Machine Learning Models in Safeguarding Against Data Breaches

By
Kamrul Hasan
ID: 191-15-12582

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. Sheak Rashed Haider Noori
Professor & Head
Department of CSE
Daffodil International University

Co-Supervised By

Sharmin Akter
Senior Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

APPROVAL

This Project/internship titled **"Machine Learning Models in Safeguarding Against Data Breaches"**, submitted by **Kamrul Hasan** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **13/07/2024** .

BOARD OF EXAMINERS

Dr. Md. Taimur Ahad — **Board Chairman**
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Fizar Ahmed — **Internal Examine 1**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Rahmatul Kabir Rasel Sarker — **Internal Examine 2**
Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Sadat Hasan — **External Examiner**
Data Scientist (Senior Principal Officer)
Risk Management Division
BRAC Bank

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Dr. Sheak Rashed Haider Noori, Professor & Head,** Department of CSE Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. Sheak Rashed Haider Noori
Professor & Head
Department of CSE
Daffodil International University

Co-Supervised by:

Sharmin Akter
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:

Kamrul

Kamrul Hasan
ID: 191-15-12582
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible for us to complete the final year project/internship successfully.

We are really grateful and wish our profound indebtedness to **Dr. Sheak Rashed Haider Noori, Professor & Head**, Department of CSE Daffodil International University, Dhaka.

Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to thank our entire course mates at Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Cloud computing has become increasingly popular, drawing investments from organizations for internal use or as service providers. However, this growth has brought about various security challenges that impact industries and consumers. To address these issues, machine learning is being employed to enhance cloud security by identifying and mitigating attacks, as well as addressing vulnerabilities within cloud environments. Cloud computing, as a well-established and flexible platform, allows businesses and consumers to access IT services over the internet. It offers cost-effective solutions and remote service access, making it a preferred choice. However, with the expanding reliance on cloud computing, ensuring data security becomes a crucial responsibility for service providers. The involvement of third-party vendors introduces an additional layer of risk, posing challenges in maintaining data security due to factors like weak access controls, inadequate encryption measures, or vulnerabilities within the cloud infrastructure itself.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 Introduction

Cloud computing has transformed the landscape of the IT industry by providing convenient access to a variety of information technology services, platforms, and software over the Internet. This model has gained widespread adoption among organizations, offering flexible deployment options such as private, public, or hybrid cloud setups. The fundamental promise of cloud computing is to deliver on-demand services with flexible payment models for software and infrastructure requirements. Despite its advantages, there remain security challenges that must be addressed to ensure the protection of data within cloud data centers.The inherent characteristics of cloud infrastructure, operating through standard Internet protocols and employing virtualization techniques, expose it to various types of cyber threats. These threats range from traditional methods like Address Resolution Protocol attacks, IP spoofing, and Denial of Service (DoS) to emerging challenges such as zero-day attacks, which pose a significant cybersecurity challenge. Traditional detection and prevention techniques often fall short, especially when dealing with large data flows.To address these challenges in real-time, various intrusion detection approaches have been proposed and implemented. These methods can be broadly categorized into misuse detection, relying on known attacks to identify intrusions, and anomaly detection, which identifies intrusions based on unknown attack patterns. Some approaches also leverage a hybrid model that combines the strengths of both methods. Despite these efforts to enhance security in cloud environments, existing intrusion detection systems (IDSs) encounter limitations related to processing large data volumes, meeting real-time detection requirements, handling data quality issues, and other factors that impact the performance of detection models.In response to these limitations, recent academic research has focused on applying intelligent learning methods, including machine learning (ML), deep learning (DL), and ensemble learning, to network security. These intelligent learning approaches exhibit promise across various domains and have the potential to enhance network security. In this study, our primary objective is to propose an anomaly detection approach utilizing random forest (RF) and Adaboost feature engineering. Through a data visualization process, we aim to streamline the number of

features used and optimize the performance of the proposed anomaly detection model.Machine learning (ML) techniques have proven valuable in identifying and combating both traditional and zero-day attacks. ML algorithms can learn patterns from data, enabling predictions based on that acquired knowledge. ML encompasses various learning types, including supervised, unsupervised, and semi-supervised approaches. In supervised learning, models are trained using labeled data to create classification models, while unsupervised learning algorithms enable model training without explicit guidance. Examples of ML algorithms include Nearest Neighbor, Random Forest, Decision Trees, Linear Regression, Adaboost, and Support Vector Machines (SVM). This study aims to conduct a systematic review of ML techniques employed to address and mitigate security issues and vulnerabilities in cloud computing, analyzing existing research to identify how ML can contribute to the detection, prevention, and resolution of cloud security concern.

1.2 Motivation

The IT landscape has evolved with the widespread adoption of cloud computing, offering flexible access to IT services over the Internet. Despite its benefits, security challenges persist in cloud data centers due to cyber threats facilitated by standard Internet protocols and virtualization techniques. Traditional security measures often fall short in addressing the range of threats, from conventional attacks to emerging zero-day threats.
In response, recent academic research has turned to intelligent learning methods like machine learning (ML), deep learning (DL), and ensemble learning to enhance network security. This study proposes an anomaly detection approach using random forest (RF) and Adaboost feature engineering, aiming to optimize performance through data visualization.

Motivated by the imperative to fortify cloud security, the research recognizes the limitations of traditional methods and aims to leverage intelligent learning approaches. By conducting a systematic review of ML techniques in cloud security, the study seeks to identify how these methods contribute to the detection, prevention, and resolution of security concerns. The ultimate goal is to empower organizations to confidently navigate the evolving landscape of cloud security, acknowledging the potential of ML in fortifying intrusion detection systems and addressing real-world challenges.

