# Enhanced Security Measures in Wireless Sensor Networks: Leveraging Random Forest and K-means Clustering for Node Replication Attack Detection

V. Anusha Sowbarnika
Mathematics and Computer Science,
Susquehanna University,
Selinsgrove 17870, Pennsylvania, USA
Email: veluswamy@susqu.edu

R. Lokeshkumar
School of Computer Science and Engineering,
Vellore Institute of Technology,
Vellore 632014, Tamil Nadu, India
Email: lokeshkumar.r@vit.ac.in

*T. Gopalakrishnan
Department of Information Technology,
Manipal Institute of Technology Bengaluru,
Manipal Academy of Higher Education, Manipal, India
Email:gopalakrishnan.ct@gmail.com
Corresponding Author

S Priya
Department of Information Technology,
Manipal Institute of Technology Bengaluru,
Manipal Academy of Higher Education, Manipal, India
Email:s.priya@manipal.edu

Gerard Deepak
Department of Computer Science and Engineering,
Manipal Institute of Technology Bengaluru,
Manipal Academy of Higher Education, Manipal, India
Email:gerard.deepak@manipal.edu

## Abstract

During data transmission, a major security threat is posed by the node replication attacks in Wireless Sensor Networks (WSNs) where data integrity gets comprised. Initially, the projected work tells the vulnerabilities related to the node replication attacks where the sensor node's identity is replicated by the intruders leading to the network disruption. A novel strategy is implemented by integrating the unsupervised learning algorithm 'K-means Clustering' with supervised learning algorithm 'Random Forest Classifier' in this research work to find the node replication attacks effectively by these machine learning methods. The partitioning model of K-means Clustering will

cluster the nodes by their similar behaviors. Along with K-means clustering, the ensemble learning approach, Random Forest is utilized to process the feature selection. The US Airforce LAN dataset is employed in this work and a reliable model for node replication attack detection is developed by Random Forest which categorizes the features as normal and intruder nodes after clustering. The accuracy of detection is evaluated by these methods after conducting several experiments on this dataset which measures the efficiency of false positives detection rate. The satisfactory execution results from the initial findings with significant improvements on conventional security methods. The improved level of endurance and precision is exhibited where risk mitigation is effectively achieved by this proposed work posed in WSN by node replication attacks.

***Keywords:*** Node Replication Attacks, WSN, Machine Learning, Security, K-means, Random Forest

## 1. INTRODUCTION

In Wireless Sensor Networks (WSNs), the major concern is node replication attacks because of the large number of collaborated sensor nodes during data collection and broadcasting [10]. These networks face a lot of significant privacy violations since they are operated in different instances with resource scarcity. Data confidentiality and availability is targeted by node replication attacks which develops as clever attack among the various threats in the network. The node replication attacks will replicate the identity of the sensor nodes and the replicas get distributed in WSN by this attack. The network containing genuine and attacker nodes are given in Figure 1. Here, N3 is the intruder and tries to replicate sensor node's identity in the network. The intruder node will act as a genuine node for network infiltration where the intruder node gets operated to gain access in the network. The integrity of the system gets damaged by the attacker nodes which will also hamper the data transmission, false data injection, and degrading the functionality of the network [7,9].

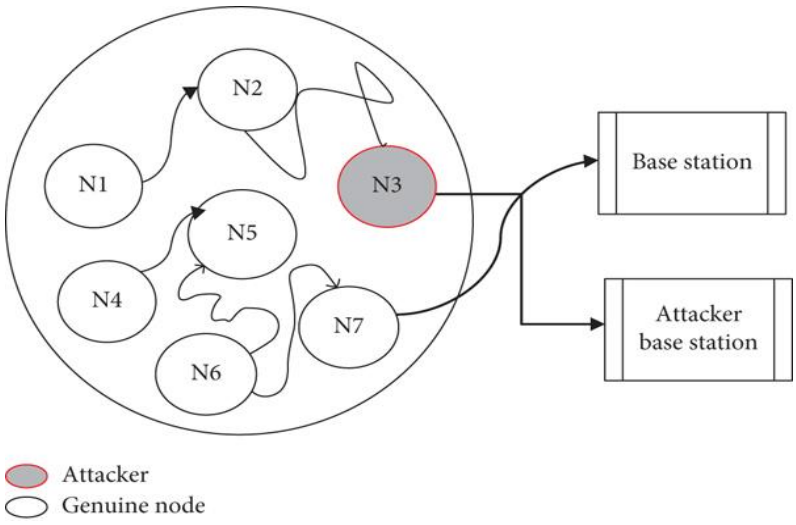


**Figure 1 Genuine and Attacker Nodes**

The vital part of Wireless Sensor networks (WSNs) is the Node replication attacks where trust is developed among the nodes and the trusted nodes are threatened by the attacker nodes [4]. The incorrect data injection, data manipulation, network resource depletion is performed by this trust exploitation. In addition to trust exploitation, it degrades the performance of the network resulting in network interruptions and depletion. A variety of approaches must be developed to counter these attacks. To create safe and authentic data communication, the unique node identifiers are utilized by physical unclonable functions (PUFs) with cryptographic methods to detect the attacker nodes by analyzing their behavior and its patterns [19]. Integrated methods are required to solve this type of data intrusion and data integrity problems created by the node replication attacks.

Preventing the Wireless Sensor Networks (WSNs) from node replication attacks has turned out to be of utmost importance in protecting the network by affording trusted and reliable data [5]. It also helps to extend the network lifetime, resource protection, and data integrity. The combination of these behavioral characteristics and cryptographic techniques underlines the importance of providing security measures to the network. This combined approach will protect the network from data penetration across the variety of working environments by assuring resilience.

Machine Learning algorithms are implemented in WSNs to classify and detect the malicious nodes. The intruder activity can be found by discovering the intruder patterns through data sets concerning security events. By employing this machine learning algorithm, similar occurrences of the event can be detected automatically by the trained models. The dataset will feed into the machine learning algorithm which helps in threat detection, and this is achieved by the assimilation of unsupervised and supervised learning algorithms.

