# Protection the NFV Net-work Using the Random Forest Classification/ Original article

**Mustafa H. Taha*[1],**

**Ibtesam Jomaa Ibrahem[2],**

**Wafaa Waheeb Abdullah Ali[3]**

1 Department of Human Resources, General Directorate of Diyala Education, Diyala, Iraq,

2 Department of Computer Science, Diyala University President, Diyala, Iraq,

3 Department of Planning, General Directorate of Diyala Education, Diyala, Iraq,

# حماية شبكة (NFV) باستخدام تصنيف الغابة العشوائية

**مصطفى حسين طه[1]*،**

**ابتسام جمعة ابراهيم[2]،**

**وفاء وهيب عبدالله[3]**

1 قسم الموارد البشرية, المديرية العامة لتربية ديالى, ديالى, العراق

2 رئاسة جامعة ديالى, جامعة ديالى, ديالى, العراق

3 قسم التخطيط, المديرية العامة لتربية ديالى, ديالى, العراق

* mustafahusseintaha@gmail.com

## Abstract

A misuse attack is the most common and dangerous type of attack that could target NFV (Network Function Virtualization). In a misuse attack, the attacker attempts to consume the NVF environment's resources by sending a large amount of traffic. To protect the NFV environment, an early and accurate misuse attack detection system has been proposed based on NFV and Random Forest Classifier. The proposed model starts with importing the dataset and analyzing it then pre-processing and feature selection using a PSO (particle swarm optimization) Test algorithm, classification is based on the most common and efficient machine learning technique, which is Random Forest and Niave Bayse used. Finally, in order to test and evaluate the performance of the proposed model, the KDD (knowledge discovery in databases) dataset is utilized. The proposed system outperformed the most comparable previous research in terms of performance, as indicated by the results, achieving an accuracy rate of 99.98% for Random Forest and 99.5% for Niave Bayse.

**Keywords: Network Function Virtualization (NFV), Machine Learning ML, Random Forest (RF), Niave Bayse, PSO Test algorithm, KDD.**

## المستخلص

هجوم سوء الاستخدام هو النوع الأكثر شيوعًا وخطورة من الهجمات التي يمكن أن تستهدف NFV المحاكاة الافتراضية لوظائف الشبكة . في هجوم سوء الاستخدام، يحاول المهاجم استهلاك موارد بيئة NVF عن طريق إرسال كمية كبيرة من حركة المرور. لحماية بيئة NFV، تم اقتراح نظام مبكر ودقيق للكشف عن هجمات سوء الاستخدام استنادًا إلى NFV و Random Forest Classifier. يبدأ النموذج المقترح باستيراد مجموعة البيانات وتحليلها ثم المعالجة المسبقة واختيار الميزات باستخدام خوارزمية اختبار PSO (تحسين سرب الجسيمي)، ويعتمد التصنيف على تقنية التعلم الآلي الأكثر شيوعًا وكفاءة، وهي Random Forest وNiave Bayse المستخدمة. وأخيرا، من أجل اختبار وتقييم أداء النموذج المقترح، تم استخدام مجموعة بيانات KDD (اكتشاف المعرفة في قواعد البيانات). تفوق النظام المقترح على الأبحاث السابقة الأكثر مقارنة من حيث الأداء، كما أشارت النتائج، حيث حقق معدل دقة 99.98% لـ Random Forest و 99.5% لـ Niave Bayse .


**الكلمات المفتاحية : شبكة NFV, التعلم الالي، تصنيف الغابة العشوائي**

## 1. Introduction

The most critical part that must be considered when developing systems, especially those that must connect to the internet, is information security. The identity risk management sector is where it belongs [1]. The usual approach is to prevent or reduce the risk of improper data usage, leakage, destruction, deletion, corruption, alteration, review, tracking, or devaluation. The main goal of Information management is to maintain data's as the confidentiality, integrity and availability, and or (CIA triad) supporting the efficiency of business [2]. Over the past decade, massive cyber-attacks against major organizations have multiplied. Among the most devastating are ransomware, which aims at encrypting the data stored on the system until the target organization pays a ransom.

At times, security threats are on the rise. A computer system that has a flaw or weakness, security tactics, internal controls, and planning, A framework's security policy can be compromised by a web vulnerability, which could lead to data theft or modification. SQL injection and cross-site scripting (XSS) attacks, as well as broken authentication and session management, are the most common types of attacks. Weak or broken authentication mechanisms may allow attackers to circumvent authentication and gain unauthorised access to the application [3]. Resources on the NFV network are vulnerable to abuse, particularly those that only a small number of users or cannot share. In these attacks, the attacker takes resources away from other users and uses them for their purposes without sharing them with other users. These attacks typically cause bottleneck issues, which cause service delays or even the interruption of NFV network services [4]. Network Function Virtualization (NFV) and Software Defined Networking (SDN) are

recognized as key technological paradigms in modern networking systems for efficiently handling emerging challenges in managing complex networks and there are unpredictable traffic patterns in various scenarios, including Internet of Things, cloud networking, and mobile data traffic in fifth generation (5G) and beyond networks [5]. Middle box functionality is progressively being virtualized and offered in software with the development of software-defined networking (SDN) and virtualisation of network functions (NFV), which can reduce energy consumption, resource use and service provider operating costs [6]. The typical traditional network is rigid and stiff. The Internet's rapid expansion has brought about challenges, including heterogeneity, scalability, and interoperability. A fundamental approach to virtualization in communication networks is Network Function Virtualization (NFV) [7]. Virtual networks that provide services through virtual parts of virtual machines are known as Network Function Virtualization (NFV). It is easy to implement and up-to-date. NFV's low costs are due to the sharing of resources. Attacks are not uncommon for NFV, just like other networks. Misuse attacks are the most common attack in NFVs because all parts share the same resources, the attack's use of one or more resources that affect all parts of the NFV is particularly noteworthy.

