big data and cognitive computing

Article

A New Approach to Data Analysis Using Machine Learning for Cybersecurity

Shivashankar Hiremath [1,2], Eeshan Shetty [1], Allam Jaya Prakash [3], Suraj Prakash Sahoo [4], Kiran Kumar Patro [5], Kandala N. V. P. S. Rajesh [6,*] and Paweł Pławiak [7,8]

[1] Department of Mechatronics, Manipal Institute of Technology, Manipal Academy of Higher Education, Udupi-Karkala Rd, Eshwar Nagar, Manipal 576104, Karnataka, India; ss.hiremath@manipal.edu (S.H.)
[2] Survivability Signal Intelligence Research Center, Hanyang University, Seongdong-gu, Seoul 04763, Republic of Korea
[3] School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Vellore 632014, Tamil Nadu, India; jayaprakash.allam@vit.ac.in
[4] School of Electronics Engineering (SENSE), Vellore Institute of Technology, Vellore 632014, Tamil Nadu, India; surajprakash.sahoo@vit.ac.in
[5] Department of ECE, Aditya Institute of Technology and Management, K Kotturu, Tekkali 532201, Andhra Pradesh, India; kirankumarpathro446@gmail.com
[6] School of Electronics Engineering, VIT-AP University, Inavolu, Amaravati 522241, Andhra Pradesh, India
[7] Department of Computer Science, Faculty of Computer Science and Telecommunications, Cracow University of Technology, Warszawska 24, 31-155 Krakow, Poland
[8] Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Baltycka 5, 44-100 Gliwice, Poland
* Correspondence: kandala.rajesh2014@gmail.com

Abstract: The internet has become an indispensable tool for organizations, permeating every facet of their operations. Virtually all companies leverage Internet services for diverse purposes, including the digital storage of data in databases and cloud platforms. Furthermore, the rising demand for software and applications has led to a widespread shift toward computer-based activities within the corporate landscape. However, this digital transformation has exposed the information technology (IT) infrastructures of these organizations to a heightened risk of cyber-attacks, endangering sensitive data. Consequently, organizations must identify and address vulnerabilities within their systems, with a primary focus on scrutinizing customer-facing websites and applications. This work aims to tackle this pressing issue by employing data analysis tools, such as Power BI, to assess vulnerabilities within a client's application or website. Through a rigorous analysis of data, valuable insights and information will be provided, which are necessary to formulate effective remedial measures against potential attacks. Ultimately, the central goal of this research is to demonstrate that clients can establish a secure environment, shielding their digital assets from potential attackers.

Keywords: analytics; cybersecurity; machine learning; Power BI; website

Citation: Hiremath , S.; Shetty, E.; Prakash, A.J.; Sahoo, S.P.; Patro, K.K.; Rajesh, K.N.V.P.S.; Pławiak, P. A New Approach to Data Analysis Using Machine Learning for Cybersecurity. *Big Data Cogn. Comput.* **2023**, *7*, 176. https://doi.org/10.3390/bdcc7040176

Academic Editors: Marco Fisichella and Antonia Russo

Received: 23 September 2023
Revised: 3 November 2023
Accepted: 8 November 2023
Published: 21 November 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Due to increased demand and increased competitiveness, the majority of operations conducted by corporate companies have become digital. In contrast to the conventional methods of corporate practices, the internet has become the sole tool for carrying out activities due to its reach and ability to attend to a larger clientele. Moreover, with the advent of new technologies like cloud computing and SQL, a large quantity of sensitive data can be stored in repositories online or in the cloud. However, the universality and versatility of the internet can prove to be fatal and a loophole for dangerous cyber-attacks, which might not only affect the sensitive data of the corporation but also affect the entire information technology structure of the organization by crippling it, halting operations, and, in turn, decelerating profits exponentially. Therefore, in order to combat this, it is essential to develop a foolproof, strong cybersecurity system to combat these loopholes and develop a remedial solution. Cyber security is one of the practices of developing machinery to combat cyber assaults on software and computerized systems. It has an application spread over virtually every sector due to its digital dependence on operations. Moreover, in the political, military, and healthcare sectors, it is all the more important due to the risks involved in the sensitive data being leaked.

Cyber-attacks come in various forms such as SQL injections [1], malware, and cross-site scripting (XSS) [2]. These attacks occur in various parts of the machinery or structure, like the database or the HTML code of the page, and manipulate the values to obtain unauthorized access to the data of clients and extort these data for monetary incentives and malicious gains. Companies lose a massive amount of data and money due to falling victim to these attacks and it is hence mandatory to invest in cybersecurity. In order to understand and develop a cyber-security strategy, it is essential to obtain and aggregate data from sources like passive scans and reports of the website's cyber health. After receiving the data, the next crucial step would be to analyze the data to obtain insights that can be provided to the clients and other departments so that they can act on the insights and formulate a new strategic plan. This is where data analytics comes into the picture. Data analytics [3] is the practice of collecting, managing, analyzing, visualizing, and presenting data. Over the years, data analytics has garnered significant popularity due to the influx of large amounts of data and due to it enabling data-driven decisions with higher chances of success. The ideal strategy of action can be planned by using the practices mentioned for cybersecurity data such as antivirus scan results, features of URLs, and the presence or absence of security features like security headers, and SSL certificates. This is why data analytics is important in the cybersecurity domain. An upgrade to data analytics and analysis is data science, which utilizes machine learning algorithms and deep learning to predict values for analysis to present predicted values to prospective clients so that clients can invest in the product. Machine learning has extensive uses in predictive analysis. It can be used to predict continuous values, like sales or values that do not have a 0 or 1 output. For continuous outputs, linear and polynomial regression models are used to predict the results. To predict categorical values of 0 or 1, classifier algorithms such as logistic regression, the KNN algorithm, the Naïve Bayes algorithm, the decision tree algorithm, and support vector machines are used. There are deep learning models on Keras, TensorFlow, PyTorch, etc., which are mainly used for emulating and programming neural networks to perform functions similar to a human brain. This would involve text detection, image classifiers, etc. However, these are only a few examples and there is not an exhaustive list.

In order to carry out the implementation of the cyber security system for protecting data, the tools most extensively used are Python 3.12 and Microsoft Power BI-V.2.119. 323.0. Visualizations and graphical representations of data are carried out by Microsoft Power BI, whereas the machine learning part is developed on the Python platform. Microsoft Power BI is a data visualization tool commonly used by analysts to make graphical depictions of the data via reports and dashboards which can be used to convey any insights to prospective clients. It has a free desktop version that allows users to make visualizations offline with no attached cost whatsoever. It also has a pro and premium version, which provides additional features like Azure Machine Learning, etc. Microsoft Power BI is the preferred tool mainly due to its ease of use, serving as a starting point for the majority of analysts. It also provides streamlined distribution and ease of working with real-time data. For the work assigned, a data set of the various cybersecurity subdomains is provided which contains all the cybersecurity characteristics of said subdomains. Power BI is used to create a report that will depict all the parameters with ease, make visualizations of these parameters, and come up with some basic conclusions and insights. Python is a dynamic universal programming tool that has applications in almost every sector of computer science and recently has become massively popular in the data science sector. It is an open-source language; hence, it does not require any form of payment to use, and it can be installed easily. In the data science domain, Python is mainly used for both data analysis and machine learning. From Python, some of the tasks can be identified as obtaining data sets from data engineers or via web scraping; removing and cleaning redundant values and modifying faulty values; conducting analysis; obtaining desired insights; and using machine learning algorithms to calculate the most accurate algorithm for predicting the likeliness of a cyber-attack.

Table 1 serves as a valuable resource for researchers and practitioners in the field of intrusion detection, offering a quick reference for the key attributes and findings of these influential studies. These studies collectively contribute to the ongoing effort to enhance the security of computer systems and networks. Researchers have explored a range of techniques, from traditional approaches like *K*-nearest neighbors (K-NN) and support vector machines (SVM) to more modern methods like random forest (RF) and Genetic Algorithms. The choice of method often reflects a trade-off between factors such as accuracy, training time, and false alarm rates.

