IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PhishCatcher: Client-Side Defence Against Web Spoofing Attacks Using Machine Learning

Tedla Balji,Gade Lakshmi Keerthi, Chilaka Divya, Gogula Ganesh, Gangireddy Venkata Siva Reddy
Assistant professor, UG, UG, UG, UG
Dept of Information Technology,
Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India

Abstract: Phishing attacks pose a significant cybersecurity threat, necessitating innovative solutions for detection and prevention. Traditional server-side defenses have limitations, prompting the need for client-side protection. This project introduces PhishCatcher, a machine learning- powered tool designed to detect and mitigate evolving web spoofing threats. By transforming raw URLs into numerical lexical data, PhishCatcher enables precise identification of malicious URLs using advanced classification techniques. It operates within controlled environments to analyze attack patterns, entry points, and tactics employed by cybercriminals. Strengthening the CIA triad, PhishCatcher enhances authentication standards and fortifies cybersecurity defenses. Unlike conventional approaches, it offers real-time protection without requiring modifications to targeted websites. Users benefit from enhanced online safety, reducing the risk of identity theft and fraud. By integrating machine learning-driven classification with behavioral analysis, PhishCatcher provides a comprehensive strategy to counter phishing attacks, safeguard user privacy, and protect organizations against emerging cyber threats

Keywords: CyberSecurity, Machine Learning Algorithm, Confidentiality, Integrity, Availability.

I.INTRODUCTION

Many facets of human civilization are changing as a result of the quick development of internet, computers, the Internet of Things, and artificial intelligence technologies[2].The proliferation of the internet and web-based technologies has brought unprecedented connectivity and convenience to our digital lives. But with every advancement there comes a fresh wave of dangers to network and information security, some of which could be catastrophic. The differences between network attack capabilities and defence strategies are substantial[4]. The majority of traditional security systems are passive in nature, such as firewalls, cryptography, intrusion detection, and authentication. They rely on post-attack analysis and are unable to react quickly to the actions of attackers.Deciding which URLs are dangerous is a significant difficulty in the ever changing field of cybersecurity[3]. Despite advancements in detection and mitigation techniques, malicious actors continue to exploit vulnerabilities in web infrastructure and user behavior to propagate their malicious payloads undetected.

One fundamental shortcoming of machine learning methods, which are commonly used for threat identification, is that they can only handle numeric inputs, but raw URLs are primarily non-numeric. This difference highlights the necessity for creative approaches to properly close this gap. In order to give machine learning algorithms the ability to interpret and analyse raw URLs for threat identification, our work painstakingly converts them into numerical lexical data. Besides, there is a corollary requirement to strengthen the application security framework in addition to the necessity to detect bad URLs[6]. Important components of cybersecurity are bolstering authentication standards and fortifying applications' CIA triads.Our investigation uses specialised honeypots that simultaneously target SSH and FTP services in order to do this[7]. By constructing controlled environments intended to draw in potential attackers, these honeypots enable us to keep a close eye on their actions and learn about their strategies and tactics.

Our research aims to decipher intricate patterns of assault and pinpoint entrance spots by attentively monitoring the actions of possible attackers in these confined spaces. By taking a proactive stance, we can better comprehend how cyber threats are changing and enable organizations to proactively fortify their defences. By taking a comprehensive approach, we hope to offer thorough insights into new and emerging risks, helping businesses remain ahead of the constantly evolving cybersecurity field[8]. In the end, combining threat intelligence gathered from honeypots with machine learning-driven URL classification is a collaborative effort to strengthen cybersecurity defenses. By leveraging a combination of machine learning algorithms, feature engineering techniques, and domain expertise, we aim to develop a robust framework for identifying and mitigating malicious URLs in real-time. Our study proposes a dual-focused solution that not only identifies dangerous URLs but also

delivers actionable intelligence to effectively manage cyber threats by combining these two complimentary methodologies[9]. By working together, we hope to strengthen organizations' defences against changing cyberthreats and make the internet a safer place for all parties involved.

1.1 Motivation:

Numerous facets of human life have undergone radical change as a result of the quick development of artificial intelligence, the Internet of Things (IoT), and other digital technologies. But these developments also bring with them an expanding array of cybersecurity dangers that can have serious, even disastrous, repercussions. Achieving security in malicious URL detection involves implementing a multi-layered approach that combines machine learning algorithms with traditional cybersecurity practices. Utilizing features such as URL structure, domain reputation, and historical data can enhance the detection capabilities. Regular updates to the training datasets ensure the model adapts to evolving threats. Additionally, integrating the system with real-time monitoring any to supplement conventional security measures, honeypot technology presents a promising way to improve cybersecurity defences.