A neural network-based detection system was proposed by them for botnets in NFV. The system was formed using data sets that were collected through traditional networks. Based on the study, their proposed detection system could detect up to 99% of botnet attacks. Various types of attacks can target Network Function Virtualization (NFV) in different forms and destinations. Network Function Virtualization (NFV) could be targeted by various types of attacks in different forms and destinations. Moreover,

numerous studies have been established to tackle this type of attack, but no one has come up with the best solution.

The main point of this work is explained in the following steps:

1) Design and implementation of detection and diagnosis of misuse attacks in NFV networks with higher accuracy based on the Random Fours algorithm and Niave Bayse of machine learning.
2) The ideal categorization of normal and offensive behaviour in the intrusion detection model is achieved by combining NFV with a random forest and Niave Bayse classifier.
3) To improve the intrusion detection model, a frequent pattern mining-based adaptive online update strategy is added to the random forest and Niave Bayse classification and detection process is able to adapt to dynamic changes in the network environment and has a high detection rate for known and unidentified attacks.
4) The accuracy criteria are used to compare the performance of machine learning algorithms, knowing the best algorithm for diagnosing and distinguishing the misuse attack requires accuracy.

Our presentation presents an adaptive intrusion detection model that is based on NFV and the random forest algorithm to reduce the impact of dynamic changes in the network environment on the precision of the pattern for detecting abusive attacks, better extract effective data characteristics and enhance the detection performance of the abusive attack detection model. This work aims to propose a system capable of detecting early and accurate NFV misuse attacks using the most efficient machine learning technique the random forest .the intrusion detection model can adapt to dynamic changes

in the network environment by introducing frequent pattern mining in the detection process.

Niave Bayse, a machine learning technique, has been added to enhance the detection of misuse attacks. To test and evaluate the performance of the proposed model, the KDD dataset is utilized.

The remainder of this work divided into the following sections. Details on related theories are presented in Section 2. The suggested adaptive misuse attack detection model and associated techniques are described in Section 3. Experimental results are presented in Section 4. Section 5 ends our study by discussing the most important points of the system and comparing our proposed model with current methods

For the purpose of swift and verifiable identification of misconfigurations breaching security characteristics in NFV, [8] proposed a new methodology called MLFM which combines machine learning (ML) efficiency with rigorous formal methods (FM). The heart of our approach is an iterative teacher-learner interaction where the teacher (FM) can progressively (over several iterations) as training data, provide more representative verification findings, and the learner (ML) can use these results to gradually generate more accurate ML models.

In [9] covers the following goals: establishing and examining the two classifiers that are most frequently used in IDS implementation (KNN and Bayes), assessing the performance of each classifier separately, and modeling a hybrid classifier based on the advantages of the two classifiers. A quantitative methodology for data collection and analysis was used in this investigation. The NSL-KDD and the original KDD 1999 datasets were used in the investigation. Evaluated the developed techniques using traffic workloads

and virtualized networked environments. While coefficients and signal shifts were employed to complete period detection, SVM was used to identify cycle numbers.[10] Created an intrusion detection system by employing a recurrent neural network. The proposed mechanism has three stages: data collection, characteristic extraction, and the deep threat classifier. The results of the studies show that Naive Bayes, K-Nearest Neighbor, Random Forest and Decision Tree are ineffective compared to the best precision of 98.18% according to the 10-fold cross-validation analysis.

In [11], network intrusion detection (NID) systems have made extensive use of data-driven methodologies. The way the datasets are gathered, however, has led to several difficulties. In comparison to regular traffic, the majority of attack classes in network intrusion datasets are regarded as the minority, and many datasets are gathered using virtual machines or other simulated settings as opposed to real-world networks. By fitting models as random forests or support vector machines to non-representative "sandbox" datasets, these problems reduce the performance of intrusion detection machine learning models.

## 2. Related Theories

In the field of our research, we relied on several modern techniques and methods to build the proposed model which are listed as follows:

### 2.1 Network Function Virtualization (NFV)

Telecommunication service providers (TSPs) have historically been forced to implement physical patented equipment for each network purpose, resulting in high construction costs and constraints when scaling a network.

This is not the case in the case of NFV [12].The role of a Network virtualization is a cloud infrastructure networking implementation that virtualizes network services that can be chained together to construct a virtualized communication infrastructure. This allows a shared, flexible, and energy-efficient network ecosystem to be developed. TSPs are referred to as "NFV providers" in this situation. As shown in Figure 1, the European Telecommunications Standards Institute (ETSI) has established an NFV structure with three key components [12, 13].

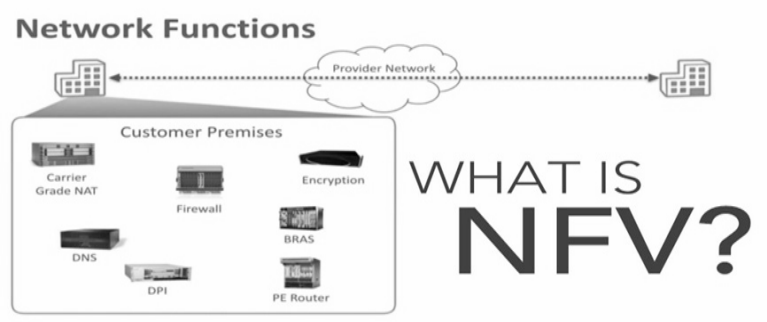



**Figure 1. General architecture of NFV [13].**

## 2.2 Exploratory data analysis (EDA)

The primary purpose of exploratory data analysis is to examine what the data can reveal outside of formal modelling or hypothesis-testing tasks. The analysis of data sets is assisted by EDA, which focuses on four essential elements to highlight their statistical characteristics: the distribution's shape, the presence of outliers, and the measures of central tendency (including the mean, mode, and median) are all taken into account, and the measures of spread (which include the standard deviation and variance). Data scientists evaluate the data and derive important insights using EDA techniques.

Effective EDA in this area also heavily relies on the abilities and domain expertise of data scientists [15].

