Journal of Business and Management Studies
ISSN: 2709-0876
DOI: 10.32996/jbms
Journal Homepage: www.al-kindipublisher.com/index.php/jbms

JBMS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Leveraging Advanced Machine Learning Algorithms for Enhanced Cyberattack Detection on U.S. Business Networks

Md Rasibul Islam[1] ✉ Md Nasiruddin[2], Mitu Karmakar[3], Rabeya Akter[4], MD Tushar Khan[5], Abdullah AL Sayeed[6] and Al Amin[7]

[12]*Department of Management Science and Quantitative Methods, Gannon University, USA*
[3]*School of Business, International American University, Los Angeles, California, USA.*
[4]*Master of science in information technology. Washington University of Science and Technology, USA*
[5]*Master of Science in Business Analytics, Trine University, USA*
[6]*Master of Business Administration in Project Management, Central Michigan University, USA*
[7]*Accounting and Information Systems, Jahangirnagar University, Dhaka, Bangladesh*
Corresponding Author: Md Rasibul Islam, **E-mail**: islam011@gannon.edu

| ABSTRACT

Cyberattacks' rising volume and sophistication have made conventional security measures, such as firewalls, signature-based intrusion detection systems, and antivirus software, increasingly inadequate. The upsurge of cyber threats has been one of the most pressing predicaments for U.S. organizations in the digital age. With the increased dependence on internet-based forums, cloud computing, and interconnected networks, companies face an advancing number of extreme cyberattacks. The chief objective of this research project is to design and deploy proven machine learning methods to enhance the detection and combating of cyberattacks on U.S. organization networks. This research project retrieved a cyber-attack dataset from Kaggle.com, which had a collection of public datasets of cyber threats. This dataset was curated precisely, offering a realistic representation of cyber-attack scenarios, making it an ideal playground for various analytical tasks. The collection was classified as per the source of the relevant information, such as host-based datasets, network traffic datasets, malware or fraud reports, or a special section for datasets that can be classified according to a specific source. The dataset comprised numerous network traffic attributes such as source and destination IP addresses, ports, protocol, payload size, and attack labels. For this research project, three machine learning algorithms were used, namely: Logistic Regression, XG-Boost and Random Forest. This research project applied performance metrics such as accuracy, precision, recall, and F1 score for the performance of the classification models were considered. The result illustrated that the random forest model was far superior in accuracy compared to the logistic regression model; particularly, it had excellent accuracy. Through the use of advanced machine learning models, organizations will be in a position to devise more dynamic and intelligent security systems that evolve with the threat landscape. These intelligent systems monitor every kind of anomaly, malicious activity, and threat response with unparalleled effectiveness. The findings of this research project have significant implications for enhancing cybersecurity in U.S. organizational networks.

| KEYWORDS

Cybersecurity, Machine Learning, Cyberattack Detection, U.S. Business Networks, Advanced Algorithms.

| ARTICLE INFORMATION

ACCEPTED: 02 October 2024 **PUBLISHED:** 26 October 2024 **DOI:** 10.32996/jbms.2024.6.5.23

1. Introduction

1.1 Background

According to Hasan (2022), the exponential escalation of cyber threats has been one of the most pressing predicaments for U.S. organizations in the digital age. With the increased dependence on internet-based forums, cloud computing, and interconnected networks, companies face an advancing number of complex cyberattacks. These sophisticated cyberattacks against organizations take the form of data breaches, ransomware, APTs, and zero-day exploits. In recent times, it has been reported that cyberattacks on US businesses have risen massively in frequency and ferocity, bringing losses amounting to millions of dollars, interference in operations, and reputations ruined. These developing threats have been hard to keep up with through conventional cybersecurity solutions because these attackers continuously seem to find ways around existing defenses. Buiya et al. (2024), contend that with cybercriminals continuing to adopt increasingly sophisticated methods, tactics, and techniques, the demand for proactive and adaptive measures to be implemented for the protection of critical business infrastructures has increased. This alarming trend shows the need for innovative approaches against cyberattacks, including the incorporation of advanced machine learning algorithms in cybersecurity frameworks.

1.2 Importance of Research

Khan et al. (2024), posit that the rising volume and sophistication of cyberattacks have made conventional security measures, such as firewalls, signature-based intrusion detection systems, and antivirus software, increasingly inadequate. Essentially, traditional methods lag in the detection of new sophisticated attacks that are constantly changing or that exploit unknown vulnerabilities. These vulnerabilities have brought into sharp focus an urgent need for enhanced cybersecurity measures to not only detect but, where possible, predict and mitigate such emerging threats in real-time. Since machine learning algorithms can process copious volumes of data, identify patterns therein, and adjust to new information constantly, the third pivotal application may well advance the art of business defense against cyberattacks. Shawon et al. (2024), assert that through the use of ML, organizations will be in a position to devise more dynamic and intelligent security systems that evolve with the threat landscape. These intelligent systems monitor every kind of anomaly, malicious activity, and threat response with unparalleled effectiveness. It also becomes critically important research to help U.S. businesses protect their networks from the growing menace of cyberattacks-essential for operational continuity and protection of sensitive data in ways that maintain customer trust in a rapidly digitizing economy.

