




## Article

# Machine Learning-Based Anomaly Detection Using K-Mean Array and Sequential Minimal Optimization

Saad Gadal <sup>1</sup>, Rania Mokhtar <sup>2</sup>, Maha Abdelhaq <sup>3</sup>, Raed Alsaqour <sup>4,\*</sup> , Elmustafa Sayed Ali <sup>5</sup>   
and Rashid Saeed <sup>2</sup> 

<sup>1</sup> Electronics Engineering Department, Sudan University of Science and Technology, Khartoum 11111, Sudan; desogi@gmail.com

<sup>2</sup> Department of Computer Engineering, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia; ramohammed@tu.edu.sa (R.M.); eng\_rashid@hotmail.com (R.S.)

<sup>3</sup> Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia; msabdelhaq@pnu.edu.sa

<sup>4</sup> Department of Information Technology, College of Computing and Informatics, Saudi Electronic University, Riyadh 93499, Saudi Arabia

<sup>5</sup> Department of Electrical and Electronics Engineering, Red Sea University, Port Sudan 33311, Sudan; elmustafasayed@gmail.com

\* Correspondence: raed.ftsm@gmail.com

**Abstract:** Recently, artificial intelligence (AI) techniques have been used to describe the characteristics of information, as they help in the process of data mining (DM) to analyze data and reveal rules and patterns. In DM, anomaly detection is an important area that helps discover hidden behavior within the data that is most vulnerable to attack. It also helps detect network intrusion. Algorithms such as hybrid K-mean array and sequential minimal optimization (SMO) rating can be used to improve the accuracy of the anomaly detection rate. This paper presents an anomaly detection model based on the machine learning (ML) technique. ML improves the detection rate, reduces the false-positive alarm rate, and is capable of enhancing the accuracy of intrusion classification. This study used a dataset known as network security-knowledge and data discovery (NSL-KDD) lab to evaluate a proposed hybrid ML technology. K-mean cluster and SMO were used for classification. In the study, the performance of the proposed anomaly detection was tested, and results showed that the use of K-mean and SMO enhances the rate of positive detection besides reducing the rate of false alarms and achieving a high accuracy at the same time. Moreover, the proposed algorithm outperformed recent and close work related to using similar variables and the environment by 14.48% and decreased false alarm probability (FAP) by (12%) in addition to giving a higher accuracy by 97.4%. These outcomes are attributed to the common algorithm providing an appropriate number of detectors to be generated with an acceptable accurate detection and a trivial false alarm probability (FAP). The proposed hybrid algorithm could be considered for anomaly detection in future data mining systems, where processing in real-time is highly likely to be reduced dramatically. The justification is that the hybrid algorithm can provide appropriate detectors numbers that can be generated with an acceptable detection accuracy and trivial FAP. Given to the low FAP, it is highly expected to reduce the time of the preprocessing and processing compared with the other algorithms.

**Keywords:** anomaly detection; hybrid algorithm; data mining; sequential minimal optimization; k-mean clustering; network security



**Citation:** Gadal, S.; Mokhtar, R.; Abdelhaq, M.; Alsaqour, R.; Ali, E.S.; Saeed, R. Machine Learning-Based Anomaly Detection Using K-Mean Array and Sequential Minimal Optimization. *Electronics* **2022**, *11*, 2158. <https://doi.org/10.3390/electronics11142158>

Academic Editors: Muhammad Salman Haleem, Liangxiu Han, Ernesto Iadanza and Baihua Li

Received: 16 June 2022

Accepted: 7 July 2022

Published: 10 July 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

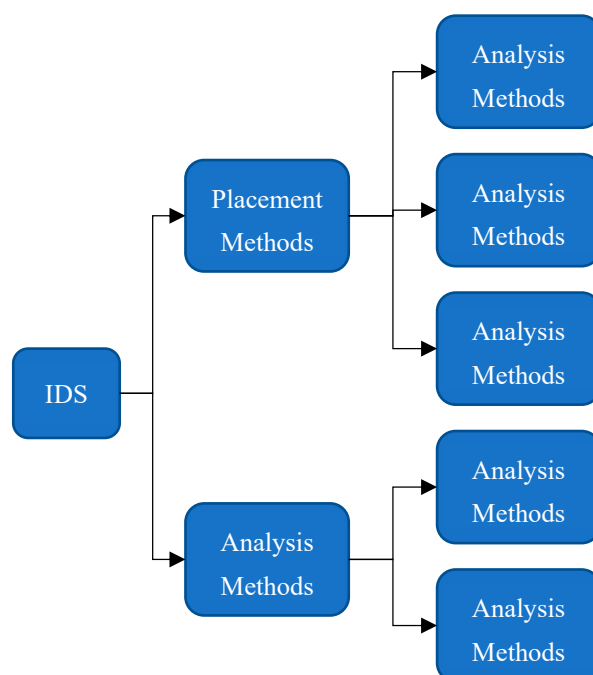


**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Computer networks have become more vulnerable to penetration and to exploits exposing information, due to the Internet being completely open to users. Recently, network attacks are becoming more sophisticated and harder the detect. The statistical studies according to the Symantec Global Internet Security Threat report indicate that intrusions

are at record levels and are increasing drastically. Given the exponential growth of the dependence on data, algorithms for data protection from threats and attacks are greatly needed to preserve the privacy, confidentiality, availability, and integrity of information systems. Hence, intrusion detection systems (IDSs) are a vital tool for protecting data. Figure 1 shows IDS categories. While detection algorithms only recognize well-known attacks, anomaly detection algorithms can recognize unidentified attacks according to users' behavior. Two of the main issues with anomaly detection are speed and efficiency [1]. If the network has high traffic, it is almost intolerable to utilize a fast sophisticated algorithm for intrusion detection (ID) in advance. Many new procedures achieve a good rate of IDS, but they need high resource allocation and are time-consuming (i.e., communication, memory energy, or another system requirement). These deficiencies may become even more complicated if the traffic is manipulated in real-time.



**Figure 1.** IDS categories.

One of the secure design principles is defense-in-depth, which implies adding multiple security mechanisms to prevent, detect, contain, and recover from attacks. Security mechanisms such as access control, multi-factor authentication, and data encoding are being utilized to act as a frontline defense to prevent potential attacks [2]. Protection procedures and tools, such as intrusion preventions, anti-viruses, firewalls, and IDSs, can monitor the activity of network systems to detect, prevent, and counter suspicious actions [3]. IDSs enable the continuous monitoring of the network traffic to detect anomalous activity in the systems, which is considered a vital method to perform network security [4]. The use of machine learning (ML) and statistical methods enable the building of an effective IDS to protect the networks [5].

Due to advances in computer network technologies, people have relied heavily on network services to obtain information over the Internet. According to the statistics provided in information security, there are a large number of threats that affect computer systems and information. Therefore, defense mechanisms are continuously being developed to preserve the integrity of computer systems and networks to ensure the confidentiality of information. An intrusion detection system (IDS) provides a major role in protecting. An effective IDS enables the protection of computer and information systems from potential intrusions as well as helps detect intrusion and misuse. The process of detected system

anomalies protects against potential new attacks or a zero-day attack, where anomalies are detected based on user behavior.

There are so many data mining techniques to classify the normal or abnormal behavior of a user. However, these data mining techniques have some limitations, so the main objective of the research is to reduce these limitations and improve the accuracy. The glossary zero-day attack was initially denoted to the day's accounts since new software programs were publicly released, so the zero-day program is attained by attacking an inventor's PC before the release. Previous studies presented several data mining (DM) techniques that are based on classifying user behavior, but they have some limitations, as this research provides solutions to reduce them and improve the accuracy of mining. The limitations related to the detection techniques are described as follows:

- Efficiency and speed are the most main issues in anomaly detection systems. One of the problems is related to the traffic volume on the network, as complex detection algorithms cannot be used at an adequate speed if the traffic is high. Many theories of advanced algorithms that rely on a high detection rate were presented, but they are very complex to apply in practice.
- The effect of missing data on the results obtained during classification. Accurate and reliable conclusions cannot be drawn if there are missing data that are important for feature selection during classification.
- One of the most important limitations of intrusion detection algorithms is real-time traffic analysis. The information system is potentially exposed to an intrusion risk if the real-time traffic detection is inaccurate.