Table 1. Literature survey data analysis using machine learning for cybersecurity.

Literature	Year	Method	Database	Number of Classes	Remarks
Swathi et al. [4]	2017	K-NN	NSL-KDD	3	Author not reported Precision and Recall
Verma et al. [5]	2018	K-NN and K-means	CIDDS-001	2	Method has low false alarm rate
Belouch et al. [6]	2018	SVM RF DT NB	UNSW-NB15	2	More training time required
Krichen et al. [7]	2017	Logistic Regression with Genetic Algorithm	UNSW-NB15	3	Less accuracy
Jabbar et al. [8]	2016	RF	NSL-KDD	3	More time for prediction

2. Background

2.1. Cybersecurity

Cybercrime is an offense that uses a computer or a device as a vector to attack another system to either sabotage the device or gain unauthorized access to data. It can also involve other activities, such as fraud, identity theft, and suspension of the system. Attackers then extort the victim for monetary gain, which can result in severe losses for companies. It is crucial to implement cybersecurity [9] strategies as most data are now cyber-oriented and stored on the internet, making them the most vulnerable to cyber crimes. In the context of the task assigned, it is crucial to analyze the trends in cybersecurity. Web servers have proven to be a very susceptible platform to cyber-attackers. Attackers deploy their hazardous code and techniques on affected servers and hence they must be diagnosed first.

Cloud computing [10] is becoming a growing norm for the majority of companies. Via the cloud, companies can create and deploy apps, manage data, store terabytes of data in their storage facilities, and perform artificial intelligence and machine learning functions. However, despite the increase in the number of features offered by the cloud, concerns are raised regarding the security features. Therefore, cloud providers must think about developing a secure system to protect the data. Today, via mobile networks, earlier issues of accessibility have been bridged. Data can be stored on compact cellular devices, making them the most popular device amongst the masses recently. However, due to this popularity, they become more susceptible to cyber-attacks. Therefore, various strategies like firewalls and other protective practices must be implemented to prevent any data breaches. In order to develop safe strategies for these devices, some cyber security techniques are used by the cyber security team of a company. Malware detection programs [11] that scan the system files to detect any flaws and viruses are used by cybersecurity engineers. Malware is a general term for a variety of attack types, some of which are viruses, worms, ransomware, and trojans, but this is not an exhaustive list [12,13]. Firewalls can be considered as a screening mechanism that protects the system from any form of hacking bugs or viruses. They screen all forms of content entering the system from the internet and filter out all content messages and commands that cannot meet a certain criterion. Furthermore, they can perform other functions, such as stateful inspection and application-layer filtering. Antivirus systems are computer software programs for diagnosing any form of malicious content. These scans can be updated and progressively implemented to obtain an output that shows the loopholes present and also to discover any new viruses which were introduced later so that is does not become redundant.

2.2. Big Data Analytics

Big data analytics is the terminology coined for the analysis of data that has a processing power in a range that surpasses conventional databases and is restricted depending on the application. In this case, the amount of data is too large and is produced at a very high speed, making the data impossible to handle. Big data has the characteristics of the four main Vs: volume, velocity, veracity, and value. In addition to these, variety, variability, validity, visualization, and vulnerability are also used in big data analytics [14,15]. To be able to store these data, different types of databases can be used. NoSQL is the abbreviated form of 'not only SQL', and NoSQL databases differ from other database tables as they do not store data in a tabular format as in relational databases. The major advantage of these databases is their ability to store large quantities of data at high speed. Moreover, they can accommodate various data types, including unstructured, semi-structured, and structured data. These data types are crucial for modern data management systems. Unstructured data, such as text and multimedia content, lack a predetermined data model. Semi-structured data, such as XML or JSON, have some organizational qualities but do not adhere to a rigorous format. Structured data, on the other hand, adhere to a predefined standard and can easily be grouped into rows and columns. By efficiently managing these many data kinds, firms can extract important insights, improve decision-making processes, and find hidden patterns or trends that might otherwise go unnoticed. Cloud storage systems have been used in recent times to store a lot of data. One of the main benefits of cloud storage is its increased scalability. In addition to this, it is very easy to retrieve data at zero startup cost. Examples of cloud providers' storage services [16] are Amazon, S3, and Azure Data Lakes. Using these tools, volumes of client data are stored and utilized by data analysts whenever necessary for analysis. After retrieving data from databases, they are then imported into a visualization tool, like Power BI or Tableau, in JSON or CSV format and then the data are analyzed further. In addition, it is noteworthy that other formats, such as XML, are also commonly used or direct database connections are created in data analysis and visualization. These data formats and communication mechanisms are supported by a wide range of visualization tools.

2.3. Machine Learning in Cybersecurity

Big data implies the storage of large volumes of data at high speeds. This provides a massive quantity of data that can be broken down into simpler insights with the help of analytical tools. Further, the quality of these insights also depends on the quality of the data and the analytical methods. Small sets of obtained data are taken for analysis using machine learning algorithms. Various outputs can be obtained which can help analysts detect potential issues. Through machine learning, a vivid look at the severity and type of attack can be obtained and hence one is able to statistically decipher the loopholes present and present the findings to prospective clients. Machine learning also helps in developing predictive models that can be used to detect the entry point of the cyber assault and then take the correct analytical approach depending on the scenario. Furthermore, machine learning approaches in cybersecurity are becoming increasingly important for improving threat detection, improving incident responses, and protecting digital systems from a variety of cyber-attacks. Incorporating big data analytics into the cybersecurity domain is a step towards positive development as it provides an accurate analysis of the predicament and also a calculated idea of various hypothetical scenarios. It is hence essential for all companies and firms to employ machine learning models in their cybersecurity strategies. Depending on the desired outcome, different kinds of algorithms can be used. (a) Regression models [17–19] are mainly used for continuous variables. To execute a regression model, existing data are taken and split into testing and training data sets. The training data are then trained with a regression algorithm and predictions are made. In the domain of cyber-security, regression algorithms are used for detecting variables like the number of fraudulent transactions and the possible location of a cyber–attack. Different types of regression models can be used depending on the arrangement of data points, like linear, polynomial, ridge, and lasso regression. (b) Classification algorithms are mainly used to determine a binary or a '0 or 1' output. In the cyber-security domain, these algorithms can be used to detect the status of the cyber health of the URL. While executing an algorithm, they can expose the data to a variety of classifier algorithms, calculate the accuracy score of these algorithms, and decide the best algorithm which must be deployed. The algorithms used here are mainly logistic regression, decision tree [20–22], the k-nearest neighbors algorithm [23], and support vector machine [24,25]. Classification is also used to segregate spam mail. For greater accuracy, increasingly larger data sets are employed in deep learning [26–28]. They can be used for various functions like text detection, image classification, or even for classification. For regression, artificial and recurrent neural networks [29–31] are used. Thus, it is ideal to incorporate big data analytics into the cybersecurity domain. With the increase in the number and sophistication of cyber-attacks, it is essential to have a data-driven strategy to make calculated decisions. Therefore, data analysis tools can be made with specific visualizations and presented to prospective clients. Moreover, with cyber-attacks constantly developing, it is essential to understand their nature and an ideal strategy must be formulated. Moreover, machine learning predictive analysis can be used to determine the location and probability of a cyber-attack. Finally, a comprehensive analysis will be carried out to provide a detailed analysis and provide insights for the authorities.

