International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 6, pp 185-194, June-2023

https://doi.org/10.47001/IRJIET/2023.706029

Enhancing Cloud Banking Security: Detection of Distributed Web Attacks through Random Forest Algorithm

[1]*Rafal Sattar Jabbar, [2]Mohamad Tawfik Hamze

[1,2]Computer Science, Faculty of Science & Literature, American University of Culture and Education, Beirut, Lebanon

***Abstract* - The advent of the digital era has brought unprecedented convenience and efficiency to the banking sector. However, it has also exposed financial institutions to a multitude of cyber threats. In particular, the increasing value of banking information in the digital realm has made it an attractive target for malicious actors. The repercussions of successful hacking attempts on these systems can be severe, ranging from financial losses to compromised customer data and erosion of trust in the banking sector. Consequently, bolstering the security of banking systems has become a paramount concern. This paper undertakes a comprehensive analysis of the security landscape surrounding financial organizations by leveraging a bank dataset comprising 15,000 samples and 38 variables. Through rigorous data analysis techniques, the dataset is utilized to train a Random Forest algorithm, which is then employed to evaluate and identify Distributed Denial-of-Service (DDoS) attacks launched against financial institutions. The results of this study are highly promising, as the Random Forest algorithm achieves an impressive accuracy rate of 99% in identifying potential security flaws. By providing valuable insights and empirical evidence, this research contributes to the existing body of knowledge in the field of cyber security, specifically concerning the detection and prevention of DDoS attacks in financial organizations.**

Keywords: Cloud Banking, Machine Learning, Random Forest, cybercriminals, Distributed Denial-of-Service.

1. Introduction

With the increasing adoption of cloud computing in the banking sector, ensuring robust security measures has become a critical concern. Cloud banking offers numerous advantages such as cost-efficiency, scalability, and improved service delivery. However, the shared nature of cloud infrastructure also introduces potential vulnerabilities that can be exploited by malicious actors, leading to distributed web attacks. These attacks, such as distributed denial of service (DDoS) attacks, can disrupt the availability and integrity of online banking services, resulting in financial losses and reputational damage for financial institutions.

To mitigate the risks associated with distributed web attacks in cloud banking, effective detection mechanisms are essential. Traditional security measures, such as firewalls and intrusion detection systems, are insufficient in handling the dynamic nature and scale of these attacks. Therefore, the development of advanced detection techniques leveraging machine learning algorithms has gained significant attention.

Data pertaining to distributed denial of service (DDoS) attacks was collected from an open-source website on February 2, 2021. The raw datasets were retrieved from an openly available online database. Pre-processing methodologies were employed to mitigate inherent challenges associated with the dataset. Outliers were identified and removed, followed by the application of suitable statistical techniques to achieve dataset balance. Subsequently, relevant characteristics were selected for utilization. The dataset was then partitioned into a training set and a test set, wherein the former was employed for training machine learning models and the latter for accuracy assessment.

In the areas of online retail and financial services, the following is a list of some of the most prominent uses 2 of cloud computing:

- Banks that Allow Public Access to Their Online Banking Platforms. Some of these services, products, and information may be offered by the bank, while others may be given by third parties, and all of them can be accessed through the next generation of commercial internet banking apps.
- These new retail-oriented Internet banking solutions will be powered by technologies based on widgets and gadgets, which will provide customers a greater degree of control over the user interface of their Internet banking as well as the kinds and amounts of information and commodities they may access.
- Stock Trading on Mobile Devices for Individual Investors: Smartphones and personal digital assistants are

two widespread examples of such portable electronic gadgets. No protection is provided for foreign exchange dealings executed on a mobile device.

- Independent fully featured commercial client applications: This breakthrough allows for the development of commercial self-service portals and tools, which in turn allow banks to give their business customers with accurate, up-to-date information, allowing for better-informed management decisions.
- Through the utilization of cloud computing, traditional banks and other types of financial institutions may be able to convert a significant one-time capital expenditure into a more controllable ongoing operational expense. There is no requirement for brand new hardware or software. Cloud computing's flexibility and scalability make it possible for financial institutions to pay only for the resources they really use, rather than paying for everything.
- The cloud service provider takes care of all server management. There is room for improvement in the areas of data security, fault tolerance, and disaster recovery among banking industry businesses. When compared to traditionally managed systems, the costs of cloud computing's redundancy and backup are much lower.
- Financial organizations can reduce the time it takes to create new products because to the flexibility of cloud-based operational models. This has the potential to improve financial organizations' ability to respond rapidly to client needs. The cloud is an on-demand service that cuts down on the cost and effort of opening a business. Cloud computing also provides the extra benefit of enabling rapid, low-cost iteration of product improvements. Cloud computing not only enables the offloading of mission-essential services, but also of less critical chores like software updates, system maintenance, and other PC-related concerns. That is why some businesses can choose to put their money into banking rather than IT.

