Optimizing Feature Selection in Intrusion Detection Using Fisher Score Algorithm: An Analytical Study

Ganesh Rathod[1], Vikrant Sabnis[2], Jay Kumar Jain[3]
[1]*Research Scholar, Faculty of Engineering & Technology, Mansarovar Global University, Bhopal, India*
[2]*Faculty of Engineering & Technology, Mansarovar Global University, Bhopal, India*
[3]*Department of Mathematics, Bioinformatics and Computer Applications,Maulana Azad National Institute of Technology, Bhopal, India*

Abstract— **Traditional intrusion detection methods often fail to meet the unique security requirements of IoT applications, raising serious security concerns due to the rapid expansion of the Internet of Things (IoT) - connectivity and its extensive property, complex structure, and complexity. Machine learning algorithms for intrusion detection have emerged as a solution, offering promise in protecting these complex systems. This paper examines a comprehensive analysis of an Intrusion Detection System (IDS) in an IoT system. Recognizing the important role of feature extraction in IDS, this study aims to help researchers by providing insights on data set selection and confirming the effectiveness of the Fisher Score algorithm. Through a careful comparative analysis using established selection methods Mutual Information, Chi-Square, Principal Component Analysis (PCA), and Recursive Feature Elimination (RFE) this study seeks to help researchers choose the most appropriate feature extraction technique in intrusion detection within the IoT framework. Using logistic regression as a classification model, this study allows a thorough analysis and comparison of different selection methods. The results highlight the importance and accuracy of the Fisher Score algorithm in selecting key features for Intrusion Detection in IoT systems. It is worth mentioning, that this research is limited to a specific dataset, NBaIot, which is considered the best and most up-to-date for Intrusion Detection. The specificity observed in this study is dependent on the number of items and data sets. Although the findings may be variable, the selected sets and item sizes have informed the identification of interventions in this study that are deemed optimal.**

Index Terms— **Machine Learning, Intrusion Detection System, Fisher Score, Mutual Information, Chi-Square, Principal Component Analysis (PCA), Recursive Feature Elimination (RFE).**

I. INTRODUCTION

The fast improvement of the Internet of Things (IoT) has modified various industries, imparting supreme connectivity and convenience in numerous components of our lives. But this technological evolution has brought serious protection issues, mainly in IoT applications. The IoT panorama poses specific challenges compared to standard internet environments, offering sizable assets, elaborate architectures, and restricted processing sources. Among the important components of IoT protection, setting up strong intrusion detection mechanisms remains a pivotal concern. The IoT atmosphere features a large community of interconnected gadgets, starting from smart home home equipment and wearables to commercial machinery and metropolis infrastructure. This interconnectedness empowers stronger functionalities and facts-pushed services, revolutionizing the way we interact with generation. However, it additionally amplifies the susceptibility to security threats, making IoT systems susceptible to unauthorized access, statistics breaches, and cyber-assaults.

Traditional safety features often fall quick in addressing the intricacies and evolving nature of IoT environments. The complicated and dynamic nature of IoT architectures makes it difficult for conventional intrusion detection structures to successfully reveal and safeguard those networks. Moreover, IoT devices regularly function with restrained computational resources, similarly complicating the implementation of strong safety features. Intrusion detection is paramount in mitigating potential threats and safeguarding IoT networks. It entails the identification of unauthorized get entry to, odd sports, and potential

vulnerabilities in the network. Traditional intrusion detection systems predominantly rely upon rule-based totally procedures or signature-based detection, which won't suffice inside the context of IoT due to their static nature and limited adaptability

To address the shortcomings of traditional methods, researchers and practitioners have increasingly more became to gadget learning (ML) techniques to support intrusion detection in IoT environments. ML models provide the ability to conform and analyze from facts, enabling extra dynamic and effective intrusion detection systems inside the IoT domain. One essential thing influencing the efficacy of intrusion detection structures is feature selection. Feature selection strategy's purpose to pick out the maximum relevant attributes within datasets that make contributions notably to detecting anomalies or malicious sports. By extracting pertinent features from giant datasets, those techniques beautify the efficiency, accuracy, and scalability of intrusion detection systems. Among numerous characteristic selection techniques, the Fisher Score algorithm has garnered interest for its effectiveness in choosing discriminative capabilities.

This algorithm evaluates the statistical variations among instructions to perceive capabilities that contribute the most to category. In the context of intrusion detection, employing the Fisher Score algorithm for feature choice can doubtlessly beautify the accuracy and effectiveness of ML-based intrusion detection systems in IoT environments. However, comparing the overall performance of feature selection algorithms in intrusion detection entails complete evaluation and assessment with different established techniques. Comparisons with algorithms which include Mutual Information, Chi-square, and Principal Component Analysis (PCA)

The subsequent sections of this paper are organized as follows: Section II presents an exploration of related work focusing on feature selection algorithms within IoT environments. Section III delves into the methodology employed for selecting the most suitable feature selection algorithm. Following this, Section IV presents a comprehensive comparative analysis, specifically highlighting the Fisher Score algorithm against established feature selection techniques like Mutual Information, Chi-square, PCA, and RFE. Finally, Section V draws conclusions.

II. LITERATURE REVIEW

IoT-based smart environments represent a revolutionary paradigm leveraging IoT devices to offer an array of services and applications, aiming to enhance human life quality. Despite their transformative potential, these systems grapple with inherent security challenges arising from the vulnerabilities and limitations intrinsic to IoT devices and protocols. The vulnerabilities expose these environments to potential risks, threatening the integrity, functionality, and privacy of the systems. Intrusion Detection Systems (IDSs) emerge as indispensable safeguards, playing a pivotal role in fortifying IoT-based smart environments against malevolent attacks that exploit these vulnerabilities. These systems serve as vigilant gatekeepers, constantly monitoring network or host activities. The paper [1], proposes a novel approach to detect cyberattacks in IoT networks using a combination of deep learning and three-level algorithms. It uses the BoT-IoT dataset, which is a realistic and challenging dataset that simulates different types of attacks in a network environment. The paper [2], proposes a novel feature-selection algorithm for IoT networks to improve the intrusion detection performance and reduce the computational cost. The algorithm uses a hybrid approach of filter and wrapper methods to select the optimal subset of features from network traffic data. It shows that the proposed algorithm can achieve comparable or better results than the full feature set with a significantly reduced number of features. The paper also identifies some features that have an unrealistically high predictive power and should be avoided in the ML models' training. The paper [3], claims that the proposed algorithm can enhance the security and efficiency of IoT networks. It proposes an energy-efficient model for IoT using compressive sensing, which is a technique that reduces the amount of data collected and transmitted by IoT devices. Many works are carried out in intrusion detection and their primary objective is the early detection and identification of any aberrant or suspicious behaviour, serving as crucial indicators of potential security breaches within the ecosystem [4]. The significance of IDSs lies in their proactive stance, enabling prompt responses to potential threats. By swiftly identifying anomalies or deviations from established behavioural patterns, IDSs act as a

frontline defence mechanism, pre-empting and mitigating the impact of security breaches that could compromise the functionality and privacy of IoT-based smart environments.

