An Intelligent Homogenous Model For Prediction Of Network Intrusion Detection Using Synthetic Minority Over Sampling Technique And Local Outlier Factor

Awujoola J. Olalekan, Francisca Ogwueleka, P.O. Odion, Martins Irhebhude
Computer Science Department, Nigerian Defence Academy
ojawujoola@nda.edu.ng

Abstract- The network intrusion detection problem inflicts inestimable problems to the research institution, organization, and industrial areas while the local intrusion prevention methods, like firewalls, access entry control or encryption, had performed below expectation in protecting the networks and systems from increasing numbers of attacks. Recently, there are persistent and continuous forms of different attacks existing on the cyber-space domain and this impel researchers to develop and design robust techniques in order to address the continuous problem. Therefore, because of high dimensionality of network traffics and ineffective performance of conventional machine learning algorithm in the field of intrusion detection, this research work proposed a novel approach of a hybrid solution with Wrapper based-Feature-Selection technique for removing irrelevant attributes, then use local outlier factor (LOF) to remove the outliers in the data and while further utilize synthetic minority over sampling technique (SMOTE) to recognize the oversampling of the minority samples in a bid to constructively increase the prediction accuracy of the minority class under the assumption that the overall distribution are unchanged and the information loss of majority samples decreased. However, this work considered using an ensemble classification algorithm approach, which combines decisions from Adaboost with J48 and also Adaboost with RandomForest. The experimental results confirmed that the proposed method diagnosed the anomalies very effectively, enhance the prediction accuracy, produced a better result in terms of detection efficiency and false alarm rate from the existing problems to other approaches implemented

Keywords: *Intrusion, Ensemble, classification, J48, Local outlier factor (LOF), Synthetic minority over sampling technique (SMOTE), Random Forest, Wrapper, Feature selection*

1. INTRODUCTION

Due to the large use and consumption of the Internet in our day to day activity, the security of our network has become the main pillar to most of the web applications, like online retail sales, auctions, file processing etc. Intrusion detection and classification helps to identify computer threats by examining different information records stored in network processes (Endorf, Schultz, & Mellander, 2004). This can be regarded as one of the outstanding procedures to effectively tackle the problems in network security. An intrusion on the internet can trade-off data security through several internet means. These days, the high growth of data transfer rate, network proliferation, and unpredictable Internet usage have adjoined more anomaly problems. Thus researchers need to develop more effective, reliable, and self-monitoring systems that are worthy to sort troubles and can carry out operations without human interaction. With this type of attempt, fatal failures of susceptible systems can be eliminated or reduced to the nearest minimum (Jabez & Muthukumar, 2015).

Internet is, however regarded as a virtual world that bestow to the people an enterprise opportunities to exercise their activities and services like education, e-commerce, entertainment, and e-business. Additionally, Internet is a domain that bottles up the massive flow of data, which represents organisation or individual privacy, as well as financial transactions of institutions. Influencing on exchange data about Availability, Confidentiality and Integrity, represents threatening aftermath on the systems and networks. At the front end of the increasing network attacks, different security devices have been developed and deployed to protect the systems against attacks such as antivirus software, encryption, authorization mechanism, firewalls, Intrusion Detection System (IDS), and Intrusion Prevention System (IPS) (Sofiane & Mohamed, 2018)

IDS is an important tool in any security infrastructure, which is used to protect a system against attacks. Furthermore, it suffers from computational complexity, response time, and storage requirements. Feature selection (FS) is among preprocessing steps, which searches to decrease the degree of these problems by reducing the number of features. Thus, FS selects the best feature subset(s), which avoid over-head classification problems (Sofiane & Mohamed, 2018).

Data mining problem on imbalanced datasets is the situation where the ratio of majority and minority classes became more cardinal as it exhibits its effect on many real domains. (Deepika, Satya, D, & Hemanth Kumar, 2015). However, the class imbalance problem arises when the instances of some classes are more than the instances of other classes. In this regard, existing classifiers tend to be overwhelmed by the classes with more instances and ignore other classes. In real-world applications, this becomes more drastic as the ratio of small to large classes will be as huge as 1 to 100, 1 to 1,000 or even 1 to 10,000.