The K-means clustering integrated with Random Forest approach creates an aggressive and complete system in detecting the node replication attacks through behavioral analysis. The intruder dataset is collected by identifying the cluster size and centroid position. The distance

between each intruder is calculated from each centroid position and it is assigned to the closest distance of the cluster based on its distance. It helps to differentiate between the normal and attacker node. On the other hand, the features are extracted after the creation of cluster labels by Random Forest algorithm which accurately detects the node replication attack through the dataset. This model is resilient in isolating the normal node, attack node features and complex patterns from the network. The intruder nodes are detected after the pattern identification on training the dataset provided. In the Random Forest method, the precision and accuracy ratio of attack detection classification can be boosted by the independent and dependent features. The variable behaviors of the nodes are detected and exhibited dynamically in WSN environment by this ensemble model.

The WSN security can be strengthened by the machine learning methodologies against node replication attacks. The novel approach "K-means Clustering and Random Forest" tells how the doors are opened for identifying the effectual defense mechanism in wireless sensor networks.

## 1.1 MOTIVATION

The node replication attacks can be prevented in Wireless Sensor Networks (WSNs) by the integration of K-means Clustering and Random Forest. This integration effectually fights against this attack and accuracy detection is powerful. These novel approaches are empowered to provide solutions to this weakness in finding the attacks and the vulnerabilities. Machine learning algorithms are powerful in enhancing network security because of their intelligence in learning the attack patterns. The combination of feature selection 'Random Forest' with cluster label creation 'K-means Clustering' based on their behaviors will improve the results of intrusion detection. The data integrity accuracy can be guaranteed and strengthened if this integration is successful in WSN applications. The security protocol advancements are given by this research to reduce network interruptions by the attackers and enhance the network performance. The development of the strong security feature is the aim against data infiltration by protecting the WSN network.

## 1.2 CONTRIBUTIONS

The key aspects of the research contributions are:

- Designing an assimilation approach by K-means clustering and Random Forest to protect the Wireless Sensor Networks (WSNs) against node replication attacks.
- Utilizing machine learning methods to enhance the node replication attack detection accuracy.
- Construction of broad defensive network by combining the behavioral analysis of K-means with Random Forest against data infiltration by strengthening the WSN security. infiltration attempts.

## 1.3 ORGANIZATION

The research work is organized as: Section 1 covers the Introduction part, motivation, and contributions of the research. Section 2 mentions the Literature review, Research

Methodology is dealt with in section 3. Section 4 talks about the Results and Discussion, and section 5 narrates the conclusion.

## 2. LITERATURE REVIEW

The technology development has transformed the way differently with the WSN creation in which they are connected. In Machine learning, automated methods are proposed to face the challenges in Wireless Sensor Networks (WSN). These challenges are addressed by the projected work given by Mohammed S. Alsahli et al (2021) which helps to prevent the intrusion system by affording firewall-based network with WSN and KDD99 dataset. Accurate testing algorithms are evolved for testing, and it recommended for WSN network security.

In WSN, the cyber threats are efficiently handled by creating a multi-layer framework for intrusion detection by Nada M. Alruhaily et al (2021) introduces a multi-layer intrusion detection framework. He introduced the classifier algorithm called "Naive Bayes" for decision making in the real time network. The detailed cloud-based analysis is implemented by the Random Forest approach. The demonstration gives the 100% accuracy rate for variety of attacks like flooding and scheduling. This integrated analysis gives accuracy about correct and incorrect attack predictions.

WSNs are vulnerable to security attacks regardless of their numerous uses of their applications in violent and unrestricted networks. Network data can be protected through several detection mechanisms introduced by Samir Ifzarne et al (2021) in WSNs. The model is developed by Online Passive Aggressive classifier and information gain in introducing a model for Denial-of-Service attacks based on information gain ratio and the online Passive Aggressive classifier where the intruder nodes are prioritized. The WSN-DS dataset is used for experimentation to prove the intruder detection accuracy ratio.

The advanced model of Intrusion Detection System is proposed by Subarna Shakya (2021) in wireless sensor networks (WSNs). Optimization is combined with machine learning method to reduce the incorrect predictions. The modified Grey Wolf Optimization (MLGWO) algorithm is integrated with machine learning approach to improve the processing rate and performance. The wolf in this model gives a positive impact on evaluating the accuracy performance on the NSL KDD'99 dataset.

Machine learning algorithms became essential for intrusion detection due to the several drawbacks of conventional security methods. This type of algorithm helps to address the needs of taking the normal and attacker patterns. making Machine Learning (ML) algorithms essential for categorizing network traffic into normal and intrusive. The performances are evaluated by Vasudeva Pai et al (2021) on the NSL-KDD dataset by machine learning methods to find the attacks like remote attacks, user to root, Denial of service attacks in automation and industrial sectors.

Automated Machine Learning (AutoML) was introduced by Abhilash Singh et al (2022) for accurate intrusion detection prediction for 'k' number of barriers. One of the superior models in machine learning is Bayesian optimization which acts as a Gaussian process regression

where its performances are merged with AutoML and correlation coefficient on the attacker found dataset. This integrated approach provides an accurate tool for displaying intrusion detection and prevention.

Ayesha S. Dina et al (2021) intended the combination of machine learning method with Signature-based and anomaly-based solutions for classifying the pivotal behavior of legitimate and intruder nodes in the network. The last ten years literature was studied by the Author which turns out to be a powerful source for the scholars who do research in this domain. The highlights of this survey include the future discovery of intrusion attacks.

The attacks like DDoS, gray hole, black hole, and wormhole are addressed by the research work given by Pratik Gite et al (2023). The data patterns are examined continuously from each node by utilizing the Base Station machine learning algorithm (BS) in the system. The attacks are prevented by sending the notification by base station to all neighbors after seeing the destructive behavior. The existing dataset is used for training the model where the results contain all the types of attacks. NS2 simulator is used for conducting the experiments which helps to obtain maximum accuracy in deriving the results after the trained data. The packet delivery ratio and energy consumption parameters help for accurate prediction and performance.