This paper provides a novel data mining (DM) technique for the IDS approach. The proposed technique is based on ML and hybrid algorithms, namely, K-mean for cluster formulation and sequential minimal optimization (SMO) for clustering and categorization [6]. The contribution of this paper is in integrating the processes of SMO and K-means clustering approaches to enhance IDS performance, accurately identify new attacks, and increase detection rate [7], in addition to reducing the false-positive alarm rate in real-time by taking advantage of attack patterns.

The remainder of the paper is organized as follows: Section 2 discusses data mining techniques, Section 3 discusses the related works, Section 4 presents the proposed machine learning-based anomaly detection algorithm, Section 5 illustrates the results and discussion, and finally, Section 6 concludes the paper.

## 2. Data Mining Techniques

DM plays an important role in IDS and is used in different data applications. Mining techniques such as classification, association rules, and clustering enable users to make sense of information about intrusions by monitoring network data. IDS categories are based on their scope from standalone PCs to network systems. The most common categorizations are hosts-based IDSs (HIDSs) and network IDSs (NIDSs). The system that monitors significant files in OS is an instance of an HIDS, while the systems that examine traffic of the received network is an instance of an NIDS. The following paragraphs classify ID techniques used in DM applications [8].

Classification is defined as the process of analyzing data by taking an instance of the dataset to be assigned to a specific class and extracting models known as classifiers that define important data categories [9]. An IDS is a server or software program that screens and monitors network traffic for malicious activities or violations of security policies. The system, which relies on the classification concept, sorts network traffic into normal or malicious. The process of data classification is divided into two parts. The first part is known as the learning period, during which a classifier is created, and from it, the data categories are predicted in the second part, which is the classification step. In classification analysis, the end-user/analyst needs to know how to define the categories in advance.

In the classification process, the main goal of the classifier is to explore the data to discover the different categories, in addition to arranging the new records into the

category [10]. Many classification techniques are used, such as decision tree induction, genetic algorithm, fuzzy logic, and Bayesian networks-nearest neighbor classifier. In general, data classification techniques have a lower impact on ID methods compared to data clustering techniques, which have a great impact on the performance of IDSs. This is because of the high amount of data required to classify the dataset into normal and abnormal categories.

The clustering approach provides an easier and faster classification process than human labeling for a large amount of data. It enables the labeling of data and grouping it into similar objects. Each group is known as a cluster and consists of several members with similar traits, and the members differ from one group to another. Clustering methods can be useful for classifying network data to detect intrusions. There are several clustering algorithms, and they are divided into five groups as follows [11]:

**Hierarchical clustering techniques:** This method creates tree-based structure classification from unclassified data assets. It can be developed with the assistance of statistical methodologies.

**Density-based techniques:** This technique strains the arguments of each cluster from a precise distribution probability. It can only be utilized for spherical-based clusters. The value of a density-based cluster considers the point's density, where density arguments should be prepared before dataset scanning.

**Grid-based techniques:** The key benefit of this algorithm is its vast calculation time, regardless of the number of data cells. The object band is quantized into a predetermined number of cells.

**Model-based techniques:** This method calculates the greatest data fit based on the hypothesis model. The number of clusters based on statistical standards can be determined repeatedly. The algorithm may construct clusters based on a modeling density probability that imitates the distribution of 3D data objects.

**Partition techniques:** In this technique, for  $n$  points datasets with hypothesis  $k$  data dividers, each point should fit precisely one cluster, and each cluster should comprise at least one point. The dividing method enhances the reiterative re-partitioning method by removing points from one cluster to another. The method of data division relies on a specific partitioning function.

Table 1 discusses the differences and comparisons between these techniques. The importance of the table lies in defining the differences among the most-studied clustering techniques and their use in improving anomaly detection in recent years so that each of them is clarified by presenting the strengths and weaknesses against the used mechanism (K-means).

**Table 1.** Differences between the various clustering techniques.

Clustering Techniques	Advantages	Disadvantages	Example
Hierarchical	Flexible for all shapes	Slow	DAINA
Density-based	Simple and fast	The number of clusters should be specified prior	K-mean
Grid-based	All data types	Difficult quality of clusters	DBSCAN
Model-based	Scalable and vast	Poor for 3D data cluster	STENG
Partition-based	Powerful and flexible	Only for abnormal	CBOWEB

Clustering technologies detect complex intrusions over various periods and act as unsupervised learning mechanisms to discover patterns in multidimensional unpaired data [12]. The patterns within the cluster are equivalent to each other but differ from one group to another. Therefore, the abnormal patterns indicate the occurrence of unusual activity, which may be pointing to the possibility of infiltration of the data or a new attack.

The importance of using the clustering mechanism helps in discovering errors and misuse in addition to reporting the possibility of an attack.

### 3. Related Work

Previous studies presented several DM techniques that are based on classifying user behavior, but they have some limitations. This research provides solutions to reduce these limitations and improve the accuracy of mining. Some of the main limitations and challenges related to the detection techniques are described below.

One of the most important limitations of ID algorithms is real-time traffic analysis. The information system is potentially exposed to an intrusion risk if real-time traffic detection is inaccurate.

Efficiency and speed are the main issues in anomaly detection systems. One of the problems is related to the traffic volume on the network, given that complex detection algorithms are used at an adequate speed if the traffic is high.

The effect of missing data on the results obtained during classification. Accurate and reliable conclusions cannot be drawn if there are missing data that are important for feature selection when carrying out classification.

According to the abovementioned limitations and challenges, this section reviews a survey of various ID models and techniques. It also presents the methodologies used to develop the IDS and the latest updated models. A study by C. Taylor et al. [13] proposed an approach known as the network analysis of anomalous traffic events (NATE), which is based on clustering and multivariate analysis. NATE can enhance the ability of IDS to deal with detection constraints and big data traffic [14]. Moreover, NATE enables performance features of limited attack scope and anomaly detection, in addition to minimizing network traffic measurement [15]. The NATE operation is based on two phases; the first is data collection and analysis for possible attacks, and the second is intrusion detection in the real-time environment [16]. The NATE classification is a cluster-based algorithm. The proposed study shows that the clustering approach enables quick updates of the new attack features for real-time traffic in the database [17].

A. Bakhtiar and G. Antonio [15] provided a production-based expert system toolset (P-BEST) to detect misuse attacks and develop a new signature mechanism. P-BEST can provide efficient IDS performance in a real-time environment [16]. The proposed mechanism allows integration with C programming for flexibility and ease of use. However, it has a low detection capability of intrusions and attacks with incomplete and uncertain data or unknown environment information [18].

C. Zheng et al. [19] presented a framework for DM to build IDS models. The proposed framework enables the automatic use of the IDS model [20]. The operations of the DM framework tiers are dependent on the ability of inductively learned computations related to relevant system features, raw audit data processing, and network-dumped data, which are all summarized into connection records and attributes [21]. This approach applies two algorithms: association rules and frequent episodes [22].

M. Saeed et al. [23] presented the use of a decision tree for multiple host-based detector combinations. The proposed idea depends on the ID measures and decision tree. The measures are considered the basis of IDS modeling [24]. The modeling measures are performed by the statistical rule-based method [25].

Minegishi, T. et al. [26]. presented a framework for data mining (DM) to build IDS models. Three tiers are reviewed data mining framework parties, classification, association rules, and frequent episodes programs. The proposed framework enables the use of the IDS model automatically [23]. The operations of the data mining framework tiers are dependent on the ability of inductively learned computations related to relevant system features, the raw audit data processing, and the network dumped data, which are all summarized into connection records attributes [23]. This approach applies two algorithms, association rules, and frequent episodes.



The author Minegishi, T. and his team also presented an anomaly detection system known as audit data analysis and mining (ADAM) in [27]. The proposed ADAM uses data mining techniques for detecting intrusions. It combines classification and association rules algorithms for discovering attacks in TCP dump [26]. ADAM can classify suspicious activities of known and unknown attack connections.

Barbará, D. et al. [28] presented the use of a decision tree for multiple host-based detectors' combination. The proposed idea depends on the intrusion detection measures and decision tree. The measures are considered the base of IDS modeling. The modeling measures are performed by the statistical rule-based method.

Another study presented by Zhang et al. [29] provides a hybrid misuse and anomaly detection approach for NIDS. The study investigates the combination of two detection methods to reduce the limitations of both when they are considered individually. The proposed hybrid detection approach is evaluated as a technique for data mining intrusion detection for the random dataset.