An oversampling technique called SMOTE (Synthetic Minority Oversampling Technique) is used to generate synthetic samples of the minority class in order to balance the dataset. The major issue faced by researchers with imbalanced problems is regarding the evaluation measures. As previously mentioned, common evaluation measures such as accuracy can yield misleading conclusions as there can also be an imbalance in costs of making different errors, which could vary per different cases. (Deepika, Satya, D, & Hemanth Kumar, 2015)

An object can be treated as an anomaly or outlier if all its properties are different from the rest properties of the data. Most of the attack records are very interesting because they usually don’t relate to the distribution of malicious software, unauthorized use of system resources, or system malfunction.

In the beginning, the outlier detection sometimes looks like a simple task because all that is needed is to find the data that does not meet the template of normal behaviour. Nonetheless, there are many different methods of detecting outlier detection, but there is still a need for further research for these several reasons.

It is cardinal and noteworthy to be aware that malicious tasks on the network are made up of very minimal compared with normal Internet or cyber network activity. This type of manner or pattern is different from normal user activity, and that’s the reason why it can be easily detected by using outlier detection methods. One of the biggest challenges in this domain is a high amount of data available for analysis and imbalanced data. Even a little false alarm rate would sweep the analyst off the ground. However, the nature of both the normal and abnormal network packets are not constantly stable.

This research work therefore set forth a novel approach for detecting network intrusion anomalies by utilizing an hybrid preprocessing technique that combined synthetic minority over sampling technique (SMOTE), local outlier factors (LOF) and Wrapper feature selection technique.

2. LITERATURE REVIEW

This section express many attempts made by researchers in the area of network intrusion detection system. However, most of the detection works used KDDCUP99 and NSL-KDD datasets. Statistical approach and expert system based are the two major techniques commonly adopted in tackling intrusion detection. Many results are generated while using the KDD cup 99 and NSL-KDD datasets for research work and some of them are briefly discussed.

Arjunwadkar & Thaksen (2015) propose machine-learning model that combines hybrids classifiers as a preprocessor of Intrusion Detection System. The model reduces the feature vector dimension and reduced greatly the time taking to build the model. This invariably reduced the noise created from the features. The proposed hybrid intrusion detection model was performed using NSL-KDD data set, and the experimental result revealed that combination of J48 and Naïve Bayes resulted in accuracy of 99.03%

Kabir, Onik, & Samad, (2017) used Wrapper feature selection approach to remove all the irrelevant features and utilize a bayesian classifier for building the intrusions detection model. The proposed framework yielded an accuracy of 98.3% with a false positive rate of 0.7%.

A hybrid based intrusion anomalous detection model is proposed by (Zohreh & Yue, 2018), where decision tree and k-NN is combined to build an hybrid model. The experiment was performed using feature selection technique for extraction of optimized information from NSL-KDD dataset. The proposed model achieved positive detection accuracy of 99.7% with false alarm rate of 0.2%

Demir & Dalkilic, 2018 presented a stacking ensemble model for identifying and detection of intrusion attacks. The model identifies correctly some considerable classes of attacks. The performance accuracy of the model is 92.55%.

The researcher Samra, Muhammad, & Xiaopeng, (2019) compared the proposed hybrid model that is developed with probabilistic BayesNet and IBK with Bayesnet, J48, JRip, IBK and SMO. The model performance was evaluated with KDDCUP99 dataset and the experimental result showed that the intrusion miner approach, output an accuracy of 96.1%.

Manal, Arwa, Asmaa, & Soad, (2018) propose an Enhanced Intrusion Detection System using Feature Selection Method and Ensemble Learning Algorithms. The performance of the proposed technique is compared and evaluated using cross validation test and testing dataset of NSL-KDD dataset. However, the accuracy performance of the model shows that 99.7984% accuracy is obtained when just 19 feature set and product probability rule was used.

The intrusion detection system using Bagging with partial decision tree base classifier proposed by Gaikwad & Ravindra, (2015) is appraised using true positives, classification accuracy, model building time and false positive. It was distinguished that the system obtained a good classification accuracy of 99.7166% with cross validation.