3. Problem Statement

The main aim of the study is to analyze a stored URL, detect its validity, and enumerate the security headers of the particular URL. This can be executed by using a Python Script created using the correct codes and functions of Python. Then, a JSON data set of cybersecurity data is used, which is imported into Microsoft Power BI for further analysis. Before using Power BI, one needs to know why one wants to analyze the cyber security data and then obtain the data to satisfy the need for analysis. Then it can begin to process and prepare the data. This would involve segregating the data, replacing erroneous values with correct values, dropping duplicates if required, dropping null rows or filling them with a particular value, etc. Then, there is the requirement of cleaning the data and making them incomprehensible for visualization. With this, various figures and facts of the data are presented to the authorities and prospective clients. In order to predict the result of the URL data supplied, machine learning algorithms are used to identify the type or category of a URL, such as whether it is related to e-commerce, news, entertainment, education, etc, and determine whether a URL is potentially harmful or malicious, such as phishing websites, malware-hosting sites, or fraudulent pages.

4. Proposed Methodology

For implementation of the cyber security system, one needs to understand the various cyber-attacks. Structured Query Language (SQL) is mainly used as a database for storing data in large volumes. SQL is a database management tool into which corporations can enter data in a structured tabular form. Using SQL queries, they can add, create, and modify values in the database. SQL is a very useful tool for analysts as they can easily work with huge data sets with simple queries. It is very important to understand the importance of keeping databases protected so that the sensitive data stored in them do not get leaked and extorted by cyber criminals. SQL injection is a type of cyber-attack where a cyber-attacker injects a malicious code that alters the sequence of the SQL queries. This can in turn affect the output of the SQL command and have fatal consequences like unauthorized access to the data. For example, an attacker can use a malicious SQL code to gain access to financial and banking details of the corporation, gain unauthorized access to the corporation's funds, and then extort the funds of the organization. Figure 1 shows the unauthorized access during an SQL injection. Unauthorized access during SQL injection attacks is a critical concern in the realm of cybersecurity. SQL injection is a well-known and widely exploited vulnerability that occurs when malicious actors inject malicious SQL code into input fields or web application parameters. The consequences of a successful SQL injection attack can be dire, as it allows unauthorized access to a database, application, or system. This section delves into the intricacies of unauthorized access during SQL injection, discussing its implications, methods of exploitation, and the measures to prevent such security breaches. To avoid such mishaps and offenses from occurring, it is important to keep the database's password encrypted and regularly analyze the SQL activity.

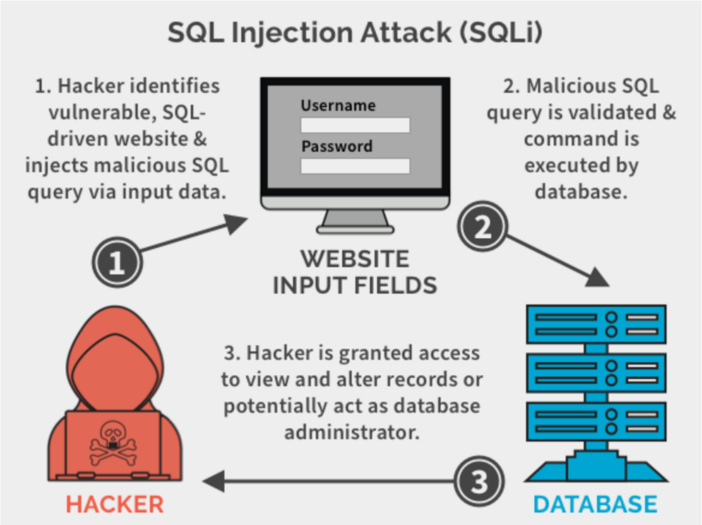

SQL Injection Attack (SQLi)
1. Hacker identifies vulnerable, SQL-driven website & injects malicious SQL query via input data.
Username
Password
2. Malicious SQL query is validated & command is executed by database.
1
2
WEBSITE INPUT FIELDS
3. Hacker is granted access to view and alter records or potentially act as database administrator.
3
HACKER
DATABASE

Figure 1. Steps occurring during a SQL injection.

Malware is the collective term used for harmful software that is deployed into computer systems to harm and damage the infrastructure of the system and in turn extort and unlawfully acquire data. Adware [32] is a type of malware that forces the user to divert to internet ads which is a platform for more forms of malware. This malware is generally found in enticing advertisements on websites or in games. Spyware [33] is used as tracking software, where malicious software is introduced into the system and secretly tracks the activities conducted by the system, secretly collecting data. An example of spyware would be stealing passwords of company accounts. Viruses contain software or code that manipulates the system to perform a malicious action before spreading to other parts of the system. Worms involve harmful software which when introduced into the system reproduces and replicates to amplify the malicious effect. Trojan is software that does not replicate itself but compels the system into unintentionally activating an action that can cause severe damage to the system, like deleting, modifying, and copying data or just disrupting the company's system operations. This software is hazardous to the system so remedial measures must be taken.

Cross-site scripting [2] is another type of injection cyber-attack, where a code is injected into the website's HTML fundamental structure to unlawfully export data from the website. As depicted, the perpetrator keeps note of the user's web activity by tracking the cookies of the visit. After discovering and locating a loophole, they develop a code to inject into the system. By injecting the code into the HTML code of the site, the attacker gains control of the visitor's session cookies which can be used to track the user's activity and extort data from the visitor. This injection affects subsequent visitor searches by tracking their activity.

These are the cyber-attacks that have significantly affected the operations of many companies and need to be attended to to prevent any further losses and cessation of operations. Therefore, it is important to have a data-driven strategy to obtain insights into the attack and predict the outcome of a cyber-attack to prepare a strategy to combat the attack before it even occurs. To do this, a Python code was executed using machine learning algorithms. Python contains predefined libraries which can be imported to satisfy certain actions. The most popular libraries are NumPy, which is mainly involved in mathematical operations and for creating multidimensional arrays to input data into; Pandas, for data manipulation and for data structures like dictionaries to create data frames and series; and Matplotlib, for visualizing data. There is another library called Seaborn which is built on Matplotlib for more unique visualizations like box plots, regression plots, scatter plots, etc. Machine learning mainly uses different libraries, which can be used for all regression and classification models. The cybersecurity system can be built in the same platform to analyze the security system in a corporate environment.

4.1. Detecting the Validity of URLs and Displaying Security Headers

The script depicts the Python code which creates a dictionary with all security headers for a particular URL in order to determine the validity of the URL by importing the Python library validators. Validators verify the syntax of the input string variable and if it satisfies the syntax of an internet URL, a 'True' value is returned. After determining the validity of the URL, in same way, one proceeds to the next step to create the dictionary. Adding more information about the request or response, HTTP headers are a crucial component of the HTTP protocol. For clients and servers to communicate securely and effectively, they are essential. X-Frame Options, X-Content Options, and Server are a few crucial HTTP headers for increased security. Finally, the function is used to detect the values from the dictionary keys 'Server', 'X-Frame Options', and 'X-Content Options', as these headers are essential for enhanced security. Algorithm 1 shows the creation of a security header and the display of the URL.