In the study by P. Yuhuai et al. [30], a method was proposed to improve the efficiency of the decision tree algorithm. They reviewed the operations of bagging and boosting and randomization techniques to generate various classifier ranges through training data manipulation. Z. Peng et al. [31] provided the ability of artificial intelligence (AI) technology to enhance the accuracy of anomaly detection. The proposed study evaluates the use of semi-supervised learning and unsupervised learning techniques to detect anomalies. The authors used the K-means clustering approach and training instances through the Euclidean distance method and then evaluated the C4.5 algorithm [32]. Their results showed that the semi-supervised training algorithm gives better performance than supervised or unsupervised algorithms.

V. Olena et al. [33] designed an effective intrusion identification system based on the fuzzy logic approach. The proposed system enables the detection of intrusion behavior in the network. It uses a mechanical method to create fuzzy rules, which are obtained from specific rules using repeating elements [34]. Through the results of the experiments, the authors concluded that the system based on fuzzy logic achieves a higher accuracy to determine whether the records are normal or offensive.

G. Azidine et al. [35] proposed the use of two algorithms for ID: the backpropagation algorithm and C4.5 algorithm. In addition to dealing with known attacks, these algorithms are mainly used to detect misuse and determine the level of deviations in normal profiles. They can also explore algorithms based on supervised ML [17]. The authors used KDD CUP99 databases and tested the datasets by the proposed algorithm containing several attack types, such as denial of service (DoS), investigation, user-to-root (U2R), and remote-to-local (R2L). Through the results obtained, the study showed that the use of neural networks provides high performance in detecting known attacks, but the use of decision trees gives a higher and more exciting performance when detecting new attacks [36].

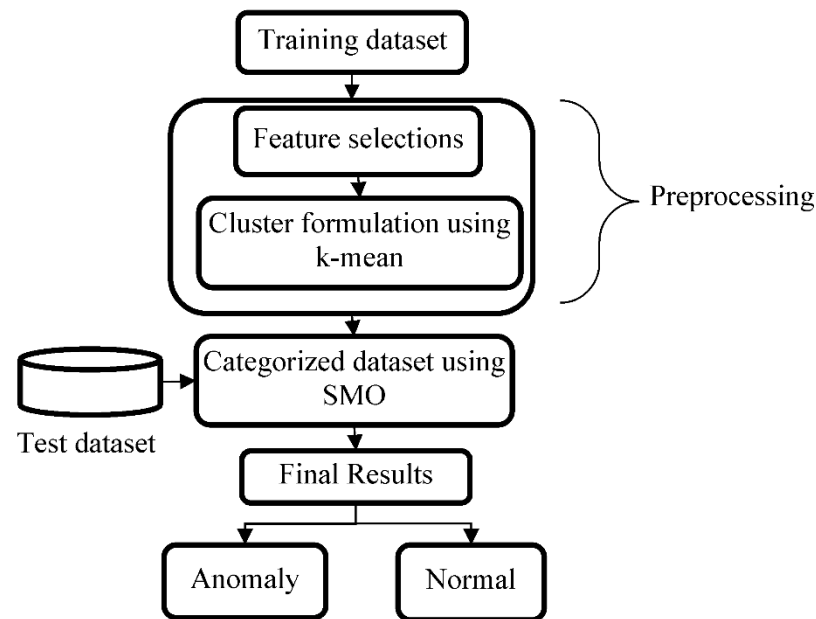
G. Mutanov et al. [37] proposed a hybrid ML technique for network ID based on a combination of K-means clustering and support vector machine classification. This research aims to reduce the rate of false-positive alarms and false-negative alarm rates and improve the detection rate. The authors used the network security-knowledge and data discovery (NSL-KDD) dataset, and the classification was performed by using a support vector machine [38]. After training and testing the proposed hybrid ML technique, the results showed that the proposed technique achieves a positive detection rate and reduces the false alarm rate.

In our work, a new anomaly detection with a hybrid DM algorithm is proposed, with the key aim of improving the detection rate and reducing the false-positive alarm rate. The study uses a dataset known as NSL-KDD to assess the hybrid K-means clustering and SMO ML technique.

#### 4. Machine Learning-Based Anomaly Detection Algorithm

First, the idea of the normality concept is introduced to obtain a suitable solution for anomaly detection in the network. The idea of the normal is linked to the development of a formal model that clarifies the relationship between the basic variables related to the system dynamics. Accordingly, the degree of deviation that appears in the behavior of the system due to the detection of any event or anomaly is measured depending on the normal state model. Our methodology depends on using a new anomaly detection method based on K-means clustering and SMO algorithms [39]. The detection process is tested in the online network to generate an appropriate number of detectors with a high detection rate and accuracy.

The proposed method is set to reduce the number of features by using feature selection algorithms in the preprocessing phase. The selection of particular features from the dataset is performed by applying consistency-subset level and genetic search algorithms. The selection method removes the irrelevant features before the cluster formulation and categorization operations and, subsequently, the K-mean cluster formulation process. The K-mean cluster formulation decreases the dataset training, processing time, and complexity. The classification process is based on SMO to enhance the detection quality. Figure 2 shows the methodology flow chart and describes the model diagram.



**Figure 2.** Model for machine learning-based anomaly detection with a hybrid data mining algorithm.

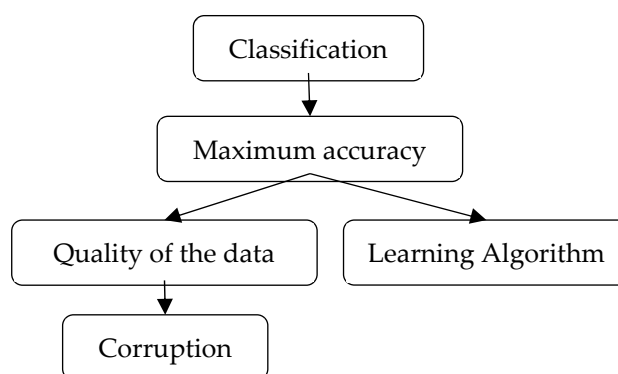
In the model diagram shown in Figure 2, normal and abnormal data are trained, taking the assumption that attack data do not occur as frequently as normal data. The training data contain both normal and abnormal data. We assume that attack data will not occur as frequently as normal data. This assumption is valid because intrusion-based attackers usually monitor the data traffic based on collected samples, where larger samples allow the attackers to have a successful intrusion. Thus, less than  $x\%$  of the data is anomalous, where  $x$  is equal to the intrusion and attacks probability. The preprocessing for the original NSL-KDD dataset helps prepare the trained data for the classification stage and reduces the data mystery. In addition, it provides accurate information to the following detection engine. The stage of the preprocessing depends on feature selections and cluster formulation-based K-means approaches. This stage helps clean the network's data by collecting and processing lost or inadequate attributes as the data are preprocessed through the following stages.

#### 4.1. Feature Selection

Feature selection is one of the vital stages in the preprocessing in our hybrid model, given that the selection output plays an important role in the ML process and the cluster formulation of the K-mean. Thereafter, it affects the final effectiveness of the DM algorithm of the ID model. Overall, the input parameters are classified for the high-dimension feature stream.

However, some features will not be relevant to the hybrid algorithm process because the classification will not involve them. Inappropriate, duplicate, or noisy data may divert the ML process if not excluded early in the process. The existence of these irrelevant data in the dataset may increase the complexity of the model and learning time, which accordingly degrades the performance of learning algorithms. Thus, removing such features will improve the performance of the clustering and classification algorithms, leading to positive effects on the entire IDS performance. It likewise helps in speeding up the detection process and enhancing the precision of results and overall security. Therefore, there is a need to identify and handle these irrelevant datasets. Some methods are specified to detect irrelevant attributes in the datasets (i.e., pairwise attribute algorithm, PANDA) to detect the most and least noisy thresholds. PANDA is useful due to its work without class knowledge [40]. Noise identification is then followed by irrelevant feature handling and removal. There are three main approaches: identify and ignore, identify and filter out (i.e., fuzzy wavelet analysis), and finally, identify and scrub (i.e., automatic repeat request, forward error correction, or hybrid techniques) [41].

Real data traffic is mostly exposed to many factors. Some of the most important factors are noise, inappropriate, and duplicate data. Data corruption, where data are severely affected during data collection and data preparation procedures, is an inevitable issue. There are two types of errors: implied noise presented by collection utilities (i.e., sensor) and random error presented by random sources as additive white Gaussian noise. Inappropriate and duplicate data are actual data that have been used incorrectly at the wrong time or allocated in the wrong way. The system throughput and efficiency seriously rely on the training data quality, and the strength is compared with the categorizer's errors. The flow is illustrated in Figure 3.



