ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2022.0370

PERFORMANCE ANALYSIS OF AN EFFICIENT FRAMEWORK FOR INTRUSION DETECTION SYSTEM USING DATA MINING TECHNIQUES

C. Amali Pushpam and J. Gnana Jayanthi
Department of Computer Science, Bharathidasan University, India

Abstract

In the midst of the COVID-19 epidemic crisis, due to the tremendous development of mobile and internet technologies, the excessive growth in cyber-crime makes networksurity a major concern. As a result, individuals and companies are gradually moving towards the use of Intrusion Detection System (IDS), as it plays a persuasive role in monitoring and detecting the traffic of a network. However, high dimensional data affect the performance of IDS by reducing prediction accuracy, increasing false positive rate and classification time. Hence the focus of this research work is to develop a novel framework by integrating Auto – Bi Level (ABL) Classification with Double Filtering Fine Tuning – Ensemble Hybrid (DFFT-EH) feature selection. The experiments are conducted using NSL- KDD a benchmark intrusion detection dataset and it is proved that the proposed framework performs well with good accuracy, less false positive rate and less classification time when compared with voting ensemble classifier and other existing standard algorithms.

Keywords:

Auto – Bi Level (ABL) classification, Intrusion Detection System (IDS), Data Mining (DM), Feature Selection, Ensemble

1. INTRODUCTION

Fast-growing mobile and digital technologies generate a large amount of high-dimensional data that move around the network with anomalous data of intruders. The penetration of intruders through different attack modes has become more dangerous and generates new vulnerabilities to networksurity. Therefore, the need for IDS for networksurity on cyber space has become an obligatory. Intrusion detection system has the potential to identify the malicious events by examining and identifying them. Though it is astounding in identifying attacks from both internal and external sources, high speed network consigns some challenges ahead of it. High speed network transmits huge volume of data with hundreds and hundreds of features. Analyzing this high dimensional data is a very big challenge to IDS [1]-[3]. IDS is unique in its ability of incorporating the strength of other techniques. Henceforth, to handle high dimensional data, data mining with feature selection has been integrated in IDS.

The main task of IDS is to classify the events as normal and abnormal. In this work, data mining concept is integrated with an IDS to identify the relevant, hidden pattern effectively with less execution time [4]. A number of classification techniques with their merits and limitations classify network traffic based on class label with significant reduction in false positive. Nowadays these classification processes are more complex and require many specializations. Therefore, instead of relying on single classifier, ensemble of classifiers is more enviable in classification. Ensemble provides better prediction accuracy, but its prediction computation effort and time are high [5]-[7]. In proposed work, ABL classification has been introduced to produce good prediction accuracy with less computation effort and less testing time.

Feature selection is one of the techniques used to reduce the dimensionality of data by selecting optimal features. Therefore, it speeds up the analysis process, increases model performance, reduces model complexity and minimizes the training and testing time [8]-[12]. Selecting relevant features is not quite easy though a number of methods such as filter, wrapper, hybrid and embedded are available in feature selection. Filter methods are less accurate but fast; and cost and time effective. They are classifier independent. In contrast to filter, wrapper methods are more accurate but slow. Wrappers are time inefficient and classifier dependent. Embedded performs learning and feature selection simultaneously and not suitable for large data due to poor generalization. As a hybrid method combines the strength of both filter and wrapper, perform well than filter and faster than wrapper. Researchers, who focus on accuracy and speed, prefer hybrid method for feature selection. However, more computation effort and classifier dependency of wrapper are to be addressed in hybrid. To meet these challenges, Double Filtering Fine Tuning – Ensemble Hybrid feature selection has been introduced in proposed work.

The proposed work incorporates the ensemble classifier and hybrid feature selection method to use the strengths of these two and reduce their limitations.

In this paper it is aimed to (i) develop an efficient framework for IDS using Auto Bi-Level (ABL) classification with DFFT-EH feature selection to enhance the prediction accuracy of framework with less false positive rate and low testing time; (ii) compare the proposed algorithm with voting ensemble classifier built with RF and DT and other existing standard algorithms by conducting experiments on NSL KDD data set. The experimental results illustrate that the proposed work produces better prediction accuracy compared to existing algorithms. It has also been proven that the proposed work achieves good accuracy with low false positive rate and less testing time.

The organization of this paper is as follows: section 2 reviews some of the related work similar to the proposed work. Section 3 outlines the proposed method. Section 4 briefs the experimental study using NSL KDD data set and shows the performance analysis of the proposed work with voting ensemble classifier and other existing standard algorithms and finally section 5 ends with conclusion.