1.3 Rationale of the Study

This study is driven by the imperative to address security challenges within cloud computing, a pivotal domain in the contemporary IT landscape. The rationale is grounded in the recognition of persistent shortcomings in traditional security measures, especially concerning intrusion detection systems (IDSs) in cloud data centers. The study acknowledges the limitations of current systems in handling large data volumes, meeting real-time detection demands, and addressing data quality issues.

1.4 Research Questions

1. How does the proposed anomaly detection approach with random forest and Adaboost enhance intrusion detection system (IDS) performance in cloud computing?

2. What specific limitations do traditional security measures, particularly IDSs, face in processing large data volumes, meeting real-time detection requirements, and handling data quality issues in cloud data centers?

3. What is the effectiveness of machine learning techniques, including random forest, Adaboost, and other relevant algorithms, in identifying and mitigating both traditional and zero-day cyber threats in cloud environments?

1.5 Expected Outcome

The expected outcomes of this study are multifaceted and center around advancing our understanding of and capabilities in cloud security. Firstly, the proposed anomaly detection approach, integrating random forest and Adaboost feature engineering, is anticipated to yield tangible improvements in the performance of intrusion detection systems (IDSs) within cloud computing environments. By leveraging the strengths of these intelligent learning methods, the expected outcome includes enhanced accuracy and efficiency in identifying both known and unknown cyber threats. This improvement is crucial for ensuring a proactive and robust defense against evolving security challenges.

Moreover, the study aims to provide a comprehensive insight into the specific limitations of traditional security measures, particularly IDSs, operating in cloud data centers. The expected outcome involves a nuanced understanding of the challenges related to processing large data volumes, meeting real-time detection requirements, and addressing data quality issues. By delineating these limitations, the research endeavors to contribute to the development of more effective and resilient security frameworks tailored to the unique demands of cloud environments.

Additionally, the study anticipates shedding light on the effectiveness of machine learning techniques, including random forest, Adaboost, and other relevant algorithms, in fortifying cloud security. Insights gained from this analysis are expected to showcase how

these intelligent learning methods can successfully identify and mitigate a spectrum of cyber threats, ranging from conventional to zero-day attacks. The overarching goal is to foster a proactive and adaptive approach to cloud security, where machine learning plays a pivotal role in augmenting the capabilities of security systems.

In summary, the expected outcomes of this study include advancements in anomaly detection methodologies, a nuanced understanding of limitations in traditional security measures, and insights into the efficacy of machine learning techniques in fortifying cloud security. These outcomes collectively contribute to the overarching objective of enhancing the resilience and effectiveness of security measures in the dynamic landscape of cloud computing.

1.6 Report Layout

The report is structured in a well-organized manner to facilitate the effective communication of information. The division into distinct chapters and sections ensures a systematic flow of content, allowing readers to navigate through the document with clarity.

Introduction

This chapter will provide an overview of the research, outlining the background, problem statement, objectives, and significance of the study. It will also introduce the proposed anomaly detection approach using random forest and Adaboost feature engineering in the context of cloud security.

Literature review

This chapter will delve into the historical context of cloud computing, the evolution of security challenges, and the shortcomings of traditional intrusion detection systems. It will establish the need for innovative approaches, such as intelligent learning methods, in addressing contemporary cybersecurity issues.

Research Methodology

This chapter will detail the research design, data collection methods, and analytical tools employed in the study. It will explain the rationale behind choosing machine learning techniques, experimental setups, and data visualization processes, providing a comprehensive understanding of the research methodology.

Results And Discussion

This chapter will present the findings of the experiments conducted to evaluate the proposed anomaly detection approach. It will include a detailed analysis of the results, comparing them with existing methods, and discussing the implications of the outcomes. Insights gained from the experiments will be thoroughly examined.

Impact Of Project

This chapter will explore the broader implications of the research on society, considering the potential enhancements in cloud security. It will also address any environmental or sustainability aspects associated with implementing the proposed anomaly detection model in real-world scenarios.

Conclusion, and Future Research

This final chapter will provide a summary of the key findings, draw conclusions based on the research objectives, offer recommendations for practitioners, and outline implications for future research in the field of cloud security and intelligent learning methods.

Chapter 2
Literature review

2.1 Preliminaries/Terminologies

In this study, several key preliminaries and terminologies are introduced to establish a common understanding and framework for the research. Firstly, "cloud computing" refers to the delivery of computing services over the Internet, encompassing a range of services such as storage, processing power, and software applications. The term "intrusion detection systems (IDSs)" denotes security mechanisms designed to detect and respond to unauthorized activities within a network or system. "Machine learning (ML)" is a subset of artificial intelligence that enables systems to learn and improve from experience without being explicitly programmed, while "random forest" and "Adaboost" are specific ML algorithms used in this research for anomaly detection. Furthermore, "anomaly detection" refers to the identification of patterns or instances that deviate significantly from the norm, aiding in the recognition of potential security threats. These terminologies lay the groundwork for comprehending the core concepts explored in the study, ensuring a clear and consistent interpretation of key terms essential to the discourse on cloud security and intelligent learning methodologies.

