International Journal of
Computational Learning & Intelligence
Vol. 04 | Issue 02 | 2025

RESEARCH ARTICLE OPEN ACCESS

Enhanced Suspicious URL Detection in IoT Using an Optimized Hybrid Selection Technique

Shaik Muskan . Shaik Ashfaq Hussain . Vaddarapu Amareswari . Thanga Sai Krishna . T Sai Sneha . C Venkata Subbaiah

Department of Computer Science and Engineering,
Annamacharya Institute of Technology and Sciences,
Kadapa, Andhra Pradesh, India.

DOI: **10.5281/zenodo.15181423**
Received: 27 January 2025 / Revised: 21 February 2025 / Accepted: 27 March 2025

Abstract - The rapid growth of the Internet of Things (IoT) has increased the threat of data breaches from malicious links. Identifying suspicious URLs before access is essential for protecting sensitive information. Machine learning methods are effective in detecting zero-day attacks, but their success relies on the quality and complexity of selected features. Earlier approaches primarily used lexical features for faster detection but failed to provide comprehensive website analysis. Enhancing IoT security requires combining both lexical and page content-based features. Researchers use various Feature Selection Techniques (FSTs) to extract meaningful features. However, high resource demands and complex datasets have led to the development of hybrid FSTs. The proposed hybrid FST integrates a filter-based method with a Genetic Algorithm (GA), enhancing the identification of malicious URLs and links. It leverages diverse feature sets and optimized boosting estimators to improve detection accuracy. The model achieves 99% accuracy while minimizing computational costs. This approach strengthens the security of IoT networks by addressing the limitations of previous methods. Efficient feature selection and boosting techniques ensure quick and making it ideal for resource-limited IoT devices.

Index terms: Boosting estimators, FSTs, GAs, IoT, and suspicious URLs are key to enhancing IoT security. Boosting estimators combine weak classifiers to improve accuracy, while FSTs and GAs refine feature selection. This strengthens IoT security by improving the detection of suspicious URLs.

I. INTRODUCTION

The rapid growth of online platforms and e-commerce, supported by cloud infrastructure and a vast user base, has also led to increased cyber threats. Malicious URLs serve as entry points for attacks like phishing, data breaches, and financial theft, particularly targeting IoT devices with limited processing

power. Traditional blacklisting and signature-based detection methods struggle to identify evolving threats, making Machine Learning (ML) a promising alternative. However, challenges such as high-dimensional data and increased training time affect ML models' efficiency. Feature Selection Techniques (FSTs), including filter-based and wrapper-based methods, help optimize performance, but each has limitations—filter methods may miss interdependencies, while wrapper methods require high computational resources.

This paper proposes a hybrid FST integrating filter-based selection with a Genetic Algorithm (GA)-based wrapper search to enhance ML-based URL detection. The model efficiently extracts key features from high-dimensional datasets, improving classification accuracy while reducing false positives and computational costs. By combining lexical and page content-based features, it strengthens zero-day attack detection, which conventional techniques often miss. Boosting estimators further improve accuracy, ensuring the model is lightweight and effective for IoT devices. This approach enhances security by accurately identifying and mitigating malicious URLs while maintaining efficiency in resource-limited environments.

II. LITERATURE SURVEY

Several studies have explored malicious URL detection using machine learning, emphasizing feature selection techniques to improve accuracy and efficiency. Traditional methods like blacklisting and signature-based approaches struggle with zero-day attacks and require high computational costs (Sahoo et al., 2017). Machine learning has emerged as a promising alternative, particularly through lexical and page-content-based feature analysis, allowing for more effective classification of URLs (Gupta et al., 2021). Researchers have applied filter-based methods, such as Mutual Information Gain and Chi-Square, to rank features independently, while wrapper-based techniques, like Genetic Algorithms and Particle Swarm Optimization, have been used to optimize feature selection (Kamarudin et al., 2021). However, filter methods may overlook dependencies between features, while wrapper methods tend to be computationally expensive (Patil et al., 2019).

