Anomalous Network Behaviour Detection in Interoperable Health Systems Using Machine Learning in Resource Limited Areas

Marseline Michael Mtey[1], Anael Elikana Sam[2], Mussa Ally Dida[3]

[1,2,3]NM-AIST, P.0.Box 447, Arusha-Tanzania

Abstract

The connection of devices in distributed environments produces and shares a vast amount of data useful for different organisational decision-making. In healthcare service organisations, for example, multiple e-health systems from different departments or facilities connect and share health data and information. During sharing, proper management is important to ensure the information is secure against intruders. Machine learning as a non-conventional security technique can be used along conventional techniques like firewalls, antivirus and intrusion detection systems to predict future network threats and other anomalies using historical backgrounds and other features. However, some machine learning algorithms have complex computation thus requiring resourceful systems in terms of network bandwidth, CPU power, memory, and storage capacity. In resource-constrained environments, therefore, special consideration is needed to ensure that the analysis of the big data is successful and that the benefits associated with them are effectively obtained.

In this paper, a Machine Learning algorithm was selected among four algorithms whose performance was compared through various performance metrics. Classification accuracy, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Relative Absolute Error (RAE) among other performance metrics were used to compare the ANN, Random forest, Decision trees, and Naïve byes classification algorithms using an extract from CICDDOS2019 dataset. Using the Weka version 3.8.6, the algorithms were compared to choose the best one to classify the data.

By using three computers with different resources, the experiments were carried out to determine the performance of those machine learning algorithms. The result revealed that the random forest produced a good average classification performance in resource-limited systems since it surpassed other algorithms in classifying the data at an average of 99 per cent with a low average mean absolute error of 0.0001. Furthermore, as an ensemble that classifies with multiple decision trees algorithm, it likewise uses reasonable time to build and test the model therefore recommended for resource-limited systems.

Keywords: Machine learning algorithms, Interoperability, Big data, Security, Anomalous behaviour, Resource limited areas, e-health system, Tanzania, Data Sharing.

1. Introduction

1.1.Background of the Study

The introduction of a diverse number of devices and technology in healthcare systems has resulted in better service provision globally. Tanzania and other Low and Middle-Income Countries use data from smart wearable devices, IoTs and sensors to monitor personal health anywhere provided that they are equipped with relevant technology [1]. Similarly, in distributed computing where several computers are connected, huge data are produced useful for scientific research, innovations, and in prediction of various diseases [2]. Healthcare centres' departments as well as branches can share data instantly when needed through their communication channels. This process improves patient attendance and the diagnosis process quicker than the siloed systems. Sharing this big data that is produced requires a sophisticated system to cope with the speed and volume of production, as well as velocity [3]; [4].

Relevant allocation of resources ensures that the analysis becomes successful in retrieving the hidden meaning of the big data. However, the health systems from the majority of facilities are limited in computation and storage resources, to manage big data. This forces the operations to be outsourced to third-party service providers like cloud computing within or outside the country. The outsourced party provides enough storage space and computation capacity thus reducing the data management burden by the health centres.

With the sharing and use of big data, there is a high possibility of intruders taking advantage to attack the systems if protective measures are not taken into account considering the sensitivity of health data. Also in distributed systems, it is likely for attackers to try to acquire an identity of one of the systems and exploit vulnerability [5]. Furthermore, data-sharing patterns with third-party service providers may be studied by intruders to trigger their attacks [6]; [7]. In addition, some attackers inspect slow networks and trigger their actions with the advantage of the delay in request-response sessions for slow connections.

Noting the possibility and availability of attacks, various solutions have been proposed to secure the systems by ensuring that any suspicious activities are disclosed and avoided [8]. The solutions employ both conventional and non-convention security techniques. The conventional security techniques include the use of antivirus, anti-malware, Intrusion Detection and Prevention Systems (IDPS) and firewalls within or outside the system. Also, common conventional techniques include the use of cryptographic measures for securing data and information [9].

Since conventional techniques cannot detect unidentified attacks in the network and cannot forecast or predict the security situation of a system based on various observed behaviours or historical data, the non-conventional technique is called upon [10]. Among the non-convention techniques used include a machine learning algorithm to detect attacks in the network by adjusting against any form of threat [11]. Various studies applied machine learning in the health domain including the HealthGuard [12], dynamic security-aware routing mechanism framework [13], an integrated system of a base station and sensor nodes in the IoT [14] and the medical image processing by deep learning models [3]. All the proposed systems aim at securing and managing health data against attacks though they did not focus on the resource consideration of the associated parties.

1.2.Challenges

Various solutions to secure data in integrated health systems are proposed including both conventional and non-conventional techniques. They aim at protecting data and information shared among the systems against all forms of attacks. It was observed that non-conventional techniques specifically machine learning algorithms are preferred over conventional techniques because they can predict the threats based on various observed behaviours or historical data.

However, some machine learning algorithms have complex computations that cannot be possible using limited resources including network bandwidth, CPU power, memory, and storage capacity. Using these systems with limited resources could result in timeout errors, or attackers may easily study the data-sharing pattern to trigger their attacks. In addition, if third-party service providers are opted for, there is a risk of disclosure of sensitive health information to unauthorised users due to the curious behaviour of third-party service providers. It is essential therefore to propose a solution that will serve as a remedy to resource-limited areas without requiring a complex mathematical operation.

This study therefore aims to propose a machine learning algorithm to detect network anomalous behaviour in integrated health systems in resource-limited areas.

1.3. Objectives

The main objective of this study is to propose a machine learning algorithm to detect network anomalous behaviour in integrated health systems in resource-limited areas.

Specifically, the study aims to:

1. Reviewing available security solutions to protect health data against attacks
2. Analyse the machine learning-based solutions for protecting health data
3. Propose machine learning algorithms for detecting anomalous behaviour in the network.

