National College of Ireland

Strengthening Proactive Cyber Defence: Innovative Approaches for Effective Cyber Threat Intelligence Gathering, Analysis and Application

MSc Research Project
Cybersecurity

Chandhiya Ramasamy
Student ID: X22105042

School of Computing
National College of Ireland

Supervisor: Jawad Salahuddin

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Chandhiya Ramasamy
Student ID:	X22105042
Programme:	MSc. Cybersecurity **Year:** 2023-2024
Module:	Research Project
Supervisor:	Jawad Salahuddin
Submission Due Date:	14th December 2023
Project Title:	Strengthening Proactive Cyber Defence: Innovative Approaches for Effective Cyber Threat Intelligence Gathering, Analysis and Application
Word Count:	5838 **Page Count:** 19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: .....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Strengthening Proactive Cyber Defence: Innovative Approaches for Effective Cyber Threat Intelligence Gathering, Analysis, and Application

Abstract

In response to the escalating threat of data poisoning assaults on machine learning-based security systems in cyber threat intelligence (CTI), this research introduces an innovative methodology. Leveraging the algorithms Isolation Forest, Logistic Regression, and Support Vector Machines (SVM), the study addresses the critical need to enhance system resilience. Through experimentation with a synthetic CTI Common Vulnerabilities and Exposures (CVE) dataset, feature selection, and rigorous model training, the study observed that Logistic Regression and Support Vector Machines (SVM) outperformed Isolation Forest. The comparative analysis of different models revealed distinct performance metrics, identifying Logistic Regression and SVM as particularly adept in identifying data poisoning threats and demonstrating resilience across a variety of conditions. The study's theoretical contribution lies in advancing anomaly detection within CTI datasets, aligning with the current state of the art while introducing a novel combination of established techniques. In practice, this research fortifies machine learning-based security mechanisms, providing tangible protection against data tampering and enhancing the reliability of CTI outputs. Remaining unresolved aspects offer avenues for future work, emphasizing hyperparameter optimization, exploring additional anomaly detection techniques, and practical deployment scenarios. These opportunities signify potential refinement and extension of the proposed methodology in the dynamic landscape of cyber threat intelligence.

Keywords: Cyber Threat Intelligence, Data Poisoning assaults, Isolation Forest, Logistic Regression, Support Vector Machines, Hyperparameter optimization, Anomaly detection techniques.

1 Introduction

In the rapidly evolving landscape of cybersecurity, Cyber Threat Intelligence (CTI) stands as a critical pillar in fortifying defenses against potential threats and attacks. CTI involves the collection, analysis, and interpretation of data from various sources **(Ranade et al., 2021)** to proactively identify and mitigate potential cyber threats. However, this indispensable aspect of cybersecurity is not immune to challenges, and one of the pervasive threats it faces is data poisoning.

Data poisoning in CTI refers to the malicious introduction of fabricated or manipulated data into machine learning-based security systems. This insidious practice compromises the integrity and reliability of CTI, raising concerns about the efficacy of security measures built upon these foundations **(Li and Liu, 2021)**. The urgency to address this issue stems from the escalating sophistication of cyber threats and the growing dependence on CTI for preemptive security measures.

Understanding and mitigating the impact of data poisoning in CTI is imperative for the resilience of modern cybersecurity frameworks. The interconnectedness of data, artificial intelligence, and machine learning in the CTI landscape makes it susceptible to adversarial manipulations **(Yaacoub et al., 2022)**. As such, there is a critical need to delve into innovative approaches that bolster proactive cyber defense, ensuring the reliability of CTI outputs and

fortifying the foundations upon which cybersecurity decisions are made **(Apruzzese et al., 2023)**. This research contributes to the scientific literature by addressing a critical gap in current cybersecurity discourse.

The focal point of this research initiative revolves around a critical query: How can machine learning (ML)-based security systems be fortified against the insidious threat of data poisoning attacks within the domain of Cyber Threat Intelligence (CTI)? In addressing this overarching question, the study embarks on a multifaceted exploration aimed at deciphering the intricate challenges posed by data poisoning in CTI. The research scrutinizes existing literature comprehensively, surveying the state of the art to unravel the nuances of data poisoning threats in the CTI landscape. With this foundational understanding, the investigation pivots towards the design and implementation of a robust methodology. Leveraging suitable anomaly detection algorithms, the research endeavors to craft an innovative approach that not only identifies but also prevents data poisoning. The subsequent evaluation phase rigorously assesses the effectiveness of the implemented algorithms, providing insights into their viability for fortifying ML-based security systems against the omnipresent threat of data poisoning in the dynamic cybersecurity realm.

The Isolation Forest algorithm, an algorithm for machine learning that shows promise in spotting odd patterns inside datasets a crucial ability in the setting of cyber threats intelligence is the subject of this inquiry. Renowned for its efficacy in identifying anomalies within high-dimensional datasets, Isolation Forest stands out as an ideal solution tailored to the intricate nature of CTI. Its decision tree-based approach facilitates swift isolation of abnormalities, making it an optimal choice to address the nuanced challenge of detecting fabricated data within Common Vulnerabilities and Exposures (CVE) datasets in CTI.

Logistic Regression is underpinned by its versatility and interpretability. It excels in binary classification tasks, making it adept at distinguishing normal and anomalous patterns within CTI datasets. Its simplicity, coupled with the ability to provide probability estimates, renders it a pragmatic choice for the nuanced task of discerning fake data within CVE datasets.

Support Vector Machines (SVM) represent a formidable algorithmic solution which is underpinned by its efficacy in high-dimensional spaces and its versatility in handling non-linear relationships within datasets. SVM's ability to delineate complex decision boundaries adds a layer of sophistication to anomaly detection tasks, making it a robust choice for identifying fake data in CTI datasets. Its capability to discern subtle patterns within the data landscape further solidifies its role in fortifying the security infrastructure against the insidious threat of data poisoning.

A significant component of the method's effectiveness is the careful selection of important features from the artificial dataset. By identifying characteristics that are suggestive of data poisoning, aberrant patterns can be detected by algorithms. By introducing the simulation to these particular features during the training phase, algorithms help the model identify and isolate cases of data corruption. This focused training prepares the groundwork for comparison with more established machine learning scenarios, namely SVM and Logistic Regression. This research thoroughly assesses the Isolation Forest algorithm's performance in comparison to traditional models like SVM and Logistic Regression. This comparison analysis shows the shortcomings of current models in facing the threat of data poisoning incidents in addition to highlighting the effectiveness of the suggested technique (**Togbe et al., 2020)**.

The subsequent sections of this report will unfold systematically. Section 2 of this paper offers a novel strategy to strengthen machine learning-based security measures against attacks involving data poisoning in the setting of CTI. Following this, the motivation for the study will be substantiated with pertinent literature. A detailed description of the creation of the CTI data set, selection requirements, and Isolation Forest model training is provided in section 3. Section 4 presents the experiment's findings, which demonstrate the model's flexibility in dealing with various situations, including hyper parameter optimization and alternative techniques for anomaly detection. The study concludes in part 5 with a discussion of its consequences for real-world cybersecurity applications. Section 6 provides an overview of the study's conclusions.

2 Related Work

As I delved into the wealth of literature, a common thread emerged: the looming threat of data poisoning in various industries. These scholarly works collectively highlight the crucial role of machine learning and artificial intelligence in tackling this pervasive challenge. The papers eloquently discuss the extensive impact of manipulated data across sectors, emphasizing the need to refine the training and detection mechanisms of ML/AI algorithms. Additionally, some studies contribute valuable insights into evaluating the performance metrics of these algorithms, enhancing our understanding of their effectiveness.