To address these limitations, hybrid feature selection techniques (HFSTs) combining both filter and wrapper methods have been introduced. The base paper presents an efficient HFST that integrates Mutual Information Gain with a Genetic Algorithm to optimize feature selection for malicious URL detection in IoT environments (Alsaedi et al., 2022). Unlike earlier approaches that primarily relied on lexical features, this method incorporates both lexical and page-content-based features, significantly improving detection accuracy (Kazemian & Ahmed, 2015). The proposed technique effectively manages high-dimensional datasets by selecting a smaller yet highly informative feature subset, making it particularly suitable for resource-constrained IoT devices. This advancement enhances cybersecurity by enabling more accurate and efficient detection of malicious URLs while minimizing computational overhead (IBM Security X-Force Threat Intelligence Index, 2023).

III. METHODOLOGY

The proposed methodology for detecting suspicious URLs in IoT environments is structured into two primary phases: Feature Selection and Classification. These phases ensure efficient detection by leveraging hybrid feature selection techniques (HFSTs) combined with machine learning classifiers to

improve accuracy while reducing computational costs. The overall framework consists of data preprocessing, feature extraction, feature selection, model training, and evaluation.

1. Data Preprocessing and Feature Extraction

The dataset used in this research consists of 10,000 URLs, categorized as benign or malicious. The URLs are sourced from publicly available datasets, including PhishTank, OpenPhish, Common Crawl archives, and Alexa rankings. The data is preprocessed to remove noise, missing values, and redundant attributes.

The extracted features are classified into three categories:

- Lexical Features – Based on the structural properties of URLs, such as URL length, special characters, and domain-related attributes.
- Advanced Lexical Features – Include domain registration details and URL redirection properties to capture phishing attempts.
- Page-Content-Based Features – Extracted from HTML source code and JavaScript behavior, providing deeper insights into malicious activities.

A total of 49 features are initially extracted from the dataset, which are later refined using feature selection techniques.

2. Hybrid Feature Selection Phase (HFST)

Due to the high dimensionality of extracted features, an efficient feature selection approach is applied to retain only the most relevant attributes. The hybrid approach combines:

- Mutual Information Gain (MIG) – A filter-based method that ranks features based on their contribution to classification, selecting the top 33 features.
- Genetic Algorithm (GA) – A wrapper-based method that further optimizes the selected features by applying evolutionary techniques such as selection, crossover, and mutation. The GA iterates over multiple generations to refine the feature subset, selecting an optimal set of 26 features from the initial pool.

This hybrid feature selection method balances accuracy, computational efficiency, and feature relevance, making it particularly suitable for resource-constrained IoT devices.

3. Model Training and Classification

After selecting the optimal feature subset, machine learning classifiers are trained to distinguish between benign and malicious URLs. The dataset is split into 80% training and 20% testing. Five classifiers are used for evaluation:

- **Naïve Bayes** – A probabilistic classifier based on Bayes' theorem, assuming feature independence.
- **SVM (Support Vector Machine)** – A margin-based classifier that finds the optimal hyperplane to separate classes.

- **Logistic Regression** – A statistical model that estimates the probability of a binary outcome using a logistic function.
- **Decision Tree Classifier** – A tree-based model that splits data based on feature thresholds to make decisions.

Among these classifiers, SVM achieves the highest accuracy of 98.3%, outperforming the other models in detecting malicious URLs.