**Figure 3.** Effect of inappropriate, duplicate, or noisy data on machine learning.

For example, the data may include inappropriate, duplicate, or noisy data, which may be excluded. In this case, redundant and irrelevant features can be introduced as noise or interfered data that divert the ML process. Feature selection reduces the number of attributes; removes inappropriate, redundant, or noisy features; and leads to positive effects on system performance output, such as speeding up the DM algorithms, enhancing ML precision, and resulting in a better security model.

The feature selection procedure is demonstrated in Figure 4. Figure 4 shows that if we have the set of features ( $F_0$  to  $F_N$ ) from the original dataset, which was collected from the network traffic, then the selected features achieved from the selection tool (i.e., feature selection toolbox (FST) in MATLAB) are ( $F_0$  to  $F_M$ ). The feature number achieved differs



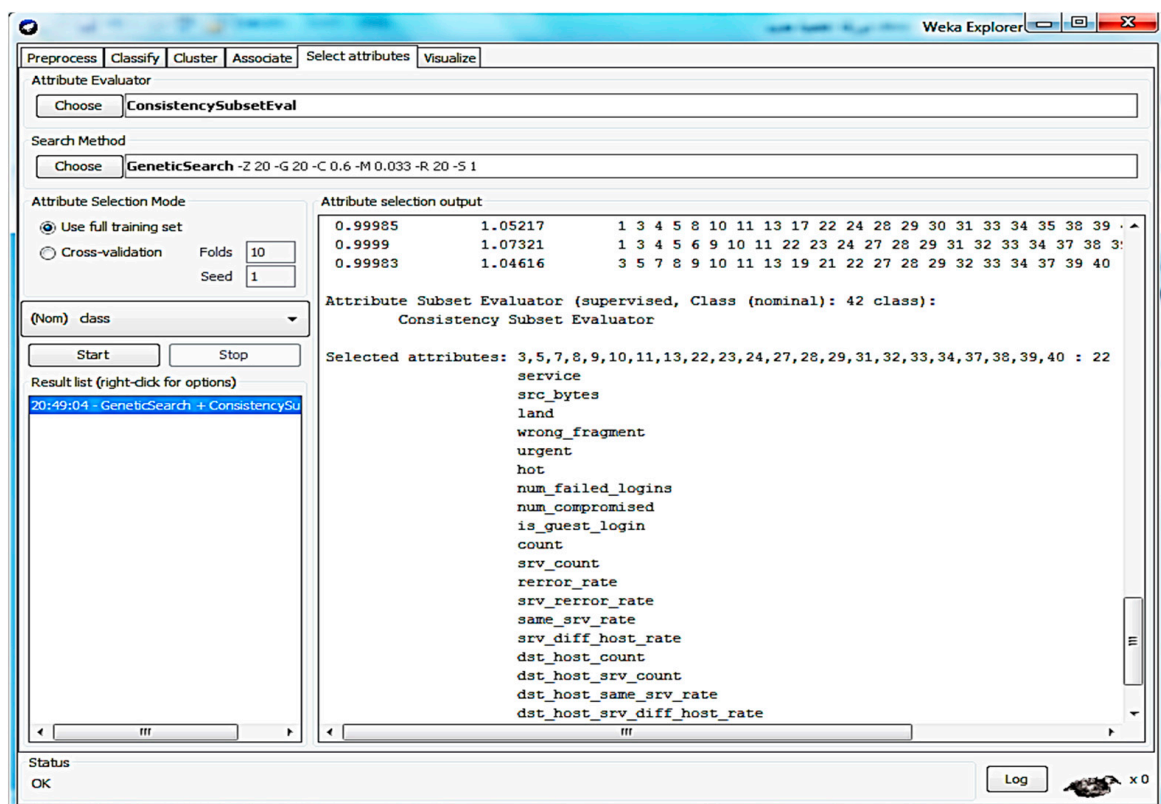
based on the utilized tool for selection and the feature's cross-correlation that has been fed to the tool. The next stage involves the elementary philosophies of the analysis of features. The number of features in the output, indicated by  $M$  in the illustration, is usually less than the number of features in the input indicated by  $N$ . However, in some special cases, the input and output may be equal.



**Figure 4.** Feature selection using the FST tool.

After this stage, the dataset will include only the most effective features to be fed to the ID engine. The selection of features aims to improve the anomaly detection ratio and reduce the false alarm probability (FAP) of ID in the network. Waikato Environment for Knowledge Analysis (WEKA 3.9.4) is an ML tool used in this study to calculate the selection of feature subsets for our hybrid DM scheme. This is performed by the categorization of the test throughput on each feature subset.

Figure 5 represents the program interface of WEKA 3.9.4 software, which provides ML tools to enable the process of feature selection. It enables setting the consistency-subset level and attribute parameters. The consistency-subset level is an element-level processor providing reiteration within a data block in the data subset. The genetic-search scheme is used to select certain features from the dataset and eliminate features that are irrelevant to the process before the cluster formulation and categorization stages.



**Figure 5.** Feature selection consistency-subset level and genetic search algorithms.

#### 4.2. Clustering Phase

The cluster formulation stage is performed by using K-means cluster formulation algorithms. Three clusters are created and tested. The method is iterated through data

training; the cluster's structures have been inherited from each other. The cluster update causes the centroid values to be modified, which affects the present cluster elements. If there are no more modifications in the cluster centroids values, that means the cluster formulation for the K-means process has become stable. After that, the clustering model is built.

The generated network traffic due to the data collection is analyzed in the testing stage. The final stage is categorization, where it is managed; that is, the SMO algorithm is utilized to categorize the database as an anomaly or normal. The K-mean approach is usually easier and simpler to be implemented, while the construction of SMO is comparatively higher in terms of computational complexity.

One of the great applications of the proposed method is in the area of IoT and WSN with a high traffic volume. One of the strengths of the proposed method is the high accuracy of intrusion detection using data mining. The proposed method was designed to deal and process a high traffic volume without affecting the performance.

## 5. Results and Discussion

Utilizing Waikato environments has developed experiments for Knowledge Analyses (WEKA®). WEKA® is a utility that is used efficiently in ML and DM; it was developed and invented in 1997 by researchers at Waikato University and in New Zealand. WEKA® [41] is a combination of DM and ML procedures that are built using JavaBean codes and data files, and it has been used for graphic user interfaces (GUIs) to exchange for comfortable human interaction. WEKA® comprises 76 categorization procedures, 15 feature assessors, 10 procedures for search and mining for feature selection, and 49 data preprocessing utilities.

There are five procedures to find association rules. WEKA® also has six GUIs, including knowledge flow, explorer, experimenter, etc. The data storage format is stored in an attribute-relation file format (arff). It comprises utilities as well for imagining. WEKA® has plenty of boards that can be utilized to perform specific threads. It also has great scalability, which can be extended and comprised of any newly developed ML algorithm. These extended or new algorithms can be linked to the database directly. The detection of attack can be measured by the following metrics: (A) false-positive (FP) or false alarm corresponds to the number of detected attacks, but these attacks are normal; (B) false negative (FN) corresponds to the number of detected normal instances, but these instances are actual attacks; that is, they are the target of IDSs; (C) true positive (TP) corresponds to the number of detected attacks that are in fact attacks; and (D) true negative (TN) corresponds to the number of detected normal instances that are normal.

In this paper, three parameters of measurement are used, namely, detection rate (DTR), false-positive rate alarm (FPR), and accuracy (AC). DTR is defined as the ratio of attacks detected to the total number of attacks. This is the best parameter to measure the performance of the model and is determined using Equation (1) [42]:

$$DTR = \frac{TP}{TP + FN} \times 100 \quad (1)$$

FPR is one of the main parameters to find out the effectiveness of various models and is a major concern during network setup. Normal data are considered abnormal or attack-type data. FPR is obtained using Equation (2):

$$FPR = \frac{FP}{TN + FP} \times 100 \quad (2)$$

AC is the proportion of the total number of the correct predictions to the actual dataset size. It is determined using Equation (3), and then the matrix of confusion can be presented as shown in Table 2.