1.4. Contributions

This study will contribute in the following areas:

1. In data security aspects it provides a secure mechanism to protect sensitive health data confidentiality and integrity as well as ensuring its availability.
2. It provides insights into the application of machine learning algorithms in securing health data records.
3. It helps in devising ways to protect personal health records and clinical data against cyber-attacks.
4. It helps to actively and passively prevent network security threats and attacks.

2. Literature Review

To simplify organizational activities, organizations' departments or branches, are linked to share data between them. The sharing process can be possible if the systems are connected by accepting sets of agreed standards in an interoperable manner. In interoperability, systems can share, exchange and use data and information for decision-making [15]. In healthcare centres patients' health records are needed at different departments for various medical decisions. The data comes from several systems and devices including IoT, wearables and sensors all equipped with networks to enable them to share data [16]. Sharing that information through the network needs proper management to avoid disclosure to unauthorized parties. Security is among the essential management aspects of health information when sharing with the public or through the network due to their sensitivity [17]. Any breach to heathcare system and data may pose a serious challenge to both the facility and the patients [18].

To have secure integrated health information systems, various solutions ranging from local to national levels have been applied. Some of the solutions include the integration of systems within a single health facility, within a specific geographic area, or throughout a country. Developed nations of Europe, aimed at having an integrated system for all the European Union (EU) countries while the effort is to have a national interoperable system in Low and middle-income countries [1].

In Asia, [19] discussed efforts to develop a nationwide China-interoperable HIS for all health facilities systems by the end of the year 2020. The authors pointed out the aim for China to have an interoperable HIS by the end of the year 2020. The system is aimed at ensuring that healthcare facilities share health data. China set a guideline for all healthcare authorities nationwide to construct a Population Health Information Platform (PHIP) aiming at attaining nationwide interoperability. To date, some states in China have successfully established their integrated systems including Shenzhen where some hospitals have already adhered to the provided guidelines [20].

Again, [21] devised an interoperable HIS focusing more on data sharing than a system-to-system integration. The authors showed the challenges that developers of interoperable systems face including focusing on machine-to-machine interoperability than on the data part. They proposed the use of bidirectional transformations (BX) based on the Clinical Document Architecture (CDA) developed by HL7 for exchanging medical records.

[22] reviewed the efforts shown by developing countries towards having an Integrated HIS based on EA for resource-limited areas. The authors reviewed the efforts in some Low- and Middle-Income Countries (LMICs) including India, Sierra Leone, Rwanda, Jordan and South Africa. In their review, they noted that there exists no tangible EA system in LMICs though there is a good effort towards design and implementation. Furthermore, [23] proposed a framework policy to be used for Low-income countries as well as Lower middle-middle-income countries. After reviewing various literatures they proposed steps for developing a policy framework including considering stakeholders' readiness, considering realistic expectations of the environment, targeting local context and performing continuous evaluation. However, the study did not provide a way forward on security specifically in resource-limited areas which is a key consideration of this study.

In interoperable health systems, since there is sharing of a massive amount of data from those devices, the cost of health service provision becomes low at the cost of security and resource burden [17]; [24]. This needs outsourcing computations to a third-party service provider which in turn increases security risks. Some attackers target the data in storage servers while others sniff the communication pattern to gain control of one node in the network. Some common network attacks include a popular *Brute Force Attack* in which an attacker can crack passwords, and discover hidden pages and content in a web application by a hit-and-try attack until it succeeds [25]. Multiple trials are done by the attackers by guessing multiple passwords or other identities until they succeed. Also, in a *Heartbleed Attack* which is a bug in the OpenSSL cryptography library, a Transport Layer Security (TLS) protocol is exploited by sending a malformed heartbeat request to the victim. This periodic signal generated by hardware or software indicates normal operation or to synchronize other parts of a computer system. Since it is sent between machines at regular intervals in the order of seconds the vulnerable part may be deceived of communicating to legitimate system.

Similarly, Denial of Services (DoS) are very common attack in networks that make efforts to block data transmission [26]. A single bot from a botnet can attack a network or multiple bots can do the same to make a system or network resource unavailable temporarily. In a DoS attack, a targeted machine or resource is flooded with superfluous requests in an attempt to overload systems and prevent some or all legitimate requests from being fulfilled. On the other hand, a *Distributed DoS (DDoS) Attack* occurs when multiple systems, flood the bandwidth or resources of a victim as a result of multiple compromised systems (for example, a botnet) flooding the targeted system with generating huge network traffic [27]; [28].

Other attacks include Web Attacks, Botnet, Infiltration Attacks and other multiple attacks being sent to the internet every day such as IP sweep, full port scan and service enumerations using Nmap [29].

Since every attack is associated with some behaviours and features of a network, similarly their detection also requires a thorough understanding of its attack behaviour. Conventional techniques like user authentication, cryptography and intrusion prevention systems like firewalls are used [30]. Similarly, blockchain technology with its distributed architecture, immutable ledger, and advanced security options can also be used to protect electronic health record sharing between different e-health systems [31].

Blockchain secure healthcare data by ensuring integrity and transparency to improve service delivery [32]. However, since some attacks are advanced and change their forms with each attack, a normal conventional technique cannot detect them because frequent update and patch is needed to detect them [33]. To overcome this, various intrusion detection systems (IDS) are developed to detect various types of malicious network traffic and computer usage, which cannot be detected by a conventional firewall [34]. They are one of the most important defence tools against sophisticated and ever-growing network attacks [29].