This paper focuses on enhancing the security of cloud banking by proposing a detection framework based on the Random Forest algorithm. Random Forest is a robust machine learning technique known for its ability to handle high-dimensional and imbalanced datasets, making it suitable for detecting distributed web attacks in the cloud banking environment.

The primary objective of this paper is to develop an accurate and efficient detection model that can identify and classify distributed web attacks in real-time. By effectively differentiating between legitimate user traffic and malicious activities, the proposed framework aims to enhance the overall security posture of cloud banking systems.

Threat to E-Commerce

The risks of internet shopping include identity theft, fraud, and security holes. When you do business online, you put yourself at risk from many different types of attackers. Some are accidental, some are on purpose, and some are just the result of carelessness on the part of the people involved. Electronic cash, data abuse, credit card and debit card fraud, and other security vulnerabilities are common dangers to the most extensively used forms of electronic payment systems 3.

The following is a list of some of the concerns that have been raised regarding cloud computing's security4:

- The process of granting access to data for more than one distinct entity while allowing those entities to share the same set of physical policies or devices and/or vitalized software resources (such as, but not limited to, memory, storage, network throughput, hardware metrics, and screen covers) is referred to as multi-tenancy.
- It is just as perilous to analyze and use this sensitive data, as it was to get it. Security precautions may be relaxed in the cloud to allow for instantaneous notification of major events like the start of a new execution process or the creation of a new file.
- It is important to derive valuable information from raw data in order to close the semantic gap. This is a prerequisite for closing the gap.
- Lack of Ability to Regain Control, When customers keep their data in the cloud, they give up whatever legal control they might have had over that data. This indicates that cloud providers have genuine access to the private data of their customers and may participate in data mining on the data of their customers, which may pose a threat to the customers' right to privacy. Even if files are erased from every cloud service that replicates data at many different data centers, there is still a risk that some residue of the data could still persist. This is the case even if the files are deleted completely. Because customers are unable to view their data directly, their customers increasingly see cloud-based firms as “Black Boxes”.
- Before more people will feel comfortable moving their operations to the cloud, a fundamental obstacle that must be addressed is trust, or the fear of giving up physical access to one's data. This anxiety prevents people from relocating their activities. Because of this, businesses are attempting to attract the patronage of clients by assuring them that the products they sell will always meet the safety and quality standards set forth by the appropriate authorities.
- The potential for an assault is the primary source of concern regarding the new virtual architecture's security.

The risk of information-driven buffer attacks is increased when the cloud computing system in question cannot be trusted.

An attempt to render regular Internet services inaccessible or unreachable on a target system (such as a computer, router, or network) is what's known as a denial-of-service attack, or DoS attack for short. Patches are often made available rather quickly after the discovery of vulnerabilities of this kind. One common tactic for a denial-of-service attack is to coerce the target into performing a large number of costly computations, such as encrypting and decrypting data or carrying out a series of calculations in secret 5. This can be accomplished in a number of ways. Although DoS attacks can take many other forms, this strategy is a common one.

The process of identifying suspicious activity within an Internet of Things network is referred to as "intrusion detection." It is reasonable to anticipate that the incursion will result in the loss of privacy, compromises to security, or waste of limited resources. The following are some examples of the several kinds of intrusions that keep system administrators up at night 6:

- Changes made without permission to system files that grant access to sensitive system or user data.
- The loss or alteration of sensitive user information or files without the owner's permission.
- Unauthorized alterations to the tables or other system information included within network components (for example, making modifications to the routing tables of a router in an intranet in order to disable access to the network).
- Use of computers and networks without permission, either by making new accounts without permission or by using existing accounts in an inappropriate manner. fraud committed using a victim's account.

Traditional safety measures are unable to provide timely and efficient protection against newly emerging diseases and dangers. A fundamental limitation of intrusion detection that is based on abuse detection is that it is unable to provide protection against unanticipated network invasions. The Internet of Things (IoT) can be protected from having its data corrupted by AI's ability to automate and offer intelligence. Preventing malicious acts from disrupting a network's normal operation may be possible with the aid of machine learning tools. Automatically collecting data from network assets, AI-enabled network security devices may detect the specific location of security breaches 7.