Potential real-world consequences of Fake Cyber Threat Intelligence

This paper (**Ranade, P et al., 2021)** addresses a critical concern in the field of cybersecurity by demonstrating how transformer-based models, specifically GPT-2, can be fine-tuned to automatically generate fake Cyber Threat Intelligence (CTI) text descriptions. The research is well-structured and clearly articulates the potential risks associated with fake CTI, emphasizing its use in data poisoning attacks on Cybersecurity Knowledge Graphs (CKG) and other AI-based cyber defense systems. The inclusion of a human evaluation study involving cybersecurity professionals adds a valuable perspective, highlighting the believability of the generated fake CTI among experts. The paper successfully showcases the adverse impacts of ingesting fake CTI, including incorrect reasoning outputs, representation poisoning, and model corruption, underscoring the need for defenses against such attacks.

Fake News Detection Using Machine Learning Algorithm

The paper by (**Chauhan, R et al., 2023)** addresses the rampant spread of fake news in the digital era and proposes a machine learning (ML) method to identify and classify fake news accurately. The approach involves training ML models such as SVMs, Random Forests, and DNNs to distinguish between real and fake news by analyzing a large dataset and extracting textual and contextual information. The study concludes with a robust automated system demonstrating excellent durability and accuracy in real-time counterfeit detection. The strength of this approach lies in its ability to efficiently recognize subtle patterns and characteristics inherent in false news, enhancing accurate identification. However, a potential limitation is the constant evolution of false news, requiring ongoing research and continuous improvements to ML models to keep pace with emerging challenges and ensure the efficacy of the fake news detection system. Despite this, the paper signifies a significant achievement in integrating ML technology to combat false information, providing a scalable and efficient strategy to tackle the escalating volume of misleading content. The proposed solution reduces reliance on manual fact-checking, saving significant time and resources in the battle against misinformation.

(Al Asaad, B., & Erascu, M. 2018) tackles the pervasive issue of fake news and disinformation in the post-truth era, employing machine learning techniques, particularly supervised learning, for detection. Using a dataset of fake and real news, the study trains a machine learning model with the Scikit-learn library in Python, employing text representation models such as Bag-of-Words, TF-IDF, and Bi-gram frequency. Two classification approaches, probabilistic and linear, are tested on the title and content, addressing clickbait/non-clickbait and fake/real distinctions. The results highlight the superiority of linear classification, particularly with the TF-IDF model in content classification, while the Bi-gram frequency model performs less accurately in title classification. The paper acknowledges the complex nature of controlling information flow online but emphasizes the importance of ongoing research to combat the spread of misinformation. The strength of this work lies in its attempt to verify news articles' credibility using a combination of classification methods and text models, yielding relatively satisfying accuracy results. However, a potential weakness is the acknowledgment that achieving higher accuracy necessitates a more sophisticated algorithm, possibly incorporating data mining technologies with big data, to enhance the dataset's inclusivity and improve accuracy scores. Future work is outlined to delve deeper into the combination of feature extraction methods and classifiers for optimal performance and to explore advanced algorithms with big data applications to further improve accuracy in detecting fake news.

Comparison and Performance Evaluation of Machine Learning Algorithms

Research by (**Leela Siva Rama Krishna, N., & Adimoolam, M. 2022)** aims to achieve precise fake news detection by employing Logistic Regression (LR) and comparing the accuracy of textual property with the Support Vector Machine (SVM) algorithm. The results indicate that the LR algorithm exhibits an accuracy of 95.12%, while the SVM algorithm achieves an accuracy of 91.68%. With a significance value of 0.079 for accuracy and 0.125 for precision, there is a statistically significant difference between the sample groups. The conclusion highlights the LR algorithm's superior accuracy in identifying fake news compared to the SVM algorithm. The study proposes the use of LR for detecting fake news, emphasizing its potential to save time and simplify the process compared to SVM. This finding contributes to the ongoing efforts to develop effective and efficient systems for combatting the proliferation of fake news.

The work by (**Reis et al., 2019)** delves into the detection of fake news stories disseminated on social media, exploring various features extracted from news stories, including source and social media posts. In addition to assessing existing features proposed in the literature for fake news detection, the paper introduces a new set of features and evaluates the prediction performance of current approaches. The results yield insights into the effectiveness and importance of features for detecting false news. The evaluation involves classic and state-of-the-art classifiers, such as k-Nearest Neighbors (KNN), Naive Bayes (NB), Random Forests (RF), Support Vector Machine with RBF kernel (SVM), and XGBoost (XGB). The effectiveness is measured using the area under the ROC curve (AUC) and the Macro F1 score, with RF and XGB classifiers exhibiting the best results, statistically tied with AUC values of 0.85 (± 0.007) and 0.86 (± 0.006), respectively. The study emphasizes the relevance of the AUC for fake news detection, allowing for control over the tradeoff between true and false positive rates. While presenting challenges and opportunities, the paper contributes to advancing the field of fake news detection through an extensive evaluation of features and classifiers, providing valuable insights for practical implementation.

This article performs an extensive evaluation of eight machine learning algorithms representing diverse paradigms for fake news detection. The paper provides valuable insights into the relative performance, efficiency, and training speed of regression, SVM, MLP, naive Bayes, random forests, decision trees, and CNNs across various datasets. The inclusion of a public Web-based application for real URL testing enhances the practical applicability of the study. The identification of convolutional neural networks (CNNs) as the best-performing algorithm, despite higher training time, adds a significant contribution to the ongoing discourse on fake news detection. While the paper (**Katsaros, D et al., 2019)** effectively contrasts the algorithms' efficiency and training speed, it could provide more nuanced insights into the interpretability of the chosen models. Additionally, discussing potential challenges or limitations associated with each algorithm's real-world deployment would strengthen the paper's practical implications. A more comprehensive exploration of the trade-offs between accuracy and training time, especially concerning CNNs, would enhance the completeness of the conclusions.

(Gupta, Vet al., 2022) addresses the contemporary challenge of fake news proliferation on online platforms by leveraging machine learning algorithms for automated classification. It contributes to the field by exploring various textual properties and employing Natural Language Processing (NLP) techniques for data pre-processing, enhancing the accuracy of the machine learning models. The study employs a comprehensive set of supervised machine learning algorithms, including Logistic Regression, Support Vector Machine, Naive Bayes, K-Nearest Neighbour, Random Forest, and Decision Tree, achieving notable accuracies. The detailed discussion on the performance metrics and the comparison of tree-based algorithms adds valuable insights. The paper also highlights potential areas for improvement, such as the exploration of additional classifiers, feature selection, and the integration of diverse feature extraction methods like bag-of-words and Word2Vec.

2.1 Problem Statement

Interestingly, a noticeable gap exists in the literature—a lack of exploration into preprocessing strategies tailored for the Common Vulnerabilities and Exposures (CVE) dataset within Cyber Threat Intelligence (CTI). My research addresses this gap, emphasizing the vital importance of deploying ML algorithms to identify and eliminate fake data within the CVE dataset. This approach aims to strengthen the proactive capabilities of cyber threat intelligence tools by ensuring the integrity of information before its integration into ML-based frameworks.