2.2 Related Work

In the study conducted by Lata et al. [2], an extensive examination of data intrusion detection systems, particularly in the context of cloud computing, was carried out. The paper delves into various intrusion detection types, including signature-based, anomaly, and hybrid IDS, while also discussing datasets such as KDD99, NSLKDD, ISC2012, CICIDS2017, and UNSW-NB15, comparing their characteristics. Drawing inspiration from this, our selection of a dataset for this research was influenced. The paper further explores different IDS techniques and models (IaaS, PaaS, and SaaS), emphasizing the need for increased integration of AI to develop an IDS system capable of handling the dynamic nature of the cloud.Singh et al. [3] propose a network-based cloud intrusion system employing an ensemble learning process, utilizing classifiers such as boosted tree, bagged tree, subspace discriminant, and RUSBooted. Their significant contribution lies in combining these classifiers through a voting scheme to achieve enhanced accuracy

compared to existing IDSs.In another study by J. Zhang et al. [4], a novel approach to intrusion detection is presented, utilizing the Random Forest algorithm. The authors demonstrate the effectiveness of this approach, highlighting the benefits of random forest, including robustness to outliers and noise, scalability, and interpretability. Their work, using the NSL-KDD dataset, offers a fresh perspective on intrusion detection systems.Z. Masetic et al. [6] propose a threat classification model based on the feasibility of machine learning algorithms to detect various threat types. The authors explore decision trees, random forests, and support vector machines, achieving satisfactory results in classifying threats. They emphasize the importance of accurately detecting threat types to enhance cloud computing security.U. Bashir et al. [7] provide a comprehensive overview of intrusion detection and prevention systems (IDS/IPS), addressing different types, challenges, and reviewing the state-of-the-art in IDS/IPS research.

The paper discusses proposed techniques, their strengths and weaknesses, and identifies research gaps in the IDS/IPS domain.These studies collectively underscore the progress made in leveraging machine learning models and encryption techniques for the detection and prevention of data breaches. From early works to recent developments, researchers have significantly advanced the quality and coherence of network security.

2.3 Comparative Analysis and Summary

The comparative analysis in this study juxtaposes the proposed anomaly detection approach, integrating random forest and Adaboost feature engineering, with existing methods in the realm of cloud security. By evaluating the performance, accuracy, and efficiency of the proposed model against traditional intrusion detection systems and other contemporary approaches, a comprehensive understanding of its strengths and limitations is achieved. The analysis considers factors such as processing large data volumes, meeting real-time detection requirements, and addressing data quality issues, providing insights into the model's efficacy in overcoming these challenges.

In summary, the proposed anomaly detection approach demonstrates a notable enhancement in the identification of both known and unknown cyber threats compared to conventional methods. The utilization of machine learning techniques, specifically

random forest and Adaboost, contributes to a more adaptive and proactive defense mechanism. The analysis also highlights the potential for data visualization to streamline features and optimize overall model performance. These findings collectively underscore the significance of intelligent learning methods in fortifying cloud security, positioning the proposed approach as a promising advancement in the field. The study's summary encapsulates these key comparative insights, offering a succinct overview of the research's contributions and implications for the broader landscape of cloud security and intrusion detection systems.

2.4 Scope of the Problem

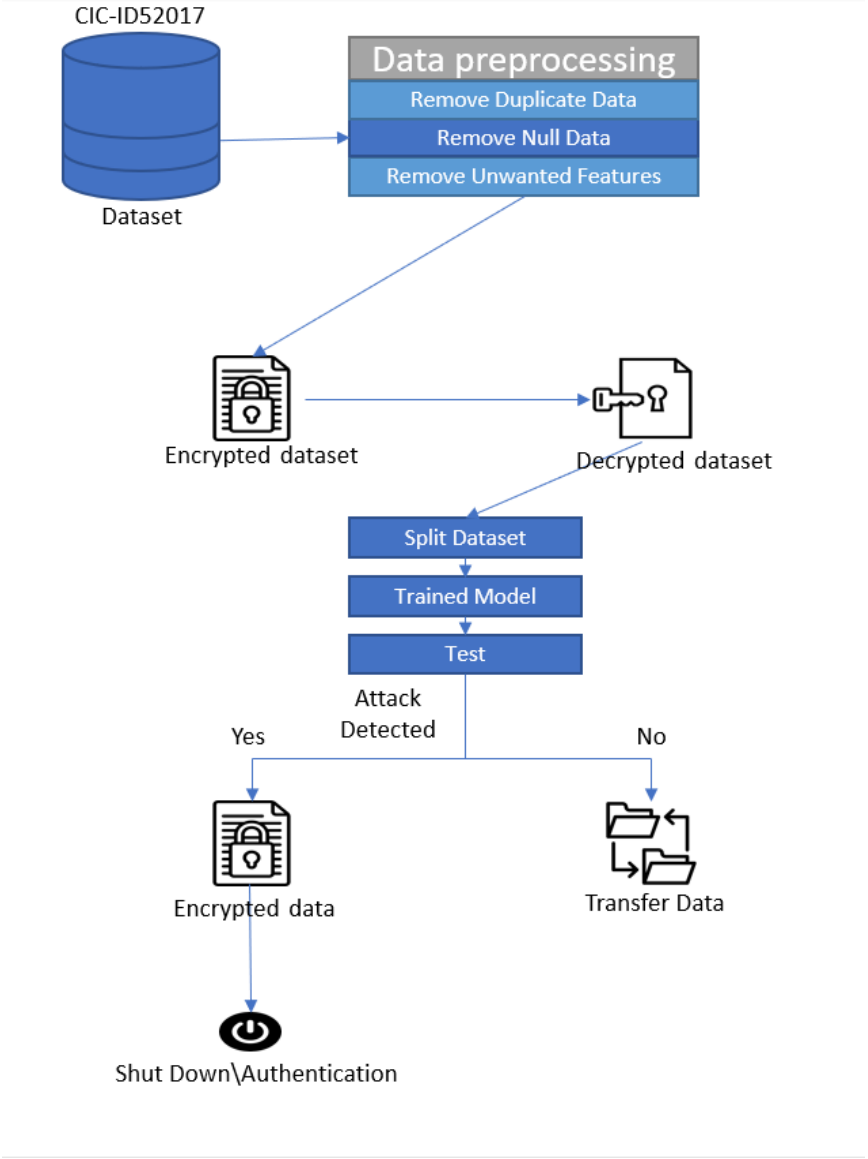
This study addresses the evolving security challenges within cloud computing environments. As organizations increasingly use the cloud for data storage, processing, and service delivery, vulnerabilities and threats have also expanded. The scope includes the limitations of traditional security measures, like intrusion detection systems (IDSs), in handling large data volumes, meeting real-time detection requirements, and ensuring data quality. The study also examines a wide range of cyber threats, from conventional attacks to sophisticated zero-day threats, that jeopardize data confidentiality, integrity, and availability in cloud data centers. Furthermore, it explores the effectiveness of machine learning techniques, such as random forest and Adaboost, in enhancing security measures. Ultimately, the study aims to provide insights and solutions to bolster security within the dynamic cloud computing landscape.