## 2.3 Feature Selection Using PSO

James Kennedy and Russell Eberhard created particle swarm optimization (PSO). It was created using a straightforward idea inspired by the movement of fish schools and bird flocks. It was created following several computer simulation-based interpretations. PSO uses several agents (particles) that come together to form swarms. In search of the best solution, this swarm is moving around in the search area. To match its own and other particles' flying experiences, it modifies each particle's "flying" in the search space.

PSO started with randomly created particles and their velocities, which represent the speed of the search. The particles are then assessed for fitness. Two major tests are conducted following this examination. The first test, known as personal best, compares a particle's experience with itself (pbest). The second test contrasts a particle's fitness with the overall swarm experience. The phrase "global best" (gbest) for the best particle is saved as a result of running these two tests. The termination criteria are then satisfied [16]. A binary string that corresponds to the feature selection case determines each particle's position. The formulas below determine the frequency of updating each particle (1, 2, and 3):

$$v_{pd}^{new} = w\,x\,v_{pd}^{old} + c_1\,x\,rand_1\,x\left(pbest_{bd} - x_{pd}^{oid}\right) + c_2\,x\,rand_2\,x\left(gbest_d - x_{pd}^{old}\right) \quad (1)$$

$$s\left(v_{pd}^{new}\right) = \frac{1}{1+e^{-v_{pd}^{new}}} \quad (2)$$

$$If\ (rand < sv_{pd}^{new}))\ then\ x_{pd}^{new} = 1; else\ x_{pd}^{new} = 0 \quad \textbf{(3)}$$

Where $v_{pd}^{new}\ and\ v_{pd}^{old}$ are the particle velocities, $x_{pd}^{old}$ is the current particle position (solution), and $x_{pd}^{new}$ is the updated particle position (solution). The values $pbest_{bd}\ and\ gbest_{d}$ defined as stated above. The two factors $rand_1\ and\ rand_2$ are random numbers between (0, 1).

## 2.4 Classification

The process of predicting the goal class of each data sample includes utilizing a supervised learning approach that determines the classes to be used beforehand. When utilizing classification algorithms, data attribute values must be utilized to determine class definitions [15]. The first stage in the classification phase is called preparation, which is necessary to construct a model for classifying new data. The model is used to forecast the class mark of the new data in the second step, which is referred to as classification. To describe a previously established set of data classes, a classifier is built during the training phase. The classification algorithm creates the classifier by acquiring knowledge from training data and related class labels. This phase can be regarded as a learning phase where the function y = f(x) can predict the associated class label for the given data, and the prediction's accuracy is the first thing to be calculated. If you use training data for computing accuracy, it will be optimistic (the classifier leads to overflow the data), so in this state, a test set is employed where the data is not dependent on the training set. If the accuracy of the classifier is acceptable when using the testing set, it can be used in the future for the data whose class label is unknown. In the controlled classification, the bulk of the initiative can be

rendered prior to the final classification. Although supervised classification is far more reliable than unsupervised classification, misclassifications can be identified because it is heavily reliant on teaching. The monitoring of classification involves careful attention to the creation of training details. Classification outcomes would be bad if training data were poor or not representative. As a consequence, controlled labeling usually takes more resources and time than unsupervised [16].

### 2.4.1 KDD Datasets

The capacity to select the best option from a range of choices to address a specific issue or accomplish a set of goals is known as decision-making. Strategic, tactical, and operational decisions may often be made in any organization or system. In any company, decision-making processes can be supported in a variety of ways. A well-known strategy employed in businesses is the decision support system (DSS). A decision support system (DSS) is an information system that supports the management, operating and planning levels of an enterprise, organization, or business and addresses a variety of issues to help decision-makers.

A well-designed DSS can assist decision-makers in gathering information from numerous sources and in making judgments by resolving issues [17]. Data mining is the main KDD sub process that uses various algorithms and techniques on databases to create predictive models and extract useful and instructive patterns. The maternal health and child immunization databases can be utilized to identify previously unidentified patterns and knowledge, which can then be used to create more effective and efficient decisions for the management of healthcare. The fundamental goal of integrating decision-

making with KDD is to reinforce and enhance DSS through the models created for KDD, especially in situations where a significant amount of historical data is accessible for knowledge discovery [18].

### 2.4.2 Random Forest Classifier

All satellite scenes from each year were classified using the random forest method [19]. This method, which is a variation of the classification tree algorithm, is used in ensemble learning techniques. Based on tree-wise randomly selected samples and subsets of the training data, individual decision trees are generated automatically by the classification tree algorithm. Many classification trees are created for a random forest model, and the classification outcome is determined by a vote of each tree. A random sample of the training data set is used to create each classification tree individually using a unique learning method. The best split is carried out at each node of classification trees using random selections of the predictor variables. Every tree grows to its maximum potential under the control of the user-set node size. There are a number of significant adjustable factors that must be set before the random forest classification can be used. The main parameters are the depth of each tree in the forest, the number of classification trees to run in the forest (ntree), and the number of randomly chosen variables to utilize in each tree's construction (mtry) [20].

### 2.4.3 Evaluation Methods

A confusion matrix is a form of visualization method that is widely used to verify the accuracy of classifiers in classification [21]. Accuracy is a parameter that tests a classifier's capacity to render a correct diagnosis [22]. The accuracy of the equation is seen in equation 4.

$$Accuracy = (\frac{TP + TN}{TP + TN + FP + FN}) \times 100 \quad \textbf{(4)}$$

TP and TN are synonymous with True Positive and True Negative, which suggests that the proportions of positive and negative conditions were properly identified. Also, FP means false positive, which refers to all the negative cases that were falsely classified as positive, and False Negative, or FN, is the term for all positive cases that were misclassified as negative [23].

## 3. The Proposed System

A machine-learning algorithm is using in the proposed system to achieve a higher accuracy rate for detecting misused attacks. The input into the system is a KDD data set, and the output is the attack rating. The proposed work consists of several main stages, namely loading the KDD dataset, preprocessing the encoded dataset using the One Hot Encoding technique, selecting the most important feature from the dataset using the PSO technique, and classifying the results using Random Forest and Niave Bayse. The diagram of the proposed model's general blocks is shown in Figure 2.