Designing and implementing Intrusion Detection Systems (IDSs) tailored for IoT-based smart environments present substantial complexities and prerequisites. Highlighted challenges and requirements, as identified by [5], encompass: The heterogeneity and diversity of IoT devices and protocols, which make it difficult to apply a uniform detection mechanism and to handle the different data formats and communication standards. The limited resources and capabilities of IoT devices, such as memory, processing power, battery life, and bandwidth, which constrain the computational and storage capacity of the IDSs and affect their performance and efficiency. The scalability and adaptability of the IDSs, which need to cope with the dynamic and large-scale nature of IoT networks and to adjust to the changing environment and traffic patterns. The accuracy and reliability of the IDSs, which need to achieve a high detection rate and a low false positive rate, and to handle the noise, uncertainty, and incompleteness of the data. The privacy and security of the IDSs, which need to protect the sensitive data and information of the IoT devices and users from unauthorized access and modification, and to prevent the IDSs themselves from being attacked or compromised. The paper [6], proposes a method to profile network traffic by using classification techniques in machine learning. The paper uses two datasets: NSL-KDD and UNSW-NB15, which contain different types of network attacks and benign traffic. It applies four classification algorithms: Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine, to classify the network traffic into normal or attack classes. It then evaluates the performance of the classifiers using accuracy, precision, recall, and F1-score metrics. Finally, this paper compares the results of the classifiers on the two datasets and discusses the advantages and limitations of each classifier. The paper concludes that Random Forest and Support Vector Machine are the best classifiers for network traffic profiling, and that the UNSW-NB15 dataset is more challenging and realistic than the NSL-KDD dataset.

Feature selection is a process of selecting a subset of relevant features from a large set of features to improve the performance and efficiency of machine learning models. Feature selection is especially important for IoT intrusion detection, as IoT networks generate massive amounts of data with high dimensionality and heterogeneity as per [7]. Feature selection can help reduce the computational cost, storage space, and communication overhead of IoT intrusion detection systems, as well as enhance the accuracy, robustness, and interpretability of the models [8][9]. Various feature selection methods have been proposed and applied to IoT intrusion detection datasets, such as UNSW-NB15, CSE-CIC-IDS2018, and ToN-IoT. These methods can be broadly classified into three categories: filter, wrapper, and embedded methods. Filter methods rank the features based on some statistical measures, such as chi-square, information gain, correlation, and Fisher score, and select the top-ranked features. Wrapper methods use a machine learning algorithm as a black box to evaluate the features and search for the optimal subset using some heuristic techniques, such as genetic algorithm, tabu search, and cellular automata. Embedded methods integrate the feature selection process into the machine learning model, such as random forest, decision tree, and neural network, and select the features based on the model's internal criteria [10].

Numerous research studies have extensively explored and compared feature extraction methodologies such as Fisher Score, Mutual Information (MI), Principal Component Analysis (PCA), Chi-square, and Recursive Feature Elimination (RFE) within the domain of intrusion detection systems, each aiming to optimize detection accuracy and efficiency. In [11], the traditional Fisher score method selects each feature independently based on their scores under the Fisher criterion, which leads to a suboptimal subset of features that may not capture the interactions and dependencies among features. The Paper [12], introduces a generalized Fisher score method that jointly selects features by maximizing the ratio of the between-class scatter and the within-class scatter in the projected feature space. It also proposes an efficient algorithm to solve the optimization problem of the generalized Fisher score method. Another study [13], evaluates the proposed method on several benchmark datasets and compares it with other feature selection methods, such as ReliefF, mRMR, and Laplacian score. It claims that the generalized Fisher score method achieves better or comparable

classification performance with fewer features than the other methods. Principal Component Analysis (PCA) has emerged as a powerful technique for feature selection and dimensionality reduction in various fields, including pattern recognition, machine learning, and intrusion detection systems. PCA aims to capture the maximum variance in data by transforming the original features into a new set of linearly uncorrelated variables, known as principal components [14]. Numerous studies have demonstrated the efficacy of PCA in reducing the dimensionality of high-dimensional datasets while preserving the most informative features. By retaining the principal components that contribute significantly to the variance, PCA aids in extracting essential information from complex data structures [15]. In the realm of intrusion detection systems, researchers have extensively explored the applicability of PCA for feature selection. PCA facilitates the identification of a reduced set of features that encapsulate the essential information for accurate intrusion detection while minimizing information loss [cite 3]. Its ability to transform the original feature space into a lower-dimensional space allows for more efficient and effective detection of anomalies or malicious activities [16]. However, while PCA offers dimensionality reduction, its limitations must be considered. The transformed principal components might not always align with the features that are most relevant for specific detection tasks. In [17], certain cases, using PCA alone for feature selection might discard discriminative information, leading to reduced detection accuracy.

III. METHODOLOGY

This study adopts a methodical approach to fortify intrusion detection in IoT settings. The methodology involves five integral steps, starting with the extraction of relevant data from diverse sources. Following this, data preprocessing refines the dataset by eliminating noise and irrelevant information. Leveraging various feature selection algorithms such as Fisher Score, Mutual Information, Chi-square, PCA, and RFE the study identifies and extracts key attributes crucial for detecting anomalies or malicious activities in IoT networks. Logistic Regression serves as the primary classification model due to its interpretability and effectiveness in binary classification. The models are rigorously evaluated using standard metrics to compare the performance of each feature selection technique in tandem with the Logistic Regression classifier. Figure 1 visually encapsulates this comprehensive methodology, showcasing how each step contributes to enhancing the performance of intrusion detection in IoT environments.

Figure 1. Feature Selection and Classification for Intrusion Detection in IoT

This symbolic representation encapsulates the sequential steps involved in the process of evaluating different feature selection methodologies followed by the training of a predictive model, ultimately leading to the computation of accuracy scores associated with each technique. Figure 2 depicts the sequential algorithm outlining the steps involved in the study's methodology.