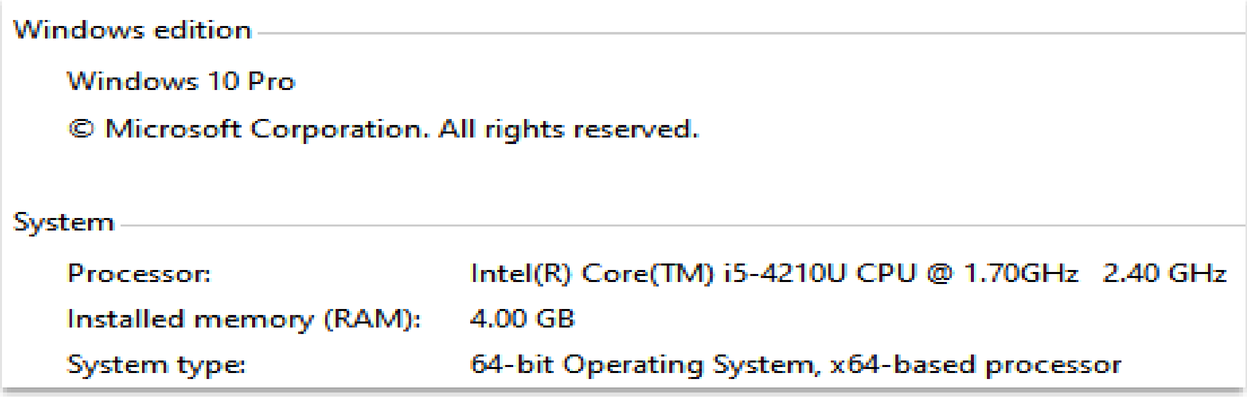
With dynamic attack behaviours, the application of machine learning techniques can serve to detect network attacks by studying those behaviours [35]. According to [36], the advancement of deep neural networks has become effectively applied in detecting anomaly behaviour in sharing multimedia data in networks. [37] pointed out that, security in networks especially anomalous behaviour can be detected with the application of machine learning. Similarly, in integrated systems especially when there is a combination of Internet of Things (IoT) devices, machine learning algorithms can detect anomalous behaviour in the sharing network. Since the network shares a large amount of data, it is therefore perfect to use machine learning algorithms, since they give a high-accuracy prediction for attacks when large data sets are used [38].

3. Methods

This study employed an experimental method to analyse the selected machine learning algorithms using some features. Three computer systems with different specifications were used in the performance comparisons. Also, four supervised machine-learning algorithms were selected from the available classification algorithms.

3.1. The Computer Systems Selection

Three computers with different capacities were used in the comparison as indicated by Figures 3.1,3.2 and 3.3.



Windows edition

Windows 10 Pro

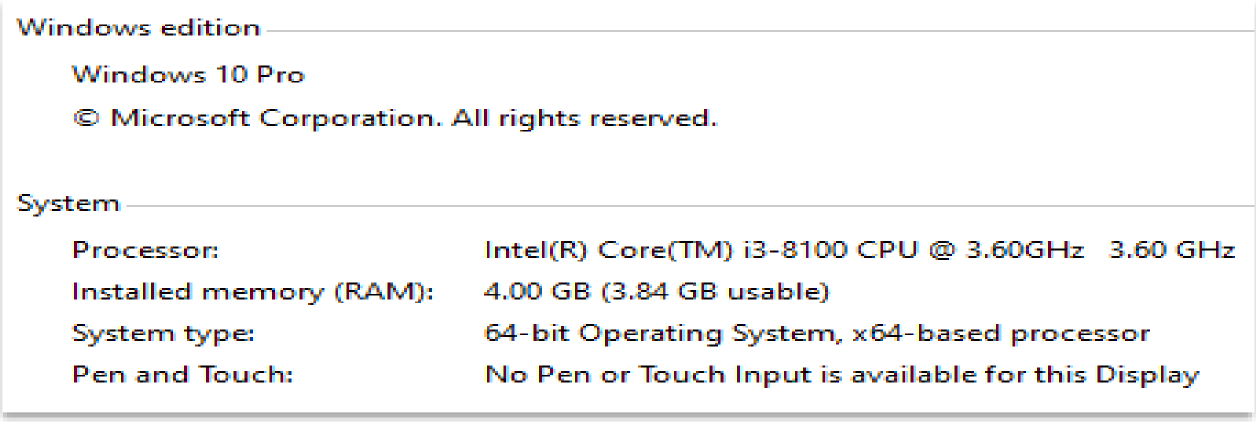
© Microsoft Corporation. All rights reserved.

System

Processor:	Intel(R) Core(TM) i5-4210U CPU @ 1.70GHz 2.40 GHz
Installed memory (RAM):	4.00 GB
System type:	64-bit Operating System, x64-based processor

Figure 3.1. The first System Specifications

The first computer had the core i5 CPU with a 2.40Ghz speed, a memory capacity of 4.00GB and a storage of 223 GB. The system was running on Windows 10.



Windows edition

Windows 10 Pro

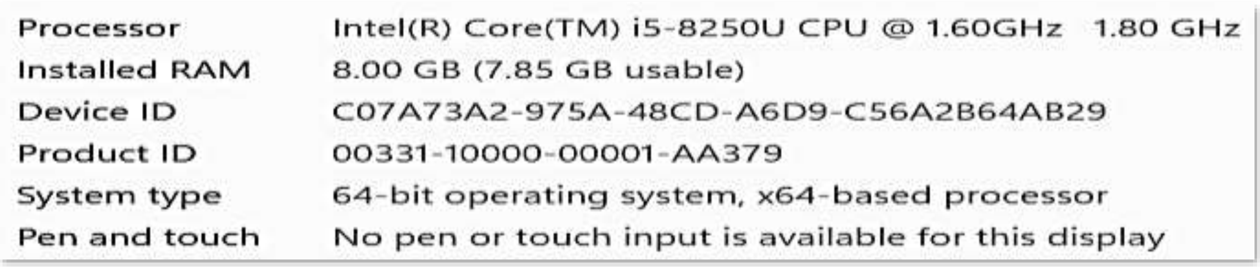
© Microsoft Corporation. All rights reserved.

System

Processor:	Intel(R) Core(TM) i3-8100 CPU @ 3.60GHz 3.60 GHz
Installed memory (RAM):	4.00 GB (3.84 GB usable)
System type:	64-bit Operating System, x64-based processor
Pen and Touch:	No Pen or Touch Input is available for this Display

Figure 3.2. The Second System Specifications

The second computer was running on Windows 10 with a CPU speed of 3.60Ghz, installed memory of 4.00GHz and Storage Capacity of 929 GB.