2. RELATED WORKS

Different approaches related to ensemble classifier with feature selection have been proposed by researchers in their works to improve intrusion detection system. Related works between 2005 and 2021 are summarized here. Celestine Iwendi et al. [13] have proposed ensemble classifier of J48, Random Forest and

Reptree with Correlation-based Feature Selection approach (CFS) for binary and multi class classification. Ngoc Tu Pham et al. [14] have applied Bagging and Boosting ensemble techniques with two different feature selection methods to improve performance of IDS. Fadi Salo et al. [6] proposed an ensemble classifier built with base classifiers such as Support Vector Machine (SVM), Multi-Layer Perceptron (MLP) and Instance-based learning algorithms (IBK) with hybrid method combining Information Gain (IG) and Principal Component Analysis (PCA) for dimensionality reduction. Longjie Li et al. [15] have introduced a two-step Intrusion Detection approach for multi-class classification. Binary classifiers and K-NN algorithm with feature selection done using Gain Ratio filter method have been applied. Zhou et al. [16] have proposed an ensemble classifier, built with base classifiers of C4.5, Random Forest (RF) and Forest by Penalizing Attributes (Forest PA). Based on the Average of Probabilities (AOP) rule, final prediction was taken. For feature selection, hybrid method has been applied. Lin et al. [17] proposed an Automatic Feature Selection and Ensemble Classifier for Intrusion Detection.

3. PROPOSED WORK

Existing works produce good prediction accuracy by combining data mining and feature selection but failed to produce good results in false positive rate and testing time. With an aim of attaining these three-performance metrics, the proposed work is designed to build an efficient framework for IDS using ABL Classification with DFFT-EH Feature Selection. For dimensionality reduction, Double Filtering Fine Tuning – Ensemble Hybrid (DFFT-EH) feature selection and for Classification of network connection records, Auto-Bi Level (ABL) Classification have been applied. The research work is depicted in Fig.1 and Fig.2.

3.1 DOUBLE FILTERING FINE TUNING-ENSEMBLE HYBRID FEATURE SELECTION (DFFT-EH)

High dimensional data which enlarge the search space of DM techniques is a major challenge to IDS. Feature selection is applied for dimensionality reduction. Different methods in feature selection such as filter, wrapper, hybrid and embedded are used to select highly relevant and non-redundant features which increase model's performance and reduce model complexity and timing. According to evaluation metrics, researchers apply different feature selection methods in their work. Each method produces different feature subset and different result. Hence instead of trusting in a particular method, choosing a hybrid is a better choice. In proposed work, hybrid method has been applied as it combines the benefits of filter and wrapper. Different possible combinations of filter and wrapper methods in hybrid are Single Filter Single Wrapper (SFSW), Single Filter Multiple Wrappers (SFMW), Multiple Filters Single Wrapper (MFSW) and Multiple Filters Multiple Wrappers (MFMW). Among these, as MFSW is simple and efficient, it has been applied in Double Filtering Fine Tuning-Ensemble Hybrid (DFFT-EH) feature selection. Two Hybrid methods of different combinations of filters and wrappers are carried out in parallel and produce two different feature subsets. Final optimal feature subset is obtained by applying inheritance property. It enhances classifier performance and gives good accuracy.

3.2 ABL CLASSIFICATION

In ensemble learning, voting ensemble classifier is common and widely used in classification. In voting ensemble, in training phase multiple classification models are prepared using training dataset. In testing phase, each classifier predicts individually through voting. As well as some additional work for taking final prediction is done. All these increase testing time; add workload and complicate the structure. As all the base classifiers are equally treated, their uniqueness is not identified and not effectively utilized. This also affects the performance of model. To meet these issues, some significant modifications have been done in proposed work i.e., Auto-Bi Level (ABL) classification method. From first phase of proposed work, RF and DT have been chosen to construct ABL classification. RF performs well than other classifiers. However, it is slow. Comparing to RF, DT is less accurate but fast. By considering their uniqueness, in ABL classification method, RF is considered as "Accuracy Classifier" (AC) and DT as "Fast Classifier" (FC). Instead of checking all network instances by both RF and DT, prediction work is distributed among RF and DT i.e first network instance is passed to FC. Based on soft voting, probability of each class of instance is estimated. Maximum of these probabilities is taken as confidence score. If confidence score is greater than threshold, that class is final prediction. Otherwise, if uncertain case arises, same instance is passed automatically without human intervention to AC. It predicts the class of that instance. This method speeds up the process, enhances performance of model by auto bi-level classification and reduces false positive rate (FPR).

