Scientific Research Publishing

Journal of Information Security, 2025, 16(4), 447-471
https://www.scirp.org/journal/jis
ISSN Online: 2153-1242
ISSN Print: 2153-1234

Attack Detection and Alarming System on IOT Facilities Using Random Forest Enabled-Correlation Based Clustering (RF-CBC) Technique

Adedayo David Adeniyi[1], Rhoda Ajayi[2], Josephine Olamatanmi Mebawondu[3]

[1]Department of Mathematical and Computer Sciences, Faculty of Science, University of Medical Sciences, Ondo, Nigeria
[2]Department of Computer Science, University of New Haven, West Haven, USA
[3]Department of Computing, Afe Babalola University, Ado-Ekiti, Nigeria
Email: aadeniyi@unimed.edu.ng, drdayoadeniyi@gmail.com, talk2rhoda@yahoo.com, jmebawondu@abuad.edu.ng

How to cite this paper: Adeniyi, A.D., Ajayi, R. and Mebawondu, J.O. (2025) Attack Detection and Alarming System on IOT Facilities Using Random Forest Enabled-Correlation Based Clustering (RF-CBC) Technique. *Journal of Information Security*, **16**, 447-471.
https://doi.org/10.4236/jis.2025.164023

Received: December 28, 2024
Accepted: August 19, 2025
Published: August 22, 2025

Abstract

In the past decade, Internet Of Things (IOT) technology has become one of the fastest-growing and most widely used technologies and is rapidly becoming a basic feature of global civilization. However, the high connectivity and diversity of these IOT devices make them complex and vulnerable to both visible and invisible security threats that are capable of causing irrecoverable damage. To alleviate these challenges, this work presents a novel and analytical hybrid machine learning model that suitably combines the Random Forest with Correlation-Based Clustering techniques, in order to report and detect potential attacks on IOT facilities. This work also showcases the development of the Single Threshold Boxplot Outlier-Based feature scaling method (STBO). The (STBO) method is used to scale down the number of attributes in order to select the best feature at the pre-processing stage of the attack detection procedures. The implementation of the present system is accomplished with the aid of an in-house Python program using XAMP/Apache HTTP as the hosting server with MySQL application for database development and management. A comparative analysis of the present model alongside ANN, Traditional Random Forest, Naïve Bayes, and the traditional Clustering method shows that the proposed system outperformed the baseline methods studied, with precision rates and attack detection quality equal to or greater than 75% in most cases, and is therefore capable of providing a useful, faster, efficient, and accurate anomalous detection online and on a real-time basis consistently with low false positive and negative rates.

Keywords

Attack Detection, IOT, Outlier, Correlation, Random Forest, Clustering

1. Introduction

In recent times, there has been an upsurge of interest in the usage of Internet of Things (IOT) worldwide. IOT extends the use of internet facilities by connecting millions of devices with the capability of interacting with one another through smart technologies such as smart cities, smart homes, smart hospitals, smart banking, smart schools, smart agriculture, etc.

Internet Of a Thing (IOT) is becoming one of the fastest growing and most widely used technologies in the past decade in both the private and business domains, with the rapid growth of internet and network connectivity [1] [2]. The production and marketing of smart devices are increasing rapidly; more hardware devices such as sensors, actuators, microcontrollers, etc., and Internet Of a Thing (IOT) software are being introduced daily by manufacturers, purposely to gain a competitive market advantage. Currently, it is estimated that the number of IOT and connected devices is over 20 billion; it is expected that the number will reach up to 50 billion by the year 2025 [3]. However, many of these devices and products failed to take into consideration the issue of security during their design, and are vulnerable to security threats (both visible and invisible). The high number of these IOT devices, coupled with their diversity and complexity, represents a huge security risk such as Denial Of Service (DOS), Data Breaches (DB), tampering, spoofing, privilege escalation, and IOT botnets, to mention just a few, and therefore is capable of causing irrecoverable damage [4].

Security is one of the cornerstones of any information society, such as the Internet Of Things (IOT). The subject of information security has been given much attention in recent years; it has become a very important research and professional topic in the field of the Internet Of Things (IOT). Therefore, the need to ensure the security of the IOT network is of great significance to the success of IOT technology. In recent times, several studies on the topic of attack detection on IOT facilities have been carried out, some of which show great capability in protecting the IOT network. However, a good number of these traditional web attack detection technologies face several challenges, thus the need to research a more viable and progressive IOT attack detection system.

In this work, a novel attack detection and alarming system for IoT networks based on hybrid Two-Phase attack detection techniques is developed by integrating Random Forest with the novel Correlation-Based Clustering (RF-CBC) machine learning techniques. Specifically, the proposed RF-CBC is capable of accepting the Uniform Resource Locator (URL) of potential users on the network traffic, analyzing the URL request in order to detect and report anomalous requests within the network. The contributions of this work are fourfold:

First, a novel hybrid, two-phase machine learning algorithm called Random Forest Enabled-Correlation Distant Based Clustering (RF-CBC) is proposed. Attention is specifically focused on combining the Random Forest technique with the Correlation Based Clustering technique. In the first phase, the Random Forest is used to classify users into different clusters based on their access credentials, while the second phase uses the correlation-based clustering techniques to analyze the browsing pattern of each visitor to the facility in order to detect any abnormality. The Correlation clustering technique uses correlation statistics to determine the similarity between a given tuple and the other tuples in the clusters; classification is done based on the closest correlation to the given tuple instead of the popular Euclidean distance. This enables the designer of the attack detection and reporting system to have more varieties of such algorithms in order to be able to select the best-performing algorithm. The proposed correlation-based measuring technique is computationally efficient and accurate for scalable implementation. It is capable of handling large datasets with no assumption about data distribution. It also has the capability to show dissimilarity between the test sample and training sample. The proposed RF-CBC-based attack detection and alarming system is capable of providing usable, consistent, efficient, faster, and accurate anomalous detection online and in real time with low false positive and negative rates.

Second, this study examines many existing intrusion detection systems on IOT devices with the aim of investigating the performance of the various algorithms used in developing the attack detection system; this is in order to arrive at a more viable and efficient way of realising the present system.

Third, a novel feature selection model referred to as Single Threshold Boxplot Outlier Based Feature Selection (STBO) feature scaling method is proposed. This method is used to scale down the number of attributes in order to select the best features at the pre-processing stage of the attack detection process before applying the proposed RF-C model. This is important in order to overcome scalability and computational complexity problems common to many existing attack detection algorithms while showing capability in handling high dimensionality and noisy data, therefore improving the accuracy of the proposed machine learning algorithm to a large extent.

Fourth, this present work proposed the construction of a specific attack detection and reporting system for Internet Of a Thing (IOT), facilitating the use of an experimental website developed using the Python programming language with XAMP/Apache HTTP being adopted as the hosting server and MySQL database management software for data acquisition, model extraction, and data management at the back end.

The proposed attack detection and reporting system will accept a potential user's browsing URL request, tokenize the URL request, store it in a data mart, and then pass the token to the feature learning model to analyze the URL request and transform it into a vector together with attached anomalous information. The proposed RF-CBC model first determines the access credentials of potential users, then classifies their URL information to determine the presence of any forms of

attack, the result of which is used for final decision making and the identified attack is reported. The classifier is updated using the update module.

This will assist IoT designers and administrators in planning an update of their IoT facilities to determine potential threats and facilitate the protection of IoT facilities against visible and invisible threats, as well as to enlighten the public at large.

The proposed classical, hybrid-double phases attack detection algorithm, the RF-CBC algorithm that serves as the basis for the development of the attack detection and alarming system on the IOT facility, will be presented alongside the proposed Single Threshold Boxplot Outlier Based Feature Scaling (STBO) dimensionality reduction technique. A comparative analysis of the present model was done alongside four other machine learning algorithms, which included the ANN, Naïve Bayesian, Traditional Random Forest, and the traditional clustering algorithm. This is to demonstrate the superior performance of the proposed RF-CBC model and to justify the rationale behind the selection of the proposed RF-CBC model. The result of the experiment shows excellent performance of the designed system over the baseline method studied, with a precision rate and attack detection quality equal to or greater than 75% in most cases. The proposed attack detection and alarming system is capable of providing usable, consistent, efficient, faster, and accurate anomalous detection online and in real-time with a low false positive and negative rate.