Processor	Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz 1.80 GHz
Installed RAM	8.00 GB (7.85 GB usable)
Device ID	C07A73A2-975A-48CD-A6D9-C56A2B64AB29
Product ID	00331-10000-00001-AA379
System type	64-bit operating system, x64-based processor
Pen and touch	No pen or touch input is available for this display

Figure 3.3. The Third System Specifications

The third system was running on Windows 11 with a CPU speed of 1.80 GHz, installed memory of 8.00GB and 1 TB storage.

To compare their performance, four machine learning algorithms were compared through all three computers. The CICDDoS2019 dataset was used to compare the algorithms whereby their average performance was recorded to select one among them with better performance.

3.2. Algorithms and Dataset Selection

Random selection of four supervised Machine learning algorithms was done since they are the classification algorithms. The aim was to classify the dataset based on benign or anomalous classes using the selected dataset. The algorithms included Naïve Bayes with multinomial text, J48 Decision Tree, Random Forest and Multilayer Perceptron algorithms. The algorithms were tested on the CICDDoS2019 dataset to compare their performance in classifying the data. The CICDDoS2019 dataset was developed by the Canadian Institute for Cybersecurity to be used for detecting anomalous network behaviour [39]. It consists of benign and common DDoS attacks, representing the true real-world data. The attacks include NetBIOS, LDAP, MSSQL, UDP, UDPLag and Syn. The algorithms' performances were measured based on various performance metrics including Recall, Precision, F-Measure, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Relative Absolute Error (RAE).

3.3. Features Selection

From more than 80 network features of the CICDDOS2019 dataset, the study randomly selected 22 features for the experiment. Some of them included:

1. Source Port
2. Protocol
3. Total Fwd Packets
4. Total Backward Packets
5. Total Length of Packets
6. Fwd IAT Min
7. Bwd IAT Total
8. Bwd IAT Max
9. Bwd IAT Min
10. Fwd PSH Flags
11. Bwd PSH Flags
12. Fwd URG Flags
13. Bwd URG Flags
14. Fwd Header Length
15. Bwd Header Length
16. Min Packet Length
17. Max Packet Length
18. FIN Flag Count
19. Label

The reason for the selection was to have a minimum number of features and representation of the dataset to ensure having the same result as when the whole dataset was used. The other reason is the limited capacity of the systems used unable to analyse the whole dataset.

3.4. Performance Metrics Selection

From the list of various performance metrics, some were purposely selected to compare the performance of the four algorithms. They included Recall, Precision, F-Measure, Measurement Errors as well as time and classification correctness.

3.4.1.Recall, Precision and F-Measure

In this performance measure, the analysis is done by comparing two performance metrics, the referencing set and a referenced set. By considering two distributions P as a reference and Q as a referencing one, precision should measure how much of Q can be generated by a “part” of P while recall should measure how much of P can be generated by a “part” of Q [40]. In this case, the dataset was divided into a test set as a referencing set which was tested against the training set as a reference set to determine the algorithms’ performance.

The F-Measure is obtained from the measure of Precision and Recall with the following formula.

$$F\text{-}Measure = (2 * Precision * Recall) / (Precision + Recall)$$

3.4.2. Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Relative Absolute Error (RAE)

MAE and RMSE are used for evaluation to measure the difference between the predicted values and actual values. Therefore, in this study, the algorithms were evaluated against the data set through a test set and training set.

They are represented by the following formula:

MAE= $$\frac{\sum_{n=1}^{N}\left|\hat{r}_n - r_n\right|}{N}$$ RMSE= $$\sqrt{\frac{\sum_{n=1}^{N}\left(\hat{r}_n - r_n\right)^2}{N}}$$

Where:

$r\hat{}_n$ is the prediction; r_n is the true value in the testing data set; while N is the number of samples.

Relative Absolute Error (RAE) is used to measure the performance of a model. It compares the ratio of a mean error (residual) to errors produced by the model.

3.5. The Results

To find the classification accuracy of the four (4) machine learning algorithms, Weka machine learning software developed by the University of Waikato, New Zealand was used. The output was based on how accurately the model identified the benign and malicious activity in the network during data sharing.

A total of 18710 instances with 22 attributes were used for the classification to compare the models' performance. The evaluation metrics for the first system are indicated in Table 3.1.

Table 3.1: Performance Classification of the Models for System 1

SNO	Algorithm	Recall	Precision	F-Measure	MAE	RMSE	RAE
1	Naïve Bayes	1.00	0.66	0.80	0.11	0.20	100
2	Decision Tree	0.99	0.99	0.99	0.001	0.02	0.98
3	Random Forest	0.99	0.99	0.99	0.000	0.01	0.68
4	ANN	0.95	0.86	0.75	0.01	0.09	13.96

The table indicate that Naïve bayes has recall of 1.00, Precision of 0.66, F-Measure of 0.80, MAE of 0.11, RMSE of 0.20 and RAE of 100. On the other hand, Decision tree algorithms has Recall, Precision and F-Measure of 0.99, MAE of 0.01, RMSE of 0.02 as well as RAE of 0.98. Also, Random Forest algorithm has Recall, Precision and F-Measure of 0.99, MAE of 0.00, RMSE of 0.01 as well as RAE of 0.68. In addition, ANN has Recall of 0.95, Precision of 0.86 and F-Measure of 0.75, MAE of 0.01, RMSE of 0.09 as well as RAE of 13.96

On the other hand, the time taken to test, train and build the model and the percentage correctness of classifying the instances in the first system for all the algorithms are shown in Table 3.2.