A deep learning framework was initialized by Abhilash Singh et al (2023). It is an innovative approach for intrusion detection where a fully connected feed-forward Artificial Neural Network (ANN) is utilized for accurate prediction of ‘k’ barriers. The Monte Carlo simulation is employed to extract the trained features which will show the efficiency and accuracy of the model. This algorithm has set the benchmark in intrusion detection accuracy. Another deep learning framework of feed-forward artificial neural network was proposed by S. Muruganandam et al (2023) for counting the ‘k’ number of barriers estimation accurately. Author also employed the Monte Carlo simulation technique in calculating the intrusion attack effectively.

## 3. RESEARCH METHODOLOGY

The node replication attacks are evaluated in structured way in this research methodology. Originally, the comprehensive survey was carried out to determine the gaps and the current approaches in identifying the intruder network. The study tells there is no satisfactory accuracy prediction of attacks in the network using machine learning methods. These gaps can be filled by creating a dataset of diverse network environment containing normal and attacker nodes. Hence the machine learning algorithm integration played an important role in classification and feature extraction of the dataset in anomaly detection by K-means Clustering and Random Forest approach.

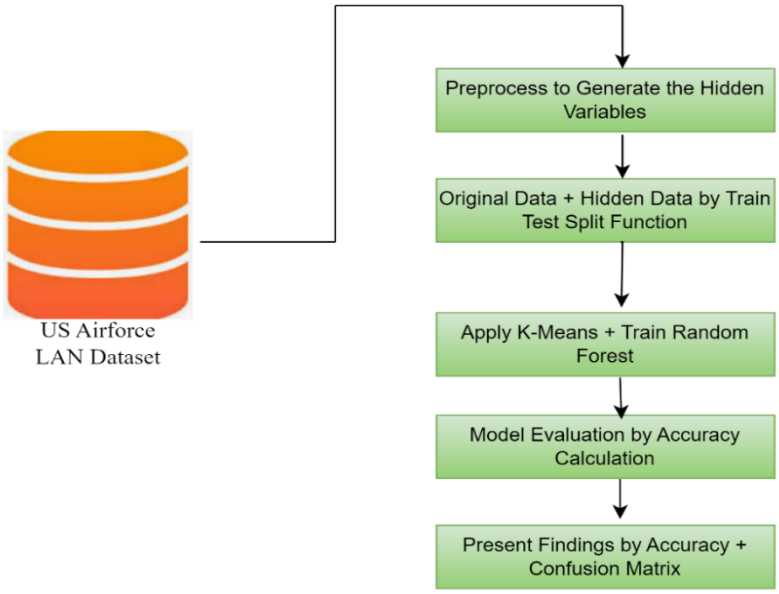


**Figure 2 Proposed Workflow Model**

After the initialization of the algorithm, the model's performance is evaluated by this assimilated method where the dataset gets divided into testing and training. The proposed workflow of the model is given in Figure 2. At the first stage, the US Airforce LAN dataset is taken where data cleaning is performed to extract the hidden variables. In the second stage, once the hidden variables are combined with original dataset, training and testing is executed. The training and testing phase helps to predict the future results by applying the integrated approach by clustering and classification. In the last stage, accuracy, F1 score, recall is used as the performance metrics for evaluation for intruder identification. The positive findings can be attained by the combination of K-means clustering and Random Forest which strengthens the security against node application attacks in WSN.

### 3.1 Creation of Cluster Labels by K-means Clustering

The cluster labels are constructed based on the number of attackers classified under the target variable. Before cluster label creation, the noise present in the dataset must be cleaned to obtain accurate results in intruder detection. The feature of the dataset is standardized or scaled by the data normalization. The shape of the US LAN Airforce dataset contains 25,192 rows and 39 columns of records.

Data normalization is the technique employed in this proposed work for data preprocessing to handle the variations among the measurements and dimensions. The equation [1] tells the data normalization in (0, 1) range by MinMaxScaler approach.

$$X_{nor} = \frac{X - \text{Xminimum}}{\text{Xmaximum} - \text{Xminimum}} \quad (1)$$

In the above equation, the normalization score is denoted as $X_{nor}$; the maximal and minimal dataset features are denoted as Xmaximum and Xminimum.

After preprocessing step, cluster label creation is executed to find specific patterns in the network by K-means clustering in WSN [3]. Due to the similar characteristics of the target variable, the data points are grouped as clusters in the network. Since the target feature in the US LAN Airforce dataset contains ‘normal’ and ’anomaly’ attackers and hence the cluster size is 2. The cluster labels are provided by K-means Clustering through continuous improvements which figures out the normal and attacker nodes.

The K-means method is iterative, and data is segregated into discrete and independent categories where the data points are assigned to the single group which is denoted by 'k' [8]. The similarity among the intra-cluster data points are maximized and this is accomplished by minimizing the sum of squared distances between data points and their respective cluster centroids [17]. The overall objective is to minimize variation within clusters, promoting homogeneity and similarity among data points within the same cluster.

The operation of the k-means algorithm involves the following steps:

- Specify K clusters, randomly selecting K initial cluster centroids to represent the initial cluster centers.
- Shuffle to choose the centroid position in the dataset and initial random points are chosen randomly.
- Assign each node to the cluster whose centroid is closest in terms of distance (usually Euclidean distance).
- Cluster’s centroid position is recalculated based on the mean of the nodes assigned to that cluster, representing the center of the cluster.
- Reassign each node to the cluster with the updated centroid, considering the new cluster distances.
- Iterate steps 3 and 4 until convergence or a predetermined number of iterations is reached. Convergence happens when centroids stabilize, and the assignment of nodes to clusters no longer changes significantly.
- Upon convergence, each node is assigned the cluster label corresponding to the cluster it belongs to.
- Nodes are distinguished by their cluster labels, forming groups with similar characteristics within the same cluster.