3 Methodology

In this research, we employ a set of anomaly detection tools, namely Isolation Forest, SVM, and Logistic Regression, as part of our methodology to identify compromised data within Cyber Threat Intelligence (CTI) before its integration into ML-based security systems. The primary goal of this proposed methodology is to enhance the resilience of ML models and improve the accuracy of CTI analysis by effectively isolating anomalies, including manipulated data. This strategic approach aims to fortify cybersecurity defenses against potential data poisoning attacks. Additionally, through a comprehensive performance evaluation, we intend to recommend the most effective approach for bolstering the security of CTI and ML models. Fig. 1 shows the flowchart for the deployment of model.

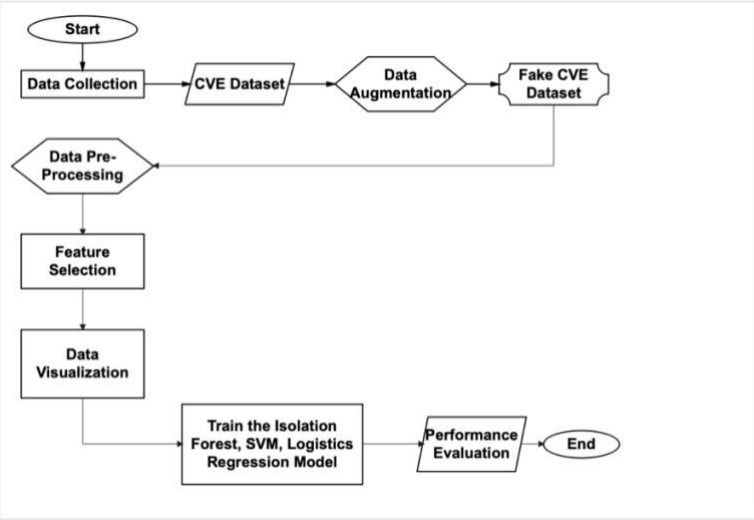

Fig. 1. Flowchart of the Model Deployment

A comprehensive description of each step is given below:

3.1 Data Collection

During the data collection phase, the Common Vulnerabilities and Exposures (CVE) dataset was sourced from Kaggle[1] as the primary dataset for this study. To enhance the dataset's diversity and simulate potential adversarial scenarios, a data augmentation method was applied.

3.2 Data Augmentation

Data augmentation approaches are utilized to improve the generalization and stability of the SVM, Logistic Regression, and Isolation Forest models. To diversify the training dataset for the Isolation Forest, artificial data points are created to indicate differences in real and fake CVEs. This guarantees that the model improves its ability to identify minute irregularities in new cases. When using SVM, methods such as bootstrapping and feature space modification are made available to the data used for training, enhancing its diversity and improving the accuracy of classification. In a similar vein, the dataset is enhanced for Logistic Regression using methods like noise introduction or oversampling, which enables the model to identify a wider variety of patterns and correlations. Together, these methods of data enhancement strengthen the models' accuracy and robustness, ensuring their effectiveness in cyber threat intelligence scenarios when their capacity to adjust to changing threats is essential. These artificially created instances were then seamlessly integrated with the originally downloaded file, resulting in an enriched dataset that encompasses both genuine and manipulated information.

3.3 Data Pre-Processing

In the data pre-processing phase for the Isolation Forest approach, CVE dataset undergoes meticulous preparation to ensure its suitability for effective machine learning analysis. To facilitate model training, the training set is subjected to a series of pre-processing steps. These steps include the normalization and scaling of features, handling missing or inconsistent data, and encoding categorical variables.

[1] Available at https://www.kaggle.com/datasets/andrewkronser/cve-common- vulnerabilities-and-exposures

3.4 Feature Selection

The selection of critical features necessary for identifying data poisoning is carefully considered throughout the compilation of the CVE dataset. Carefully selected features, such as anomalous trends and inconsistencies, are suggestive of data tampering. These characteristics could include a range of elements, such as the degree of vulnerability, exploitation patterns, time-related factors, and background data on the CVE entries. To improve the Isolation Forest algorithm's ability to detect and lessen the effects of competitive data corruption in later cybersecurity applications, the selection process tries to give it a prejudiced set of parameters that allow it to distinguish between real and controlled entries throughout the training phase **(Xiao et al., 2015)**. The selection of features step fortifies the system's capacity to recognize and reduce data poisoning dangers in the CTI domain by identifying critical attributes that differentiate authentic from manipulated entries.

3.5 Data visualization

When collecting textual elements like CVE explanations, data vectorization is an essential pre-processing step. In this case, the label encoding method is used. Label encoding is a technique where categorical data, such as textual labels, is converted into numerical vectors. This numeric representation facilitates the conversion of textual data into a format suitable for machine learning algorithms, which often require numerical input.

3.6 Dataset splitting

The dataset undergoes a division into training and testing sets utilizing the 'train_test_split' function from scikit-learn. This function facilitates a systematic partitioning of the dataset, ensuring a comprehensive approach to model training and evaluation.

The function is equipped with parameters named X, y, test_size, and random_state, each serving a specific purpose:

- ‘X’: Represents the feature matrix containing the input data.

- ‘y’: Represents the target variable or labels corresponding to each instance in the feature matrix.

- test_size=0.1:

This parameter determines the proportion of the dataset allocated for testing. In this case, ‘test_size=0.1’ indicates that 10% of the data will be used for testing, while the remaining 90% will be used for training. A 90-10 split (90% for training and 10% for testing) is a common and reasonable choice, especially when dealing with relatively large datasets. It provides sufficient data for training while reserving a separate portion for evaluating the model's performance.

- random_state=42:

The ‘random_state’ parameter is set to 42, serving as the seed for the random number generator. This ensures that the randomness introduced in the Isolation Forest algorithm is reproducible. By using a fixed seed, you can obtain consistent results when running the code multiple times. The value '42' is a commonly used standard seed in the algorithm, ensuring consistency and reproducibility across different runs.

3.7 Train the algorithms

3.7.1 Isolation Forest

The Isolation Forest method is trained to discriminate between genuine and bogus CVEs. During this procedure, the method is fitted to the prepared dataset. It is expected that bogus CVEs will have shorter paths in the tree topology than real entries. By giving bogus CVEs shorter routes in the decision trees during training, the Isolation Forest technique requires use of the inherent variations in anomalous behaviour **(Laskar *et al.*, 2021)**. This leads to the generation of unique decision trees for every false CVE, which helps the system detect anomalies in the dataset. In the field of cyber threat intelligence, this training methodology guarantees that the Isolation Forest algorithm is proficient in identifying the distinct attributes linked to data poisoning risks, consequently enhancing its resistance against cyber-attacks.