Adebola, Stephen, & Oluwabukola, (2019) propose a stacking ensemble using Naïve Bayes, random forest and J48 classifiers as base learners while SVM is used as the meta learner. The model achieved detection accuracy of 99.5% and a false positive rate of 0.6%

Pham, Foo, Suriadi, Jeffrey, & Lahza, (2018) combined two feature selection techniques Feature vitality-based reduction method (FVBRM) and gain ratio to extract relevant features from NSL-KDD dataset..They further combined two ensemble classifiers such as bagging and boosting with different decision trees as the base learners. The experiments revealed that bagging with J48 had an accuracy of 84.25% and a very high false alarm rate of 2.79%

3. Research Methodology of the proposed Model

A homogenous hybrid approach is proposed for this IDS and the objective behind this approach is to bring forth mechanisms to ameliorate the classification and detection of network intrusion. In order to achieve this goal, we first deal with the problem of feature selection by extracting all the irrelevant features in the dataset and then performed the classification experiment with the ensembles and single classifiers. Table 1 shows the accuracy result of the first experiment. We then address the problem of an imbalanced dataset by subjected the preprocessed wrapper dataset to SMOTE and LOF. The result is shown in table 2. Therefore, this proposed model adopted a hybrid preprocessing technique by combining both Synthetic Minority Oversampling Technique (SMOTE) and local outlier factor (LOF) as the preprocessor. The proposed model methodology flow is shown in Figure1.

3.1 Local Outlier Factor

The local outlier factor is built on the idea of a local density, where a spot is given by k nearest neighbors where distance is used to evaluate the density. While juxtaposing the local density of an object to the local densities of its neighbors, therefore we can pinpoint the area of similar density, and points that have a substantially lower the density than their neighbors. These are considered to be outliers. Equation 1 represents the fundamental idea of LOF while figure 2 depicts the outlier.

3.1.1 Outlier Detection

Mostly, outliers are described as objects that do not comply with general behavior or model of the data (Han & Kamber, 2011). But as a result of variety of the data been processed in data mining, it is difficult to objectively define outlier. Outlier can therefore be identified quite easily because of substantially different (e.g. numerical) values compared to other non-outlying objects.

$$\boldsymbol{LOF(p)} = \left(\frac{\sum_{q\in nn(p)}\frac{\boldsymbol{LRD(q)}}{\boldsymbol{LRD(p)}}}{\|\boldsymbol{k-neighbourhood}\|}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \mathbf{1}$$

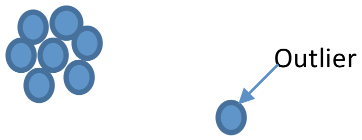

Figure 2: Outlier

3.2 SMOTE (Synthetic Minority Oversampling Technique

Datasets can be referred to as imbalanced if their classification classes are not uniformly distributed. However, network datasets majorly made up of permissible traffics with only a few percentages of illegitimate traffics. Therefore, oversampling by reduplication of subsisting minority samples can lead to a more specific neighborhood in feature space. This can however result in overfitting on minority class samples.

SMOTE is usually used to address those problems by generating synthetic examples in which it operates in “feature space” instead of operating in a “data space”. SMOTE algorithm generally appraises the minority class instances and oversamples it by producing synthetic instances joining all of the k minority class nearest neighbors. However, significance of k largely depends on the total amount of oversampling to be done. SMOTE process always commences by determining some point y_i and selecting its nearest neighbors y_{i1} to y_{ik}.