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (3)$$

**Table 2.** Matrix of confusion.

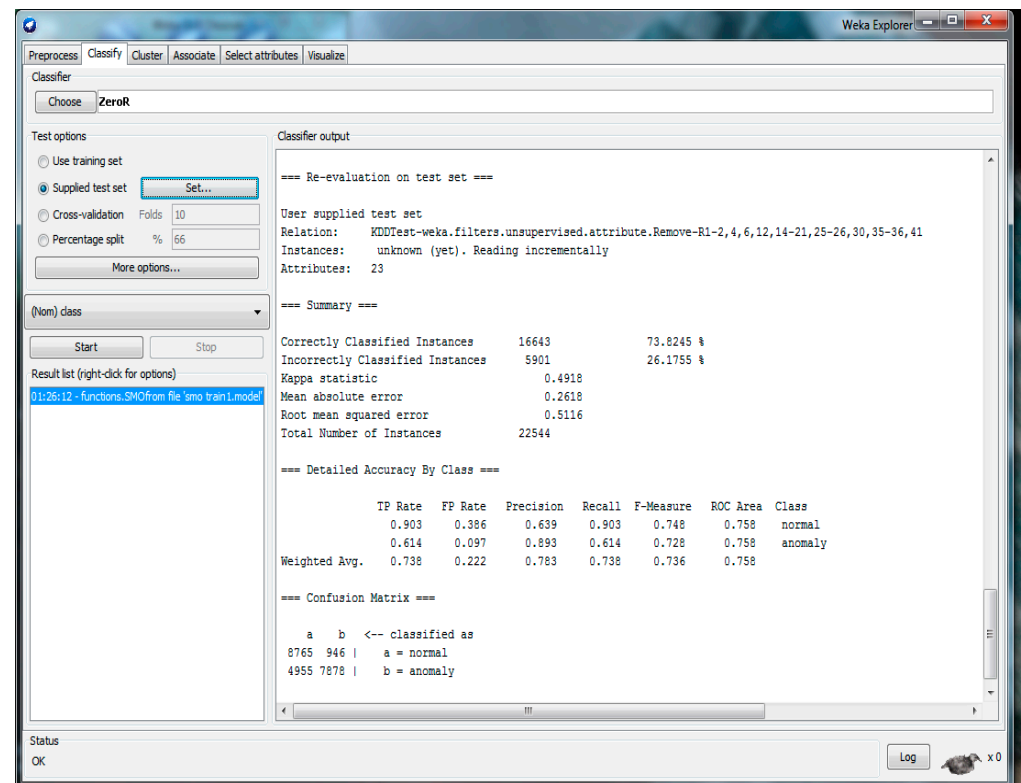
Clustering Techniques	Predicted: Abnormal	Predicted: Normal
Actual: Abnormal	True Positive	False Negative
Actual: Normal	False Positive	True Negative

### 5.1. Accuracy Measure of Individual Algorithms (SMO)

To perform the data analysis and prediction of algorithms that were used to build our model, firstly, we applied the SMO algorithm for the NLS-KDD dataset with 22 attributes (with feature selection) using WEKA, as shown in Figure 6. The details of accuracy parameters for SMO are shown in Table 3. The details of the confusion matrix are shown in Table 4. Finally, the measurement parameters for SMO are calculated and presented in Table 5, whereby using Equations (1)–(3), we easily can calculate AC, FPR, and DTR. From Table 4, one can calculate:

Correctly Classified Instances (CCI 16643) = 73.8245%

Incorrectly Classified Instances (ICI 5901) = 26.1755%

**Figure 6.** SMO result using a dataset with 22 attributes.**Table 3.** Demonstrate details of accuracy parameters for SMO.

TP Rate	FP Rate	Precision	Recall	F-Measure	F-Measure	Class
0.903	0.386	0.639	0.903	0.748	0.758	normal
0.614	0.097	0.893	0.614	0.728	0.758	anomaly
0.738	0.222	0.783	0.738	0.736	0.758	Weighted Avg

**Table 4.** Confusion matrix for SMO.

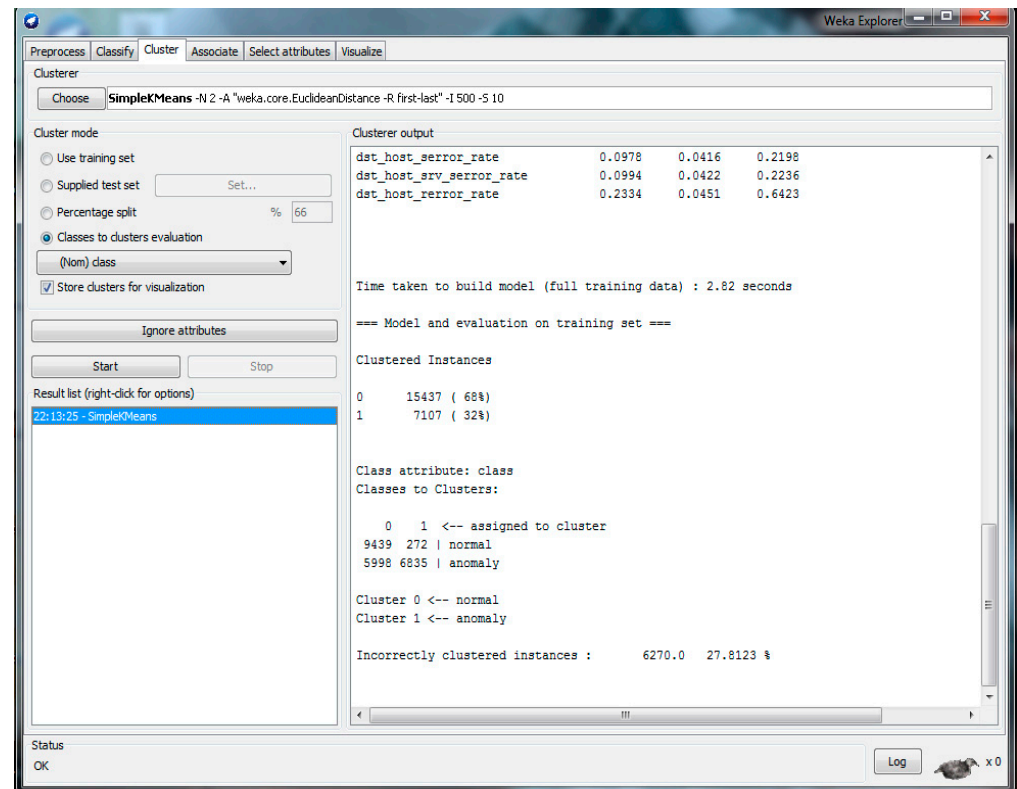
a	b	Classified as
TN = 8765	FP = 946	a = normal
FN = 4955	TP = 7878	B = anomaly

**Table 5.** SMO parameters measurement.

Algorithm	DTR	FPR	AC
SMO	61.39	9.7	73.82

### 5.2. K-Mean Algorithm Implementation

The result of the NLS-KDD dataset with 22 attributes and feature selection using WEKA is shown in Figure 7. Then, the CCI was 72.1877% and ICI was 27.8123%. From the achieved results in Figure 7, the details of the confusion matrix are presented in Table 6. Then, Table 7 demonstrates measurement K-mean parameters where, by using Equations (1)–(3), AC, FPR, and DTR were calculated.