The adopted process of feature selection is the approach with multiple levels to combine the strengths of integrated methods. Initially, in the filter method stage, the selection of features was based on correlation analysis. It was employed to identify crucial attributes by assessing multi-collinearity in the data. This method involves calculating the class and attribute’s relationship, where the relevance of a feature $f_i$ to the class is determined if and only if there exists some $f_i$ and corr for which the probability as outlined in Eq. 2:

$$P(F_i = f_i) > 0 \quad (2)$$

In the provided equations, CORR represents a given data, corr represents subsets, $F_i$ represents a feature candidate, and $f_i$ denotes the subset feature. The predictability of the attribute and the class correlation is assessed, as expressed in Eq. 3:

$$P (CORR = corr \mid F_i = f_i) \neq p (CORR = corr) \quad (3)$$

The assessment of each attribute-class association involved the application of the genetic search technique, which yields the selected features by considering the maximum fitness value (Eq. 4).

$$\text{Fitness } (X) = \frac{3}{4} B + \frac{1}{4} = (1 - \frac{S+F}{2}) \quad (4)$$

A rule-based strategy was devised to extract a subset feature 'S with minimal features while preserving an equivalent fitness level to subsets with maximum features. In simpler terms, if two feature subsets exhibit the same fitness values, the rule evaluator prioritizes the subset with fewer features.

***Algorithm 1 Integration of K-Means Clustering and Random Forest Model***
***Input***: US Airforce LAN.CSV, US Airforce LAN1.CSV [Training Data, Testing Data]
***Output***: Model Evaluation by Accuracy Calculation
***while*** *preprocess_data(s)* **do**
    s ← load(features(x), labels(y))
    x ← Input vectors
    y ← Target vectors
    z = preprocess_data_minmaxscaler(s)
***Apply*** K-Means Clustering
    C ← clusters
    ***Initialize***(c)
    C ← *Kmeans_clustering*(z)
***Perform*** Feature Engineering
    Feature_set ←*combine_features_clusters*(x, C)
***Determine*** Train_Test Split Function
    X_train, X_test, y_train, y_test ← *split_data*(Feature_set, y)
***Integrate*** Random Forest Model
    ***Initialize*** (RF_Model)
    RF_Model ← *train_random_forest*(X_train, y_train)
***Predict*** Trained Model
    pred_y ← *RF_Model.predict*(X_test)
***Evaluate*** Model
    Accuracy ← *calculate_accuracy*(y_test, pred_y)
***End while***

In the research on detecting instances of node replication attacks in WSNs, these procedures facilitated the generation of cluster labels to differentiate between normal and replicated instances. Subsequently, these cluster labels served as input features for the subsequent

integration with Random Forest, enhancing the overall efficacy of the security framework. The steps for combining K-means clustering and Random Forest are outlined in algorithm 1.

### 3.2 Random Forest in Classifying the Node Replication Attacks

During the second phase of the features in multi-level selection, the method of forward selection is employed sequentially to further narrow down the selection of highest features constructed on the baseline classifier's accuracy. This decision was motivated by the fact that filter methods remain unaffected by the classifier, prompting the need for an integration method. The method starts with an empty set and systematically adds the next optimal feature, ensuring that the combination with previously chosen features yields the maximal baseline classifier accuracy.

In this dataset, the training data is given 70% of the dataset and testing data is given 30% of the dataset. The data for training and testing is allocated by initializing the random size function. The dataset is shuffled by the size mentioned in the random size method and allocation is performed. The data featuring and data analysis is effectively carried out by integrating Random Forest after K-means Clustering method. The target features give the cluster labels and the features are served as the input to Random Forest.

The random forests algorithm is a powerful ensemble approach used for both classification and regression, making it one of the most effective data mining techniques [15]. Widely applied in various domains, it has found uses in prediction and probability estimation. Surprisingly, the algorithm has not been previously employed in automatic node replication attack detection. In this proposed system, the misuse component leverages the random forests algorithm for node replication attack detection classification, while the anomaly component relies on the outlier detection mechanism inherent in the algorithm. Random forest constructs multiple decision trees and predictions are combined to achieve a stable outcome. The work is represented in figure 3.