1.3 Objectives

The prime objective of this study is to design and deploy proven machine learning methods to enhance the detection and combating of cyberattacks on U.S. organization networks. This operation includes such algorithmic designs that would tend to analyze a large volume of network data for surveillance over anomalous activities with high accuracy in differentiating between normal and malicious behavior. The research also aims to reduce false positives that have been a nagging nuisance with traditional cybersecurity solutions, leading to superfluous alerts and wasted resources. Regarding this aspect, the integration of advanced Machine Learning models will contribute to rendering the cybersecurity framework resilient against known threats and adapting to new, unseen attack vectors. Ultimately, the successful deployment of these machine learning techniques will deliver the tools to U.S. businesses needed to protect their networks against the ever-growing array of cyber threats.

2. Literature Review

2.1 Existing Work:

Almajed (2022), articulates that current cybersecurity techniques and technologies revolve around a wide array of solutions aimed at preventing, detecting, and responding to cyberattacks. Classic mechanisms of defense include firewalls, antivirus software, intrusion detection systems, and intrusion prevention systems; traditionally these have used signature-based techniques, which unfortunately detect only known threats. The firewalls block incoming and outgoing traffic, respectively, according to predefined rules. Antivirus software searches for malware by finding the signature of known malware, while IDS/IPS uses network activity monitoring and identifies suspicious behavior by finding a pattern of predefined rules.

Behiry & Aly (2024), argues that besides signature-based methods, heuristic and behavioral analysis techniques have also emerged as effective means to identify complex-level threats. These approaches focus more on the file and network traffic behavior analysis and on the detection of various anomalies that may point to a potential threat. For example, anomaly detection systems compare any current network activity against historical baselines to flag deviations that could be malicious. Machine learning and AI have been some of the latest tools in the cybersecurity domain. With ML algorithms, large volumes of such datasets can be treated, and patterns identified to predict cyber threats based on past behavior. These integrated techniques include deep learning, reinforcement learning, and unsupervised learning applied to cybersecurity tools to enhance real-time detection and response capabilities.

Furthermore, Vaiyapuri et al.(2024), posit that endpoint detection and response (EDR) technologies have evolved to offer greater visibility into network endpoints. EDR solutions continuously monitor devices for suspicious activities and provide forensic analysis necessary for understanding the scope of potential breaches. Cloud-based security solutions leverage the scalability of cloud infrastructures and are increasingly set up to protect businesses from DDoS attacks and provide secure access in remote work environments. These tools also often incorporate automated threat intelligence feeds, which allow systems to be updated in real time based on newly identified threats across the globe.

2.2 Gaps and Challenges:

Considering all the inventions and innovations within cybersecurity technologies, there are still several gaps and challenges that persist. One noteworthy limitation of conventional signature-based defenses is their overreliance on known threat patterns. These traditional systems are ill-equipped to detect zero-day attacks or polymorphic malware, where attackers frequently modify their tactics to avoid detection [Dixit et al. 2021]. In that case, the system won't identify the threat because it simply does not match any predefined signature or rule or leaves the breach for newly appearing and morphing attack vectors.

Zeeshan et al. [2024], hold that another significant challenge is that anomaly detection systems frequently have varying accuracy. While they tend to be pretty good at recognizing uncommon patterns, they tend to be prone to high false-positive rates. Normal network activities that avoid historical baselines by even a little propagate systems to fire alerts, which overwhelm cybersecurity teams and could thereby result in alert fatigue. This makes prioritizing actual threats much more difficult, thus slowing down the response. Besides that, because of the imbalanced dataset problem, where the number of normal instances largely dominates the malicious ones, machine learning models applied in cybersecurity currently face challenges. Such skewed models usually have poor detection performance on minority classes, which are typically made of rare but critical cyberattacks.

Besides that, the dynamic and adversarial nature continuing within the cyberspace threat landscape creates a challenge with every passing moment. The threats keep changing and adapting to bypass the security mechanisms, so it is tough for the static defense mechanisms to keep up the pace. Besides, APTs mainly invite stealthy attackers that remain in the network for a pretty long period, which complicates the detection challenge [Zhou et al. 2018]. But while machine learning models are very promising in this area, these models are not immune to adversarial-type attacks in which small, intentional modifications to the input data can deceive the model into making a wrong classification.

Equally important, Delplace [2020], argues that there is a persisting challenge of integrating and consolidating cybersecurity tools in a complex, diverse network. Most of the time, various organizations use commoditized, uncoordinated security tools operating separately from each other, which leads to fragmented visibility and delayed response times for incidents. The need for a more holistic and automated way in which one views cybersecurity-mashup real-time data and communicates cross-system-ensures these gaps are met and a resilient defense system is achieved.

3. Dataset Description

This research project retrieved a cyber-attack dataset from Kaggle.com, which had a collection of public datasets of cyber threats. This dataset was curated precisely, offering a realistic representation of cyber-attack scenarios, making it an ideal playground for various analytical tasks. The collection was classified as per the source of the relevant information, such as host-based datasets, network traffic datasets, malware or fraud reports, or a special section for datasets that can be classified according to a specific source. The dataset comprised numerous network traffic attributes such as source and destination IP addresses, ports, protocol, payload size, and attack labels [Pro-AI-Robikul, 2024]. The chief objective was to accurately classify the network traffic into attack groups using algorithms like Logistic Regression and Random Forest.