Because distributed web attacks in cloud banking might entail complicated and dynamic attack patterns, machine learning is an efficient methodology for identifying these attacks. In order to spot patterns and uncover assaults that could otherwise go undetected, machine-learning models can be trained on massive volumes of data. Traditional techniques of data management and analysis may struggle under the sheer volume of data produced by cloud banking systems. In addition, Real-time detection of attacks is possible with the use of machine learning models.

The supervised ensemble-learning model known as Random Forest has found widespread use in the fields of classification and regression analysis 8. When classifying a new object, the Random Forest makes use of a huge number of decision trees, each of which utilizes the same input vector as the others. Using this method, the researcher demonstrates an accurate and stable forecast with high performance. When it comes time to anticipate the outcome of each randomly produced tree, RF employs feature testing to choose the most well liked categorization rules to use. Some reasons for employing the Random Forest algorithm are listed below:

- It needs less time to train than alternative algorithms.
- It works efficiently on a huge dataset and makes highly accurate predictions.
- When a considerable amount of data is absent, it can still function accurately.

By utilizing a random forest model that has been trained on the banking dataset, the aim of this paper is to identify distributed denial-of-service attacks on the cloud-based monitoring system that is used by the banking industry. Following this section will be a discussion of relevant literature, an implementation of the model, a discussion and evaluation of the results, and lastly a summary of the research.

State of the Art

For many years now, Intrusion Detection Systems (IDS) have been considered among the most reliable means of securing the IoT digital infrastructures against a wide range of cyber-attacks. Security architectures for today's Internet of Things (IoT) networks typically IDS installed on the network itself. Signature or abuse-based IDS, anomaly-based IDS, and hybrid techniques are three of the most well-known security strategies for keeping tabs on Internet of Things (IoT) devices for signs of intrusion or other anomalies 9.

Besides presenting a highly efficient algorithm for periodicity mining, the research 10 has also offered a way for constructing location-independent Internet of Things s models. K-Means and BIRCH Clustering were applied, and the resulting 96.3 % accuracy was quite satisfying.

ISSN (online): 2581-3048
Volume 7, Issue 6, pp 185-194, June-2023
https://doi.org/10.47001/IRJIET/2023.706029

In addition, using an ANN classification approach on the Kaggle dataset DS2OS traffic traces yields an accuracy of 99.4 percent [11].

For effective anomaly identification, Ahmad et al. [12] examined many deep learning (DL) models, including a convolutional neural network, a recurrent neural network, and a long short-term memory. They achieved it by utilizing data from the IoT-Botnet 2020 project.

The use of artificial intelligence (AI) has been implemented in a wide variety of cutting-edge techniques for discovering IoT cyberattacks. Eighty papers published between 2016 and 2021 are analyzed systematically [13].

Meidan et al. [14] used random forests to develop a method for IoT authentication in large companies that makes use of white list and ensemble learning, extraction of features, and item classification based on traffic data from the network. Across nine different devices, this method was found to have an average accuracy of 99% in the tests.

Based on support vector machines (SVMs), which see attack detection as a classification problem, the authors 15developed a DDoS detection system. Information from the flow table of the SDN switch, such as IP source, source port, flow entrance speed, standard deviation of flow packets, variation of flow bytes, and pair-flow ratio, is collected and used as characteristic values for SVM classification of DDoS attacks.

The Bank of Russia detected the Bespalova virus in 2017 [16], and it was discovered to cause cash withdrawals after a code was typed into an ATM. Cross-site scripting assaults, also known as XSS, are one method that could be used in an attack to install harmful scripts into a website that is otherwise genuine. In the case that an XSS attack is successful, it has the capability of undermining the security of internet of things devices and may potentially bring the system to a grinding halt if it collects credentials from users who are genuinely allowed to use it. In addition, it has the potential to compromise the security of devices that are not connected to the internet.

Attackers may be able to compromise the system through its software and hardware components if adequate security and verification procedures are not in place. This would give them the ability to steal critical information. Because of this, it is necessary to discover breaches as soon as they occur so that attacks can be thwarted and the security of the network can be preserved. Attacking a system using a denial-of-service (DoS) or distributed denial-of-service (DDoS) is one of the methods that cybercriminals have at their disposal [17].