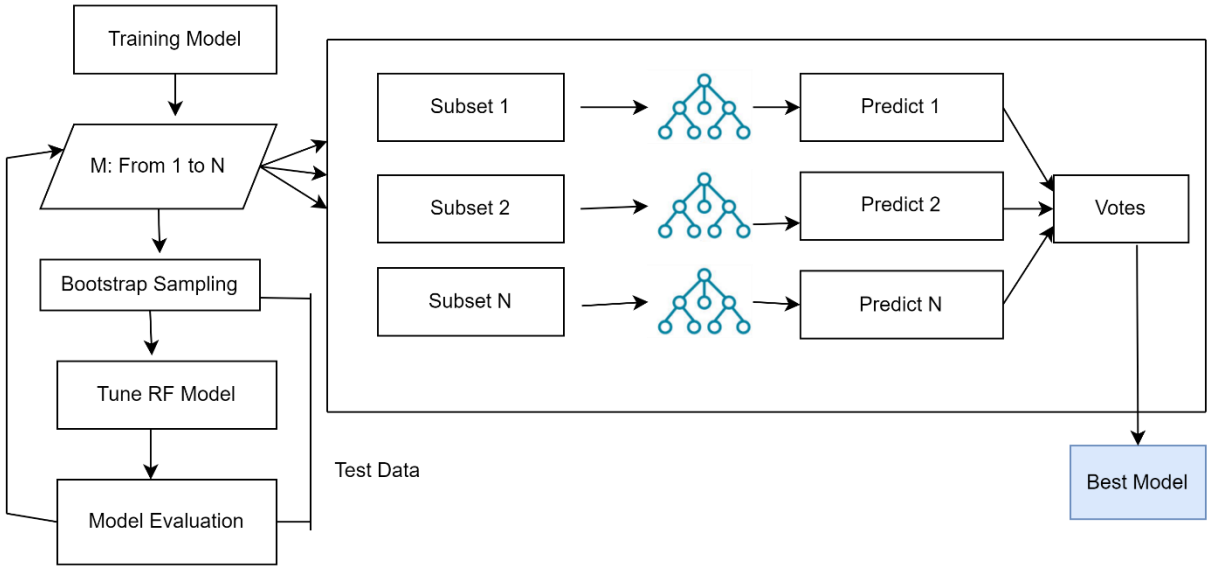


**Figure 3 Working of Random Forest**

The Decision Tree algorithm, belonging to the supervised learning family, aims to create a model for training that predicts the target variables of target or class by decision rules from existing data. (training data) [22]. It offers a straightforward understanding compared to other classification algorithms. The tree representation is taken by the algorithm to address the problems, by associating the internal node to attribute, and associating the leaf node to a class label. Disjunctive Normal Form, major problems are faced by decision Trees in identifying the root node's attribute at each level, called attribute selection.

Random Forest emerges as a robust classifier for identifying node replication attacks in distributed systems [14]. Its strength lies in effectively handling intricate, non-linear relationships within the feature space. The aggregation of predictions from numerous decision trees addresses overfitting concerns and enhances overall robustness. Specifically, the high dimensional data can be handled by Random Forest wherein the detection precision is improved by the feature selection theory. The architecture of Random Forest enhances its prediction performance and is more susceptible to noise and outliers. Additionally, its inherent parallelization capability accelerates training processes, especially on extensive datasets. The algorithm includes a built-in feature important metric, facilitating the identification of crucial factors contributing to node replication detection. The distributed systems can be defended by Random Forest from node replication attacks with high precision and versatility.

### 3.3 Node Replication Attack Detection by the Integration of K-means Clustering and Random Forest Algorithm

The execution involves the emulation of a wireless sensor network operating under normal conditions and anomalies. The outcomes, comprising traffic and delays, are subjected to clustering and classification using Python. K-means clustering is employed to compute clusters for both traffic and delay data.

Random Forest is then efficiently utilized to enhance classification accuracy and determine respective performance metrics and threshold values. Subsequently, these thresholds are applied to identify anomalies in the network. The analysis is bifurcated into two parts: one using traffic data to establish threshold values for replicated nodes, and the other using delay data to determine threshold values for replicated nodes. The integrated mechanism is depicted in figure 4.

Two types of traffic data are utilized: traffic which is sent and received traffic. In the analysis of traffic that is transmitted, the focus is on detecting replicated nodes. Conversely, in traffic received analysis, the objective is to identify nodes collaborating with node replication attackers, because they might demonstrate receiving elevated levels of traffic during an attack. The nodes with elevated traffic are categorized as node replication attacker nodes. To ascertain which node collaborates with a particular node replication attacker node by examining the range of communication. Collaboration is deemed possible if a node falls within the range; otherwise, collaboration is considered impractical.