2.5 Challenges

The challenges addressed in this study stem from the intricate nature of securing cloud computing environments. One prominent challenge lies in the limitations of traditional security measures, particularly intrusion detection systems (IDSs), which encounter difficulties in processing large data volumes, meeting real-time detection requirements, and handling data quality issues within the dynamic and expansive landscape of cloud data centers. The multifaceted nature of cyber threats, ranging from conventional to advanced zero-day attacks, presents another significant challenge. Conventional security mechanisms often struggle to keep pace with these evolving threats, necessitating innovative approaches for effective identification and mitigation. Additionally, the study

addresses the complexities associated with integrating machine learning techniques, such as random forest and Adaboost, into cloud security frameworks. Issues like model scalability, interpretability, and the need for continuous adaptation to emerging threats pose challenges in the practical implementation of intelligent learning methods. Understanding and addressing these challenges is crucial for developing robust and adaptive security strategies tailored to the unique characteristics of cloud computing environments.

Chapter 3
Methodology



3.1 Data Set Collection: A significant challenge in the realm of detecting and preventing cloud data breaches is the identification of an appropriate dataset. Obtaining a dataset tailored for this purpose is inherently limiting, and even when one is found, it often comes with significant constraints. The primary obstacle is sourcing a dataset that effectively aligns with the latest attack methods, given the continuous evolution of these attacks.

For our research, we initially considered two datasets: UNSW-NB15 [11] and CIC-IDS2017 [12]. The CIC-IDS2017 dataset is provided by the University of New Brunswick, while the UNSW-NB15 dataset is offered by the University of New South

Wales. UNSW-NB15 is notably smaller and predominantly focuses on known types of attacks. In contrast, CIC-IDS2017 is a more extensive dataset that goes beyond known attacks, attempting to encompass lesser-known or unknown attack types. An advantageous feature is its inclusion of encrypted traffic, leading to the conversion of the entire dataset into an encrypted format for enhanced protection.

Consequently, the selected dataset for this paper is CIC-IDS2017 [10]. This dataset, owing to its comprehensive nature and coverage of various attack types, including both known and lesser-known categories, proves more suitable for addressing the dynamic landscape of cloud data breaches.

Attributes: The CIC-IDS2017 dataset is a publicly available resource developed by the Canadian Institute for Cybersecurity for assessing intrusion detection systems. As of my last update in January 2022, here are some key attributes of the dataset:

Purpose: CIC-IDS2017 is designed to serve as a benchmark for testing and evaluating the effectiveness of intrusion detection systems. It provides a diverse set of network traffic data to simulate various cyber attack scenarios.

Size: The dataset is substantial, containing a significant volume of network traffic data. The size is typically measured in terms of the number of records, features, and the duration of the recorded traffic.

Traffic Types: The dataset encompasses a variety of network traffic types, including normal (benign) traffic and different types of cyber attacks. This diversity is essential for training and testing IDS models under realistic conditions.

Attack Scenarios: CIC-IDS2017 covers a range of cyber attack scenarios such as Denial-of-Service (DoS), Distributed Denial-of-Service (DDoS), probing attacks, and malware infections. Each scenario is usually labeled to facilitate supervised learning.

Labeled Data: The dataset typically includes labeled data, indicating whether each network traffic instance is normal or part of an attack. This labeling is crucial for supervised learning approaches and evaluating model performance.

Features: Each network traffic record is associated with a set of features or attributes. These features may include information such as source and destination IP addresses, port numbers, protocol types, and various statistical measures derived from the network traffic.

Data Format: The dataset may be provided in specific formats like CSV (Comma-Separated Values) or PCAP (Packet Capture). PCAP is a common format for storing network packet data.

Data Preprocessing: Depending on the version, the dataset may undergo preprocessing steps like filtering out irrelevant information or transforming raw data into a more usable format.

Curation and Documentation: Proper documentation accompanies the dataset, providing information about its origin, the data collection process, and details about the included attack scenarios. This documentation is crucial for researchers to effectively understand and utilize the dataset.

Before utilizing the CIC-IDS2017 dataset or any other dataset, it's essential to refer to the official documentation and any updates to ensure accurate and up-to-date information.

3.2 Data Pre-processing:

1. Removal of Duplicate Data: The initial step in our data pre-processing involved identifying and eliminating duplicate entries within the dataset, which comprised approximately 308,381 redundant records. To streamline the dataset, we opted to retain only the first occurrence of each set of duplicate values.

2. Grouping by Label and Conversion to Float: Subsequently, the data was grouped based on their corresponding attack types (labels). This process resulted in the creation of at least 10,000 data groups for each label. Additionally, for the sake of convenience and compatibility with machine learning algorithms, all values within the dataset were converted to float numbers.