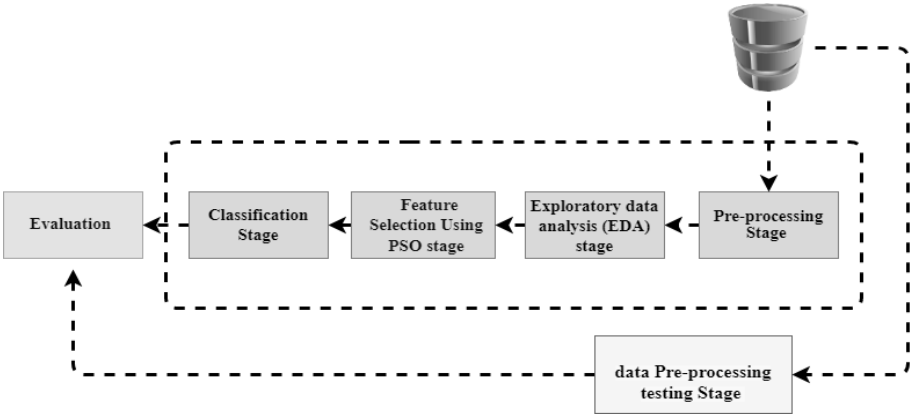



**Figure 2. The Proposed Model's General Block Diagram.**

### 3.1 Pre-processing Stage

Data preprocessing is an important stage that has been taken into consideration before the further processing stage. In this work, One Hot Encoding is used for dataset preprocessing. In order to convert categorical variables to numerical values in an interpretable format, the most well-known technique, one-hot encoding, has been, in addition to the job of quantifying categorical data in machine learning. Additionally, One Hot Encoding generates a vector whose length may equal the number of feature categories in the test dataset. Except for the ith component, which is given a value of 1, all components of this vector are assigned the value 0 if the data point falls into the ith category. Following categories logically in terms of numbers is made possible by this. However, it is quickly fixed by the statement that approximately one hot encoded has exactly one position in its array that is labeled as a (1).

### 3.2 Exploratory data analysis (EDA)

Techniques for data visualization and analysis are widely employed. Following is a discussion of these methods:

#### I. Data Exploration

That is the first step in analyzing the data. This is where we can learn more about the content and characteristics of the data set. The size of the data is exposed. Locating the missing value of the data is possible. We can see if there are any potential connections between the data. By using tabular data and comprehending its properties, data visualization is accomplished.

### II. Data Cleaning

To complete the process, one must identify inaccurate data, eliminate unnecessary information from the data, and replace the incorrect data. The real data cleaning process is error elimination and data validation. Cross verification of data can help you find and correct errors. Validating the data is a way to resolve the problem.

### III. Model Building

To describe the behavior of a variable, a statistical model or machine learning model is utilized. It is possible to have both supervised and unsupervised models. A classification or regression model can be used to obtain the results. The outcome can be seen by using a model. After that, the model must be evaluated.

### IV. Present Result

By using charts, graphs, and tables, we may view large amounts of complex data. Graphs can help people digest information. The idea can be explained with a simple approach. The areas that need improvement can be identified by it. It effectively clarifies the component [24].

## 3.3 Feature selection Using PSO algorithm

The population of the particles in the first swarm is typically spread randomly over the search space. In every iteration, both "best" values, $p_{best}$ and $g_{best}$, are used to update each particle. In the issue space, the coordinates of each particle are recorded that correspond to the best solution (fitness) it has found so far. The name $p_{best}$ is given to this fitness value that is kept. The

highest value is a global "best" value and is referred to as $g_{best}$ when a particle uses all the population as his topological neighbor. . The pseudo-code of the proposed PSO model algorithm is given in several steps, as follows:

Step1: find variance γ2 of the data by using standard deviation γ.

Step2: Determine the null and alternate hypothesis where N0: no difference in variances and $N_a$ : difference in variances

Step3: Find $F_{calc}$ using equation (1), $F_{calc} = \gamma_1^2 / \gamma_2^2$ , where $F_{calc}$ = Critical F-value, $\gamma_1^2 \& \gamma_2^2$ = difference of the two samples.

Step 4: Find the degrees of freedom of the two samples using equation (2), $d_f = n_s - 1$, where: $d_f$ = Degrees of freedom of the sample, $n_s$ = Sample size.

Step 5: Find $F_{table}$ value using d1and d2, obtained in Step4 from the F-distribution table using learning rate a = 0.03, $d_1 = d_f$ of the bigger sample with numerator variance,$d_2 = d_f$ of the smaller sample with denominator variance.

Step 6: Interpret the results using $F_{calc}$ and $F_{table}$. where Max referred to the best value the feature takes , Min referred to lowest value the feature takes ,range = $|Max - Min|$ , D represents discrete values of feature ,and C represents continuous values of feature.

### 3.4 Classification Stage

The Proposed Misuse Attack Detection Based on Data mining in NFV

### 3.4.1 Random Force classification algorithm

The dataset was divided in the proposed model into two parts using the Random Force classifier. 70% of the dataset is in the first section, which

is also, where the suggested system is trained. The second half, which makes up 30% of the dataset, is utilized to test the suggested system. Then, assess the suggested system's accuracy in classifying the abuse feature algorithm, which is offered in multiple steps as follows: (Matrix Data Set as Input; Best Tree as Output)

1: in the first step the samples will be selected from the dataset.

2: After construction, a decision tree for all samples will provide a prediction result.

2.1 evaluate the previous entropy.

2.2: evaluate the information Expanded for the attribute.

2.3: evaluate the attribute value then partition the data set based on these values.

2.4: Repeat Steps 1- 3 per Tree using the relevant partition.

3: in the last step the voting for every expected result will be completed.