**Figure 7.** K-mean clustering result using a dataset with 22 attributes.**Table 6.** Confusion matrix for K-mean.

0	1	Classified as
TN = 9439	FP = 272	0 = normal
FN = 5998	TP = 6835	1 = anomaly

**Table 7.** Measurement of K-mean parameters.

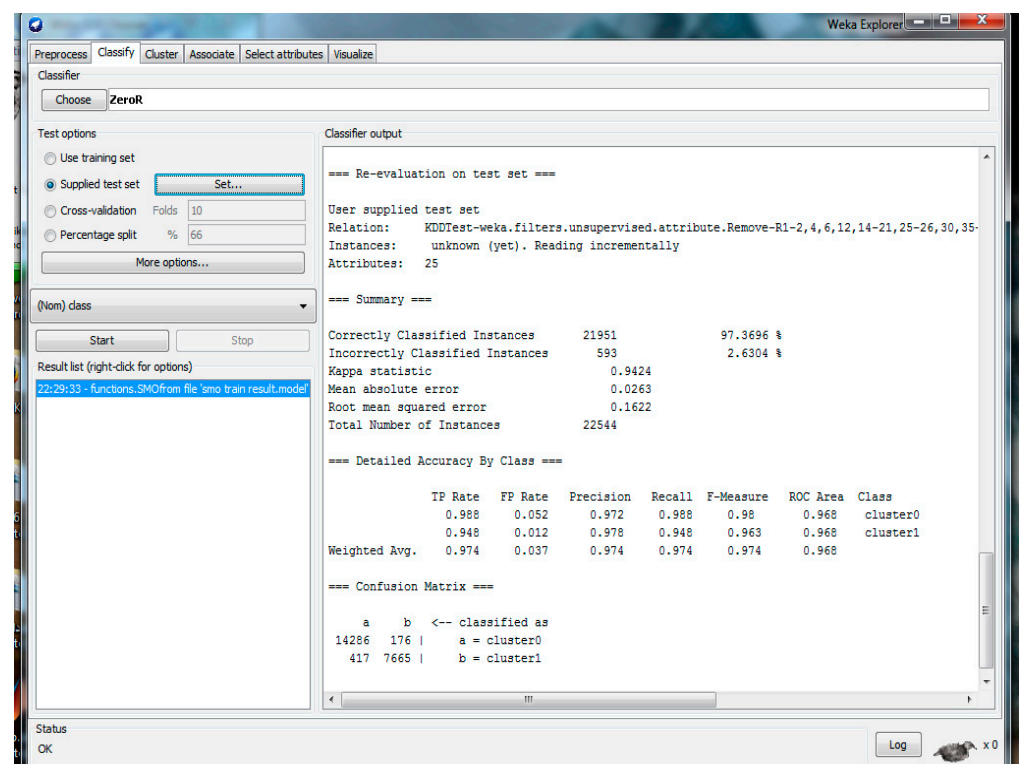
Algorithm	DTR	FPR	AC
K-mean	53.26	2.8	72.188

### 5.3. Hybrid Technique, i.e., K-mean and SMO Implementation

The resultant NLS-KDD dataset with 22 attributes and feature selection using WEKA is shown in Figure 8. Based on the results in Figure 8, the accuracy details are presented in Table 8, while the confusion matrix is presented in Table 9. One can then calculate

Correctly Classified Instances (CCI 21951) = 97.369%

Incorrectly Classified Instances (ICI 593) = 2.6304%

**Figure 8.** Results of using hybrid technique, i.e., K-mean, SMO, and dataset with 22 attributes.**Table 8.** Accuracy details for hybrid technique, i.e., K-mean and SMO parameters.

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.988	0.052	0.972	0.988	0.98	0.968	normal
0.948	0.012	0.978	0.948	0.963	0.968	anomaly
0.974	0.037	0.974	0.974	0.974	0.968	Weighted Avg

**Table 9.** Confusion matrix for the hybrid K-mean and SMO method.

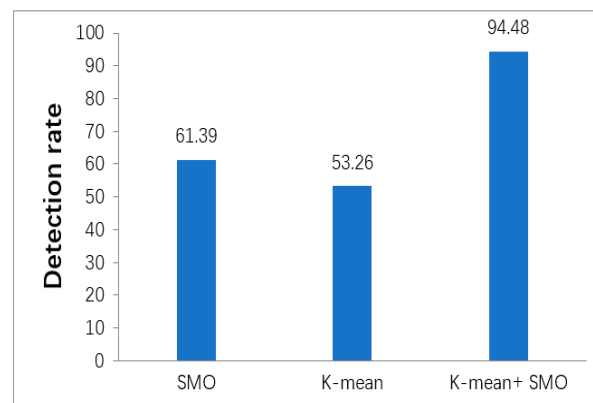
a	b	Classified as
TN = 14286	FP = 176	a = normal
FN = 417	TP = 7665	b = anomaly

Table 10 demonstrates the measurement of K-mean parameters where, by using Equations (1)–(3), AC, FPR, and DTR were calculated.

**Table 10.** Measurement parameters for the hybrid K-mean and SMO method.

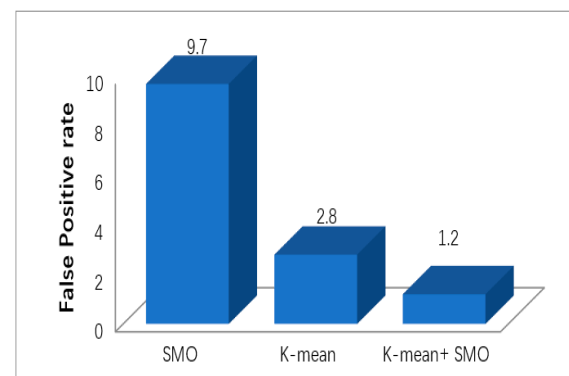
Algorithm	DTR	FPR	AC
Hybrid K-mean and SMO	94.48	1.2	97.3695

Figure 9 shows the comparison of the detection rate for K-mean, SMO, and SMO jointly with the K-mean DM. The detection ratio represents the correctness of a model for detecting intrusion. The experimental result shows that the proposed algorithm performs better in terms of correctness in detecting intrusion (94.48), while other individual DM techniques perform as follows: SMO = 61.39 and K-mean = 53.26.

**Figure 9.** Comparison of detection rate for K-mean, SMO, and SMO jointly with K-mean data mining.

The improvement obtained by the proposed method was the result of an increase in the detection rate and a decrease in the false alarm rate in the network ID. Owing to these two enhancement features, K-means enables testing of the classification performance on each feature set, while the supervised algorithm (SMO) improves the detection quality by reducing the number of features in the dataset.

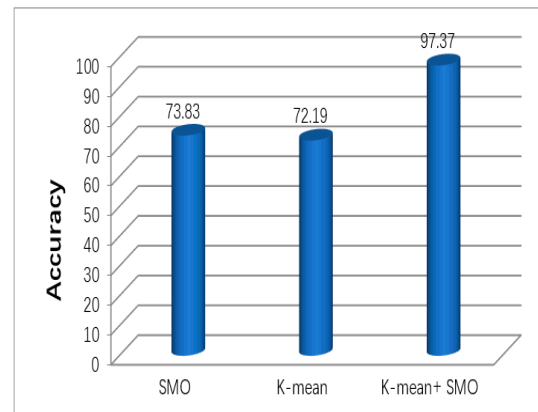
Figure 10 shows the comparison of false-positive rates for K-mean, SMO, and SMO jointly with K-mean DM. The false-positive rate of the proposed model performs better (1.2) compared to other individual models of DM techniques (SMO = 9.7 and K-mean = 2.8). This parameter is a very important measure to evaluate the performance of a model. Hence, the results show that the proposed model performs better than other models.

**Figure 10.** Comparison of false-positive rates for K-mean, SMO, and SMO jointly with K-mean data mining.

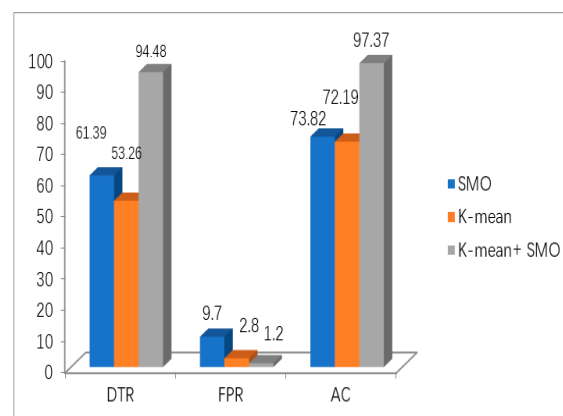
According to the results presented in Figures 11 and 12, the experimentation results show that the proposed SMO jointly with the K-mean model is more accurate than the



other individual DM techniques (SMO = 73.82 and K-mean = 72.188). Figure 12 shows that the accuracy of the proposed SMO jointly with the K-mean model is 97.3695. The SMO jointly with the K-mean model accomplishes a better performance than the SMO and K-mean when applied separately.