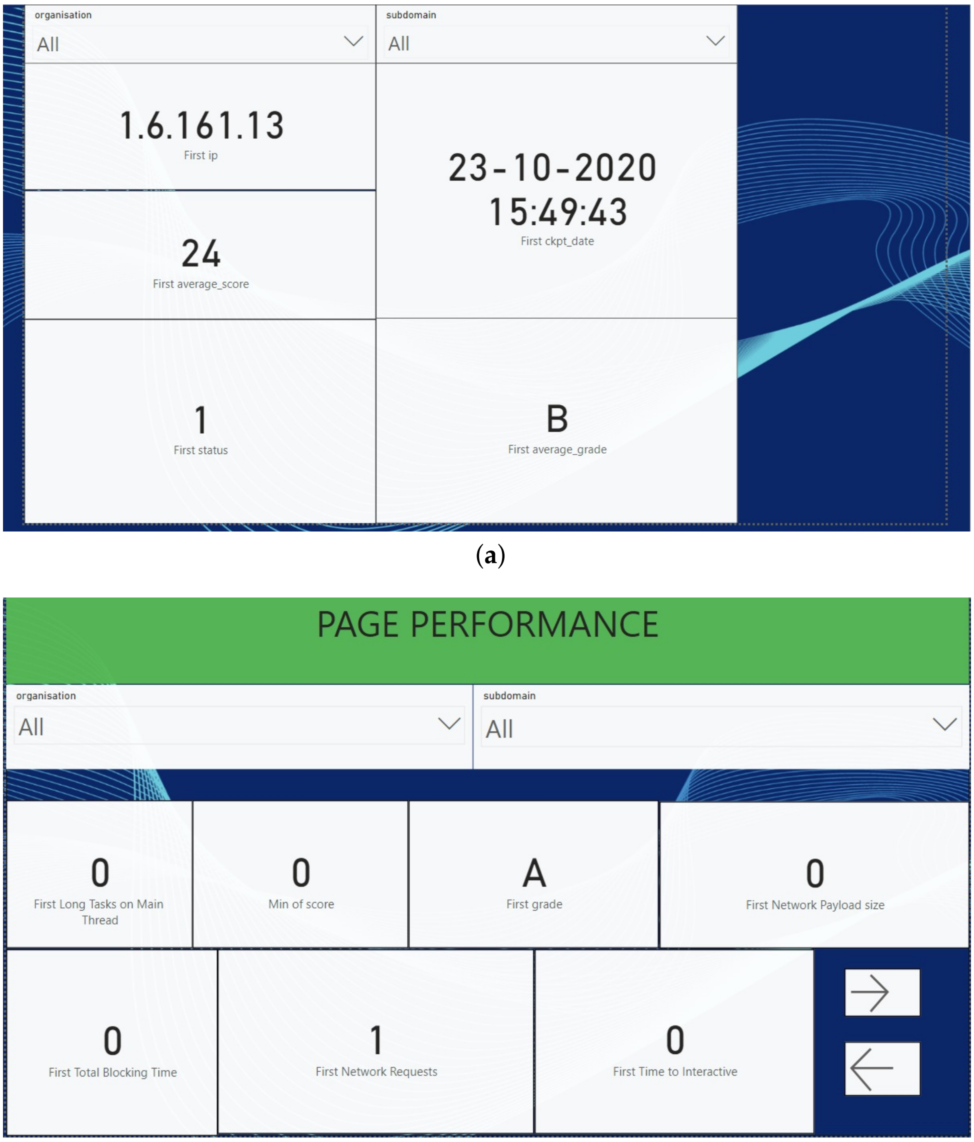
Algorithm 1: Function created for the security header and displaying the URL

```
Def check key (dicts, key): initialization;
if key in dicts. keys (): then
    print ("The following key is present:", key);
    value = dicts.get(key);
    z = print (key,":", value);
    return z
end
```

4.2. Analysis of URL Data via Power BI

Power BI is an interactive tool used for the visualization of data. The collected and filtered data are visualized to obtain the display insights that are sought. In order to conduct the visualization, the following steps of the data analysis cycle need to be undertaken. Initially, the need for data analysis needs to be understood and data should be collected from different sources. Then, filtration of data and modification of null and redundant data are carried out. Further, the data are analyzed and visualized in different graphical formats. In the end, the data are presented to the prospective clients and the result is decided.

The importance of the data is related to knowing how effectively the data were created and analyzed. This involves having a discussion with the client to understand their requirements and needs before one may proceed. The collection of data takes time. Knowing where the data needs to be collected from and how much time is required is essential for the processing of all data. This can be determined via web scraping or by obtaining the data from data engineers. Data are generally obtained in CSV or JSON formats. The collected data contain various information about subdomains of major corporate URLs. Once the data are procured, filtering with the Power Query Editor to remove redundant and null data is required. Furthermore, the type of filtration can be set. In the columns and rows of the editor are text filters, number filters, date filters, and more. With the Power Query Editor, data can be filtered and modified with the help of dropdowns as well as by adding new columns and rows. Then, the data are more comprehensible and available for analysis. The analysis can be performed by creating a visualization to obtain visual insights into the data.



organisation: All

subdomain: All

1.6.161.13 First ip

24 First average_score

1 First status

23-10-2020 15:49:43 First ckpt_date

B First average_grade

(**a**)

PAGE PERFORMANCE

organisation: All

subdomain: All

0 First Long Tasks on Main Thread

0 Min of score

A First grade

0 First Network Payload size

0 First Total Blocking Time

1 First Network Requests

0 First Time to Interactive

(**b**)

Figure 2. Examples of (**a**) a page of a dashboard displaying data and (**b**) graphical visualizations made via Power BI.

Visualizations of the data may be performed for any subdomain level by simply changing values in the dashboard. This may predict information like score grade, the IP address, etc., as shown in Figure 2. Furthermore, they can be used for comparative analysis of different subdomains and to construct a graphical representation. Power BI dashboard pages are a central element in the world of data analytics and business intelligence. They provide a visually engaging and interactive platform for users to explore, analyze, and gain insights from data. The combination of data tables and graphical visualizations, along with various interactive features, makes Power BI a valuable tool for organizations aiming to harness the power of their data for better decision-making. Whether they are monitoring key performance indicators, tracking sales trends, or assessing operational efficiency, Power BI dashboard pages play a crucial role in transforming data into actionable intelligence.

A graphical plot of the scores of all the subdomains can keep count of the other subdomain grades via graphical visualizations (Figure 2b). This eases the process of analysis tenfold and hence is a popular tool amongst data analysts. One of the most important aspects of cybersecurity is predicting online attacks. Algorithms for machine learning can be used to identify and stop different kinds of assaults and other harmful behaviors directed at web services. Logistic regression, KNN, decision tree, support vector machine, XGboost, SK-learn neural network, Naïve Bayes, and random forest are machine learning techniques used for this purpose. Normal and assault examples are included in the labeled data sets used to train these machine-learning methods. The cybersecurity data for machine learning models were compiled using relevant cybersecurity data from a variety of sources, such as network logs, security devices, intrusion detection systems, and so on. Inconsistent or useless data were deleted from the data sets. Then, the relevant properties that are required for the analysis were determined. Important characteristics were extracted from raw data and new features that can improve the model's prediction were produced. Categorical data were converted to numerical representations using label encoding techniques. Numerical data were normalized or scaled so that all features were on the same scale. Finally, the data set was divided into training, validation, and testing sets to appropriately evaluate the machine learning model's performance. A variety of variables, such as request patterns, IP addresses, user-agent data, request frequencies, and other pertinent contextual information, were retrieved from web traffic to create the training data. During the training phase, these features help the algorithms learn to differentiate between legitimate and malicious online traffic. Potential website assaults can be identified and stopped by using trained models to forecast and categorize incoming online requests in real-time.

4.3. Machine Learning with Python

The implementation of a security system for malware was carried out using machine learning algorithms on the Python platform. Here, two variables, X and Y, are used, where X is the independent variable and comprises all numerical values leaving the type of column, and Y is the dependent variable or the variable to be predicted. These data types are split into X and Y as training and test data sets, respectively. The Scikit-learn library was used for initializing the value, and then random forest, support vector machine, logistic regression, and gradient boosting machine were used for implementing the machine learning algorithms. Classifier algorithms were used to determine the classes of threat. The main steps for formulating a machine learning algorithm involve fitting the training data values and then predicting the X-test values. To improve the accuracy of ML models, exploratory data analysis was conducted, like outlier treatment, box plots, scatter plot correlation matrices, etc., to obtain more general data insights and to fine-tune the machine learning model. After using all the classification models, the final accuracy was calculated using a confusion matrix. Then, the results were analyzed and an appropriate conclusion was formed.