Finally, the experimental results were thoroughly presented, and the proposed system will be implemented online and in real-time on the web server of the University of Medical Sciences, Ondo City, Nigeria.

2. Review of Related Work

This section examines many existing related works on intrusion detection systems in IoT facilities and the methods adopted, with the aim of investigating the performance of such algorithms used. This is done in order to arrive at more reliable and efficient ways of realizing the present system.

Several scholars in the field of attack detection on IoT facilities have carried out several studies on the topic, some of which show high potential in protecting the IoT facilities. However, despite the promising results, challenges still persist in securing the IoT facilities. Scalability problems due to the high connectivity of IoT networks, the novelty of attack techniques, and the interpretability of many of the existing attack detection models remain a concern [5]-[8].

Maheswari *et al.* [9], in their work, carried out research on a web attack detection system for IoT using ensemble classification. The results of their experiment show significant improvement in terms of accuracy when compared with the baseline methods used. However, their system was only able to detect some selected common types of attack, *i.e.*, their system can only detect SQL injection and cross-site scripting, hence neglecting other types of attack. The present system is designed to take care of various types of attacks as they relate to IoT facilities.

Yavuz, Unal, and Gul [10] proposed a deep learning-based machine learning

approach for the detection of routine attacks. The results of their experiment show high accuracy and precision on their data set. However, their systems are limited by the number of attacks that can be detected. A system that could be used to detect multiple attack types is needed, hence the need for this present work.

Learning Vector Quantization (LVQ) and K-NN were used by Naorum and Al-Sultani [11] for intrusion detection; their results record a good detection rate. The challenge with their approach is time complexity since the LVQ requires a long time to be trained, which requires the size of the classes to be equally likely, which is not the case with the present system. The present system is faster and more scalable in its operation.

In the work of Jawhar and Mehrotra [12], a hybrid intrusion detection system was proposed using fuzzy logic, neural network, and clustering algorithm with multi-layered perception to detect normal users and four attack types. The results of their experiment show a high detection rate of about 99.9%. However, their system is marred by the challenges of the distribution of the records in the training set not being close to equal between classes and the inability to detect multiple attack types.

Kouassi, Monsan, and Adou [13], in their work, explore the effectiveness of long-term memory neural networks (LSTMs) and Deep Neural Network (DNN) models for detecting attacks in IoT networks. The results of their experiment show that their models prove to be more effective for detecting attacks in IoT networks, particularly for sophisticated attacks.

Sasia, Lashkaria, Lua, Xiong, and Iqbal [14] carried out a survey on various IoT attacks by categorizing attacks in the taxonomy according to various factors such as attack domains, attack threat type, attack executions, etc. These were accompanied by their respective remedies. Their study revealed several open research areas pertinent to the subject of IoT attacks.

Siraparapu and Azad [15], in their work, carried out a comprehensive review of systems for securing IoT devices in the digital era. It explores the role of IoT secure systems in Industry 4.0, optimizing manufacturing processes and supply chain management. It emphasizes the significance of IoT secure systems, discussing challenges, limitations, and benefits to organizations. Their analysis reveals emerging trends in IoT security standards and identifies critical gaps in current regulations, offering a forward-looking perspective on ensuring integrity and privacy across diverse domains.

The present system is capable of overcoming the identified challenges of most of the existing systems studied, with the capability to handle multiple attack types due to its ability to monitor various clients to the website using their login credentials.

Evidence from the available literature shows that the major challenges faced by many existing machine learning algorithms for prediction or classification are scalability problems, inability to handle noisy data, computational complexity, and low-dimensional attributes; these usually result in classification and predictive errors. These occur when the number of features and instances is too large [16]-[19]. As

a result of this, many scholars in the field of machine learning have come up with a number of techniques for scaling down the number of attributes and data reduction [20].

Hegde *et al.* [21], in their work, perform feature selection using a symmetrized feature selection algorithm in combination with stacked generalization-based metaheuristic techniques on chronic disease datasets from the Kaggle repository. The result of their experiment shows better accuracy than the baseline methods used. However, this method is applicable to small subsets of problems and configurations. The present system is capable of handling large datasets.

Domingo and Hulten [22] used Hoeffding approaches to scale up machine learning algorithms. The method, which can be applied to choose among a set of discrete models or to estimate a continuous parameter, is capable of minimizing the time bound through the number of samples used, subject to the target limits on the loss of performance when using a subset of the data set. Their method was reported to produce very interesting results; however, the need to derive these bounds makes it difficult for many algorithms to use. The present system uses a single threshold bound, which makes it simple and easy to apply in many algorithms.

Flores *et al.* [23] trained the ARMA model using a statistical analysis technique to eliminate irrelevant features; their results show a good level of accuracy but with the risk of eliminating useful information. The application of the single threshold plot box method is capable of overcoming these challenges.

Sebban and Mock [24] carry out feature reduction using both filter and wrapper methods on a small subset of a dataset; this method performs well on their dataset. This method, however, is marred by the challenges of being computationally expensive with a high run time if used on a large dataset. These challenges are overcome by the present method of feature selection.

3. Methodology

This section presents a description of a series of processes and methods utilised to achieve the objectives of the proposed system.

3.1. Experimental Design

Our proposed attack detection and alarming system for IoT facilities adopts a four-module architecture, viz:

(a) The feature scaling and selection module: This module adopts the novel Single Threshold Boxplot Outlier Based Feature scaling method (STBO) to reduce the number of attributes in order to select the best and most useful feature for the proposed RF-CBC model at the pre-processing stage of the attack detection and reporting processes.

(b) The profiling and classification module: This module accepts useful URL information and builds a user profile online and in real time in order to determine their different access credentials using random forest techniques. These are stored in a data mart/database created, leading to the creation of different clusters of us-

ers based on the profile of their access credentials.

(c) The decision module: This module is designed to use the correlation statistic to analyze the user's browsing pattern alongside their access credentials and the attack information in order to arrive at a decision for detecting the anomaly. This is achieved by computing the correlation between the current user's URL information and the historical browsing information in the clusters stored in the data mart.

(d) The update module: The update module is designed to fine-tune the update of the anomaly detection system; here, all raw URL requests and normalized data together with the intrusion detection result are stored in the database for further analysis in a feedback mechanism. This is to improve the robustness and reliability of the RF-CBC model as well as to facilitate further analysis by security experts.

3.2. Data Collection

In this study, a click log history of 15,374 anonymous users of the University of Medical Sciences (UNIMED) official website, who signed into their accounts over a period of 12 months from 12 May, 2023 to 3rd April, 2024, was selected randomly. The selected click log is made up of about 30,572 anonymous visitors to the UNIMED website server's URL address database. Figure 1 shows a sample of extracted users' browsing history's HTTP from the UNIMED official website located at http://unimed.edu.ng.

	Hits		Referrer
1	679951	10.07%	- (Direct Request)
2	312946	4.64%	https://www.unimed.edu.ng/
3	61564	0.91%	https://www.google.com/search
4	39624	0.59%	https://www.unimed.edu.ng/portal/studentlogin.php
5	38066	0.56%	https://www.unimed.edu.ng/portal/studentInformation.php
6	36974	0.55%	https://www.unimed.edu.ng/departments.php
7	34255	0.51%	https://www.unimed.edu.ng/faculty.php
8	29303	0.43%	https://www.unimed.edu.ng/portal/paymentupnew.php
9	23552	0.35%	https://www.unimed.edu.ng/portal/formcss2.css
10	20358	0.30%	https://www.unimed.edu.ng/staff/profile.php
11	17480	0.26%	https://www.google.com/
12	16993	0.25%	https://www.unimed.edu.ng/unimedNews.php
13	14571	0.22%	https://unimed.edu.ng/
14	14011	0.21%	https://www.unimed.edu.ng/services_page.php
15	12922	0.19%	https://www.unimed.edu.ng/portal/printForm.php
16	10654	0.16%	https://www.unimed.edu.ng/styles1.css
17	9689	0.14%	https://www.unimed.edu.ng/portal/reprint-invoice.php
18	8885	0.13%	https://www.unimed.edu.ng/postutme/howtoapply.php
19	8579	0.13%	https://www.unimed.edu.ng/portal/courseregistration.php
20	8308	0.12%	https://www.unimed.edu.ng/portal/resultChecker.php
21	7787	0.12%	https://www.unimed.edu.ng/index.php
22	7463	0.11%	https://www.unimed.edu.ng/staff/students_registration.php
23	7095	0.11%	https://www.unimed.edu.ng/themes/1/js-image-slider.css
24	7061	0.10%	https://www.unimed.edu.ng/portal/
25	6631	0.10%	https://www.unimed.edu.ng/postutme/programmes.php
26	6381	0.09%	https://www.unimed.edu.ng/tabcontent.css
27	5442	0.08%	https://www.unimed.edu.ng/portal/paymentupverify.php
28	5145	0.08%	https://www.unimed.edu.ng/postutme/postutmelogin.php
29	4837	0.07%	https://www.unimed.edu.ng/updp/updplogin.php
30	4652	0.07%	https://www.unimed.edu.ng/postutme/
31	4254	0.06%	https://unimed.edu.ng/departments.php
32	4130	0.06%	https://www.unimed.edu.ng/portal/payment_pgd_others.php

Figure 1. Sample historical HTTP request to the UNIMED official website.