ALGORITHM

Start

Step 1: Load the dataset ('Final_dataset.csv')

Step 2: Split the dataset into features (X) and the target (y)

Step 3: Normalize the features using MinMaxScaler

Step 4: Split the dataset into training and testing sets (80-20 ratio)

Step 5: Define feature selection methods:
- Fisher Score with SelectKBest
- Mutual Information with SelectKBest
- Chi-square with SelectKBest
- Principal Component Analysis (PCA) with PCA
- Recursive Feature Elimination (RFE) with Logistic Regression

Step 6: For each feature selection method:
- Apply the method to the training set
- Transform the test set accordingly

Step 6.1: If using PCA or RFE:
- Apply the method directly to the training and testing sets

Step 6.2: Otherwise:
- Fit the method on the training set with target labels
- Transform both training and testing sets

Step 6.3: Train a Logistic Regression model on the transformed training set

Step 6.4: Predict on the transformed test set

Step 6.5: Evaluate accuracy using accuracy score with true test labels and predictions

Step 6.6: Display or store the accuracy for the specific feature selection method

Step 7: End

Figure 2. Algorithm

A. Mathematical Model for Feature Selection and Classification in Intrusion Detection

In a mathematical model, each step is symbolically represented to demonstrate the logical flow of the process. Here's a representation of the steps using mathematical symbols:

-Let D represent the dataset.

-Let X denote the features, and by the target.

-Let MinMaxScaler normalize the features.

-Let Train and Test represent the training and testing sets, respectively (split 80-20).

-Let FS, MI, CS, PCA, RFE denote feature selection methods.

-Let LR represent the Logistic Regression classifier.

-Let ACC_{FS}, ACC_{MI}, ACC_{CS}, ACC_{PCA}, ACC_{RFE}, represents accuracy scores for Fisher Score, Mutual Information, Chi-Square, Principal Component Analysis and Recursive Feature Elimination selection methods.

-Let $Pred_{FS}$, $Pred_{MI}$, $Pred_{CS}$, $Pred_{PCA}$, $Pred_{RFE}$, represents predicted outcomes generated by the Logistic Regression model specifically trained using the features selected via the Fisher Score, Mutual Information, Chi-Square, Principal Component Analysis and Recursive Feature Elimination selection methods respectively.

$D \rightarrow X,y$

$X \rightarrow$ MinMaxScaler

$X \rightarrow Train,Test(80-20\ split)$

$FS(X,y) \rightarrow Train_{FS}, Test_{FS}$

$MI(X,y) \rightarrow Train_{MI}, Test_{MI}$

$CS(X,y) \rightarrow Train_{CS}, Test_{CS}$

$PCA(X,y) \rightarrow Train_{PCA}, Test_{PCA}$

$RFE(LR,X,y) \rightarrow Train_{RFE}, Test_{RFE}$

$LR(Train_{FS}) \rightarrow Pred_{FS}$

$LR(Train_{MI}) \rightarrow Pred_{MI}$

$LR(Train_{CS}) \rightarrow Pred_{CS}$

$LR(Train_{PCA}) \rightarrow Pred_{PCA}$

$LR(Train_{RFE}) \rightarrow Pred_{RFE}$

$ACC_{FS} = Accuracy\ (Pred_{FS}, Test_{FS})$

$ACC_{MI} = Accuracy\ (Pred_{MI}, Test_{MI})$

$ACC_{CS} = Accuracy\ (Pred_{CS}, Test_{CS})$

$ACC_{PCA} = Accuracy\ (Pred_{PCA}, Test_{PCA})$

$ACC_{RFE} = Accuracy\ (Pred_{RFE}, Test_{RFE})$

This symbolic representation encapsulates the sequential steps involved in the process of evaluating different feature selection methodologies followed by the training of a predictive model, ultimately leading to the computation of accuracy scores associated with each technique. Figure 2 illustrates the sequential steps involved in study.

B. Data Extraction

The study focused on collecting datasets specifically tailored for Intrusion Detection in IoT (Internet of Things) environments. Among various sources, particular emphasis was placed on the N-BaIoT dataset due to its comprehensive nature and relevance in addressing IoT security challenges. The dataset was selected for its multifaceted network traffic data, which plays a pivotal role in assessing the efficacy of intrusion detection systems within IoT networks. The N-BaIoT dataset, sourced from diverse IoT devices, stands out due to its multivariate, sequential nature, comprising a substantial 7062606 instances. Notably, the dataset contains a broad spectrum of data, facilitating anomaly detection and multi-class

classification tasks. It includes benign and malicious traffic data, enabling the classification of ten distinct attack classes alongside a 'benign' class.

The dataset characteristics underscore its utility for various machine learning tasks, particularly classification and clustering. With real attributes totalling 115 and no missing values, the dataset maintains completeness and reliability, vital for analysis and model development. The study's focus on anomaly detection involved training and optimizing deep autoencoders using two-thirds of each device's benign data. The test data consisted of the remaining benign data and all the malicious data, which allowed for applying the trained autoencoders as effective anomaly detectors. This approach achieved a remarkable 100% True Positive Rate (TPR) in detecting anomalies, validating the dataset's effectiveness for evaluating intrusion detection systems in IoT networks. N-BaIoT dataset's richness and diversity provide an excellent foundation for exploring and developing robust intrusion detection systems in IoT environments, ensuring thorough evaluation and advancement of security measures.

C. Data Preprocessing

The data preprocessing stage in this study aimed to meticulously refine the N-BaIoT dataset, a comprehensive and multi-faceted dataset specifically tailored for intrusion detection within IoT (Internet of Things) networks. Addressing the intricate challenges inherent in IoT security, this phase focused on refining and preparing the dataset to ensure its quality, reliability, and suitability for subsequent analysis and model training. The dataset, sourced from diverse IoT network traffic data, holds paramount importance in evaluating and enhancing the efficacy of intrusion detection systems, thus necessitating a thorough preprocessing regime to curate high-quality data for accurate model training and evaluation The initial steps in data preprocessing cantered on identifying and mitigating uncertainties present within the dataset. Missing values and noise, common pitfalls in data analysis, were meticulously addressed to preserve data integrity and reliability. Mean imputation and outlier removal techniques were strategically employed to handle missing values and reduce noise, ensuring the dataset's quality was uncompromised. This pivotal step aimed at fortifying the dataset against inaccuracies and inconsistencies, thereby establishing a robust foundation for subsequent analysis.