4. EXPERIMENTAL ANALYSIS

Thisstion analyzes the results obtained when testing the performance of the proposed ABL Classification with DFFT-EH Feature selection.

4.1 EXPERIMENTAL SETUP

The proposed work is implemented using Pyhon 3.7.10 programming language with the environment of an improved version of KDD CUP'99 which struggle due to redundancy. After removing duplicate records, NSL KDD has sufficient and reasonable number of instances maintaining the same quality. Collection of files in different format such as .CSV and .ARFF are available for experimental study. K KDDTrain+ and KDDTest+ are full NSL-KDD train and test set containing 125973 training records and 22544 testing records.

KDDTrain+_20Percent, KDDTest-21 are subset of above files having 25192 training and 11850 testing records. Each connection record of NSL KDD having 43 features of different data types like nominal, binary and numeric, is categorized as either normal or any specified attack. NSL KDD dataset covers a wide variety of intrusions simulated in a network environment. Those intrusions are grouped into four categories namely DoS, Probe, R2L and U2R. Learning model should be trained with training data containing all possible malicious traffic records. Then only it can detect all types of attacks without any discrimination.

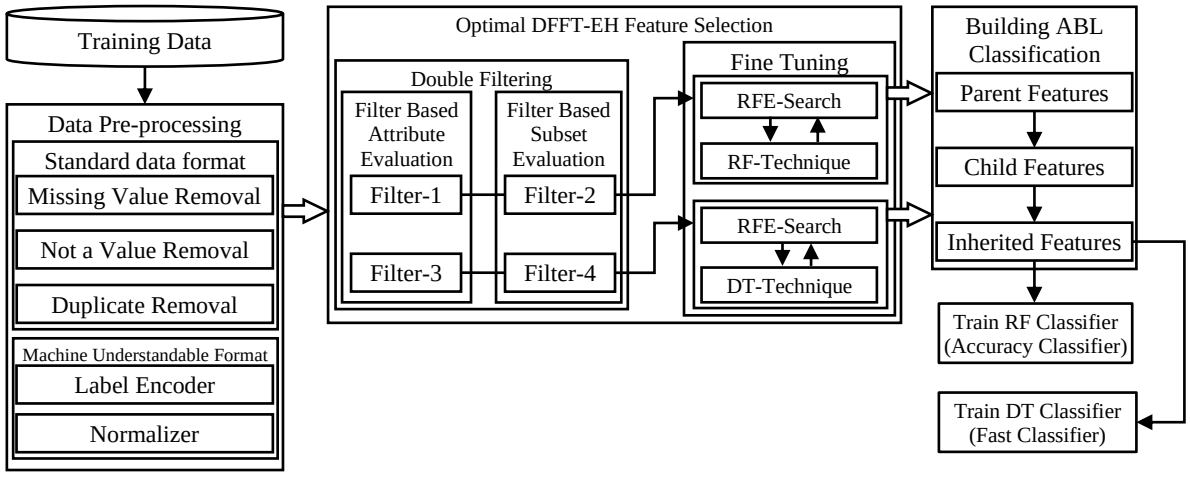

Fig.1. Framework of Proposed Work for Training

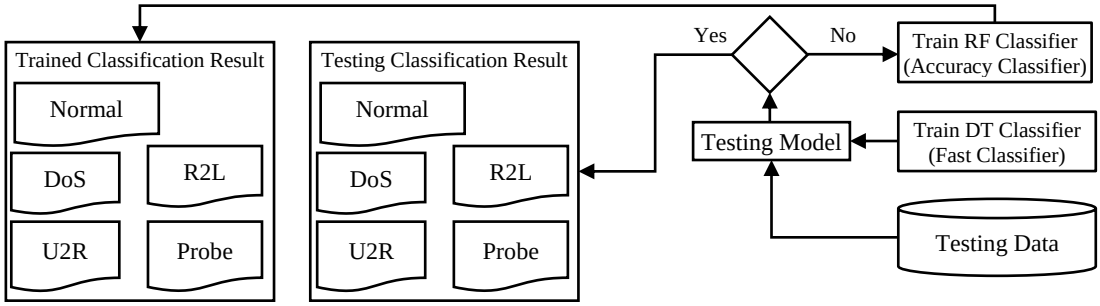

Fig.2. Framework of Proposed Work for Testing

In our work, model is trained with Network instances of KDDTrain+ _20percent.