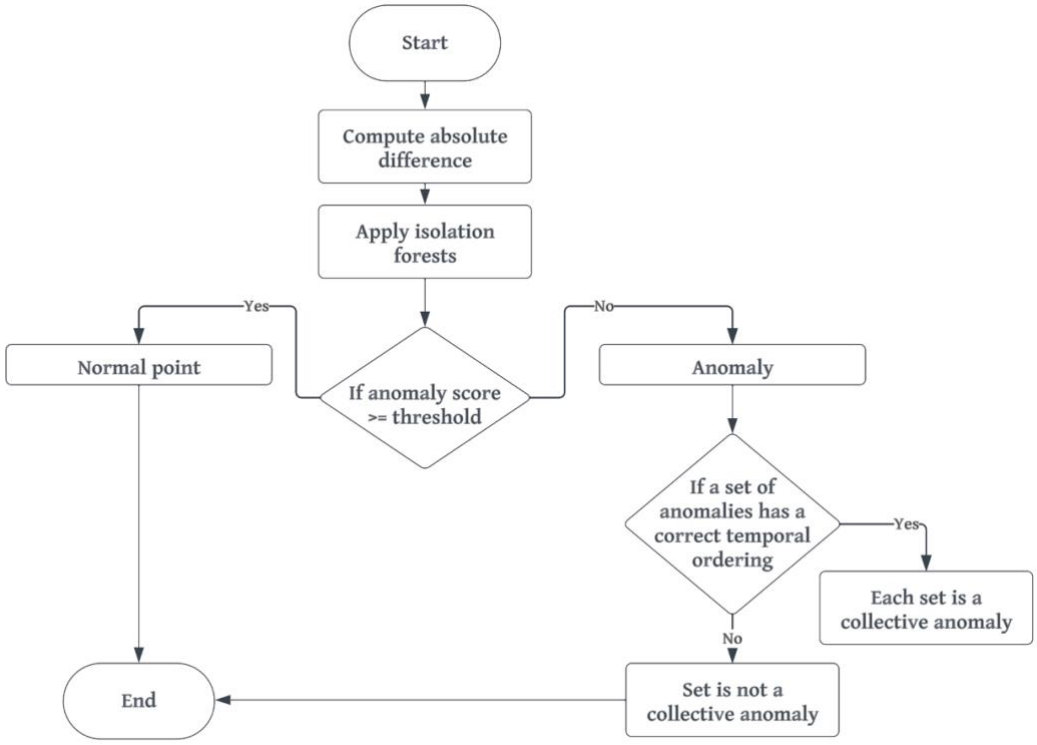

Fig. 2. Flowchart of the Isolation Forest Model

3.7.2 SVM

By identifying the best hyperplane in the feature space that optimally differentiates real and fake CVEs, the SVM is trained to identify patterns in the dataset. The SVM's parameters are iteratively adjusted during the training process to reduce classification errors while preserving the greatest possible difference between classes. Effective generalization to fresh, unused data is ensured by this margin **(Blanco, Japón and Puerto, 2020)**. The SVM can efficiently classify and differentiate between real and altered CVEs after it has been trained on the previously created dataset.

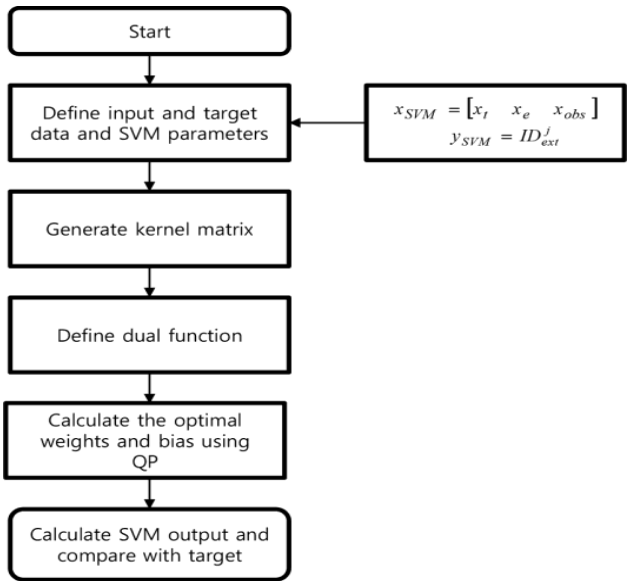

Fig. 3. Flowchart of SVM

3.7.3 Logistics Regression

A supervised learning approach called logistic regression models the likelihood of an outcome that is binary, in this instance, the categorization of real or fake CVEs. After initializing the model, the logistical loss function is minimized iteratively by using the training set to update the biases and weights. By using this optimization procedure, the logistic regression approach is guaranteed to discover the underlying connections and trends in the dataset **(Adeyemo, Wimmer and Powell, 2019).** Determining the ideal parameters that determine the decision boundaries between various classes is the focus of the training phase. After having been trained, the logistic regression algorithm can forecast fresh data and provide a probabilistic assessment of how likely it is that an entry is real or fraudulent.

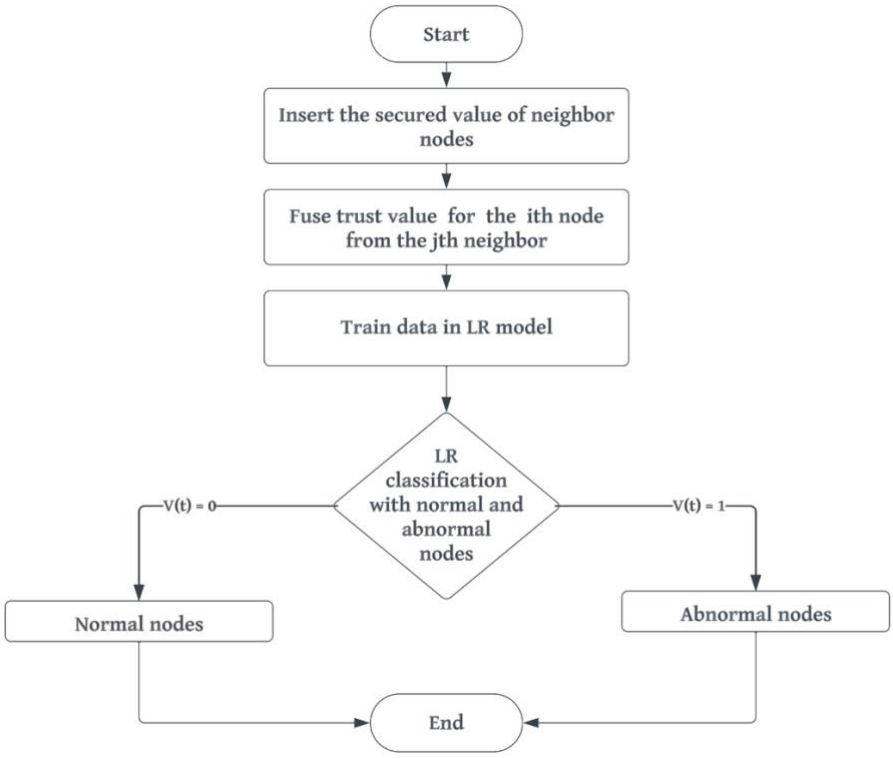

Fig. 4. Flowchart of the Logistics Regression Model

3.8 Performance Evaluation

The last phase of this method will utilize standard performance criteria for anomaly detection, including accuracy, precision, recall, and F1-score, to investigate the efficacy of the approach in detecting data poisoning (**Deepa et al., 2022)**.

Precision signifies the proportion of correctly identified instances among those retrieved. It is computed using the following formula:

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

Recall represents the proportion of relevant instances that have been successfully retrieved. It is calculated as follows:

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

Accuracy represents the proportion of correct predictions in the test dataset and is determined by the following formula:

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

F1 score is the weighted average of Recall and Precision.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The algorithm's efficiency improves as these evaluation metrics attain higher values.

4 Implementation

4.1 Tools and Test Data

Having obtained the Real Common Vulnerabilities and Exposures (CVE) Dataset from Kaggle[2]. It is a public-domain dataset without rights (**Kronser, 2020**). The CVE dataset contains details on known security exposure and vulnerabilities. It gives information regarding the vulnerability, how dangerous it is, and links to any easily available patches or solutions. All the methods were utilized from the scikit-learn library included with the Python programming language (Scikit-learn, 2023).

4.2 Correlation Matrix

Correlation matrices are used in data analysis to understand relationships between different variables in a dataset.