Further fortifying the dataset, the preprocessing phase encompassed data transformation and normalization procedures. Leveraging Min-Max scaling techniques, the dataset's features were normalized to a standardized range, mitigating scale differences that could otherwise skew model predictions. Additionally, log transformations were judiciously applied to stabilize variance, ensuring the dataset's suitability for accurate model training and evaluation. The intent was to foster a dataset that not only meets the stringent criteria of accuracy but also offers a reliable representation of IoT network behaviours. Feature engineering played a pivotal role in this phase, employing a suite of feature selection techniques including Fisher Score, Mutual Information, Chi-square, Principal Component Analysis (PCA), and Recursive Feature Elimination (RFE). These techniques were instrumental in discerning and retaining the most influential attributes crucial for accurate intrusion detection. The overarching aim was to strike a balance between reducing dimensionality and preserving critical information essential for accurate classification. The dataset underwent an 80-20 split, partitioning it into distinct training and testing subsets. This segregation facilitated robust model training on a substantial portion of the dataset while reserving a separate subset for unbiased model evaluation. Mitigating data skewness was another pivotal aspect addressed during preprocessing. Transformative techniques, such as log transformations or Box-Cox transformations, were implemented to normalize data distributions, ensuring a more accurate representation of patterns and behaviours within the IoT network data.

Data preprocessing phase embarked upon a meticulous journey of refining the N-BaIoT dataset, alleviating uncertainties, transforming features, and optimizing the dataset for intrusion detection within IoT environments. By undertaking an array of preprocessing techniques, the dataset was fortified to ensure subsequent model training and evaluation were performed on high-quality, reliable data. This robust preprocessing methodology not only underpins the credibility of subsequent analyses but also serves as a cornerstone in enhancing the efficacy of intrusion detection systems in IoT realms.

D. Feature Selection Algorithms

The process of feature selection stands as a cornerstone in the quest for optimizing intrusion detection systems within IoT environments. This pivotal stage aims to identify and retain the most discriminative and relevant attributes from the dataset, enhancing the precision and efficacy of the subsequent intrusion detection models.

The Fisher Score, a fundamental feature selection technique, originates from the statistical concept of Fisher Discriminant Ratio, devised by Sir Ronald A. Fisher. This method is prominent for its efficacy in discerning features that contribute most substantially to distinguishing between different classes within a dataset. It operates by evaluating the discriminatory power of individual features through the variance analysis between and within classes. The Fisher Score algorithm, integrated within the SelectKBest method, functions by computing the variance between classes and within classes for each feature present in the dataset. It measures the separation between different classes by assessing the ratio of between-class variance to within-class variance. Features with higher scores indicate stronger discriminatory power in distinguishing between classes, thereby signifying their significance in characterizing and differentiating the data.

The core principle behind Fisher Score revolves around the notion of maximizing the variance between classes while minimizing the variance within classes. This criterion aims to identify features that exhibit considerable variability concerning different classes while maintaining consistency within each class. Consequently, the algorithm selects attributes that effectively capture the intrinsic characteristics contributing to class separability, thus facilitating accurate and robust classification or detection tasks. In the context of intrusion detection within IoT networks, Fisher Score plays a vital role in selecting attributes that encapsulate relevant behavioural patterns or distinctive characteristics associated with benign and malicious network activities. By evaluating the discriminatory power of features, Fisher Score aids in the identification of critical attributes essential for precise detection of anomalies or security threats within the intricate and diverse IoT ecosystem. The utilization of Fisher Score in the realm of intrusion detection systems inside IoT networks aligns with the goal of as it should be distinguishing among normal and anomalous network behaviours. It allows for the identification of pertinent features that contribute notably to intrusion detection, ultimately improving the efficiency and effectiveness of security features deployed in IoT environments.

The Mutual Information-based Feature selection method embedded within the SelectKBest method measures the diploma of dependency or statistics shared between variables, particularly that specialize in their relevance in conveying essential information about magnificence labels. It assesses the power of association among every characteristic and the goal variable by way of quantifying how a whole lot statistics about the goal variable is contained inside a selected feature. Mutual Information operates on the principle of information gain, evaluating the amount of information obtained about the target variable when considering a particular feature. It quantifies the extent to which knowledge about the feature reduces uncertainty in predicting the target variable. Features exhibiting higher Mutual Information scores signify stronger relationships or dependencies with the target variable, highlighting their significance in predicting or characterizing the class labels. In the domain of intrusion detection within IoT environments, Mutual Information serves as a crucial tool for selecting features that effectively contribute to distinguishing between normal and anomalous network behaviours. By systematically analysing the relevance and dependency of features on the target variable (such as identifying malicious activities), Mutual Information aids in retaining attributes that are highly informative and influential in detecting security threats within IoT networks. Utilizing Mutual Information for feature selection in intrusion detection structures inside IoT environments enhances the system's functionality to discern important attributes that encapsulate relevant data related to security breaches or anomalous behaviours. It contributes to building greater robust and accurate intrusion detection fashions tailor-made to the precise demanding situations posed by the complex and evolving IoT environment.

The Chi-square test is a statistical degree used to take a look at the independence of specific variables. In the context of feature choice, Chi-square statistics compare the relationship between categorical capabilities and the goal variable, determining the importance of these associations in distinguishing one-of-a-kind instructions or labels. Chi-square operates

through measuring the discrepancy among observed and expected frequencies of categorical variables, enabling the identity of features that show off non-random institutions with the target variable. It calculates a Chi-square statistic for each feature, quantifying the extent of dependency or independence among the characteristic and the magnificence labels. In intrusion detection within IoT environments, Chi-square-based feature selection aids in identifying attributes that significantly contribute to distinguishing between normal and anomalous network behaviours. By assessing the strength of association between categorical features and class labels, Chi-square selects features crucial for effective intrusion detection in the dynamic IoT landscape. Utilizing Chi-square statistics within feature selection techniques enhances the capability to discern vital attributes that play a pivotal role in detecting security threats within IoT networks. This approach contributes to building extra powerful and focused intrusion detection fashions, enhancing safety features inside IoT environments.