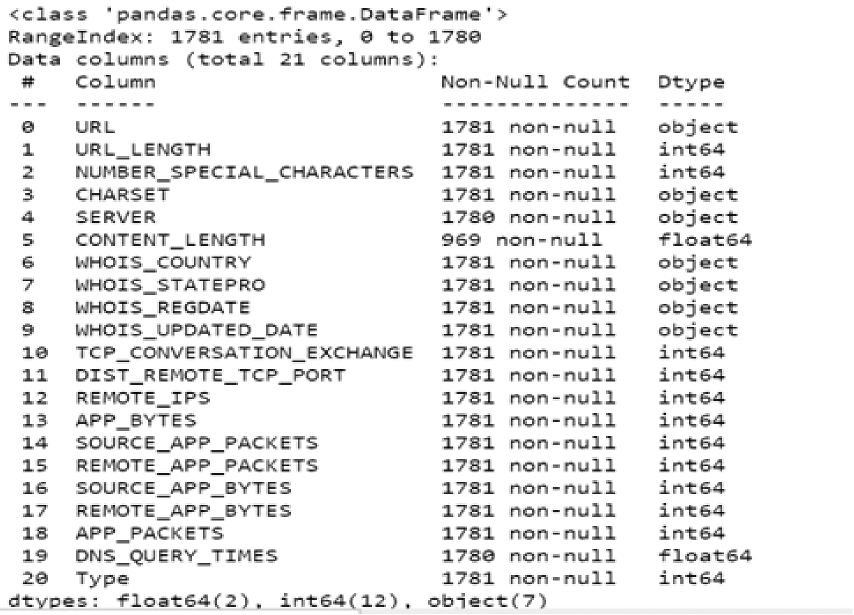
5. Results and Discussion

The machine learning approach helps detect threats to corporations and provides a security system for them. Different approaches have been developed to predict the severity of the cyber-attack based on the score created in the data for the sake of analysis. The analytical model uses a neural network platform to form a regression and classification model. Then, the data were analyzed based on an accuracy score and confusion matrix.

The task was carried out by importing different libraries like NumPy, pandas, scikit learn, and Keras 2.10.0 to visualize the data. Seaborn and Matplotlib were used to generalize the data set information. Figure 3 shows a total of 14 numerical data types and 7 object data types. It is evident that the variable 'CONTENT_LENGTH' has several null variables that need to be eliminated for analysis. It was also determined that there are no duplicate rows for the data. This data pre-processing leads to a statistical characterization of the data at hand, which is depicted in Table 2.

Table 2. Summary of statistics of the given data set.

	Count	Mean	Std	Min	25%	50%	75%	Max
URL_LENGTH	967.0	55.5604	25.5724	16.0	39.0	49.0	66.5	249.0
NUMBER_SPECIAL_CHARACTERS	967.0	10.8086	4.1243	6.0	8.0	10.0	12.0	43.0
CONTENT_LENGTH	967.0	11,733.6970	36,429.1315	0.0	324.0	1853.0	11,324.5	649,263.0
TCP_CONVERSATION_EXCHANGE	967.0	16.9648	45.1398	0.0	0.0	8.0	22.0	1194.0
DIST_REMOTE_TCP_PORT	967.0	3.4694	6.8890	0.0	0.0	0.0	4.0	58.0
REMOTE_IPS	967.0	3.1882	3.3698	0.0	0.0	2.0	5.0	17.0
APP_BYTES	967.0	1720.4095	3925.8009	0.0	0.0	762.0	2332.5	99,843.0
SOURCE_APP_PACKETS	967.0	19.3526	46.0724	0.0	0.0	10.0	26.0	1198.0
REMOTE_APP_PACKETS	967.0	19.3826	50.0361	0.0	0.0	10.0	25.0	1284.0
SOURCE_APP_BYTES	967.0	17,021.9462	79,423.8336	0.0	0.0	822.0	9382.0	2,060,012.0
REMOTE_APP_BYTES	967.0	1904.9679	4017.9624	0.0	0.0	880.0	2765.5	100,151.0
APP_PACKETS	967.0	19.3526	46.0724	0.0	0.0	10.0	26.0	1198.0
DNS_QUERY_TIMES	967.0	2.3805	2.8218	0.0	0.0	0.0	4.0	20.0
TYPE	967.0	0.109617	0.3125	0.0	0.0	0.0	0.0	1.0