4.2 EXPERIMENTAL STUDY

The experiments are performed in two phases; (i) DFFT-EH Feature selection (ii) ABL Classification

4.2.1 DFFT-EH Feature Selection:

DFFT-EH consists of two hybrid methods having a combination of double filters and single wrapper. In hybrid-I, Mutual Information (MI) and minimum redundancy and Maximum Relevance (mRMR) have been applied to select relevant and non-redundant features. This features subset is input to wrapper where, Recursive Feature Elimination (RFE) is used for best features subset selection.

Table.1. Feature Selected by Hybrid-1

Feature Selection Method	No. of Features Selected	Number of Selected Features
MI	18	3,4,5,6,12,23,25,26,29,30,32,33, 34,35,36,37,38,39
mRMR	16	3,4,5,12,23,25,26,29,30,32,34,35, 36,37,38,39
RFE+RF	10	3,4,5,23,29,30,34,35, 36, 39

Comparing to various algorithms in wrapper methods such as Genetic Algorithm (GA), Meta heuristic algorithms which are complex in nature, RFE is an efficient and simple. RFE eliminates feature(s) from entire set through iteration and generates different feature subsets. These subsets are evaluated by Random Forest classifier. Features selected by Hybrid-I are listed in Table.1.

In Table.2, Fisher Score and Fast Correlation Based Feature Selection (FCBF) have been applied to obtain relevant and non-redundant features subset. In wrapper, Recursive Feature Elimination (RFE) with Decision Tree (DT) classifier is used for best features subset selection. The Table.1 and Table.2 produce two features subset with a time of 68.47s and 55.93s respectively. As hybrid-II produces feature subset faster than hybrid-I, features subset selected by hybrid-II is acting as “Child”, and another one is “Parent”.

Table.2. shows Features selected by Hybrid -II Feature Selection

Feature Selection Method	No. of Features Selected	Number of selected Features
F-Score	29	1,2,3,4,6,8,9,10,12,14,22,23, 25,26,27,28,29,30,31,32,33, 34,35,36,37,38,39,40,41
FCBF	06	3,4,8,12,14,37
RFE +DT	05	3,4,8,12,37

To obtain final optimal features subset, instead of applying any combiner, parent features are inherited into child. Performance of these three features subset is evaluated by standard classifiers namely SVM, RF, DT, K-NN and NB and their results are presented in tables; Table.3-Table.5 and graphically represented in Fig.3-Fig.5.

Table.3. Child Features vs. Performance of Classifiers

Model	Accuracy	Precision	Recall	F1-Score
SVM	80.97%	71%	81%	75%
RF	93.83%	94%	94%	93%
DT	93.81%	94%	94%	93%
KNN	91.21%	91%	91%	91%
NB	74.36%	77%	74%	70%

Table.4. Parent Features vs. Performance of Classifiers

Model	Accuracy	Precision	Recall	F1-Score
SVM	51.34%	29%	51%	35%
RF	99.20%	99%	99%	99%
DT	98.91%	99%	99%	99%
KNN	97.89%	98%	98%	98%
NB	83.98%	82%	84%	82%

Table.5. Inherited Features vs. Performance of Classifiers

Model	Accuracy	Precision	Recall	F1-Score
SVM	51.34%	29%	51%	35%
RF	99.33%	99%	99%	99%
DT	98.95%	99%	99%	99%
KNN	97.93%	98%	98%	98%
NB	83.72%	81%	84%	82%

Performance analysis on DFFT-EH Feature Selection method is done between classifiers and feature subsets based on accuracy:

Among five classifiers (SVM, RF, DT, KNN and NB), RF, DT and KNN produce better accuracy 99.22%, 98.90% and 97.89% respectively with inherited feature subset. SVM performs better with child feature subset and produces an accuracy of 80.73% and doesn't show any changes in accuracy (51.37%) while working with parent and inherited feature subset. NB performs well with parent feature subset and produces an accuracy of 83.81%. Overall, the performance of RF, compared to other classifiers, is better with three feature subsets and produces better accuracy with inherited feature subset. It is given in Table.6.