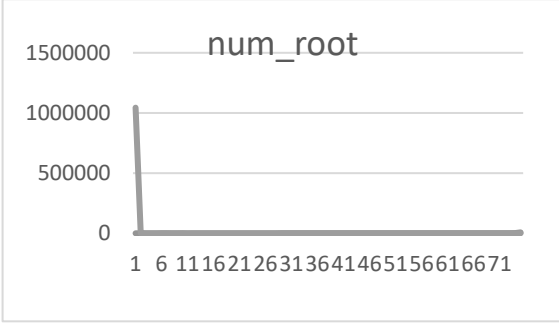
### 3.4.2 Testing Dataset

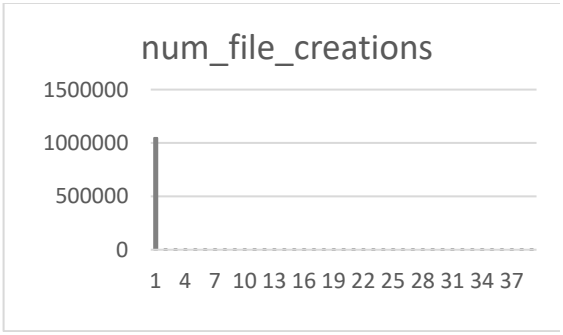
The KDD dataset contains approximately one hour of anonymized traffic traces from the misuse attack and Distributed Denial-of-Service (DDoS) attack. The server is able to connect to the internet and its computer capabilities by using all of the network bandwidth. This attack aims to prevent users from accessing the targeted server. The KDD dataset contains more than 4 million requests and connections. Table 1 shows that each connection had 10 fields of information. It is also the analysis value in Table 2. This dataset has 972,517 connections from misuse attacks, and the remaining belong to DDoS traffic and normal traffic.

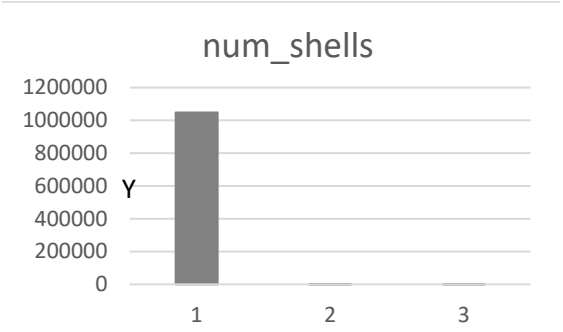
**Table 1. Misuse Dataset 42 Features.**

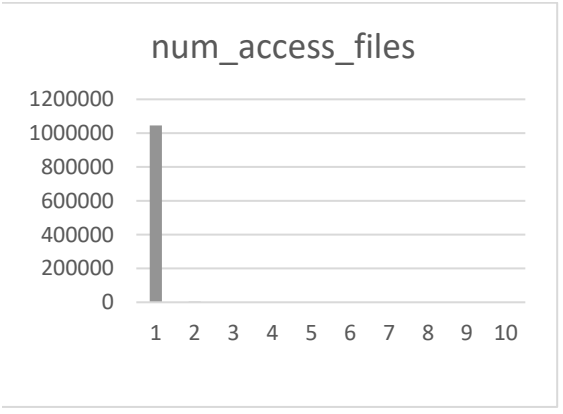
| No. class | Feature name | Total number of attack |
|---|---|---|
| 1 | Root_ num | The root in the connection is done as a number of operations or the number of 'root' accesses |
| 2 | File_creations_num | The file creation operations were performed in the connection of the number |
| 3 | Shells_num | Shell prompts of the number |
| 4 | Files_access_num | Control the number of operations performed on accessing files. |
| 5 | Num_outbound_cmds | Commands that are sent out in an FTP session |
| 6 | Is_hot_login | If the login is on the 'hot' list, like root or admin, then it will be 1; otherwise, it will be 0 |
| 7 | Is_guest_login | 1 if the login is a “guest” login; 0 otherwise |
| 8 | Count | The number of connections completed in the last two seconds with the same host as the actual connection destination. |
| 9 | Srv_count | The number of connections made to the same service (port number) within the preceding 2 seconds as the current connection |
| 10 | Serror_rate | The proportion of connections among those aggregated in count that have triggered the flags s0, s1, s2, or s3 (23) |