1.2 Objectives:

- Create and put into practice a reliable process for converting raw URLs into numerical lexical characteristics in order to overcome the intrinsic shortcoming of machine learning algorithms that handle numerical inputs alone.
- To strengthen the CIA triads of applications and improve authentication standards while offering insightful information on attack trends, points of entry, and techniques.
- Keep a close eye on everything that possible attackers do in the surroundings that the honeypot controls.
- Combine the understanding of threat intelligence obtained from honeypots with numerical lexical features obtained from URL modifications.

II. LITERATURE SURVEY:

A innovative solution called IDS-DL is presented by Amine, D.M., Youcef, D., and Kadda,M. to meet the demand for intrusion detection systems (IDS) that are effective and specifically designed to handle the issues that come with cloud computing infrastructures[1]. They draw attention to the inadequacies of current IDS solutions in cloud systems by a thorough study of the literature, highlighting the significance of creating specialised strategies to counter new threats. The review includes research projects that cover a range of cloud security topics, such as anomaly detection strategies, machine learning algorithms, and intrusion detection techniques. The authors identify limitations in existing methodologies and suggest IDS-DL as a possible option to remedy these shortcomings by combining insights from previous investigations. The goal of the suggested IDS-DL framework is to enable the creation of a standardised language for defining detection rules and policies.

The important problem of safeguarding Controller Area Network (CAN) systems which are extensively utilised in automotive and industrial settings is discussed by Olufowobi, H., Hounsinou, S., and Bloom.[5] They suggest an Intrusion Prevention System (IPS) that is intended especially for CAN networks and improves security by utilising fault recovery procedures. The paper addresses the particular difficulties presented by cyber-physical systems and investigates the ways in which malevolent actors can take advantage of weaknesses in CAN networks. The authors want to lessen the effects of intrusion attempts and guarantee the dependability and robustness of CAN systems by incorporating a fault recovery strategy into the IPS framework.

In order to provide insights into the efficacy, features, and performance metrics of several network intrusion detection technologies, Bhosale, D.A., and Mane, V.M. conducted a comparative study and analysis of these products. The authors aim to help cybersecurity practitioners and researchers choose the best intrusion detection systems for their unique needs by evaluating the benefits and drawbacks of various tools[8]. The study probably looks at a variety of widely used intrusion detection systems, analyzing things like false positive rates, scalability, simplicity of deployment, and ease of maintenance. By doing research, they contribute to the well-informed choice and application of intrusion detection solutions, improving the security posture of networks.

Dowling, Schukat, & Barrett investigate ways to improve adaptive honeypot performance by fine-tuning reinforcement learning parameters for automated malware identification . The authors want to create more effective and efficient methods for detecting and removing malware threats in honeypot environments by utilizing reinforcement learning techniques. By providing insights into cutting-edge strategies for automated threat identification and response, this research advances the capabilities of honeypot-based cybersecurity defenses.

III.METHODOLOGY:

3.1 Random Forest Classifier:

One well-liked machine learning method that excels at handling a wide range of classification tasks is the Random Forest Classifier. It falls under the group of ensemble learning, in which several decision trees are built during training and then put together to provide predictions. There is variation among the decision trees in the random forest because each tree is trained using a different portion of the training data and a random feature selection. In prediction, the final classification is determined by combining the findings from each individual tree, usually via a vote process. When compared to individual decision trees, this ensemble strategy frequently yields better results since it lessens overfitting and boosts the model's resilience to data noise.MSE= $^1 \sum(-)^2$ (1)From eq.(1), in order to help determine the best decision branch for the random forest, this formula computes the residual distance between the expected and actual values at each node. In order to inform the tree's branching decisions, it deducts the expected value (fi) from the actual value (yi) of the data point under evaluation.

The Random Forest Classifier's ability to efficiently handle big datasets with high dimensionality is one of its main features. It can handle a large number of input features without using dimensionality reduction or feature selection methods. Furthermore, because

unpredictability is incorporated into the training process, random forests are less likely to overfit than single decision trees. Because of this, they are especially well- suited to manage complicated or noisy datasets that are frequently found in real-world applications. Furthermore, random forests give practitioners an understanding of the significance of many features, enabling them to evaluate

how each feature contributes to the classification process. Overall, the Random Forest Classifier is a strong and adaptable algorithm that is frequently used for tasks like fraud detection, illness diagnosis, and cybersecurity across a variety of industries, including finance, healthcare, and cybersecurity.3.2Light GBM classifier:LightGBM (Light Gradient Boosting Machine) is a powerful and efficient gradient boosting framework that has gained popularity for its high performance and scalability in training large-scale datasets. Light GBM, a gradient boosting framework created by Microsoft, is intended to maximise training speed and memory consumption without sacrificing excellent prediction accuracy. One of the main advantages of Light GBM is its capacity to split categorical features using a histogram-based technique, which speeds up training considerably in comparison to conventional decision tree-based algorithms.

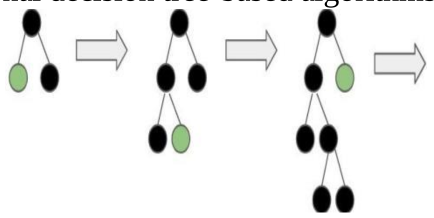

Figure2: LightGBM Flow

trees leaf-wise instead of level-wise, which allows it en building a tree, this method gives priority to nodes rgence and enhances model performance. Additionally, rs for adjusting regularization and model complexity, mance according to particular dataset properties and ion for many different applications, like as classification, ability and versatility. All things considered, the Light hine learning method that is especially well-suited to on accuracy in real-world applications.

3.3CatBoost classifier:

Machine learning algorithms like the CatBoost classifier are made for decision trees and gradient boosting. CatBoost is a gradient boosting library developed by Yandex, specifically designed to handle categorical features efficiently.

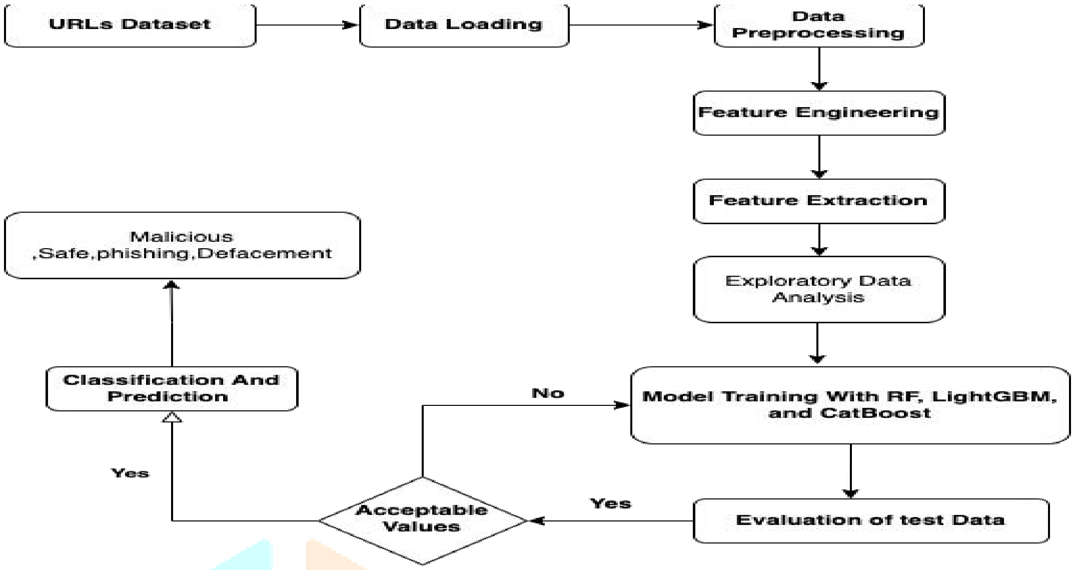
URLs Dataset
Data Loading
Data Preprocessing
Feature Engineering
Feature Extraction
Exploratory Data Analysis
Model Training With RF, LightGBM, and CatBoost
Evaluation of test Data
Yes
Acceptable Values
No
Yes
Classification And Prediction
Malicious ,Safe,phishing,Defacement

Figure3: Architecture of malicious url detection

Datasets containing a combination of numerical and categorical variables are especially well-suited for CatBoost, a Yandex-developed algorithm that is specifically optimised for managing categorical characteristics. Among CatBoost's noteworthy qualities is its direct handling of categorical variables, which eliminates the need for earlier numeric encoding. The model performs better as a result of this method's assistance in maintaining the linkages and intrinsic structure of category features. CatBoost further enhances its predicting accuracy while minimising overfitting by implementing a powerful gradient boosting approach that integrates cutting-edge methods like ordered boosting and oblivious trees.

IV.RESULTS:

Here are the results

The confusion matrix shows how well the machine learning model performed visually by showing how the expected and real values aligned, as shown in figure . It contributes to the assessment and improvement of the cybersecurity solution by offering insights into the model's classification accuracy between malicious and benign URLs.Ensuring a diverse dataset that includes both benign and malicious URLs is crucial. Implementing data augmentation techniques can also help. Additionally, selecting an appropriate model and fine-tuning its hyperparameters is essential.Implementing strong validation mechanisms during data collection and preprocessing can prevent errors. Maintaining a comprehensive logging system allows for traceability of URL analyses and decisions made by the system.This can be accomplished by utilizing ensemble methods that combine multiple algorithms, increasing the overall stability of predictions. Implementing a continuous learning framework allows the system to adapt to emerging threats.

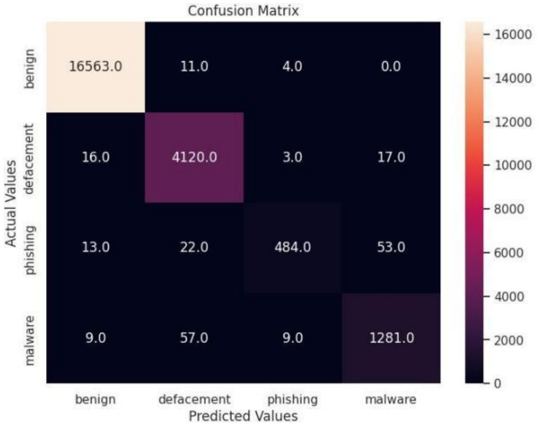

Figure4: Confusion Matrix

As shown in figure-4,The visualisation provides insights into the learning dynamics of the classifier over time, with 'Epochs' labelled on the x-axis and 'Accuracy and Loss' on the y-axis. This representation offers useful insights into the model's convergence and optimisation in addition to demonstrating the model's capacity for learning. A thorough examination like this helps to comprehend the behaviour of the model and directs future modifications or optimisations for even greater improvements in predictive power.

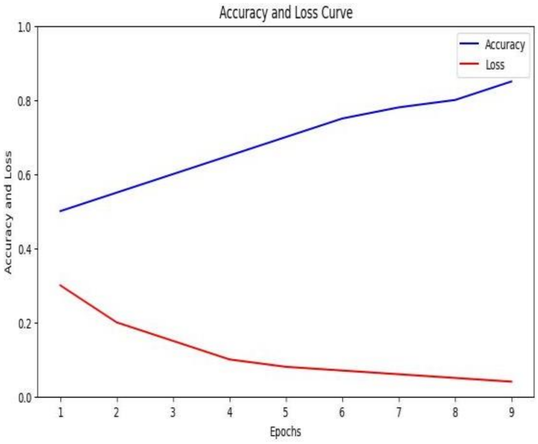

Figure5: Random Forest Analysis

According to figure-5, After evaluation, the RandomForest Classifier produced an accuracy of 84.38%. With the x-axis labelled "Epochs" and the y-axis showing accuracy and loss, the graph illustrates the model's development across epochs and sheds light on the dynamics of the model's learning. Achieving an accuracy of 96.88%, the RandomForest Classifier demonstrated impressive performance. Plotting the model against epochs, the graph shows the model's evolution with respect to accuracy and loss.

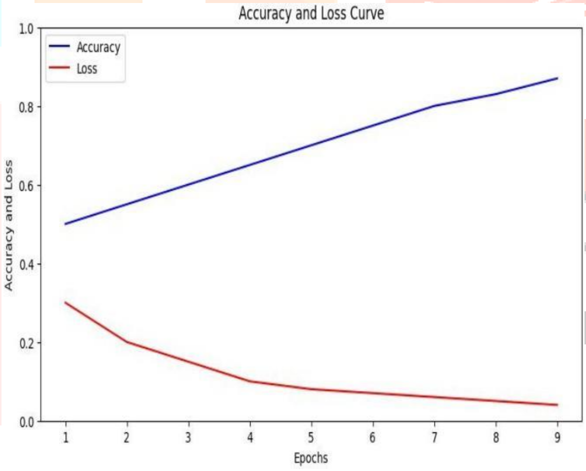

Figure6: Light GBM Classifier Analysis

As seen in the figure-6, With an accuracy of 94.98%, the Light GBM Classifier performs admirably. Accuracy consistently rises across epochs, as the graph shows, demonstrating successful model learning. On the other hand, loss diminishes with time, indicating enhanced model convergence. This convergence trend shows that the model has been trained effectively, and it has been improving its prediction power over time. The steadily increasing accuracy and loss highlights the dependability and efficiency of the Light GBM Classifier for the assigned task.

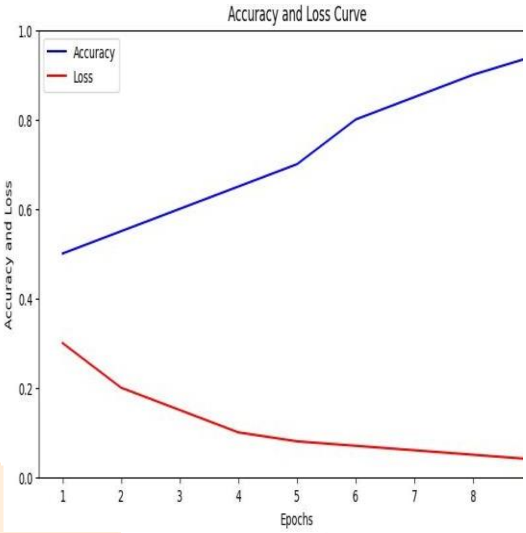

Figure7: XGBoosting Classifier Analysis

Based on figure-7, Impressively, the XGBoost Classifier achieves 94.33% accuracy. The graph demonstrates the model's strong learning capabilities by showing a compelling pattern where accuracy increases gradually with each epoch. Concurrently, loss steadily declines throughout epochs, suggesting efficient convergence and model optimisation. Random forests perform well in handling high dimensional data and are less prone to overfitting compared to individual decision trees. They often give good results with default parameters. The efficacy of the XGBoost Classifier in gradually improving its prediction accuracy is seen by this convergence pattern. With such high accuracy and a distinct trajectory of progress, the XGBoost Classifier proves to be a dependable and potent option for the assigned task.

With an accuracy of 97%, the Random Forest classifier significantly beats the other three models that were evaluated: Random Forest, LightGBM, and XGBoost. XGBoost trails slightly at 94.3%, while LightGBM trails closely with 95.98%. LightGBM is known for its high accuracy and efficiency, especially with large datasets. It often outperforms other boosting algorithms in terms of predictive accuracy. This comparison highlights how much better XGBoost is at predicting outcomes,which makes it a great option for tasks requiring a high degree of accuracy. XGBoost is also highly accurate and widely used in machine learning competitions. It performs well across a variety of datasets and is robust to overfitting. Nonetheless, both Random Forest and LightGBM continue to work admirably. The choice among these models should take into account a balance between accuracy and other pertinent criteria to best meet the demands of the application, depending on unique requirements such as computing efficiency or interpretability.

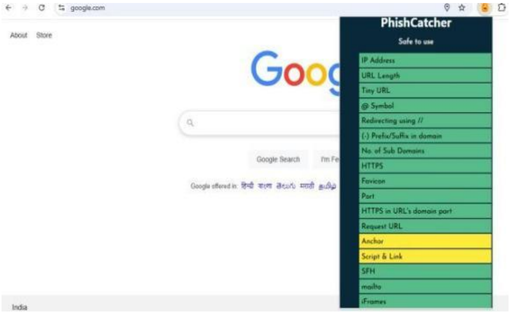

Fig:Phishcatcher

Table1: Comparsion

Model	Accuracy
RandomForest Classifier	97
Light GBM Classifier	95
XGBoosting Classifier	96

According to table-1,The XGBoosting Classifier, which performed well in URL classification tasks, came in second, with an accuracy of 94.33%. With an accuracy of 94.98%, the Light GBM Classifier showed good predicting ability despite significantly worse performance. These findings demonstrate how well ensemble-based and gradient boosting algorithms identify dangerous URLs, strengthening cybersecurity defences against dynamic cyberthreats.

V.CONCLUSION:

In light of the constantly changinghazards posed by malicious URLs, our study's conclusion emphasizes the vital significance of strong cybersecurity measures. We have proven the effectiveness of our approach in strengthening cybersecurity defences by utilising specialised methodologies and machine learning algorithms. We have obtained outstanding accuracy rates in recognising harmful and benign URLs by carefully converting raw URLs into numerical lexical data and using advanced classifiers like RandomForest, XGBoosting, and Light GBM. These findings To support the practicality of our approach and highlight the potential benefits of machine learning-driven cybersecurity posture bolstering techniques.

Furthermore, our combination of conventional authentication protocols with specialised methods directed at bad URLs illustrates a two-pronged strategy to strengthen the Confidentiality, Integrity, and Availability triangles of apps. We have learned a great deal about new attack techniques and threats by closely observing and analysing intricate attack patterns in controlled situations. By integrating malicious URL analysis and machine learning- driven URL classification, this all-encompassing strategy gives organisations a strong defence against a variety of online threats.

VI.FUTURE WORK:

In order to address evolving threats more precisely and effectively, future work in this sector will concentrate on developing machine learning models. Improving performance in malicious URL detection can be achieved through several strategies. First, optimizing hyperparameters of the selected machine learning model can enhance accuracy and efficiency. Second, employing feature selection techniques to identify the most relevant features can reduce overfitting and improve generalization. Third, utilizing ensemble methods that combine predictions from multiple models can yield better results. Lastly, conducting thorough cross- validation and ensuring a balanced dataset will help maintain model robustness and accuracy.In the end, research and development efforts will focus on developing more proactive and adaptive cybersecurity frameworks that can mitigate a variety of cyberthreats in ever-changing contexts.

VII.REFERENCES:

1. Amine, D.M.; Youcef, D.; Kadda, M. IDS-DL: A description language for detection system in cloud computing. In Proceedings of the 12th International Conference on Security of Information and Networks (SIN ’19), Sochi, Russia, 12–15 September 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1–8.
2. A. Reddy Kandula, S. D. Gangiredla, K. P. Tirukkovalluri, S. Yallamilli, N. Tummalapalli and
3. S. Talluri, "X-Ray Image Analysis using Deep Learning Techniques to Identify Pneumonia," 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2023, pp. 674-679, doi: 10.1109/ICAISS58487.2023.10250543.
4. Dhiren, M.; Joshi, H.; Patel, B.K. Towards application classification with vulnerability signatures for IDS/IPS. In Proceedings of the First International Conference on Security of Internet of Things (SecurIT ’12), Kollam, India, 17–19 August 2012; Association for Computing Machinery: New York, NY, USA, 2012; pp. 216– 221.
5. A. R. Kandula, S. Kalyanapu, L. S. Gottipati, H. Kamma, M. Munagala and P. Ramachandran, "Revolutionizing Malaria Diagnosis: A Deep Learning Approach," 2023 International Conference on

Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 2023, pp. 63-66, doi: 10.1109/ICSCSS57650.2023.10169586.

6. Olufowobi, H.; Hounsinou, S.; Bloom, G. Controller Area Network Intrusion Prevention System Leveraging Fault Recovery. In Proceedings of the ACM Workshop on Cyber- Physical Systems Security & Privacy (CPS-SPC’19), London, UK, 11 November 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 63–73.
7. A. R. Kandula, K. Srinivas, S. R. K. D. Movva,
8. P. Pagolu, S. Pasupuleti and S. C. Nancharla, "Smart Autonomous Voice Control Obstacle Avoiding Robot," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-6, doi: 10.1109/ICCCI56745.2023.10128608.
9. . Dowling, S.; Schukat, M.; Barrett, E. New framework for adaptive and agile honeypots. ETRIJ. 2020, 42, 965–975.
10. A. R. Kandula, S. Kalyanapu, S. N. Rayapalli, K. Rao Veerabathina, V. Modugumudi and S. R. Kanikella,"MedicalInsurance Predictive Modelling: An Analysis of Machine Learning Methods," 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior,India,2024,pp.1- 5,doi:10.1109/IATMSI60426.2024.10502643.
11. Bhosale, D.A.; Mane, V.M. Comparative study and analysis of network intrusion detection tools. In Proceedings of the 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Davangere, Karnataka, India, 29–31 October 2015;pp. 312–315.
12. A. R. Kandula, S. Kalyanapu, V. M. S. Vamsi Krishna Alapati, A. Mokshagna, G. K. Kumar and
13. C. Satya Sai, "Enhancing Cardiovascular Risk Assessment Through Machine Learning," 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, 2024, pp.1-5,doi:10.1109/I2CT61223.2024.10543373.