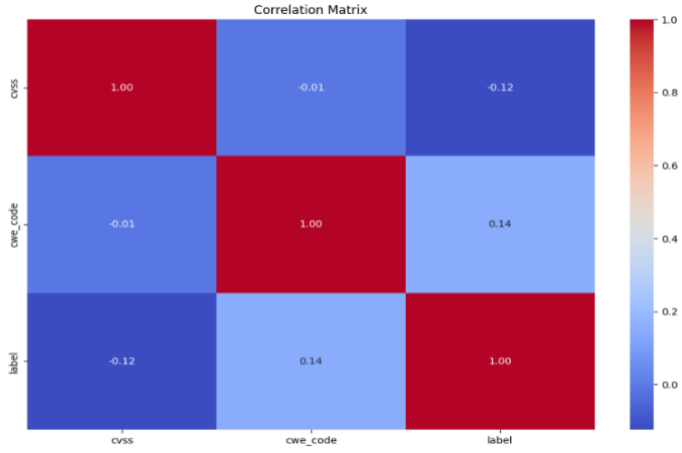

Fig. 5. Correlation Matrix

[2] Available at https://www.kaggle.com/datasets/andrewkronser/cve-common- vulnerabilities-and-exposures

The correlation matrix analysis revealed that features with low or near-zero correlation with the target variable (Label) might not contribute substantially to distinguishing between real and controlled entries. Therefore, I can prioritize features exhibiting significant relationships (values close to 1) with the target variable while steering clear of multicollinearity. Fig.5 exemplifies one of several correlation matrices I utilized to identify pivotal features for my model where it visually represents the correlation matrix involving the features cvss, cwe_code, and the target variable (label). Similarly, when analysing all other features against target variable (label), I found out that cwe_code, cwe_name and access_authentication features play crucial role in my work to distinguish between real and controlled entries.

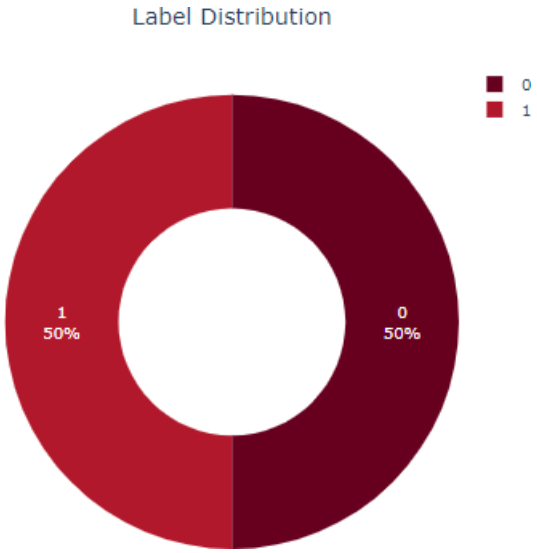

Fig. 6. Label Distribution

The Label Distribution fig. 6 gives a visual depiction of the imbalances or balance between the various groups by showing the frequency or distribution of distinct classification labels throughout a dataset.

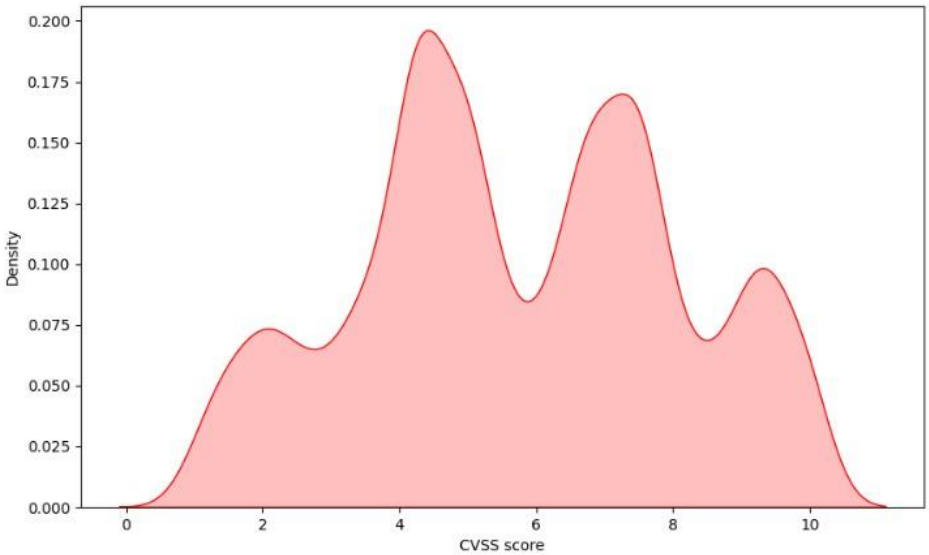

Fig. 7. CVSS Score Vs Density

The distribution of vulnerabilities according to severity scores is shown in Fig. 7, which also shows the connection among density and Common Vulnerability Scoring System (CVSS) scores. The purpose of this figure is to demonstrate how vulnerabilities are concentrated in different CVSS score varies and the way this affects the whole system danger.

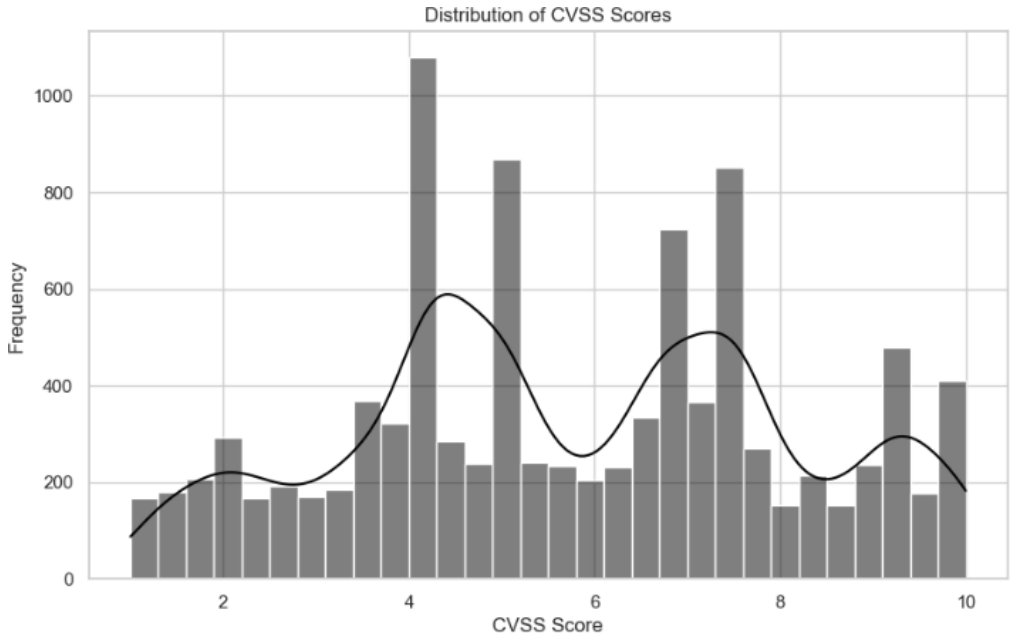

Fig. 8.Distribution of CVSS Scores

A graphical representation of the severity levels is given by Fig. 8, which shows the distribution of CVSS scores among vulnerabilities. The graph also helps to comprehend the general vulnerabilities landscapes by showing the frequency and range of CVSS scores.