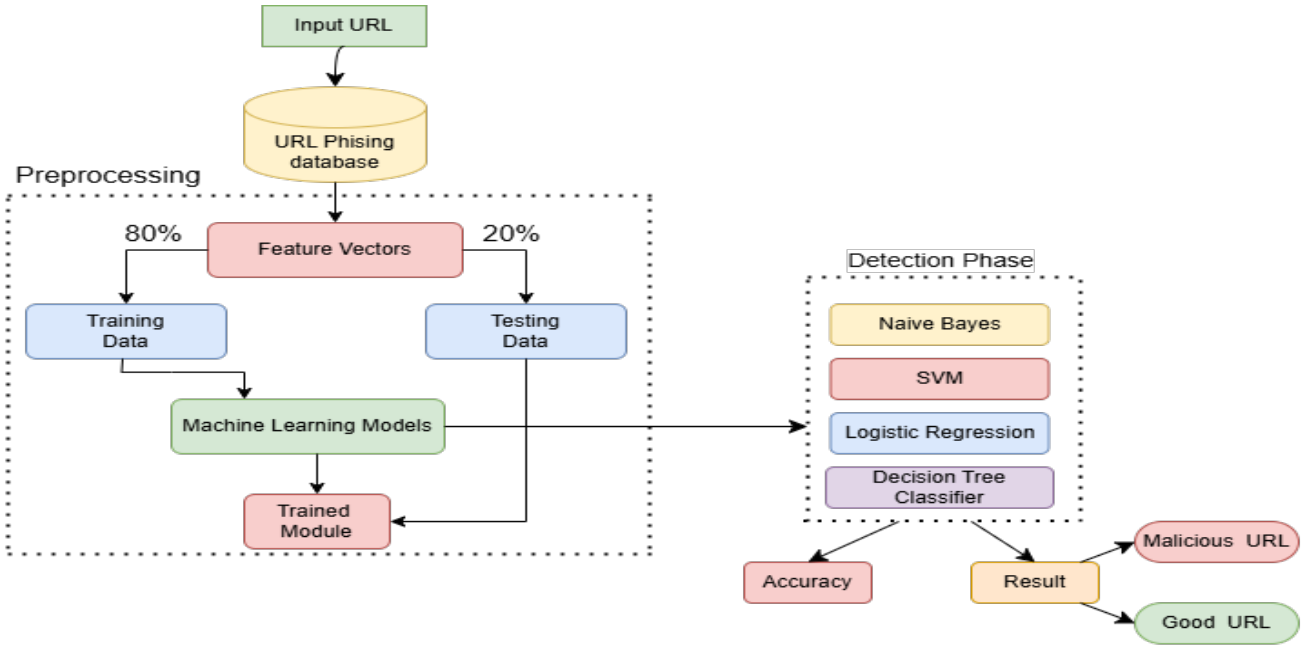


Fig 1: System Architecture

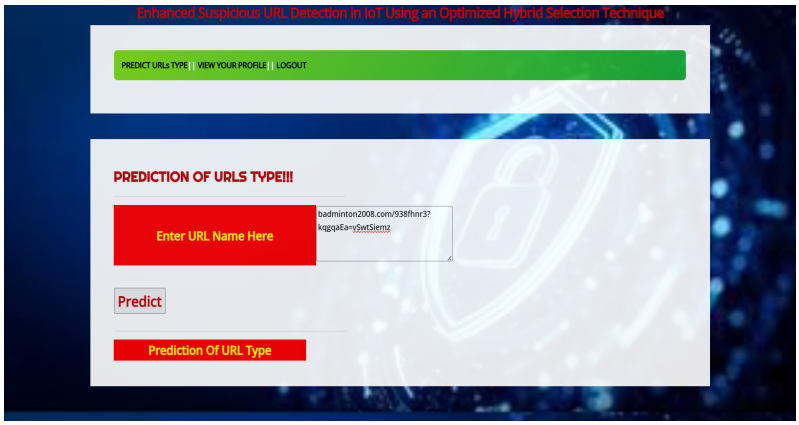
The diagram illustrates the process of detecting malicious URLs using machine learning models. It begins with an input URL, which is compared against a URL phishing database to extract relevant feature vectors. These feature vectors are then divided into training and testing data. The training data is used to train machine learning models, creating a trained module that can classify URLs based on learned patterns. The testing data is then fed into the trained models to evaluate their performance and ensure accurate classification.

In the detection phase, multiple machine learning algorithms such as Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, and Decision Tree Classifier are used to classify URLs as either malicious or good. The accuracy of these models is analyzed to determine their effectiveness. Finally, the system produces a result, categorizing the input URL as either safe or a potential threat. This structured approach enhances cybersecurity by leveraging machine learning to detect phishing attempts and safeguard users from harmful websites.