**The analysis value of 10 Features of Dataset explain in figure 2**

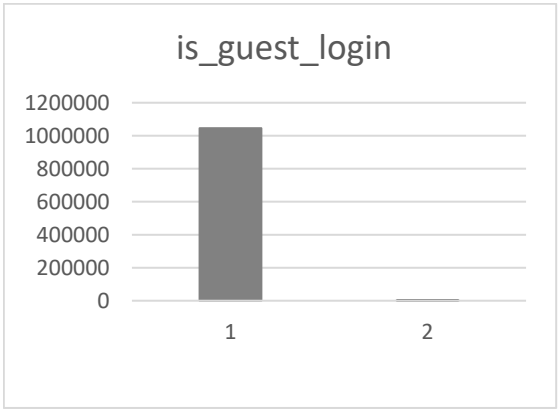


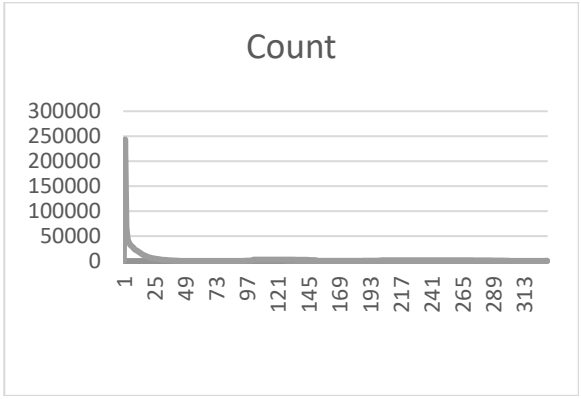

num_file_creations
1500000
1000000
500000
0
1 4 7 10 13 16 19 22 25 28 31 34 37

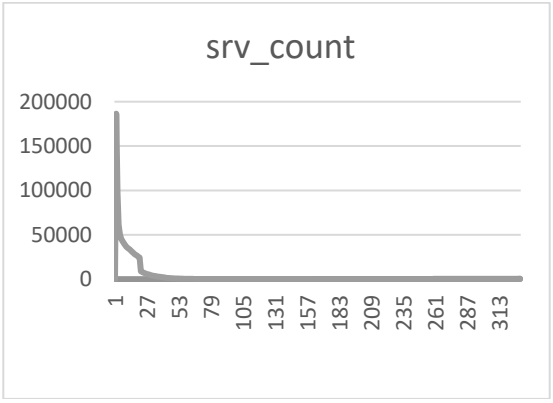

num_shells
1200000
1000000
800000
600000
400000
200000
0
Y
1
2
3

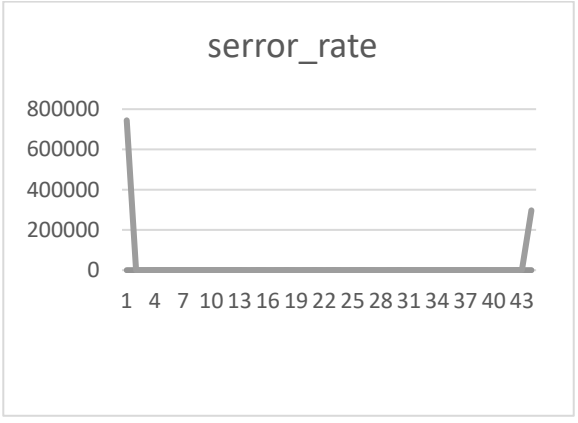

num_access_files
1200000
1000000
800000
600000
400000
200000
0
1 2 3 4 5 6 7 8 9 10


is_host_login
1200000
1000000
800000
600000
400000
200000
0
1
2


is_guest_login
1200000
1000000
800000
600000
400000
200000
0
1
2


Count
300000
250000
200000
150000
100000
50000
0
1
25
49
73
97
121
145
169
193
217
241
265
289
313



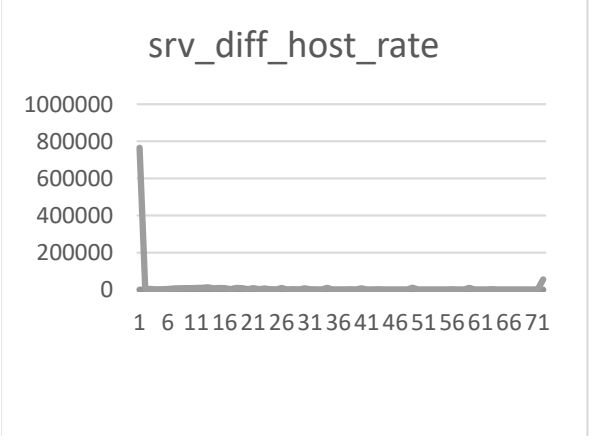




**Figure 3. Analysis value of 10 Features of Dataset.**

# 4.The Results

This section, results from the proposed system that has been designed to defend against misuse attacks targeting NFV are presented and discussed. The outcomes of each stage of the proposed model are shown below:

## 4.1 Import Dataset

The first step proposes a system load data set and divides the dataset into parts (70% training and 30% testing), where the dimensions of the training set are (211979, 50) and the dimensions of the test set are (325498, 33).

## 4.2 Data Preprocessing using One Hot Encoding Technique

Furthermore, according to the previous chapters' discussions, Hot Encoding is employed to convert all features into binary. Here is the One-Hot encoding procedure:

- Input [matrix of integers]
- Processing [Denoting values of features]
- Output [Sparse matrix, one feature]

In case of we assume the input features take values in range (0, n_ values). The first step in one hot encoding technique is to identify categorical features, as illustrated in Table 2. There are only four features that have symbols or text, which are protocol type, flag, service, and label.

**Table 2. First step of one hot encoding algorithm.**

| # | Feature | Number of categories | |
|---|---|---|---|
| | | training | testing |
| 0 | protocol type | 3 | 3 |
| 1 | Service | 70 | 65 |
| 2 | Flag | 11 | 10 |
| 3 | label' | 23 | 2 |

However, quite a range of factors has been taken into consideration, such as Data Quality and Quantity, Data Availability, and Gaps in the Data.

## 4.3 Feature Selection using PSO-Test

After using the feature one hot encoding and normalization technique, the number of features is Train: Dimensions of Misuse (1431357, 117) and Test: Dimensions of Misuse (1431357, 117). Table 3 displays the results of selecting features.

**Table 3. PSO F-test results for feature selection.**

| |
|---|
| [logged_in, |
| Count, |
| Serror_rate, |
| Stv_serror_rate, |
| Same_ Stv_rate, |
| Dst_host_stv_Count, |
| Dst_host_ Same_ Stv_rate, |
| Dst_host_ Serror_rate, |
| Dst_host_ Stv_serror_rate, |
| Service_private, |
| Flag_S0, |
| Flag_SF] |

## 4.4 Results of Classification

The training stage's misused datasets contributed to 70% of the data in the classification stage, and the testing stage was where 30% came from. The proposed system uses the Random Fours classification algorithm, and the results are based on the accuracy ratio, so this section illustrates the results of the random forest classification (4.4.1) and the analysis and discussion in the subsections (4.4.2).

### 4.4.1 Results of Random Forest Classification

This section will show the results of the random forest classification algorithm as illustrated in algorithm 2, based on values of accuracy rate using equation (11), with all cases provided in the proposed system as shown in figure 4.