Table 1: Exhibits Key Attributes and Features

S/No	Key Attribute/ Feature	Description
1.	Attack ID	Unique Identifier for each attack.
2.	Source Country, Destination Country	Nations correlated with the IP addresses.
3.	Source IP, Destination Country	Countries associated with the IP addresses.
4.	Source Port, Destination Port	Port numbers associated with the connection.
5.	Protocol	Network protocol (TCP, UDP, etc.).
6.	Attack type	The specific type of cyber-attack.
7.	Payload Size (bytes):	Size of the data packet involved in the attack.
8.	Detection Label	Indicates whether the traffic was detected as an attack.
9.	Confidence Score	Probability score of the detection.
10.	Affected System	The System impacted by the attack.
11.	ML Model	The machine algorithm is used for the detection.
12.	Time Stamp	Time at which the scenario occurred.
13.	Port Type	Type of Port utilized in the communication.

3.1 Data Pre-Processing and Cleaning

Step 1: Missing Values Check- The count of missing values in each column was computed for Data.Frame df. An appropriate method identified missing values in the data frame; sum() then counted the occurrences of such values in each column.

Step 2: Dropping unimportant columns- with the computed data frame [df], the relevant columns list entailed the column names that were considered essential to the analysis or project. This was based on the project requirements and was curated as such. Further code snippets formed a new data frame consisting only of the columns listed within the relevant columns lists, thus dropping whatever columns were not pegged within the list.

Step 3: Encoding Categorical Variables - An ideal code snippet was executed for the data transformation to encode categorical variables into numerical labels and standardize the numeric features. This step was instrumental in designing and affirming a reliable machine-learning model.

4. Methodology

4.1 Data Analysis

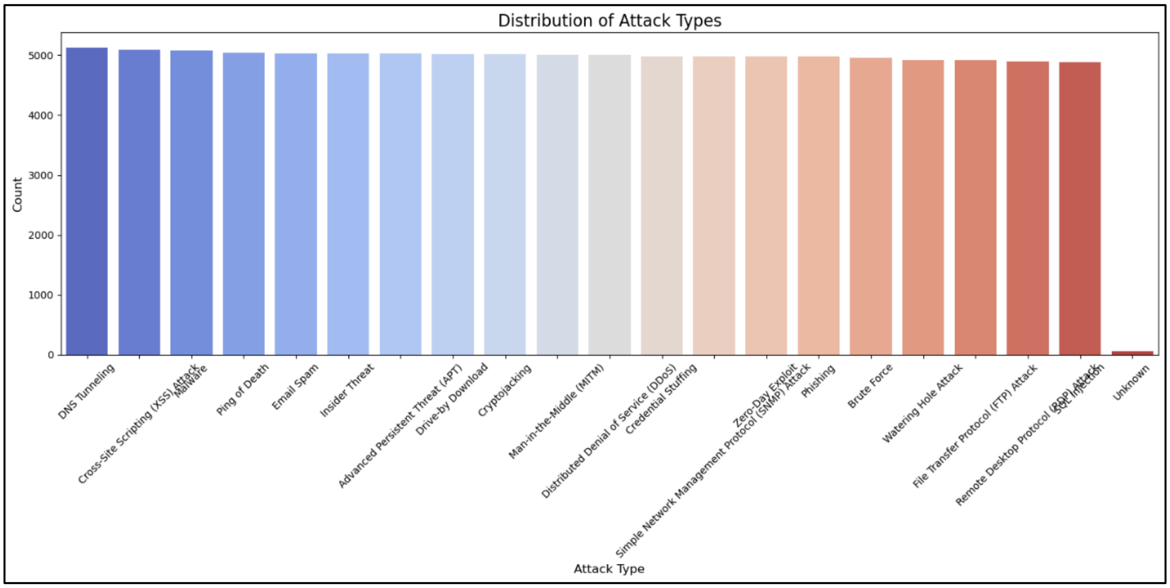


Figure 1: Showcases Distribution of Attack Types

The bar chart above provides a visual representation of the frequency of various cyberattacks. The chart shows DNS Tunneling, Cross-Site Scripting (XSS), and Ping of Death as the most frequent attacks. Other notable attacks listed on this chart include Email Spam and Insider Threat. Other noteworthy cyber-attacks entailed APT, Drive-by Download, Cryptojacking, and Man-in-the-Middle. DDoS attacks, Credential Stuffing, and Zero-Day Exploits also appear on the list. Of more interest was to note that the trends showing fewer attacks are those of the simple Network Management Protocol-SNMP attacks, Phishing, Brute Force, Watering Hole Attacks, FTP Attacks, RDP attacks, and Unknown attacks.

