Faculty of Engineering
University of Nigeria, Nsukka
22nd- 25th June, 2021

Proceedings of the
2021 Sustainable Engineering and
Industrial Technology Conference

# A Multi-Step Adaptive Synthetic Oversampling and Random Forest Cascaded Model for Multi-Class Intrusion Detection

H.A. Isah, O.A. Abisoye, K. Lawal

*Department of Computer Science Federal University of Technology Minna, Nigeria.*

***Emails:*** *Hashimabdul51@yahoo.com (HAI), o.abisoye@futminna.edu.ng (OAA), Kehinde.lawal@futminna.edu.ng (KL)*

**Abstract**

Hackers have developed better and smart traditions to attack WSN in sequence when data are transfer in systems. The harm, hackers can carry out upon thorough a WSNs is well understood. A reasonable damage scenario can be envisaged where a state intercepting encrypted financial data gets hacked. Logical cyber security systems have become without doubt significant for improved security against malicious threats. The proposed multi-step adaptive synthetic oversampling and random forest cascaded model for intrusion detection system (IDS) using big data, The NSL-KDD dataset used as a benchmark to evaluate the feasibility and effectiveness of the proposed architecture. Simulation results demonstrate the potential of our proposed IDS system, performance better compared to existing methods,

**Keywords:** wireless sensor network, instruction detection, data transfer

## 1. INTRODUCTION

With the quick development of wireless sensor technology, embedded computing technology, wireless communication technology, and distributed information processing technology, wireless sensor networks have very broad application prospects, like national defense, ecological observation, environmental monitoring, medical security, space exploration, volcano observation, architecture, and city management, [1]. Wireless sensor networks can realize real-time monitoring, sensing and acquired information of various environments through the cooperation of various integrated micro-sensors.

### *1.1. Problem Statement*

Problems of intrusion detection in wireless sensor networks is the imbalance of attack types, such as Denial of Service (DoS) attacks, they have more connections than probing attacks, user to root (U2R) attacks, and root to local (R2L) attacks model.

### *1.2. Aim and Objectives*

1. design a multi-step oversampling algorithm for multi-class minority oversampling using ADASYN.
2. Apply the algorithm in (1) on the KDD cup'99 dataset using MATLAB programming environment.
3. Design a cascaded random forest classification model and apply the model on the dataset in (2).
4. Evaluate the performance of the model in (3) using RMASE, MAE, accuracy, precision and F-measure

## 2. MATERIALS AND METHODS

This chapter presents the materials and methods that were deployed to achieve the aim of this research work and the methodology deployed to achieve each objective

### *2.1. Multi-Step ADASYN Algorithm Design*

The NSL-KDD 99 10% training dataset was used for this experiment. The dataset contains five class of intrusion attacks. Normal, Dos, probe, R2L and U2R attacks. From the dataset, the total number of samples is 125973 comprising of 67343 normal, 45927 Dos, 11656 Probe, 995 R2L and 52 U2R. A total of 39 features and one categorical class makes up the features of the dataset.

The multi-step ADASYN algorithm for minority multi-class sampling

### *2.2. Cascaded Random Forest Model Design*

Figure ?? shows the cascaded Random Forest model for multi-class classification of intrusion data. It comprises of four models each designed for a one versus one scenario. The NvDos RF model was designed to classify the Normal versus Dos attack class. NvProbe model is for Normal versus probe attack, NvR2L is for normal versus R2L attack while the NvU2R model is for normal versus U2R attack.

The designed models were implemented using Weka machine learning environment. The results obtained were evaluated and compared with cascaded

ISBN 0000-0000

Table 1: Data Oversampling Result.

| S/N | Intrusion Type | Unbalanced Values | Balanced Values |
|---|---|---|---|
| 1 | Normal | 67343 | 67343 |
| 2 | Dos | 45927 | 67343 |
| 3 | Probe | 11656 | 67343 |
| 4 | R2L | 995 | 67343 |
| 5 | U2R | 52 | 67343 |

Table 2: Confusion Matrix of Balanced Cascaded model.

| Models | True Positive (TP) | True negative (TN) | False Positive (FP) | False Negative (FN) |
|---|---|---|---|---|
| NvDos | 67333 | 67330 | 13 | 10 |
| NvProbe | 67325 | 67293 | 50 | 18 |
| NvR2L | 67331 | 67333 | 10 | 12 |
| NvU2R | 67337 | 67334 | 9 | 6 |

Table 3: Confusion Matrix of unbalanced Cascaded model.

| Models | True Positive (TP) | True negative (TN) | False Positive (FP) | False Negative (FN) |
|---|---|---|---|---|
| NvDos | 45907 | 67331 | 12 | 20 |
| NvProbe | 11599 | 67322 | 21 | 57 |
| NvR2L | 995 | 67340 | 3 | 40 |
| NvU2R | 26 | 67341 | 2 | 26 |

Table 4: Performance Evaluation of the Balanced Model.

