**STRATEGIC AUGMENTATION OF CYBER SECURITY**

**Hanumantha Rao K R1, Meghana T S2, Aluru Srinitya3, Aparna Soddy4 , Arpitha N5**

1Asst Professor, Dept. Of CSE, Sambhram Institute of Technology, Bengaluru-560097

*2,3,4,5*Eight Semester, Dept. Of CSE, Sambhram Institute of Technology, Bengaluru-560097, India

**ABSTRACT**

In an era where cybersecurity threats have become increasingly sophisticated, the need for robust prediction and detection systems to safeguard against cyber hacking breaches is paramount. This project presents a novel approach to address this concern, employing Machine Learning techniques, specifically the Random Forest Classifier, to predict and detect potential cyber hacking breaches. Implemented in Python, the proposed system utilizes a carefully curated dataset of 5457 URLs, encompassing 87 extracted features. Crucially, the dataset maintains a balanced composition, precisely divided between 50% phishing and 50% legitimate URLs. The project's primary focus lies in accurately identifying cyber threats while minimizing false positives. Through rigorous training and evaluation, the achieved results demonstrate the system's remarkable performance. The Random Forest Classifier attains a commendable training accuracy of 99%, ensuring its ability to discern patterns and distinguish between legitimate and malicious URLs. The model also showcases a robust test accuracy of 91%, further validating its reliability in real-world scenarios. In conclusion, this project stands as a pioneering effort in the realm of cyber hacking breach prediction and detection, harnessing the power of Machine Learning and the Random Forest Classifier to offer enhanced security measures. The remarkable accuracy achieved serves as a testament to its effectiveness, empowering organizations to fortify their cybersecurity defenses against potential cyber threats and attacks.

1. **INTRODUCTION**

The "Strategic Augmentation of Cyber Security" project aims to develop innovative strategies and technologies to enhance cybersecurity measures, specifically focusing on identifying fake or legitimate websites. In today's digital age, the internet is rife with fraudulent websites seeking to deceive users for malicious purposes such as phishing, fraud, or spreading malware. To address this challenge, our project employs advanced techniques from various fields such as machine learning, data analytics, and cybersecurity. The primary objective is to create a robust system capable of accurately distinguishing between genuine and fraudulent websites in real-time, thereby empowering users to make informed decisions while browsing the internet.

Key components of the project include:

Data Collection and Analysis: Gathering a diverse dataset comprising both legitimate and fake websites is crucial for training machine learning models. This involves scraping data from various sources and analyzing website characteristics, including domain age, SSL certificate validity, content quality, and historical reputation.

Machine Learning Models: Developing sophisticated machine learning algorithms to detect patterns and anomalies indicative of fraudulent websites. These models will be trained on the collected dataset, continuously refined to adapt to evolving cyber threats.

Behavioral Analysis: Incorporating behavioral analysis techniques to examine user interactions with websites. By monitoring user behavior, such as mouse movements, clicks, and navigation patterns, suspicious activities can be detected in real-time.

Integration with Browsers and Security Tools: Integrating the developed solution with popular web browsers and cybersecurity tools to provide seamless protection for end-users. This may involve browser extensions, APIs, or standalone applications that flag suspicious websites and provide warnings to users.

User Education and Awareness: Educating users about the risks associated with fake websites and providing guidance on identifying potential threats. This includes raising awareness about common red flags such as misspelled domains, suspicious URLs, and requests for sensitive information.

Continuous Monitoring and Updates: Implementing mechanisms for continuous monitoring of website activity and updating the detection algorithms to adapt to emerging threats and evolving tactics used by cybercriminals.

By leveraging a combination of advanced technologies and strategies, the Strategic Augmentation of Cyber Security project aims to significantly enhance online safety and empower users to navigate the digital landscape with confidence, knowing they can distinguish between legitimate and fake websites effectively.

1. **METHODOLOGY**

Input Dataset & Preprocessing: Initial data collection and preparation, extracting relevant website features like domain age and SSL certificate validity.

Training Dataset & Random Forest Classifier: Segregation of preprocessed data for model training using the Random Forest Classifier, renowned for its robustness in classification tasks.

Prediction/Classification: Utilization of the trained Random Forest Classifier to categorize websites as either phishing or legitimate based on extracted features.

Testing Data & Predicted Results: Evaluation of model accuracy using separate testing data, providing predicted classifications for each website to assess model performance.