Principal Component Analysis (PCA) is a dimensionality discount approach extensively used to discover critical styles or systems inside high-dimensional datasets. In function selection, PCA reconfigures the original capabilities into a reduced set of primary components, making sure minimal loss of records at the same time as condensing the dataset's dimensions. PCA operates by transforming the original features into a new set of orthogonal variables called principal components. These components are ordered by the amount of variance they capture in the dataset, with the first few components retaining the maximum variance. By selecting a subset of these components, PCA achieves dimensionality reduction while preserving the essential characteristics of the data. In intrusion detection systems for IoT environments, PCA's dimensionality reduction capabilities streamline the analysis of complex network data. By condensing the feature space into a smaller set of principal components, PCA simplifies the representation of network behaviours while preserving critical information related to security threats. Utilizing PCA for feature choice in intrusion detection enhances the performance of models by lowering computational complexity and minimizing the risk of overfitting. It permits for a more focused analysis of vital functions, permitting extra correct and efficient detection of anomalies inside IoT networks.

Recursive Feature Elimination (RFE) is a function selection technique that systematically prunes the feature set via putting off much less important attributes iteratively. This iterative procedure evaluates the contribution of each feature to the version's performance, removing the least impactful ones in a stepwise way.RFE starts by training the model on the entire feature set and assessing the importance of each feature based on predefined criteria, often model coefficients or feature weights. It then eliminates the least significant feature(s) and repeats the process until the specified number of features remains or until a performance threshold is reached.

In the realm of intrusion detection within IoT networks, RFE plays a crucial role in identifying and retaining the most relevant attributes essential for distinguishing normal network behaviour from potential security threats. By progressively eliminating less informative features, RFE streamlines the feature space, allowing intrusion detection models to focus on the most discriminative attributes. Utilizing RFE for feature selection in intrusion detection systems enhances the efficiency and accuracy of models by reducing computational complexity and overfitting risks. It aids in constructing more refined and effective intrusion detection systems specifically tailored for the complexities of IoT environments.

E. Classification model and Evaluation

The study utilizes Logistic Regression, which stands as the model of choice for classification, primarily due to its interpretability and strong performance in binary classification tasks. Its ability to estimate the probability of an instance belonging to a specific class aligns well with the binary nature of intrusion detection in IoT networks, offering a clear understanding of feature contributions in distinguishing between normal and anomalous activities. In evaluating model performance, the primary metric utilized is accuracy. Accuracy measures the overall correctness of the model's predictions by calculating the ratio of correctly predicted instances to the total instances evaluated. It serves as a fundamental yardstick to assess the model's effectiveness in correctly identifying both normal and

intrusive network behaviours within IoT environments.

The dataset utilized for this analytical study is N_BaIoT [24], segmented into training and testing subsets with an 80-20 division. This setup allocates a substantial portion for training to ensure adequate model learning while maintaining a separate portion for unbiased testing. The use of cross-validation further validates the model's robustness by iteratively validating its performance across multiple training and validation subsets. This approach mitigates overfitting concerns and ensures the reliability of accuracy as a performance measure. The study conducts a rigorous comparative analysis, evaluating diverse feature selection methods like Fisher Score, Mutual Information, Chi-square, PCA, and RFE in conjunction with Logistic Regression. The assessment primarily revolves around accuracy, aiming to discern the most influential attributes for intrusion detection within IoT environments. The meticulous analysis provides insights into the effectiveness of various feature selection strategies in accurately identifying and differentiating between normal and anomalous behaviours

IV. RESULTS AND DISCUSSIONS

The investigation focused on assessing the performance of feature selection algorithms in intrusion detection within IoT environments, specifically by altering the 'k' values within each algorithm. Five distinct algorithms Fisher Score, Mutual Information, Chi-square, PCA, and RFE were tested, each configured to select different numbers of 'k' features. This analysis aimed to discern the influence of 'k' values solely on accuracy, the chosen performance metric, with

evaluation restricted to Logistic Regression as the classifier. The 'k' parameter played a pivotal role in the feature selection algorithms employed in this study. For Fisher Score, Mutual Information, and Chi-square, the 'k' value determined the number of top features selected based on their respective scoring methods. In contrast, PCA aimed to reduce the dataset's dimensionality to 'k' components, while retaining critical information. RFE iteratively eliminated features until reaching the desired number of 'k' features. Performance evaluation primarily relied on accuracy, emphasizing the reliability of Logistic Regression as the classifier. This limitation to accuracy was grounded in its relevance as a key metric in evaluating the model's ability to correctly predict the class labels in intrusion detection scenarios.

The following sections provide a comprehensive comparison of accuracy across five prominent feature selection algorithms as 'k' values varied from 40 to 8. These algorithms were chosen for their widespread usage and effectiveness in classification problems, underscoring their prominence in the realm of intrusion detection within IoT networks.

A. Accuracy comparison for ‘k=40’

The comparison of accuracy across different feature selection algorithms for 'k=40' offers valuable insights into their relative performance:

Table I . Accuracy Comparison for ‘k=40’

Name of the Algorithm	Accuracy
Fisher Score	0.8245
Mutual Information	0.8268
Chi-square	0.6986
Principal Component Analysis	0.8241
Recursive Feature Elimination	0.8232

The Table I. showcases the accuracy values obtained for different feature selection algorithms applied to intrusion detection within IoT environments. Mutual Information stands out with the highest accuracy of 0.8268, closely followed by Fisher Score at 0.8245. Both of these algorithms demonstrate strong performance in selecting relevant features for intrusion detection tasks in IoT networks.

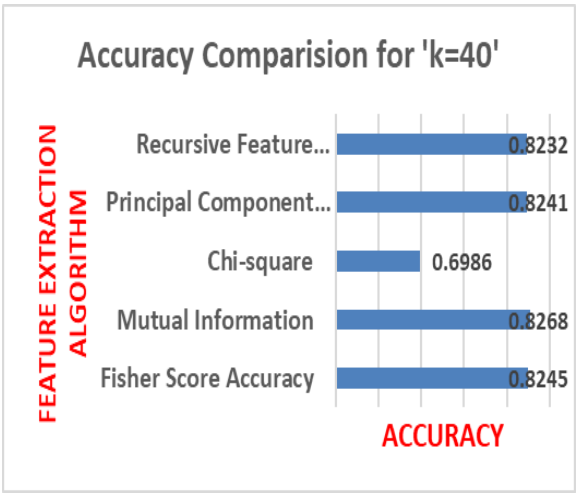

Figure 1. Graph for k=40

Principal Component Analysis (PCA) also exhibits notable accuracy at 0.8241, reinforcing its effectiveness in dimensionality reduction while retaining essential information. Recursive Feature Elimination (RFE) maintains a competitive accuracy of 0.8232, positioning it as another robust feature selection technique for intrusion detection. However, Chi-square lags behind significantly, achieving an accuracy of 0.6986. This lower accuracy suggests limitations in capturing crucial features for intrusion detection when using Chi-square compared to the other methods evaluated in this context. Fisher Score and Mutual Information emerge as the top performers in this assessment, showcasing their potential as effective feature selection algorithms for intrusion detection in IoT environments, followed closely by PCA and RFE. Chi-square, while showing lower accuracy, still provides some insight into feature relevance but might be less suitable for this specific intrusion detection task within IoT network.