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1781 entries, 0 to 1780
Data columns (total 21 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   URL                        1781 non-null   object
 1   URL_LENGTH                 1781 non-null   int64
 2   NUMBER_SPECIAL_CHARACTERS  1781 non-null   int64
 3   CHARSET                    1781 non-null   object
 4   SERVER                     1780 non-null   object
 5   CONTENT_LENGTH             969 non-null    float64
 6   WHOIS_COUNTRY              1781 non-null   object
 7   WHOIS_STATEPRO             1781 non-null   object
 8   WHOIS_REGDATE              1781 non-null   object
 9   WHOIS_UPDATED_DATE         1781 non-null   object
 10  TCP_CONVERSATION_EXCHANGE  1781 non-null   int64
 11  DIST_REMOTE_TCP_PORT       1781 non-null   int64
 12  REMOTE_IPS                 1781 non-null   int64
 13  APP_BYTES                  1781 non-null   int64
 14  SOURCE_APP_PACKETS         1781 non-null   int64
 15  REMOTE_APP_PACKETS         1781 non-null   int64
 16  SOURCE_APP_BYTES           1781 non-null   int64
 17  REMOTE_APP_BYTES           1781 non-null   int64
 18  APP_PACKETS                1781 non-null   int64
 19  DNS_QUERY_TIMES            1780 non-null   float64
 20  Type                       1781 non-null   int64
dtypes: float64(2), int64(12), object(7)
```

Figure 3. The pre-processing data of the security system.

Outlier treatment is another important task to filter data to obtain optimum parameters. Outliers are values that are way above or beyond the permissible range of values for the data. They can be problematic because they may affect the efficacy of the model. Hence, to avoid handling inefficiencies, outliers need to be treated before proceeding with machine learning model creation. A box plot showing outliers of columns in the data is shown in Figure 4. Box plots are a valuable tool for summarizing and visualizing the distribution of numerical data. They provide a clear and concise representation of the data set's central tendency, variability, and potential outliers. Box plots are particularly useful when comparing multiple data sets or when identifying the overall distribution characteristics of a single data set. It is evident that CONTENT_LENGTH and SOURCE_APP_BYTES have large numbers of outliers, which require treatment. To initiate treatment, the upper quartile, lower quartile, and interquartile ranges were calculated. Obtaining the upper and lower threshold values may give an insight into the outlier ranges. Outliers are data points that exhibit values that are notably different from the central tendency of the data. They can be caused by various factors, including measurement errors, natural variability, or genuine extreme observations. Identifying and managing outliers in data variables are fundamental steps in data analysis and statistical modeling. Outliers can significantly affect the results and interpretation of analyses, and, therefore, it is crucial to employ appropriate methods for their detection and handling. Understanding the nature and impact of outliers allows researchers and analysts to make more accurate and reliable inferences from their data, ultimately improving the quality of decision-making and research outcomes.

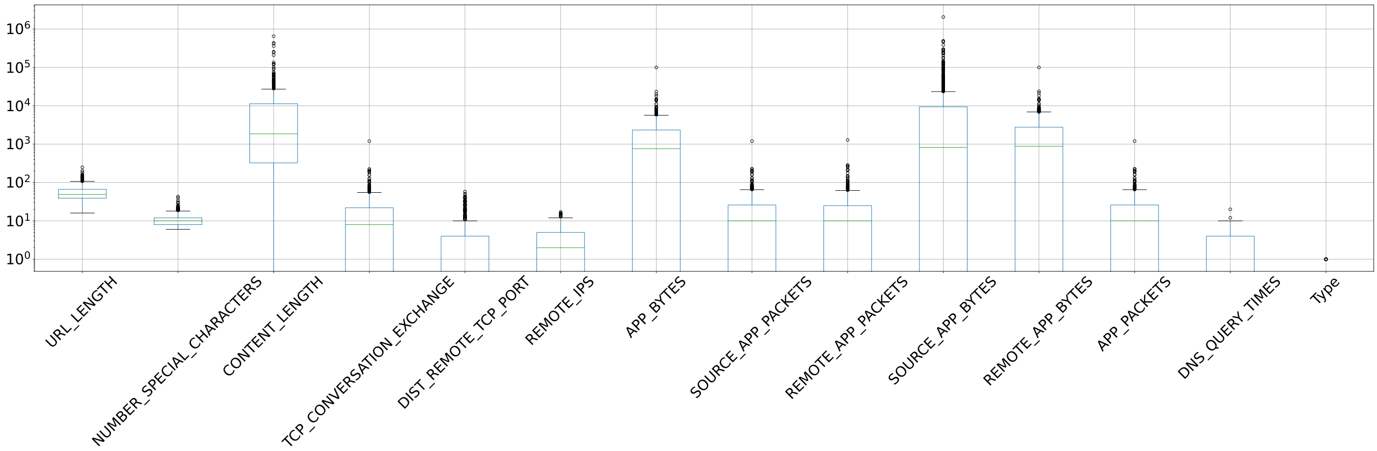


Figure 4. Outliers of the different data variables.

Once the outliers were treated, we could proceed to the univariate and bivariate analyses of the data. Here, the variant information was visualized using a histogram of all the data values. This information provides the frequency of all the values shown in Figure 5. Histograms are valuable graphical representations for understanding the distribution of data sets. They provide a visual summary of the frequency or count of data points falling within specific intervals or bins. Histograms provide a visual and intuitive way to understand the frequency distribution of data sets. By constructing and interpreting histograms for different data sets, researchers and analysts can make informed decisions, identify patterns, and gain valuable insights into the nature of their data. The ability to visualize data in this manner is a crucial skill in data analysis, statistics, and decision-making processes. A correlation matrix shows the correlation coefficients between several variables. Furthermore, it measures the strength and direction of the linear relationship between two variables. Here, the range varies from -1 to 1, where -1 means a perfect negative correlation, 0 means no correlation, and 1 means a perfect positive correlation. The correlation data of the security model are shown in Figure 6. A correlation matrix is a valuable tool for exploring and quantifying the relationships between multiple variables within a data set. It aids in identifying dependencies and associations that can be useful for data-driven decision making, model building, and understanding the underlying patterns in data. A well-constructed correlation matrix is essential in the process of data analysis and statistical exploration. Based on the data set, the security system needs to be analyzed using the URL content. Here, different sets of URLs were run in the system, such as Google, Facebook, Amazon, etc. The security headers task showed that Google has an X-Frame header as well as a server header in their code, depicting that the website is well protected from a variety of cyber-attacks and has good cyber health. It was also evident that bigger companies' URLs had a larger number of security headers in comparison to smaller competitors.

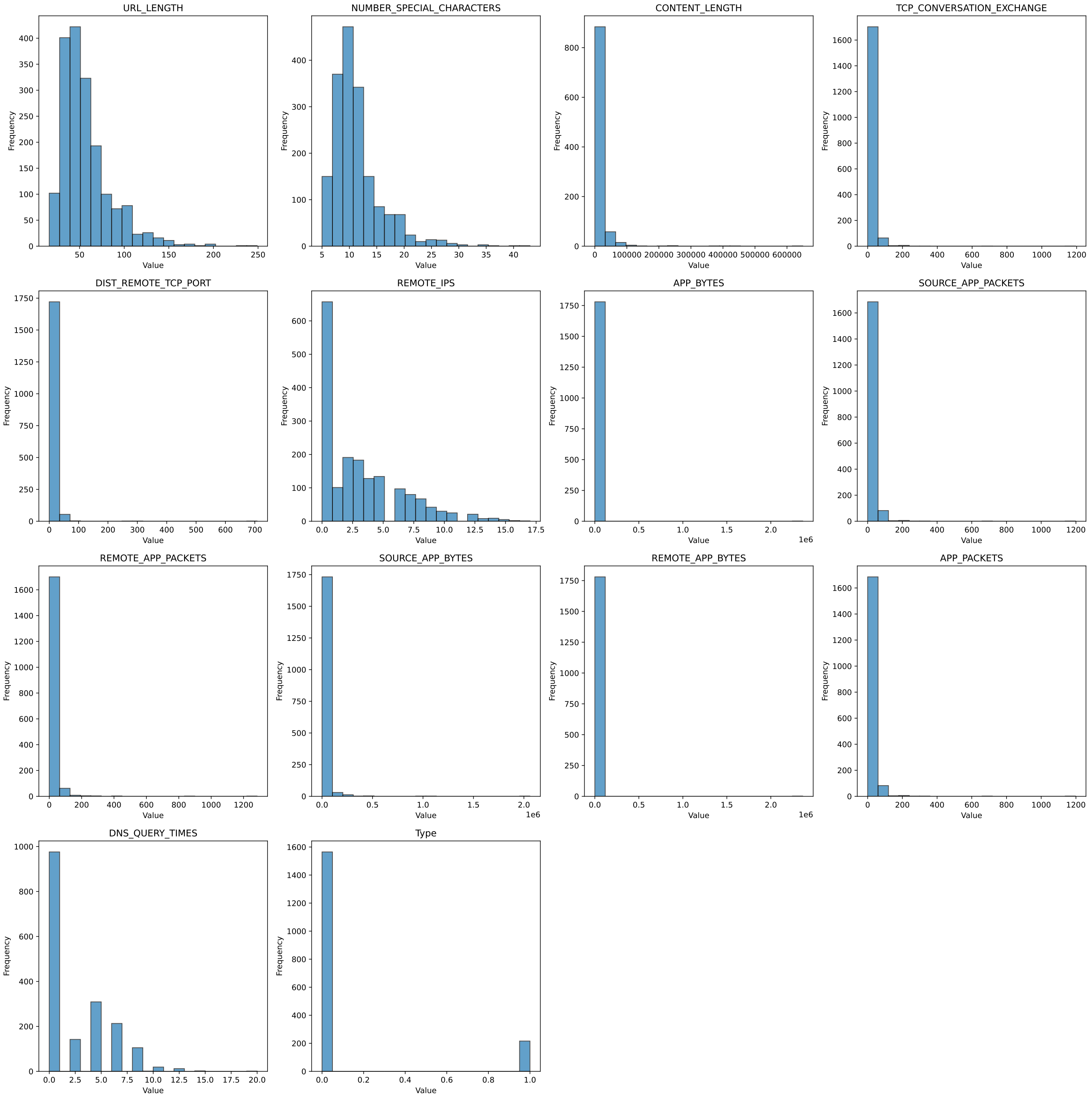


Figure 5. Histogram depicting the frequency of different data sets.

From the task involving data analysis with Power BI, there was a diverse distribution of scores amongst different subdomains as well as grades. There were not many variations in the other parameters of the dashboard. The fundamental structure for most of the companies involved in the analysis is machine learning. Thus, the following pre-processing tasks were examined before analysis. The analysis was conducted for 14 numeric columns and 7 object columns. Duplicate rows in any of the columns were avoided. Then, there were 812 null values in the content length column, which were filtered out. The outliers were treated via the quartile method. The correlations between variables were extracted via a correlation matrix. Finally, using SKlearn on the pre-processed data set, the optimum model for classification was determined.

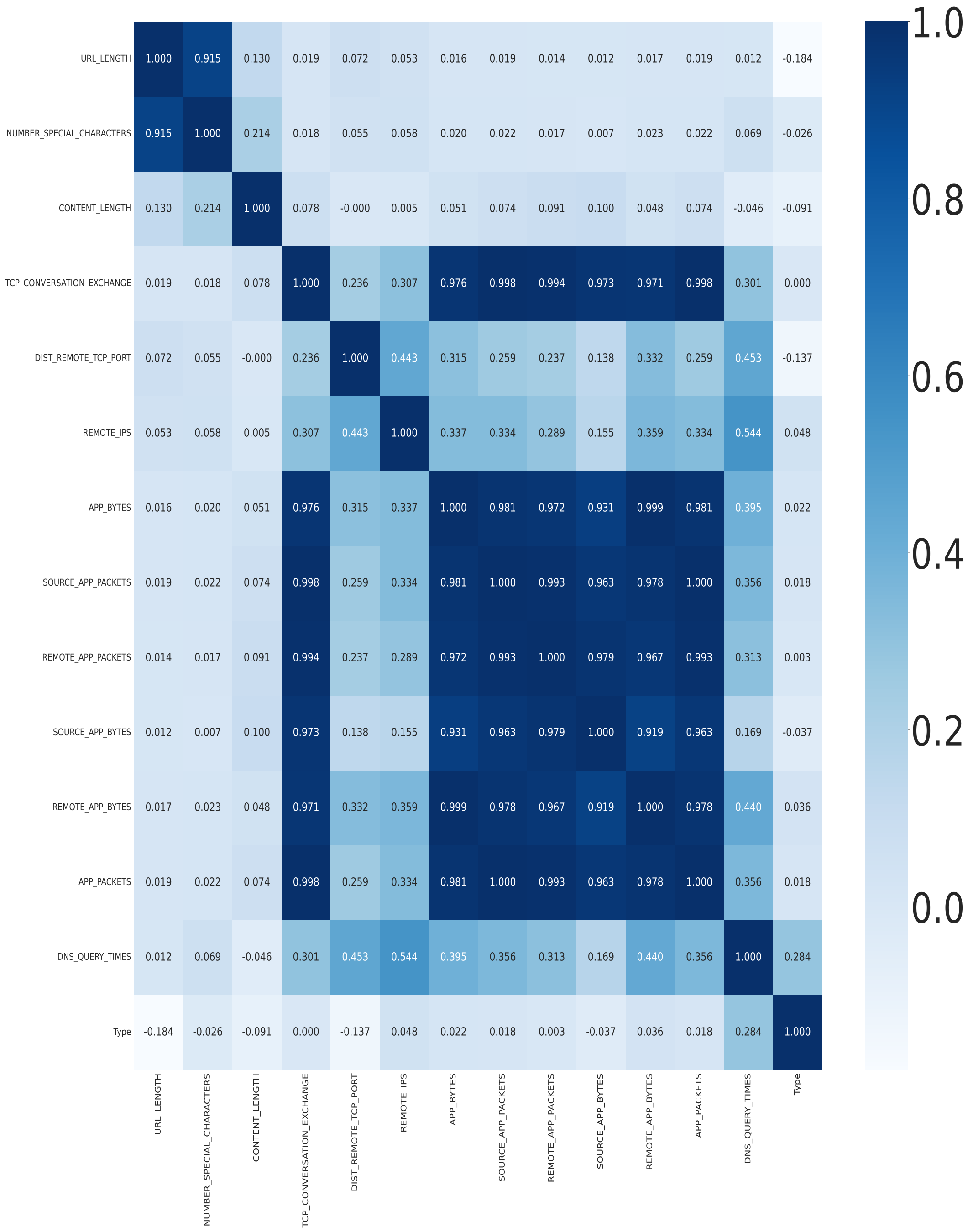


Figure 6. Correlation matrix between data.

Security data were collected from different domains and then used in different algorithms. The prediction model with the highest accuracy and the most realistically feasible model (keeping overfitting and underfitting in mind) were used for the analysis. The accuracy of the different machine learning algorithms is reported in Table 3. A comprehensive evaluation of various machine learning algorithms for cybersecurity is provided, highlighting their accuracy as a key performance metric. The results are indicative of the efficacy of these algorithms in addressing cybersecurity challenges. Among the algorithms examined, XGBoost and random forest stand out, with impressive accuracy rates of 95.87% and 95.53%, respectively. These algorithms have shown a high level of accuracy in identifying and mitigating cybersecurity threats, making them strong candidates for real-world applications. The decision tree also exhibited a competitive accuracy of 92.78%, underscoring its potential in cybersecurity tasks. On the other hand, the support vector machine and K-nearest neighbor (KNN) achieved accuracy rates of 88.31% and 89.34%, respectively, indicating their capability but with slight room for improvement. Logistic regression, Naive Bayes, and Scikit-learn neural networks achieved accuracy rates in the range of 83.84% to 91.40%, showing that they too have potential but might require further optimization to reach the levels of XGBoost and random forest. This comprehensive analysis provides valuable insights for selecting an appropriate machine-learning approach to enhance cybersecurity measures. In order to determine the prediction performance of the model, a confusion matrix was drawn, which is shown in Figure 7. The 249 predictions were predicted correctly as safe. Then, 18 malignant URLs were predicted as safe, and 8 safe URLs were predicted as malignant. Finally, 16 predictions were malignant. Therefore, the total accuracy (ratio of correct predictions to the total samples) reported by our final algorithm is 91.06%. These results underscore the significance of employing advanced machine-learning techniques in cybersecurity applications. XGBoost and random forest, with accuracy rates exceeding 95%, demonstrate robustness in identifying and countering cyber threats. The high accuracy of decision tree at 92.78% also positions it as a strong contender for cybersecurity tasks, while support vector machine and K-nearest neighbor show respectable accuracy rates of 88.31% and 89.34%; they could benefit from fine-tuning to further enhance their performance. Logistic regression, Naive Bayes, and Scikit-learn neural networks, with accuracy rates ranging from 83.84% to 91.40%, have demonstrated potential but may require parameter optimization and data preprocessing to fully harness their capabilities in the cybersecurity domain. These findings not only contribute to a deeper understanding of the strengths and limitations of various machine learning algorithms in cybersecurity but also provide practical insights for cybersecurity practitioners and researchers seeking to enhance their data analysis approaches. In conclusion, this research suggests that XGBoost, random forest, and decision tree are promising choices for cybersecurity tasks, given their high accuracy rates, but during the selection of the most suitable algorithm, one should consider the specific context and requirements of the cybersecurity application.

Table 3. Performance comparison of various machine learning algorithms in terms of accuracy.

Algorithm	Accuracy (%)
Logistic Regression	91.40
KNN	89.34
Decision Tree	92.78
Support Vector Machine	88.31
XGBoost	95.87
Sklearn Neural Network	83.84
Naive Bayes	84.53
Random Forest	95.53

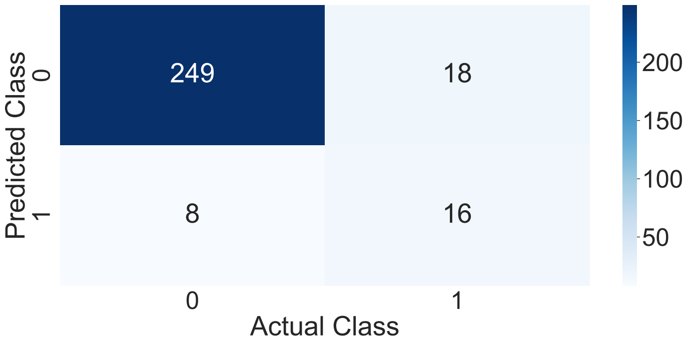


Figure 7. Confusion matrix of the proposed security system model.

6. Conclusions

Securing important data from an attacker is crucial nowadays. Major attacks can occur from clicking messages and URLs. This may lead to a big loss for corporations. To protect them from these attacks, there needs to be some precautions adopted. Initially, the validity of various URLs needs to be proven and security headers need to be found. The URL validity was assessed using Power BI to acquire insights into the data. The algorithms were run in a machine learning platform to predict the safeness of the URLs. In order to achieve the prediction of safe zones, a sequential process was carried out to analyze data. The collected data from various sources were cleaned and filtered using Microsoft Power BI. Machine learning algorithms were used to predict the accuracy safe zone. In the present study, different URL-based data sets were used to predict the best security system and analyze its performance. The logistic regression model was chosen for the analysis, with an accuracy of 91.40%, which demonstrates that it is a better prediction model for identifying security systems.

Author Contributions: Conceptualization: S.H.; methodology: S.H., and E.S.; software, E.S.; validation: S.H., and E.S.; formal analysis: A.J.P.; investigation: K.K.P.; data curation: S.P.S.; writing—original draft preparation: E.S.; writing—review and editing: A.J.P., K.K.P., K.N.V.P.S.R. and S.P.S.; supervision: P.P. and A.J.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset used in this study is public and available at this portal (https://www.kaggle.com/datasets/hassan06/nslkdd), accessed on 10 August 2023.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shar, L.K.; Tan, H.B.K. Defeating SQL injection. *Computer* **2012**, *46*, 69–77. [CrossRef]