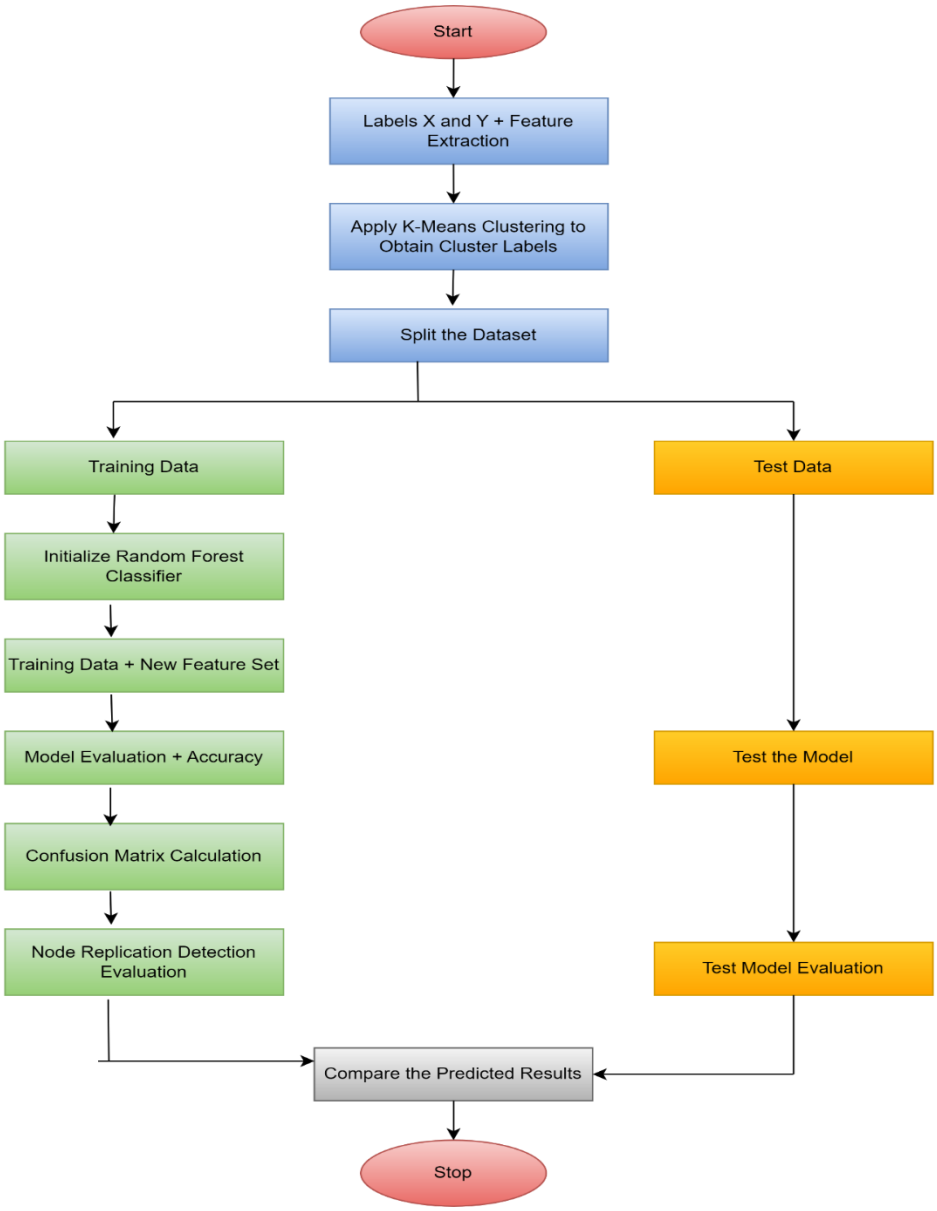


**Figure 4 Integrated Model of the Projected Work**

If a traffic transmitted under attack and node is identified as a replicated node, then it should be nonzero. Conversely, in normal scenario, the traffic sent should be zero.

$Traffic_{sent}$ - 0, Normal

$Traffic_{sent}$ - 1, Under Anomaly Attack

In the receiver part, the calculated amount of traffic it receives should be below the value of threshold. For instance, if the received traffic's threshold value is denoted as y, then in normal cases, it should be equal to or less than this specified value.

$Traffic_{recv}$ <=y, Normal

$Traffic_{recv}$ > y, Node Replication Attack

In this scenario, the traffic received for a node comprises the sum of its own traffic and the transmitted traffic by a node replication attacker node. If C represents the node and D is the node replication attacker node, the traffic received can be expressed as follows:

Traffic[C→received] = Traffic[C→received + Traffic[D→received]

***Algorithm 2 Node Replication Attack Detection***

***Input***: US Airforce LAN.CSV, US Airforce LAN1.USV (Training Data, Testing Data)
***Output***: Node Replication Attack Detection

// $\boldsymbol{Traffic_{sent}}$ is the data transmitted by a node
//***T[a]*** is the duration for a node to transmit the packets
//***T[a]_TH*** is the threshold value for packet transmission
// $\boldsymbol{Traffic_{recv}}$ is the data received by a node
// $\boldsymbol{TH_a}$ is the threshold level for the traffic that a node receives.
//$\boldsymbol{dis_{ab}}$ is the distance of communication range between a and b nodes
//$\boldsymbol{R_{del}}$ is the boundary of communication range between a and b nodes

```
Hybrid_NodeReplication_Detection()
{
        At every node in between
        for a = 1 to m
        nodereplication_detection()
          {
              At node a:
               if(Traffic_sent == 0)
                    if(T[a] > T[a]_TH)
                     Node[a] ← NodeReplication_Attack_Node
                     Communication_range()
                     {
                      if(dis_ab < Rdel)
                         if(Traffic_recv > TH_a)
                              Nodes [a] and [b] are in collaboration
                         else
                               Move to transitional nodes
                       else
                           Nodes [a] and [b] are not in the range of communication
                      }
                 else
                     No_Attacker()
}
```

The proposed technique demonstrates the capability to identify the anomalies in WSNs. This method effectively identifies normal nodes and node replication attacker nodes through the analysis characterized by their failure to forward traffic and absorbing all packets, are detected by comparing their typically zero traffic sent values. Node replication attacker nodes, responsible for redirecting traffic and potentially causing increased delay, sometimes resulting in infinity, are identified through the analysis of traffic delay parameter values. Experimental results indicate the successful detection of the introduced hybrid anomaly by the proposed scheme, and it is depicted in algorithm 2.

## 4. RESULTS AND DISCUSSION

The work results highlight the efficacy of utilizing K-means Clustering and Random Forest for the detection of node replication attacks in wireless sensor networks. Across a range of diverse datasets, this hybrid approach consistently exhibited high accuracy in effectively identifying instances of node replication.

The typical US Airforce LAN dataset was chosen in the proposed work and tested on the dataset in Python. The dataset consists of 25192 rows and 39 columns. The projected work helps to enhance the node replication attack detection by creating cluster labels and feature selection. The K-means Clustering method is used to create the cluster labels and Random Forest is used for classification.

The process of clustering was essential to make the clusters by data partitioning and this is effectually achieved by K-means Clustering where replicated nodes are pointed by highlighting the anomaly samples. The class is the target vector that helps in creating the cluster labels and the cluster labels are grouped by the names ‘Normal’ and ‘Anomaly’. Initially, the cluster creation was executed. After this execution, it acts as an input for the next step called feature selection. Secondly, the feature selection method is effectively carried out by Random Forest algorithm. Finally, data splitting is based on training and testing data: a) For training - 70% of the data and b) For testing - 30% of the data.

The instance classification with high level of accuracy is demonstrated by utilizing the capabilities of the ensemble learning model called Random Forest which reduces the amount of false positive ratio. For the node replication attacks, enhanced level of attack detection accuracy and creating a strong barrier against various levels of complications is attained through this integrated approach. Moreover, the tolerance towards noise and fluctuations is exhibited in various network scenarios by this method indicates its flexibility. A confusion matrix is the tabular form for examining the performance of the classification approach. The performance extensive summary is outlined by the correct and incorrect predictions. The predictions are presented in Table 1.

| Cluster Labels | Normal | Anomaly |
|---|---|---|
| **Normal** | 4044 (TN) | 7 (FN) |
| **Anomaly** | 117 (FP) | 3390 (TP) |