We carried out a number of pre-processing operations on the raw user's URL address database extracted by cleansing it, in order to eliminate irrelevant or noisy entries. After this, we developed a data mart of the log data and partitioned it into sessions [25]. After the session identification stage, we then applied the proposed Single Threshold Boxplot outlier-based feature scaling and feature selection algorithm to select the best feature for the proposed RF-CBC model. The proposed Random Forest algorithm was then used to group users into clusters based on similarities in their access.

Credentials, browsing HTTP logs, and attack information, while the correlation-based Clusters algorithm was used for the final decision as to whether a user is an attacker or a normal user based on their access credentials and similarities in their search behaviour to that of the identified clusters and the attack information, the result of which is used to update the data mart for further analysis. The overall architecture of the entire RF-CBC attack detection and reporting system is shown in **Figure 2**.

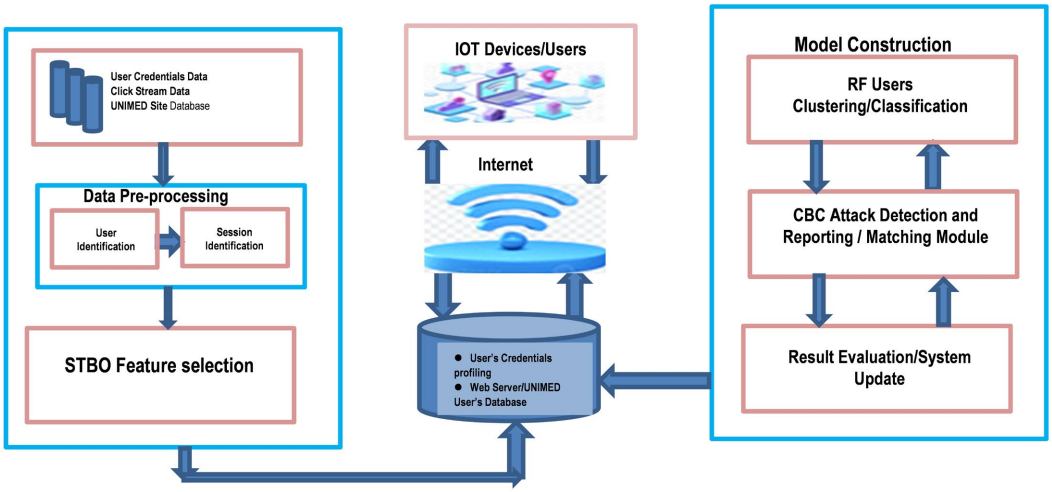

Figure 2. The overall architecture of the entire RF-CBC attack detection and reporting system.

3.3. Data Pre-Processing Feature Selection

Some of the problems with many machine learning algorithms are scalability and computational complexity. Noisy data and low-dimensional attributes, caused by too many features and instances, usually result in prediction or classification errors. To overcome these challenges, this work proposes a novel feature selection technique called the Single Threshold Boxplot Outlier Based Feature Scaling method (STBO) to scale down the number of attributes and select the best features for the proposed attack detection processes [17] [23] [26]. In recent times, researchers in machine learning have proposed a number of feature selection techniques, such as Gini-index, χ^2 statistics, information gain, Wrapper method, and correlation technique, etc., [17] [23] [26]. Some of these performed well on their data sets. Some of these techniques have been studied, and the pros and cons of each method have been thoroughly understood before proposing the current STBO technique.

3.3.1. Single Threshold Boxplot Outlier Based Feature Scaling Method (STBO)

In this work, the click log history of anonymous visitors to the UNIMED official website was selected as described in Section 3.2. The terms that occur in the documents are presented as parameters, features, attributes, tuples, or variables. After session identification, the term count was taken. Table 1 shows the term document matrix over all documents with the total term count. Figure 3 shows the statistics of the occurrence of each attribute in the different documents extracted.

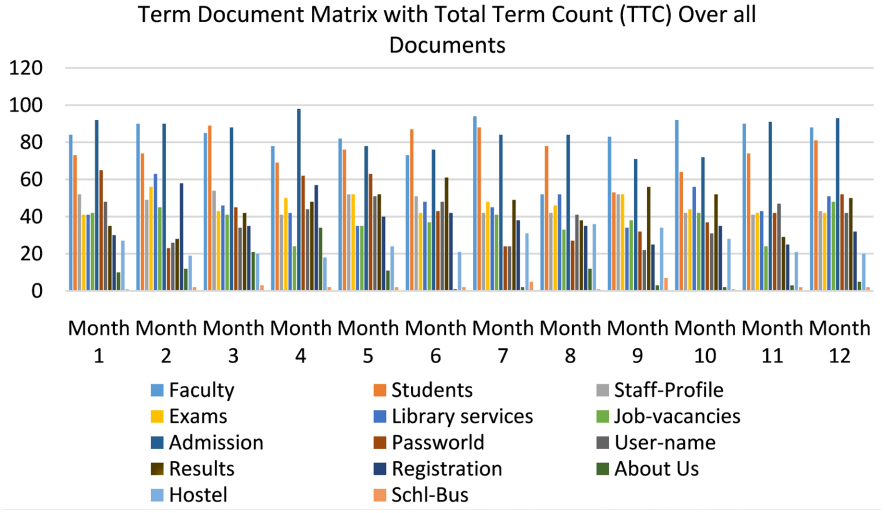

Figure 3. The statistics of the occurrence of each attributes in the different documents extracted.

Table 1. Attributes documents matrix for 12 months with Total Term Count (TTC) over all documents using UNIMED users' logs database.

TD-X	*Faculty*	*Students*	*Staff-Profile*	*Exams*	*Library services*	*Job-vacancies*	*Admission*	*Passworld*	*User-name*	*Results*	*Registration*	*About Us*	*Hostel*	*Schl-Bus*
Month 1	84	73	52	41	41	42	92	65	48	35	30	10	27	1
Month 2	90	74	49	56	63	45	90	23	26	28	58	12	19	2
Month 3	85	89	54	43	46	41	88	45	34	42	35	21	20	3
Month 4	78	69	41	50	42	24	98	62	44	48	57	34	18	2
Month 5	82	76	52	52	35	35	78	63	51	52	40	11	24	2
Month 6	73	87	51	42	48	37	76	43	48	61	42	1	21	2
Month 7	94	88	42	48	45	41	84	24	24	49	38	2	31	5
Month 8	52	78	42	46	52	33	84	27	41	38	35	12	36	1
Month 9	83	53	52	52	34	38	71	32	22	56	25	3	34	7
Month 10	92	64	42	44	56	42	72	37	31	52	35	2	28	1
Month 11	90	74	41	42	43	24	91	42	47	29	25	3	21	2
Month 12	88	81	43	42	51	48	93	52	42	50	32	5	20	2
TTC	991	906	561	558	556	450	1017	515	458	540	452	116	299	30

In this work, we propose the Single Threshold Boxplot Outlier Based Feature scaling method of eliminating features that diverge significantly from the general pattern of the users' click logs using the Boxplot, also called box-and-whisker plots. The Boxplot is a statistical method that uses five (5) summary statistics *i.e.* Minimum, Maximum, First Quartile (Q_1), Second Quartile (Q_2), and Third Quartile (Q_3). Each quartile represents 25% of the data points [27].