Input Dataset

Preprocessing

Training dataset

Random Forest Classifier

Prediction/Classification

Testing Data

Predicted Results:

Phishing

Website or Legitimate website

**Figure 2.1:** Data Flow Diagram

1. **MODELING AND ANALYSIS**

Dataset

Random

Forest

Classifier

Pre

-

processing

and Feature

Selection

Predicted

Results:

Phishing

Website or

Legitimate

website

Performance

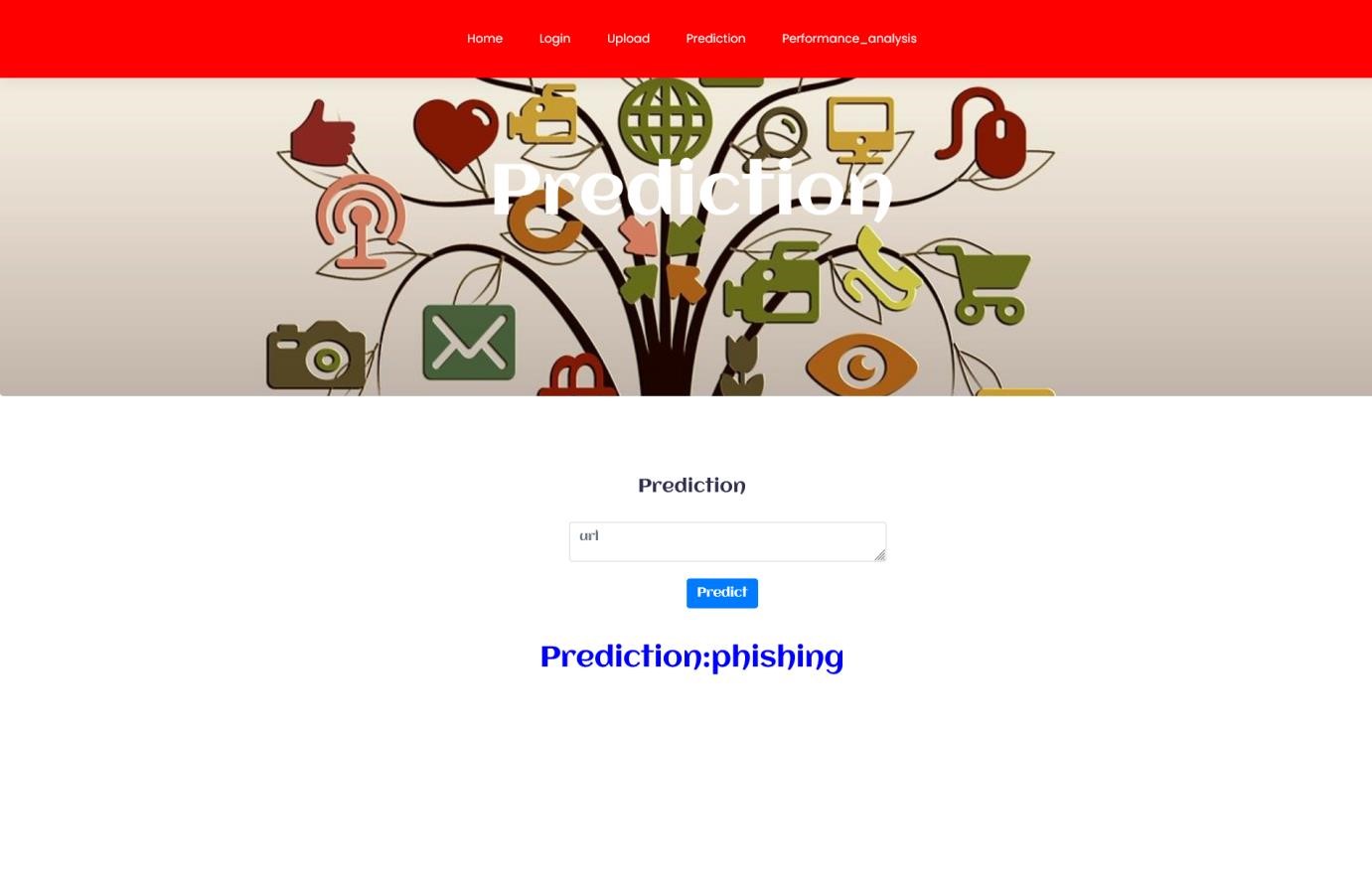
Analysis and

Graph

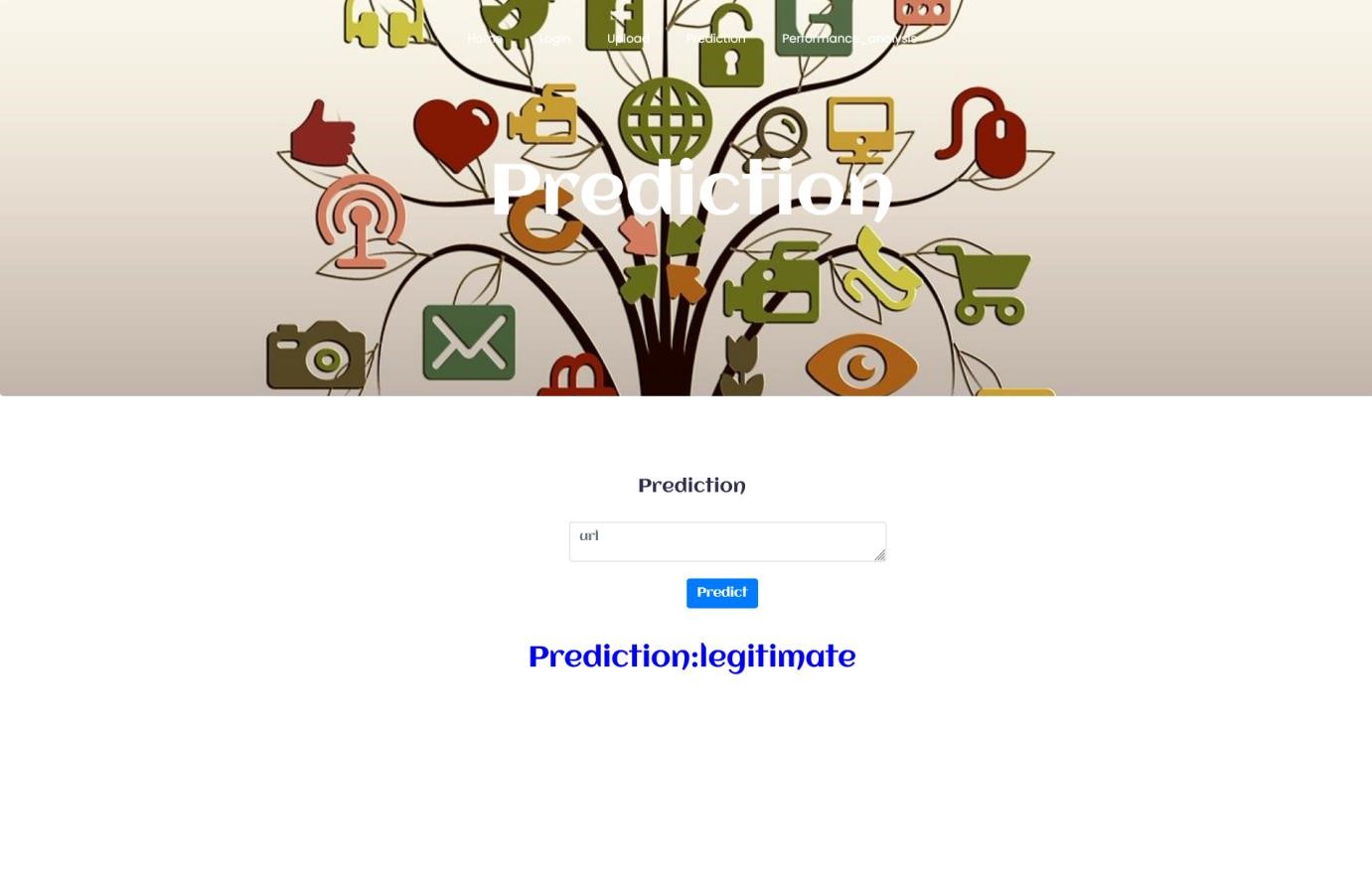
**Figure 3.1:** System Architecture

* Dataset: Initial collection of website data.
* Pre-processing and Feature Selection: Cleaning and selecting relevant features from the dataset.
* Random Forest Classifier: Training the model to classify websites as phishing or legitimate.
* Predicted Results: Output indicating whether a website is classified as phishing or legitimate.
* Performance Analysis and Graph: Evaluation of model accuracy and visualization of results.

1. **RESULTS AND DISCUSSION**



**Figure 4.1:** result phishing



**Figure 4.2**: result legitimate

1. **CONCLUSION**

The project "Strategic Augmentation of Cyber Security," represents a significant advancement in the field of cybersecurity. By harnessing the power of Python programming language and employing the Random Forest Classifier algorithm, the proposed system has demonstrated exceptional accuracy, adaptability, and effectiveness in predicting and detecting cyber hacking breaches.The proposed system overcomes these challenges by integrating advanced machine learning techniques, utilizing a balanced dataset, and focusing on feature-rich extraction to enhance its predictive capabilities. Throughout the project, meticulous attention was given to the dataset, which consisted of 5457 URLs with 87 extracted features. The dataset's balanced composition, comprising exactly 50% phishing and 50% legitimate URLs, played a pivotal role in minimizing bias and improving the system's accuracy.