Table 3.2: Classification Accuracy and Time for System 1

Sno	Algorithm	Test Time (s)	Build Time (s)	% Classification Correctness	% Incorrect classification
1	Naïve Bayes	0.40	0.08	66.70	33.30
2	Decision Tree	0.24	0.86	99.73	0.27
3	Random Forest	0.64	2.82	99.88	0.11
4	ANN	0.18	47.9	95.35	4.65

The table indicate that Naïve Bayes algorithm has the lowest build time of 0.08 compared to 0.86, 2.82 and 47.7 for Decision tree, Random forest and ANN respectively. However, Rando forest algorithm has the highest classification correctness of 99.88 in comparison to 66.70 for Naïve Bayes, 99.73 for Decision Tree and 95.35 for ANN.

For the second system Table 3.3 shows the evaluation metrics for the algorithms while Table 3.4 shows the time taken to test, train and build the model as well as the percentage correctness.

Table 3.3: Performance Classification of the Models for System 2

SNO	Algorithm	Recall	Precision	F-Measure	MAE	RMSE	RAE
1	Naïve Bayes	0.96	0.96	0.96	0.01	0.11	13.56
2	Decision Tree	0.99	0.98	0.99	0.002	0.04	2.47
3	Random Forest	0.99	0.99	0.99	0.002	0.03	2.12
4	ANN	0.98	0.96	0.94	0.006	0.06	6.18

Table 3.3 indicates that Naïve bayes has Recall, Precision and F-Measure of 0.96, MAE of 0.01, RMSE of 0.11 and RAE of 13.56. Decision tree algorithms has Recall, and F-Measure of 0.99, Precision of 0.98, MAE of 0.02, RMSE of 0.04 and RAE of 2.47. Random Forest algorithm has Recall, Precision and F-Measure of 0.99, MAE of 0.02, RMSE of 0.03 as well as RAE of 2.12. on the other hand, ANN has Recall of 0.98, Precision of 0.96, F-Measure of 0.94, MAE of 0.006, RMSE of 0.06 and RAE of 6.18

Table 3.4: Classification Accuracy and Time for System 2

Sno	Algorithm	Test Time (s)	Build Time (s)	% Classification Correctness	% Incorrect classification
1	Naïve Bayes	0.34	0.11	95.79	4.21
2	Decision Tree	0.03	0.42	99.44	0.56
3	Random Forest	0.19	2.67	99.55	0.45
4	ANN	0.08	46.44	98.42	1.57

For classification accuracy and time, Decision tree has the lowest test time of 0.03 while Naïve bayes has the lowest build time of 0.11. However, Random forest has the highest classification correctness of 99.55 percent compared to ANN which has the lowest classification correctness of 98.42 percent.

For the third system, Table 3.5 shows the evaluation metrics for the algorithms while Table 3.6 shows the time taken to test, train and build the model as well as the percentage correctness.

Table 3.5: Performance Classification of the Models for System 3

SNO	Algorithm	Recall	Precision	F-Measure	MAE	RMSE	RAE
1	Naïve Bayes	0.96	0.96	0.95	0.013	0.11	13.56
2	Decision Tree	0.99	0.98	0.99	0.002	0.04	2.47
3	Random Forest	0.99	0.99	0.99	0.002	0.03	2.12
4	ANN	0.98	0.96	0.94	0.010	0.06	6.19

The table indicates that Naïve bayes has Recall, Precision and F-Measure of 0.96, MAE of 0.013, RMSE of 0.11 and RAE of 13.56. Also, Decision tree algorithms has Recall, and F-Measure of 0.99, Precision of 0.98, MAE of 0.02, RMSE of 0.04 and RAE of 2.47. Random Forest algorithm has Recall, Precision and F-Measure of 0.99, MAE of 0.02, RMSE of 0.03 as well as RAE of 2.12. On the other hand, ANN has Recall of 0.98, Precision of 0.96, F-Measure of 0.94, MAE of 0.010, RMSE of 0.06 and RAE of 6.19

Table 3.6: Classification Accuracy and Time for System 3

Sno	Algorithm	Test Time (s)	Build Time (s)	% Classification Correctness	% Incorrect classification
1	Naïve Bayes	0.66	0.20	95.79	4.21
2	Decision Tree	0.03	0.72	99.44	0.56
3	Random Forest	0.25	3.64	99.55	0.45
4	ANN	0.11	86.14	98.42	1.58

The time to test and build the model shows that Decision Tree has the lowest test time of 0.03 while Naïve Bayes has the lowest build time of 0.20. However, though the Random Forest has the highest test time of 0.25 and second highest build time of 3.64 next to ANN with 86.14, it has the highest classification accuracy of 99.55 compared to ANN with 98.42

The average performance classification of the four algorithms are indicated by Table 3.7.

Table 3.7: Algorithms average score

Algorithm	Recall	Precision	F-Measure	MAE	RMSE	RAE
Naïve bayes	0.97	0.86	0.90	0.044	0.14	42.37
Decision tree	0.99	0.98	0.99	0.002	0.03	1.97
Random forest	0.99	0.99	0.99	0.001	0.02	1.64
ANN	0.97	0.93	0.88	0.009	0.07	8.78

The table indicate that Random Forest has the highest Recall, Precision and F-Measure of 0.99 with lowest MAE, RMSE and RAE of 0.001, 0.02 and 1.64 respectivelly. Naïve Bayes has the lowest Recall, Precision and F-Measure of 0.97, 0.86 and 0.90 respectivelly with the highest errors at 0.044 for MAE, 0.14 for RMSE and 42.37 for RAE.

On the other hand, Table 3.8 shows the time taken to test and build the model and classification correctness.

Table 3.8: Average Time and Classification Correctness

Algorithm	Test Time (s)	Build Time (s)	Classification Correctness (%)	Incorrect classification (%)
Naïve Bayes	0.5	0.13	86.09	13.9
Decision Tree	0.1	0.67	99.54	0.5
Random Forest	0.4	3.04	99.66	0.3
ANN	0.1	60.16	97.39	2.6

The table indicate that, Decision Tree and ANN have the lowest test time of 0.1 compared to Naïve Bayes with the highest test time of 0.5 and Random Forest had 0.4. For the build time, Naïve Bayes has the lowest build time of 0.13 compared to 60.16 seconds for ANN. However, the Random Forest algorithm had the highest classification accuracy of 99.66 compared to 86.09 for Naïve Bayes.