We first arranged the data elements, *i.e.*, The features' total term count for each feature/tuple for all the documents was extracted in ascending order. We then calculate the Q_1, Q_2 and Q_3; we later calculate the interquartile range (IQR), *i.e.*, IQR = (Q_3 − Q_1). We also calculate the lower and upper bounds for the outlier, which will be included in the non-outlier zone.

Lower Bound (LB) = Q_1 − 1.5 * IQR

Upper Bound (UB) = Q_3 + 1.5 * IQR. [24]

However, since our interest is to eliminate only irrelevant features, we eliminate only the data points that fall below the lower bound and consider them as outliers; hence, we consider only a single threshold, rather than the usual two thresholds.

The attributes within the lower bound are to be used by the proposed RF-CBC algorithm.

Algorithm listing for the Single Threshold Boxplot Outlier Based Feature Scaling method (STBO) is shown in **Algorithm Listing 1**.

3.3.2. Algorithm Listing

Algorithm Listing 1. The single threshold boxplot outlier based feature scaling algorithm (STBO).

```
Input: ATM (Attributes Document Matrix of user log-in to UNIMED website
Output: List of attributes above the lower bound
1.  N= Number of document extracted (D1,D2,D3, ……, Dn)
2.  K= Number of attributes, at1. at2, at3, …….atk
3.  nij = number of times a given attribute ati occur in document D2
4.  Begin
        Input n of unknown document
        Input K of unknown Document
        i=j=1
            { for each sample document (n) do;
                For j=1 to n
                    For i= 1 to k
                    If (nj=0) then
                        Rf(ati: Dj) =0
                    Else
                        Sum i,j= sum i,j
                    End if

                End for i; j=j+i; end for j}
        end do.
// Section to eliminate the outlier attributes
Attributes sum list(ati,dj)
Sort (attributes sum list(ati,dj))
Get the minimum value of attributes sum, Get the maximum value of attributes sum
Q2 ie. Median count = ½(N+1)th
Finge count= ½ (1+median count), the result of which must be an integer
Q1= count from the beginning of the sorted attributes sum list, the number derived from the finge count
Q3= count from the end of the sorted attributes sum list, the number derived from the finge count
 IQR=Q3-Q1
Compute the Lower Bound (LB) = Q1-1.5 * IQR
Compute the Upper Bound (UB) = Q3+ 1.5 * IQR
Eliminate any attributes that fall below the lower bound and treat them as outliers
End.
```

3.3.3. Application of the Single Threshold Boxplot Outlier Based Feature Scaling Method (STBO)

This section presents the demonstration of the present STBO feature selection techniques using our collected data sets from the UNIMED official website, as shown in Table 1. We applied the STBO feature scaling algorithm to eliminate irrelevant attributes and retain only relevant attributes that fall above the lower bound for our attack detection system. We compute the total term count for all the selected attributes and then sort them in descending order of their TTC, as shown in Table 2. We compute Q_1, Q_2, Q_3, the IQR, the lower and the upper bound. We then minimize the number of attributes by selecting attributes that fall above the lower bound from the data set extracted. This is done in order to increase the accuracy and efficiency of our RF-CBC machine learning algorithm. Given our data set according to Table 1 and the sorted attributes according to their TTC, as shown in Table 2.

Table 2. The sorted attributes according to their TTC.

TD-X	*School-Bus*	*About Us*	*Hostel*	*Job-vacancies*	*Registration*	*User-name*	*Pass-world*	*Results*	*Library services*	*Exams*	*Staff-Profile*	*Students*	*Faculty*	*Admission*
TTC	30	116	299	450	452	458	515	540	556	558	561	906	991	1017

The Median count $Q_2 = 1/2(N+1)^{th}$

$=1/2(14+1)^{th} = (15/2)^{th} = 7.5^{th}$, we use 7^{th} and the 8^{th} element, *i.e.*

$=(515+540)/2 = 1055/2$

$=527.5$, therefore $Q_2 = 527.5$

Finge count = 1/2(1 + median count), the result of which must be an integer = $1/2(1+7.5)^{th} = 1/2(8.5)^{th} = 4.25$, fringe must be an integer therefore = 4.

Q_1 = count from the beginning of the sorted attributes sum list, the number derived from the finge count, *i.e.*, 4.

i.e. $Q_1 = 450$

Q_3 = count from the end of the sorted attributes sum list, the number derived from the finge count, *i.e.*, 4.

i.e. $Q_3 = 561$

$$\text{IQR} = Q_3 - Q_1 = 561 - 450 = 111$$

$$\begin{aligned}\text{Compute the Lower Bound}(\text{LB}) &= Q_1 - 1.5 * \text{IQR} \\ &= 450 - 1.5 * 111 \\ &= 450 - 166.5 \\ &= 283.5\end{aligned}$$

Any attributes TCT that fall below the lower bound of 283 are treated as outliers and eliminated.

3.3.4. Evaluation of the (STBO) Feature Selection Technique

This section presents the evaluation of our proposed STBO feature selection technique in two different ways. First, we carried out feature selection on the extracted historical browsing pattern data from the University of Medical Sciences (UNIMED) website, using two baseline techniques which include Information Gain (IG), CFS, and the proposed STBO. Table 3 shows the number of features selected by each technique. The experimental results show that the IG and the present STBO selected fewer features, with STBO selecting fewer features in comparison with the baseline methods studied.

Table 3. The number of features selected by CFS, IG and STBO feature selection techniques.

Total number of available attributes	Number of Attributes selected by each technique		
	CFS	IG	STBO
20	8	8	5
25	14	12	7
30	18	16	8
60	25	19	10
80	28	21	12
90	32	23	15
112	43	28	18

To further evaluate the accuracy and effectiveness of the present method, we tested the degree of accuracy of our RF-CBC algorithm with the STBO, IG, and CSF feature selection algorithms. The accuracy of the RF-CBC was determined using the expression:

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

where:

AC = Accuracy: This measures the proportion of correctly detected attacks and normal instances.

TP = True Positive: This is the number of correctly detected attack instances,

TN = True Negative: This is the number of correctly detected normal instances,

FP = False Positive: This is the number of incorrectly detected attack instances, and

FN = False Negative: This is the number of incorrectly detected normal instances.

The result of the experiment, as shown in Figure 4, indicates that the proposed RF-CBC performed well, with the STBO technique significantly improving the accuracy of the RF-CBC algorithms in comparison with the IG and the CFS techniques at different levels of predictions and under the same experimental settings.

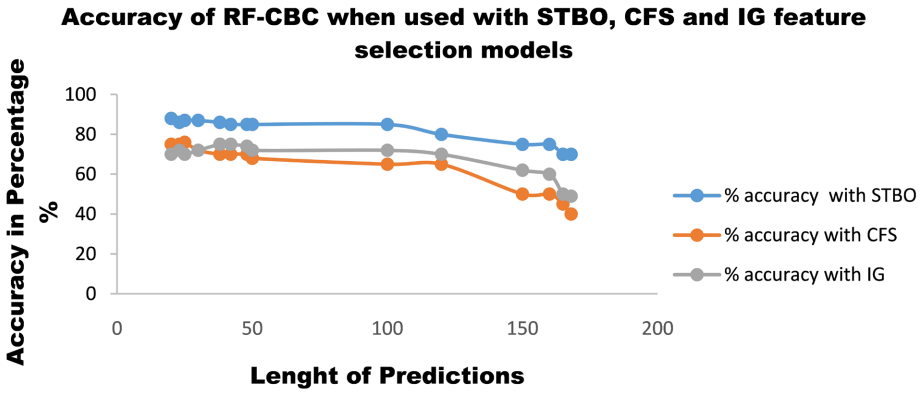

Figure 4. The RF-CBC accuracy with the STBO, CFS and the IG feature selection methods at different level of prediction.

3.4. The Proposed Random Forest Enabled-Correlation Based Clustering (RF-CBC) Technique

Random forest is an ensemble decision tree in which each tree depends on a collection of random variables. It is a supervised machine learning algorithm widely used in classification and regression problems. Random forest builds decision trees using different samples, then takes the majority votes of the trees for classification or averaging in the case of regression. Leo Breiman was believed to have first come up with the idea of random forest [28].

Given an n-dimensional random vector $X=\left(X_1,X_2,X_3,\cdots,X_n\right)^{\mathrm{T}}$ where $X_1,X_2,X_3,\cdots,X_n$ represent the real valued input with Y representing the predictive class (voting) [29].