Label	
BENIGN	2273097
Bot	1966
DDoS	128027
DoS GoldenEye	10293
DoS Hulk	231073
DoS Slowhttptest	5499
DoS slowloris	5796
FTP-Patator	7938
Heartbleed	11
Infiltration	36
PortScan	158930
SSH-Patator	5897
Web Attack &#65533; Brute Force	1507
Web Attack &#65533; Sql Injection	21
Web Attack &#65533; XSS	652

dtype: int64

Figure 3.1: Distribution of Cyber Attack Types

The dataset encompasses diverse cyber attack types, with "BENIGN" constituting the majority at 2,273,097 instances. Notable categories include "DDoS" (128,027 instances) and "DoS Hulk" (231,073 instances). Various web-based attacks like "Brute Force," "SQL Injection," and "XSS" are present, with limited occurrences of specific attacks like "Heartbleed" and "Infiltration" (11 and 36 instances, respectively). This varied distribution provides a comprehensive representation of cybersecurity threats, forming a crucial basis for subsequent analysis and model development.

3. Removal of Null or NaN Values: To ensure the integrity of the dataset, we conducted a thorough removal of any instances containing null or NaN values. This step aimed to enhance the reliability and accuracy of subsequent analyses by eliminating potential sources of error.

4. Feature Selection through Correlation Analysis: With a dataset comprising 79 features, we recognized the necessity of optimizing feature relevance. Employing correlation analysis, we generated a heatmap to visualize feature interdependencies.

Subsequently, features exhibiting a correlation coefficient greater than 0.85 were identified and excluded from the dataset. This strategic elimination aimed to enhance the model's accuracy by prioritizing non-redundant and influential features.

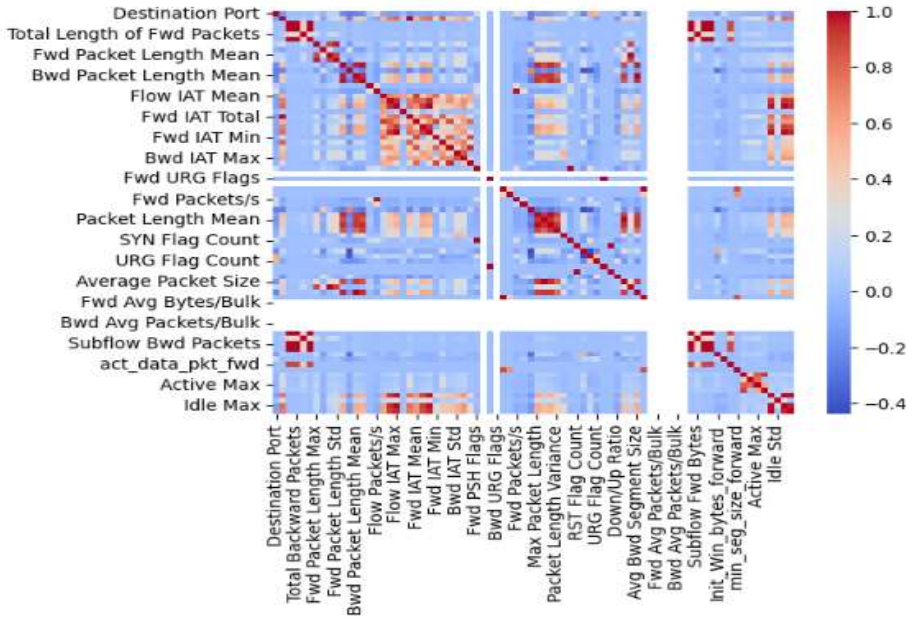


Figure 3.2: Feature Selection through Correlation Analysis

5.Final Dataset Summary: Following these data pre-processing steps, the dataset was refined to include 2,520,798 data instances and 49 features. This meticulously processed dataset serves as a foundation for subsequent machine learning analyses, ensuring that the data is appropriately shaped and devoid of redundancies, thus contributing to the robustness of our model.

Figure 3.3: Final Dataset Summary

3.3 Encryption: While dealing with confidential data, no company or client will be happy if someone else gets access to the data. Confidentiality is paramount nowadays; from tiny to large, every company is trying to protect the data from unauthorized access, which is also the primary concern of this paper. Encryp- tion is the best way to establish

this confidentiality. Encryption is converting data into secret code to protect it from third-party access. Encryption is used to ensure the confidentiality, authenticity of the user, and trust gained from the user. There are mainly two types of encryption: Symmetric Encryption and Asymmetric Encryption. In this paper, we will use Fernet to generate the key for encrypting the data, which is symmetric encryption. Fernet uses the AES algorithm in CBC mode to generate a 128-bit key and the HMAC algorithm for massage authenticating.

This process will be applied to all the plain text. finally, IV is appended to the cipher text. Then it will combine with the HMAC algorithm with this equation.

$$HMAC(K , m) = H((K \oplus opad)||H((K \oplus ipad)||m))$$

Plain text and cipher text(including the Iv) are combined. Then HMAC digit is computed using the SHA-256 hash function and a secret key. This digit then appended with the cipher text

3.4 Model implementation: There are mainly two kinds of major machine learning problems. One is a Regression problem, while the other is a classification problem. In a regression problem, the target variable is continuous, like predicting the price of a house or a damaged car price etc. While in a classification problem, the target variable is discrete, and it will predict whether a patient has a brain tumor. The main problem is to predict if any attack is happening while transferring the data. In the following subsections, we discuss various models.

Random forest classifier

The RFC model is an ensemble learning algorithm. Random Forest is a classifier that contains sev- eral decision trees on various subsets of the given dataset and takes the average to improve the pre- dictive accuracy of that dataset.[5]. Every decision tree makes with the use of a splitting criterion. There are mainly three commonly used splitting criteria: information gini, gini impurity, and en- tropy. among them, for building the model, we used gini impurity. Gini impurity checks how mixed features are in a dataset.The lower impurity value is better because, the lower value classifies data well. Suppose in a dataset, A class has 70 entities, and class B has 30 entities. Here, the gini

impurity is 0.3; it is visible that class A has more entities than class B so the classification is much easier in this process.