Table.6. Classifiers vs. Accuracy on different features subsets of DFFT-EH

Features Set	SVM	RF	DT	KNN	NB
Parent	51.34	99.20	98.91	97.89	83.98
Child	80.97	93.83	93.81	91.21	74.36
Inherited	51.34	99.33	98.95	97.93	83.72

Inherited features subset is optimal features subset as it produces good accuracy than parent and child features subset with three classifiers such as RF, DT and KNN out of five. Henceforth, proposed feature selection method is not bias to particular classifier. Therefore, inherited features subset is the final optimal features subset of proposed work. Among five classifiers, RF and DT perform well with good accuracy. Hence these two classifiers have been selected to construct ABL classification in next phase.

4.2.2 Auto Bi-Level (ABL) – Classification – Phase-II

RF and DT selected from Phase-I have been applied in ABL - classification and is trained with 13 features. The Proposed work produces good accuracy of 99.20 % and FPR of 0.06% with less classification time of 1.97s. Also, higher TPR (0.998) and lower FPR (0.0006) prove that the performance of model is noticeable. In multi classification, it not only classifies attacks, but gives details of attacks. Most probably all attacks fall into four categories namely DoS, Probe, R2L and U2R. Proposed work, through multi classification, identifies attack-class also.

The performance of proposed research work is evaluated in terms of precision, recall and F1-Score for each class. ABL Classification report for five classes is described in Table.7. It explains the efficiency of proposed work in detecting individual classes namely Normal, DoS, Probe, R2L and U2R. From the result, the proposed work is highly competent in detecting DoS, normal and Probe and also its detection performance is good in rare attack R2L. In case of another low frequency attack, U2R, it achieves good score in recall.

Table.7. ABL-classification report for five classes

Class	Precision	Recall	F1-Score
DoS	1.00	1.00	1.00
Normal	0.99	0.99	0.99
Probe	0.99	0.99	0.99
R2L	0.96	0.93	0.94
U2R	0.62	1.00	0.77
Macro-Average vs. Weighted-Average			
Macro-Average	0.91	0.98	0.94
Weighted-Average	0.99	0.99	0.99

Average scores of precision, recall and F1-score of all classes are calculated in two ways such as Macro_Avg and weighted_Avg. To handle data imbalance (no equal distribution of all five classes in data) and giving importance to some prediction more (based on their proportion), Weighted_Avg has been chosen to calculate average of all classes.

For performance analysis, the proposed work is compared with Voting classifier which is constructed with same base classifiers namely RF and DT of proposed work. Ensemble concept is applied in both voting classifier and ABL classifier. But ABL classifier applies this with significant modification in testing phase. Performance evaluation is done on three important metrics such as prediction accuracy, classification time and FPR. Experimental results prove that the proposed work produces better accuracy (99.20%) than the voting classifier (98.95), and the classification time is also laudably reduced by 5.42 times compared to the voting classifier. Also, FPR of proposed work is 0.0006. Compared to other standard individual classifiers such as RF and DT, Random Forest classifier which is an ensemble of trees gives good accuracy of 99.33%. However, compared to testing time, testing time is high. But the proposed work has achieved similar accuracy in less testing time. In case of Decision Tree, as DT is fast, it produces 98.95% accuracy with less testing time. Hence it is clear that the performance of these classifiers (RF, DT and Voting Classifier) is not standard in all metrics. But the proposed work gives acceptable and commendable results in

all the three metrics such as accuracy, False Positive Rate (FPR) and classification time. The proposed work gives a better detection rate for low frequency attacks R2L and U2R respectively.

Table.8. Comparison with other standard algorithms

Classifier	Accuracy	FPR	TPR	Testing Time (s)
RF	99.33	0.0006	0.9986	13.2952
DT	98.95	0.0012	0.9989	0.0956
Voting Classifier	98.95	0.0012	0.9989	13.3084
ABL Classifier	99.20	0.0006	0.9989	1.9701

Inherited features produced by DFFT-EH Feature selection method gives good accuracy. Out of five classifiers, it performs well with three classifiers namely RF, DT and K-NN. i) Hence this optimal feature subset is not classifier dependent. ii) As search space for wrapper is reduced by double filtering, time taken by individual wrapper method is reduced iii) ABL classification method is also tested with inherited features subset and gives good accuracy. Hence ABL classification method produces good accuracy with optimal feature subset.