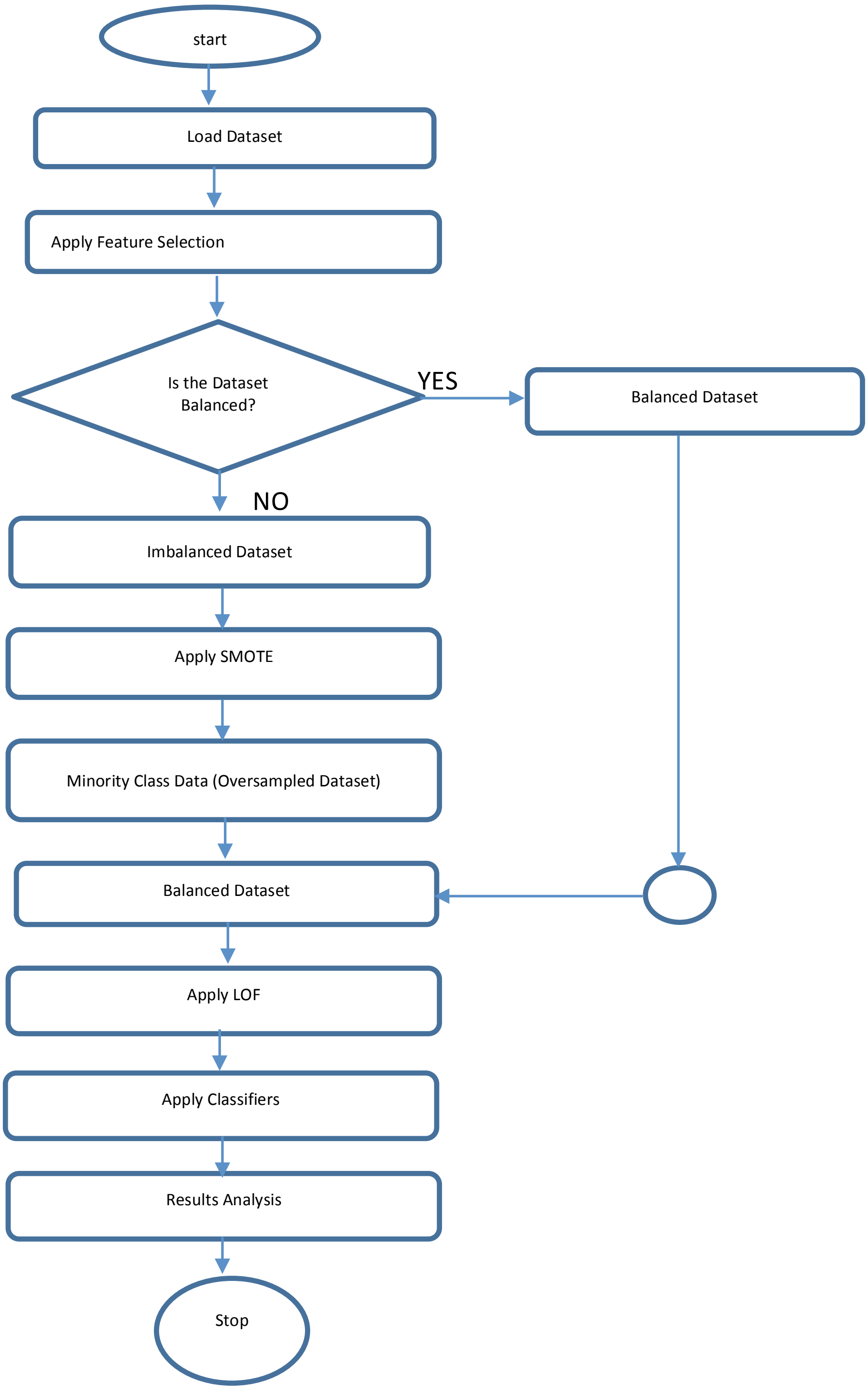

Figure 1: Proposed Flow of methodology. Source: (Researcher's Model, 2020)

Therefore, numbers from r1 to rk are produced by randomized interpolation of the chosen nearest neighbors (Deepika, Satya, D, & Hemanth Kumar, 2015).

3.3 Ensemble classification

The ensemble contains quite a numbers of learners that are mostly produced from the training set with the help of a base learner classifier (Awujoola, Francisca, & Odion, 2020). These classifiers are very critical and powerful to solve the same problem and all together achieve a forecasting result with higher stability and accuracy by creating multiple independent models and combining them (Li & Sun, 2013). The classical reasons for employing ensemble classifiers to improve the effectiveness and representational issue, a statistical classifier is not qualified to obtain the best representation in the hypothesis space, and therefore, it is necessary to combine independent classifiers to improve the predictive performance.

The thought of ensemble classification algorithm is to discourage the use of one single classifier but combining a set of classifiers called an ensemble of classifiers, then combine their predictions or forecast for the classification of unseen data (Awujoola, Francisca, & Odion, 2020).