**Figure 11.** Comparison of accuracy for K-mean, SMO, and SMO jointly with K-mean data mining.



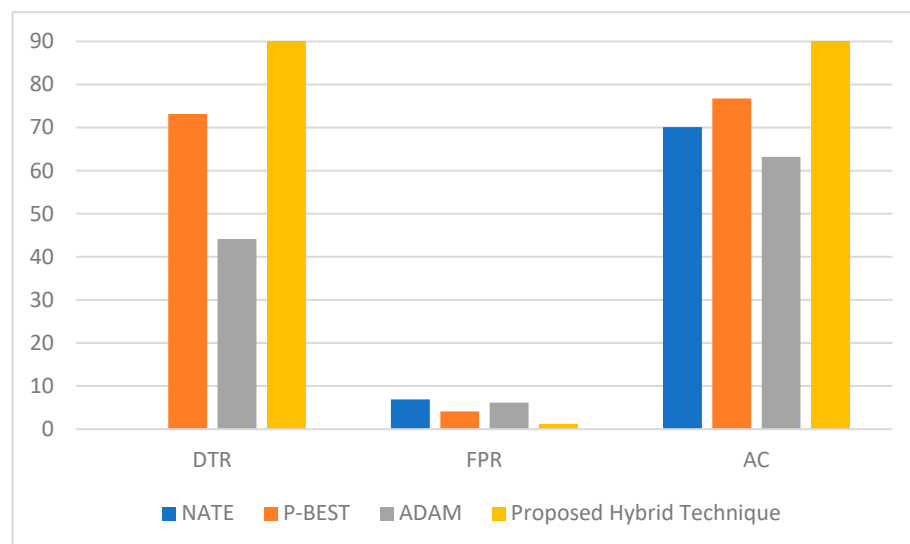
**Figure 12.** Benchmarking for K-mean, SMO, and SMO jointly with the K-mean data mining of measurement parameter.

In terms of anomaly detection probability or anomaly detection ratio, which is defined as the model correctness for ID, the computer experimental output presents that the SMO jointly with the K-mean algorithm also achieves better performance in terms of ID correctness (94.48), while other separated DM techniques perform as follows: SMO = 61.4 and K-mean = 53.3.

Figure 12 shows the performance of the DTR, FPR, and AC of the proposed hybrid model compared with SMO and K-means. The proposed model gives 1.5-times better performance than other models without the hybrid DM techniques (SMO = 9.7 and K-mean = 2.8). In addition, the accuracy and detection rates for the hybrid model are enhanced by approximately 74.97% and 60.6%, respectively, compared to other models. FAP is quite a significant parameter for the simulation design to assess the performance of the proposed approach. Therefore, the results presented that the SMO jointly with K-mean model performance was better than the other related models in the literature.

From the discussion and experimentation outcomes, it was shown that the application for various scenarios with evaluation parameters of the proposed algorithm, SMO jointly with the K-mean model, achieved acceptable results. By using the hybrid algorithm of DM dataset scenarios, the anomaly detection probability was enhanced significantly. For most of the scenarios, the main aim was to enhance the anomaly detection probability, but FAP was also reduced greatly, and detection accuracy was maximized.

Figure 13 shows the benchmarking between the proposed hybrid algorithm and the other related algorithms in terms of DTR, FPR and AC. Similar parameters were used for all algorithms wherever applicable. The hybrid proposed algorithm shows better performance with about 18% from P-BEST with acceptable processing complexity.



**Figure 13.** Comparison between the proposed hybrid algorithm and other three related works.

## 6. Conclusions and Future Work

This study proposed a hybrid method for anomaly detection utilizing K-mean cluster formulation and SMO categorization. The methodology precisely addressed topics that arise in the framework of large-band databases. SMO utilizes feature selection in preprocessing stages to improve the dataset. The consistency-subset level and genetic search algorithm were used to choose certain features from the NLS-KDD dataset and eliminate the features that are inappropriate for the process before the clustering and categorization stages. Then, K-means was used for clustering to eliminate the training of the training datasets while keeping the processing time under a certain threshold.

The administered categorization algorithm known as SMO was adopted to enhance the detection quality. A benchmarking was conducted for the contributed approach using SMO jointly with K-mean DM with other related algorithms. The results present that the proposed algorithm outperformed recently and closely related works, i.e., NATE, ADAM, and P-BEST, using similar parameters and the environment by approximately 14.48%; the FAP was reduced by 12%, and a high accuracy of 97.4% was reached.

The proposed algorithm can be considered for anomaly detection in future DM systems, where online processing time is highly likely to be reduced. The justification is that the joint algorithm provides appropriate numbers of detectors to be generated with an acceptable accuracy detection and trivial FAP. Owing to a low FAP, it is highly expected to reduce the time of the preprocessing and processing.

However, a few challenges should be resolved, such as patterns and anomaly detection in massive datasets in real-time, and achieving a practically unlimited number of variables and processing power, which must be addressed by extensive research and development. The machine learning-based anomaly detection using k-mean array and sequential minimal optimization shows a significant efficiency and speed. However, with the increase of the real-time traffic volume on the network, it needs high-performance processing to maintain acceptable performance, especially in real-time analysis.

With data becoming a great business, any disturbance in enterprise data may cause serious outages that lead to exorbitant costs. Future works in ML-based data anomaly detection will be toward proactive schemes rather than reactive ones, where the detection of

anomalies can be handled in near real-time. This makes ML algorithms have great potential in the near future.

In future work, the proposed approach will be evaluated on other standard training datasets to ensure its high performance. In addition, some other feature selection algorithms can be used that can select the more significant feature and make the system more effective. Additionally, the proposed method classifies the dataset into two classes. Future research can classify dataset into five classes: DoS, probe, U2R, R2L, and normal. Finally, a mathematical model for the proposed algorithm, using ML with K-mean SMO, is significant research work that deserves to be explored.

**Author Contributions:** Conceptualization, S.G.; methodology, R.M.; software, S.G.; validation, R.M., E.S.A. and R.S.; formal analysis, E.S.A.; investigation, R.A.; resources, R.A.; data curation, R.S.; writing—original draft preparation, S.G.; writing—review and editing, M.A. and R.A.; visualization, S.G.; supervision, R.M. and R.S.; project administration, R.M.; funding acquisition, M.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project Number (PNURSP2022R97), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

**Acknowledgments:** This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project Number (PNURSP2022R97), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Joseph, M.V. Significance of data warehousing and data mining in business applications. *Int. J. Soft Comput. Eng.* **2013**, *1*, 329–333.
2. Tellis, V.M.; Souza, D.J.D. Detecting anomalies in data stream using efficient techniques: A review. In Proceedings of the 2018 International Conference on Control, Power, Communication and Computing Technologies (ICCPCT), Kannur, India, 23–24 March 2018; pp. 296–298.
3. Zhang, L.; Chen, Y.; Liao, S. Algorithm optimization of anomaly detection based on data mining. In Proceedings of the 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Changsha, China, 10–11 February 2018; pp. 402–404.
4. Xie, J.; Wu, D.; Liao, T. Method of anomaly detection of temperature data in vacuum thermal test based on data mining. In Proceedings of the Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), Harbin, China, 19–21 July 2018; pp. 1040–1045.
5. Cai, S.; Sun, R.; Hao, S.; Li, S.; Yuan, G. An efficient outlier detection approach on weighted data stream based on minimal rare pattern mining. *China Commun.* **2019**, *16*, 83–99. [[CrossRef](#)]
6. Ali, E.S.; Hasan, M.K.; Hassan, R.; Saeed, R.A.; Hassan, M.B.; Islam, S.; Nafi, N.S.; Bevinakoppa, S. Machine Learning Technologies for Secure Vehicular Communication in Internet of Vehicles: Recent Advances and Applications. *Secur. Commun. Netw.* **2021**, *2021*, 8868355. [[CrossRef](#)]
7. Yang, Z.; Ding, W.; Zhang, Z.; Li, H.; Zhang, M.; Liu, C. A Service selection framework for anomaly detection in IoT stream data. In Proceedings of the International Conference on Service Science (ICSS), Xining, China, 24–26 August 2020; pp. 155–161.
8. Rehman, A.; Hassan, M.F.; Hooi, Y.K.; Qureshi, M.A.; Chung, T.D.; Akbar, R.; Safdar, S. Context and machine learning based trust management framework for internet of vehicles. *Comput. Mater. Contin.* **2021**, *68*, 4125–4142. [[CrossRef](#)]
9. Zhang, L.; Liu, C.; Chen, Y.; Lao, S. Abnormal detection research based on outlier mining. In Proceedings of the 11th International Conference on Intelligent Computation Technology and Automation (ICICTA), Changsha, China, 22–23 September 2018; pp. 5–7.
10. Anandharaj, A.; Sivakumar, P.B. Anomaly detection in time series data using hierarchical temporal memory model. In Proceedings of the 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 12–14 June 2019; pp. 1287–1292.
11. Elmubark, M.A.; Saeed, R.A.; Elshaikh, M.A.; Mokhtar, R.A. Fast and secure generating and exchanging a symmetric keys with different key size in TVWS. In Proceedings of the International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering (ICCNEEE), Khartoum, Sudan, 7–9 September 2015; pp. 114–117.
12. Qin, Y.; Lou, Y. Hydrological time series anomaly pattern detection based on isolation forest. In Proceedings of the IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 15–17 March 2019; pp. 1706–1710.
13. Sun, W.; Zhang, G.; Zhang, X.; Zhang, X.; Ge, N. Fine-grained vehicle type classification using lightweight convolutional neural network with feature optimization and joint learning strategy. *Multimed. Tools Appl.* **2021**, *80*, 30803–30816. [[CrossRef](#)]