They are able to achieve their objectives by denying their targets service and preventing their targets from transmitting or receiving traffic with an overwhelming amount of their targets' computational and network resources. The repercussions of a security breach in the Internet of Things can be felt across both long-range networks (like NB-IoT and LoRa) and short-range networks (like ZigBee, Wi-Fi, and so on), making both types of networks susceptible to the effects of the breach. Users of Internet-connected things (IoT) devices have access to a wide variety of services thanks to the Internet. Despite this, the communication infrastructure that is based on TCP/IP is susceptible to a number of privacy and security issues, such as unauthorized access, replay attacks, and identity theft. These risks are exacerbated by the lack of a centralized authority.

2. Materials and Methods

Antivirus, intrusion, malware, and denial-of-service (DDoS) protection are the classifications that can be applied to device security. Solutions based on artificial intelligence (AI) need to be able to differentiate between trustworthy and untrusted devices, as well as typical and abnormal patterns of activity.

We made use of data regarding distributed denial of service attacks that we collected from an open-source website (which we obtained on February 2, 2021). The raw forms of the datasets were retrieved from an online database that was openly available to anyone. We utilized pre-processing methods so that the dataset would be less challenging for us to deal with. Following the removal of the outliers, the dataset was statistically balanced using the methodologies available. Once we've decided on the features to use, we'll split the data into a training set and a test set. Machine learning models are "trained" on the training set before being "tested" on the testing set for accuracy.

The raw datasets were collected from a publicly available web database. We employed pre-processing methods to clean up the data and make it more manageable. After the outliers were removed, statistical methods were used to bring the dataset into balance. We will divide the data into a training set and a test set after we have selected the features to utilize.

Data from a publicly available database detailing DDoS attacks has been used in the study. The Banking Dataset monitors network intrusions collection has 15000 samples with 38 attributes. In the data set, DDoS attacks appear 38\% of the time. The pie chart reveals that there are seven distinct categories represented by the labels, each of which must be identified by the classifier Fig. 1.

ISSN (online): 2581-3048
Volume 7, Issue 6, pp 185-194, June-2023
https://doi.org/10.47001/IRJIET/2023.706029

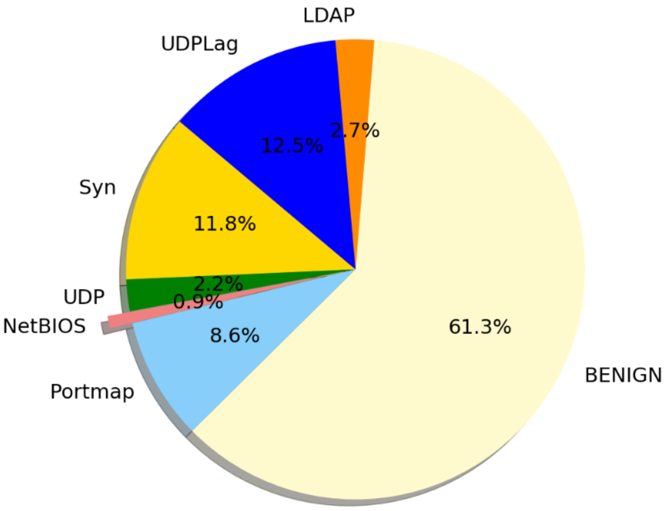

Figure 1: Frequency of the target class in the entire dataset

Data Pre-processing

IoT datasets are generated from a diverse range of sensors employed across various industries. The effectiveness of the solutions developed using these datasets is directly influenced by the quality of the training data used in their development. However, the original dataset often presents challenges such as missing information and skewed data, which hinder its usability. In order to overcome these obstacles and progress to subsequent stages of the process, it becomes necessary to perform data mining, gain expertise in data distribution, and undertake preparatory actions such as error correction and data imputation.

To enhance the dataset's quality and applicability, several steps were taken:

- Removal of all instances of possible zeros from the dataset was carried out to eliminate potential bias caused by incomplete or erroneous data points.
- Label encoding was employed to convert categorical labels, such as the Label, Flow ID, and Time Stamp columns, into a numerical representation that can be processed by computers. This enables effective utilization of the encoded information in subsequent analysis and modeling tasks.
- To focus on the most informative features, the dataset was further refined by selecting the top 20 significant features using the Extra Trees Classifier. This feature selection technique helps to prioritize the attributes that contribute most to the desired outcomes, enhancing the efficiency of subsequent modeling processes.
- The dataset was split into a training set, comprising 80% of the data, and a testing set, comprising 20% of the data. Both sets are utilized during the training and validation phases of the Random Forest model. This division allows for independent evaluation of the model's performance on unseen data, enabling accurate assessment of its generalization capabilities.