3.4 NSL-KDD Intrusion Dataset

The records in the NSL-KDD were carefully selected based on the shortcoming of KDD99. However, records of different classes are balanced in the NSL-KDD, which avoids the classification bias problem. The NSL-KDD also removed duplicate and redundant records; therefore, it contains only a moderate number of records. Therefore, the experiments can be implemented on the whole dataset, and the results from different papers are consistent and comparable. The NSL-KDD alleviates the problems of data bias and data redundancy to some degree. The NSL-KDD does not include new data; thus, minority class samples are still lacking, and its samples are still out-of-date. (Hongyu & Bo, 2019). Every instance of this dataset has 41 features and is labeled as either normal or attack. The attack types fall into the following four classes:

Denial of Service (DoS) attacks: In this type of attack, the intruder tries to keep the network busy by exploiting the bandwidth that results in denial of service for legitimate requests.

User to Root (U2R) attacks: By logging in as a normal user, the intruder tries to access the system with root privilege.

Remote to Local (R2L) attacks: The attacker tries to locate vulnerability to access the system remotely.

Probing: The intruder tries to gather information about a network in order to bypassing its security policy.

4. Results and Analysis

Experiments are carried out using Dell Microsoft Windows 10 system with Intel® Core ™ i7 CPU, 2.30 Ghz processor, and 8.00 GB RAM. However, WEKA open source program is used to evaluate the method, build the model, and perform feature selection. 10-fold cross-validation is adopted to perform on the training dataset so as to choose the optimal training parameters.

Table 1. Classification Accuracy Result with Wrapper Approach technique

Classifiers	**Correctly classified %**	**Incorrectly classified**	**MAE**	**RMSE**	**RAE %**	**RRSE**	**KAPPA STATISTICS**
J48	99.7182	0.2818	0.004	0.0504	0.7978	10.107	0.9943
RandomForest	99.7737	0.2263	0.0044	0.042	0.8879	8.425	0.9955
Adaboost(J48)	99.8134	0.1866	0.002	0.0424	0.3974	8.493	0.9963

Classifiers	**TP-RATE**	**FP RATE**	**PRECISION**	**RECALL**	**F-MEASURE**	**ROC**	**CONFUSION MATRIX**
J48	0.997	0.003	0.997	0.997	0.997	0.998	a b 13424 25 46 11697
RandomForest	0.998	0.002	0.998	0.998	0.998	1.000	a b 13429 20 37 11706
Adaboost(J48)	0.998	0.002	0.998	0.998	0.998	1.000	a b 13436 13 34 11709

Adaboost Ensemble with J48 classifier has the highest classification accuracy in table 1 with performance accuracy of 99.8134% with Kappa statistic of 0.9963, while Random Forest and J48 classifiers followed with 99.7737% and 99.7182 respectively. Also confusion matrix showed that the misclassified class is smaller

Table 2. Prediction Result with Wrapper approach, SMOTE and LOF (Proposed)

Classifiers	**Correctly classified**	**Incorrectly classified**	**MAE**	**RMSE**	**RAE**	**RRSE**	**KAPPA STATISTICS**
J48	99.6372	0.3628	0.0049	0.0578	1.0512	12.0111	0.9922
Random Forest	99.8051	0.1949	0.0044	0.0412	0.9515	8.5554	0.9958
Adaboost (J48)	99.7699	0.2301	0.0023	0.046	0.5056	9.5675	0.995

Classifiers	**TP-RATE**	**FP RATE**	**PRECISION**	**RECALL**	**F-MEASURE**	**ROC**	**CONFUSION MATRIX**
J48	0.996	0.004	0.996	0.996	0.996	0.998	a b 13400 49 85 23401
RandomForest	0.998	0.002	0.998	0.998	0.998	1.000	a b 13428 21 51 23435
Adaboost(J48)	0.998	0.002	0.998	0.998	0.998	1.000	a b 13422 27 58 23428

Table 2 shows significant improvement in Random Forest performance accuracy with 99.8051% against 99.7737 in table 1. This is as a result of the combination of SMOTE and LOF. Therefore Random forest can be enhanced with LOF and SMOTE for better performance on intrusion detection

5. Conclusion

The most important contributions of this work are the integration of wrapper dimensionality reduction approach and the ensemble technique that resulted in the construction of efficient and accurate attack detection. Also, the hybrid preprocessing technique adopted for the imbalance of the dataset using SMOTE and detection of the LOF resulted in accurate detection/ classification of the attack in the dataset using Random Forest.