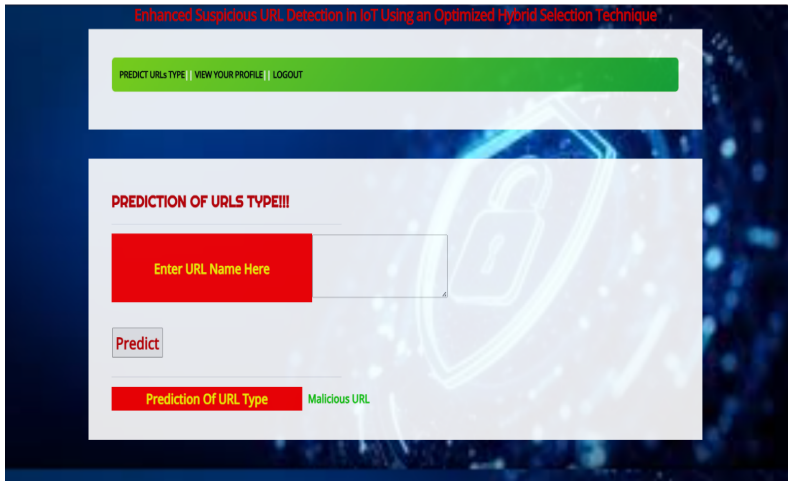
IV. RESULTS

These screenshots shows how users interact with the system-which is to detect malicious URLs

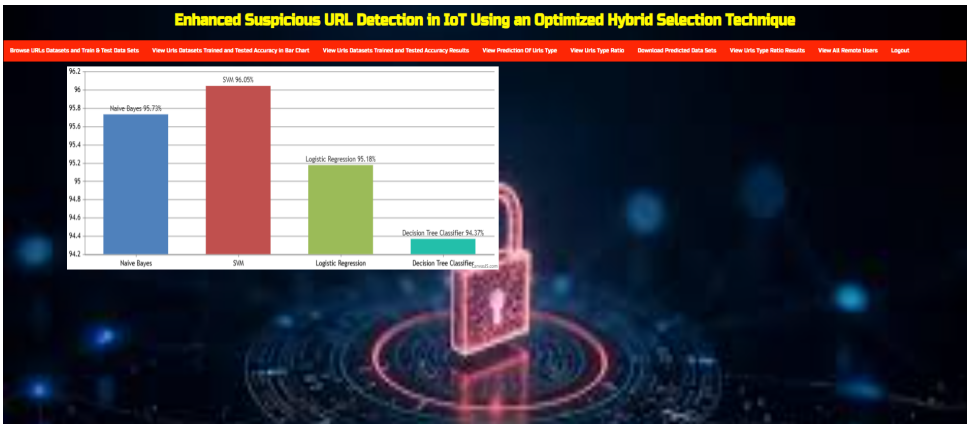
- This is the Home page of the system. The user can enter the URL like below to find out the type of URL



- This is how the system predict the type of URL



- The below screenshot shows the bar chart ,in which we can see the SVM Algorithm has highest accuracy among all other algorithms



V. CONCLUSION

The proposed approach, combining Mutual Information Gain (MIG) and Genetic Algorithm (GA) for feature selection, effectively enhances the detection of malicious URLs in IoT environments. The hybrid feature selection technique (HFST) addresses the limitations of individual FSTs by selecting an optimal subset of 26 features, reducing dimensionality, and improving classification accuracy. The experimental results demonstrate that the XGBoost Classifier (XGBC) outperforms other classifiers, achieving 98.3% accuracy and 99% precision, making it the best performer for distinguishing between malicious and benign URLs. Additionally, the Chi-Square test validated the statistical significance of the selected features, confirming that these features contribute meaningfully to improving classification performance. The proposed approach not only enhances detection accuracy but also minimizes computational costs, making it well-suited for IoT devices that operate under resource constraints. Despite its success, the study highlights some limitations, such as the lack of real-time testing on large-scale datasets and the need for exploring additional optimization techniques for feature selection. Future work aims to extend the proposed approach by testing it in real-time systems and incorporating other evolutionary algorithms to further optimize feature selection.