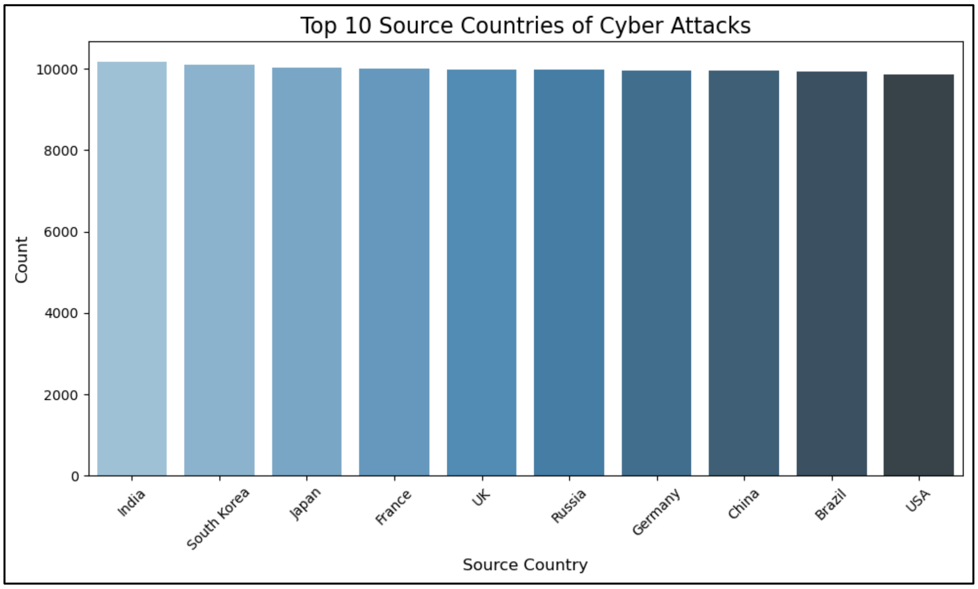


Figure 2: Displays Top 10 Source Countries of Cyber Attacks

The bar chart above visually represents the frequency of cyberattacks originating from different countries. From the analysis, it was evident that India topped the list as a source of cyber-attacks, while South Korea and Japan came second and third, respectively. Other sources of cyber-attacks include France, the UK, Russia, Germany, China, Brazil, and the USA. While the chart pegs India as a source for most cyber threats, this could depend on many factors: from simply having more Internet access, to cybersecurity infrastructures in place, to how much is indeed reported in each country.

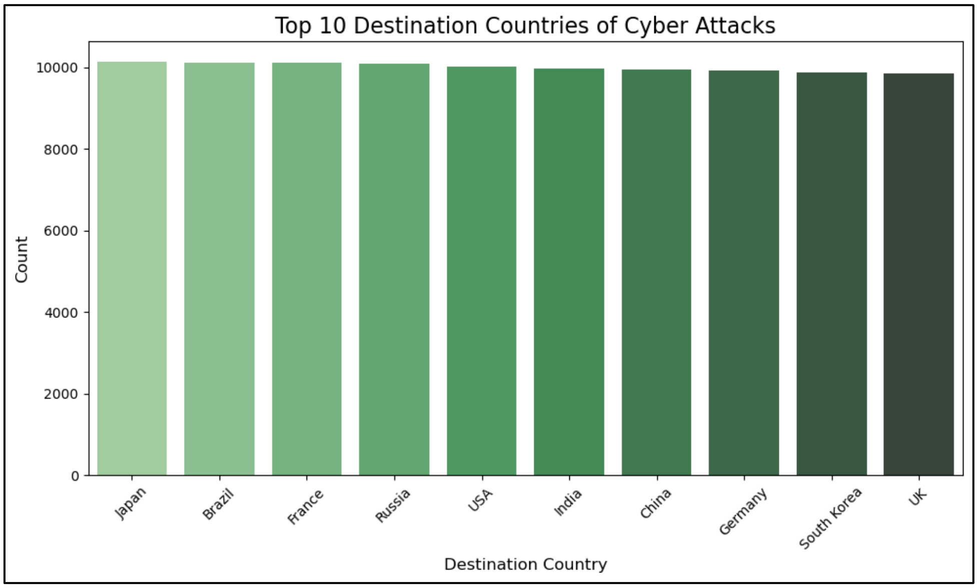


Figure 3: Depicts Top 10 Destination Countries of Cyber Attacks

The bar chart above shows the destination of various cyberattacks. On the x-axis, it enumerates the destination countries, and on the y-axis, it gives the count of the attacks. From this chart, one observes that Japan is the most targeted country, followed by Brazil and France. Other notable destinations include Russia, the USA, India, China, Germany, South Korea, and the UK. Although the leading target of cyber threats is depicted to be Japan, various factors may be considered that could influence the appeal of different countries to any cyber attackers; some of these factors are economic development technological infrastructure, and geopolitical considerations.