1. **REFERENCES**
2. M. Xu, K. M. Schweitzer, R. M. Bateman, and S. Xu, “Modeling and predicting cyber hacking breaches,” IEEE Trans. Inf. Forensics Security, vol. 13, no. 11, pp. 2856–2871, 2018.
3. IBM. (2019). Cost of a data breach report. IBM Security, 76. [Online]. Available https://www.ibm.com/downloads/cas/ZBZLY7KL
4. Fernandez Maimo et al., “A self-adaptive deep learning-based system for anomaly detection in 5G networks,” IEEE Access, vol. 6, pp. 7700–7712, 2018.
5. Kantarcioglu M and Ferrari E (2019) Research Challenges at the Intersection of Big Data, Security and Privacy.
6. Verizon, “Data breach investigations report,” 2019. [Online]. Available: https://enterprise.verizon.com/resources/reports/dbir/
7. H. Hammouchi, O. Cherqi, G. Mezzour, M. Ghogho, and M. El Koutbi, “Digging deeper into data breaches: An exploratory data analysis of hacking breaches over time,” Procedia Computer Science, vol. 151, pp. 1004–1009, 2019.

1. rack T. Majority of malware analysts aware of data breaches not disclosed by their employers. http://www.threattracksecurity.com/press-release/majority-of-malware-analystsaware-of-databreaches-not-dis closed-by-their-employers.aspx
2. K. Pujitha , Kattamanchi Prem Krishna , K. Amala , Annavarapu Yasaswini , Sivakumar Depuru , Kopparam Runvika, “Development of Secured Online Parking Spaces”, Journal of Pharmaceutical Negative Results, vol. 13, no. 4, pp. 1010–1013, Nov. 2022.
3. Sivakumar Depuru , Anjana Nandam , P.A. Ramesh , M. Saktivel , K. Amala , Sivanantham. (2022). Human Emotion Recognition System Using Deep Learning Technique. Journal of Pharmaceutical Negative Results, 13(4), 1031–1035. https://doi.org/10.47750/pnr.2022.13.04.141 (Original work published November 4, 2022) [10] S. Depuru, P. Hari, P. Suhaas, S. R. Basha, R. Girish and P. K. Raju, "A Machine Learning based Malware Classification Framework," 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2023, pp. 1138-1143, doi: 10.1109/ICSSIT55814.2023.10060914
4. S. Depuru, K. Vaishnavi, B. Manogna, K. J. Sri, A. Preethi and C. Priyanka, "Hybrid CNNLBP using Facial Emotion Recognition based on Deep Learning Approach," 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 2023, pp. 972-980, doi: 10.1109/ICAIS56108.2023.10073918.
5. Ayyagari, R. (2012). An exploratory analysis of data breaches from 2005-2011: Trends and insights. Journal of Information Privacy and Security
6. Algarni, A. M., Malaiya, Y. K. (2016, May). A consolidated approach for estimation of data security breach costs. In 2016 2nd International Conference on Information Management (ICIM) (pp. 26-39). IEEE.
7. Kafali, Jones, J., Petruso, M., Williams, L., Singh, M. P. (2017, May). How good is a security policy against real breaches? A HIPAA case study. In 2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE) (pp. 530-540). IEEE.
8. Sen, R., Borle, S. (2015). Estimating the contextual risk of data breach: An empirical approach. Journal of Management Information Systems, 32(2), 314-341.
9. Bertino, E., & Ferrari, E. (2018). Big data security and privacy,”. In A comprehensive guide through the Italian database research over the last 25 years (pp. 425–439). Springer.

Gray, J., Gerlitz, C., & Bounegru, L. (2018).

1. Smith, T. T. (2016). Examining Data Privacy Breaches in Healthcare
2. A. Bachar, N. E. Makhfi, O.E. Bannay, "Towards a behavioral network intrusion detection system based on the SVM model", in 2020 1st international conference on innovation research in applied science engineering and technology (IRASET), Meknes, Morocco, 2020, pp. 1-7
3. A. Bachar, N. E. Makhfi, O.E. Bannay, "Towards a behavioral network intrusion detection system based on the SVM model", in 2020 1st international conference on innovation research in applied science engineering and technology (IRASET), Meknes, Morocco, 2020, pp. 1-7
4. L. Bilge, Y. Han, and M. Dell’Amico, "Riskteller: Predicting the risk of cyber incidents", in Proc. of the 2017 ACM SIGSAC conf. on Computer and Communications Security, 2017, pp. 1299–1311.
5. S. Sarkar, M. Almukaynizi, J. Shakarian, and P. Shakarian, "Predicting enterprise cyber incidents using social network analysis on dark web hacker forums", The Cyber Defense Review, pp. 87–102, 2019

1. M. Lopez-Martin, B. Carro, J. I.Arribas, and A. Sanchez-Esguevillas, "Network intrusion detection with a novel hierarchy of distances between embeddings of hash IP addresses",Knowledge-based Syst.,vol 219,2021