**Table 1 Confusion Matrix Table**

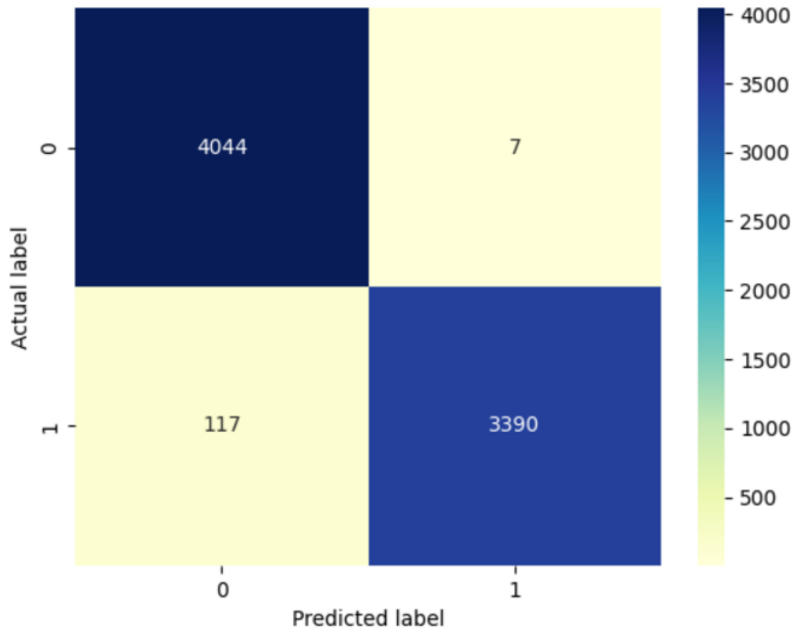


**Figure 5 Confusion Matrix**

The four sections in the confusion matrix describes the following:

- True Positive (TP): Signifying instances where the work accurately predicted the positive class, such as correctly identifying node replication attacks.
- True Negative (TN): Representing instances where the work correctly predicted the negative class, for example, accurately identifying instances that are not node replication attacks.
- False Positive (FP): Occurring when the work incorrectly predicts the positive class, like falsely identifying a normal node as a replication attack.
- False Negative (FN): Arising when the work incorrectly predicts the negative class, such as failing to identify a node replication attack.

In Figure 5, the incorrect and incorrect predictions are stated by the confusion matrix. The confusion matrix plays an effective role in evaluating the performance metrics like sensitivity, precision F1 score, recall and accuracy. The delicate vision is enriched into the excessive classes which are strength and weaknesses of the model. The consequences related to the different types of mistakes are evaluated by the researchers and scholars based the thorough analysis and improvements or changes in the scenario can be made effectively by making the well-informed decisions. The accuracy in predicting the node replication assaults is made effective by the construction of confusion matrix in wireless sensor networks constructively. The effectualness of the model determines the classification accuracy among the anomaly and normal nodes.

| SNO | Methods | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| 1 | K-Means | 0.86 | 0.87 | 0.86 | 0.85 |
| 2 | Random Forest | 0.94 | 0.93 | 0.92 | 0.94 |
| 3 | K-Means + Random Forest | 0.98 | 0.98 | 0.97 | 0.98 |

**Table 2 Node Replication Detection – Performance Metrics**

The hybrid strategy "K-means with Random Forest Classifier" mentioned in Table 2 and Figure 6 has repeatedly showed superior results for F1 score, precision, accuracy, and recall. The efficiency of the model is evaluated by the extensive set of measurements by studying the ability of the model and the accuracy of positive identifications in capturing the entire scope of the instances of node replication attacks in wireless sensor networks.

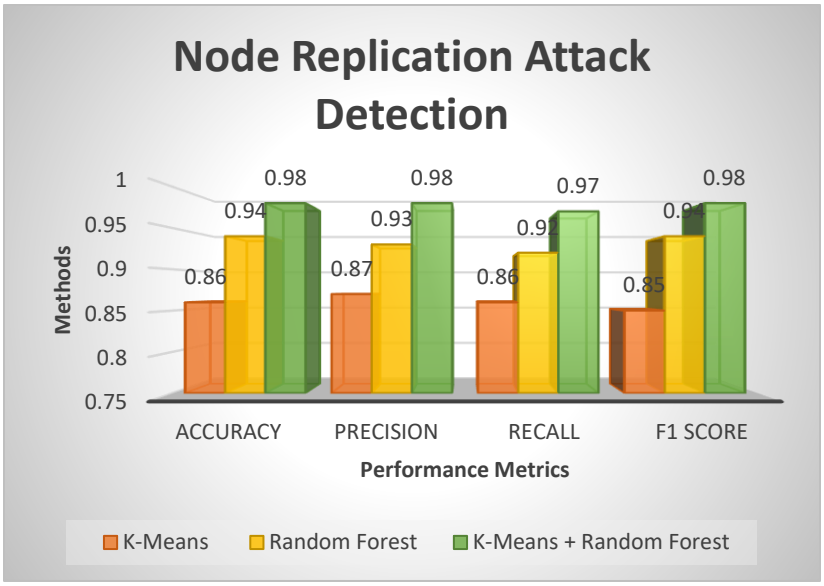


**Figure 6 Node Replication Detection – Performance Metrics**

The synergistic interaction is highlighted by the projected works "K-means Clustering and Random Forest", wherein the cluster formation is added by the unsupervised learning method K-means and decision-making process is refined by the ensemble method of Random Forest classifier. The detailed defense mechanism against node replication attacks is facilitated by this integrated approach by the behavior of network analysis and this paves way as a solution for the node replication attacks. The research outcomes focus on how the security measures can be improved by the in-depth investigation of node replication attacks in wireless sensor networks. This affords strong and robust defense mechanism for the upcoming threats in the wireless networks.

## 5. CONCLUSION

A powerful security strategy for identifying the node replication attacks is presented in wireless sensor networks by the combined process of K-means clustering and Random

Forest methods. The work collaboration of these two machine learning methods offers a great level of security. Firstly, the clusters are formed by data partitioning with the help of K-means Clustering which helps to recognize the similarities, outlier detection by identifying and isolating the malicious attackers. Secondly, anomaly detection and obscure patter discovery is made reliable by the capabilities of ensemble learning model "Random Forest" for the endeavors with node replication identification. A thorough and effectual protection mechanism is achieved through this amalgamation in wireless sensor networks against node replication attacks. The posture of security to certify the compliance threats developing is improved by these machine learning methods which serve as the constructive method for maintaining the data availability and integrity in wireless sensor networks. The implication of utilizing the complicated machine learning methods are underlined in this research for upbeat malicious attack detection This research underscores the significance of employing sophisticated machine learning approaches for proactive threat detection and moderation in the active landscape of security features of wireless sensor network.