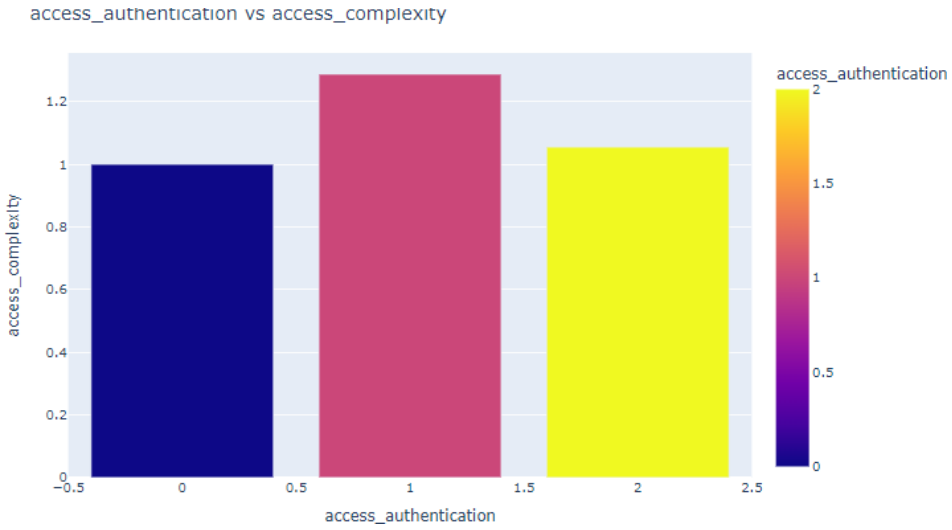

Fig. 9. Access_authentication Vs Access_complexity

The purpose of Fig. 9 is to illustrate the connection among access authentication and access complexity in the setting of cybersecurity by visualizing the associations or patterns among these two factors. The figure offers insights into the interaction between authentication demands and system access complexity, which helps to provide an in-depth comprehension of possible vulnerabilities in security.

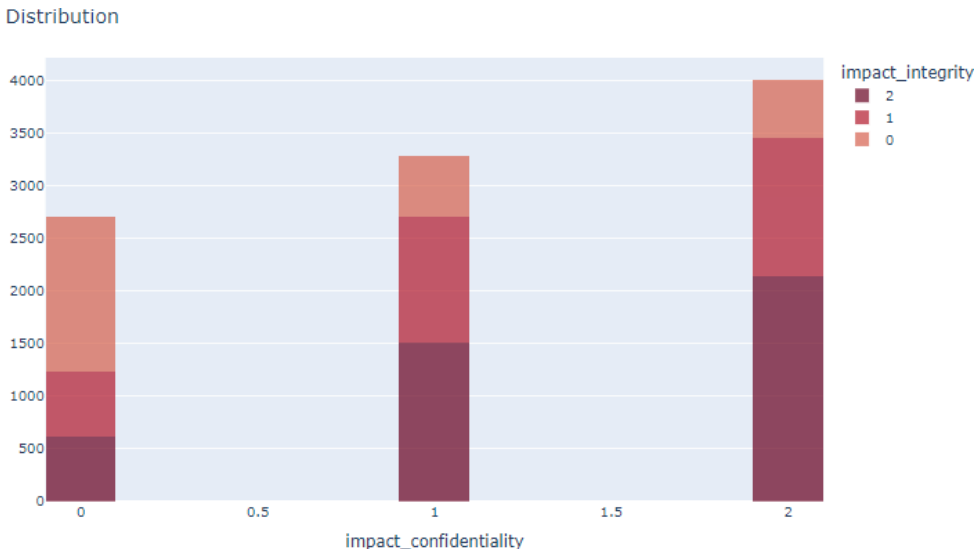

Fig. 10. Impact Confidentiality Vs Distribution

Knowing the possible scope and effects of security threats on private data is made easier by looking at Fig. 10, which illustrates the connection among their effect on privacy and distribution of vulnerabilities. This relationship provides an understanding of the way safety incidents affect the privacy of data according to their distribution.

4.3 Confusion Matrix

A useful tool for assessing efficiency is the confusion matrix, which offers a succinct overview of the classification model's output. The tabulation of counts for false positive, true positive, true negative, and false negative forecasts provides a clear picture of the model's performance in class distinction. This matrix is useful for evaluating a classification model's performance measures, such as accuracy, recall, and precision, which helps to provide a thorough knowledge of the model's advantages and disadvantages while handling different categories within a dataset.

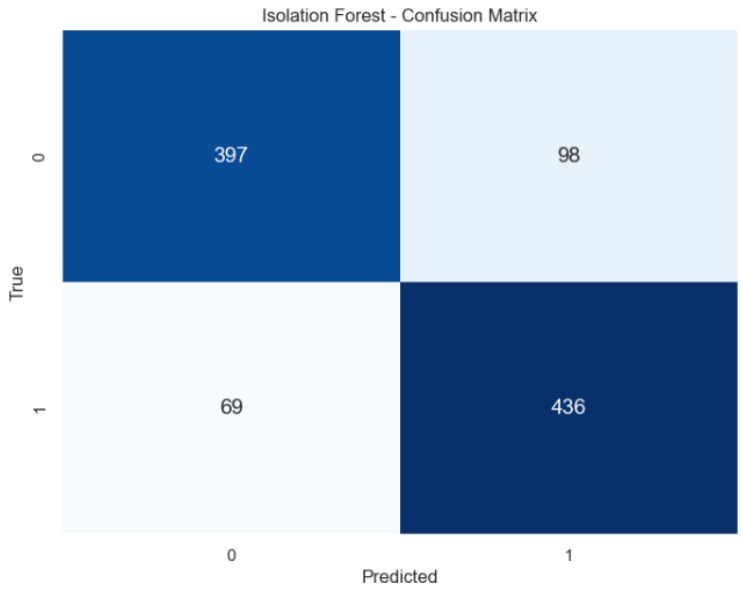

Fig. 11. Confusion Matrix of Isolation Forest

Fig. 11 shows False positives instances when genuine records are mistakenly classified as anomalies, true negatives demonstrate situations where anomalies are mistakenly classified

as genuine entries, and true positives indicate situations where anomalies have been correctly recognised as false entries in the confusion matrix for Isolation Forest. This matrix provides a visual representation of how well the Isolation Forest performed in identifying real data from fake within the CTI collection.

Fig. 12. Confusion Matrix of SVM

A Support Vector Machine's (SVM) confusion matrix of Fig. 12 shows the right and wrong forecasts of positive and negative examples, thereby summarising the model's efficacy in binary classification.

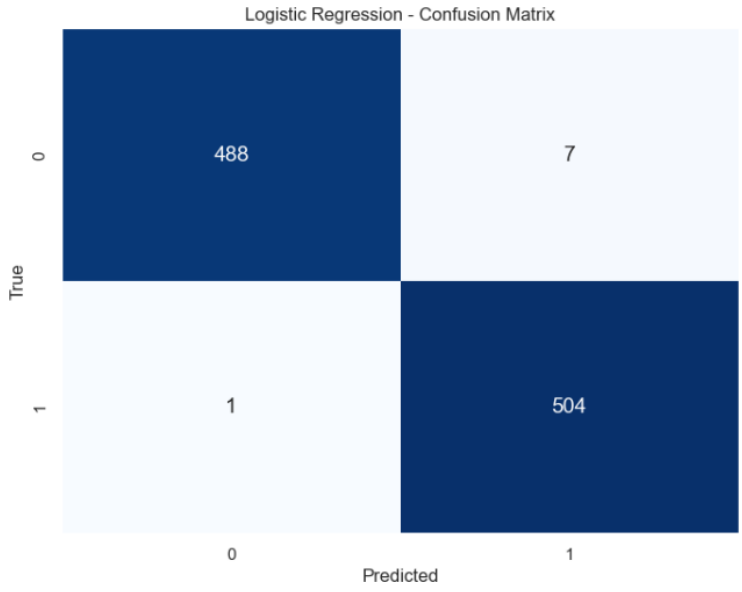

Fig. 13. Confusion Matrix of Logistics Regression

Fig. 13. Gives the confusion matrix for Logistic Regression includes true negatives, false positives, true positives, and false negatives, providing a concise summary of the model's performance in binary classification scenarios.