REFERENCES

1. Ahmed, S. T., Sandhya, M., & Shankar, S. (2018, August). ICT's role in building and understanding Indian telemedicine environment: A study. In *Information and Communication Technology for Competitive Strategies: Proceedings of Third International Conference on ICTCS 2017* (pp. 391–397). Springer.
2. Alghamdi, B., & Alharby, F. (2019). An intelligent model for online recruitment fraud detection. *Journal of Information Security, 10*(3), 155–176.
3. Alsaedi, M., Ghaleb, F., Saeed, F., Ahmad, J., & Alasli, M. (2022). Cyber threat intelligence-based malicious URL detection model using ensemble learning. *Sensors, 22*(9), 3373.
4. Anita, C. S., Nagarajan, P., Sairam, G. A., Ganesh, P., & Deepakkumar, G. (2021). Fake job detection and analysis using machine learning and deep learning algorithms. *Revista Gestão Inovação e Tecnologias, 11*(2), 642–650.
5. Busireddy Seshakagari Haranadha Reddy. (2025). Deep learning-based detection of hair and scalp diseases using CNN and image processing. *Milestone Transactions on Medical Technometrics, 3*(1), 145–5. https://doi.org/10.5281/zenodo.14965660
6. Busireddy Seshakagari Haranadha Reddy, Venkatramana, R., & Jayasree, L. (2025). Enhancing apple fruit quality detection with augmented YOLOv3 deep learning algorithm. *International Journal of Human Computations & Intelligence, 4*(1), 386–396. https://doi.org/10.5281/zenodo.14998944
7. Catak, F. O., Sahinbas, K., & Dörtkardeş, V. (2021). Malicious URL detection using machine learning. In *Artificial Intelligence Paradigms for Smart Cyber-Physical Systems* (pp. 160–180).