By performing these data preprocessing steps and employing appropriate techniques, we aim to ensure that the training data used in developing IoT solutions is of high quality, representative, and capable of yielding reliable and accurate results. These efforts contribute to the overall effectiveness and performance of the models and algorithms applied in IoT applications.

Random Forest Classifier

One of the most common types of machine learning algorithms used for classification work is known as the Random Forest Classifier. By merging the results of many different decision trees, it hopes to improve the accuracy of the predictions that may be made as shown in Fig. 2.

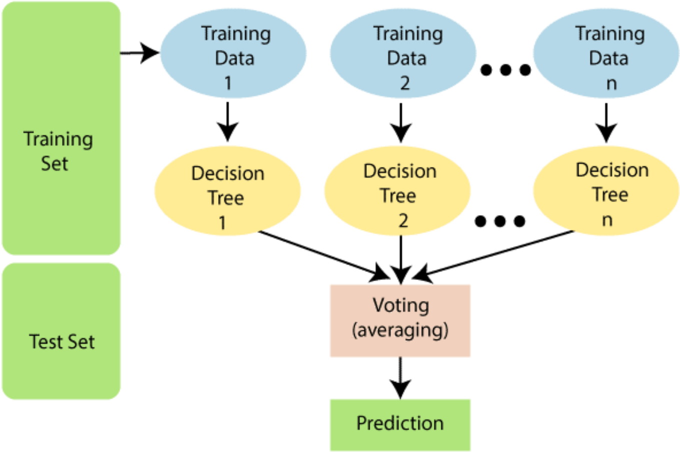

Figure 2: The working of the Random Forest algorithm

A decision tree is a hierarchical structure that is made up of nodes that represent various choices or actions depending on characteristics or qualities of the input data. This kind of structure is also known as a decision matrix. There are two distinct kinds of nodes that may be found inside the framework of a decision tree:

Decision nodes and leaf nodes: These nodes are very important to the Random Forest Classifier's decision-making process and play a significant part in it.

Decision nodes are the nodes in a network that are in charge of making choices based on certain rules or circumstances. They are comprised of several branches, each of which corresponds to a distinct alternative conclusion or option. Each branch illustrates a potential value or a range of potential values for a certain characteristic or quality.

ISSN (online): 2581-3048
Volume 7, Issue 6, pp 185-194, June-2023
https://doi.org/10.47001/IRJIET/2023.706029

The decision tree's leaf nodes, on the other hand, are said to reflect the ultimate outcomes or forecasts of the tree. They are not connected to any other branches or choices and do not have any of their own. Instead, the classification or regression result is provided by the leaf nodes, and it is dependent on the choices that were made along the route leading from the root node to the leaf node. Within the context of the decision tree, leaf nodes represent the repercussions of prior decisions or the outcomes that were anticipated from those decisions.

The Random Forest Classifier creates an ensemble of decision trees by training each tree on a unique subset of the training data. This allows the Random Forest Classifier to provide more accurate results. This method, which is referred to as bootstrapping, involves training each decision tree on a randomly chosen portion of the initial dataset while simultaneously replacing the samples. In addition, an arbitrary selection of traits is taken into consideration for splitting at each decision node of the decision trees. This helps to ensure that there is sufficient variety among the trees.

Once all of the decision trees in the ensemble have been trained, the Random Forest Classifier will integrate the results of the predictions made by each individual tree. In classification problems, it makes use of a voting system called majority voting, in which the category that receives the most number of votes across all of the trees is chosen as the final forecast. When doing tasks involving regression, the values that have been predicted by the several trees are averaged to provide the final forecast.

The Random Forest Classifier is well-known for its durability, scalability, and capacity to handle datasets with a high dimension. By combining the results of many different decision trees into a single model, it is possible to lower the likelihood of overfitting and achieve better generalization performance. In addition, the Random Forest method can capture complicated correlations between characteristics and target variables, which make it appropriate for a broad variety of classification problems across a variety of domains. This makes the Random Forest algorithm a good choice.

Training and Testing

During the training and testing phase, the models are trained by utilizing the data that has been supplied. throughout the process of training, one may get insights on the learning behavior of the model by assessing the value of the loss function or studying the learning curve. Both of these activities take place throughout the training process. The loss function computes the percentage of error that exists between the model's predictions and the values that actually exist. If one keeps an eye on the loss function, they will be able to determine how well the model is fitting the training data.

