Classifying Cybersecurity Threats in URLs Using Decision Tree and Naive Bayes Algorithms: A Data Mining Approach for Phishing, Defacement, and Benign Threat Detection

Deshinta Arrova Dewi[1,*], Tri Basuki Kurniawan

[1]Faculty of Science and Technology, Universitas Bina Darma, Indonesia

[2]Faculty of Data Science and Information Technology, INTI International University, Malaysia

ABSTRACT

This research focuses on the application of data mining techniques to classify URLs into multiple cybersecurity threat categories, including phishing, defacement, and benign URLs. Accurate classification of URLs is crucial in the current digital landscape, where cyber threats are increasing in both frequency and complexity. This study employs two popular machine learning algorithms, Decision Tree and Multinomial Naive Bayes, to analyze and classify URL data based on their textual content. The URLs were transformed using Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, allowing the models to learn distinctive patterns within the URL strings that signify different threat types. The dataset used comprises 24,800 labeled URLs, representing a realistic mix of common and rare cyber threat categories. Both models demonstrated strong classification performance, with the Decision Tree achieving an accuracy of 94.01% and Naive Bayes reaching 92.36%. While both classifiers performed well on the dominant categories such as phishing and benign URLs, challenges remained in accurately detecting less frequent classes due to class imbalance. The Decision Tree model showed a slightly better ability to handle these imbalances and provided interpretability through feature importance analysis, highlighting key URL tokens influencing classification decisions. Naive Bayes, although efficient and effective for the majority classes, exhibited lower recall for minority classes. The results indicate that machine learning models can effectively support automated threat detection systems by classifying URLs with high accuracy, thereby enhancing cybersecurity defenses. Future work may explore advanced modeling techniques, such as ensemble methods or deep learning, alongside improved feature engineering and data augmentation to address class imbalance and improve detection of rare threats. Additionally, incorporating multi-source data could further strengthen threat classification. Overall, this research contributes valuable insights into URL-based cyber threat classification using accessible and interpretable machine learning approaches, supporting the development of proactive and scalable cybersecurity solutions.

Keywords URL Classification, Phishing Detection, Decision Tree, Naive Bayes, Cybersecurity Threat Detection

Submitted 2 January 2025
Accepted 12 February 2025
Published 15 March 2025

*Corresponding author
Deshinta Arrova Dewi,
deshinta.ad@newinti.edu.my

Additional Information and Declarations can be found on page 187

DOI: 10.63913/jcl.v1i2.10

Introduction

In the digital age, cybersecurity has emerged as a pivotal concern for both individuals and organizations due to the extensive reliance on digital systems. Among the myriad of threats that proliferate in this landscape, URL threats, particularly phishing, website defacement, and benign-looking sites that harbor

How to cite this article: D.A.Dewi and T.B.Kurniawan, "Classifying Cybersecurity Threats in URLs Using Decision Tree and Naive Bayes Algorithms: A Data Mining Approach for Phishing, Defacement, and Benign Threat Detection," *J. Cyber. Law.*, vol. 1, no. 2, pp. 175-189, 2025.

malicious intent, pose significant risks that warrant thorough investigation and understanding. Phishing attacks have become one of the most prevalent forms of cyber threats. These attacks exploit human psychology, deceiving users into providing sensitive information by masquerading as credible organizations or communications. According to Andriu [1], phishing has evolved over the decades in sophistication, which challenges traditional detection methods and underscores the need for improved techniques utilizing artificial intelligence to bolster email security against these deceptive practices. The increasing complexity of phishing attacks is highlighted further by a report indicating that 1 in every 10 URLs could be considered malicious, emphasizing the urgency of these threats amidst a backdrop of rising incidents [2]. This statistic not only accentuates the vulnerability of users but also stresses the necessity for enhanced awareness and education around these threats.

Website defacement represents another critical cybersecurity concern. As outlined by Al-Quayed et al [3], the interconnectedness of today's digital infrastructure necessitates robust protective measures to safeguard against malicious attacks that could compromise website integrity and trust. Researchers like Chowdhury et al [4] have focused on identifying vulnerabilities using multifaceted approaches that leverage machine learning to detect and mitigate various web application threats, including defacement and other forms of website exploitation. Such advances in detection technologies are critical as they address not only phishing but also broader vulnerabilities that can lead to site defacement, highlighting the necessity of proactive cybersecurity strategies. Furthermore, the existence of benign sites that can harbor malicious exploits underscores the complexity of cybersecurity threats in the digital age. The increased sophistication of attackers in disguising harmful intentions under the guise of legitimacy makes it imperative for both individuals and organizations to cultivate a strong culture of cybersecurity awareness. As emphasized by Ghazali et al [5], fostering digital literacy, especially among non-technical users, is fundamental in preparing individuals to recognize and respond to potential cybersecurity threats effectively, thereby contributing to an overall heightened security posture within organizations. This proactive approach may serve to mitigate the risks posed by seemingly innocent sites that can ultimately endanger security.

Detecting malicious URLs presents a formidable challenge in the realm of cybersecurity, primarily due to the rapid evolution of threats and the complexity of URL structures. As cybercriminals continually adapt their tactics, distinguishing between benign and harmful URLs has become increasingly sophisticated. The dimensions of this challenge can be organized into categories regarding classification accuracy and features employed for detection. One significant aspect of the detection challenge stems from the imbalanced nature of datasets used in training classification algorithms. Butnaru et al [6] noted that the dataset in their study included a significant disparity between benign and phishing URLs, reflecting real-world web traffic dynamics wherein benign URLs overwhelmingly outnumber malicious ones. This imbalance can lead to classifiers that are biased towards identifying benign URLs correctly while underperforming in recognizing phishing attempts, thereby affecting overall detection efficacy. Furthermore, traditional reliance on methods such as blacklists and signature matching encounters limitations due to the dynamic nature of URL creation. Blacklisting struggles to keep up with the high turnover of malicious URLs, given that new threats can emerge daily, creating

gaps in detection capabilities [7].

Current methodologies for detecting malicious URLs can primarily be classified into two categories: behavior-based and rule-based detection systems. Jaiswal et al [8] emphasize the use of rules or markers to flag malicious URLs quickly, highlighting the evolving tactics of attackers that necessitate enhanced behavioral analysis for more comprehensive detection. This duality in detection approaches illustrates the need for classifiers to maintain flexibility in feature extraction processes. Moreover, the inherent complexities tied to multi-class classification scenarios, wherein distinguishing between various forms of malicious URLs introduces added layers of difficulty, are compounded by the need for thorough data gathering and feature engineering [9]. Recent advancements in machine learning techniques provide promising avenues for improving URL classification. Tiryaki et al [10] have demonstrated the effectiveness of artificial intelligence models in mitigating these challenges by implementing more nuanced detection strategies that can adapt to the changing threat landscape. Techniques involving deep learning and optimization, like those presented by Hilal et al [11], suggest that semantic and lexical characteristics of URLs can be leveraged for improved classification accuracy. Such models can potentially discern subtle patterns that characterize malicious URLs, thereby enhancing detection rates.

The primary objective of this research is to leverage data mining techniques to classify URLs into distinct categories based on their threat level, such as phishing, defacement, or benign. With the increasing prevalence of cyberattacks that exploit malicious URLs, accurate classification has become a critical step in enhancing cybersecurity measures. This research aims to improve the detection accuracy of these threats by analyzing URL patterns and characteristics through machine learning algorithms. Accurate classification not only helps in timely identification of potential security risks but also contributes to the development of automated systems capable of mitigating cyber threats before they cause significant harm to users and organizations. This study focuses specifically on two well-known data mining algorithms: Decision Tree and Naive Bayes. These algorithms are selected due to their proven efficiency and interpretability when dealing with textual data such as URLs, which are complex and unstructured in nature. The Decision Tree algorithm offers the advantage of clear decision-making rules that can be easily interpreted by cybersecurity professionals, while Naive Bayes is valued for its speed and effectiveness in probabilistic classification. By comparing the performance of these two algorithms, this research intends to identify the most suitable model for classifying URL threats, taking into account factors like accuracy, computational speed, and ease of implementation in real-world cybersecurity systems. In terms of scope, the study is confined to analyzing a labeled dataset of URLs, where each URL is categorized into predefined threat types. This limitation allows for a focused approach in training and evaluating the machine learning models, ensuring reliable and measurable outcomes. The analysis of URL data will consider features extracted through text vectorization methods such as TF-IDF, enabling the models to learn from the lexical and structural attributes of URLs. The ultimate goal is to develop a classification framework that can be integrated into early warning systems or security software, thereby strengthening the ability to detect and prevent cyber threats effectively. This research contributes to the broader field of cybersecurity by providing data-driven insights and practical solutions for URL threat classification using

accessible and interpretable machine learning techniques.

Literature Review

Overview of Cybersecurity Threats

Cybersecurity threats continue to evolve, representing significant challenges for organizations and individuals in the digital landscape. Among the most prevalent forms of these threats are phishing and website defacement, each with distinct characteristics and implications for security. Phishing is a form of cyberattack that uses deceptive tactics to manipulate users into divulging sensitive information, such as login credentials or financial data. These attacks often take the form of fraudulent emails or messages that appear legitimate, leading users to malicious websites designed to steal their data. Research indicates that falling prey to phishing can have devastating consequences, including significant financial losses and reputational damage to organizations [12]. The sophistication of phishing techniques has increased over the years, employing advanced social engineering tactics that enhance the likelihood of success, thus compounding the challenge of detection and prevention.

As outlined by Nifakos et al [13], human factors play a critical role in phishing susceptibility, as individuals may lack adequate training to recognize phishing attempts. This highlights the importance of implementing comprehensive user training and education programs that emphasize the identification of phishing schemes, thereby bolstering an organization's overall cybersecurity posture. Website defacement represents another serious threat, characterized by unauthorized alterations to web pages. Attackers may replace legitimate content with false information or propaganda, as seen in various instances where popular sites have been compromised [14]. There are two main types of defacement: substitutive defacement, where original content is replaced, and additive defacement, where malicious links are overlaid on the existing content. These attacks can severely damage an organization's reputation and reliability while potentially leading to financial losses and data breaches [15]. Given the rapid growth and evolution of phishing and website defacement, organizations must adopt dynamic, proactive cybersecurity measures. These include advanced detection systems utilizing artificial intelligence, continuous monitoring for suspicious activities, and thorough education programs to prepare employees against a range of cyber threats. By understanding the mechanisms and implications of these common threats, organizations can implement comprehensive strategies to minimize their vulnerability in the ever-changing landscape of cybersecurity.

Data Mining in Cybersecurity

Data mining has become an integral tool in the field of cybersecurity, particularly for classifying threats such as malicious URLs. Various studies have applied data mining methodologies to improve detection rates and enhance the understanding of cybersecurity threats. This review synthesizes existing literature on the application of data mining techniques for classifying such threats, focusing on the classification of URLs. Rehman et al [16] conducted a comprehensive analysis of a hands-on cybersecurity dataset using diverse data mining approaches. Their work explored the effectiveness of simple, temporal, and sequential association rules, demonstrating the potential of data mining to derive meaningful insights from practical cybersecurity exercises. This foundational understanding highlights data mining's applicability not only in

identifying threats but also in assessing training and behavior in cybersecurity contexts. Jenni and Shankar [17] provided a thorough review of various methods for phishing detection, emphasizing the essential role of data mining in identifying deceptive patterns. They identified key techniques such as clustering, decision trees, and machine learning strategies, which help in discovering and classifying phishing attacks. By employing sophisticated techniques such as artificial neural networks and Bayesian networks, the authors underscored how data mining enables the extraction of relevant features from datasets to enhance phishing detection capabilities.

Similarly, Taluja [18] presented an overview of data mining techniques relevant to cybersecurity, noting the increasing prevalence of machine learning approaches for detecting and preventing cyber-attacks. They elaborated on the algorithms employed, such as support vector machines and naive Bayes classifiers, emphasizing their significance in effectively managing the complexities of cybersecurity. This paper contributes to an understanding of the landscape of available methods and the need for continuous evaluation and adaptation in cybersecurity analytics. Li et al [19] discussed the application of the Apriori algorithm to mine association rules in the context of global cyberspace security issues. Their research highlighted how association rule mining can unveil hidden information and relationships within extensive datasets, providing a framework for understanding and mitigating security threats. This approach reinforces the notion that data mining can assist not only in real-time detection but also in predictive analytics.

Wu and Yang [20] focused on the use of machine learning algorithms in data mining for identifying network security hazards. They emphasized the importance of extracting latent data and constructing mining models to enhance detection methodologies. Their findings suggest that the integration of machine learning with data mining offers a robust mechanism for addressing the evolving nature of cyber threats, particularly in URL classification. Ullah and Babar [21] analyzed big data tools essential for cybersecurity analytics, proposing that leveraging these technologies provides significant advantages in gathering and analyzing large volumes of security-related data. Their work highlights how a well-integrated big data approach can enhance the effectiveness of data mining techniques in cybersecurity. This perspective underlines the importance of employing advanced analytics capabilities to improve threat detection and

Machine Learning Algorithms

In the context of cybersecurity, machine learning algorithms serve as crucial tools for detecting and classifying threats, particularly with respect to URL-based attacks. This discussion focuses on two widely-used algorithms in this domain: Decision Trees and Naive Bayes, highlighting their relevance and effectiveness in URL threat detection. Decision Trees are a popular model in the field of machine learning due to their simplicity and interpretability. They operate by recursively partitioning the dataset into subsets based on feature values, creating a tree-like structure where each node represents a decision point. This method is particularly effective for classification tasks because it can handle both categorical and numerical data effectively [22]. In the context of URL threat detection, Decision Trees can leverage features such as URL length, the presence of suspicious keywords, and domain age to classify URLs as either benign or malicious. The model's intuitive nature allows cybersecurity professionals to easily interpret the logic behind its classifications. Furthermore,

Decision Trees are robust against overfitting when paired with techniques such as pruning, making them suitable for practical applications in cybersecurity. Recent studies have demonstrated that when integrated into larger frameworks, Decision Trees can significantly improve the accuracy of classification tasks by serving as a fundamental building block.

Naive Bayes is another frequently employed algorithm in cybersecurity, particularly for text classification and URL threat detection. This probabilistic classifier is based on Bayes' theorem and presumes independence among predictor features, which simplifies computation significantly [23]. It has a relatively low computational overhead, making it efficient in processing large datasets common in cybersecurity [24]. The strength of Naive Bayes in URL classification lies in its ability to classify based on the frequency of features present in the URL. For example, it can analyze textual components to discern patterns and predict whether a URL falls into a malicious or benign category. Studies indicate that Naive Bayes can achieve high accuracy rates, particularly when used with feature selection optimizations, such as the Particle Swarm Optimization technique demonstrated by Dulhare [25]. Furthermore, it has been found that Naive Bayes often outperforms more complex models when dealing with categorical input data, making it particularly suitable for applications in URL threat detection where speed is critical.

Method

Figure 1 outlines the complete project workflow, which begins with data loading, preparation, and exploratory analysis; proceeds to data preprocessing and modeling using Decision Tree and Naive Bayes; and concludes with feature importance analysis and the saving of the final models and tools.

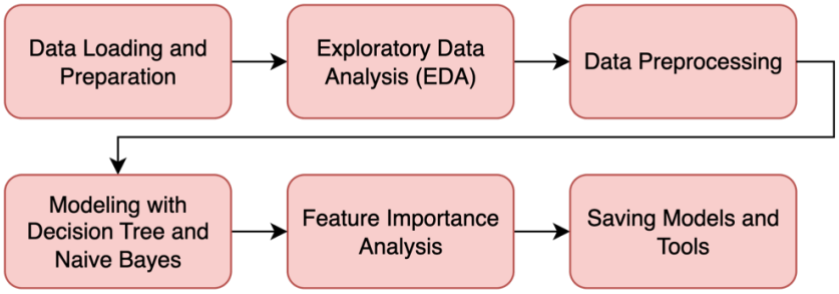

Figure 1 Research Method Flowchart

Dataset Loading and Preparation

The first step in this research involved loading the dataset containing URLs and their corresponding threat labels using the pandas library. The dataset was stored in CSV format and read into a DataFrame for ease of manipulation. To ensure data integrity, any rows containing missing values in critical columns such as 'url' or 'type' were dropped. This cleaning process is essential to prevent errors during model training and evaluation and to maintain the reliability of the results. By removing incomplete data entries, the study focused on high-quality, fully labeled examples that accurately represent the problem domain.

Exploratory Data Analysis (EDA)

Before proceeding with modeling, an exploratory data analysis was conducted

to gain insights into the dataset's structure and characteristics. Initial inspection included displaying the basic information about the dataset such as the number of records, column data types, and sample entries to understand the nature of the data. A class distribution plot was generated using seaborn's countplot function to visualize the frequency of each URL type—phishing, defacement, and benign. This helped identify potential class imbalances, which are critical to consider for classification tasks. Additionally, a word cloud visualization was created to capture the most frequently occurring terms and patterns in the URLs, providing an intuitive understanding of common tokens or URL segments that might differentiate threat categories. These visualizations assisted in formulating hypotheses about features relevant for classification.

Data Preprocessing

To prepare the dataset for machine learning algorithms, categorical labels in the 'type' column were converted into numeric values using label encoding. This transformation enables algorithms that require numerical input to process the class labels effectively. Since URLs are unstructured text data, it was necessary to convert them into a structured numeric form. This was achieved using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique. TF-IDF captures the importance of each token in a URL relative to its frequency across all URLs, thus highlighting distinctive features. To optimize computational efficiency and reduce dimensionality, the feature space was limited to the top 5,000 tokens. Following vectorization, the dataset was split into training and testing subsets, with 80% allocated for model training and 20% reserved for performance evaluation. This stratified split ensures that the models are evaluated on unseen data to assess their generalization capability.

Modeling with Decision Tree and Naive Bayes

Two supervised machine learning models were developed: the Decision Tree classifier and the Multinomial Naive Bayes classifier. The Decision Tree model was selected for its interpretability and ability to model complex, non-linear decision boundaries by learning hierarchical rules from the input features. On the other hand, the Multinomial Naive Bayes model was chosen for its computational efficiency and proven effectiveness in text classification, leveraging probabilistic assumptions to estimate the likelihood of each class given the features. Both models were trained on the vectorized training dataset, learning to differentiate URL threat categories based on the underlying feature patterns. After training, the models were saved as checkpoints using the joblib library, facilitating reproducibility and future deployment without retraining.

Model Evaluation

The performance of the trained models was rigorously evaluated on the test dataset. Key evaluation metrics computed included accuracy, precision, recall, and F1-score for each URL category. Accuracy provided a general measure of the overall correct classifications, while precision and recall assessed the models' ability to correctly identify true positives and minimize false positives and false negatives, respectively. The F1-score, as the harmonic mean of precision and recall, offered a balanced metric accounting for both error types. Confusion matrices were also plotted to visualize classification outcomes, enabling identification of specific misclassification patterns among URL types. These comprehensive evaluation tools allowed a detailed assessment of each

model's strengths and weaknesses in detecting phishing, defacement, and benign URLs.

Feature Importance Analysis

Understanding which features most influence the classification decisions is critical for interpretability and improving model transparency. For the Decision Tree classifier, feature importance scores were extracted to quantify the contribution of individual URL tokens to the prediction outcomes. The top 20 features with the highest importance scores were visualized in a horizontal bar chart, revealing key tokens and URL segments that strongly differentiate threat categories. This analysis provides actionable insights into the lexical elements of URLs that security practitioners should monitor closely and aids in refining future feature engineering strategies.

Saving Models and Tools

To ensure that the trained models and preprocessing tools could be reused efficiently, the Decision Tree and Naive Bayes models, along with the TF-IDF vectorizer and label encoder, were saved to disk using joblib. Saving these checkpoints allows future research, validation, or deployment to proceed without the need for retraining, saving time and computational resources. It also ensures consistency in model behavior over time, which is vital for production cybersecurity systems where stable and predictable performance is required.

This methodological framework combining thorough EDA, robust preprocessing, effective model training, detailed evaluation, and interpretability measures forms a comprehensive approach to URL threat classification using data mining techniques.

Result and Discussion

Dataset Overview

The dataset utilized in this research comprises a total of 24,800 URL records, each annotated with corresponding threat labels relevant to cybersecurity, including categories such as phishing, benign, defacement, and potentially other threat types. The dataset structure consists of three columns: an index column (Unnamed: 0), a URL string column (url), and the threat classification label (type). Notably, there were no missing values detected in the essential fields, ensuring that the data was complete and reliable for model training and evaluation purposes. A preliminary inspection of the dataset showed that URLs vary significantly in format and domain, reflecting a diverse range of examples encountered in real-world cyber environments. This diversity is crucial to developing robust classification models capable of handling various URL characteristics. The class distribution was found to be imbalanced, with some threat categories having significantly more examples than others, which poses typical challenges in machine learning classification tasks.

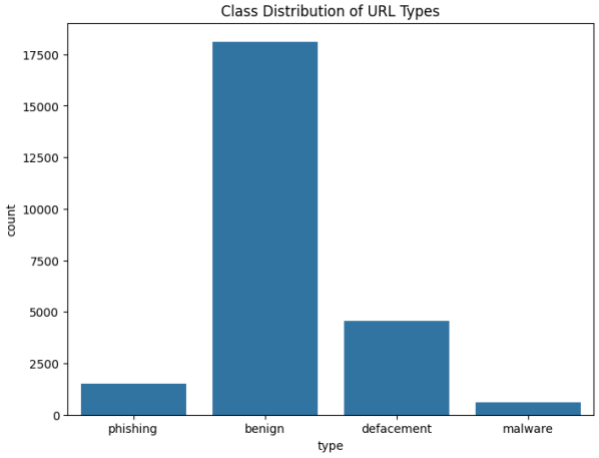

Figure 2 Distribution of URL Types

Figure 2 illustrating the class distribution of URL types reveals a pronounced imbalance among the four categories. “Benign” URLs dominate the dataset, numbering approximately 18,000 entries, which demonstrates that the majority of websites in the collection pose no threat. The second-largest category, “defacement,” encompasses around 4,500 URLs, indicating a substantial but much smaller segment than benign sites. Far fewer entries belong to the “phishing” category, which appears to comprise roughly 1,500 URLs, while the smallest group is “malware,” with an estimated count of about 600. This skewed distribution highlights the challenge of training models to accurately detect rarer threats, as classifiers can easily become biased toward the overwhelmingly prevalent benign class.

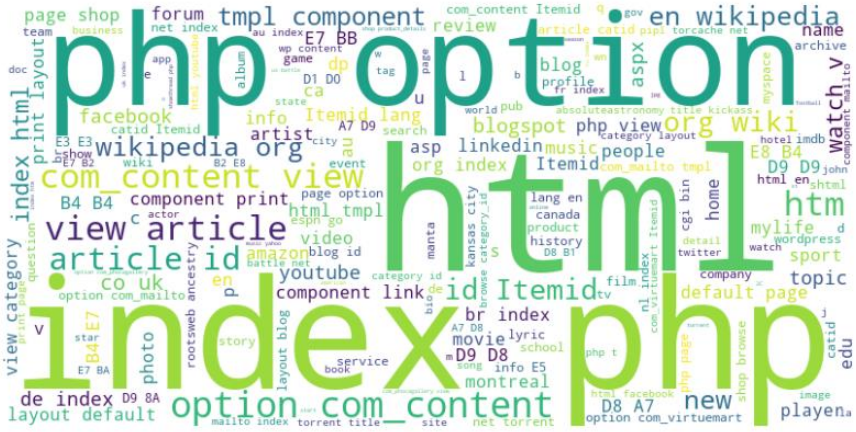

Figure 3 Word Cloud of Most Frequent Tokens

Figure 3 provides an intuitive glimpse into the most frequent tokens within the URLs across all categories. The largest and most prominent words—such as “php,” “html,” and “index”—underscore the prevalence of PHP-based web pages and common file-naming conventions in this collection of URLs. Other sizable tokens like “view,” “com_content,” and “article” suggest that many URLs originate from content management systems or dynamically generated pages. Smaller but still notable words, including “youtube,” “facebook,” and “wp,” hint at the presence of popular platforms or WordPress-based sites. The word cloud’s visual emphasis on these terms underscores how often certain components

appear in URLs, which in turn informs the feature extraction process during TF-IDF vectorization.

Decision Tree Model Performance

The Decision Tree classifier trained on this dataset demonstrated strong overall performance, achieving an accuracy of 94.01% on the unseen test set. The detailed classification report highlights the model's ability to effectively discriminate between different URL threat categories. For the most prevalent classes—such as phishing and benign URLs—the model achieved exceptional precision and recall scores, both exceeding 0.95, which indicates a low rate of false positives and false negatives for these categories. This level of performance suggests that the Decision Tree is highly capable of capturing distinguishing features that separate common threat types from safe URLs.

However, for less frequent categories in the dataset, the model's performance declined noticeably. The recall rate for the smallest class dropped to 0.47, with an F1-score of 0.56, showing that almost half of the instances in this class were misclassified. This is a common issue with imbalanced datasets, where the model tends to be biased toward majority classes due to insufficient examples of minority classes during training. The macro average F1-score of 0.81 reflects this imbalance, showing a moderate decline in performance when treating all classes equally. Meanwhile, the weighted average F1-score of 0.94 indicates that the model's overall accuracy is heavily influenced by the dominant classes in the dataset. Confusion matrices further revealed specific misclassification patterns, particularly between defacement and smaller threat categories, suggesting the need for improved feature engineering or sampling techniques to better capture subtle differences.

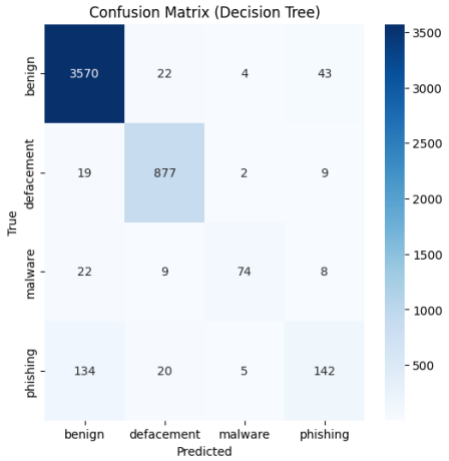

Figure 4 Confusion Matrix of Decision Tree

Examining the confusion matrix for the Decision Tree model (figure 4) reveals high accuracy for the two major classes—"benign" and "defacement"—with 3,570 benign URLs correctly classified out of 3,639 and 877 defacement URLs correctly identified out of 907. However, performance on "malware" and "phishing" is notably weaker: only 74 of 113 malware URLs were correctly identified, and just 142 of 301 phishing URLs were correctly classified. Many malware instances were misclassified as benign (22) or defacement (9), and phishing URLs were often mistaken for benign (134) or other threat types (25). Overall, the Decision Tree's ability to distinguish the two dominant classes is

excellent, but it struggles to separate rarer threats from the prevalent ones—an outcome that mirrors the underlying class imbalance.

Naive Bayes Model Performance

The Multinomial Naive Bayes model also demonstrated commendable performance with an overall accuracy of 92.36%, slightly lower than the Decision Tree but still competitive in the context of URL classification. Similar to the Decision Tree, Naive Bayes achieved very high precision (0.93) and recall (0.99) for the largest threat class (phishing), indicating robust detection capability for the most common and critical threats. This confirms Naive Bayes' suitability for text-based classification tasks, leveraging its probabilistic approach to identify URLs with typical phishing signatures.

Nevertheless, Naive Bayes showed considerable weaknesses when classifying less common URL threat categories. For one of the minority classes, the recall dropped sharply to 0.30, and the F1-score was just 0.45, meaning the model struggled to correctly identify many instances from this group. This suggests that Naive Bayes is less effective at generalizing across imbalanced data where distinctive features of minority classes may not be well captured by its underlying assumptions. The macro average F1-score of 0.74 reflects this inconsistency, while the weighted average of 0.91 suggests the model performs well when considering class proportions but lacks robustness for rare classes. The confusion matrix visualizations highlighted specific areas where Naive Bayes misclassified URLs, often confusing defacement with other minor threat types, underscoring the need for tailored solutions in handling imbalanced and nuanced data.

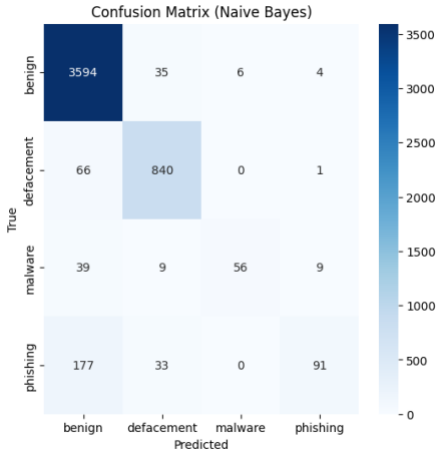

Figure 5 Confusion Matrix of Naive Bayes

In contrast, the Naive Bayes classifier's confusion matrix (figure 5) shows a similar pattern but with slight variations. For "benign" URLs, Naive Bayes correctly identified 3,594 out of 3,639, demonstrating robust performance for the majority class. "Defacement" detection also remains strong, with 840 correct classifications out of 907. However, the model's performance on "malware" and "phishing" is further diminished: only 56 of 113 malware URLs and 91 of 301 phishing URLs were accurately detected. In particular, numerous phishing URLs were misclassified as either benign (177) or defacement (33). These results reflect Naive Bayes' tendency to favor the majority classes, since it assumes feature independence and relies on token frequencies, making it less effective

at capturing subtle patterns that distinguish rarer threats.

Comparative Insights and Implications

When comparing the two models, the Decision Tree classifier showed a modest but consistent advantage in handling class imbalances and providing more balanced classification results across all threat categories. Its interpretability, through feature importance scores, also offers practical benefits for cybersecurity analysts aiming to understand which URL characteristics most influence classification decisions. Conversely, Naive Bayes provides faster training times and competitive accuracy for dominant classes but falls short in recognizing less frequent threats. These findings suggest that while both models are effective for automated URL threat classification, further enhancements could be achieved by integrating advanced techniques such as ensemble learning, synthetic minority oversampling, or neural network approaches to better manage data imbalance and improve detection rates for rare but critical threat categories. The high accuracy on common threats ensures that these models can form the basis of real-time cybersecurity monitoring systems, but attention to minority classes remains essential to avoid blind spots in threat detection.

Discussion

When comparing the performance of the Decision Tree and Naive Bayes algorithms, both models demonstrated strong capabilities in classifying URL threats with relatively high accuracy. The Decision Tree achieved a slightly higher overall accuracy of 94.01% compared to Naive Bayes' 92.36%, indicating its better ability to capture complex patterns within the data. Moreover, the Decision Tree offers greater interpretability, allowing for easy visualization of feature importance and decision paths, which is valuable for cybersecurity experts seeking to understand the reasoning behind classifications. In contrast, while Naive Bayes is computationally efficient and performs well on common classes, its probabilistic assumptions limit its flexibility, especially when handling nuanced or less frequent threat categories.

Despite these strengths, both algorithms faced significant challenges related to the inherent class imbalance present in the dataset. The majority classes, such as phishing and benign URLs, were classified with high precision and recall, but minority classes suffered from lower detection rates and increased misclassification. This imbalance led to reduced recall and F1-scores for smaller threat categories, which poses a critical limitation as these rare threats might be overlooked in practical cybersecurity applications. Handling such imbalance is challenging, as the models tend to bias towards the majority classes during training, making it difficult to achieve consistent performance across all categories.

Another challenge involved the subtlety and complexity of URL features that differentiate closely related threat types. Some URLs shared similar patterns or token distributions, leading to confusion between classes such as defacement and other minor threats. Both models struggled to fully capture these fine-grained differences, resulting in misclassifications reflected in the confusion matrices. These limitations highlight the need for further research into feature engineering, data augmentation, or advanced modeling techniques to improve detection rates for less common threats and enhance the overall robustness of URL classification systems in cybersecurity.

Conclusion

This study demonstrated that both Decision Tree and Naive Bayes algorithms are effective in classifying URL threats, achieving high accuracy in distinguishing phishing, benign, and defacement categories. The Decision Tree model slightly outperformed Naive Bayes, especially in handling class imbalances and providing more interpretable results through feature importance analysis. While both models showed strong performance on prevalent classes, challenges remained in accurately identifying less frequent threat types, highlighting the complexity of URL-based threat classification. The findings of this research have important implications for real-world cybersecurity applications, particularly in automating the detection and classification of malicious URLs. Implementing such models in security systems can enhance early threat detection, reduce reliance on manual inspection, and enable faster response to emerging cyber threats. The interpretability of models like Decision Trees also supports cybersecurity professionals in understanding and refining detection criteria, which is crucial for maintaining robust defense mechanisms against evolving online threats. For future research, there are several promising directions to improve upon this work. Integrating more complex machine learning models such as ensemble methods or deep learning architectures could enhance classification accuracy and better manage imbalanced data. Additionally, advanced feature engineering and data augmentation techniques could be explored to capture subtle patterns within URLs more effectively. Expanding the scope to include other types of cyber threats and leveraging multi-modal data sources, such as network traffic or user behavior, could also provide a more comprehensive and resilient approach to cybersecurity threat detection.

Declarations

Author Contributions

Conceptualization: D.A.D.; Methodology: T.B.K.; Software: D.A.D.; Validation: D.A.D.; Formal Analysis: T.B.K.; Investigation: T.B.K.; Resources: T.B.K.; Data Curation: D.A.D.; Writing Original Draft Preparation: T.B.; Writing Review and Editing: T.B.K.; Visualization: D.A.D.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A.-V. ANDRIU, "Adaptive Phishing Detection: Harnessing the Power of Artificial Intelligence for Enhanced Email Security," *Romanian Cyber Secur. J.*, vol. 5, no. 1, pp. 3-9, 2023, doi: 10.54851/v5i1y202301.

[2] M. W. Twain, "Assessing the Effectiveness of Cybersecurity Measures in the United States Telecommunications Industry: A Comprehensive Analysis," *J. Mark. Commun.*, vol. 6, no. 1, pp. 1-10, 2023, doi: 10.53819/81018102t4159.

[3] F. Al-Quayed, Z. Ahmad, and M. Humayun, "A Situation Based Predictive Approach for Cybersecurity Intrusion Detection and Prevention Using Machine Learning and Deep Learning Algorithms in Wireless Sensor Networks of Industry 4.0," *Ieee Access*, vol. 12, no. 3, pp. 34800-34819, 2024, doi: 10.1109/access.2024.3372187.

[4] M. A. Chowdhury, M. Rahman, and S. Rahman, "Detecting Vulnerabilities in Website Using Multiscale Approaches: Based on Case Study," *Int. J. Electr. Comput. Eng. Ijece*, vol. 14, no. 3, p. 28414, 2024, doi: 10.11591/ijece.v14i3.pp2814-2821.

[5] N. Ghazali, I. Suryani, and S. Z. Syed Idrus, "Challenges and Opportunities of Cybersecurity Education for Non-Technical Majors," *Jcsi*, vol. 6, no. 1, pp. 47-55, 2024, doi: 10.58915/jcsi.v6i1.873.

[6] A. M. Butnaru, A. Mylonas, and N. Pitropakis, "Towards Lightweight URL-Based Phishing Detection," *Future Internet*, vol. 13, no. 6, p. 154, 2021, doi: 10.3390/fi13060154.

[7] K. Veena, "Malicious URL Detection Using Machine Learning," *Interantional J. Sci. Res. Eng. Manag.*, vol. 7, no. 14, pp. 1-10, 2023, doi: 10.55041/ijsrem18973.

[8] M. Jaiswal, A. B. Raut, and M. T. Scholar, "Detection of Malicious URLs Using Classification Algorithms," *J. Res. Sci. Eng.*, 2022, vol. 4, no 11, pp. 1-12, doi: 10.53469/jrse.2022.04(11).19.

[9] S. Sankaranarayanan, A. T. Sivachandran, A. S. Mohd Khairuddin, K. Hasikin, and A. R. Wahab Sait, "An Ensemble Classification Method Based on Machine Learning Models for Malicious Uniform Resource Locators (URL)," *Plos One*, vol. 19, no. 5, p. 0302196, 2024, doi: 10.1371/journal.pone.0302196.

[10] F. Tiryaki, Ü. Şentürk, and İ. Yücedağ, "Developing and Evaluating an Artificial Intelligence Model for Malicious URL Detection," *Eur. J. Sci. Technol.*, vol. 2023, no. 17, pp. 13-17, 2023, doi: 10.31590/ejosat.1234556.

[11] A. M. Hilal *et al.*, "Malicious URL Classification Using Artificial Fish Swarm Optimization and Deep Learning," *Comput. Mater. Contin.*, vol. 74, no. 1, pp. 607-621, 2023, doi: 10.32604/cmc.2023.031371.

[12] K. Singh, P. Aggarwal, P. Rajivan, and C. González, "Training to Detect Phishing Emails: Effects of the Frequency of Experienced Phishing Emails," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 63, no. 1, pp. 453-457, 2019, doi: 10.1177/1071181319631355.

[13] S. Nifakos *et al.*, "Influence of Human Factors on Cyber Security Within Healthcare Organisations: A Systematic Review," *Sensors*, vol. 21, no. 15, p. 5119, 2021, doi: 10.3390/s21155119.

[14] C. Urcuqui, M. G. Peña, J. L. Osorio Quintero, and A. Navarro, "Antidefacement," *Sist. Telemática*, vol. 14, no. 39, pp. 9-27, 2016, doi: 10.18046/syt.v14i39.2341.

[15] M. Albalawi, R. Aloufi, N. Alamrani, N. Albalawi, A. Aljaedi, and A. R. Alharbi, "Website Defacement Detection and Monitoring Methods: A Review," *Electronics*, vol. 11, no. 21, p. 3573, 2022, doi: 10.3390/electronics11213573.

[16] M. U. Rehman, H. Bahşi, B. P. Linas, and B. J. Knox, "Exploring Trainees' Behaviour in Hands-on Cybersecurity Exercises Through Data Mining," *Eur. Conf. Cyber Warf. Secur.*, vol. 23, no. 1, pp. 585-593, 2024, doi:

10.34190/eccws.23.1.2141.
[17] R. Jenni and S. K. Shankar, “Review of Various Methods for Phishing Detection,” *Eai Endorsed Trans. Energy Web*, vol. 5, no 20, p. 155746, 2018, doi: 10.4108/eai.12-9-2018.155746.
[18] R. Taluja, “The Role of Data Mining in Cybersecurity: An Overview of Techniques and Challenges,” *Turk. J. Comput. Math. Educ. Turcomat*, vol. 10, no. 2, pp. 1056-1062, 2023, doi: 10.17762/turcomat.v10i2.13625.
[19] Z. Li, X. Li, R. Tang, and L. Zhang, “Apriori Algorithm for the Data Mining of Global Cyberspace Security Issues for Human Participatory Based on Association Rules,” *Front. Psychol.*, vol. 11, no. 2, pp. 1-10, 2021, doi: 10.3389/fpsyg.2020.582480.
[20] L. Wu and C. Yang, “Research on Data Mining of Network Security Hazards Based on Machine Learning Algorithms,” *Appl. Math. Nonlinear Sci.*, vol. 9, no. 1, pp. 1-12, 2024, doi: 10.2478/amns-2024-0074.
[21] F. Ullah and M. A. Babar, “Architectural Tactics for Big Data Cybersecurity Analytics Systems: A Review,” *J. Syst. Softw.*, vol. 151, no. 1, pp. 81-118, 2019, doi: 10.1016/j.jss.2019.01.051.
[22] B. Yu, F. Tang, D. Ergu, R. Zeng, B. Ma, and F. Liu, “Efficient Classification of Malicious URLs: M-Bert—a Modified BERT Variant for Enhanced Semantic Understanding,” *Ieee Access*, vol. 2024, no. 1, pp. 13453 - 13468, 2024, doi: 10.1109/access.2024.3357095.
[23] A. Meidina and Z. Abidin, “Diagnosis of Heart Disease Using Optimized Naïve Bayes Algorithm With Particle Swarm Optimization and Gain Ratio,” *Recursive J Inform.*, vol. 1, no. 2, pp. 47-54, 2023, doi: 10.15294/rji.v1i2.67278.
[24] M. Koca, İ. AVCI, and M. A. Shakir AL-HAYANİ, “Classification of Malicious URLs Using Naive Bayes and Genetic Algorithm,” *Sak. Univ. J. Comput. Inf. Sci.*, vol. 6, no. 2, pp. 80-90, 2023, doi: 10.35377/saucis...1273536.
[25] U. N. Dulhare, “Prediction System for Heart Disease Using Naive Bayes and Particle Swarm Optimization,” *Biomed. Res.*, 2018, vol. 29, no. 12, pp. 1-10, doi: 10.4066/biomedicalresearch.29-18-620.