4. Discussion of the findings

In this part, the findings from the results of the conducted experiments to compare the performance of the selected machine learning algorithms using the three systems are discussed.

The average score of the four algorithms based on Recall, Precision, F-Measure, MAE, RMSE and RAE were used to determine the one with better score. Figure 4.1 indicates the trend of the average results of the four algorithms based on those performance metrics.

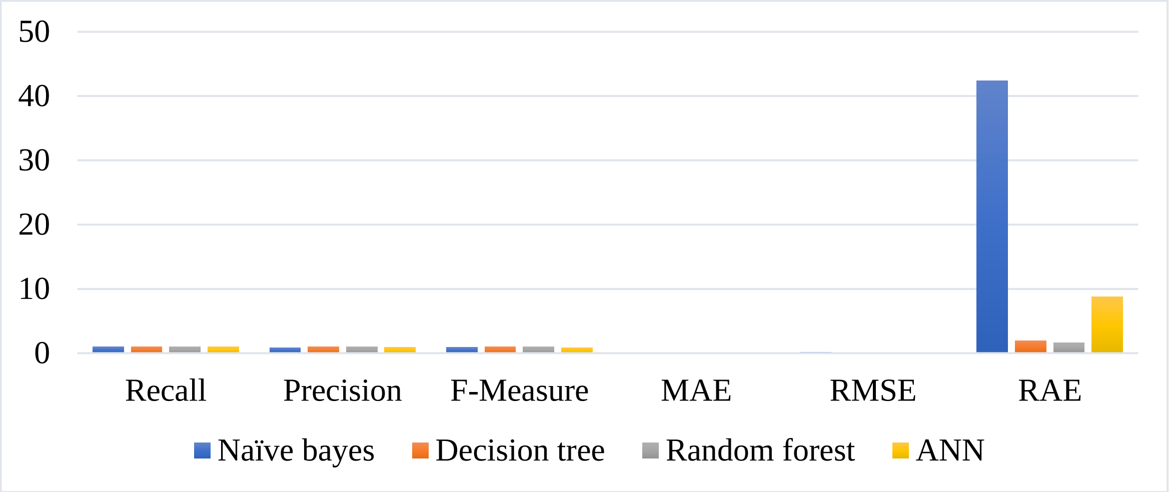


Figure 4.1: Perfomance Metrics for the Algorithms

Figure 4.1, indicates that Naive Bayes and ANN algorithms shared the same value for Recall of 0.97. Similarly, Decision tree and Random Forest algorithms shared the same value of Recall of 0.99. Recall indicates the value of correct positive predictions out of all positive predictions, thus clarify that Decision tree and Random Forest to be better compared to others in prediction. On the other hand, Precision which quantifies the number of correct positive predictions made, indicates that the random forest has a Precision of 0.99 in comparison to others of 0.86, 0.98 and 0.93 for Naive Bayes, Decision tree and ANN respectively.

Recall and Precision determines the F1 measure for the algorithms. Decision tree and Random forest shared the same valu of F-Measure of 0.99. However, by showing a closer similar perfomance between Random Forest and Decision tree, Random Forest has the lowest classification errors with MAE of 0.001, RMSE of 0.02, and RAE of 1.64. This indicate that random forest has better positive predictions with small errors compared to other algorithms though the decision tree has closer positive prediction.

Furthermore, Figure 4.2 indicates the trend of the algorithms classification accuracy.

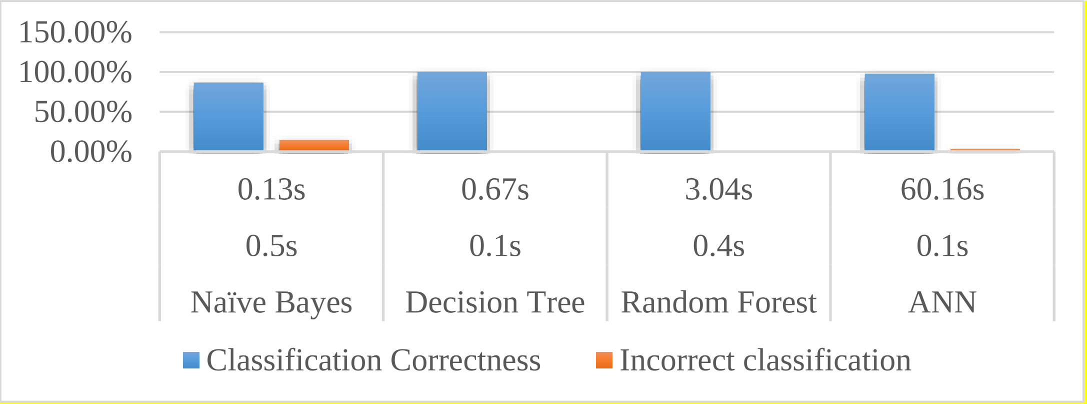


Figure 4.2:Classification Accuracy

The Figure indicate that Random Forest algorithm has the highest classification correctness of 99.66 percent compared to others at 86.09 by Naive Bayes, 99.54 by Decision Tree and 97.39 by ANN. This figure indicate that Random Forest is superior to other algorithms in classifying the network traffic based on benign and anomalous activities.

Based on the results and findings , Random Forest is an algorithms that can perform well in resource limited situations based on the used systems and the output from the experiments. This study therefore, recommend a Random Forest algorithm to be used in resource constrained areas for low errors and higher classification correctness. This concur with Ullah & Mahmoud, (2020) who claimed that random forest classifier produces good results on predictions, especially on anomalous activity in IoT networks.

5. Conclusions and Future Work

5.1. Conclusion

In detecting anomalous network behaviour using a machine learning algorithm in resource-limited areas, an algorithm that classifies precisely, with less error rate, that uses a reasonable time in building and testing the model is the preferred one.