4.4 ROC Curve

The efficacy of a model for binary classification over different discriminating thresholds is shown graphically by the Receiver Operating Characteristic (ROC) curve for binary classification. As the discrimination threshold varies, it shows the trade-off between the true

positives rate and the false positive rate. Graphing these rates at various threshold values creates the curve, which serves as a visual aid for evaluating the model's capacity to discern among positive and negative occurrences. Higher performance of models, with higher specificity and sensitivity, is typically indicated by a steeper ROC curve and more AUC.

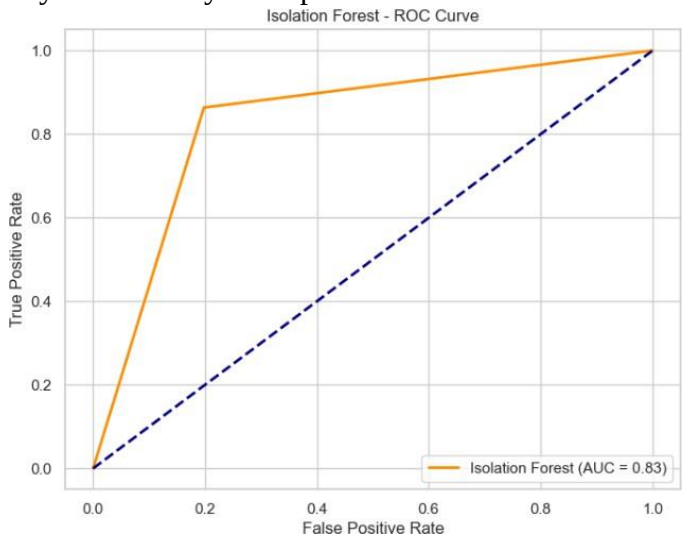

Fig. 14. ROC Curve of Isolation Forest

In Fig. 14, the ROC curve for Isolation Forest provides a visual depiction of the algorithm's capacity to discriminate between anomalies and real records in the cyber threat intelligence dataset.

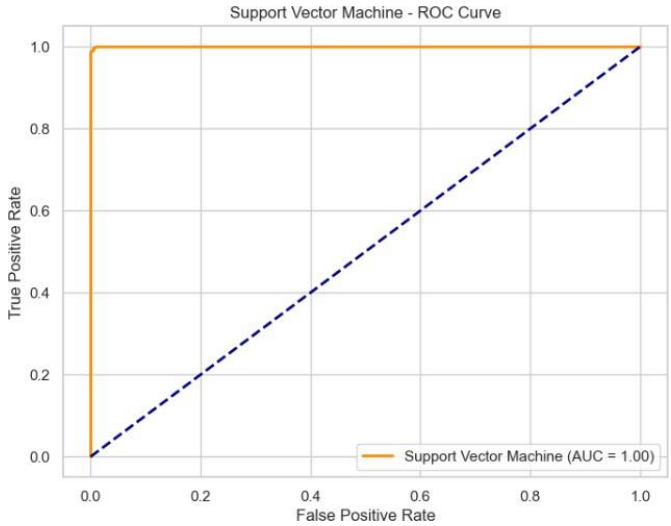

Fig. 15. ROC Curve for SVM

Fig. 15 offers a visual evaluation of the model's capacity to differentiate among positive and negative examples in binary classification situations. The ROC curve for SVM illustrates the balance among true positive rate with false positive rate.

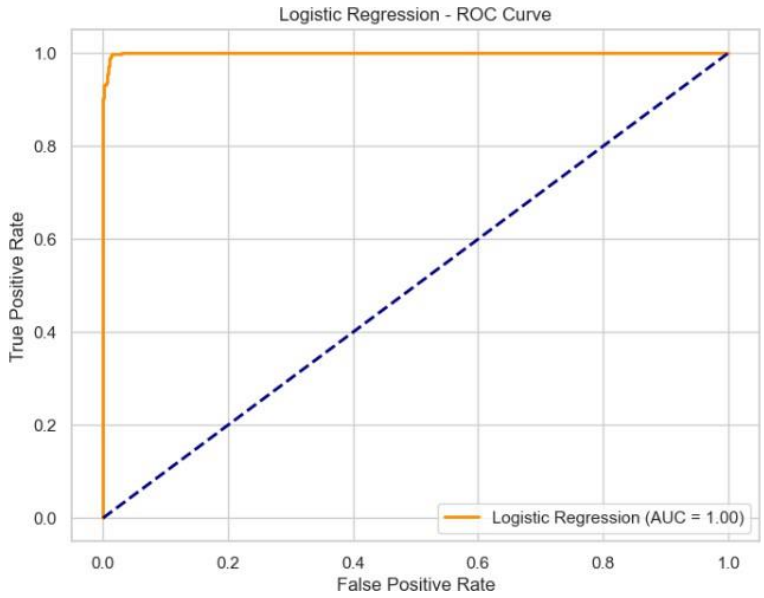

Fig. 16. ROC Curve of Logistics Regression

The Logistic Regression ROC curve of Fig.16 provides a visual depiction of the model's discriminating power in scenarios involving binary classification.

5 Evaluation

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Isolation Forest	83.3	82	86	84
SVM	99.6	99.6	99.6	99.6
Logistics Regression	99.2	98.63	99.8	99.21