B. Accuracy comparison for ‘k=30’

The examination of accuracy among various feature selection algorithms with 'k=30' provides useful insights into how they perform relative to each other

Table II. Accuracy comparison for 'k=30'

Name of the Algorithm	Accuracy
Fisher Score	0.8241
Mutual Information	0.8250
Chi-square	0.6609
Principal Component Analysis	0.8241
Recursive Feature Elimination	0.8232

In analysing the accuracy values derived from various feature selection algorithms for IoT intrusion detection, it's evident that Mutual Information performs exceptionally well, achieving the highest accuracy of 0.825. Not far behind is Fisher Score, demonstrating a commendable performance of 0.8241, indicating its proficiency in selecting relevant features for intrusion detection within IoT environments.

Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) exhibit comparable accuracies, both hovering around 0.8241 and 0.8232, respectively. This suggests their capability

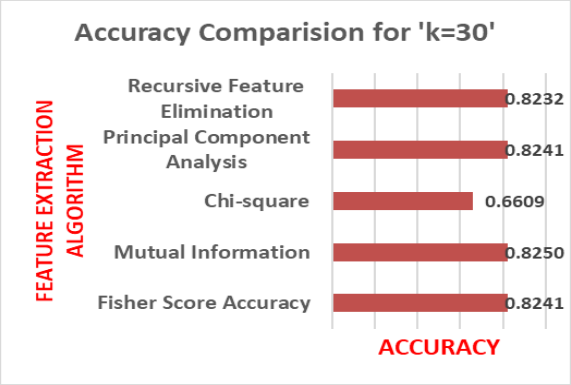

Figure 4. Graph for k=30

to effectively capture essential information while reducing dataset dimensionality, rendering them valuable for this task. However, Chi-square notably lags behind, displaying an accuracy of 0.6609. This considerably lower accuracy implies limitations in identifying critical features necessary for effective intrusion detection, highlighting potential challenges in its application within this specific IoT context.

Here, Mutual Information emerges as the top-performing algorithm, closely followed by Fisher Score. PCA and RFE demonstrate competitive accuracies, underscoring their effectiveness in feature selection for IoT intrusion detection. Conversely, Chi-square appears less effective in identifying crucial features within this specific intrusion detection scenario

C. Accuracy comparison for ‘k=20’

The comparison of accuracy across different feature selection algorithms for 'k=20' offers following valuable insights in terms of accuracy

Table III. Accuracy comparison for ' k=20'

Name of the Algorithm	Accuracy
Fisher Score	0.8264
Mutual Information	0.8240
Chi-square	0.6405
Principal Component Analysis	0.8240
Recursive Feature Elimination	0.8232

Table III, shows the accuracy values obtained from the various feature selection algorithms having Fisher Score outperforming the others, achieving the highest accuracy of 0.8264. This indicates its effectiveness in identifying and selecting relevant features crucial for intrusion detection within IoT environments.

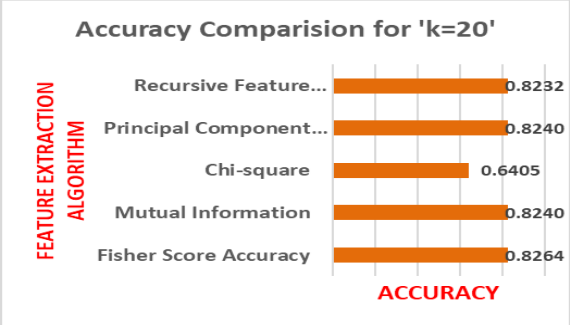

Figure 5. Graph for k=20

Mutual Information closely follows with an accuracy of 0.8240, showcasing its competitive performance in discerning pertinent attributes for intrusion detection tasks. Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) display comparable accuracies, both at 0.8240% and 0.8232, respectively. This similarity in performance suggests their efficacy in capturing essential information while reducing dataset dimensionality, proving beneficial for intrusion detection in IoT scenarios. However, Chi-square significantly lags behind, recording an accuracy of 0.6405. This substantial discrepancy indicates limitations in identifying critical features necessary for effective intrusion detection in IoT environments when using Chi-square as a feature selection method.

In comparison, Fisher Score stands out as the top-performing algorithm, closely trailed by Mutual Information. PCA and RFE exhibit competitive accuracies, while Chi-square shows notable limitations in selecting vital features for IoT intrusion detection based on accuracy alone

D. Accuracy comparison for ‘k=15’

The analysis of accuracy among various feature selection algorithms with 'k=15' provides following beneficial insights into how they perform relative to each other.

Table IV. Accuracy comparison for k=15

Name of the Algorithm	Accuracy
Fisher Score	0.8260
Mutual Information	0.8231
Chi-square	0.6095
Principal Component Analysis	0.8241
Recursive Feature Elimination	0.8231

In this analysis from above Table IV, Fisher Score continues to demonstrate its efficacy, achieving an accuracy of 0.8260. This performance solidifies its position as a top-performing feature selection method for intrusion detection in IoT environments, showcasing its ability to identify and select crucial attributes. PCA maintains a competitive standing, displaying an accuracy of 0.8241. While not surpassing Fisher Score, its performance indicates its effectiveness in capturing essential information while reducing dataset dimensionality, making it a valuable tool for feature selection in IoT intrusion detection.

Mutual Information and RFE closely follow with accuracies of 0.8231, showing consistent performance in identifying relevant attributes for intrusion detection tasks within IoT networks. However, Chi-square continues to exhibit limitations, recording an accuracy of 0.6095%. This considerable disparity highlights challenges in leveraging Chi-square as an effective feature selection method for IoT intrusion detection based on accuracy metrics alone.

Overall, Fisher Score maintains its lead as a robust feature selection algorithm, while PCA, Mutual Information, and RFE continue to demonstrate competitive performances. Conversely, Chi-square trails significantly behind in accuracy, indicating potential limitations in selecting vital features for intrusion detection in IoT environments.

