Intrusion Detection System as Audit in IoT Infrastructure using Ensemble Learning and SMOTE Method

by Moch Arief Soeleman

Submission date: 21-Apr-2020 10:05AM (UTC+0700)

Submission ID: 1303307085

File name: paper_Intrusion_Detection..SMOTE_Arief_Soeleman.pdf (278.27K)

Word count: 4796

Character count: 26483

Intrusion Detection System as Audit in IoT Infrastructure using Ensemble Learning and SMOTE Method

1st Aldhi Ari Kurniawan
Faculty of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
aldhiarikurniawan@gmail.com

2nd Heru Agus Santoso
Faculty of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
herezadi@gmail.com

3rd M.Arief Soeleman
Faculty of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
arief22208@gmail.com

4rd Ahmad Zainul Fanani
Faculty of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
a.zainul.fanani@dsn.dinus.ac.id

***Abstract*—With the industrial revolution 4.0, the use of IoT-based systems is increasing, both in the field of health manufacturing, urban planning, housing, and even automotive. Therefore, the security of the IoT system needs to be considered, this is related to data integrity, privacy, service stability. Through intrusion detection, activities on the IoT system will be able to be analyzed whether there are suspicious activities that can interfere with or threaten IoT services. In several previous studies in the literature, the approach used to detect intrusions in the IoT system has a high false alarm rate. This research proposes an approach through machine learning, specifically the ensemble learning approach and the synthetic minority over-sampling technique (SMOTE) method as a method of detecting intrusions in the IoT system which is expected to produce better performance. The results of this study indicate that the proposed approach is able to detect intrusion and classify into five types of intrusion including normal intrusion, probe, dos, r2l, u2r. Based on the evaluation results, the proposed approach can improve the performance of intrusion detection in terms of accuracy to 97.02%, detection rate of 97%, false alarm rate 0.16%, compared to base learning and approaches in previous studies used as intrusion detection methods, but in processing time performance have not shown satisfying results.**

Keywords*—*IoT, Intrusion Detection, Audit, Machine learning, Ensemble learning.

I. INTRODUCTION

Industry 4.0 is very closely related to the Internet of Things (IoT), this is because the main element of the 4.0 industrial revolution is IoT[1]. IoT has influence on various industries such as manufacturing, health, logistics, housing, urban planning, agriculture, and even the automotive industry[2]. In the field of manufacturing IoT can be used as a link between production machines to run efficiently, besides that it can also function as management to monitor production flow. In addition, the inventory of goods has also used IoT, thus making the efficiency of information flow of goods[3]. With the presence of IoT, in the automotive field, the car industry has also implemented autonomous driving, it is possible that in the future many car manufacturers will use IoT to exchange information between cars.[4].

In the world of business data is an important asset[5] The data that is processed will be able to provide information for humans or machines, which of course invites people to try to hack the data. Therefore data security in the industrial revolution era 4.0 is a formidable challenge[5]. The use of IoT in Industry 4.0 is a new concept for many people[6]. The original concept of IoT is intelligence intelligence and automation control. The government and various agencies are now keen to introduce the use and benefits of smart technology. Despite its sophistication and technological intelligence, IoT still has a gap. Of 1,150 respondents in the Asia Pacific region surveyed by Aruba, a subsidiary of Hewlett Packard Enterprise, 88 percent reported having experienced IoT-related security breaches, from government, health, manufacturing and retail sectors, there were thousands of data security cases. Another case occurred in Tiongok, 89 percent of health services there had been broken into[7]. Nearly 6,000 videos of sick newborns from a hospital in eastern China leaked to the public. This baby video is actually intended for parents to be able to monitor their children from anywhere.