It is possible to adjust a variety of variables, such as the learning rate and the number of iterations, in order to enhance the performance of the model. The learning rate is what establishes the amount of the changes that the model makes to its parameters while it is being trained, while the number of iterations is what decides how many times the model cycles through the training data. By making adjustments to these parameters, you may assist the model overcome possible problems and enhance its performance.

After the model has been educated, it is critical to evaluate how well it performs when applied to data that it has not before seen. The performance of the model is examined using a distinct testing dataset in order to do this. Instances that were not used during the training phase have been included in the testing dataset. We are able to quantify the model's capacity to generate accurate predictions on data that it has not seen by evaluating it using the testing dataset.

During the training phase, it is of the utmost importance to address the issues of under-fitting as well as over-fitting. Under-fitting happens when a model fails to capture the underlying patterns in the data and performs poorly on both the training dataset and the testing dataset. This may happen when the model is not calibrated properly. On the other side, over-fitting occurs when the model grows too complicated and begins to memorize the training data. This results in poor generalization when applied to data that has not been seen before.

In order to address these concerns, it is required to make consistent modifications to the model's parameters. During this stage of the process, the model is fine-tuned by choosing the proper hyper parameters, which may include regularization methods, model architecture, and optimization algorithms. Monitoring and adjusting these parameters on a consistent basis helps avoid under-fitting and over-fitting, so guaranteeing that the model performs to its full potential on both the training dataset and the testing dataset.

3. Results and Discussion

We are able to evaluate the effectiveness of the model based on a variety of criteria. In particular, in classification tasks, many evaluation indicators are utilized to assess the accuracy and consistency of the model's predictions and generalizations. This ensures that the tasks are completed correctly. This contributes to ensuring that the data are interpreted in the correct manner. In the process of evaluation, some of the measurements that are utilized include area under the curve (AUC), accuracy, and precision. The F1 score is also utilized.

Confusion Matrix

The whole scope of the confounding is represented as a matrix, demonstrating the model's efficacy 18. Any classification problem's efficacy in both the training and testing phases can be evaluated using the rates in table:

- TP: The proportion of optimistic predictions that actually come true.
- TN: is a sample in which all the tests come back negative.
- FP: This probability measures how often a positive result actually is a false positive.
- FN: That is the probability that something that was expected to be negative ends up being helpful in actuality.

Fig. 3 shows confusion matrix for fandom forest classifier.

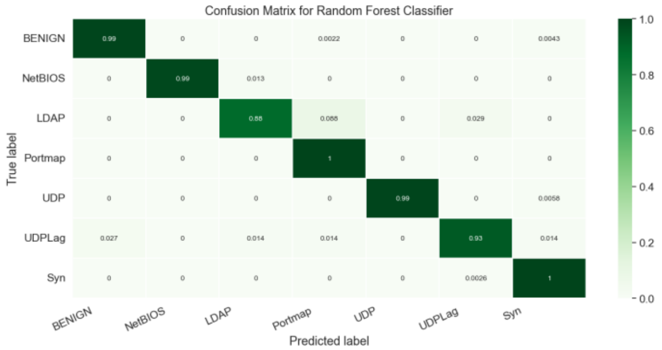

Figure 3: Confusion Matrix

Accuracy

When evaluating the effectiveness of a classification model, accuracy is a parameter that is often put to use as a measurement tool. It is the ratio of accurate predictions produced by the model to the total number of predictions produced by the model. When dealing with a classification issue that involves more than one class, accuracy takes into account both the true positive (TP) and the true negative (TN) predictions.

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

Precision

The dependability of a classification model is evaluated using a performance measure known as precision, which focuses on the model's accuracy in terms of the positive predictions it generates. It does this by calculating the percentage of accurate positive predictions that the model has produced in comparison to the total number of positive predictions. When the goal is to reduce the number of instances in which the model incorrectly identifies positive examples, precision is useful since it displays the frequency with which the model properly identifies positive instances.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Recall

Recall is a performance measure that is used to assess the completeness of a classification model's predictions, especially with regard to the positive examples that are present in the dataset. It is also known as sensitivity or the true positive rate. It determines the percentage of true positive cases that have been appropriately detected in relation to the total number of instances that are positive.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

F1 Score

By analyzing the F1 Score, we can find the sweet spot between the model's accuracy and recall. High precision/low recall rates provide remarkable accuracy, but at the cost of maybe missing a few predicted, hard-to-identify situations.

$$F1\,Score = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Table 1 presents the evaluation results of the classification model using various metrics for different classes. The metrics include Precision, Recall, F1-score, and Support. These metrics provide insights into the model's performance for each class, allowing us to assess its accuracy, completeness, and ability to correctly classify instances.