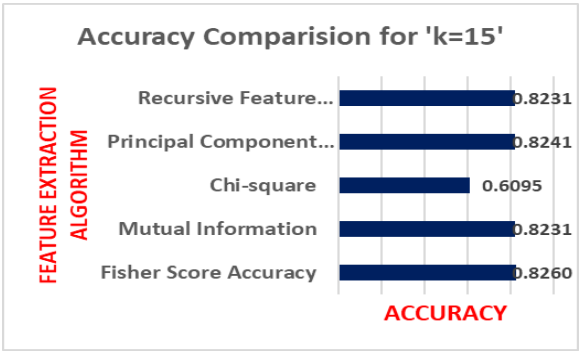

Figure 6. Graph for k=15

E. Accuracy comparison for ‘k=8’

The investigation of accuracy among various feature selection algorithms with 'k=8' provides important insights into how they perform relative to each other.

Table V. Accuracy comparison for k=8

Name of the Algorithm	Accuracy
Fisher Score Accuracy	0.7105
Mutual Information	0.5882
Chi-square	0.3718
Principal Component Analysis	0.8218
Recursive Feature Elimination	0.8241

In Table V, Recursive Feature Elimination (RFE) stands out with the highest accuracy at 0.8241, showcasing its efficiency in identifying critical attributes for intrusion detection in IoT networks.

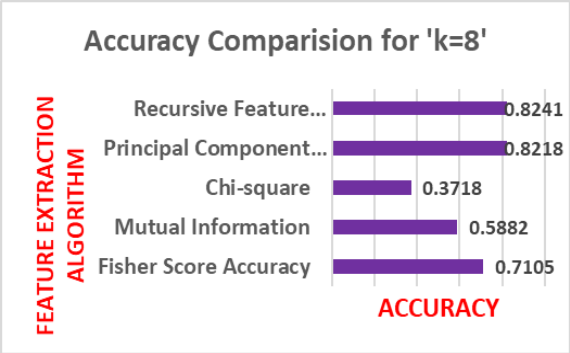

Figure 7. Graph for k=8

Principal Component Analysis (PCA) demonstrates a respectable accuracy of 0.8218, indicating its ability to capture relevant information while reducing dataset dimensionality. In contrast, Fisher Score, Mutual Information, and Chi-square exhibit significantly lower accuracies, at 0.7105, 0.5882, and 0.3718 respectively. These lower performances may suggest limitations in these methods' effectiveness in selecting important features for IoT intrusion detection based on the accuracy metric alone

F. Findings

The evaluation of 'k' parameter variations on feature selection algorithms underscores the notable performance of the Fisher Score method in the context of intrusion detection within IoT environments. Across different 'k' values, Fisher Score consistently maintains competitive accuracy, particularly at 'k=20', recording an accuracy of 82.6%. This resilience suggests Fisher Score's effectiveness in selecting critical features relevant to intrusion detection tasks. Comparatively, while Mutual Information briefly surpasses Fisher Score at 'k=30', Fisher Score maintains a consistent performance across different 'k' values, exhibiting its reliability. On the other hand, Chi-square demonstrates a noticeable decline in accuracy as 'k' decreases, highlighting potential limitations in capturing crucial attributes at lower 'k' values. Principal Component Analysis (PCA) and Recursive Feature Algorithm (RFE) showcase stability in accuracy across varied 'k' values, though PCA consistently performs slightly better than RFE.

Table VI. Important Findings

'k' Parameters	Fisher Score Accuracy	Mutual Information Accuracy	Chi Square Accuracy	PCA Accuracy	RFE Accuracy
k=40	0.825	0.827		0.824	0.823
k=30	0.824	0.825	0.661	0.824	0.823
k=20	0.826	0.824	0.640	0.825	0.823
k=15	0.826	0.823	0.610	0.824	0.823
k=8	0.710	0.588	0.372	0.822	0.824

In the above Table VI, RFE exhibited stable performance throughout the various 'k' values, maintaining a consistent accuracy without drastic fluctuations. Its stability could be interpreted as an indication of its reliability in feature selection. However, compared to Fisher Score, while RFE maintained consistency, it didn't outperform Fisher Score in terms of accuracy across different 'k' values. This stability might imply that RFE is consistent in selecting features even with variations in 'k' parameters, but it might not excel in capturing the most informative attributes for intrusion detection as effectively as Fisher Score did. Therefore, while RFE showcases stability, the comparative analysis suggests that Fisher Score might offer a more competitive edge in selecting crucial features for intrusion detection tasks within IoT environments due to its higher and more consistent accuracy across varying 'k' parameters.

The robustness of Fisher Score, maintaining accuracy levels amidst 'k' value variations, positions it as a promising choice for feature selection in intrusion detection for IoT networks. This stability, coupled with competitive accuracy, underscores its reliability in extracting pertinent attributes, solidifying its potential recommendation for intrusion detection systems in IoT environments. However, further exploration is warranted to ascertain the optimal 'k' value that maximizes Fisher Score's efficacy in feature selection for this specific domain

V. CONCLUSION

The research delved into the realm of IoT security, emphasizing the significance of effective Intrusion Detection Systems (IDS) within these intricate

ecosystems. Highlighting the essential role of feature selection in fortifying IDS, this study aimed to offer valuable insights into dataset selection and validate the efficacy of the Fisher Score algorithm. Through a meticulous comparative analysis involving established feature selection techniques of Machine Learning—Mutual Information, Chi-square, PCA, and RFE, the objective was to assist researchers in choosing the most appropriate algorithm for feature extraction in Intrusion Detection within IoT frameworks. By employing logistic regression as the classification model, this research rigorously evaluated and compared various feature selection methods. The findings conclusively underscored the prominence of the Fisher Score algorithm in extracting pivotal features essential for Intrusion Detection in IoT environments. However, it's essential to note that this research was conducted within the context of a specific dataset, N_BaIoT, renowned for its relevance in Intrusion Detection. The accuracy observed in this study is contingent upon the chosen datasets and feature quantities for intrusion detection.

For future advancements, expanding the scope to encompass diverse datasets and exploring different feature quantities across various classification models could provide deeper insights into the versatility and robustness of feature selection methodologies. This expanded exploration would not only enhance the understanding of different algorithms' performances under varied dataset conditions but also contribute significantly to a more comprehensive comprehension of intrusion detection in IoT settings. Furthermore, extending this research to include other performance evaluation metrics such as F1 Score, False Positive Rate, and additional classification algorithms beyond logistic regression could offer a more holistic view of the efficacy of feature selection methods in enhancing IDS within IoT landscapes. This evolution would further fortify the applicability and generalizability of these methodologies in real-world IoT security scenarios.