In the manufacturing sector burglary has reached 82 percent because of security system problems. Of all the manufacturing companies that experienced security breakdowns related to IoT, half were related to malware, while 40 percent were caused by human error. In the retail sector 76 percent have experienced security breaches, taking into account retailers who have suffered losses due to IoT-related attacks due to malware issues, it is clear that these businesses need to find a middle ground between providing a seamless and integrated shopping experience by protecting their networks from any attack. In the IoT infrastructure detection of intrusion is very important[8]. With the rapid development of IoT, safety factors have been considered and become the most challenging topics in the framework of developing IoT infrastructure[9]. Hackers and viruses can thwart data exchange and integrity. In addition, data in security can specifically reduce the security of the entire IoT system and carry risks in its use. All activities, behavior, and use of the external are monitored by the IoT on the entire network in the system[8]. Every company in the fields of manufacturing, health, urban planning, automotive, must be able to guarantee the security of legality, as well as the quality of information and services used.

Audit on the system is a task that must be completed with a system created by the company. Logs on an IoT-based system have a record of every operation performed on that IoT device. Finding anomalies in IoT-based systems Required expert staff capable of analyzing log data systems. In the audit system to find or comment on the manual will require a long time, this requires a system log data that contains data with very large dimensions. Intrusion Detection system is a method for analyzing activities carried out on the system through log data on the system. In addition to detecting intrusions from outside, the intrusion detection system also checks the behavior of users from inside who are doing suspicious activity[9]. To ensure the security of using IoT-based services, a system is needed to detect anomalies or intrusions.

From the IoT security problems there are approaches that might be used to solve them, one of which is through the approach of intrusion detection systems based on machine learning. Machine learning is a knowledge-based learning method obtained from training data which is then used as a model for classifying, classifying, associating, and predicting. To detect intrusion of the IoT system it is necessary to do a learning process on the IoT log data to obtain knowledge that can be used to be able to classify based on the type of intrusion on the IoT system.

II. Related Research

Several studies relating to safety issues in previous IoT, Pajouh et al. Research conducted by hamed et al, with a reduction module and two levels of classification they detect intrusion in the two-layer dimension, it is intended for the detection of User to Root (U2R) activities and Remote to Local (R2L). Components of analysis and analysis of linear discrimination are used to reduce high to low dimensions, and then do the classification with two levels through Naive Bayes, Certainty Factor, from K-NN to detect intrusion in the system. In the study of hamed et al showed that the proposed method can outperform other methods in previous studies[10].

Khreich et al. Detection of system anomalies when running at the host level is a challenge in system security analysis. When detecting an anomaly in a large scale in previous studies, it still obtains a high level of false alarms. Using the one-class detection approach on supports vector machines (OC-SVM) and merging frequencies with temporal information in logs. the system as well as through feature extraction techniques, this research tries to suppress the false alarm level from system intrusion detection. In the extraction feature approach that has been proposed, the process starts from log segmentation on the system into several n-gram partitions based on variable lengths and then mapped into vectors, from the extraction feature then used as training on OC-SVM detectors. From the proposed approach, the results show that the features of n-grams in vectors are able to outperform the performance of vectors using the most commonly used burglary methods used in related research. A high accuracy rate and a low false alarm rate are obtained obtained with the Markov approach along with n-gram in the anomaly detection system, in addition to the approach to detect anomalies through OC-SVM and gausian karnel in vectors, can obtain higher detection accuracy[11].

Mohamudally et al. The multiphase aspect is shown in this study in IoT applications and networks with real usage and problems that arise from the use of anomaly detection machines in perspective on network convergence and also in software. Based on various models of comparative time series that are used to detect an anomaly in the system that has been done, shows that the alternative plug and play is not an appropriate size, besides that in machine learning the unsupervised approach is the most flexible and considered approach more efficient as an approach for analyzing the IoT system[12].

Li et al. Based on the approach through deep learning for intrusion detection systems in smart cities and feature extraction is done on this research. The schema of the approach through deep learning and feature extraction was introduced in this study. The dataset used uses KDD CUP99 to conduct experiments, using 10% of the dataset as training data. By comparing the proposed approach with the approach that has been carried out in previous studies, it is obtained that the proposed approach is able to produce an intrusion detection system that is faster and more efficient than the high detection results.[13].