Given an unknown Joint distribution $P_{xy}\left(X,Y\right)$

$F\left(x\right)$ is a function to predict Y, which is determined by a loss function $L\left(Y,F\left(X\right)\right)$, the loss function is defined to minimize the expected value of loss, which is denoted by Equation (1). [29].

$$E_{xy}\left(L\left(Y,f\left(x\right)\right)\right) \tag{1}$$

The subscript here denotes expectation with respect to joint distribution of X and Y

$L\left(Y,f\left(x\right)\right)$ measures how close $f\left(x\right)$ is to Y, we choose zero-one-loss for classification, we have

$$L\left(Y,f\left(x\right)\right)=L\left(Y\neq f\left(x\right)\right)=\begin{cases}0, & \text{if } y=f\left(x\right)\\ 1, & \text{otherwise}\end{cases} \tag{2}$$

If the set of possibilities value of Y is denoted by $\mathcal{Y}$, to maximise $E_{xy}\left(L\left(Y,f\left(x\right)\right)\right)$ for zero-one-loss we have

$$f\left(x\right)=\frac{\arg\max n\left(Y=y\middle|X=x\right)}{Y\in\mathcal{Y}} \tag{3}$$

This is also known as Bayes' rule.

In classification $f\left(x\right)$ represent the most frequently predicted class popularly known as ("votting") where $h_i\left(x\right),\cdots,h_j\left(x\right)$ is the "base learning" combination

that gives the "ensemble" prediction $f(x)$ [29].

$$f(x) = \frac{\arg\max}{Y \in \mathcal{Y}} \sum_{j=1}^{j} I\left(y = h_j^{(x)}\right) \tag{4}$$

In Random forest, the J_{th} based learner denotes a tree represented by $h_j(x, \theta_j)$ where θ_j denotes collection of random variables and θ_j's are independent where $j = 1, \cdots, j$. The random forest algorithm is used to classify each client into their different users' categories based on their access credentials. Based on majority votes, these are stored in a data mart/database created, leading to the creation of different clusters of users based on the profile of their access credentials.

3.4.1. The Data Clustering

In the early days of data mining/machine learning, humans adopted manual labeling of data. However, with the increase in the volume of data available these days, manual labeling of data has become difficult, tedious, and expensive. Hence, there is a need for automatic labeling. Clustering can be described as a method of grouping a set of data objects into different classes of similar objects [30]. Available literature shows that there are a number of clustering methods for machine learning, which include: K-modes, K-means, K-median, genetic K-means, intelligent K-means, etc. [31]

This present work adopts the K-modes clustering techniques. The K-Modes is a non-parametric algorithm capable of handling categorical data and optimizing a matching metric. The loss function (L_o) is used without explicitly applying any distance metric. Here, similar trees are categorized into the same cluster based on majority voting. We select K-initial modes, then form k clusters by assigning all the data points to the cluster with the nearest mode (vote) using the matching metrics. We later recompute the modes of the clusters until the convergence criteria are met.

3.4.2. The Correlation Statistics Distance Measurement Technique

The correlation statistic technique is used to analyze the users' browsing patterns alongside their access credentials and the attack information to arrive at a decision for detecting the anomaly. This is achieved by computing the correlation between the client's URL information and the historical browsing information in the clusters stored in the data mart.

A review of different distance measurement techniques shows that a good number of the existing techniques are marred with various challenges ranging from inaccuracy, susceptibility to noisy data, computational complexity, to scalability challenges, etc.

The present correlation Distance measurement technique is capable of overcoming these challenges while providing scalable, efficient, accurate, and simple distance measurement for any machine learning algorithm.

The correlation distance measurement technique is used to measure the relation and association between phenomena. The correlation relationship between two tuples X and Y can be expressed using Equation (5) below [29].

$$r(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (5)$$

where $\bar{x}$ is the mean of Tuple, and $\bar{y}$ is the mean of tuple y.

X = Training tuple

Y = test Tuple

Correlation can have a value of 1, 0 or −1

If the correlation is 1, then this implies that there is a perfect positive correlation between variable X and Y

If the correlation is 0, that implies no correlation (the value of X and Y, don't seem linked at all)

If the correlation is −1, that implies there is a perfect negative correlation [32].

3.4.3. Algorithm Listing

The random forest enabled Correlation based clustering (RF-CBC) algorithm listing is shown in **Algorithm Listing 2**.

Algorithm Listing 2. The random forest enabled Correlation based clustering (RF-CBC) algorithm.

Input: Data D={ } where I – 1 to n
Output: Class Label{Yes-> attacker / No ->legitimate User
Let D= {(X_1,Y_1) ……X_n,Y_n)} represents the training data with x_i = (x_{i1}, ……. X_{ip})^T, for j=1 to J
// Section to predict users categories using Random forest technique based on Users credentials and browsing URL information

1. Take a sample D, of size n from D
2. Use the sample D_1, as the training data
3. Fit a tree using binary recursive petitioning
   a. Begin with all observation in a single node
   b. Repeat
      i. Take M prediction at random from P available prediction
      ii. Find the best binary split among all binary split on the M prediction from step i
      iii. Split the node into two descendant nodes using the split from step ii
   c. Until the stopping criterion is met
4. For prediction at a new point x use the expression in the equation 4

$f(x)=\frac{argMax}{|Y\epsilon Y}\sum_{j=1}^{j}I(y=h_j^{(x)})$ ……where, $h_j^{(x)}$ is the response variable prediction at x using j^{th} tree.

// section to classify the user/variables into clusters using K- mode clustering techniques.

5. Select the K initial nodes
6. Repeat
   a. Create K cluster by assigning all the data point to the cluster of the nearest modes using the matching metric
   b. Re-compute the modes of the cluster
7. Until convergence criterion is met

// Section to detect and report attack using correlation distance measurement techniqu

8. Let Y be an input user`s credential and browsing URL` information

   Y= { (c_1,c_2-----c_n), (q_1, q_2-----q_n)},
9. Function
10. Input: training data set x:{ (c_{i1}, c_{i2},-------c_{in}), (q_{i1}, q_{i2},-------q_{in}) }, I =1,2,3------K
11. Let I =1
12. Repeat until (perfect correlation)
13. Retrieve available users credential and browsing case history Yi from the user`s database
14. Compute the correlation between the given new user`s case using the expression

$$r(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})((y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

15. If the correlation between x_i and Y_i =1 (i.e if r =1) then
16. Let X_i be a member of Y_i
17. Go to 24
18. else
19. If the correlation between x_i and y_i = -1 then
20. Let xi be an attacker (treat user xi as an attacker)
21. Report user X_i as an attacker
22. Else
23. Increment I by 1
24. Section to adapt the test case to fit in the new case to the correlated cluster
25. Evaluate the result
26. If (result fits well) then
27. Retain the new user browsing history and credentials
28. Otherwise (discard the new case)
29. End if; end if
30. End function

3.4.4. Application of the Proposed RF-CBC Technique to Detect and Report Attack and Abnormality

This section demonstrates the application of the present RF-CBR technique to detect and report abnormality on the experimental website, *i.e.*, the UNIMED website, using our extracted user browsing history data set of the university website.

Considering our experimental website *i.e*, the UNIMED website users' credentials and click stream data are considered as a vector with attributes: user's category, access type, allowed operation, (Click $\log_1$, Click $\log_2$, Click $\log_3$ … Click $\log_n$), with clients/users represented by X_1, X_2, X_3, X_4, ⋯, X_n as class labels, as presented in Table 4.

Given an unknown user's X_5, determine the class of user X_5. The random forest classifier will first be used to classify all the users into clusters based on similarities in their users' credentials and their click logs. We then compute the correlation distance between users X_5 and all other vectors in each of the clusters by applying Equation (5), *i.e.*

$$r(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Table 4. The UNIMED website data mart class label training tuples and user's credentials click logs.

Users	*Access type*	*Operation type*	*Log*$_1$	*Log*$_2$	*Log*$_3$	*Credentials type/ Class Label*
X_1	Full access	Unrestricted	Index	Staff profile	Exams and Record	Administrator
X_2	Privileged	Staff privileged	Index	Staff profile	Add courses	HOD
X_3	Restricted	Basic operation	Admission	faculty	Programmes	Visitor
X_4	Limited	Student operation	Registration	Courses	Fees payment	Student
X_6	Privileged	Staff privileged	Index	Staff profile	Approved courses	Tutor
⋮						
X_5	Restricted	Basic operation	Staff profile	Exams and records	Upload Credentials	?