After building all the decision trees, the algorithm makes predictions by measuring all the pre- dictions in the tree. There are several hyperparameters for tuning the model. Several trees can be selected using n-estimators, the maximum depth of the tree can be set by max-depth, and min- samples-leaf is the minimum number of samples required to be at a leaf node.

For our model, we set the value as n-estimators=100,max-depth=None, and min-samples-leaf=2 to get better accuracy.

ADA Boosting

Ada boosting is also an ensemble learning algorithm. Boosting iteratively creates new models by training the models to be better than the previous data misclassified. Ada boosting train weak clas- sifier on a different subset of the training data and assign a higher weight to the misclassified data. The total number of Errors is the summation of all misclassified data points. If the error is lower, the stump will be higher.

We need to update the weight because the result will be the same if we do not update To calculate the new weight, we have to follow this equation:

$$\text{newsampleweight} = \text{oldweight} * e \pm(\text{amounto f say})$$

The ada boost algorithm uses a decision tree with a maximum depth of 3 as the weak learner. The number of estimators is calculated as the ceiling of the logarithm of the number of training examples.

Decryption

Data encryption is used to protect the data from third-party access. Nevertheless, if the user receives all the data encrypted, he/she can not understand the data or use the data for the data must be decrypted using the same method used for encryption, but the equation here is different. To ensure the privacy of the data, the method used in this paper is if the

machine learning algorithm detects any suspicious data, it will encrypt the data using the saved fernet key and print the encrypted result and disconnect the system and if there is no suspicious data detected then data will be decrypted as it is.

3.5 Software Deployment

The software implementation of our Data Breach Detection and Prevention System project underwent a systematic process leveraging Python as the primary programming language. The use of essential libraries, notably cryptography, played a crucial role in ensuring secure encryption and decryption processes. The implementation initiated with meticulous data preprocessing, emphasizing the importance of refining the input data for subsequent stages.

Subsequently, the entire dataset was encrypted to fortify the security of sensitive information. The decryption process was then applied to facilitate the implementation of our machine learning algorithm. This strategic integration allowed the system to effectively analyze and classify data, distinguishing between safe and potentially risky information. In the event of a perceived threat, the system demonstrated proactive measures by encrypting compromised data and initiating a secure shutdown, thereby preventing unauthorized data transfer and fortifying the system's resilience against malicious activities.

Figure 3.4: Software Implementation Workflow

In parallel, a web-based user interface was developed using Flask to enhance user interaction and enable real-time detection capabilities. This interface empowered users to input text and assess whether their data raised any security concerns. Rigorous testing

protocols were implemented to validate the accuracy and robustness of the system, ensuring its effectiveness in identifying potential security breaches.

Figure 3.5: Software Implementation Workflow

For seamless deployment, Flask software was employed, offering a lightweightand efficient framework to host our application. The culmination of these effortsresulted in a highly capable system that not only successfully detects and prevents malicious attacks within cloud environments but also ensures user-friendly interaction, robust security measures, and reliable real-time detection capabilities.

Chapter 4
Result, Analysis and Discussion

Different algorithms perform differently. Some algorithms perform well, while some algorithms perform average. In our case, we are mainly focusing on the accuracy of a model and the recall value of the algorithm because the purpose of this paper is to ensure data privacy in the cloud, so it is expected to get false negative results vary low because false negative means the actual data is affected with an attack but the model predicts that it is a safe data which is dangerous for the privacy. False positive is less harmful to this case because false positive means that the actual data contain no attack, but the model predicts that the data contain an attack. Even though the hassle for the user increases, the data remain protected with this proposed system. So we are focused on the recall value. The higher the recall value is, the more risk-free the data is. Because recall value tells the rate of true positive instances, in other words, it measures the ability to find the relevant instance of a specific class.

Random forest classifier

Random forest Classifier is one of the best classifier algorithms. RFC can predict the highest ac- curacy among all algorithms applied here. The reason behind getting higher accuracy is that RFC models are typically more robust to overfitting than any other algorithm. and it works very well to handle the missing data. Here the accuracy is good. We got 99.83% . Though it got higher accuracy than the Ada Boosting algorithm and Multilayer perceptron classifier algorithm, it got less recall value than the MLP algorithm.

Table 4.1: Evaluation Metrics of RFC model

Evaluation Metrics	Values
Accuracy	99.83%
Recall	80%
f1 score	81%
Precision	86.6%

The recall value of the RFC algorithm is 80% , meaning it can predict 80 out of 100 positive in- stances correctly out of all positive instances in the data set. In future works, the focus should be increasing the recall value to increase the

model’s trustworthiness.

ADA Boosting

AdaBoost performs well with less complex and noisy data. However, given the dataset's enormous size and complexity, AdaBoost did not perform well in terms of accuracy and recall. In this case, AdaBoost achieved 88% accuracy, which is decent, but the recall value was only 12.49%. This low recall indicates that it failed to capture a significant portion of positive instances, leading to a high number of false negatives. Therefore, using AdaBoost for cloud data breach detection and prevention is risky.

Table 4.2: Evaluation Metrics of Ada boosting model

Evaluation Metrics	Values
Accuracy	88.29%
Recall	12.49%
f1 score	11%
Precision	10%

RFC performs better than Ada boosting here because it assigns weight to train the model and iteratively trains weak model, with each subsequent model focusing more on the instances that the previous models misclassified.