Table 1: Evaluation Metrics

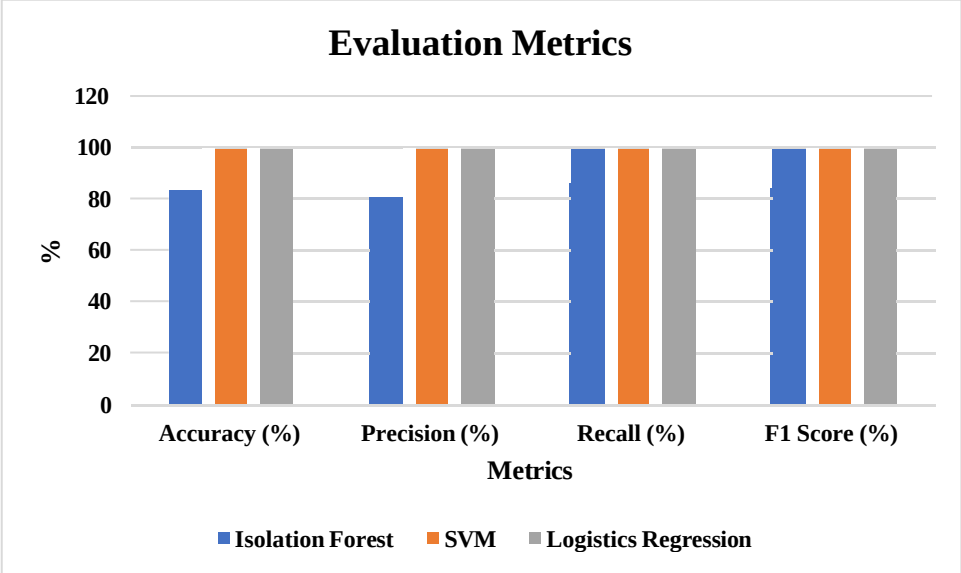

Fig. 17. Evaluation Metrics

The assessment metrics for the three distinct anomaly detection techniques Isolation Forest, SVM, and Logistic Regression are shown in Table 1. SVM and Logistic Regression beat Isolation Forest in terms of accuracy, with respective scores of 99.6% and 99.2%, as opposed to Isolation Forest's 83.3%. SVM and Logistic Regression show remarkable precision (99.6%) and recall (99.6% and 99.8%, respectively), while Isolation Forest shows somewhat lower precision (82%) and recall (86%) values. F1 results also demonstrate the higher performance of SVM and Logistic Regression over Isolation Forest, taking into account the trade-off between precision and recall. Comparing the results in Fig. 17 to Isolation Forest, it appears that SVM and Logistic Regression are more accurate and durable in identifying anomalies in the cyber threat intelligence dataset.

6 Discussion

While many research papers have showcased the effectiveness of the isolation forest method in accurately identifying false data, this study focused on classifying false CVE information within cyber threat intelligence and found that it exhibits lower performance. This observation underscores the fact that algorithm performance varies across different contexts and applications. However, the resilience exhibited in controlled trials highlights the possibility of enhancing machine learning-based safety mechanisms through the suggested method. The increased protection this study offers against dangers of data manipulation in CTI is one noteworthy result.

The Logistic Regression and SVM approach demonstrate its ability to identify abnormalities by isolating erroneous entries during training, which greatly increases the durability of security measures. This directly affects cybersecurity, a subject where it's critical to identify and mitigate altered data. To sum up, this research emphasizes how important the research findings are for strengthening machine learning-based security systems' resistance against data poisoning attacks in CTI.

7 Conclusion and Future Work

The objective of this study was to enhance the resilience of machine learning-driven security systems to attacks including data poisoning within the framework of cyber threat intelligence (CTI). The proposed method, which employs SVM, Logistic Regression, and the Isolation Forest technique, demonstrated potential in pinpointing anomalies more specifically, inaccurate entries in CTI datasets. Through an exacting feature selection procedure, individual method model training, and thresholder anomaly evaluation, the algorithm successfully identified CVE inputs as either safe or potentially harmful for data poisoning.

Subsequent research endeavours ought to concentrate on enhancing the Isolation Forest model via hyper parameter optimization in order to guarantee peak performance under various circumstances. Furthermore, investigating different anomaly detection methods and evaluating their relative merits will expand on our knowledge of anomaly detection in cyber threat intelligence. Putting the suggested methodology into practise in real-world cybersecurity situations is a critical step in confirming its effectiveness outside of controlled environments. These research directions optimization, exploration, and practical application are essential to bolstering machine learning-based safety measures against the dynamic array of cyber threats, particularly those related to data manipulation and tampering in the context of cyber threat intelligence.

References

Adeyemo, A., Wimmer, H. and Powell, L.M. (2019) 'Effects of normalization techniques on logistic regression in data science', *Journal of Information Systems Applied Research*, 12(2), p. 37.

Aghaei, E., Shadid, W. and Al-Shaer, E. (2020) 'Threatzoom: CVE2CWE using hierarchical neural network', *arXiv preprint arXiv:2009.11501* [Preprint].

Apruzzese, G. *et al.* (2023) 'The role of machine learning in cybersecurity', *Digital Threats: Research and Practice*, 4(1), pp. 1–38.

Berndt, A. and Ophoff, J. (2020) 'Exploring the value of a cyber threat intelligence function in an organization', in *Information Security Education. Information Security in Action: 13th IFIP WG 11.8 World Conference, WISE 13, Maribor, Slovenia, September 21–23, 2020, Proceedings 13*. Springer, pp. 96–109.

Blanco, V., Japón, A. and Puerto, J. (2020) 'Optimal arrangements of hyperplanes for SVM-based multiclass classification', *Advances in Data Analysis and Classification*, 14(1), pp. 175–199.

Jabeen, R., Singh, Y. and Sheikh, Z.A. (2022) 'Machine Learning for Security of Cyber-Physical Systems and Security of Machine Learning: Attacks, Defences, and Current Approaches', in *The International Conference on Recent Innovations in Computing*. Springer, pp. 813–841.

Kronser (2020) 'CVE (Common Vulnerabilities and Exposures).' Available at: https://www.kaggle.com/datasets/andrewkronser/cve-common- vulnerabilities-and-exposures?datasetId=500243.

Laskar, M.T.R. *et al.* (2021) 'Extending isolation forest for anomaly detection in big data via K-means', *ACM Transactions on Cyber-Physical Systems (TCPS)*, 5(4), pp. 1–26.

Li, Y. and Liu, Q. (2021) 'A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments', *Energy Reports*, 7, pp. 8176–8186.

Ranade, P. *et al.* (2021) 'Generating fake cyber threat intelligence using transformer-based models', in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–9.

Scikit-learn (2023) 'Isolation Forest example.' Available at: https://scikit-learn.org/stable/auto_examples/ensemble/plot_isolation_forest.html#sphx-glr-auto-examples-ensemble-plot-isolation-forest-py.

Togbe, M.U. *et al.* (2020) 'Anomaly detection for data streams based on isolation forest using scikit-multiflow', in *Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part IV 20*. Springer, pp. 15–30.

Xiao, H. *et al.* (2015) 'Is feature selection secure against training data poisoning?', in *international conference on machine learning*. PMLR, pp. 1689–1698.

Yaacoub, J.-P.A. *et al.* (2022) ‘Robotics cyber security: Vulnerabilities, attacks, countermeasures, and recommendations’, *International Journal of Information Security,* pp. 1–44.

Chauhan, R., Kumar, P., Uniyal, S., & Kumar, M. (2023). "Fake news detection using machine learning algorithm." *In 2023 International Conference on Sustainable Emerging Innovations in Engineering and Technology (ICSEIET)*. IEEE.

Al Asaad, B., & Erascu, M. (2018). "A Tool for Fake News Detection." In *Proceedings of the 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC).* IEEE.

Leela Siva Rama Krishna, N., & Adimoolam, M. (2022). "Fake News Detection System using Logistic Regression and Compare Textual Property with Support Vector Machine Algorithm." In *Proceedings of the 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS).* IEEE.

Reis, J. C. S., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). "Supervised Learning for Fake News Detection." *IEEE Intelligent Systems,* 34(2), 76-81.

Katsaros, D., Stavropoulos, G., & Papakostas, D. (2019). "Which Machine Learning Paradigm for Fake News Detection?" In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE.

Gupta, V., Mathur, R. S., Bansal, T., & Goyal, A. (2022). "Fake News Detection using Machine Learning." In *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*. IEEE.

Ranade, P., Piplai, A., Mittal, S., Joshi, A., & Finin, T. (2021). "Generating Fake Cyber Threat Intelligence Using Transformer-Based Models." *In 2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE.

Deepa, Shetty, C., N, R., & Hegde, P. (2022). "Fake News Detection Model using Machine Learning Techniques." In *2022 International Conference on Artificial Intelligence and Data Engineering (AIDE)*. IEEE.