In conclusion, while this study validates the proficiency of the Fisher Score algorithm in feature extraction for Intrusion Detection within IoT environments, there remains a vast landscape for exploration and enhancement. The findings serve as a robust foundation, encouraging further research and exploration in this ever-evolving domain of IoT security.

VI. COMPETING INTERESTS

Declarations:

1. Funding: No funding was received to assist with the preparation of this manuscript.
2. Employment: None of the authors have any employment affiliations or conflicts of interest to disclose.
3. Financial Interests: The authors declare they have no financial interests.
4. Non-Financial Interests: None

VII. DATA AVAILABILITY STATEMENTS

Declarations:

1. The dataset utilized for the analysis in this study, sourced from the N_BaIoT Dataset available at https://doi.org/10.24432/C5RC8J -dataset, is openly accessible for review and replication.
2. The results and findings derived from the above dataset are thoroughly detailed and included within the paper for reference and verification purposes.

REFERENCES

[1] Alosaimi, Shema, and Saad M. Almutairi. "An Intrusion Detection System Using BoT-IoT." Applied Sciences 13, no. 9 (2023): 5427.

[2] Nazir, Anjum, Zulfiqar Memon, Touseef Sadiq, Hameedur Rahman, and Inam Ullah Khan. "A Novel Feature-Selection Algorithm in IoT Networks for Intrusion Detection." Sensors 23, no. 19 (2023): 8153.

[3] Jain, Jay Kumar, and Dipti Chauhan. "An Energy-Efficient Model for Internet of Things Using Compressive Sensing." Journal of Management Information and Decision Sciences 24 (2021): 1-7.

[4] Kalnoor, Gauri, and S. Gowrishankar. "IoT-based smart environment using intelligent intrusion detection system." Soft Computing 25, no. 17 (2021): 11573-11588.

[5] Khraisat, Ansam, and Ammar Alazab. "A critical review of intrusion detection systems in the internet of things: techniques, deployment strategy, validation strategy, attacks, public datasets and challenges." Cybersecurity 4 (2021): 1-27.

[6] Chauhan, Dipti, and Jay Kumar Jain. "Profiling Network Traffic by Using Classification Techniques in Machine Learning." International Conference on Smart Trends in Computing and Communications. Singapore: Springer Nature Singapore, 2023.

[7] Meera, Akhil Jabbar, MVV Prasad Kantipudi, and Rajanikanth Aluvalu. "Intrusion detection system for the IoT: A comprehensive review." In Proceedings of the 11th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2019) 11, pp. 235-243. Springer International Publishing, 2021.

[8] Sarhan, Mohanad, Siamak Layeghy, and Marius Portmann. "Feature analysis for machine learning-based IoT intrusion detection." arXiv preprint arXiv:2108.12732 (2021).

[9] Alani, Mohammed M., and Ali Miri. "Towards an explainable universal feature set for IoT intrusion detection." Sensors 22, no. 15 (2022): 5690.

[10] Mohy-eddine, Mouaad, Azidine Guezzaz, Said Benkirane, and Mourade Azrour. "An efficient network intrusion detection model for IoT security using K-NN classifier and feature selection." Multimedia Tools and Applications (2023): 1-19.

[11] Gupta, A. "Feature selection techniques in Machine Learning (updated 2023)." Analytics Vidhya (2023).

[12] Gu, Q., Z. Li, and J. Han. "Generalized fisher score for feature selection. arXiv 2012." arXiv preprint arXiv:1202.3725.

[13] Xu, Jiucheng, Kanglin Qu, Kangjian Qu, Qincheng Hou, and Xiangru Meng. "Feature selection using neighborhood uncertainty measures and Fisher score for gene expression data classification." International Journal of Machine Learning and Cybernetics (2023): 1-18.

[14] Omuya, Erick Odhiambo, George Onyango Okeyo, and Michael Waema Kimwele. "Feature selection for classification using principal component analysis and information gain." Expert Systems with Applications 174 (2021): 114765.

[15] Camacho, José, Jesús Picó, and Alberto Ferrer. "Data understanding with PCA: structural and variance information plots." Chemometrics and Intelligent Laboratory Systems 100, no. 1 (2010): 48-56.

[16] Arivudainambi, D., Varun Kumar KA, and P. Visu. "Malware traffic classification using principal component analysis and artificial neural network for extreme surveillance." Computer Communications 147 (2019): 50-57.

[17] Arivudainambi, D., Varun Kumar KA, and P. Visu. "Malware traffic classification using principal component analysis and artificial neural network for extreme surveillance." Computer Communications 147 (2019): 50-57.

[18] Chauhan, Dipti, and Jay Kumar Jain. "A Journey from IoT to IoE." International Journal of Innovative Technology and Exploring Engineering (IJITEE) 8.11 (2019).

[19] Soliman, Sahar, Wed Oudah, and Ahamed Aljuhani. "Deep learning-based intrusion detection approach for securing industrial Internet of Things." Alexandria Engineering Journal 81 (2023): 371-383.

[20] de Souza, Cristiano Antonio, Carlos Becker Westphall, Renato Bobsin Machado, Leandro Loffi, Carla Merkle Westphall, and Guilherme Arthur Geronimo. "Intrusion detection and prevention in fog based iot environments: A systematic literature review." Computer Networks 214 (2022): 109154.

[21] AlGhamdi, Rayed. "Design of Network Intrusion Detection System Using Lion Optimization-Based Feature Selection with Deep Learning Model." Mathematics 11, no. 22 (2023): 4607.

[22] JEMILI, Farah, Rahma MEDDEB, and Ouajdi KORBAA. "Intrusion Detection based on Ensemble Learning for Big Data Classification." (2023).

[23] Mati, William Peter. "Transferability of Intrusion Detection Systems Using Machine Learning between Networks." PhD diss., University of Windsor (Canada), 2022.

[24] Meidan,Yair, Bohadana,Michael, Mathov,Yael, Mirsky,Yisroel, Breitenbacher,Dominik,,Asaf, and Shabtai,Asaf. (2018). detection_of_IoT_botnet_attacks_N_BaIoT. UCI Machine Learning Repository. https://doi.org/10.24432/C5RC8J