14. Nurelmadina, N.; Hasan, M.K.; Memon, I.; Saeed, R.A.; Zainol Ariffin, K.A.; Ali, E.S.; Mokhtar, R.A.; Islam, S.; Hossain, E.; Hassan, M.A. A Systematic Review on Cognitive Radio in Low Power Wide Area Network for Industrial IoT Applications. *Sustainability* **2021**, *13*, 338. [[CrossRef](#)]
15. Amen, B.; Grigoris, A. A Theoretical study of anomaly detection in big data distributed static and stream analytics. In Proceedings of the IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Exeter, UK, 28–30 June 2018; pp. 1177–1182.
16. Cao, N.; Lin, C.; Zhu, Q.; Lin, Y.R.; Teng, X.; Wen, X. Voila: Visual Anomaly Detection and Monitoring with Streaming Spatiotemporal Data. *IEEE Trans. Vis. Comput. Graph.* **2018**, *24*, 23–33. [[CrossRef](#)] [[PubMed](#)]
17. Guezaz, A.; Asimi, Y.; Azrou, M.; Asimi, A. Mathematical validation of proposed machine learning classifier for heterogeneous traffic and anomaly detection. *Big Data Min. Anal.* **2021**, *4*, 18–24. [[CrossRef](#)]
18. Zhao, Z.; Zhang, Y.; Zhu, X.; Zuo, J. Research on time series anomaly detection algorithm and application. In Proceedings of the IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 20–22 December 2019; pp. 16–20.
19. Chen, Z.; Yu, X.; Ling, Y.; Song, B.; Quan, W.; Hu, X.; Yan, E. Correlated anomaly detection from large streaming data. In Proceedings of the IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 982–992.
20. Ergen, T.; Kerpiçi, M. A novel anomaly detection approach based on neural networks. In Proceedings of the 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2–5 May 2018; pp. 1–4.
21. Lee, J.; Park, S. Mobile memory management system based on user's application usage patterns. *Comput. Mater. Contin.* **2021**, *68*, 4031–4050. [[CrossRef](#)]
22. Mei, L.; Zhang, F. A Novel distributed anomaly detection algorithm for low-density data. In Proceedings of the IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Dalian, China, 25–27 August 2020; pp. 197–201.
23. Saeed, M.M.; Saeed, R.A.; Saeid, E. Identity division multiplexing based location preserve in 5G. In Proceedings of the International Conference of Technology, Science and Administration (ICTSA), Taiz, Yemen, 22–24 March 2021; pp. 1–6.
24. Elfahal, M.O.; Mustafa, M.; Mustafa, M.E.; Saeed, R.A. A framework for Sudanese Arabic—English mixed speech processing. In Proceedings of the International Conference on Computing and Information Technology (ICCIT-1441), Tabuk, Saudi Arabia, 9–10 September 2020; pp. 1–6.
25. Provotar, O.I.; Linder, Y.M.; Veres, M.M. Unsupervised Anomaly detection in time series using LSTM-based autoencoders. In Proceedings of the IEEE International Conference on Advanced Trends in Information Theory (ATIT), Kyiv, Ukraine, 18–20 December 2019; pp. 513–517.
26. Minegishi, T.; Niimi, A. Detection of fraud use of credit card by extended VFDT. In Proceedings of the World Congress on Internet Security (WorldCIS-2011), London, UK, 21–23 February 2011; pp. 152–159.
27. Minegishi, T.; Ise, M.; Niimi, A.; Konishi, O. Extension of decision tree algorithm for stream data mining using real data. In Proceedings of the Fifth International Workshop on Computational Intelligence & Applications, Hiroshima, Japan, 10–12 November 2009; pp. 208–212.
28. Barabara, D.; Couto, J.; Jajodia, S.; Wu, N. ADAM: A testbed for exploring the use of data mining in intrusion detection. *ACM Sigmod Rec.* **2001**, *30*, 15–24. [[CrossRef](#)]
29. Zhang, J.; Zulkernine, M. A hybrid network intrusion detection technique using random forests. In Proceedings of the First International Conference on Availability, Reliability and Security (ARES'06), Vienna, Austria, 20–22 April 2006.
30. Peng, Y.; Tan, A.; Wu, J.; Bi, Y. Hierarchical Edge Computing: A Novel Multi-Source Multi-Dimensional Data Anomaly Detection Scheme for Industrial Internet of Things. *IEEE Access* **2019**, *7*, 111257–111270. [[CrossRef](#)]
31. Zhan, P.; Xu, H.; Luo, W.; Li, X. A novel network traffic anomaly detection approach using the optimal  $\phi$ -DTW. In Proceedings of the IEEE 11th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 16–18 October 2020; pp. 1–4.
32. Saeed, R.A.; Saeed, M.M.; Mokhtar, R.A.; Alhumyani, H.; Abdel-Khalek, S. Pseudonym mutable based privacy for 5G user identity. *Comput. Syst. Sci. Eng.* **2021**, *39*, 1–14. [[CrossRef](#)]
33. Vynokurova, O.; Peleshko, D.; Bondarenko, O.; Ilyasov, V.; Serzhantov, V.; Peleshko, M. Hybrid machine learning system for solving fraud detection tasks. In Proceedings of the IEEE Third International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 21–25 August 2020; pp. 1–5.
34. Jwo, D.-J.; Wu, J.-C.; Ho, K.-L. Support Vector Machine Assisted GPS Navigation in Limited Satellite Visibility. *CMC-Comput. Mater. Contin.* **2021**, *69*, 555–574. [[CrossRef](#)]
35. Ng, R.T.; Han, J. Efficient and Effective clustering methods for spatial data mining. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94), San Francisco, CA, USA, 12–15 September 1994; pp. 144–155.
36. Ahmed, Z.E.; Hasan, M.K.; Saeed, R.A.; Hassan, R.; Islam, S.; Mokhtar, R.A.; Khan, S.; Akhtaruzzaman, M. Optimizing Energy Consumption for Cloud Internet of Things. *Front. Phys.* **2020**, *8*, 358. [[CrossRef](#)]
37. Mutanov, G.; Karyukin, V.; Mamykova, Z. Multi-class sentiment analysis of social media data with machine learning algorithms. *Comput. Mater. Contin.* **2021**, *69*, 913–930. [[CrossRef](#)]

38. Dridi, A.; Boucetta, C.; Hammami, S.E.; Afifi, H.; Moun gla, H. STAD: Spatio-Temporal Anomaly Detection Mechanism for Mobile Network Management. *IEEE Trans. Netw. Serv. Manag.* **2021**, *18*, 894–906. [[CrossRef](#)]
39. Alsolami, F.; Alqurashi, F.A.; Hasan, M.K.; Saeed, R.A.; Abdel-Khalek, S.; Ishak, A.B. Development of Self-Synchronized Drones' Network Using Cluster-Based Swarm Intelligence Approach. *IEEE Access* **2021**, *9*, 48010–48022. [[CrossRef](#)]
40. Chang, H.; Feng, J.; Duan, C. HADIoT: A Hierarchical Anomaly Detection Framework for IoT. *IEEE Access* **2020**, *8*, 154530–154539. [[CrossRef](#)]
41. Sun, W.; Chen, X.; Zhang, X.; Dai, G.; Chang, P.; He, X. A Multi-Feature Learning Model with Enhanced Local Attention for Vehicle Re-Identification. *CMC-Comput. Mater. Contin.* **2021**, *69*, 3549–3561. [[CrossRef](#)]
42. Mansour, R.F.; Alfar, N.M.; Abdel-Khalek, S.; Abdelhaq, M.; Saeed, R.A.; Alsaqour, R. Optimal deep learning based fusion model for biomedical image classification. *Expert Syst.* **2022**, *39*, e12764. [[CrossRef](#)]