The data from our experimental website shown in Table 4 are categorical in nature; this means they are non-numeric attributes. To be able to use them for numerical calculation, we adopt the nominal scale technique [30] [33]. The nominal scale involves assigning numbers to a category of data, such that no category is greater than or less than the other categories; it involves labeling data in a particular group according to the relevant attribute possessed, in no special or specific order or magnitude. It is strictly for identification purposes.

Therefore, for a set of attributes, the users' category, which consists of (Admin, HOD, Tutor, Student, Visitor), will be coded as 1 for Admin, 2 for HOD, 3 for Tutor, 4 for Student, and 5 for Visitor. Likewise, for the set of attributes, access type, which consists of (Full access, privileged, limited, and restricted), will also be coded as 1 for full access, 2 for privileged, 3 for limited, and 4 for restricted. For the set of attributes, operation type (unrestricted, staff privilege, basic operation, student privilege) will be coded as 1 for unrestricted, 2 for privilege, 3 for student

privilege, and 4 for basic operation. For the attribute click log, which is made up of possible user clicks on the UNIMED website and includes: index (1), Staff portal (2), Admission (3), Student registration (4), Library (5), Exams portal (6), Faculty (7), Programmes (8), Add Courses (9), pay portal (10), Upload result (11), Upload credentials (12), etc., these are also labeled numerically in ascending order, such as 1, 2, 3, 4, …, n, where n represents the total number of possible clicks on the UNIMED website. The number in parentheses after each data tuple represents the nominal scale assigned to the tuples.

The complete nominal scale representation of the given training tuple, as presented in Table 4, is shown in Table 5.

Table 5. The UNIMED website data mart class label training tuples and user's credentials click logs.

Users	*Access type*	*Operation type*	*Log*$_1$	*Log*$_2$	*Log*$_3$	*Credentials type/ Class Label*	*Status*
X_1	1	1	1	2	6	Administrator	Normal
X_2	2	2	1	2	9	HOD	Normal
X_3	4	4	3	7	8	Visitor	Normal
X_4	3	3	4	9	10	Student	Normal
X_6	2	2	1		9	Tutor	Normal
X_7	4	4	1	6	9	Visitor	Attacker
⋮							
X_5	4	4	2	6	12	?	?

The correlation between user X_1 and user x_5 can now be computed as follows:

$$r(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$X_1=(1,1,1,2,6),\quad X_5=(4,4,2,6,12)$$

The correlation between user X_5 and user X_1 is −1; this indicates that there is no relationship between user X_1 and user X_5. The process is repeated between the unknown user X_5 and every other user. If the correlation is 1, then we categorize user X_1 into the cluster of user Y_1 or any other matching cluster. However, if the correlation matches that of the attacker's cluster or if the correlation is −1 throughout, then user X_1 is suspected to be an attacker and is therefore reported as an attacker.

4. System Evaluation and Result Analysis

We evaluate the proposed RF-CBC attack detection and reporting system using the University of Medical Sciences (UNIMED) website users' browsing history data set. Being a real-world data set, the data set was pre-processed and relevant features were extracted using the proposed STBO feature scaling technique before

applying the proposed RF-CBC to detect any abnormality on the website. To this effect, in-house software was developed using Python, with XAMP/Apache HTTP as the hosting server and MYSQL DBMS for data mart creation.

4.1. System Evaluation

As part of our experiment, we carried out a performance evaluation of our system through performance comparison of the present system with some baseline methods, which are the traditional random forest, the traditional clustering method, the Naïve Bayes, and the Artificial Neural Network techniques on the same datasets and environment. An experimental online attack detection and reporting system was developed to implement the RF-CBC model. In the developed attack detection system, the users enter their basic information, which is used to build their profile and credentials online and in real time. The attack detection report is triggered when the system observes abnormalities in the users' browsing patterns. The source code for the developed application is available on request. The proposed system can be implemented online by uploading the present application to a web server, which can be invoked anytime online on any web browser.

4.2. Data Set for Evaluation

In order to evaluate our RF-CBC model, we used the extracted historical browsing history of the UNIMED website dataset. The click log history of 15,374 anonymous users of the UNIMED official website, who signed into their accounts over a period of 12 months from 12th May, 2023 to 30th April, 2024, was randomly selected, which is made up of about 30,572 sessions/SQL queries. The selected records were divided into five parts, four of which were used as training sets and the remaining part was used as the testing set. The class/clusters of the training part were considered known while those of the testing part were considered unknown; the training part is used to infer the unknown part.

The software developed was used to run both the proposed RF-CBC and the baseline methods, which include: the Traditional Random Forest (TRF), the Naïve Bayesian (NBY), the Artificial Neural Network (ANN), and the Traditional Clustering (TCL) methods, for about sixty times. The results are used as a dataset for evaluation purposes. We calculate the Accuracy (ACC), True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Precision (P), and Recall (R).

Table 6. The confusion matrix for our attack detection system.

	Relevant	Relevant
Predicted	TP	FP
Not relevant	FN	TN

where:

AC = Accuracy: This measures the proportion of correctly detected attacks and

normal instances.

TP = True Positive: This is the number of correctly detected attack instances,

TN = True Negative: This is the number of correctly detected normal instances,

FP = False Positive: This is the number of incorrectly detected attack instances, and

FN = False Negative: This is the number of incorrectly detected normal instances.

We used the F1-Measure technique as presented in a confusion matrix shown in Table 6.

As presented in the confusion matrix shown in Table 6, to evaluate the detection quality of the RF-CBC model, where F-Measure is the harmonic mean of precision and recall, *i.e.*

$$F = 2 \cdot \frac{\text{Precision} \cdot \text{recall}}{\text{Precision} + \text{recall}}$$

This is referred to as the F1- measure, since recall and precision are weighted.

Precision is the number of correct predictions divided by the number of all returned predictions, *i.e.*

$$\text{Precision}(P) = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall is the number of correct predictions divided by the number of all known interest supposed to be discovered, *i.e.*,

$$\text{Recall}(R) = \frac{\text{TP}}{\text{TP} + \text{FN}}$$ [17] [30].

Table 7 shows our experimental results on the UNIMED browsing history database with ACC, TP, TN, FP, TP, Precision, and Recall. Figure 5 shows the result of our conducted experiment in F1-Measure using our browsing history click logs database.

Table 7. Experimental Results from our attack detection system

Model	TP	TN	FP	FN	*P*	*R*	F1-measure
RF-CBC	8	2	1	0	0.888888889	1	0.941176471
TRF	6	2	2	1	0.75	0.857142857	0.8
ANN	6	3	1	1	0.857142857	0.857142857	0.857142857
NBY	6	2	1	2	0.857142857	0.75	0.8
TCL	5	3	2	1	0.714285714	0.833333333	0.769230769

4.3. Presentation of Results and Discussion

We recorded the results for both the TP, FP, TN, TF. We computed the Precision, the Recall, and the F1-Measure and compared the precision rate as shown in Table

7. Table 6 shows the number of documents failing in each category using the confusion matrix. Our experimental result is shown in Figure 5. The result shows the excellent performance of the present RF-CBC model over the baseline methods studied. We established that the different algorithms studied generally have a peak point, though with significant differences in performance. The F1-measure value initially increases for each algorithm studied before their respective peak points, then gradually goes down after the peak point, meaning that the precision is nearly stable and recall increases before the peak, after which the precision decreases and recall is almost stable. Figure 5 shows the F1-Measure of the proposed RF-CBC, the TRF, NBY, ANN, and the TCL algorithms at different lengths of prediction/detection using the UNIMED official website users' click log dataset. The result shows the superiority of the present RF-CBC model over the baseline methods studied. The experimental result shows that the TRF, NBY, ANN, and TCL techniques recorded lower F1-Measure when run on our dataset, as shown in Figure 5; they performed poorly compared with the RF-CBC algorithm. The RF-CBC algorithm has the highest F1-Measure when used on our dataset. The experiment was carried out for over 60 different lengths of prediction using the same datasets and experimental settings. The Naïve Bayesian, the traditional Random Forest, and the traditional clustering algorithms perform a little better at short lengths of prediction below 15, but poorly at longer and worst at lengths of prediction longer than 25; this therefore results in a limited number of positive detections. However, the ANN performs a bit better than the TRF, NBY, and TCL techniques, but also only at shorter lengths and poorly at longer lengths of prediction and worst at lengths above 25, thereby resulting in a few positive detections. The proposed RF-CBC has over 80 F1-Measures for lengths of prediction between 1 and 20, after which it maintains about 75% F1-Measure at longer lengths of prediction as shown in Figure 5. The experimental result shows that the RF-CBC demonstrates higher potential in detecting and reporting abnormalities on our experimental website; hence, it remains the clear winner in this case.