In this study, after conducting the experiments to compare the efficiency of different machine learning models on attack scenarios in resource-limited areas, it was concluded that random forest showed a reasonable response in comparison to others. The models that were challenged include Naïve Bayes with multinomial text, J48 Decision Tree, Random Forest and multilayer perceptron. A total of 18710 instances with 22 attributes were used to check traffic behaviours during data sharing.

Random forest was therefore recommended for use, when sharing data in resource-limited areas or systems, to detect anomalous behaviour and thus to prevent the sharing system from being compromised.

5.2. Future Work

In the future, integration of the algorithm with another mode of attack detection systems is recommended. Furthermore, more than 22 (twenty-two) features are to be used in further studies to ensure there is a more precise and more accurate result.

Acknowledgement

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References.

1. M. M. Mtey and M. A. Dida, "Towards interoperable e-Health system in Tanzania: analysis and evaluation of the current security trends and big data sharing dynamics," *Int. J. Adv. Technol. Eng. Explor.*, vol. 6, no. 59, pp. 225–240, 2019, doi: 10.19101/ijatee.2019.650057.
2. S. Zehra *et al.*, "Machine Learning-Based Anomaly Detection in NFV: A Comprehensive Survey," *Sensors*, vol. 23, no. 11, pp. 1–26, 2023, doi: 10.3390/s23115340.
3. V. Mareeswari, R. Vijayan, E. Sathiyamoorthy, and E. P. Ephzibah, "A narrative review of medical image processing by deep learning models: Origin to COVID-19," *Int. J. Adv. Technol. Eng. Explor.*, vol. 9, no. 90, pp. 623–642, 2022, doi: 10.19101/IJATEE.2021.874887.
4. M. Kim, B. Jeon, H. Yoo, and K. Chung, "Abnormal Behavior Detection Using Deep-Learning-Based

Video Data Structuring," *Intell. Autom. Soft Comput.*, vol. 37, no. 2, pp. 2371–2386, 2023, doi: 10.32604/iasc.2023.040310.

5. X. Gansel, M. Mary, and A. Van Belkum, "Semantic data interoperability , digital medicine , and e-health in infectious disease management : a review," *Eur. J. Clin. Microbiol. Infect. Dis.*, vol. 38, no. 6, pp. 1023–1034, 2019.
6. S. D. Nguyen, M. Mimura, and H. Tanaka, "Slow-Port-Exhaustion DoS Attack on Virtual Network Using Port Address Translation," in *2018 Sixth International Symposium on Computing and Networking (CANDAR)*, 2018, pp. 126–132.
7. Z. Chiba, A. El Omri, N. Abghour, K. Moussaid, and M. Rida, "Smart Approach to Build A Deep Neural Network Based IDS for Cloud Environment Using an Optimized Genetic Algorithm," in *In Proceedings of the 2nd International Conference on Networking, Information Systems & Security. ACM*, 2019, p. (p. 60).
8. R. S. Sheikh, S. M. Patil, and M. R. Dhanvijay, "Framework for deep learning based model for human activity recognition (HAR) using adapted PSRA6 dataset," *Int. J. Adv. Technol. Eng. Explor.*, vol. 10, no. 98, pp. 37–66, 2023, doi: 10.19101/IJATEE.2021.876325.
9. H. B. Mahajan *et al.*, "Integration of Healthcare 4.0 and blockchain into secure cloud-based electronic health records systems," *Appl. Nanosci.*, vol. 13, no. 3, pp. 2329–2342, 2022, doi: 10.1007/s13204-021-02164-0.
10. Z. Zhu and M. Zhu, "A Novel Method for Anomaly Detection in the Internet of Things using Whale Optimization Algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 5, pp. 765–773, 2023, doi: 10.14569/IJACSA.2023.0140581.
11. P. Aggarwal and S. K. Sharma, "Analysis of KDD Dataset Attributes - Class wise for Intrusion Detection," *Procedia Comput. Sci.*, vol. 57, pp. 842–851, 2015, doi: 10.1016/j.procs.2015.07.490.
12. A. I. Newaz, A. K. Sikder, M. A. Rahman, and A. S. Uluagac, "HealthGuard: A Machine Learning-Based Security Framework for Smart Healthcare Systems," *2019 6th Int. Conf. Soc. Networks Anal. Manag. Secur. SNAMS 2019*, pp. 389–396, 2019, doi: 10.1109/SNAMS.2019.8931716.
13. S. Sengan, O. I. Khalaf, P. Vidya Sagar, D. K. Sharma, L. Arokia Jesu Prabhu, and A. A. Hamad, "Secured and privacy-based IDS for healthcare systems on e-medical data using machine learning approach," *Int. J. Reliab. Qual. E-Healthcare*, vol. 11, no. 3, 2022, doi: 10.4018/IJRQEH.289175.
14. G. Pachauri and S. Sharma, "Anomaly Detection in Medical Wireless Sensor Networks using Machine Learning Algorithms," *Procedia Comput. Sci.*, vol. 70, pp. 325–333, 2015, doi: 10.1016/j.procs.2015.10.026.
15. L. W. Ndungu, "Replication as a way to achieve interoperability in Healthcare," 2020.
16. A. M. Maina and U. G. Singh, "Why National eHealth Strategies Matter - An Exploratory Study of eHealth Strategies of African Countries," in *International Conference on Electrical and Electronics Engineering, ICE3 2020*, 2020, pp. 670–675. doi: 10.1109/ICE348803.2020.9122831.
17. A. Rejeb *et al.*, "The Internet of Things (IoT) in healthcare: Taking stock and moving forward," *Internet of Things (Netherlands)*, vol. 22, no. February, p. 100721, 2023, doi: 10.1016/j.iot.2023.100721.
18. J. Pool, S. Akhlaghpour, F. Fatehi, and A. Burton-Jones, "A systematic analysis of failures in protecting personal health data: a scoping review," *Int. J. Inf. Manage.*, vol. 74, p. 102719, 2024.
19. H. Zhang, B. T. Han, and Z. Tang, "Constructing a nationwide interoperable health information system in China: The case study of Sichuan Province," *Heal. policy Technol.*, vol. 6, no. 2, pp. 142–151, 2017.