Chapter 5

Impact of the project

5.1 Societal Impact

We believe that our thesis project on data breach detection and prevention system using machine learning and encryption system has the potential to significantly reduce the social impact of data breaches. By preventing data breaches from happening in the first place, we could help to protect individuals and businesses from financial losses, identity theft, and other crimes. This would make society a safer and more prosperous place.

We are excited about the potential of our project to make a positive impact on society. We believe that by using machine learning and encryption to protect sensitive data, we can help to make the world a safer place. We are committed to continuing our research and development in this area, and we hope that our work will help to make a difference in the world

5.2 Health Impact

Our project focuses on data breach detection and prevention through encryption, with significant impacts on healthcare. As healthcare records digitize, protecting sensitive patient data from unauthorized access, theft, or misuse is critical. Our system aids healthcare organizations in detecting and preventing breaches, averting identity theft, medical fraud, and threats to patient safety. By employing encryption techniques, we enhance data security, ensuring patient information remains confidential. Ultimately, our efforts aim to bolster the integrity of healthcare data systems, fostering trust and improving patient outcomes in a safer healthcare environment.

5.3 Safety and Legal Impact

Our project on data breach detection and prevention using machine learning and encryption has significant safety and legal impacts. By providing an additional layer of security to healthcare data, we can help prevent identity theft, medical fraud, and compromised patient safety. Additionally, our system can help healthcare organizations comply with legal and regulatory requirements, po- tentially mitigating the legal risks associated with data breaches. Overall, our project has the potential to significantly

enhance the safety and legal compliance of healthcare data systems, ultimately benefiting both healthcare organizations and patients.

5.4 Cultural Impact

Our project on data breach detection and prevention using machine learning and encryption has significant cultural impacts. By promoting data privacy and security in healthcare, we can help improve patient confidence in healthcare organizations and foster a culture of trust and trans- parency. Additionally, our system can help promote the adoption of best practices in data security across healthcare organizations, ultimately improving the overall culture of data privacy and secu- rity. Overall, our project has the potential to significantly impact the cultural norms surrounding data privacy and security in healthcare, leading to a more trustworthy and patient-centered health- care environment.

5.5 Environmental Impact

While our project on data breach detection and prevention using machine learning and encryption may not have a direct environmental impact, we recognize the potential for indirect contributions to environmental sustainability. By improving the security and privacy of healthcare data, we can potentially reduce the amount of paper and other resources used to address the consequences of data breaches. Furthermore, if our system leads to improved patient outcomes and reduced health- care costs, we can potentially free up resources for other environmental initiatives. As a result, we remain committed to exploring how our project can contribute to broader environmental sustain- ability efforts.

5.6 Sustainability

Sustainable practices can be supported by the project's capacity to improve information retrieval and decision-making processes. It can make it easier for people and organizations to adopt sustain- able methods in a number of different industries, including energy, agriculture, and urban planning, by facilitating efficient access to pertinent information. More convenient access to knowledge that has been condensed can help people make better decisions that put environmental preservation and resource management first.

Chapter 6
Conclusion and Future Work

This work uses machine learning models to detect data breaches in cloud computing. It will help to prevent data theft and unauthorized access to personal or confidential data. Here we used the encryption-decryption method for safety purposes. The University of New South Wales provided the data set we worked on. We used Random Forest Classifier and ADA Boosting Algorithms. Both of them worked great for the data set we worked on. All the results are shown in the 'result and discussion' section.

The future goal is to implement re-authentication if any suspicious data is found. Regarding data communication, it is nearly impossible to stop a data breach. However, we can take proper safety measures to minimize the number. Initially, we plan that if any suspicious data is found, the transfer will be held immediately and redirected to authentication. This will prevent any unauthorized entry.

Appendix

The Appendix serves as a comprehensive repository of supplementary materials to augment the primary content of the report. It includes an expansive dataset that provides a detailed view of the refined information, promoting transparency and facilitating a more thorough examination. Visual aids, such as additional charts and graphs, accompany the dataset, offering a visual context to complement the textual analysis and represent patterns or correlations not explicitly covered in the main report. Additionally, the appendix incorporates any pertinent information that adds depth to the study, including detailed algorithm descriptions or extended explanations of analytical processes. By encompassing these supplementary materials, the appendix aims to provide a complete and detailed perspective on the research process, fostering transparency, reproducibility, and offering valuable insights for readers, researchers, and analysts seeking a deeper understanding of the study's foundations and outcomes.

References

[1] Miro.Medium, available at << https://rb.gy/b2ktm>>, last accessed on 06-06-2024 at 12:00 PM.

[2] Lata, S. and Singh, D., 2022. Intrusion detection system in cloud environment: Literature sur- vey & future research directions. International Journal of Information Management Data In- sights, 2(2), p.100134.

[3] Singh, P. and Ranga, V., 2021. Attack and intrusion detection in cloud computing using an en- semble learning approach. International Journal of Information Technology, 13, pp.565-571.

[4] J. Zhang, M. Zulkernine and A. Haque, "Random-Forests-Based Network Intrusion Detection Systems," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 38, no. 5, pp. 649-659, Sept. 2008, doi: 10.1109/TSMCC.2008.923876.

[5] Machine Learning Random Forest Algorithm - Javatpoint. (n.d.). www.javatpoint.com.

[6] Z. Masetic, K. Hajdarevic and N. Dogru, "Cloud computing threats classification model based on the detection feasibility of machine learning algorithms," 2017 40th International Con- vention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2017, pp. 1314-1318, doi: 10.23919/MIPRO.2017.7973626

[7] U. Bashir and M. Chachoo, "Intrusion detection and prevention system: Challenges & oppor- tunities", 2014 International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2014, pp. 806-809, doi: 10.1109/IndiaCom.2014.6828073.