REFERENCES

Adebola, A., Stephen, M., & Oluwabukola, A. (2019). An Ensemble of classification techniques for Intrusion Detection Systems . *International Journal of Computer Science and Information Security (IJCSIS),* , 24-33.

Arjunwadkar, N. M., & Thaksen, J. P. (2015). An Intrusion Detection System, (IDS) with Machine Learning (ML) Model Combining Hybrid Classifiers. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, 647-651.

Awujoola, O. J., Francisca, O., & Odion, P. O. (2020). Effective and Accurate Bootstrap Aggregating (Bagging) Ensemble Algorithm Model for Prediction and Classification of Hypothyroid Disease. *International Journal of Computer Applications, 176*(39), 40 - 48.

Deepika, D., Satya, D. B., D, A., & Hemanth Kumar, R. D. (2015). REVIEW ARTICLE IMBALANCED DATASETS. *International Journal of Current Research, 7*(04).

Demir, N., & Dalkilic, G. (2018). Modified stacking ensemble approach to detect network intrusion. *Turkish Journal of Electrical Engineering & Computer Sciences*, 418-433.

Endorf, C., Schultz, E., & Mellander, J. (2004). *Intrusion detection and prevention.* California: Mc Graw-Hill.

Gaikwad, D., & Ravindra, C. T. (2015). Intrusion Detection System Using Bagging with Partial Decision TreeBase Classifier. *Procedia Computer Science, Elsevier*, 92-98.

Han, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (3rd edition, ed.). San Francisco: Morgan Kaufmann Publishers Inc.

Hongyu, L., & Bo, L. (2019). Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey. *Applied Sciences, 9*(4396), 1-28. doi:doi:10.3390

Jabez, J., & Muthukumar, B. (2015). Intrusion Detection System (IDS): Anomaly Detection using Outlier Detection Approach. *International Conference on Intelligent Computing, Communication & Convergence (ICCC-2015)* (pp. 338 – 346). Odisha, India: Elsevier ScienceDirect.

Kabir, M., Onik, A., & Samad, T. (2017). A network detection framework based on bayesian network using wrapper approach. *International Journal of Computer Applications, 166*(4), 13-17.

Li, H., & Sun, J. (2013). Predicting business failure using an rsf-based case-based reasoning ensemble forecasting method. *Journal of Forecasting*, 180–192.

Manal, A., Arwa, A., Asmaa, B., & Soad, A. (2018). Enhanced Intrusion Detection System using FeatureSelection Method and Ensemble Learning Algorithms. *International Journal of Computer Science and Information Security (IJCSIS), 16*(2), 48-55.

Pham, N., Foo, E., Suriadi, S., Jeffrey, H., & Lahza, H. F. (2018). Improving performance of intrusion detection system using ensemble methods and feature selection. *Australasian Computer Science Week*, 1-6.

Samra, Z., Muhammad, K., & Xiaopeng, H. (2019). Intrusion-Miner: A Hybrid Classifier for Intrusion Detection using Data Mining. *International Journal of Advanced Computer Science and Applications (IJACSA), 10*(4), 329-336.

Sofiane, M., & Mohamed, T. (2018). Feature Selection Algorithms in Intrusion Detection System: A Survey. *KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS, 12*(10), 5079 - 5099. Retrieved from http://doi.org/10.3837/tiis.2018.10.024

Yizhou, Y., & Lei, C. (2017). DistributedLocalOutlierDetectioninBigData. *KDD 2017 Research Paper*.

Zohreh, A. F., & Yue, L. (2018). Intrusion Detection System by Using Hybrid Algorithm of Data Mining Technique. *7th International Conference on Software and Computer Applications* (pp. 119–123). ACM Digital Library.