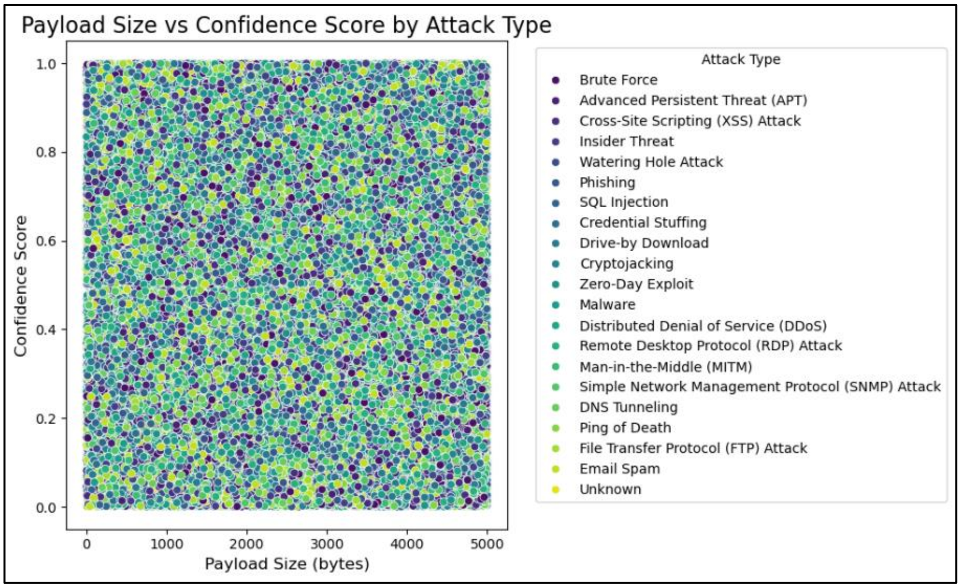


Figure 4: Portrays Payload Size vs. Confidence Score by Attack Type

The scatter plot presents a relationship between payload size and confidence score for different cyberattacks. The x-axis is the payload size in bytes, and the y-axis is the confidence score. Each dot on this plot corresponds to some instance of an attack, with a colour corresponding to an attack type. The plot indicates indeed that there is a general trend of having a larger payload size for higher confidence scores, thereby indicating that attacks with more data might be easier to detect and attribute. On the other hand, however, multiple examples of smaller payload sizes with high confidence scores are present, indicating that not all of these attacks depend on large volumes of data. Moreover, the plot shows that different attack types are characterized by different patterns in both payload size and confidence score. This serves to show how diverse cyber threats can be, and how detection and response mechanisms can be just as important and tailored.

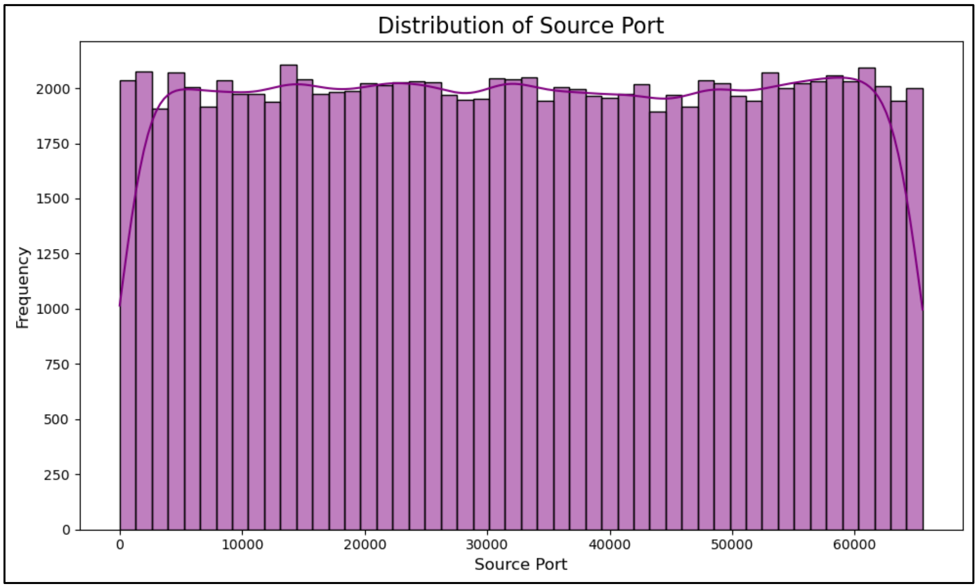


Figure 5: Exhibits the Distribution of Source Port

The histogram above displays how frequently certain source ports appear in the data. The x-axis shows the number of the source port, and the y-axis shows the frequency of the port. The graph showcases the clear distribution in the plot peak around the range from 10,000 to 20,000 and gradually declines when the port number increases. This pattern is supported by the superimposed density curve, which indicates a normal distribution with a slight skew toward higher values. In sum, this chart underlines source port usage concentration in the low ranges and enlightens the current network communication pattern.

4.2 Model Development

For this research project, three machine learning algorithms were used, namely: Logistic Regression, XG-Boost and Random Forest. Logistic regression is a type of statistical model that creates the possibility for one or more predictor variables of either of the two possible outcomes. This has to be applied in case simplicity along with interpretability is wanted. Another ensemble learning model that was used is Random Forest which fits a huge set of trees and then returns the mode for predictions independently provided by the decision trees. XG-Boost is an ensemble technique using the gradient boosting framework. It is popular due to its speed and high performance right out of the box for many problems, specifically on sparse data, apart from regularization; it prevents the overfitting problem [Pro-AI-Robikul, 2024]. Model development began by importing necessary libraries for model selection, and evaluation metrics. The dataset was preprocessed and ready, the code split the data into features and target variables. An 80/20 train-test split was performed to separate the data for training and testing.

4.3 Performance Metrics

This research project applied performance metrics such as accuracy, precision, recall, and F1 score for the performance of the classification models were considered. Accuracy denotes the number of correctly predicted instances on the total number of cases. Precision is the true positive among the total positives expected and measures how well the model can avoid false positives. Recall is the ratio of true positives to actual positives, showing how much interest rate the model has been able to catch. The F1 score is the harmonic mean of precision and recall, balancing them [Pro-AI-Robikul, 2024].