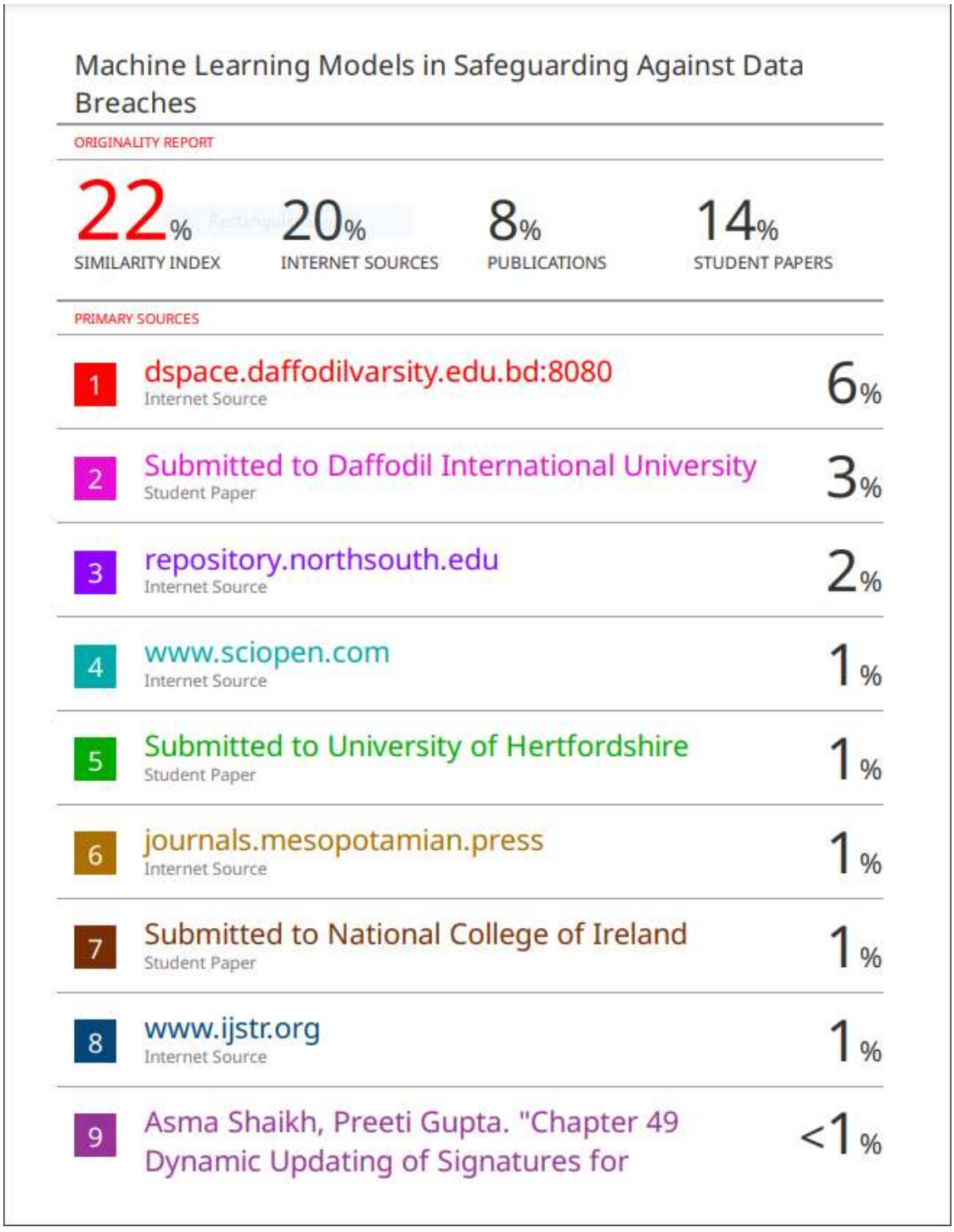
[8] F. I. Shiri, B. Shanmugam and N. B. Idris, "A parallel technique for improving the perfor- mance of signature-based network intrusion detection system," 2011 IEEE 3rd International Conference on Communication Software and Networks, Xi'an, China, 2011, pp. 692-696, doi: 10.1109/ICCSN.2011.6014986.

[9] M. A. Ferrag, L. Shu, O. Friha and X. Yang, "Cyber Security Intrusion Detection for Agriculture 4.0: Machine Learning-Based Solutions, Datasets, and Future Directions," in IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 3, pp. 407-436, March 2022, doi: 10.1109/JAS.2021.1004344.

[10] mahendradata/cicids2017-ml: The purpose of this repository is to demon- strate the steps of processing CICIDS2017 dataset using machine learning algorithms,https://github.com/mahendradata/cicids2017-ml,GitHub,July 3, 2023

[11] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detec- tion systems (UNSW-NB15 network data set)," 2015 Military Communications and Informa- tion Systems Conference (MilCIS), Canberra, ACT, Australia, 2015, pp. 1-6, doi: 10.1109/Mil- CIS.2015.7348942.

[12] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani, “Toward Generating a New In- trusion Detection Dataset and Intrusion Traffic Characterization”, 4th International Confer- ence on Information Systems Security and Privacy (ICISSP), Portugal, January 2018



Machine Learning Models in Safeguarding Against Data Breaches

ORIGINALITY REPORT

22%	20%	8%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	Submitted to Daffodil International University Student Paper	3%
3	repository.northsouth.edu Internet Source	2%
4	www.sciopen.com Internet Source	1%
5	Submitted to University of Hertfordshire Student Paper	1%
6	journals.mesopotamian.press Internet Source	1%
7	Submitted to National College of Ireland Student Paper	1%
8	www.ijstr.org Internet Source	1%
9	Asma Shaikh, Preeti Gupta. "Chapter 49 Dynamic Updating of Signatures for	<1%