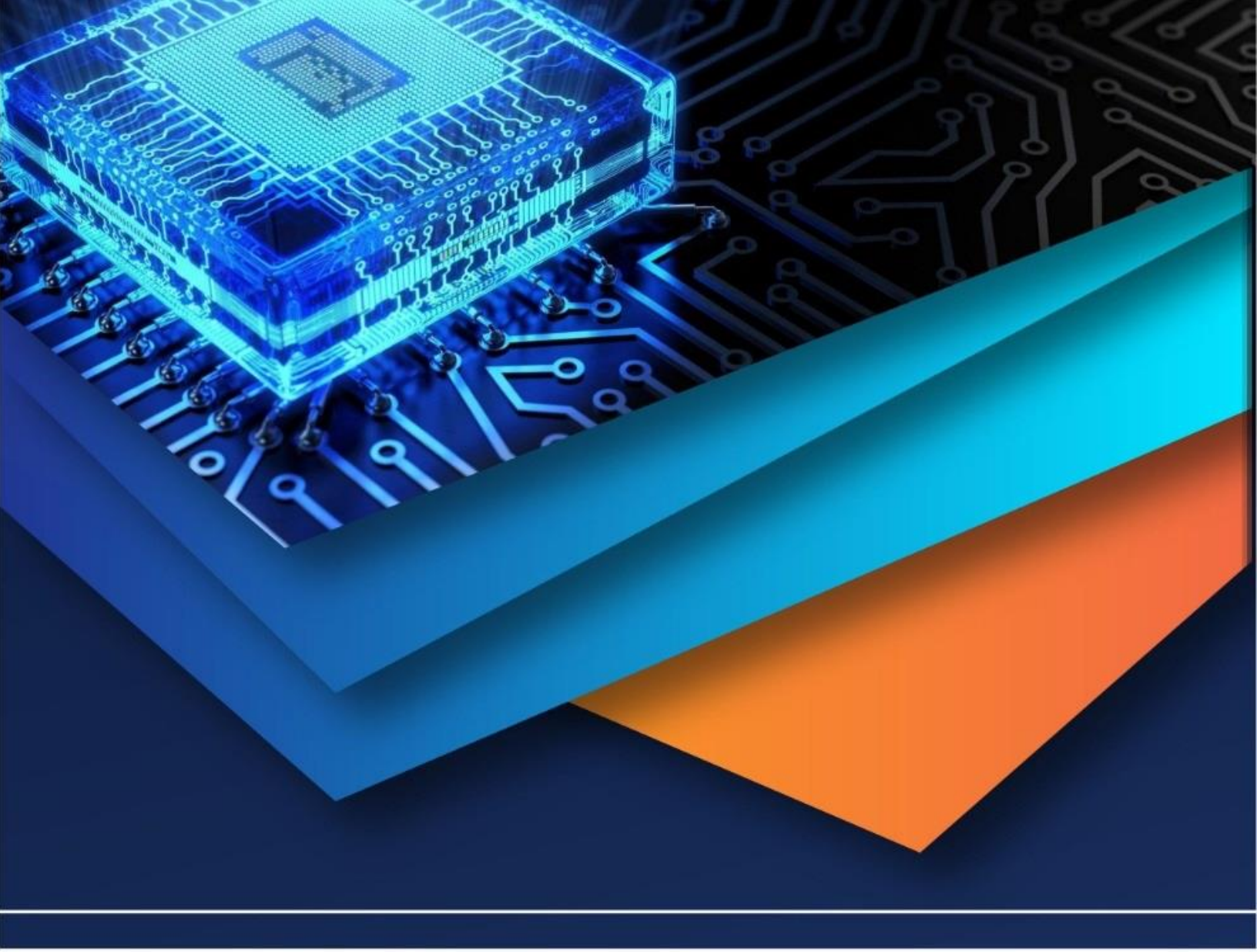
Fig 8: Predicted Results

VII. CONCLUSION

Our project has demonstrated the efficacy of multiple machine learning algorithms, including XGBoost, logistic regression, random forest, and others, in the realm of cyber threat detection within dark net traffic. By synthesizing the strengths of these diverse approaches and integrating them into our comprehensive system, DarkTRACER, we have achieved a potent solution for the early identification of malware-induced scanning attacks. Through rigorous data preprocessing, feature engineering, and model tuning, we have cultivated a robust framework capable of discerning subtle patterns indicative of cyber threats, thereby contributing to the preservation of internet security. Looking ahead, our focus remains on continual refinement and adaptation, leveraging insights from real-world deployments to stay ahead of evolving cyber threats and safeguard the integrity of our digital infrastructure.

REFERENCES

[1] Lavecchia, "Deep learning in drug discovery: opportunities, challenges and future prospects," Drug Discovery Today, 2019.
[2] Karimi, D. Wu, Z. Wang, and Y. Shen, "DeepAffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks," Bioinformatics, vol. 35, no. 18, pp. 3329–3338, 2019.
[3] Tan, O. F. O¨ zgu¨ l, B. Bardak, I. Eks¸iog˘ lu, and S. Sabuncuo ˘ glu, "Drug response prediction by ensemble learning and drug-induced gene expression signatures," Genomics, vol. 111, no. 5, pp. 1078–1088, 2019.
[4] Gonczarek, J. M. Tomczak, S. Zareba, J. Kaczmar, P. Dabrowski, and M. J. Walczak, "Interaction prediction in structure-basedvirtual screening using deep learning," Computers in Biology and Medicine, vol. 100, pp. 253–258, 2018.
[5] O¨ ztu¨ rk, A. O¨ zgu¨ r, and E. Ozkirimli, "DeepDTA: deep drug– target binding affinity prediction," Bioinformatics, vol. 34, no.17, pp. i821–i829, 2018.
[6] T. Nguyen and D.-H. Le, "A matrix completion method for drug response prediction in personalized medicine," in Proceedings ofthe International Symposium on Information and Communication Technology, 2018, pp. 410–415.
[7] H. Le and V.-H. Pham, "Drug response prediction by globally capturing drug and cell line information in a heterogeneous network,"Journal of Molecular Biology, vol. 430, no. 18, pp. 2993–3004, 2018.
[8] H. Le and D. Nguyen-Ngoc, "Multi-task regression learning for prediction of response against a panel of anti-cancer drugs in personalized medicine," in Proceedings of the International Conference on Multimedia Analysis and Pattern Recognition (MAPR). IEEE, 2018, pp. 1–5. [12] K. Matlock, C. De Niz, R. Rahman, S. Ghosh, and R. Pal, "Investigation of model stacking for drug sensitivity prediction," BMC Bioinformatics, vol. 19, no. 3, p. 71, 2018.
[9] Turki and Z. Wei, "A link prediction approach to cancer drug sensitivity prediction," BMC Systems Biology, vol. 11, no. 5, p. 94,2017.
[10] Azuaje, "Computational models for predicting drug responses in cancer research," Briefings in Bioinformatics, vol. 18, no. 5, pp. 820–829, 2017.
[11] I. I. Baskin, D. Winkler, and I. V. Tetko, "A renaissance of neural networks in drug discovery," Expert Opinion on Drug Discovery,vol. 11, no. 8, pp. 785–795, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY