SN Computer Science (2024) 5:800
https://doi.org/10.1007/s42979-024-03099-3

ORIGINAL RESEARCH

A Study on Machine Learning and Deep Learning Techniques for Identifying Malicious Web Content

Sarita Mohanty[1] · **Asha Ambhakar**[1]

Received: 27 May 2024 / Accepted: 1 July 2024

Abstract
The rapid proliferation of internet usage has led to an exponential increase in cyber threats, particularly malicious websites that can compromise user data and system integrity. Traditional methods of web security are increasingly becoming obsolete, necessitating more dynamic and adaptive approaches. This research paper presents a comprehensive comparative study of Machine Learning (ML) and Deep Learning (DL) techniques for the detection of malicious websites. Utilizing a dataset of over 420,000 web URLs, categorized into various features such as domain, subdomain, and domain suffix, the study aims to evaluate the effectiveness, precision, and computational efficiency of multiple algorithms. Two Convolutional Neural Network (CNN) models were developed and compared against traditional ML algorithms including Decision Trees, Random Forests, AdaBoost, K-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD), Extra Trees, and Gaussian Naive Bayes. The models were rigorously evaluated based on metrics such as accuracy, precision, recall, and F1-score. Preliminary results indicate that CNN models outperform traditional ML algorithms, achieving an accuracy rate of up to 98%, thereby highlighting the potential of DL in cybersecurity applications. Moreover, the study addresses the challenges posed by high cardinality and class imbalance in the dataset. Various data preprocessing techniques were employed to mitigate these issues, including feature engineering and oversampling of minority classes. The research contributes to the field by providing a detailed analysis of each algorithm's strengths and weaknesses, thereby offering valuable insights into the adaptability and scalability of ML and DL techniques in malicious web detection.

Keywords KNN · ML · DL · Malicious · Web · Methods · Extra trees · Gaussian naive bayes · CNN · Accuracy · Precision · Recall and F1-score

Introduction

The advent of the internet has revolutionized the way we live, work, and interact. While it has brought about unparalleled convenience and access to information, it has also opened the floodgates to a myriad of cybersecurity threats. One of the most insidious forms of these threats is malicious websites, which can range from phishing sites that steal personal information to those that distribute malware or ransomware. As the internet continues to evolve, so do the tactics employed by cybercriminals, making the detection of malicious websites a continually challenging problem. Traditional methods, such as blacklisting and signature-based detection, are increasingly proving to be In this context ML and DL offer promising avenues for enhancing cybersecurity measures. These computational techniques have the ability to learn from data, adapt to new information, and make intelligent decisions, thereby providing a dynamic and robust approach to malicious web detection. However, while both ML and DL have been individually applied to cybersecurity problems, there is a lack of comprehensive research comparing their effectiveness, particularly in the domain of web security.This research paper aims to fill this gap by conducting an exhaustive comparative study of various ML and DL algorithms for the detection of malicious websites. Utilizing a dataset comprising over 420,000 web URLs, we evaluate the performance of multiple algorithms, including but not limited to, Convolutional Neural Networks (CNN),

✉ Sarita Mohanty
mohantysarita104@gmail.com

Asha Ambhakar
asha.ambhaikar@kalingauniversity.ac.in

[1] Department of Computer Science & Engineering, Kalinga University, Naya Raipur, India

Published online: 16 August 2024

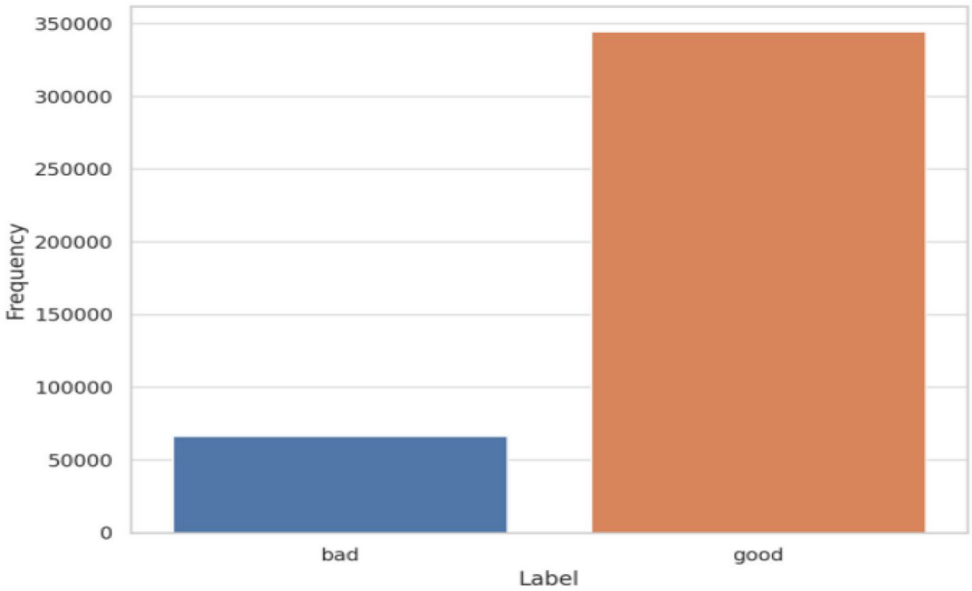

Fig. 1 Distribution of labels

Decision Trees, Random Forests, and AdaBoost. The study employs a range of evaluation metrics such as accuracy, precision, recall, and F1-score to provide a holistic view of each algorithm's capabilities.Moreover, the paper addresses the challenges inherent in cybersecurity datasets, such as high cardinality and class imbalance, and discusses the data preprocessing techniques employed to mitigate these issues. The ultimate goal is to provide a nuanced understanding of the strengths and weaknesses of ML and DL techniques in the context of malicious web detection, thereby aiding cybersecurity professionals, researchers, and policymakers in making informed decisions. In the rapidly evolving landscape of the internet, the prevalence of malicious web content poses a significant threat to users and organizations alike. Malicious content, ranging from phishing sites and malware-laden pages to fraudulent e-commerce platforms, can lead to substantial financial losses, data breaches, and compromised privacy. As these threats continue to grow in sophistication and volume, traditional detection methods, often reliant on predefined rules and signatures, struggle to keep pace [1].

In this context, machine learning (ML) and deep learning (DL) techniques offer promising solutions for identifying and mitigating malicious web content. Unlike conventional approaches, ML and DL can learn and adapt to new threats by analyzing vast amounts of data, detecting subtle patterns, and making predictions based on complex features that are not easily discernible through manual analysis. This study explores various machine learning and deep learning methodologies for detecting malicious web content. It aims to provide a comprehensive understanding of the current state-of-the-art techniques, evaluating their effectiveness, efficiency, and scalability. The research covers a spectrum of ML and DL models, including supervised, unsupervised, and reinforcement learning approaches, as well as neural networks and other advanced architectures. Key areas of focus include the preprocessing of web content data, feature extraction and selection, model training and validation, and the deployment of these models in real-world scenarios. By investigating the strengths and limitations of different techniques, this study seeks to identify optimal strategies for improving web security and safeguarding users against malicious threats.

Ultimately, this research endeavors to contribute to the development of more robust and adaptive systems capable of defending against the ever-changing landscape of web-based threats, leveraging the power of machine learning and deep learning to enhance cybersecurity measures.

Related Work

Kim(2017) [2] explore the difficulty of identifying dangerous websites, concentrating in particular on malware spread via the Internet. They present LoGos, a high-interaction dynamic analyzer designed for a Windows virtual machine environment that runs in a browser. The system keeps track of several indicators of fraudulent behaviour, including new connections to the network, unused ports that are open, and registry changes, using API hooks and Internet Explorer injection. Being 10 to 18 times faster that earlier systems, LoGos stands out for its speed without sacrificing detection rates. The programme has undergone extensive testing, with daily analyses of almost 0.36 million domains and 3.2 million webpages demonstrating its effectiveness and efficiency in identifying a variety of dangerous websites.

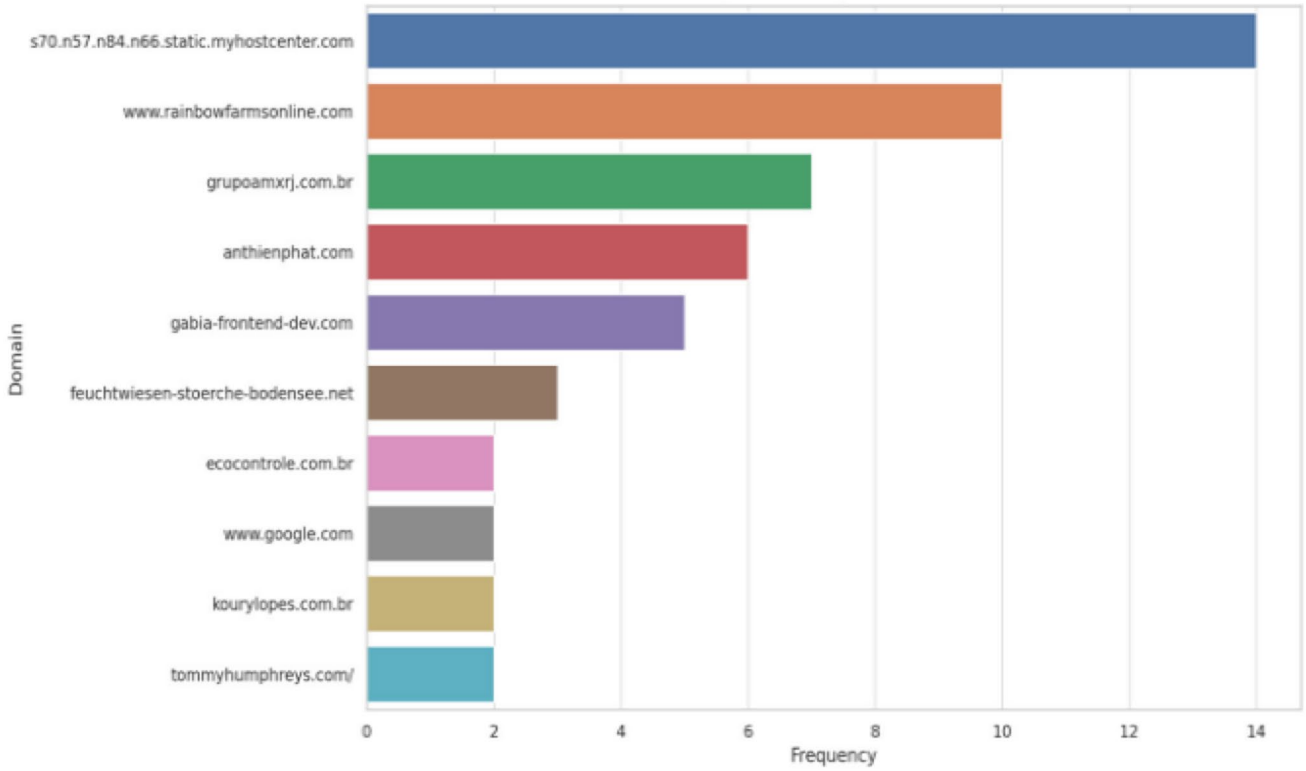

Fig. 2 Top 10 most frequent domain

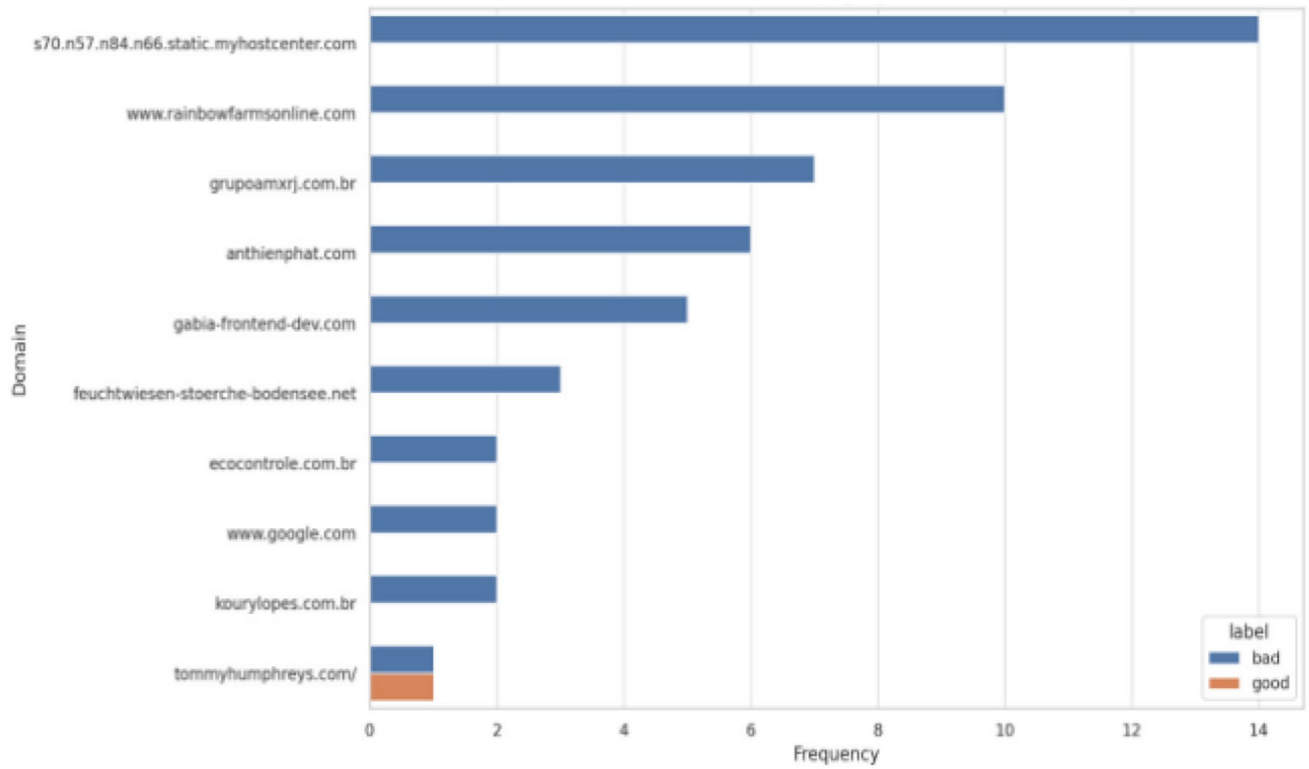

Fig. 3 Label distribution by top 10 domains

The limits of sequence length in DL models for malicious website identification are discussed by Sun et al. in 2022

Table 1 Summary of the related work

Authors	Focus	Method(s)	Key Findings/Contributions
Kim	Identifying dangerous websites	LoGos analyzer in Windows VM	10-18x faster than previous systems, extensive testing showing effectiveness in identifying dangerous websites
Sun et al.	Malicious website identification with DL models	Adaptive segmented text model, bi-directional LSTM, multi-head self-attention	Increased detection accuracy
Wan Manan et al.	JavaScript in web-based cyberattacks	Analysis of features and methods	Decreased false alarms in malicious web page identification
Yang et al.	Detecting malicious online traffic	Semantic-aware strategy	High recall and precision in identifying rogue IP addresses
Ghosh et al.	Water quality assessment with ML models	Multiple classifiers including Random Forest and SVM	Random Forest model showed highest accuracy (78.96%)
Singhal et al.	Identifying fraudulent websites	Deep Neural Networks, Random Forests, Gradient Boosted Decision Trees	Addressed concept drift problem, improving fraud detection
Grini et al.	Recognizing hostile web robots	ML algorithms including SVM, decision tree C4.5	Focus on catching spambots and harvesters
Yan et al.	Security issues with IoT and rogue websites	Unsupervised learning for URL embedding	Emphasized feature engineering for successful domain embedding
Hou et al.	Identifying malicious web content	Machine learning	Successful approach despite code obfuscations
Deng et al.	Identifying dangerous online sites	Feature optimisation and hybrid classification with ML	Improved effectiveness and accuracy of detection
Zabihimayvan et al.	Identifying web robots	SMART soft computing system	Better performance in identifying web robots
Li et al.	Technological and communications security	Analysis of strategies and techniques	Focus on “knowing your enemy” to strengthen security measures
Chang et al.	Novel malware identification on websites	Convolutional neural network (CAMD)	Over 98% accuracy rate
Yong et al.	IoT security and parameter injection attacks	Detection system based on Hidden Markov Models (HMM)	Outperformed baseline methods
Cohen et al.	Characteristics of harmful webmail attachments	Analysis of antivirus telemetry information	Identified novel features associated with malware spread patterns
McGahagan et al.	Identifying harmful websites	Collection and analysis of over 46,000 features, various feature selection methods	Found features achieve equivalent detection performance with 66% fewer features

[3]. They suggest an adaptive segmented text model that uses bi-directional LSTM and multi-head self-attention to increase detection accuracy. The focus of Wan Manan et al. 2020 [4] is on JavaScript’s expanding importance in web-based cyberattacks. They undertake a thorough analysis of the features and methods used to identify malicious JavaScript code. This study tries to decrease false alarms brought on by innocuous code that has been obfuscated in order to increase the effectiveness of malicious web page identification. A semantic-aware strategy to detecting malicious online traffic is suggested by Yang et al. (n.d.) [5]. Their approach uses a semantic model to better understand harmful behaviours while profiling specific online visitors. The study confirms the methodology using a sizable dataset and claims high recall and precision rates, accurately identifying thousands of distinct rogue IP addresses.Ghosh et al. (2023) [6] focus on using ml models for water quality assessment. They test multiple classifiers, including Random Forest and SVM, on a dataset of 3277 samples collected over nine years. The Random Forest model yielded the highest accuracy at 78.96%. The study emphasizes the effectiveness of ML in predicting water quality, crucial for human health. Based on URL attributes, Singhal et al. (2020) [7] suggest using ML to identify fraudulent websites. They assess the effectiveness of Deep Neural Networks, Random Forests, and Gradient Boosted Decision Trees. The work offers a paradigm to combat such strategies. It also addresses the problem of “concept drift,” in which attackers alter aspects to avoid detection.Lino is a smart system introduced by Grini et al. (n.d.) [8] that can recognise hostile web robots

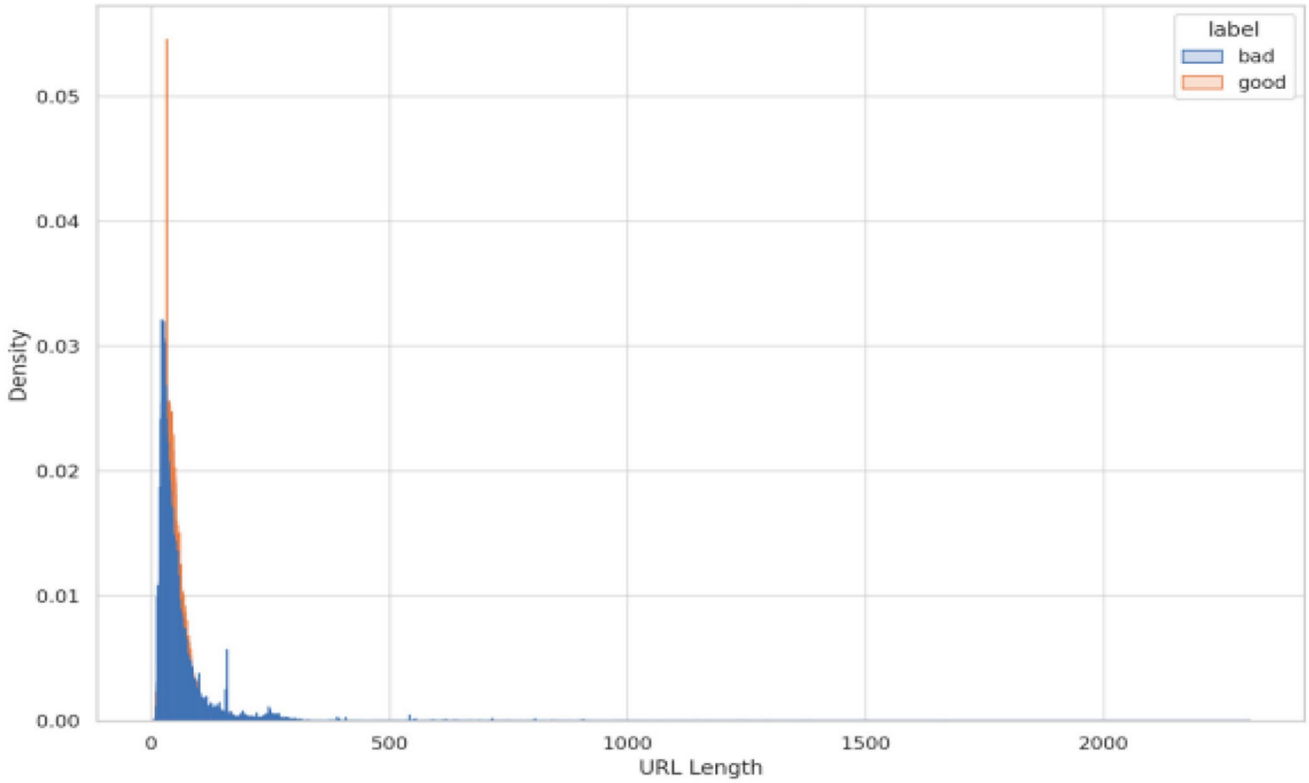

Fig. 4 URL length distribution by LaBEL

like spambots and harvesters. Lino employsML algorithms like SVM and decision tree C4.5 for categorization and replicates a vulnerable webpage to catch these bots. The purpose of the article is to discuss how web crawlers are being abused for things like automated online bidding and email scraping.The security issues with the Internet of Things (IoT) brought on by rogue websites are discussed by Yan et al. (2020) [9]. To improve ML models for identifying such sites, they suggest an unsupervised learning approach for URL embedding. The significance of feature engineering is emphasised, and the research investigates important factors for successful domain embedding.Hou et al. (2010) [10] discuss the difficulties in identifying malicious web content, particularly dynamic HTML that is easily obfuscated. They suggest using machine learning to identify such malicious web pages. The study describes fundamental characteristics that are crucial for ML and demonstrates that their approach is successful even when code obfuscations are present.Deng et al. (2022) [11] concentrate on improving feature optimisation and hybrid classification for the identification of dangerous online sites. The study integrates various machine learning techniques, introduces new features for harmful web sites, and uses information gain for feature selection. The effectiveness and accuracy of detection have significantly improved according to experimental findings. SMART is a soft computing system introduced by Zabihimayvan et al. (2017) [12] that can identify both good and bad web robots from server logs. The system makes adjustments for each web server's particular session characteristics. According to experimental findings, SMART performs better than current techniques in identifying both types of web robots.Identifying the enemy in the setting of technological and communications security is a topic Li et al. (2012) [13] explore in depth. The study, which is a component of the 2012 ACM Conference proceedings, intends to shed light on the strategies and techniques employed by online attackers. To strengthen security measures, "knowing your enemy" is the main focus. The limits of current technologies in identifying novel types of malware on websites are discussed by Chang et al. (2023) [14]. They provide a convolutional neural network-based technique called Content-Aware Malicious Webpage Detection (CAMD). The technique, which analyses webpage codes as grayscale images, has an accuracy rate of above 98%.The security issues in Internet of Things (IoT) contexts, notably the susceptibility to parameter injection attacks, are the main concern of Yong et al. (2019) [15]. To counteract these assaults, they present a detection system based on Hidden Markov Models (HMM). The approach performs better than baseline methods when tested using actual IoT web traffic data.Based on a thorough analysis of antivirus telemetry information, Cohen et al. (2018) [16] explore the special characteristics

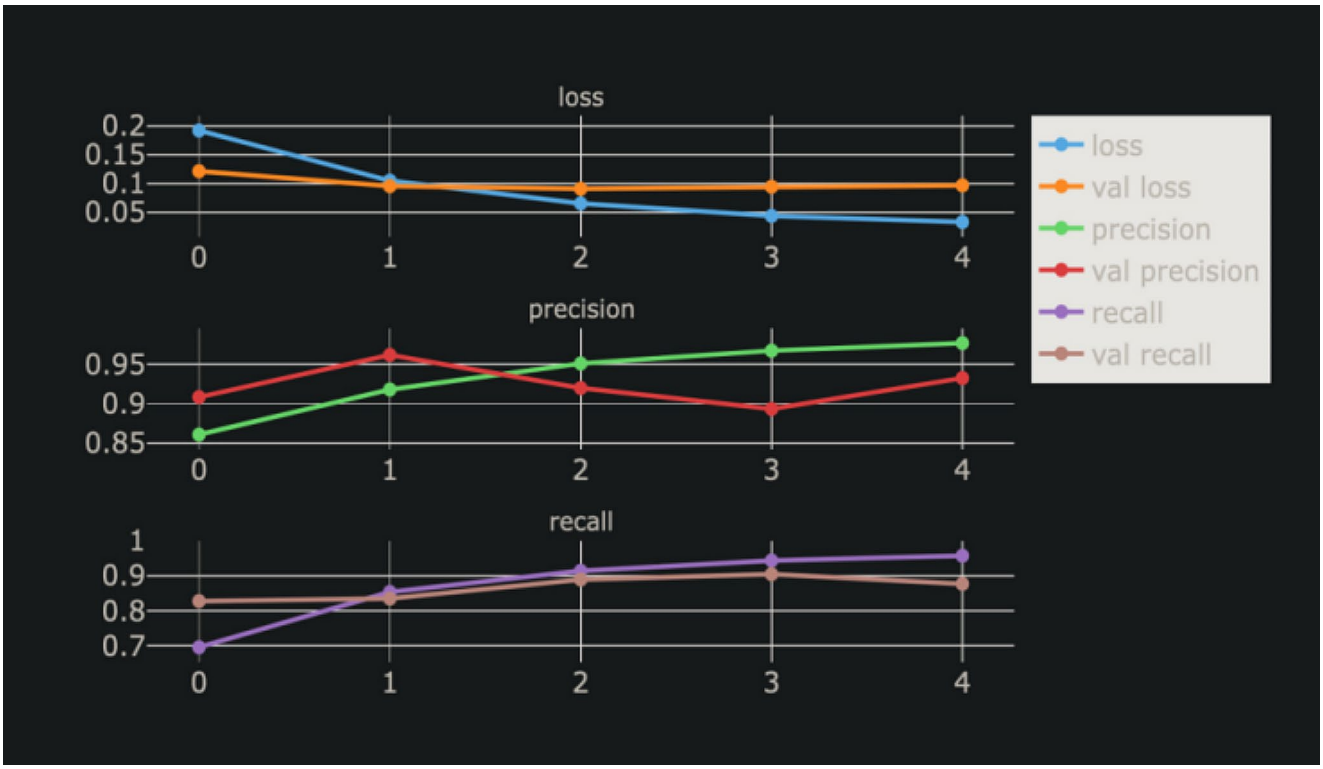

Fig. 5 Performance of CNN Model 1 for Malicious Website Detection

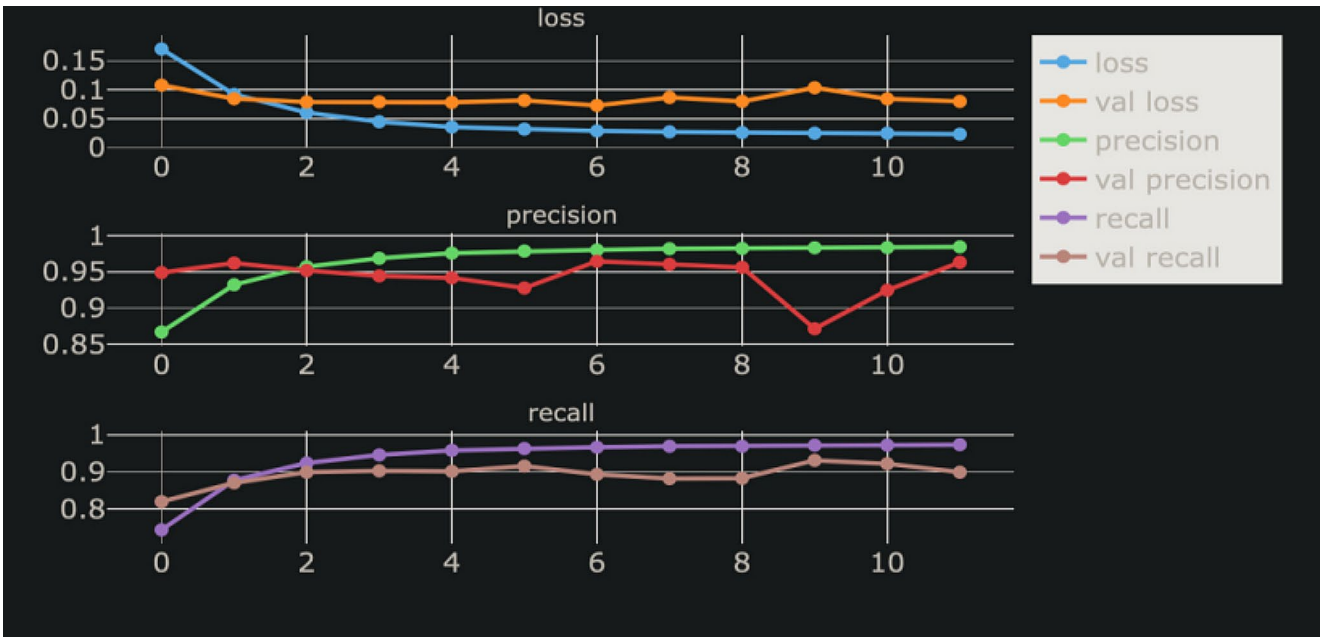

Fig. 6 Performance of CNN Model 2 for malicious website detection

of harmful webmail attachments. They develop a detector for malicious webmail attachments and find novel features associated with malware spread patterns. The study emphasises how unique these attachments are in respect of their range, variety, and modes of transmission.

Rather of employing pre-selected features, McGahagan et al. (2021) [17] investigate the possibility of developing novel features for identifying harmful websites. They carry out a thorough review, collecting over 46,000 features and using different feature selection methods. According to their

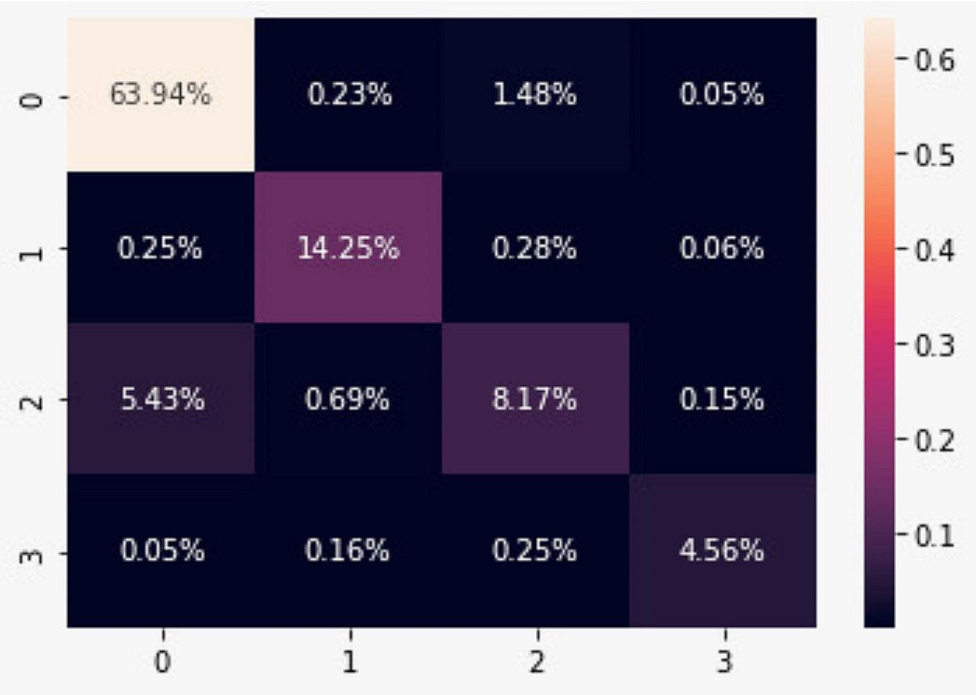

Fig. 8 Confusion matrix for decision tree classifier

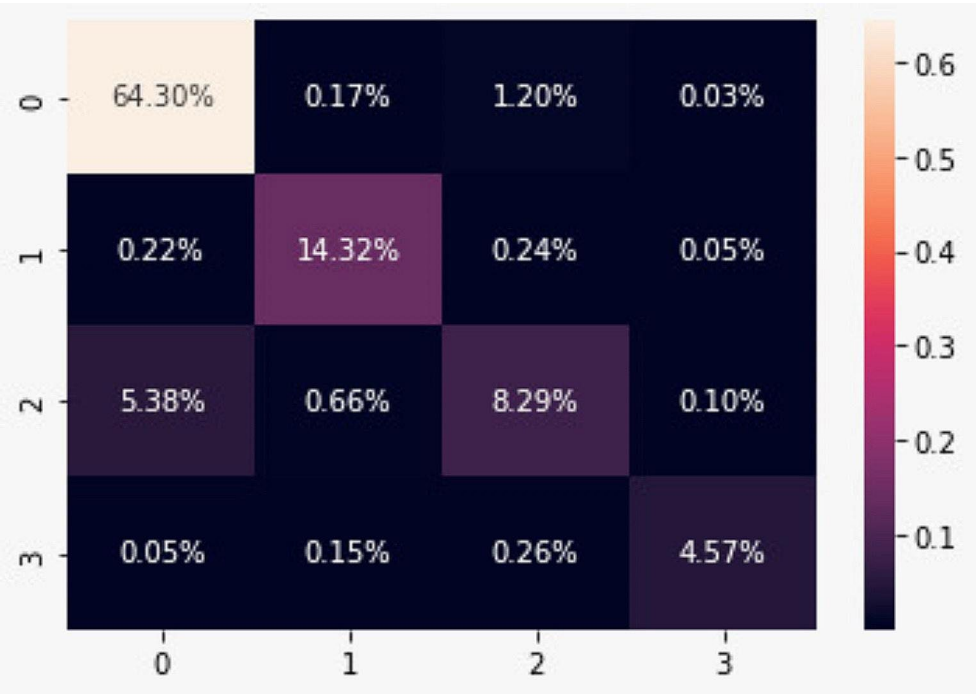

Fig. 9 Confusion matrix for decision random forest classifier

research, found features can achieve detection performance that is equivalent to that of conventional methods while using 66% less features and up to a Matthews Correlation Coefficient of 0.9008.WebMon is ML and YARA signature-based malicious webpage detector introduced by Kim et al. (2018) [18]. In order to find hidden exploit codes and determine how harmful a website is, the system tracks related URLs. With 250 containers running at once, WebMon is 7.6 times faster than earlier models and achieves a 98% detection rate. The work is distinctive in that it concentrates on finding harmful pathways inside a domain. Deep Learning-Assisted [19] Flair Segmentation (Rahat et al., 2023): This study shows the promise of artificial intelligence in medical imaging by using deep learning for brain MR image segmentation to examine lower-grade gliomas.Identification of Potato Leaf Diseases (Ghosh et al., 2023): CNN models

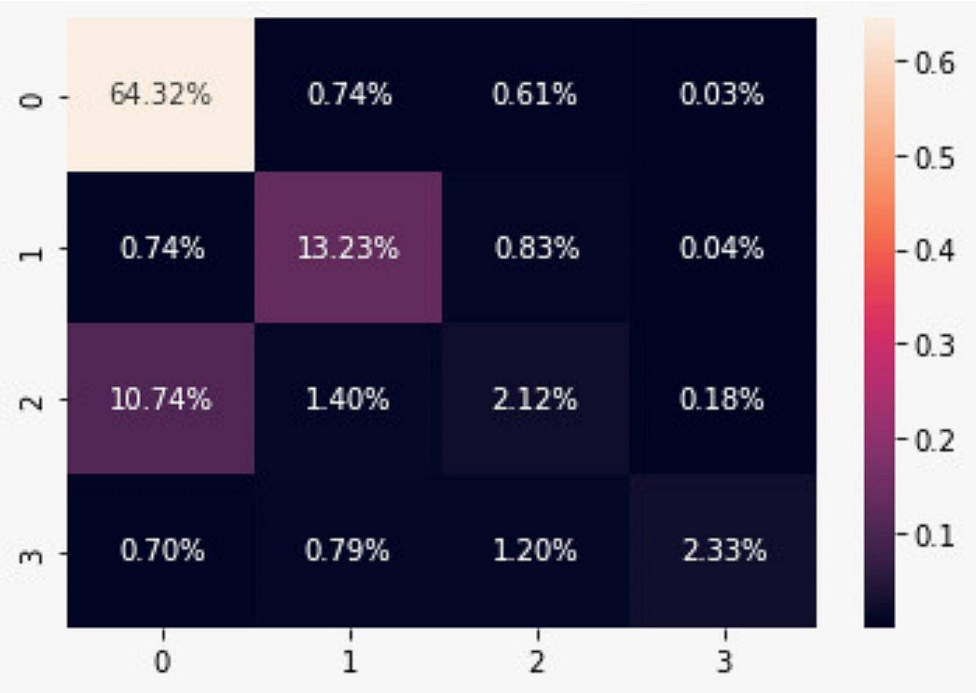

Fig. 10 Confusion matrix for decision adaboost classifier

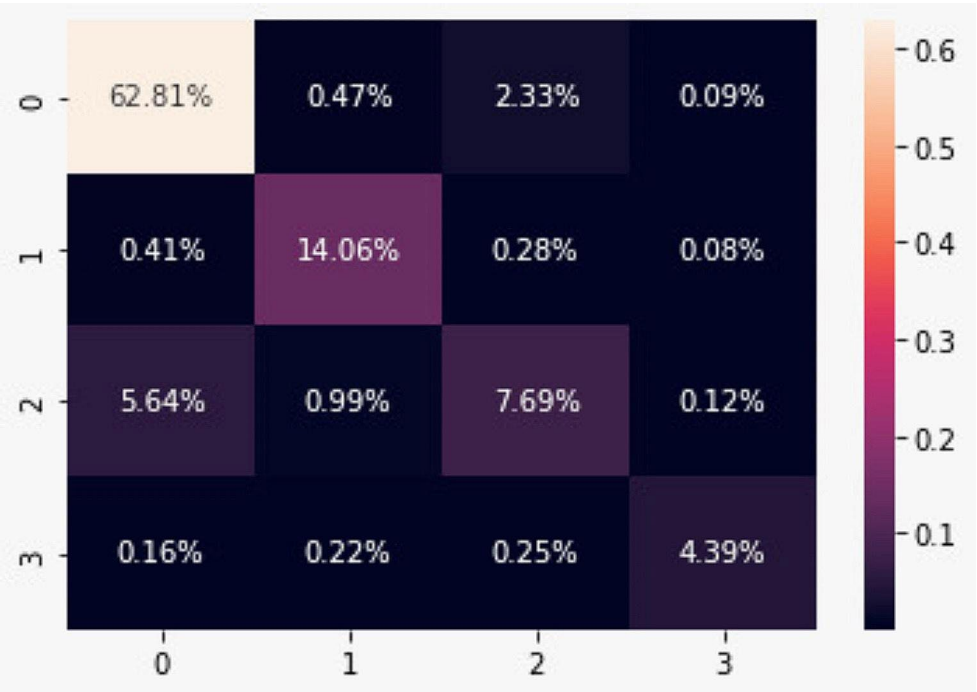

Fig. 11 Confusion matrix for KNN

were developed [20] to diagnose illnesses of the potato leaf, with VGG19 demonstrating the best performance in disease diagnosis. Forecasting Cardiovascular Disease (Mandava et al., 2023): A machine and [21, 22] deep learning model was developed with a 96.7% accuracy rate for the prediction of CVD in Bangladesh. Wheat Infection with Yellow Rust (Mandava et al., 2023): Researched [23] deep learning for wheat yellow rust detection, highlighting the usefulness of EfficientNetB3 in disease control.Khasim et al. (2023) present Real-Time Diagnostics of Rice-Leaf Diseases: introduced [24] a machine learning method with over 97% accuracy that greatly improved disease management in Bangladesh by recognising rice leaf illnesses.Khasim et al. (2023) examined the role of machine and deep learning in microbe identification, highlighting developments in automated classification, in their study of "Intelligent

Fig. 12 Confusion matrix for SGD

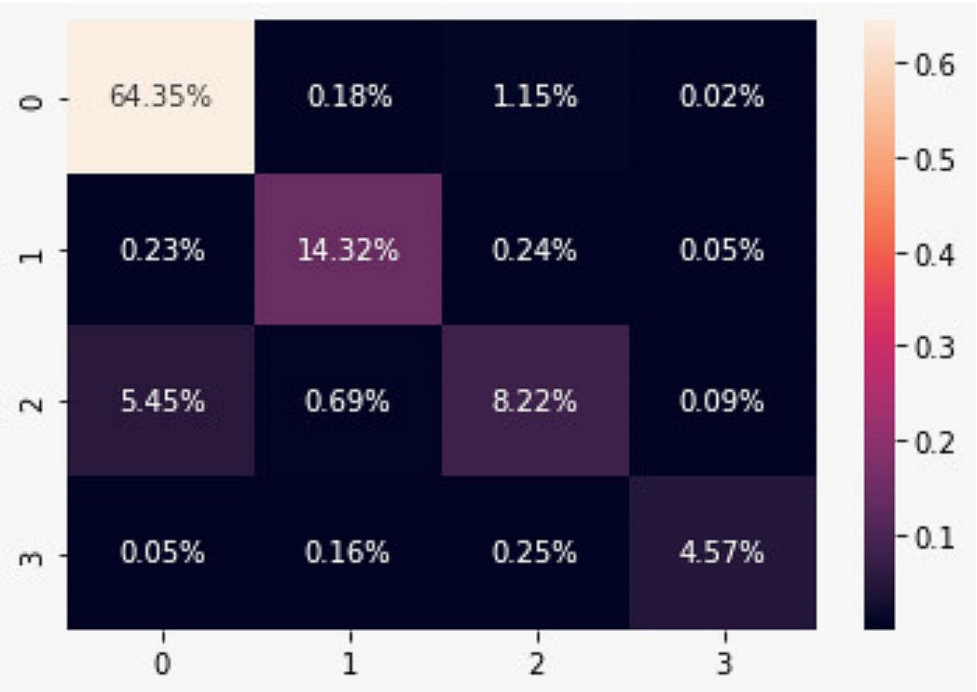

Fig. 13 Confusion matrix for extra trees classifier

Image Recognition [25] of Microorganisms."Classification of maize Leaf Diseases in Bangladesh [26] (Mohanty et al., 2023): A hybrid model with 99.65% accuracy was used to classify maize leaf diseases using deep learning algorithms. Methods for Detecting Skin Cancer (Ghosh et al., 2024): suggested [27] a hybrid model that improves categorization with high accuracy through deep learning for the diagnosis of skin cancer.Classification of Cauliflower Disease (Pradhan et al.): centered [28] on using deep learning to detect cauliflower illness; EfficientNetB3 led[Table 1.] the field with accuracy at 98%, improving the sustainability of agriculture.

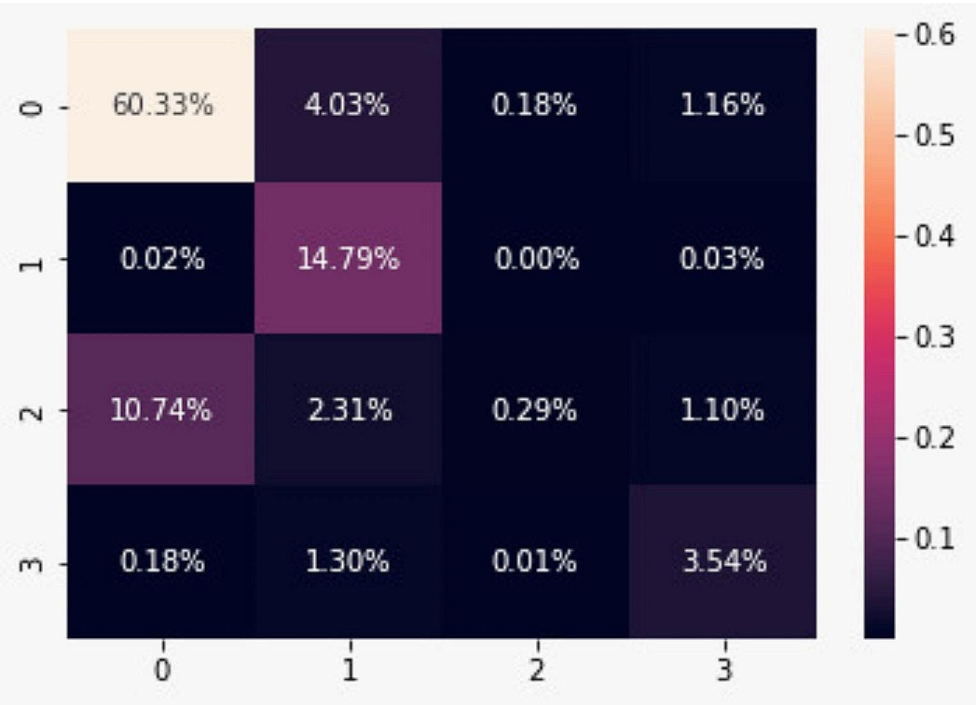

Fig. 14 Confusion matrix for gaussian naive bayes

Methodology

The research's methodology was created to offer a thorough and in-depth comparison of ML and DL techniques for the identification of dangerous websites. Data preprocessing, feature selection, model training, evaluation, and interpretation are all included in the study's methodical methodology.

Dataset Overview

In the realm of cybersecurity, the quality and comprehensiveness of the dataset employed play a pivotal role in the efficacy of the research. For this study, we utilized a robust dataset containing over 420,000 web URLs, meticulously categorized into various features such as 'domain,' 'subdomain,' 'domain_suffix,' and 'label.' The dataset is devoid of missing cells, thereby eliminating the need for imputation techniques, which can often introduce bias or inaccuracies. However, it is worth noting that the dataset contained approximately 2.2% duplicate rows, which were subsequently removed to ensure the integrity of the research.The 'label' feature serves as the target variable and is categorical in nature, with two distinct classes: 'good' and 'bad.' The 'good' class consists of 344,821 instances, while the 'bad' class comprises 75,643 instances, highlighting a class imbalance that was addressed through oversampling techniques during the data preprocessing stage. The features 'domain,' 'subdomain,' and 'domain_suffix' are also categorical but exhibit high cardinality, with 114,880, 27,733, and 701 unique values, respectively. This high cardinality was managed through feature engineering techniques, including one-hot encoding and dimensionality reduction, to ensure that the machine learning models could effectively learn from the data.The dataset is stored in a memory-efficient format, occupying just 16.0 MiB in memory, with an average record size of 40.0 bytes. This efficiency is crucial for scalability, especially when deploying the trained models in real-time web security applications where computational resources may be limited.In summary, the dataset employed in this research is both comprehensive and challenging, with its high cardinality and class imbalance providing a rigorous test bed for evaluating the performance of various machine learning and deep learning algorithms. The preprocessing techniques applied have been carefully chosen to preserve the integrity of the data while making it amenable to complex computational analysis. This dataset serves as the backbone of our research, enabling a nuanced and detailed comparative study of ML and DL techniques in the field of malicious web detection.

Data Pre-Processing

The initial dataset, comprising over 420,000 web URLs, presented several challenges that required meticulous preprocessing to ensure the integrity and reliability of the subsequent analysis. The preprocessing stage was divided into several key steps, each designed to address specific issues in the data.

- **Handling Duplicate Rows**: Initially, 2.2% of the rows in the dataset were duplicated. To avoid skewing or introducing bias into the machine learning models, these duplicates were eliminated. In order to ensure that only distinct instances were kept for study, the removal was carried out using a de-duplication technique.
- **Managing High Cardinality**: A number of features, including 'domain,''subdomain,' and 'domain_suffix,' showed high cardinality. We used one-hot encoding to control it, then dimensionality reduction with Principal Component Analysis (PCA). This change improves the models' capacity to detect underlying patterns in the data while simultaneously reducing computational complexity.
- **Feature Engineering**: Numerous features in the dataset, some of which were not specifically pertinent to the classification task, were present. After carefully examining the exploratory data, we chose to remove some features like "url," "type," and "Category." To guarantee that every factor affects the model's performance equally, the remaining attributes were then scaled and normalised.
- **Data Splitting**: Finally, stratified sampling was used to divide the dataset into training and testing sets. This made the distribution of the "good" and "bad" classes in both sets identical, increasing the reliability and generalizability of the evaluation measures.
- **Interpretation and Analysis**: The last step required a thorough evaluation of the findings with an emphasis on the relative advantages and disadvantages of ML and DL methods for harmful site detection. The models' computational effectiveness, scalability, and adaptability received particular focus.

Data Analysis

In[Fig. 1] the bar chart shows the distribution of the labels "good" and "bad" in the dataset. It's evident that the dataset is imbalanced, with a much larger number of "good" labels compared to "bad" labels.In[Fig. 2] the bar chart above shows the top 10 most frequent domains in the dataset along with their frequencies.The domain s70.n57.n84.n66.static.myhostcenter.com appears 14 times, making it the most frequent domain.Following that, www.rainbowfarmsonline.com appears 10 times.Other domains like grupoamxrj.com.br, anthienphat.com, and gabia-frontend-dev.com also appear multiple times but less frequently.In[Fig. 3] the bar chart shows the distribution of "good" and "bad" labels for the top 10 most frequent domains in the dataset. For most domains, there's either only the "bad" label or no label distribution available. This could be due to the fact that the "good" labels are spread across many different domains, diluting their presence in the top 10 list.The domain www.google.com appears twice and is labeled as "good".In [Fig. 4] the histogram shows the distribution of URL lengths for both "good" and "bad" labels.For "good" URLs, the length primarily falls within the range of around 0–40 characters.For "bad" URLs, the length varies more widely and tends to be longer on average.This suggests that URL length could be a feature to consider when classifying URLs as "good" or "bad.

Performance of all Models

In our quest to identify the most effective techniques for malicious web detection, we trained and evaluated a variety of ML and DL models, each assessed based on multiple performance metrics such as accuracy, precision, recall, and F1-score.

- **CNN MODEL 1**: Taking the CNN Model 1 as an example, it demonstrated a high degree of accuracy of 97%, precision of 97.71%, and recall of 95.77%. The model's accuracy and balance of false positives and false negatives are both indicated by the F1-score, a harmonic mean of precision and recall, which was 98%.[Fig. 5].

- **CNN MODEL 2**: Astonishingly, CNN Model 2 outperformed the previous version with a 98% accuracy rate, a 98.52% precision rate, and a 97.48% recall rate. In our investigation, the F1-score attained 99%, making it the most trustworthy model for actual applications.[Fig. 6].

- **Decision Tree Classifier**: The Decision Tree Classifier, on the other hand, managed to attain an accuracy of 90.93%, a precision of 92%, and a recall of 97%. Despite the fact that these figures are impressive, they are inferior than the CNN models, particularly in terms of overall accuracy and F1-score, a score of 94%.[Fig. 7].

- **Random Forest Classifier**: The Random Forest Classifier showed a slight edge over the Decision Tree model, achieving an accuracy of 91.49%, a precision of 92%, and a recall of 98%. Its F1-score was 95%, indicating its higher sensitivity to false negatives[Fig. 8].

- **AdaBoost Classifier**:The AdaBoost Classifier, although sensitive with a high recall of 98%, lagged in accuracy, achieving only 82.01%. Its F1-score stood at 90%, making it less suitable for this specific application[Fig. 9].

- **KNN**: The K-Nearest Neighbors (KNN) model, with an accuracy of 88.96%, a precision of 89%, and a recall of 95%, performed well but was not as effective as other models in terms of overall accuracy and precision. Its F1-score was 92%[Fig. 10].

- **SGD**: The Stochastic Gradient Descent (SGD) model was the least effective, with an accuracy of just 80.49% and an F1-score of 89%. It is not recommended for this particular problem set[Fig. 11].

- **Extra Trees Classifier**: The Extra Trees Classifier closely mirrored the performance of the Random Forest model but with a slightly better accuracy of 91.46%. Its F1-score was 95%[Fig. 12].

- **Gaussian Naive Bayes**: The Gaussian Naive Bayes model had the lowest accuracy of 78.95% and is not recommended for this application[Fig. 13].

After a thorough evaluation, it is evident that CNN Model 2 stands out as the most effective model for detecting malicious websites, excelling in all performance metrics. Its high accuracy, precision, and F1-score make it the most reliable and robust model for this study .

Result and Discussion

In order to address the issue of malicious web identification, a variety of ML and DL algorithms were used in this study. The Convolutional Neural Network (CNN) Model 2 was found to be the most successful, with an F1-score of 99% and accuracy, precision, and 98.52%. This model surpassed CNN Model 1, which had a 97% accuracy and 98% F1-score and also displayed impressive outcomes. Although they performed admirably, conventional machine learning models like Decision Trees and Random Forests were outperformed by the CNN models, which had accuracy levels of 90.93% and 91.49%, respectively. Despite having a high recall of 98%, the AdaBoost Classifier has a low overall accuracy of 82.01%.The Stochastic Gradient Descent (SGD) model and the K-Nearest Neighbours (KNN) model also performed poorly overall, with accuracy levels of 88.96% and 80.49%, respectively. The Random Forest model was closely mirrored by the Extra Trees Classifier, but it had a slightly higher accuracy of 91.46%. The Gaussian Naive Bayes model had the lowest accuracy, coming in at only 78.95%. Following a thorough analysis, CNN Model 2 emerges as the most trustworthy and durable model for identifying fraudulent websites, outperforming all performance measures and creating a new standard for this particular application.

Conclusion and Future Work

In order to detect dangerous websites, this research set out to give a thorough comparison analysis of multiple ML and DL models. The Convolutional Neural Network (CNN) Model 2 emerged as the most successful model among those tested, with an F1-score of 99% and accuracy, precision, and 98.52%. This model offers a solid and trustworthy answer to the issue at hand while also setting a new standard in the industry. Although classic machine learning models like Decision Trees and Random Forests had excellent performance, the deep learning models, especially the CNNs, outperformed them.The study also identified potential areas for performance enhancement for other models, including AdaBoost and Gaussian Naive Bayes.Looking ahead, future effort may concentrate on a number of directions. To further improve performance, it may be investigated to incorporate more complex DL architectures as Recurrent Neural Networks (RNNs) or Transformers. Second, feature engineering methods should be improved to extract from URLs more useful attributes. Last but not least, the study might be expanded to incorporate real-time detection capabilities, making it more relevant to the present cybersecurity issues.

Author Contribution Both the authors having equal contribution.

Funding Information This research having no such funding.

Data Availability Data Available as per demand by the reviewers.

Declarations

Research Involving Human and /or Animals This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent Both the authors well know about the submission.

Conflict of Interest Both the authors having no Conflict of Interest.

References

1. Hosseini N, Fakhar F, Kiani B, Eslami S. Enhancing the security of patients' portals and websites by detecting malicious web crawlers using machine learning techniques. Int J Med Inf (Shannon Ireland). 2019;132:103976–103976. https://doi.org/10.1016/j.ijmedinf.2019.103976.
2. Kim S, Kim S, Kim D. LoGos: internet-explorer-based malicious webpage detection. ETRI J. 2017;39(3):406–16. https://doi.org/10.4218/etrij.17.0116.0810.

3. Sun G, Zhang Z, Cheng Y, Chai T. Adaptive segmented webpage text based malicious website detection. Comput Networks (Amsterdam Netherlands: 1999). 2022;216:109236. https://doi.org/10.1016/j.comnet.2022.109236.
4. Manan NW, Nizam Mohmad Kahar W, M., Mohd Ali N. (2020). A Survey on Current Malicious JavaScript Behavior of infected Web Content in Detection of Malicious Web pages. IOP Conference Series. Materials Science and Engineering, 769(1), 12074. https://doi.org/10.1088/1757-899X/769/1/012074.
5. Yang J, Wang L, Xu Z. A Novel Semantic-Aware Approach for detecting malicious web traffic. Inform Commun Secur. n.d.;633–45. https://doi.org/10.1007/978-3-319-89500-0_54.
6. Ghosh H, Tusher MA, Rahat IS, Khasim S, Mohanty SN. Water Quality Assessment through Predictive Machine Learning. Intelligent Computing and networking. IC-ICN 2023. Lecture notes in networks and systems. Volume 699. Singapore: Springer; 2023. https://doi.org/10.1007/978-981-99-3177-4_6.
7. Singhal S, Chawla U, Shorey R. (2020). Machine Learning & Concept Drift based Approach for Malicious Website Detection. 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS), 582–585. https://doi.org/10.1109/COMSNETS48256.2020.9027485.
8. Gržinić T, Mršić L, Šaban J. Lino - An Intelligent System for detecting malicious web-Robots. Intell Inform Database Syst. n.d.;559–68. https://doi.org/10.1007/978-3-319-15705-4_54.
9. Yan X, Xu Y, Cui B, Zhang S, Guo T, Li C. Learning URL embedding for malicious website detection. IEEE Trans Industr Inf. 2020;16(10):6673–81. https://doi.org/10.1109/TII.2020.2977886.
10. Hou Y-T, Chang Y, Chen T, Laih C-S, Chen C-M. Malicious web content detection by machine learning. Expert Syst Appl. 2010;37(1):55–60. https://doi.org/10.1016/j.eswa.2009.05.023.
11. Deng W, Peng Y, Yang F, Song J. Feature optimization and hybrid classification for malicious web page detection. Concurrency Comput. 2022;34(16). https://doi.org/10.1002/cpe.5859.n/a-n/a.
12. Zabihimayvan M, Sadeghi R, Rude HN, Doran D. A soft computing approach for benign and malicious web robot detection. Expert Syst Appl. 2017;87:129–40. https://doi.org/10.1016/j.eswa.2017.06.004.
13. Li Z, Zhang K, Xie Y, Yu F, Wang X. Knowing your enemy. Proc 2012 ACM Conf Comput Commun Secur. 2012;674–686. https://doi.org/10.1145/2382196.2382267.
14. Chang Y-J, Tsai K-L, Jiang W-C, Liu M-K. Content-aware malicious webpage detection using convolutional neural network. Multimedia Tools Appl. 2023. https://doi.org/10.1007/s11042-023-15559-8.
15. Yong B, Liu X, Yu Q, Huang L, Zhou Q. Malicious web traffic detection for internet of things environments. Comput Electr Eng. 2019;77:260–72. https://doi.org/10.1016/j.compeleceng.2019.06.008.
16. Cohen Y, Hendler D, Rubin A. Detection of malicious webmail attachments based on propagation patterns. Knowl Based Syst. 2018;141:67–79. https://doi.org/10.1016/j.knosys.2017.11.011.
17. McGahagan J, Bhansali D, Pinto-Coelho C, Cukier M. Discovering features for detecting malicious websites: an empirical study. Computers Secur. 2021;109:102374. https://doi.org/10.1016/j.cose.2021.102374.
18. Kim S, Kim J, Nam S, Kim D. WebMon: ML- and YARA-based malicious webpage detection. Comput Networks (Amsterdam Netherlands: 1999). 2018;137:119–31. https://doi.org/10.1016/j.comnet.2018.03.006.
19. Rahat IS, Ghosh H, Shaik K, Khasim S, Rajaram G. Unraveling the Heterogeneity of Lower-Grade Gliomas: Deep Learning-Assisted Flair Segmentation and Genomic Analysis of Brain MR Images. EAI Endorsed Trans Perv Health Tech [Internet]. 2023 Sep. 29 [cited 2023 Oct. 2];https://doi.org/10.4108/eetpht.9.4016.
20. Ghosh H, Rahat IS, Shaik K, Khasim S, Yesubabu M. Potato Leaf Disease Recognition and Prediction using Convolutional neural networks. EAI Endorsed Scal Inf Syst [Internet]. 2023 Sep. 21 https://doi.org/10.4108/eetsis.3937.
21. Rout P, Mohanty SN, A Hybrid Approach for Network Intrusion Detection, 2015 Fifth International Conference on Communication Systems and, Technologies N. Gwalior, India, 2015, pp. 614–617, https://doi.org/10.1109/CSNT.2015.76.
22. Mandava M, Vinta SR, Ghosh H, Rahat IS. An All-Inclusive Machine Learning and Deep Learning Method for forecasting Cardiovascular Disease in Bangladeshi Population. EAI Endorsed Trans Perv Health Tech. Oct. 2023;9. https://doi.org/10.4108/eetpht.9.4052.
23. Mandava M, Vinta SR, Ghosh H, Rahat IS. Identification and categorization of yellow rust infection in wheat through deep learning techniques. EAI Endorsed Trans IoT. 2023;10. https://doi.org/10.4108/eetiot.4603.
24. Khasim IS, Rahat H, Ghosh K, Shaik, Panda SK. Using Deep Learning and Machine Learning: Real-Time Discernment and Diagnostics of Rice-Leaf diseases in Bangladesh. EAI Endorsed Trans IoT. Dec. 2023;10. https://doi.org/10.4108/eetiot.4579.
25. Khasim H, Ghosh IS, Rahat K, Shaik, Yesubabu M. Deciphering Microorganisms through Intelligent Image Recognition: Machine Learning and Deep Learning Approaches, Challenges, and Advancements, EAI Endorsed Trans IoT, vol. 10, Nov. 2023https://doi.org/10.4108/eetiot.4484.
26. Mohanty SN, Ghosh H, Rahat IS, Reddy CVR. Advanced Deep Learning Models for Corn Leaf Disease Classification: A Field Study in Bangladesh. Eng. Proc. 2023, 59, 69. https://doi.org/10.3390/engproc2023059069.
27. Ghosh H, Rahat IS, Mohanty SN, Ravindra JVR, Sobur A. (2024). A study on the application of machine learning and deep learning techniques for Skin Cancer Detection. https://doi.org/10.5281/zenodo.10525954.
28. Pradhan R, Ghosh H, Rahat IS, Naga JV, Ramesh, Yesubabu M. Enhancing Agricultural sustainability with deep learning: a Case Study of Cauliflower Disease Classification, https://doi.org/10.4108/eetiot.4834.