Mohammadi et al. The demand for network protection and security against cyber attacks is increasing, due to the current widespread network connectivity. Network security systems can be supported by intrusion detection systems. Through the filter and wrapper approach to clustering and the selection of features this study was carried out to support the IDS system. The filter and wrapper methods are named based on the feature grouping of the linear correlation coefficient algorithm (FGLCC) and the Cuttle Fish Algorithm (CFA) algorithm, respectively. Cuttle Fish algorithm or commonly called CFA and FGLCC which is a grouping of linear correlation coefficient features is the method chosen in the wrapper and filter approach. The dataset used by KDD CUP99 to be used as training and testing, based on the decision tree is used as a cluster. The results of the study showed that accuracy was better with a lower false alarm rate of (1.65%) compared to the method in previous studies[8]

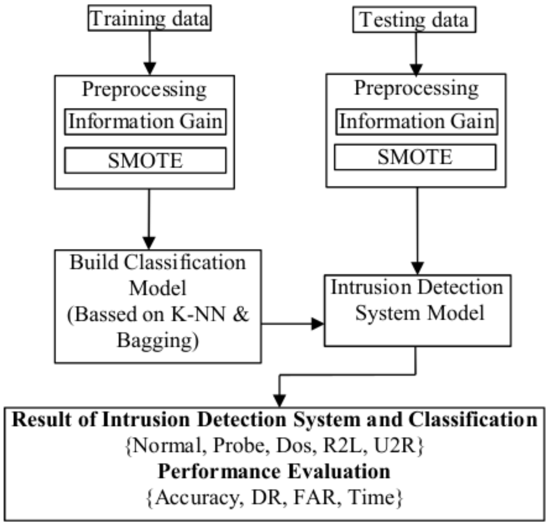
From some of these studies, it is known that intrusion detection on IoT is important to ensure system security, data, and connectivity, so this research will try to find a more effective and efficient approach from the model in some previous studies. Basically, several machine learning methods applied to intrusion detection have weaknesses with large volumes of datasets, and imbalance data. Large dataset has an influence on the long time when training data that will become a learning model, besides the Imbalance dataset will also affect accuracy. Through the selection of features and methods based on Ensemble Learning and Synthetic Minority Over-sampling Technique (SMOTE) approach, it is hoped that the learning model can be more effective and efficient.

Dataset with large dimensions is likely to be reduced through feature selection. In addition the ensemble learning approach makes it possible to improve the performance of learning models and the smote technique has the ability to deal with class imbalances by synthesizing new samples from minority classes so as to balance the dataset.

III. Purposed Method

In this study, the researcher proposes an ensemble learning-based intrusion detection method, synthetic minority over-sampling technique (SMOTE) and selection features as an approach to improve the performance of the clasifier. So that it is expected to be able to build an intrusion detection model by training and testing data that is fast, strong performance against imbalance data and has high accuracy. Details of the proposed method will be explained in detail in the following explanation:

→ Training dataset
→ Feature selection (Information Gain)
→ Resample dataset (SMOTE)
→ Ensemble Learning (Bagging)
→ Clasifier (K-NN)
→ Testing
→ Performance Evaluation



Training data
Testing data
Preprocessing
Information Gain
SMOTE
Preprocessing
Information Gain
SMOTE
Build Classification Model (Bassed on K-NN & Bagging)
Intrusion Detection System Model
Result of Intrusion Detection System and Classification
{Normal, Probe, Dos, R2L, U2R}
Performance Evaluation
{Accuracy, DR, FAR, Time}

Fig 1. Proposed Method

From the picture above it can be explained that the dataset which consists of training data and testing data in the first stage is carried out preprocessing with the Information Gain method as a feature selection method, to determine the relevant attributes to be used as learning. Then the data is balanced using the SMOTE method. The next stage is the formation of intrusion detection models through Ensemble Learning (Bagging) with K-NN as a classification method and learning of training data. After the intrusion detection model is formed, then it is tested by using preprocessing testing data. The results of the intrusion detection model testing will later produce a classification based on the type of intrusion including Normal, Probe, Dos, R2L, U2R, then Performance Evaluation is carried out with several parameters including Accuracy, Detection Rate, False Alarm Rate and Processing Time. From the results of several evaluation parameters to determine whether there is an increase in performance, a comparison is made with based learning and some intrusion detection methods that have been done previously in the literature.