Table 1: Evaluation the results using metrics

	Precision	Recall	F1-score	Support
BENIGN	1.00	0.99	1.00	1849
LDAP	1.00	0.99	0.99	77
NetBIOS	0.94	0.88	0.91	34
Portmap	0.97	1.00	0.98	236
Syn	1.00	0.99	1.00	344
UDP	0.97	0.93	0.95	74
UDPLag	0.97	1.00	0.98	386

The table suggests that the classification model performs well for most of the classes, with high precision, recall, and F1-scores. The model shows strong performance in accurately identifying instances of the BENIGN, LDAP, Portmap, Syn, UDP, and UDPLag classes. However, the NetBIOS class has slightly lower precision, recall, and F1-score values, indicating that the model may struggle to correctly identify instances of this class. These results provide valuable insights into the model's performance for each class, aiding in understanding

ISSN (online): 2581-3048

Volume 7, Issue 6, pp 185-194, June-2023

https://doi.org/10.47001/IRJIET/2023.706029

its strengths and weaknesses in classifying different categories.

Table 2 provides overall evaluation results for the classification model using metrics such as Precision, Recall, F1-score, Support, and Accuracy. These metrics give a comprehensive overview of the model's performance across all classes and provide a summary of its effectiveness in classifying instances.

Table 2: Results

	Precision	Recall	F1-score	Support
Accuracy			0.99	3000
Macro Avg	0.98	0.97	0.97	3000
Weighted Avg	0.99	0.99	0.99	3000

Table 2 demonstrates that the classification model achieves high accuracy and performs well in terms of precision, recall, and F1-score. The macro-average and weighted-average results further reinforce the model's effectiveness in classifying instances across different classes, taking into account both equal and weighted contributions. These results indicate a reliable and robust performance of the model in predicting and classifying instances in the given dataset.

Area Under Curve (AUC - ROC)

The Area Under Curve (AUC) is a statistic that is often used in machine learning and classification tasks, especially in situations involving binary classification issues. It is primarily used to assess the performance of models by using ROC curves (Fig. 4), which is an abbreviation for Receiver Operating Characteristic.

The area under the receiver operating characteristic (ROC) curve is a graphical depiction that highlights the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for different categorization criteria. TPR, which also goes by the names recall and sensitivity, is a measurement that indicates the percentage of properly detected positive cases in comparison to the total number of real positive instances.

On the other hand, the FPR is a measurement that determines the percentage of incorrectly detected negative cases in relation to the total number of genuine negative instances.

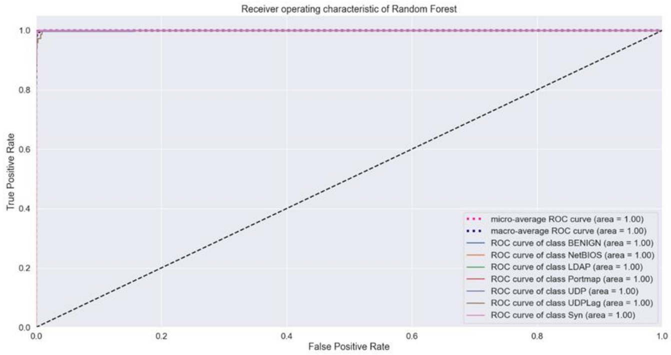

Figure 4: ROC Curve of Random Forest Classifier

The area under the ROC curve (AUC) is a metric that may be used to quantify the overall performance of a classification model over all conceivable criteria for classification. It is the likelihood that the model would rank a randomly picked positive instance higher than a randomly selected negative instance. This probability is expressed as a percentage. The area under the curve (AUC) may be anything from 0 to 1, with greater values indicating better performance. An area under the curve (AUC) of 1 indicates that the classification is flawless, whereas an AUC of 0.5 indicates that the categorization is random and is no better than chance.

Fig.4 displays flawless performance on the ROC curve, with not a single error present. The accuracy score for random was 99.16666666666667 based on the test data.