5. Implementation

Logistic Regression

```
# Required Libraries
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# Assuming df is already preprocessed and ready for modeling
# Splitting the dataset into features (X) and target (y)
X = df.drop('Detection Label', axis=1)  # Features (drop the target column)
y = df['Detection Label']  # Target

# Train-Test Split (80% training, 20% testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, 
random_state=42)

# Logistic Regression Model
logistic_model = LogisticRegression(max_iter=1000)
logistic_model.fit(X_train, y_train)

# Predictions
y_pred = logistic_model.predict(X_test)

# Evaluation Metrics
print("Logistic Regression Accuracy:", accuracy_score(y_test, y_pred))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

Table 2: Showcases the Logistic Regression Modeling

This Python snippet shows the implementation of a logistic regression model for binary classification. It starts by importing necessary libraries for model selection, linear models, and evaluation metrics. The dataset had been preprocessed and was ready; the code split it into features and target variables. Further, it conducts an 80/20 train-test split to divorce the data for training and testing. This is followed by instantiating the logistic regression model and training it with the training data. Then, use the model to make predictions on the test data and calculate the accuracy; other evaluation metrics that may be used include a confusion matrix and classification report, which would perform the model. Therefore, this code provided the roadmap for logistic regression model development and testing for any binary classification task.

Output:

```
Classification Report:
               precision    recall  f1-score   support

           0       0.81      0.83      0.82     10025
           1       0.82      0.80      0.81      9957
           2       0.00      0.00      0.00        18

    accuracy                           0.81     20000
   macro avg       0.54      0.54      0.54     20000
weighted avg       0.81      0.81      0.81     20000
```

Table 3: Depicts Logistic Regression Classification Report

This classification report represents the performance of a binary classification model with three classes, namely 0, 1, and 2. Class 2 seems to be a rare case because there are only 18 samples, while there are approximately 10,000 samples for classes 0 and 1. The model performs similarly well for classes 0 and 1, with precision and recall scores of about 0.81-0.83, thus yielding balanced f1-scores of 0.82 and 0.81, respectively. However, it completely fails to predict class 2, showing 0.00 across all metrics, likely due to the severe class imbalance. Overall model accuracy is 0.81, with a weighted average matching this score also, though the macro average would treat all classes with equal importance regardless of size much lower at 0.54, thereby highlighting the poor performance of the minority class.

Random Forest

```
# Required Libraries
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# Random Forest Model
random_forest_model = RandomForestClassifier(n_estimators=100, random_state=42)
random_forest_model.fit(X_train, y_train)
# Predictions
y_pred_rf = random_forest_model.predict(X_test)

# Evaluation Metrics
print("Random Forest Accuracy:", accuracy_score(y_test, y_pred_rf))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred_rf))
print("\nClassification Report:\n", classification_report(y_test, y_pred_rf))
```

Table 4: Portrays the Random Forest Modelling

The above code snippet is the implementation of a random forest classifier machine learning model in code, using sci-kit-learn. It initiates a model with 100 decision trees, having n_estimators=100, and a random state of 42 for reproducibility. After importing all the needed libraries, the code fits the model to some training data, X_train, and y_train; it then makes some predictions on the test set, X_test. Its performance is evaluated according to its accuracy score, confusion matrix, and a detailed classification report.

Output:

```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     10025
           1       1.00      1.00      1.00      9957
           2       0.00      0.00      0.00        18

    accuracy                           1.00     20000
   macro avg       0.67      0.67      0.67     20000
weighted avg       1.00      1.00      1.00     20000
```

Table 5: Displays the Random Forest Classification report

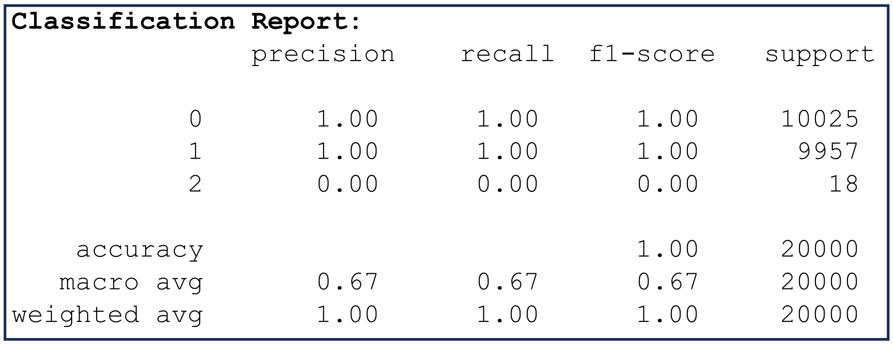
This classification report shows great performance on the majority classes 0 and 1, perfectly seated at 1.00 for precision, recall, and f1-score, showing that the model has classified all instances of these classes correctly and thus contains 10,025 and 9,957 samples, respectively. However, as is often seen, the model fails in predicting class 2, which only contains 18 samples, hence recording 0.00 across all metrics due to extreme class imbalance. In this model, the overall accuracy is 1.00, and the weighted average is 1.00, as one would expect from dominant performance in the majority classes, but the macro average is 0.67, lower due to the poor performance in class 2. This signals that while for the main classes, the model works very well, the performance of the minority class is rather bad. The latter may need some care in terms of, for example, oversampling or class weighting in the case when the use case requires decent performance on class 2.

XGboost

```
# Required Libraries
from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# XGBoost Model
xgb_model = XGBClassifier(use_label_encoder=False, eval_metric='logloss',
random_state=42)
xgb_model.fit(X_train, y_train)

# Predictions
y_pred_xgb = xgb_model.predict(X_test)

# Evaluation Metrics
print("XGBoost Accuracy:", accuracy_score(y_test, y_pred_xgb))
print("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred_xgb))
print("\nClassification Report:\n", classification_report(y_test, y_pred_xgb))
```

Table 6: Showcases the XG-Boost Modelling

The code above executes an XGBoost Classifier, another powerful approach for ensemble learning. The model is initiated with some specific parameter values: use_label_encoder is set to False, to avoid deprecation warnings; eval_metric is 'logloss', to set logarithmic loss as the evaluation metric; and random_state is set to 42 for reproducibility. After importing XGBoost and sci-kit-learn metrics, the rest of the code is the same: fit the model on the training data (X_train, y_train), predict on the test set (X_test), and evaluate the performance of the model using accuracy score, confusion matrix, and classification report. XGBoost has a reputation for being very performant and fast. In many cases, it outperforms traditional algorithms on structured or tabular data.