A. Information Gain for Features Selection

Information Gain (IG) is widely used in high dimensional data to measure the effectiveness of features in classification. This relates to the amount of information expected, namely the reduction in entropy[14].

$$\text{Info}(D) = -\sum_{\text{i}=1}^{v} p_i \log_2(\text{pi}) \qquad \text{Eq. (1)}$$

The formula description is:
c: number of values that exist in the target attribute (number of class classifications)
pi: the number of samples for class i

$$\text{Info}_A(D) = \sum_{\text{j}=1}^{v} \frac{D_j}{D} \text{xInfo}(D_j) \qquad \text{Eq. (2)}$$

The formula description is:
A: attribute
| D | : total number of database samples
| Dj | : number of samples as value j
v: value for attribute A

Furthermore, the information gain value used as a measure of the relevance or effectiveness of an attribute in the data classification is calculated using the formula below:

$$\text{Gain}(A) = \left|\text{Info}(D) - \left(\text{Info}_A(D)\right)\right| \qquad \text{Eq. (3)}$$

Getting higher information means better discriminatory power for decision making[8] Information acquisition is a good measure to determine the relevance of classification features. The importance of features to decision making in our model is done by evaluating them by measuring information acquisition[15]. Not all data attributes are created equal and not all contribute equally in decision making. Because of that its attributes can be sorted in the order of their contribution in decision making by listing features in the order of information acquisition score reduction[15].

B. SMOTE Method

The SMOTE algorithm calculates the distance between training points of minority classes aimed at defining the environment, and then an example is chosen to make points from new synthetics. The distance calculation can be calculated through manhattan or ecludean distance calculation[16]. There are two important steps to solve dataset problems that have high data dimensions: First, through distance calculation using euclidean [17]. Furthermore, Euclidean distance will assume that each attribute in the dataset is equally important to be defined in the environment of the SMOTE algorithm, but datasets that have high dimensions often have a percentage of data redundancy and variables that are irrelevant and also have noise. In the SMOTE engineering approach, synthesized child generations can be defined as follows:

$$x^{syn} = x^i + (x^j - x^x i .* y) \qquad \text{Eq. (4)}$$

Where x^i is a minority class that is considered, x^j is a randomly chosen derivative from the k-neighbor nearest minority from x^i; and γ is a vector where each element is a random number from [0, 1]; as well as symbols ". " shows the multiplication of elements[18].

C. Ensemble Learning

Ensemble learning is a machine learning paradigm in which some basic classifications are trained and then aggregated by building a final classifier. The ability to generalize an ensemble learning has proven that can achieve better performance than the basic classification when there are significant differences between the basic classifiers[19].

Bagging, as another ensemble learning method, is divided into two stages:

First, it produces a single learning model, by emphasizing diversity, in the second step, the model is combined, generally by using a merging function[19].

Bagging algorithm.

```
1. Model generation
2. Initialize the parameter
        → the ensembel $\varepsilon=\emptyset$
        → the number $n$ of instance of the training data;
        → the number $m$ of learner
        → the algorithm $A$
        → a set of prediction $P_1,\ldots,P_q$;
For j = 1...m;
        → generate a bootstrap sample $D_i$ instances from the training set;
        → builde the learner $C_I$ training A on $D_i$;
        → add the learner $C_i$ to the learner set, $\varepsilon = \varepsilon \cup C_i$;→
        return $\varepsilon$
Testing
Apply $C_i \ldots C_m$ on the new instance i calculating $C_j$ (I)
Calculate the ensemble by:
        $P\varepsilon = arg \frac{max}{k} \sum_{j=1}^{n} x\left(c_j \middle| i\right) = P_k$    Eq. (5)
```

D. IBk Classifier

IBk Classifier - In the K-Nearest Neighbor (K-NN) classifier, predictions are made based on the relative node spacing of instances of each class. There are no fixed values of K that are suitable for all domains, and the algorithm uses cross-validation K to select the appropriate value. The advantage of using the Ibk Clasifier is that the preferred fundamental perspective taken in utilizing lazy learning strategies, for example, case-based thinking, is that objective capacity will be estimated locally. Because objective capacities are estimated locally for each question in the framework, a sluggish learning framework at the same time can deal with a variety of problems and settings effectively with changes in the problem area[20].