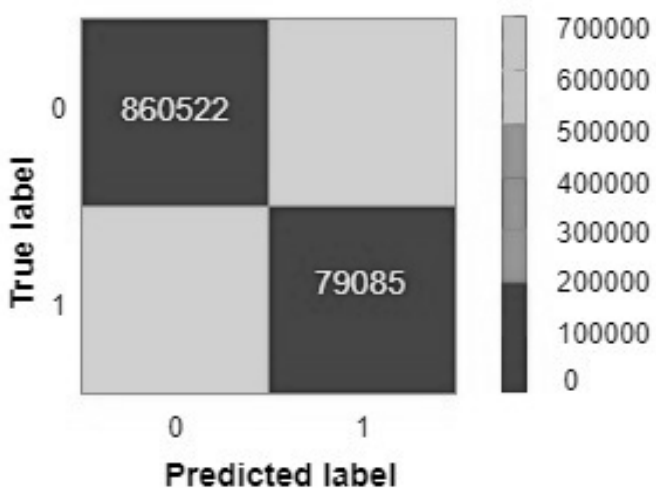



**Figure 4. Results of accuracy rate for Random forest classification.**

## 5. Analysis and discussion

The control and prevention of incoming abnormal traffic are too difficult a task that could target information security developers. Normally, an accurate and rapid detection system for controlling the attack traffic is required. Even after many methods were used and developed to identify and control the attack traffic. Most of these are not detectable with the best accuracy. Furthermore, NFV may be the victim of several types of attacks, and flood attacks are the most frequent and the most dangerous. Flooding attacks are divided into two main types: misuse and DDoS. Both of their attempts to overwhelm the NFV with traffic and consume system resources failed. This work, emphasizes misuse attack detection in the early stages.

The proposed model used the most common and effective machine learning techniques, which is random forest and Niave Bayse. The proposed system has been tested and evaluated by using the KDD dataset.

In the literature, the majority of methods indicate categorization accuracy ranges between 60% and 92%. Those with better accuracy were typically tested using a particular attack traffic type. When other traffic kinds are introduced or other sorts of attacks, such as DDoS, are taken into consideration, those solutions cannot continue to function as well.

Additionally, the proposed model based on uses the KDD benchmarking dataset to discriminate, extract, and identify the elements of the misuse attack, as well as to distinguish its traffic from other forms of traffic. The design, implementation, testing, and evaluation of effective machine learning approaches that been used for intrusion detection systems make up the second part. These formulas belong to the Random Forest family and Niave Bayse. Furthermore, the machine-learning algorithm's performance has

been tested and assessed using the KDD dataset. The dataset is divided into two segments, 30% of the data is used for testing, while 70% is used for training.

According to the data in Table 4, the proposed method detects more misuse attacks in the KDD dataset than other methods in multi-classification task detection. The suggested method's detection accuracy is higher than the most recent techniques; it is capable of adapting to changes in the network environment, Regardless of whether it is a two-class or multi-class.

**Table 4. Compare our Proposed Model with previous Work.**

| Model | Techniques | Accuracy% |
|---|---|---|
| Goel *et al.* [22] | Classification-Based Association (CBA) | 97.4 |
| Papamartzivanos *et al.* [23] | Self-adaptive and autonomous misuse of IDS | 84 |
| Wen-Tao Hao *et al.*[25] | convolutional neural network (CNN) and C5.0 classifier | 98 |
| Shukla *et al.* [26] | opposition self-adaptive grasshopper optimization | 99.35 |
| Bhavsar *et al.*[27] | a system for detecting intrusions | 98 |
| Akgn *et al.*[28] | The Long Short-Term Memory (LSTM) model | 99.2 |
| Sambagi *et al.*[29] | Utilized information gain and regression analysis to learn the ensemble of feature selection | 97.86 |
| Sahoo *et al.*[30] | Linear regression, KNN, Nave Bayes, Decision Tree, Random Forest, SVM, and ANN are all available | 98.64 |
| Mustapha *et al.*[31] | The IDS can be improved by adding another LSTM model that is responsible for blocking adversarial samples | 91.75 |
| our proposed model | Random Forest<br>Niave Bayse | 99.98<br>99.5 |

## 6. Conclusions

The suggested misuse attack detection model improves the ability of the misuse attack detection model in a Network Function Virtualization network to be able to adapt to the dynamic changes in the network environment. It is based on an NFV and a Random Forest, Niave Bayse classifier. Many algorithms are covered in this work to secure and protect the NFV against the most common and dangerous attacks that could target it.

The proposed system is applicable in NFV architecture, Cloud Computing, IoT environment, Web server and employed to discloser the Misuse attack traffics features that mislaid among the other traffics and features. The main focus of this work has been as follows:

1) This work proposed an early and accurate system that has the ability to defense against Misuse attack targeting.
2) The great challenge of this work is to distinguish between Abuses trafficking, which mislead and normal trafficking.
3) Random Forest and Niave Bayse were used to base the proposed system on the most effective machine learning algorithms.
4) The dataset is the most important part of testing and evaluating the performance of the proposed system. The proposed system's performance was tested and evaluated using the KDD dataset in this work.
5) According to experimental findings, the machine-learning algorithms has a strong detection performance in terms of detection accuracy, since 99.9% accuracy is achieved for Random Forest and since 99.5% for Niave Bayse .The suggested method's detection accuracy is higher than the most recent techniques,

whether they are two-class or multi-class, and it can adapt to changes in the network environment.

6) The proposed system's results are compared with related work, and the best accuracy is recorded with our proposed system.
7) The KDD dataset was used to test and evaluate the proposed system based on performance metrics of accuracy.
8) The proposed model is capable of defending against the most common and dangerous attack, which could be NFV, which is a misuse attack.

## References

[1] R. Diesch, M. Pfaff, & H. Krcmar" A comprehensive model of information security factors for decision-makers" Computers & Security, Vol. 92, 2020.

[2] A. Tchernykh, U. Schwiegelsohn,E.G. Talbi & M. Babenko, "Towards understanding uncertainty in cloud computing with risks of confidentiality, integrity, and availability" Journal of Computational Science, Volume 36, September 2019, 100581.

[3] A. A. Ahmed & N. B. Al Dabbagh," Web Attacks and Defenses", Journal of Education & Science, Vol, 32, No: 2, 2023 (114-127.

[4] N. Alhebaishi, L. Wang & S. Jajodia," Modeling and Mitigating Security Threats in Network Functions Virtualization (NFV)", IFIP Annual Conference on Data and Applications Security and Privacy (pp. 3-23). Springer, June , 2020.