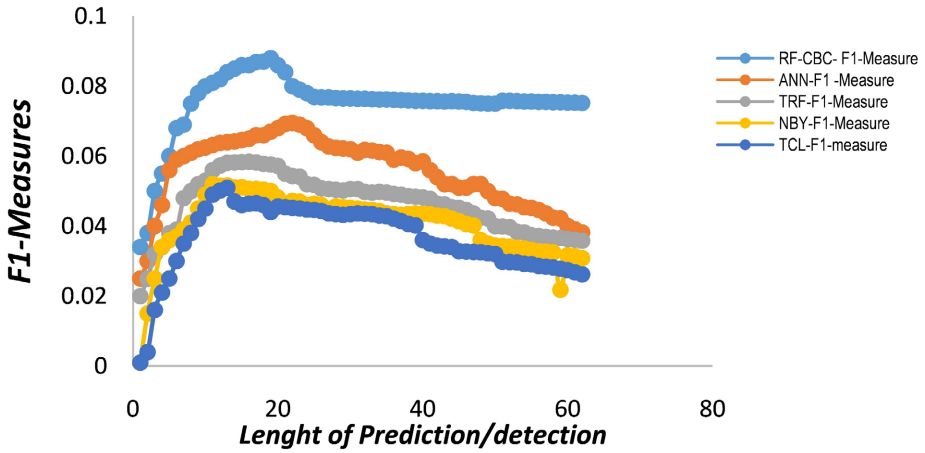

Figure 5. F1-Measures of RF-CBC, TRF, NBY, TCL and ANN at different length of prediction/detections.

Furthermore, we authenticate the potential of the present RF-CBC model over the baseline method by comparing their respective execution speeds in seconds. We recorded the run time of each algorithm at different prediction lengths using our experimental datasets and under the same experimental settings. The experimental results indicate that the RF-CBC has the lowest run time and executes faster than the baseline methods studied in all cases.

Though the NBY and the ANN also show low runtime compared to the TRF and the TCL technique, generally, the baseline methods' runtime increases rapidly as the prediction length increases, as shown in Figure 6. The proposed RF-CBC recorded a lower runtime at all lengths; therefore, it achieves better results with large datasets and longer prediction lengths. The outstanding performance of our system may be due to the adoption of multiple techniques for the detection and reporting of intruders on the experimental website, such as the introduction of the STBO feature selection technique to select the best features that eliminate noisy data, user clustering based on their different credentials and types of allowed operation by the different categories of users on the website, the use of correlation distance measurement technique, etc. All these factors have made the RF-CBC the most appropriate method for this study.

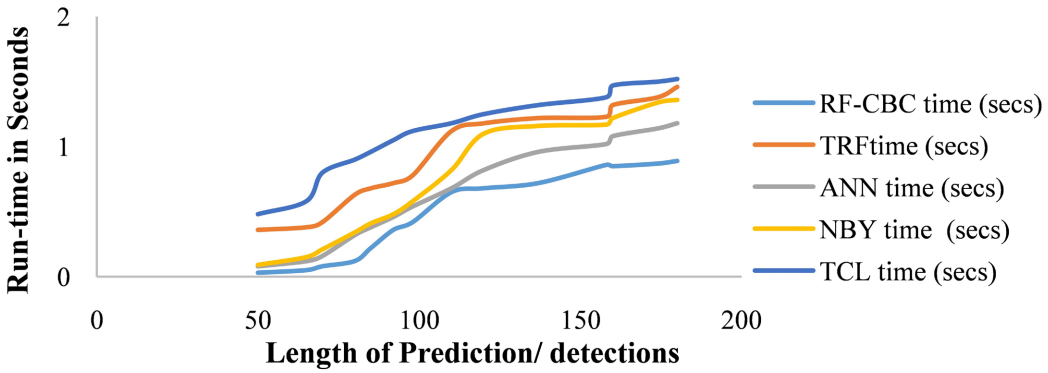

Figure 6. Run time of RF-CBC, ANN, TRF and TCL model at different length prediction/detection.

Finally, the RF-CBC model shows potential for detecting and reporting many forms of attack on any IoT device. The model is specifically suitable for domains with large volumes of data and any number of relevant attributes. The RF-CBC also shows the capability to overcome some of the challenges of many existing attack detection systems, such as noisy data, scalability, poor distance measurement functions, and computational complexity. Therefore, it establishes a flexible, scalable, transparent, accurate, faster, computationally efficient, easy to understand, and easy to implement method of attack detection and reporting systems. Our experimental results show that the RF-CBC can outperform the traditional Random Forest, Naïve Bayesian, traditional Clustering, and the Artificial Neural Network techniques in a very difficult attack detection task and in a very large dataset online and in real time consistently.

5. Summary, Conclusion, and Recommendation

The aims of this work are to develop an efficient, faster, scalable, robust, flexible, accurate, consistent, and easy-to-use attack detection and alarming system. This is achieved through the design and implementation of a novel attack detection algorithm, commonly referred to as the Random Forest enabled correlation-based clustering (RF-CBC) algorithm. To this effect, a click log history dataset of anonymous users of the UNIMED official website, who signed into their accounts over a period of twelve months, was randomly selected, and the extracted data was preprocessed. The data was cleansed to eliminate noisy or irrelevant entries, sessions were identified, and a data mart was created. A feature scaling technique called the Single Threshold Box Plot Outlier (STBO) algorithm was developed and used to select the best features before applying the proposed RF-CBC model for attack detection and reporting purposes.

The results of the experiment conducted were presented and analysed. We also carried out a performance comparison of the developed system and four other baseline methods, which are the TRF, NBY, TCL, and ANN algorithms, to demonstrate the superiority of our system over the baseline methods studied. The results of this comparison show that the present system outperformed the baseline techniques in terms of speed and accuracy. This is aimed at assisting web developers and administrators to have a wider variety of algorithms from which the best performing can be selected, and to plan updates and improvements to the security of their websites and IOT facilities through the adoption of the present system, while also assisting IOT facilities users to be more secure and protected while conducting their legitimate business online.

In conclusion, this work provides a basis for IoT facilities attack detection and an alarming system. The system collects the active users' personal information and click stream information, builds a personalized profile and access credentials that will be used to monitor and determine the user's activities and the types of operations he or she can perform on the site. The collected click stream and profile information are matched with similar clusters in the data mart in order to determine the user's credentials, upon which the system determines whether the user is an attacker or a normal user. The results of our experiment show that our attack detection and reporting engine, powered by the RF-CBC algorithm, is capable of producing an efficient, faster, accurate, and scalable attack detection and reporting online and in real time consistently at any time while overcoming some of the challenges of the existing method studied.