Stages in the K-NN classification:

1. Determine the parameter values from the nearest neighbor.

2. Calculate the distance between the Manhattan training data with testing data.

E. Performance Based Evaluation

Three performance measures are used to evaluate the performance of the proposed approach. Some of these performance measures include the Detection Rate (DR), Accuracy Rate (DA), False Alarm Rate (FAR) which are defined in the following equation[20].

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad \text{Eq. (6)}$$

$$DR = \frac{TP}{TP+FN} \quad \text{Eq. (7)}$$

$$FAR = \frac{FP}{FP+TN} \quad \text{Eq. (8)}$$

where true positive (TP) is the number of intrusions that have been classified correctly, for true negative (TN) is the amount that comes from normal notes that have been properly classified, false positive (FP) is the amount that comes from classified normal notes as intrusion, while for false negative (FN) is the amount derived from intrusion that has been classified as normal[18].

IV. EXPERIMENT AND RESULT

A. Dataset Description

The NSL-KDD dataset is a dataset consisting of logs on a network system that has 41 attributes including (eg protocol types, flags and services), which are normally labeled or some of 24 types of system intrusion classes (eg R2L, Probe, U2R , and Dos)[10]. NSL-KDD has a training set and set as a test[10]. Catatan anomali pada dataset NSL-KDD dikategorikan ke dalam empat jenis :

• DoS: Denial of Service attacks such as Teardrop, Smurf, and Neptune.
• Probe: another type of attack that is sometimes called Probing such as Portsweep, and Saint.
• U2R: Attacks from Users to Root like Rootkit, Buffer_overflow, and Module load.
• R2L: Local long-range attacks such as Xsnoop, Httptunnel, and Password.

Before the dataset is used for experiments, data transformation is first performed, this data transformation is applied to the protocol_type feature, to the nominal data feature, consisting of (tcp, udp, icmp) each nominal feature is changed to (TCP = 1, UDP = 2, ICMP = 3). The dataset used has 125973 records for training and 1112 for testing.

B. Eksperiment Result and Discussion

The results of feature selection through information gain have 28 best features out of 41 features, here are the selected features and their weight values.

TABLE I. RESULTS OF SELECTION FEATURES WITH INFORMATION GAIN

Feature Selected	Information Gain
src_bytes	1.0322102
service	0.8190565
diff_srv_rate	0.7304628
flag	0.7026954
dst_bytes	0.6621512
2me_srv_rate	0.6601568
dst_host_diff_srv_rate	0.6498302
count	0.6498302
dst_host_srv_count	0.5978336
dst_host_same_srv_rate	0.5762663
dst_host_serror_rate	0.5733752
serror_rate	0.5522771
st_host_srv_serror_rate	0.5385294
srv_serror_rate	0.5149715
logged_in	0.4419606
2t_host_srv_diff_host_rate	0.3760204
dst_host_same_src_port_rate	0.3389140
dst_host_count	0.3001916
srv_count	0.2372335
srv_diff_host_rate	0.2097353
dst_host_rerror_rate	0.1388035
protocol_type	0.1199937
dst_host_srv_rerror_rate	0.1198405
rerror_rate	0.1095781

duration	0.0816534
srv_rerror_rate	0.0784914
hot	0.0340066
is_guest_login	0.0163462

The classification results are based on the 28 best features selected with gain information compared to before feature selection obtained as follows:

TABLE II. COMPARISON AFTER FEATURE SELECTION

Porposed method	**Accuracy**	**DR**	**FAR**	**Time**
No Feature selection	96.66%	96.7%	0.33%	43.78
With feature selection	**96.75%**	**96.8%**	**0.32%**	**34.64**

From table II it is known that the results of the selection feature have an impact on the first few indicators in terms of classification accuracy increased by 0.09% compared to the results of the classification without going through the selection feature, the detection rate increased by 0.1%, false alarm rate improved by 0.01%, and processing time experiencing an acceleration of 9.14 seconds. Next is the difference in results from the proposed method which has been through the information gain, smote, ensemble learning approach can be seen in table 4