5. CONCLUSION

In this paper, an efficient framework using ABL Classification with DFFT-EH Feature selection for network intrusion detection system is proposed. The proposed work is evaluated using NSL – KDD dataset. In ABL classification, instead of treating base classifiers equally, their uniqueness is identified and prediction work is distributed among them according to their strength. This classification process is supported by relevant and non-redundant features selected by DFFT-EH Feature selection method. The experimental results prove that proposed work has achieved good accuracy and appreciable reduction in FPR and classification time. Compared with voting classifier and standard algorithms, it outperforms.

REFERENCES

[1] Saurabh Mukherjee and Neelam Sharma, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction", *Procedia Technology*, Vol. 4, pp. 119-128, 2012.

[2] Shina Sheen and R Rajesh, "Network Intrusion Detection using Feature Selection and Decision Tree Classifier", *Proceedings of IEEE Region Conference on TENCON*, pp. 1-4, 2008.

[3] Y. Saeys, T. Abeel and Y. Van De Peer, "Robust Feature Selection using Ensemble Feature Selection Techniques", *Proceedings of IEEE Region Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 1-13, 2008.

[4] T. Hamed, R. Dara and S.C. Kremer, "An Accurate, Fast Embedded Feature Selection for SVMs", *Proceedings of IEEE International Conference on Machine Learning and Applications*, pp. 135-140, 2014.

[5] Nazrul Hoque, Mihir Singh and Dhruba K. Bhattacharyya, "EFS-MI: An Ensemble Feature Selection Method for Classification", *Complex and Intelligent Systems*, Vol. 4, pp. 105-118, 2017.

[6] F. Salo, A. Nassif and A. Essex, "Dimensionality Reduction with IG-PCA and Ensemble Classifier for Network Intrusion Detection", *Computer Networks*, Vol. 148, pp. 164-175, 2019.

[7] S.S. Ahmadi, S. Rashad and H. Elgazzar, "Efficient Feature Selection for Intrusion Detection Systems", *Proceedings of IEEE International Conference on Ubiquitous Computing, Electronics and Mobile Communication*, pp. 1029-1034, 2019.

[8] E. Karabulut, S. Ozel and T. İbrikci, "A Comparative Study on the Effect of Feature Selection on Classification Accuracy", *Procedia Technology*, Vol. 1, pp. 323-327, 2012.

[9] M.C. Rekha Preethi and R. Chetan, "Least Square Support Vector Machine based IDS, using Feature Selection Algorithm", *International Journal of Emerging Trends and Technology in Computer Science*, Vol. 6, No. 3, pp. 64-68, 2017.

[10] S.L. Shiva Darshan and C.D. Jaidhar, "Performance Evaluation of Filter-based Feature Selection Techniques in Classifying Portable Executable Files", *Procedia Computer Science*, Vol. 125, pp. 346-356, 2018.

[11] M. Cherrington, F. Thabtah, J. Lu and Q. Xu, "Feature Selection: Filter Methods Performance Challenges", *Proceedings of International Conference on Computer and Information Sciences*, pp. 1-4, 2019.

[12] S.Vanaja and K. Ramesh kumar, "Analysis of Feature Select ion Algorithms on Classification: A Survey", *International Journal of Computer Applications*, Vol. 96, No. 17, pp. 28-35, 2014.

[13] C. Iwendi, S. Khan, J. Anajemba, M. Mittal, M. Alenezi and M. Alazab, "The Use of Ensemble Models for Multiple Class and Binary Class Classification for Improving Intrusion Detection Systems", *Sensors*, Vol. 20, No. 9, pp. 2559-2567, 2020.

[14] N. Pham, E. Foo, S. Suriadi, H. Jeffrey and H. Lahza, "Improving Performance of Intrusion Detection System using Ensemble Methods and Feature Selection", *Proceedings of the Australasian Week Multi-Conference on Computer Science*, pp. 1-6, 2018.

[15] L. Li, Y. Yu, S. Bai, Y. Hou and X. Chen, "An Effective Two-Step Intrusion Detection Approach Based on Binary Classification and k -NN", *IEEE Access*, Vol. 6, pp. 12060-12073, 2018.

[16] Y. Zhou, G. Cheng, S. Jiang and M. Dai, "Building an Efficient Intrusion Detection System based on Feature Selection and Ensemble Classifier", *Computer Networks*, Vol. 174, pp. 1-24, 2020.

[17] C. Lin, A. Li and R. Jiang, "Automatic Feature Selection and Ensemble Classifier for Intrusion Detection", *Journal of Physics: Conference Series*, Vol. 1856, No. 1, pp. 12067-12078, 2021.