| Models | RMSE | MAE | Accuracy | Precision | F-Measure |
|---|---|---|---|---|---|
| NvDos | 0.0139 | 0.0009 | 99.9829 | 1 | 1 |
| NvProbe | 0.0241 | 0.0026 | 99.9495 | 0.999 | 0.999 |
| NvR2L | 0.0161 | 0.0016 | 99.9837 | 1 | 1 |
| NvU2R | 0.011 | 0.0005 | 99.9889 | 1 | 1 |

Table 5: Performance Evaluation of the Unbalanced Models.

| Models | RMSE | MAE | Accuracy | Precision | F-Measure |
|---|---|---|---|---|---|
| NvDos | 0.0156 | 0.001 | 99.9717 | 1 | 1 |
| NvProbe | 0.0293 | 0.003 | 99.9013 | 0.998 | 0.997 |
| NvR2L | 0.0236 | 0.0019 | 99.9371 | 0.997 | 0.978 |
| NvU2R | 0.0172 | 0.0007 | 99.9585 | 0.929 | 0.650 |

model trained with unbalanced data. Also, Performance Evaluation The performance of the developed models were evaluated using the following;

*2.2.1. Root mean Squared Error (RMSE)*
*2.2.2. Mean Absolute Error (MAE)*
*2.2.3. Accuracy*
*2.2.4. Precision*
*2.2.5. F-measure*

## 3. RESULTS AND DISCUSSION

The results obtained in order to achieve the aim and objectives of this research work are presented and discussed in this section.

### *3.1. Multistep Oversampling Result*

Table 1 shows the result of oversampling the inbalance intrusion detection of the NSL-KDD dataset for the Normal, Dos, Probe, R2L and U2R intrusion attack types respectively.

Figure ?? is a bar chart that shows the comparison between the balanced and unbalanced class for all the intrusion attacks. This shows that the U2R attack requires the most samples to balance its class followed by R2L, Probe and Dos respectively.

### *3.2. Cascaded Intrusion Detection Model Result*

The result of the cascaded intrusion detection using Random Forest (RF) algorithm and ADASYN balanced dataset for each intrusion class is presented in this section. Table 2 and Table 3 shows the confusion matrices of the four developed models for balanced and unbalanced dataset respectively.

Respectively when compared to unbalance cascaded models.

Table 4 and Table 5 shows the performance evaluation results for the balanced and unbalanced cascaded models respectively.

The results show that the unbalanced cascaded NvU2R model produced the lowest F-measure, precision due to the high imbalance between the positive and negative class. Similarly, the cascaded balanced NvU2R model generated the lowest RMSE and MAE, while it produced the highest Accuracy. This is attributed to the appropriate balancing between the classes of the positive and negative cases.

### *3.3. Performance Evaluation*

Figure ?? and Fig. ?? shows the RMSE and MAE performance evaluation and comparison between the balanced and unbalanced cascaded models respectively. From Fig. ??, the RMSE curve indicate lower RMSE values for balanced models shown in blue curve when compared with the RMSE of the unbalanced models shown in the red curve. The curve also show that the NvU2R model produced the lowest RMSE followed by NvDos, NvR2L and NvProbe.

Figure ??, the MAE for the balanced model shown in the blue curve is lower for all models than the unbalanced models shown in the red curve. This is also as a result of the balancing effect on the dataset for each class.

Figure ??, Fig. ?? and Figure 4.6 are the accuracy, precision and F-measure of the balanced and unbalanced cascaded models. The blue curve indicated higher accuracy for all the balanced models when compared with the unbalanced models in the red curve. Similarly, the precision and F-measure curves for the balanced models shown in the blue curves of Fig. ?? and ?? indicate higher values when compared with the unbalanced models shown in the red curve. The results shown in the graphs in Fig. ??, Fig. ?? and Fig. ?? indicated that balancing the dataset of each class significantly improves the intrusion detection of each class and specifically, the U2R and R2L class where the imbalance between the class are very high.

### *3.4. Performance Validation*

The precision of the proposed model is higher, especially, on the U2R model with higher imbalance. The unbalanced cascaded model produced the second-best result followed by the normal model. For F-measure, the proposed model result is better than that of normal and unbalanced model.

## 4. Conclusion

This proposed method complements the assets of automatic feature deep learning and physical statistical determined improved feature production founded on human in the loop and big data conception that helps the learning model to better correlate the input and output relationship. Explicitly, we introduced a method analysis determined the improved WSN system for intrusion detection. The NSL-KDD dataset was employed as standard to identify the imbalance attack on network traffic patterns. The most Statistical Analysis motivated used or Improved are multi step adaptive synthetic oversampling and random forest cascaded model for multi class Intrusion Detection.

## References

[1] T. Lu, G. Liu, and S. Chang. Energy-efficient data sensing and routing in unreliable energy-harvesting wireless sensor network. *Wirel. Netw.*, 24:611–625, 2018.

## Query

1. Provide Figures 1 – 6.
2. Cite the references in the main work.