We are of the opinion that the current work can be improved further and therefore recommend that other scholars in the field explore alternative machine learning techniques and compare the results with the present work, to determine the most effective ways of solving similar problems in the future. Researchers could also further investigate the UNIMED official website users' URL TP address database or similar websites on a continuous basis, so as to be abreast of new methods of attack in the near future.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Luo, C., Tan, Z., Min, G., Gan, J., Shi, W. and Tian, Z. (2021) A Novel Web Attack Detection System for Internet of Things via Ensemble Classification. *IEEE Transactions on Industrial Informatics*, **17**, 5810-5818. https://doi.org/10.1109/tii.2020.3038761

[2] Andročec, D. and Vrček, N. (2018) Machine Learning for the Internet of Things Security: A Systematic Review. *Proceedings of the* 13*th International Conference on Software Technologies*, Porto, 563-570. https://doi.org/10.5220/0006841205630570

[3] Priya, V., Sumaiya Thaseen, I., Reddy Gadekallu, T., Aboudaif, M.K. and Abouel Nasr, E. (2021) Robust Attack Detection Approach for IIoT Using Ensemble Classifier. *Computers, Materials & Continua*, **66**, 2457-2470. https://doi.org/10.32604/cmc.2021.013852

[4] Alotaibi, B. and Alotaibi, M. (2020) A Stacked Deep Learning Approach for IoT Cyberattack Detection. *Journal of Sensors*, **2020**, Article ID: 8828591. https://doi.org/10.1155/2020/8828591

[5] Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A.V. and Rong, X. (2015) Data Mining for the Internet of Things: Literature Review and Challenges. *International Journal of Distributed Sensor Networks*, **11**, 1-14. https://doi.org/10.1155/2015/431047

[6] Mahdavinejad, M.S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P. and Sheth, A.P. (2018) Machine Learning for Internet of Things Data Analysis: A Survey. *Digital Communications and Networks*, **4**, 161-175. https://doi.org/10.1016/j.dcan.2017.10.002

[7] Koay, A.M.Y., Ko, R.K.L., Hettema, H. and Radke, K. (2022) Machine Learning in Industrial Control System (ICS) Security: Current Landscape, Opportunities and Challenges. *Journal of Intelligent Information Systems*, **60**, 377-405. https://doi.org/10.1007/s10844-022-00753-1

[8] Isaac, S., Ayodeji, D.K., Luqman, Y., Karma, S.M. and Aminu, J. (2024) Cyber Security Attack Detection Model Using Semi Supervised Learning. *FUDMA Journal of Sciences* (*FJS*), **8**, 92-100. https://doi.org/10.33003/fjs-2024-0802-2343

[9] Maheswari, L.C.U., Srivalli, G., Shivani, G., Nikitha, G.S. and Kaveri, K. (2023) A Novel Web Attack Detection System for Internet of Things via Ensemble Classification. *Turkish Journal of Computer and Mathematics Education*, **14**, 834-845.

[10] Yavuz, F.Y., Ünal, D. and Gül, E. (2018) Deep Learning for Detection of Routing Attacks in the Internet of Things. *International Journal of Computational Intelligence Systems*, **12**, 39-58. https://doi.org/10.2991/ijcis.2018.25905181

[11] Al-Sultani, Z.N. (2012) Learning Vector Quantization (LVQ) and k-Nearest Neighbor for Intrusion Classification. *World of Computer Science and Information Technology Journal*, **2**, 105-109.

[12] Jawhar, M. and Mehrotra, M. (2010) Design Network Intrusion Detection System Using Hybrid Fuzzy-Neural Network. *International Journal of Computer Science and Security*, **4**, 285-294.

[13] Kouassi, B.M., Monsan, V. and Adou, K.J. (2024) Intelligent Detection and Identification of Attacks in IoT Networks Based on the Combination of DNN and LSTM

Methods with a Set of Classifiers. *Open Journal of Applied Sciences*, **14**, 2296-2319. https://doi.org/10.4236/ojapps.2024.148153

[14] Sasi, T., Lashkari, A.H., Lu, R., Xiong, P. and Iqbal, S. (2024) A Comprehensive Survey on IoT Attacks: Taxonomy, Detection Mechanisms and Challenges. *Journal of Information and Intelligence*, **2**, 455-513. https://doi.org/10.1016/j.jiixd.2023.12.001

[15] Siraparapu, S.R. and Azad, S.M.A.K. (2024) Securing the IoT Landscape: A Comprehensive Review of Secure Systems in the Digital Era. *e-Prime—Advances in Electrical Engineering, Electronics and Energy*, **10**, Article ID: 100798. https://doi.org/10.1016/j.prime.2024.100798

[16] Anbarasi, M., Anupriya, E. and Iyengar, S.N. (2010) Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. *International Journal of Engineering Science and Technology*, **2**, 5370-5376.

[17] Adeniyi, A.D., Ajoge, N.S. and Sulaiman, U.I. (2021) Design and Realization of Pre-Ordered Feature Ranking Filtering (PFRF) Feature Selection Method for Machine Learning Algorithms. *International Journal of Engineering and Technology Research*, **21**, 82-100.

[18] Kulkarni, S.A., Gurupur, V.P. and King, C. (2022) Impact Analysis of Stacked Machine Learning Algorithms Based Feature Selections for Deep Learning Algorithm Applied to Regression Analysis. *SoutheastCon* 2022, Mobile, 26 March-3 April 2022, 269-275. https://doi.org/10.1109/southeastcon48659.2022.9764105

[19] Mebawondu, O.J., Adetunmbi, A.O., Mebawondu, J.O. and Alowolodu, O.D. (2021) Feature Weighting and Classification Modeling for Network Intrusion Detection Using Machine Learning Algorithms. In: Misra, S. and Muhammad-Bello, B., Eds., *Information and Communication Technology and Applications*, Springer, 315-327. https://doi.org/10.1007/978-3-030-69143-1_25

[20] Mebawondu, O.J. (2024) Enhancing Intrusion Detection Systems with Efficient Deep Learning Techniques. 2024 *IEEE* 5*th International Conference on Electro-Computing Technologies for Humanity* (*NIGERCON*), Ado Ekiti, 26-28 November 2024, 1-5. https://doi.org/10.1109/nigercon62786.2024.10927178

[21] Hegde, S.K., Hegde, R., Hombalimath, V., Palanikkumar, D., Patwari, N. and Dankan Gowda, V. (2023) Symmetrized Feature Selection with Stacked Generalization Based Machine Learning Algorithm for the Early Diagnosis of Chronic Diseases. 2023 5*th International Conference on Smart Systems and Inventive Technology* (*ICSSIT*), Tirunelveli, 23-25 January 2023, 838-844. https://doi.org/10.1109/icssit55814.2023.10061062

[22] Domingos, P. and Hulten, G. (2002) Learning from Infinite Data in Finite Time. In: Dietterich, T.G., Becker, S. and Ghahramani, Z., Eds., *Advances in Neural Information Processing Systems* 14, The MIT Press, 673-680. https://doi.org/10.7551/mitpress/1120.003.0091

[23] Flores, J.J., Rodriguez, H. and Graff, M. (2010) Reducing the Search Space in Evolutive Design of ARIMA and ANN Models for Time Series Prediction. In: Sidorov, G., Hernández Aguirre, A. and Reyes García, C.A., Eds., *Advances in Soft Computing*, Springer, 325-336. https://doi.org/10.1007/978-3-642-16773-7_28

[24] Sebban, M. and Nock, R. (2002) A Hybrid Filter/Wrapper Approach of Feature Selection Using Information Theory. *Pattern Recognition*, **35**, 835-846. https://doi.org/10.1016/s0031-3203(01)00084-x

[25] Antonelli, D., Baralis, E., Bruno, G., Cerquitelli, T., Chiusano, S. and Mahoto, N. (2013) Analysis of Diabetic Patients through Their Examination History. *Expert Systems with Applications*, **40**, 4672-4678. https://doi.org/10.1016/j.eswa.2013.02.006

[26] García-Pedrajas, N. and de Haro-García, A. (2012) Scaling up Data Mining Algorithms: Review and Taxonomy. *Progress in Artificial Intelligence*, **1**, 71-87. https://doi.org/10.1007/s13748-011-0004-4

[27] Mazarei, A., Sousa, R., Mendes-Moreira, J., Molchanov, S. and Ferreira, H.M. (2024) Online Boxplot Derived Outlier Detection. *International Journal of Data Science and Analytics*, **19**, 83-97. https://doi.org/10.1007/s41060-024-00559-0

[28] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[29] Adele Cutler, D., Cutler, R. and Stevens, J.R. (2011) Ensemble Machine Learning: Methods and Applications. Springer.

[30] Han, J. and Kamber, M. (2006) Data Mining Concept and Techniques. 2nd Edition, Morgan Kaufmann Publishers, 285-350.

[31] Aggarwal, C.C. and Reddy, C.K. (2016) Data Clustering Algorithms and Applications. Chapman and Hall. https://doi.org/10.1201/9781315373515

[32] Zaid, M.A. (2015) Correlation and Regression Analysis. The Statistical, Economic and Social Research and Training Centre for Islamic Countries (SESRIC) Kudüs Cad. https://www.sesric.org

[33] Nworgu, B.G. (1991) Educational Research: Basic Issues and Methodology. Wisdom Publishes Ltd.