20. F. Gong, G. Hu, H. Lin, X. Sun, and W. Wang, "Integrated healthcare systems response strategies based on the luohu model during the covid-19 epidemic in shenzhen, china," *Int. J. Integr. Care*, vol. 21, no. 1, pp. 1–7, 2021, doi: 10.5334/ijic.5628.
21. J. H. Weber and J. Ho, "Applying Bidirectional Transformations to the Design of Interoperable EMR Systems," *J. Healthc. Informatics Res.*, vol. 4, no. 2, pp. 138–150, 2020.
22. S. Higman *et al.*, "Designing interoperable health information systems using Enterprise Architecture approach in resource-limited countries: A literature review," *Int. J. Health Plann. Manage.*, vol. 34, no. 1, pp. e85–e99, 2019, doi: 10.1002/hpm.2634.
23. S. A. Mengiste, K. Antypas, M. R. Johannessen, J. Klein, and G. Kazemi, "eHealth policy framework in Low and Lower Middle - Income Countries ; a PRISMA systematic review and analysis," *BMC Health Serv. Res.*, vol. 23, no. 1, pp. 1–15, 2023, doi: 10.1186/s12913-023-09325-7.
24. P. Galetsi, K. Katsaliaki, and S. Kumar, "Big data analytics in health sector: Theoretical framework, techniques and prospects," *Int. J. Inf. Manage.*, vol. 50, no. 2020, pp. 206–216, Feb. 2019, doi: 10.1016/j.ijinfomgt.2019.05.003.
25. D. Wang, J. Ming, T. Chen, X. Zhang, and C. Wang, "Cracking IoT Device User Account via Brute-force Attack to SMS Authentication Code," in *Proceedings of the First Workshop on Radical and Experiential Security*, 2018, pp. 57–60.
26. J. Liu, T. Yin, M. Shen, X. Xie, and J. Cao, "State estimation for cyber-physical systems with limited communication resources , sensor saturation and denial-of-service attacks," *ISA Trans.*, 2018, doi: 10.1016/j.isatra.2018.12.032.
27. Y. Meidan *et al.*, "N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders," *IEEE Pervasive Comput.*, vol. 17, no. 3, pp. 12–22, 2018.
28. D. Santana, S. Suthaharan, and S. Mohanty, "What we learn from learning-Understanding capabilities and limitations of machine learning in botnet attacks," *arXiv Prepr. arXiv1805.01333*, 2018.
29. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization.," in *ICISSP*, 2018, pp. 108–116.
30. Z. A. Foroushani and Y. Li, "Intrusion detection system by using hybrid algorithm of data mining technique," *ACM Int. Conf. Proceeding Ser.*, pp. 119–123, 2018, doi: 10.1145/3185089.3185114.
31. C. Kombe, A. Sam, M. Ally, and A. Finne, "Blockchain Technology in Sub-Saharan Africa: Where does it fit in Healthcare Systems: A case of Tanzania.," *J. Health Inform. Dev. Ctries.*, vol. 13, no. 2, p. 1, 2019, [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=140226561&lang=pt-br&site=eds-live&scope=site
32. M. U. Tariq, "Revolutionizing health data management with blockchain technology: Enhancing security and efficiency in a digital era," in *Emerging Technologies for Health Literacy and Medical Practice*, IGI Global, 2024, pp. 153–175.
33. A. R. Khan, M. Kashif, R. H. Jhaveri, R. Raut, T. Saba, and S. A. Bahaj, "Deep Learning for Intrusion Detection and Security of Internet of Things (IoT): Current Analysis, Challenges, and Possible Solutions," *Secur. Commun. Networks*, vol. 2022, pp. 171–176, 2022, doi: 10.1109/ICIoT48696.2020.9089524.
34. W. Lin, S. Ke, and C. Tsai, "CANN : An intrusion detection system based on combining cluster centers and nearest neighbors," *Knowledge-Based Syst.*, no. February, pp. 1–9, 2015, doi: 10.1016/j.knosys.2015.01.009.

35. B. J. Radford, L. M. Apolonio, A. J. Trias, and J. A. Simpson, "Network Traffic Anomaly Detection Using Recurrent Neural Networks," *arXiv Prepr. arXiv1803.10769.*, pp. 1–7, 2018.
36. Y. Fei, C. Huang, C. Jinkun, M. Li, Y. Zhang, and C. Lu, "Attribute Restoration Framework for Anomaly Detection," *IEEE Trans. Multimed.*, p. 1, 2020, doi: 10.1109/TMM.2020.3046884.
37. Y. Zhong *et al.*, "HELAD: A novel network anomaly detection model based on heterogeneous ensemble learning," *Comput. Networks*, vol. 169, p. 107049, 2020, doi: 10.1016/j.comnet.2019.107049.
38. M. Rela, S. N. Rao, and P. R. Reddy, "Performance analysis of liver tumor classification using machine learning algorithms," *Int. J. Adv. Technol. Eng. Explor.*, vol. 9, no. 86, pp. 143–154, 2022, doi: 10.19101/IJATEE.2021.87465.
39. CIC, "Intrusion Detection Evaluation Dataset (CICIDS2017)," *Canadian Institute for Cybersecurity*, 2019. https://www.unb.ca/cic/datasets/ids-2017.html (accessed May 27, 2019).
40. S. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, "Assessing Generative Models via Precision and Recall for Learning Systems," in *32nd Conference on Neural Information Processing Systems*, 2018, no. NeurIPS, pp. 1–10. [Online]. Available: http://papers.nips.cc/paper/7769-assessing-generative-models-via-precision-and-recall.pdf
41. I. Ullah and Q. H. Mahmoud, "A two-level flow-based anomalous activity detection system for IoT networks," *Electron.*, vol. 9, no. 3, pp. 1–18, 2020, doi: 10.3390/electronics9030530.