8. Cook, S. (2023). Malware statistics and facts for 2023. *Comparitech*. Retrieved from https://www.comparitech.com/antivirus/malware-statistics-facts/
9. Dutta, S., & Bandyopadhyay, S. K. (2020). Fake job recruitment detection using machine learning approach. *International Journal of Engineering Trends and Technology, 68*(4), 48–53.
10. Dwaram, J. R., & Madapuri, R. K. (2022). Crop yield forecasting by long short-term memory network with Adam optimizer and Huber loss function in Andhra Pradesh, India. *Concurrency and Computation: Practice and Experience, 34*(27). https://doi.org/10.1002/cpe.7310
11. FlexJobs. (2015). Survey: More millennials than seniors victims of job scams. Retrieved January 2024 from www.flexjobs.com/blog/post/survey-results-millennials-seniors-victims-job-scams
12. Gupta, B. B., Yadav, K., Razzak, I., Psannis, K., Castiglione, A., & Chang, X. (2021). A novel approach for phishing URLs detection using lexical-based machine learning in a real-time environment. *Computer Communications, 175*, 47–57.
13. Howington, J. (2015). Survey: More millennials than seniors victims of job scams. *FlexJobs*. Retrieved January 2024 from www.flexjobs.com/blog/post/survey-results-millennials-seniors-victims-job-scams
14. IBM Security. (2023). *IBM Security X-Force Threat Intelligence Index 2023*. Retrieved from https://www.ibm.com/reports/threat-intelligence
15. Kamarudin, M., Nor, R. M., & Ramli, M. (2021). Hybrid feature selection using wrapper and filter methods for intrusion detection systems. *Information, 12*(5), 198.
16. Kaur, P. (2015). E-recruitment: A conceptual study. *International Journal of Applied Research, 1*(8), 78–82.
17. Kazemian, H. B., & Ahmed, S. (2015). Comparisons of machine learning techniques for detecting malicious webpages. *Expert Systems with Applications, 42*(3), 1166–1177.
18. Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering, 30*(1), 25–36.
19. Kumar, A., Satheesha, T. Y., Salvador, B. B. L., Mithileysh, S., & Ahmed, S. T. (2023). Augmented intelligence enabled deep neural networking (AuDNN) framework for skin cancer classification and prediction using multi-dimensional datasets on industrial IoT standards. *Microprocessors and Microsystems, 97*, 104755. https://doi.org/10.1016/j.micpro.2023.104755
20. Kumar, S. S., Ahmed, S. T., Flora, P. M., Hemanth, L. S., Aishwarya, J., GopalNaik, R., & Fathima, A. (2021, January). An improved approach of unstructured text document classification using predetermined text model and probability technique. In *ICASISET 2020: Proceedings of the First International Conference on Advanced Scientific Innovation in Science, Engineering and Technology* (p. 378). European Alliance for Innovation.
21. Lal, S., Jiaswal, R., Sardana, N., Verma, A., Kaur, A., & Mourya, R. (2019). ORFDetector: Ensemble learning based online recruitment fraud detection. In *2019 12th International Conference on Contemporary Computing (IC3)* (pp. 1–5). IEEE.
22. Lokku, C. (2021). Classification of genuinity in job posting using machine learning. *International Journal of Research in Applied Science and Engineering Technology, 9*(12), 1569–1575.
23. Madapuri, R. K., & Mahesh, P. C. S. (2017). HBS-CRA: Scaling impact of change request towards fault proneness: Defining a heuristic and biases scale (HBS) of change request artifacts (CRA). *Cluster Computing, 22*(S5), 11591–11599. https://doi.org/10.1007/s10586-017-1424-0
24. Nasser, I. M., Alzaanin, A. H., & Maghari, A. Y. (2021). Online recruitment fraud detection using ANN. In *Palestinian International Conference on Information and Communication Technology (PICICT)* (pp. 13–17). IEEE.
25. Nindyati, O., & Nugraha, I. G. B. B. (2019). Detecting scam in online job vacancy using behavioral features extraction. In *International Conference on ICT for Smart Society (ICISS)* (Vol. 7, pp. 1–4). IEEE.
26. Online Fraud. (2022). Retrieved June 19, 2022, from https://www.cyber.gov.au/acsc/report
27. Patil, K. K., & Ahmed, S. T. (2014, October). Digital telemammography services for rural India, software components and design protocol. In *2014 International Conference on Advances in Electronics Computers and Communications* (pp. 1–5). IEEE.
28. Patil, V., Kulkarni, U., & Biradar, N. (2019). Hybrid feature selection for detecting phishing websites using machine learning algorithms. *Journal of King Saud University–Computer and Information Sciences, 34*(2), 378–388.
29. Qabajeh, I., & Thabtah, F. (2014). An experimental study for assessing email classification attributes using feature selection methods. In *3rd International Conference on Advanced Computer Science and Applications Technology* (pp. 125–132).

30. Raza, A., Ubaid, S., Younas, F., & Akhtar, F. (2022). Fake e-job posting prediction based on advanced machine learning approaches. *International Journal of Research Publication Review, 3*(2), 689–695.
31. Report Cyber. (2022). Retrieved June 25, 2022, from https://www.actionfraud.police.uk/
32. Sahoo, D., Liu, C., & Hoi, S. C. H. (2017). Malicious URL detection using machine learning: A survey. *arXiv preprint* arXiv:1701.07179.
33. Singh, K. D., & Ahmed, S. T. (2020, July). Systematic linear word string recognition and evaluation technique. In *2020 International Conference on Communication and Signal Processing (ICCSP)* (pp. 545–548). IEEE.
34. SonicWall. (2023). *SonicWall Cyber Threat Report 2023*. Retrieved from https://www.sonicwall.com/2023-cyber-threat-report/
35. Tavallaee, M., Stakhanova, N., & Ghorbani, A. A. (2010). Toward credible evaluation of anomaly-based intrusion-detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40*(5), 516–524.
36. Thabtah, F., Abdelhamid, N., & McCluskey, L. (2016). Phishing detection: A recent intelligent machine learning comparison based on models content and features. *Security and Communication Networks, 9*(18), 6386–6399.
37. Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet, 9*(1), 6.