2. Fang, Y.; Li, Y.; Liu, L.; Huang, C. DeepXSS: Cross site scripting detection based on deep learning. In Proceedings of the International Conference on Computing and Artificial Intelligence, Sanya, China, 21–23 December 2018.
3. Tsai, C.-W.; Lai, C.-F.; Chao, H.-C.; Vasilakos, A.V. Big data analytics: A survey. *J. Big Data* **2015**, *2*, 21. [CrossRef]
4. Rao, B.B.; Krishna, K.V.; Swathi, K. A Fast KNN Based Intrusion Detection System For Cloud Environment. *J. Adv. Res. Dyn. Control. Syst.* **2018**, *10*, 1509–1515.
5. Verma, A.; Ranga, V. Statistical analysis of CIDDS-001 dataset for network intrusion detection systems using distance-based machine learning. *Procedia Comput. Sci.* **2018**, *125*, 709–716. [CrossRef]
6. Belouch, M.; El Hadaj, S.; Idhammad, M. Performance evaluation of intrusion detection based on machine learning using Apache Spark. *Procedia Comput. Sci.* **2018**, *127*, 1–6. [CrossRef]
7. Khammassi, C.; Krichen, S. A GA-LR wrapper approach for feature selection in network intrusion detection. *Comput. Secur.* **2017**, *70*, 255–277. [CrossRef]
8. Farnaaz, N.; Jabbar, M. Random forest modeling for network intrusion detection system. *Procedia Comput. Sci.* **2016**, *89*, 213–217. [CrossRef]
9. Bhardwaj, M.; Alshehri, K.; Kaushik, K.; Alyamani, H.; Kumar, M. Secure framework against cyber-attacks on cyber-physical robotic systems. *J. Electron. Imaging* **2022**, *31*, 061802. [CrossRef]
10. Armbrust, M.; Fox, A.; Griffith, R.; Joseph, D.; Katz, R. *Above the Clouds: A Berkeley View of Cloud Computing*; Technical Report EECS-2009-28; University of California: Berkeley, CA, USA, 2009.
11. AlOmari, H.; Yaseen, Q.M.; Al-Betar, M.A. A Comparative Analysis of Machine Learning Algorithms for Android Malware Detection. *Procedia Comput. Sci.* **2023**, *220*, 763–768. [CrossRef]
12. Karajeh, H.; Maqableh, M.; Masa'deh, R. Privacy and security issues of cloud computing environment. In Proceedings of the 23rd IBIMA Conference, Valencia, Spain, 13–14 May 2020; pp. 1–15.
13. Jouini, M.; Rabai, L. A security framework for secure cloud computing environments. In *Cloud Security: Concepts, Methodologies, Tools, and Applications*; IGI Global: Hershey, PA, USA, 2019; pp. 249–263.
14. Mathrani, S.; Lai, X. Big data analytic framework for organizational leverage. *Appl. Sci.* **2021**, *11*, 2340. [CrossRef]
15. Joshi, N.; Kadhiwala, B. Big data security and privacy issues—A survey. In Proceedings of the 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, India, 21–22 April 2017; pp. 1–5.
16. Pedchenko, Y.; Ivanchenko, Y.; Ivanchenko, I.; Lozova, I.; Jancarczyk, D.; Sawicki, P. Analysis of modern cloud services to ensure cybersecurity. *Procedia Comput. Sci.* **2022**, *207*, 110–117. [CrossRef]

17. Ma, J.; Saul, L.K.; Savage, S.; Voelker, G.M. Beyond blacklists: Learning to detect malicious websites from suspicious URLs. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 28 June–1 July 2009; pp. 1245–1254.
18. Xu, L.; Zhan, Z.; Xu, S.; Ye, K. Cross-layer detection of malicious websites. In Proceedings of the ACM Conference on Data and Application Security and Privacy, San Antonio, TX, USA, 18–23 February 2013; pp. 141–152.
19. Wang, D.; Navathe, S.B.; Liu, L.; Irani, D.; Tamersoy, A.; Pu, C. Click traffic analysis of short URL spam on Twitter. In Proceedings of the IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, Austin, TX, USA, 20–23 October 2013; pp. 250–259.
20. Chiba, D.; Tobe, K.; Mori, T.; Goto, S. Detecting malicious websites by learning IP address features. In Proceedings of the IEEE/IPSJ International Symposium on Applications and the Internet, Izmir Turkey, 16–20 July 2012; pp. 29–39.
21. Cao, J.; Li, Q.; Ji, Y.; He, Y.; Guo, D. Detection of forwarding-based malicious URLs in online social networks. *Int. J. Parallel Program.* **2016**, *44*, 163–180. [CrossRef]
22. Marchal, S.; François, J.; State, R.; Engel, T. PhishStorm: Detecting phishing with streaming analytics. *IEEE Trans. Netw. Serv. Manag.* **2014**, *11*, 458–471. [CrossRef]
23. Choi, H.; Zhu, B.B.; Lee, H. Detecting malicious web links and identifying their attack types. In Proceedings of the 2nd USENIX Conference on Web Application Development (WebApps 11), Portland, OR, USA, 15–16 June 2011.
24. Huang, H.; Qian, L.; Wang, Y. A SVM-based technique to detect phishing URLs. *Inf. Technol. J.* **2012**, *11*, 921–925. [CrossRef]
25. Nepali, R.; Wang, Y.; Alshboul, Y. Detecting Malicious Short URLs on Twitter. In Proceedings of the 21st Americas Conference on Information Systems, Fajardo, Puerto Rico, 13–15 August 2015; pp. 1–6.
26. Canali, D.; Cova, M.; Vigna, G.; Kruegel, C. Prophiler: A fast filter for the large-scale detection of malicious web pages. In Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India, 28 March–1 April 2011; pp. 197–206.
27. Hemalatha, J.; Roseline, S.A.; Geetha, S.; Kadry, S.; Damaševičius, R. An efficient densenet-based deep learning model for malware detection. *Entropy* **2021**, *23*, 344. [CrossRef] [PubMed]
28. Ahsan, M.; Gomes, R.; Chowdhury, M.M.; Nygard, K.E. Enhancing machine learning prediction in cybersecurity using dynamic feature selector. *J. Cybersecur. Priv.* **2021**, *1*, 199–218. [CrossRef]
29. Saxe, J.; Berlin, K. eXpose: A character-level convolutional neural network with embeddings for detecting malicious URLs, file paths and registry keys. *arXiv* **2017**, arXiv:1702.08568.
30. Wang, H.H.; Yu, L.; Tian, S.W.; Peng, Y.F.; Pei, X.J. Bidirectional LSTM Malicious webpages detection algorithm based on convolutional neural network and independent recurrent neural network. *Appl. Intell.* **2019**, *49*, 3016–3026. [CrossRef]
31. Yang, W.; Zuo, W.; Cui, B. Detecting malicious URLs via a keyword-based convolutional gated-recurrent-unit neural network. *IEEE Access* **2019**, *7*, 29891–29900. [CrossRef]
32. Alani, M.M.; Awad, A.I. AdStop: Efficient flow-based mobile adware detection using machine learning. *Comput. Secur.* **2022**, *117*, 102718. [CrossRef]
33. Qabalin, M.K.; Naser, M.; Alkasassbeh, M. Android spyware detection using machine learning: A novel dataset. *Sensors* **2022**, *22*, 5765. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.