[5] F. Paganelli, P. Cappanera & G. Cuffaro, " Tenant-defined service function chaining in a multi-site network slice", Future Generation Computer Systems, Volume 121, August 2021, Pages 1-18.

[6] S. Cherrared, I. Sofiane, F. Eric, G. Gregor, & G. B. Y. Imen, "A survey of fault management in network virtualization environments: Challenges and solutions." IEEE Transactions on Network and Service Management 16, no. 4 (2019): 1537-1551.

[7] I. Alam, K. Sharif, F.Li , Z. Latif, M. M. Karim, N.Nour, S. Biswas, & Y. Wang, "IoT virtualization: A survey of software definition & function virtualization techniques for internet of things." arXiv preprint arXiv, 2019, 1902.10910.

[8] Oqaily, A., Jarraya, Y., Wang, L., Pourzandi, M., & Majumdar, S. (2022). MLFM: Machine Learning Meets Formal Method for Faster Identification of Security Breaches in Network Functions Virtualization (NFV). In European Symposium on Research in Computer Security (pp. 466-489). Springer, Cham.

[9] Saeed, M. M. (2022). A real-time adaptive network intrusion detection for streaming data: a hybrid approach. Neural Computing and Applications, 34(8), 6227-6240.

[10] Latif, S., Zou, Z., Idrees, Z., & Ahmad, J. (2020). A novel attack detection scheme for the industrial internet of things using a lightweight random neural network. IEEE Access, 8, 89337-89350.

[11] Chou, D., & Jiang, M. (2020). Data-driven network intrusion detection: A taxonomy of challenges and methods. arXiv preprint arXiv:2009.07352.

[12] Alnaim, A., Alwakeel, A., & Fernandez, E. B. (2019, August). A Misuse Pattern for Compromising VMs via Virtual Machine Escape in NFV. In Proceedings of the 14th International Conference on Availability, Reliability and Security (pp. 1-6).

[13] Alwakeel, A. M., Alnaim, A. K., & Fernandez, E. B. (2019, May). Toward a Reference Architecture for NFV. In 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS) (pp. 1-6). IEEE.

[14] Fagroud, F. Z., Ajallouda, L., Toumi, H., Achtaich, K., & El Filali, S. (2020). IOT search engines: exploratory data analysis. Procedia Computer Science, 175, 572-577.

[15] Patel, H., Guttula, S., Mittal, R. S., Manwani, N., Berti-Equille, L., & Manatkar, A. (2022, August). Advances in exploratory data analysis, visualisation and quality for data centric AI systems. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 4814-4815).

[16] Almomani, O. (2020). A feature selection model for network intrusion detection system based on PSO, GWO, FFA and GA algorithms. Symmetry, 12(6), 1046.

[17] Shastri, Sourabh, and Vibhakar Mansotra. "Kdd-based decision making: a conceptual framework model for maternal health and child immunization databases." Advances in computer communication and computational sciences. Springer, Singapore, 2019. 243-253.

[18] Belgrana, F. Z., Benamrane, N., Hamaida, M. A., Chaabani, A. M., & Taleb-Ahmed, A. (2021, January). Network intrusion detection system using neural network and condensed nearest neighbors with selection of NSL-KDD influencing features. In 2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS) (pp. 23-29). IEEE.

[19] Dashpurev, B., Bendix, J., & Lehnert, L. W. (2020). Monitoring oil exploitation infrastructure and dirt roads with object-based image analysis and random forest in the Eastern Mongolian Steppe. Remote Sensing, 12(1), 144.

[20] Noshad, Z., Javaid, N., Saba, T., Wadud, Z., Saleem, M. Q., Alzahrani, M. E., & Sheta, O. E. (2019). Fault detection in wireless sensor networks through the random forest classifier. Sensors, 19(7), 1568.

[21] Moorthy, U., & Gandhi, U. D. (2021). A novel optimal feature selection technique for medical data classification using ANOVA based whale optimization. Journal of Ambient Intelligence and Humanized Computing, 12(3), 3527-3538.

[22] Yin, Y., Jang-Jaccard, J., Sabrina, F., & Kwak, J. (2022). Improving Multilayer-Perceptron (MLP)-based Network Anomaly Detection with Birch Clustering on CICIDS-2017 Dataset. arXiv preprint arXiv:2208.09711.

[23] Papamartzivanos, D., Mármol, F. G., & Kambourakis, G. (2019). Introducing deep learning self-adaptive misuse network intrusion detection systems. IEEE Access, 7, 13546-13560.

[24] Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using Python. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(12), 2019.

[25] Hao, W. T., Lu, Y., Dong, R. H., Shui, Y. L., & Zhang, Q. Y. (2022). Adaptive Intrusion Detection Model Based on CNN and C5.0 Classifier. International Journal of Network Security, 24(4), 648-660.

[26] Shukla, A. K. (2021). Detection of anomaly intrusion utilizing self-adaptive grasshopper optimization algorithm. Neural Computing and Applications, 33(13), 7541-7561.

[27] Bhavsar, M., Roy, K., Liu, Z., Kelly, J., & Gokaraju, B. (2022). Intrusion-Based Attack Detection Using Machine Learning Techniques for Connected Autonomous Vehicle. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 505-515). Springer, Cham.

[28] AKGUN, Devrim; HIZAL, Selman; CAVUSOGLU, Unal. A new DDoS attacks intrusion detection model based on deep learning for cybersecurity. Computers & Security, 2022, 118: 102748.

[29] SAMBANGI, Swathi; GONDI, Lakshmeeswari. A machine learning approach for DDoS (distributed denial of service) attack detection using multiple linear regression. In: Proceedings. MDPI, 2020. p. 51.

[30] SAHOO, Kshira Sagar, *et al.* A machine learning approach for predicting DDoS traffic in software defined networks. In: 2018 International Conference on Information Technology (ICIT). IEEE, 2018. p. 199-203.

[31] MUSTAPHA, Ali, *et al.* Detecting DDoS attacks using adversarial neural network. Computers & Security, 2023, 127: 103117.