TABLE III. COMPARISON OF THE RESULT OF THE PROPOSED METHOD WITH THE BASIC METHOD IN THE IMBALANCE DATA

Porposed method	**Accuracy**	**DR**	**FAR**	**Time**
Ibk Clasifier	96.66 %	96.7%	0.33%	43.78
Porposed Method	**97.02%**	**97%**	**0.16%**	**299.6**

Comparison of the results of the proposed method with base learning on the imbalance dataset in table III, it is known that there are changes in the indicators including accuracy has increased 0.46% better than based learning, Detection Rate increased 0.33% better, false alarm rate improved than before to 0.16%, but the processing time has slowed to 299.6 seconds. To get the best results, we performed tests on several algorithms, a comparison of test results can be seen in table IV.

TABLE IV. COMPARISON OF PORPOSED METHODS WITH SEVERAL METHOD

With Smote, Bagging	**Accuracy**	**DR**	**FAR**	**Time**
RandomForest (RF)	96.66%	96.7%	0.29%	1131.35
BayesianNetwork (BN)	94.86%	94.9%	0.26%	33.9
Naive Bayes (NB)	60.75%	60.8%	0.98%	16.19
Porposed(Ibk Clasifier)	**97.02%**	**97%**	**0.16%**	**299.6**

The results of the comparison of the proposed method with several other methods can be seen that the Ibk Clasifier is superior in terms of accuracy with a value of 97.02%, a detection rate of 97% and a false alarm rate of 0.29%, compared to the Random Forest method with an accuracy of 96.66%, Naive Bayes 60.75%, And Bayesian Network is 94.86%, but in terms of processing speed slower than Naive Bayes, and Bayesian Networks with a processing time of 299.6 seconds. Here is a comparison of the detection rates of the four methods based on the type of intrusion.

TABLE V. DETECTION ACCURACY FOR EACH INTRUSION

Type of Intrusion	**Detection Rate**			
	Proposed Method	RF	NB	BN
Normal	**98%**	98.8%	48.1%	95.8%
Dos	**97.7%**	98%	94.6%	96.6%
R2L	**77.6%**	67.3%	55.1%	69.4%
Probe	**98%**	92.9%	11.2%	95.9%
U2R	**98.2%**	97.3%	74.2%	94.7%

Table V shows that the proposed method has the best intrusion detection accuracy in U2R intrusion of 98.2% while in other types of intrusion the proposed method produces an accuracy of 97.7% Dos, 77.6% R2L, 98% Probe and 98% normal. In the random forest intrusion detection method produced 98.8% Normal, 98% Dos, 67.3% R2L, 98% Probe, 98.2 U2R. Naive Bayes 48.1% Normal, 94.6% Dos, 55.1% R2L, 11.2% Probe, 74.2% U2R, while Bayesian networks produce 95.8% Normal accuracy, 96.6% Dos, 69.4% R2L, 95.9% Probe, and 94.7% U2R. From some of these methods Naive Bayes has very low accuracy compared to other methods, especially in the detection of Probe, Normal and R2L intrusions.

TABLE VI. DETECTION RESULT FROM THE TEST DATASET

Intrusion Detection	**Normal**	**Dos**	**R2L**	**Probe**	**U2R**
True Detection	582	343	31	96	19
False Detection	13	4	9	14	1

From table VI of the test dataset totaling 1112 through the proposed method, it is known that 582 detected normal intrusions with 13 false detections, 343 detected as Dos intrusions with 4 false detections, 31 detected as R2L intrusions with 9 incorrect detections, 96 detected as intrusions Probe with 14 false detections, and 19 detected as U2R intrusions with 1 wrong detection.