## 6. REFERENCES


[1] Abhilash Singh, J. Amutha, Jaiprakash Nagar, Sandeep Sharma & Cheng-Chi Lee (2022), "AutoML-ID: automated machine learning model for intrusion detection using wireless sensor network", Scientifc Reports, (2022) 12:9074.

[2] Abhilash Singh, J. Amutha, Jaiprakash Nagar, Sandeep Sharma (2023), "A deep learning approach to predict the number of *k*-barriers for intrusion detection over a circular region using wireless sensor networks", Expert Systems With Applications 211 (2023) 118588.

[3] Amar Meryem, Bouabid EL Ouahidi (2020), "Hybrid intrusion detection system using machine learning", Network Security, Volume 2020, Issue 5, 2020, pp: 8-19.

[4] Anusha Sowbarnika V, Dr. M. Balasubramani, Dr. K. Kavitha (2023), "The security-based Optimization Algorithm for Enhancing the Energy Efficiency of Wireless Sensor Networks", International Journal of Communication Systems, Wiley Online Library, Vol. 36, Issue 16, e5584.

[5] Ashwini B, S. Abhale, and S. Manivannan (2020), "Supervised Machine Learning Classification Algorithmic Approach for Finding Anomaly Type of Intrusion Detection in Wireless Sensor Network," Optical Memory and Neural Networks, vol. 29, no. 3, pp. 244-256, 2020.

[6] Ayesha S. Dina, D. Manivannan (2021), "Intrusion detection based on Machine Learning techniques in computer networks", Internet of Things, 16 (2021), 100462.

[7] Çavuşoğlu, Ü (2019), "A new hybrid approach for intrusion detection using machine learning methods", Appl Intell 49, pp: 2735–2761.

[8] Junwen Chen, Xuemei Qi, Linfeng Chen, Fulong Chen, Guihua Cheng (2020), "Quantum-inspired ant lion optimized hybrid k-means for cluster analysis and intrusion detection", Knowledge-Based Systems, Volume 203, 106167.

[9] Liu, Hongyu, and Bo Lang (2019), "Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey" *Applied Sciences* 9, no. 20: 4396.

[10] Mishra P, V. Varadharajan, U. Tupakula and E. S. Pilli (2019), "A Detailed Investigation and Analysis of Using Machine Learning Techniques for Intrusion Detection," in *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 686-728.

[11] Mohammed S. Alsahli , Marwah M. Almasri , Mousa Al-Akhras, Abdulaziz I. Al-Issa, Mohammed Alawairdhi (2021), "Evaluation of Machine Learning Algorithms for Intrusion Detection System in WSN", International Journal of Advanced Computer Science and Applications, Vol. 12, No. 5, pp: 617 – 626.

[12] Muruganandam S, Rahul Joshi, P. Suresh, N. Balakrishna, Kakarla Hari Kishore, S.V. Manikanthan (2023), "A deep learning based feed forward artificial neural network to predict the K-barriers for intrusion detection using a wireless sensor network", Measurement: Sensors 25 (2023) 100613.

[13] Nada M. Alruhaily, Dina M. Ibrahim (2021), "A Multi-layer Machine Learning-based Intrusion Detection System for Wireless Sensor Networks", International Journal of Advanced Computer Science and Applications, Vol. 12, No. 4, pp: 281 – 288.

[14] Negandhi, P., Trivedi, Y., Mangrulkar, R. (2019), "Intrusion Detection System Using Random Forest on the NSL-KDD Dataset" Emerging Research in Computing, Information, Communication and Applications,Advances in Intelligent Systems and Computing, vol 906. Springer, Singapore.

[15] Paulo Angelo Alves Resende and André Costa Drummond (2018), "A Survey of Random Forest Based Methods for Intrusion Detection Systems", ACM Comput. Surv, Vol 51, Issue 3, 36 pages.

[16] Pratik Gite, Kuldeep Chouhan, K. Murali Krishna, Chinmaya Kumar Nayak, Mukesh Soni, Amit Shrivastava (2023), "ML Based Intrusion Detection Scheme for various types of attacks in a WSN using C4.5 and CART classifiers", Materials Today: Proceedings 80 (2023) 3769–3776.

[17] Saba Karim, Rousanuzzaman, Patel Ayaz Yunus, Patha Hamid Khan, Mohammad Asif (2019), "Implementation of K-Means Clustering for Intrusion Detection", International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Vol. 5, Issue 2, pp: 1232 – 1241.

[18] Samir Ifzarne, Hiba Tabbaa, Imad Hafidi, Nidal Lamghari (2021), "Anomaly Detection using Machine Learning Techniques in Wireless Sensor Networks", Journal of Physics: Conference Series, The International Conference on Mathematics & Data Science (ICMDS), 1743 (2021) 012021.

[19] Saroj Kr. Biswas (2018), "Intrusion Detection Using Machine Learning: A Comparison Study", International Journal of Pure and Applied Mathematics, Volume 118 No. 19, pp: 101-114.

[20] Subarna Shakya (2021), “Modified Gray Wolf Feature Selection and Machine Learning Classification for Wireless Sensor Network Intrusion Detection”, J. Sustain. Wireless Syst., vol. 03, no. 2, pp. 118-127.

[21] Vasudeva Pai, Devidas, Adesh N. D. (2021), “Comparative analysis of Machine Learning algorithms for Intrusion Detection”, IOP Conf. Series: Materials Science and Engineering, 1013 (2021) 012038.

[22] XuKui Li, Wei Chen, Qianru Zhang, Lifa Wu (2020), “Building Auto-Encoder Intrusion Detection System based on random forest feature selection”, Computers & Security, Volume 95, 101851.