```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     10025
           1       1.00      1.00      1.00      9957
           2       0.00      0.00      0.00        18

    accuracy                           1.00     20000
   macro avg       0.67      0.67      0.67     20000
weighted avg       1.00      1.00      1.00     20000
```

Table 7: Presents XG-Boost Classification Report

This classification report gives some performance metrics of the XG-Boost model. While the model is perfectly doing great, yielding a precision, recall, and F1-score of 1.00 on classes 0 and 1, on class 2 it yields precision, recall, and F1-score all 0.00, hence not being able to classify instances of that class. The macro average produces an overall accuracy value of the model at 0.67, where it classifies 67% correctly. Weighted average gives a perfect score of 1.00 on all metrics of precision, recall, and F1-score since it considers class imbalance. Finally, the Support column shows the number of instances for each class in the test set: 10,025 instances for class 0, 9,957 instances for class 1, and 18 for class 2. These results hint that most likely, the model is doing great on the majority classes but poor concerning the minority class, which may be an area for further investigation and improvement.

6. Results and Analysis

Table 8: Showcases Models Performance Summary

Performance Metric	Random Forest	Logistic Regression
Accuracy	99.90%	81.39%
Precision [class 0]	1.00	0.81
Precision [class 1]	1.00	0.82
Precision [class 2]	0.00	0.00
Recall [class 0]	1.00	0.83
Recall [class 1]	1.00	0.80
Recall [class 2]	0.00	0.00

6.1 Comparative Analysis

The result illustrated that the random forest model was far superior in accuracy compared to the logistic regression model; particularly, it had an amazing accuracy of 99.90%, while the accuracy compared to 81.39% for the Logistic Regression. Precision and recall for class 0 are perfectly reflected by both models, concluding that the perfect identification of class 0 is possibly done by these models along with the avoidance of false positives. However, class 2 was tricky for both algorithms, as there the precision and recall were 0.00, indicating their inability to classify correctly or even to avoid false positives for this class. Overall, the comparison underlines that, regarding this particular classification task, the random forest model performs better, probably because it manages to grasp complex relationships and non-linear interactions within the data.

7. Discussion

7.1 Implications of the Findings

The findings of this research project have significant implications for enhancing cybersecurity in U.S. organizational networks. Firstly, the prevalence of DNS Tunneling, Cross-Site Scripting (XSS), and Ping of Death attacks establishes a dire need for appropriate network security measures that prevent these common threats. Firewalls, intrusion detection systems, and web application firewalls can thus be installed to cut down the chances of those risks occurring. The identification of India as one of the major sources of attack origination points to the need for collaboration between the different countries and the sharing of information about global cyber threats. Thirdly, U.S. businesses should work in concert with international partners to develop and implement advanced algorithms such as the Random Forests in cybersecurity. Interestingly, attack destinations indicate that Japan is a frequent target, meaning that organizations in Japan may have special security measures to defend against targeted attacks.

Furthermore, the findings on payload size and confidence scores are indicative of supporting advanced security detection technologies that correctly identify and prioritize potential threats before declaring their characteristics. Organizations in the U.S. should consider applying machine learning and AI analytic solutions for monitoring network traffic to detect anomaly-based traffic showing signs of malicious activity. It is also indicative of the distribution of the source ports, which may suggest that some port ranges might be more susceptible to attacks. Segmentation of the network and access control are other approaches organizations can take to limit traffic to only necessary or approved ports to minimize the attack surface.

7.2 Study's Limitations

Even though the research findings on the aspect of cyber threats and their implications are of great value, some limitations have to be put into consideration. First, the data used might not be comprehensive enough to capture all the cyber-attacks occurring in business networks in the U.S. Second, though this study focused on specific attack types and source/destination countries, other emerging threats or regions may not be included. Thirdly, it does not probe into the detailed methodologies of the attackers, which would add even more value to the development of specific countermeasures.

To combat these limitations, future studies could consider expanding the dataset to include a wider range of cyberattacks and regions. Analyzing in detail the specific techniques attackers employ may provide valuable information for the development of better security measures. Finally, developing real-time threat intelligence and threat-hunting techniques may contribute to improving its threat detection and response against upcoming threats.

7.3 Future Directions

Based on the findings from the research project, different directions of further research in different cybersecurity measures are open for consideration. Advanced threat detection technologies can be developed with accuracy in threat identification and prioritization in real-time. This may look at research using Advanced algorithms such as XG-Boost or SVM in consolidation with natural language processing to analyze network traffic, find anomalies, and allow correlation of indicators of compromise.