The following is a comparison of the performance of intrusion detection methods in previous research which will be presented in table VII

TABLE VII. COMPARISON WITH PREVIOUS STUDIES

Researcher	Method	Accuracy	DR	FAR
Mohammadi et al	FGLCC-CFA	95.03%	95.23%	1.65%
Guo et al	KNN-Kmeans	93.29%	91.26%	0.78%
Pajouh et al	TDTC-KNN	N/A	84.86%	4.86%
Al-Yaseen et al	SVM-ELM-Kmeans	95.75%	95.17%	1.87%
Proposed Method		**97.02%**	**97%**	**0.16%**

From table 8 it is known that from the comparison of intrusion detection results with previous research conducted by several researchers with different methods, the approach proposed in this study is superior in terms of accuracy, detection rate, and false alarm rate.

V. CONCLUSIONS

With the IoT-based industrial revolution as its infrastructure, what needs to be considered is data, information security and stable services, intrusion on IoT infrastructure can jeopardize data integrity, user privacy, and stability of IoT-based services. Intrusion detection on IoT Infrastructure will be very useful in the industrial revolution, aiming to protect IoT-based services. The results of research conducted to detect intrusion with the ensemble learning approach, synthetic minority over-sampling technique, can increase the accuracy of intrusion detection to 97.02% and detection rate is 97% with a false alarm rate of 0.16%. In addition to the ensemble learning approach, Ibk Clasifier, smote techniques and selection features it is known that the approach produces effective intrusion detection compared to other approaches such as random forest, naive bayes, bayesian networks. Compared with some previous studies in the literature, the method proposed in this study results in better performance in terms of accuracy, detection rate, and false alarm rate, but the performance in terms of processing time has not shown satisfactory results.

REFERENCES

[1] H. Bauer, F. Brandl, C. Lock, and G. Reinhart, "Integration of Industrie 4.0 in Lean Manufacturing Learning Factories," *Procedia Manuf.*, vol. 23, no. 2017, pp. 147–152, 2018.

[2] Q. Wang, X. Zhu, Y. Ni, L. Gu, and H. Zhu, "Blockchain for the IoT and industrial IoT: A review Qin," *Blockchain IoT Ind. IoT A Rev. Qin*, no. xxxx, 2019.

[3] D. Mourtzis, E. Vlachou, and N. Milas, "Industrial Big Data as a Result of IoT Adoption in Manufacturing," *Procedia CIRP*, vol. 55, pp. 290–295, 2016.

[4] *C. Xia, X. Jin, L. Kong, C. Xu, and P. Zeng, "Lane scheduling around crossroads for edge computingbased autonomous driving," J. Syst. Archit., vol. 95, no. October 2018, pp. 1–8, 2019.*

[5] Z. Pan, S. Hariri, and J. Pacheco, "Context Aware Intrusion Detection of Building Automation Systems," *Comput. Secur.*, 2019.

[6] S. Hajiheidari, K. Wakil, M. Badri, and N. J. Navimipour, "Intrusion detection systems in the Internet of things: A comprehensive investigation," *Comput. Networks*, vol. 160, pp. 165–191, 2019.

[7] A. Badri, B. Boudreau-trudel, and A. S. Souissi, "Occupational health and safety in the industry 4.0 era_ A cause for major concern?," *Saf. Sci.*, vol. 109, no. May, pp. 403–411, 2018.

[8] S. Mohammadi, H. Mirvaziri, M. Ghazizadeh-ahsaee, and H. Karimipour, "Cyber intrusion detection by combined feature selection algorithm," *J. Inf. Secur. Appl.*, vol. 44, pp. 80–88, 2019.

[9] M. Aly, F. Khomh, M. Haoues, A. Quintero, and S. Yacout, "Enforcing Security in Internet of Things Frameworks: A Systematic Literature Review," *Internet of Things*, p. 100050, 2019.

[10] H. H. Pajouh, R. Javidan, R. Khaymi, A. Dehghantanha, and K. Raymond, "A Two-layer Dimension Reduction and Two-tier Classification Model for Anomaly-Based Intrusion Detection in IoT Backbone Networks," vol. 6750, no. c, pp. 1–11, 2016.

[11] W. Khreich, B. Khosravifar, A. Hamou-lhadj, and C. Talhi, "An anomaly detection system based on variable N-gram features and one-class SVM," *Inf. Softw. Technol.*, vol. 0, pp. 1–12, 2017.