4. Conclusion

The cloud-computing paradigm may involve unforeseeable or unfavorable circumstances, which puts financial institutions in jeopardy over the long run and exposes them to the possibility of suffering losses. These risks can have an immediate and far-reaching impact on an institution's profitability, reputation, and the possibility of incurring heavy penalties from regulators. In-house IT security teams at banks often design and implement a comprehensive security architecture that is constantly monitored and modified in response to new and evolving security risks. In order to classify DDOS attacks, this dissertation suggests utilizing the Banking Dataset and the Random Forest method. It examined the effects of varying the training parameters on the accuracy of the classification findings. The study used real-world data to derive metrics including the ROC Curve, the Confusion Matrix, and the Accuracy score. The accuracy of Random Forest was 99.1666% on test data. The future of this industry is wide open due to the enormous amounts of both structured and unstructured data that must be preprocessed in an efficient manner before being clubbed together to extract information that has been concealed.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to all those who have contributed to the completion of this research project.

Author's Declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are mine/ours. Furthermore, any Figures and images, that are not mine/ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in University of American University of Culture and Education.

Author's Contribution Statement

Rafal Sattar Jabbar and Mohamad Tawfik Hamze contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

REFERENCES

[1] European network for network and information security. Good practices and recommendations for secure use of cloud computing in the finance sector. December 2015. Available from: [link].

[2] Cloud Security Alliance. The Treacherous 12: Cloud Computing Top threats in 2016. Available from: [link]

[3] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in python. Journal of machine learning research. 2011;12(Oct):2825-2830.

[4] Achar S. Security of Accounting Data in Cloud Computing: A Conceptual Review. Asian Accounting and Auditing Advancement. 2018; 9(1):60-72.

[5] Abdelrafe Elzamly NM, Doheir M, Mahmoud A, AbdSamad Bin Hasan Basari NA, Ali Al-Shami SS. Adoption Of Cloud Computing Model For Managing E Banking System In Banking Organizations. International Journal Of Advanced Science And Technology. 2019; 28(1):318-326.

[6] Moustafa N, Slay J. The evaluation of network anomaly detection systems: Statistical analysis of the unsw-nb15 data set and the comparison with the kdd99 data set. Information Security Journal: A Global Perspective. 2016; 25(1-3):18-31.

[7] Panda M, Patra MR. Network intrusion detection using naive bayes. International journal of computer science and network security. 2007; 7(12):258-263.

[8] Sommer R, Paxson V. Outside the closed world: On using machine learning for network intrusion detection. In: 2010 IEEE symposium on security and privacy. IEEE; 2010. p. 305-316.

[9] Ranjit Bose XRL, Liu Y. The Roles of Security And Trust: Comparing Cloud Computing And Banking. In: The 2nd International Conference on Integrated Information 2013.

[10] Sharma DPSN, Cloud Computing Security Through Cryptography for Banking Sector. In: Proceedings of the 5th National Conference; Indiacom-2011 Computing For National Development. March 10-11, 2011. BharatiVidyapeeth's Institutes Of Computer Applications And Management, New Delhi. Copy Right Indiacom-2011.

[11] Gangal SRAA. Security Issues of Banking Adopting The Application Of Cloud Computing. International Journal Of Information Technology And Knowledge Management. 2011; 5(2):243-246.

[12] Gwara GOMS, Kimwele M. A Framework for Assessing Cloud Computing Security for Cloud Adoption In Microfinance Banks. International Journal of Advances in Computer Science and Technology. 2014; 3(1):34-38.

[13] Tesema DH. Cloud Computing Adoption Challenge In Case Of Commercial Bank Of Ethiopia. International Journal Of Development Research. January 2020; 10:33562-33565.

[14] Bawany NZ, Shamsi JA, Salah K. DDoS attack detection and mitigation using SDN: methods, practices, and solutions. Arabian Journal for Science and Engineering. 2017; 42(2):425-441.

[15] Xiao P, Qu W, Qi H, Li Z. Detecting DDoS attacks against data center with correlation analysis. Computer Communications. 2015; 67:66-74.

[16] Sommer R, Paxson V. Outside the closed world: On using machine learning for network intrusion detection. In: 2010 IEEE symposium on security and privacy. IEEE; 2010. p. 305-316.

[17] Wang B, Zheng Y, Lou W, Hou YT. DDoS attack protection in the era of cloud computing and software-defined networking. Computer Networks. 2015; 81:308-319.

[18] Tesema DH. Cloud Computing Adoption Challenge In Case Of Commercial Bank Of Ethiopia. International Journal of Development Research. January 2020; 10:33562-33565.

Citation of this Article:

Rafal Sattar Jabbar, Mohamad Tawfik Hamze, "Enhancing Cloud Banking Security: Detection of Distributed Web Attacks through Random Forest Algorithm" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 7, Issue 6, pp 185-194, June 2023. Article DOI https://doi.org/10.47001/IRJIET/2023.706029

****