Another line of research can be conducted on how different time series algorithms such as ARIMA are effective in countering some sort of attack. This may involve a series of controlled experiments or simulations with the applications of various security technologies: firewalls, intrusion detection systems, or encryption techniques. Further, researchers may work on the development of new security technologies, such as quantum cryptography, given the threats and vulnerabilities emerging.

Lastly, the research into cyber attackers' behavioral analysis can provide significant insights useful in developing effective countermeasures. This would cover the analysis of patterns of attack to find common tactics and motives of the attackers. Deeper insights into the threats and their performers will lead organizations to better-targeted and more effective security.

8. Conclusion

The prime objective of this study was to design and deploy proven machine learning methods to enhance the detection and combating of cyberattacks on U.S. organization networks. This research project retrieved a cyber-attack dataset from Kaggle.com, which had a collection of public datasets of cyber threats. This dataset was curated with precision, offering a realistic representation of cyber-attack scenarios, making it an ideal playground for various analytical tasks. The collection was classified as per the source of the relevant information, such as host-based datasets, network traffic datasets, malware or fraud reports, or a special section for datasets that can be classified according to a specific source. The dataset comprised a myriad of network traffic attributes such as source and destination IP addresses, ports, protocol, payload size, and attack labels. For this research project, three machine learning algorithms were used, namely: Logistic Regression, XG-Boost and Random Forest. This research project applied performance metrics such as accuracy, precision, recall, and F1 score for the performance of the classification models were considered. The result illustrated that the random forest model was far superior in accuracy compared to the logistic regression model; particularly, it had excellent accuracy. The findings of this research project have significant implications for enhancing cybersecurity in U.S. organizational networks.

References

[1] Almajed, R, Amer I, Abedallah Z A, Nahia M, and Faris A. A. (2022) Using machine learning algorithm for detection of cyber-attacks in cyber-physical systems. *Periodicals of Engineering and Natural Sciences* 10, no. 3 (2022): 261-275.

[2] Behiry, M. H., & Aly, M. (2024). Cyberattack detection in wireless sensor networks using a hybrid feature reduction technique with AI and machine learning methods. *Journal of Big Data*, *11*(1), 16.

[3] Buiya, M. R., Laskar, A. N., Islam, M. R., Sawalmeh, S. K. S., Roy, M. S. R. C., Roy, R. E. R. S., & Sumsuzoha, M. (2024). Detecting IoT Cyberattacks: Advanced Machine Learning Models for Enhanced Security in Network Traffic. Journal of Computer Science and Technology Studies, 6(4), 142-152.

[4] Dutta, V., Choraś, M., Pawlicki, M., & Kozik, R. (2020). A deep learning ensemble for network anomaly and cyber-attack detection. Sensors, 20(16), 4583.

[5] Dixit, P., Kohli, R., Acevedo-Duque, A., Gonzalez-Diaz, R. R., & Jhaveri, R. H. (2021). Comparing and analyzing applications of intelligent techniques in cyberattack detection. *Security and Communication Networks*, *2021*(1), 5561816.

[6] Delplace, A., Hermoso, S., & Anandita, K. (2020). Cyber attack detection thanks to machine learning algorithms. *arXiv preprint arXiv:2001.06309*.

[7] Hasan, M. R. (2022). Cybercrime Techniques in Online Banking. *Journal of Aquatic Science.* Retrieved from https://www.journal-aquaticscience.com/article_158883.html.

[8] Khan, M. A., Debnath, P., Al Sayeed, A., Sumon, M. F. I., Rahman, A., Khan, M. T., & Pant, L. (2024). Explainable AI and Machine Learning Model for California House Price Predictions: Intelligent Model for Homebuyers and Policymakers. Journal of Business and Management Studies, 6(5), 73-84.

[9] Pro-AI-Rokibul. (2024). *Detection-Of-CyberAttack-On-Different-Countries-With-Advanced-Machine-Learning-Algorithms/Model/main.ipynb at main · proAIrokibul/Detection-Of-CyberAttack-On-Different-Countries-With-Advanced-Machine-Learning-Algorithms*. GitHub. https://github.com/proAIrokibul/Detection-Of-CyberAttack-On-Different-Countries-With-Advanced-Machine-Learning-Algorithms/blob/main/Model/main.ipynb

[10] Shawon, R. E. R., Rahman, A., Islam, M. R., Debnath, P., Sumon, M. F. I., Khan, M. A., & Miah, M. N. I. (2024). AI-Driven Predictive Modeling of US Economic Trends: Insights and Innovations. Journal of Humanities and Social Sciences Studies, 6(10), 01-15.

[11] Vaiyapuri, T., Shankar, K., Rajendran, S., Kumar, S., Gaur, V., Gupta, D., & Alharbi, M. (2024). Automated cyberattack detection using optimal ensemble deep learning model. *Transactions on Emerging Telecommunications Technologies*, *35*(4), e4899.

[12] Zeeshan, M. A. F., Sumsuzoha, M., Chowdhury, F. R., Buiya, M. R., Mohaimin, M. R., Pant, L., & Shawon, R. E. R. (2024). Artificial Intelligence in Socioeconomic Research: Identifying Key Drivers of Unemployment Inequality in the US. Journal of Economics, Finance and Accounting Studies, 6(5), 54-65.

[13] Zhou, Y, Meng H, Liyuan L, Jing S H, and Yan W. (2018) Deep learning approach for cyberattack detection. In *IEEE INFOCOM 2018-IEEE conference on computer communications workshops (INFOCOM WKSHPS)*, 262-267. IEEE, 2018.