[12] N. Mohamudally and M. Peermamode-mohaboob, "Building An Anomaly Detection Engine (ADE) For IoT Smart Applications," *Procedia Comput. Sci.*, vol. 134, pp. 10–17, 2018.

[13] D. Li, L. Deng, M. Lee, and H. Wang, "IoT data feature extraction and intrusion detection system for smart cities based on deep migration learning," *Int. J. Inf. Manage.*, no. October 2018, pp. 0–1, 2019.

[14] S. Thaseen and C. A. Kumar, "Intrusion Detection Model Using fusion of Chi-square feature selection and multi class SVM," *J. KING SAUD Univ. - Comput. Inf. Sci.*, no. 2016, 2015.

[15] I. S. Thaseen and C. A. Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 29, no. 4, pp. 462–472, 2017.

[16] J. Sun, J. Lang, H. Fujita, and H. Li, "Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates," vol. 425, pp. 76–91, 2018.

[17] S. Maldonado, J. López, and C. Vairetti, "An alternative SMOTE oversampling strategy for high-dimensional datasets," *Appl. Soft Comput. J.*, 2018.

[18] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," *Pattern Recognit.*, 2017.

[19] C. Zucco, "Multiple Learners Combination: Bagging," pp. 1–6, 2018.

[20] L. Li, H. Zhang, H. Peng, and Y. Yang, "Nearest neighbors based density peaks approach to intrusion detection," vol. 110, pp. 33–40, 2018.

Intrusion Detection System as Audit in IoT Infrastructure using Ensemble Learning and SMOTE Method

ORIGINALITY REPORT

4%	2%	4%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Swati Jadhav, Hongmei He, Karl Jenkins. "Information gain directed genetic algorithm wrapper feature selection for credit rating", Applied Soft Computing, 2018 Publication	1%
2	Siva S. Sivatha Sindhu, S. Geetha, A. Kannan. "Evolving optimised decision rules for intrusion detection using particle swarm paradigm", International Journal of Systems Science, 2012 Publication	1%
3	Muhammad Najih Muhammad Najih, De Rosal Ignatius Moses Setiadi, Eko Hari Rachmawanto, Christy Atika Sari, Setia Astuti. "An improved secure image hiding technique using PN-sequence based on DCT-OTP", 2017 1st International Conference on Informatics and Computational Sciences (ICICoS), 2017 Publication	1%
4	Kristiawan Nugroho, Edy Noersasongko, Purwanto, Muljono, Ahmad Zainul Fanani,	<1%

Affandy, Ruri Suko Basuki. "Improving Random Forest Method to Detect Hatespeech and Offensive Word", 2019 International Conference on Information and Communications Technology (ICOIACT), 2019

Publication

5 Eri Eli Lavindi, Edi Jaya Kusuma, Guruh Fajar Shidik, Ricardus Anggi Pramunendar, Ahmad Zainul Fanani, Pujiono. "Neural Network based on GLCM, and CIE L*a*b* Color Space to Classify Tomatoes Maturity", 2019 International Seminar on Application for Technology of Information and Communication (iSemantic), 2019 <1%

Publication

6 www.ncbi.nlm.nih.gov <1%

Internet Source

7 wrap.warwick.ac.uk <1%

Internet Source

8 www.scribd.com <1%

Internet Source

9 Anirut Suebsing, Nualsawat Hiransakolwong. "Feature Selection Using Euclidean Distance and Cosine Similarity for Intrusion Detection Model", 2009 First Asian Conference on Intelligent Information and Database Systems, 2009 <1%

Publication

10	Z. Muda. "Intrusion detection based on k-means clustering and OneR classification", 2011 7th International Conference on Information Assurance and Security (IAS), 12/2011 Publication	<1%
11	Farah Zakiyah Rahmanti, Novita Kurnia Ningrum, Prajanto Wahyu Adi, Mauridhi Hery Purnomo. "A comparison of plasmodium falciparum identification from digitalization microscopic thick blood film", 2016 1st International Conference on Biomedical Engineering (IBIOMED), 2016 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		