ISSN: 2229-6948(ONLINE)
DOI: 10.21917/ijct.2021.0380


# JAA IDS - FRAMEWORK DESIGN FOR AN EFFICIENT INTRUSION DETECTION SYSTEM

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

***Abstract***

***Network security has become a very hot research area as its importance is heavily realized in various fields. Various mechanisms and tools are available to support this. But they do not meet the challenges imposed by fast growing technologies. Massive amounts of high dimensional data are one of the challenges. Data with a large number of features is entering and moving around the network. The Intrusion Detection System is a new mechanism that faces this challenge with the support of data mining and feature selection. In data mining, the ensemble is more preferable than a single method. In an ensemble, during the testing phase, all base classifiers are treated equally and individually participate and vote. To take a final decision, some extra effort has to be made. All these increase computation effort and time. To overcome these, this research paper proposes a new framework for intrusion detection systems using the Auto Bi-Level (ABL) classification technique with Double Filtering Fine Tuning–Ensemble Hybrid method.***

***Keywords:***

***Network Security, Features selection, Data Mining, Intrusion Detection System, Ensemble***

## 1. INTRODUCTION

Recent developments in current technologies such as IoT devices and mobile technologies produce an enormous amount of data. Valuable information from this data is broadly used in various applications. Extracting valuable information from data is a tedious challenge in the hi-tech world. Data Mining gives a solution to this problem by extracting valuable and hidden patterns or correlations from data.

Data mining is a data analytic tool. It selects and integrates data from different sources and analyzes them in order to find patterns. Data mining is used in various applications in banking, business, science, education, medicine, intrusion detection, agriculture, government, etc. By detecting intrusions, the Intrusion Detection System (IDS) strengthens network security. By incorporating data mining with IDS, the performance of IDS is enhanced [1].

A number of data mining algorithms are available. They are classified into four categories, namely, classification, clustering, regression, and association rules. Generally, classification algorithms, which are supervised machine learning, are broadly used in intrusion detection systems. Features which are associated with data play an imperative role in the data analysis process. Selecting optimal features reduces the load of data mining models. The optimum feature subset reduces the complexity of the model and enhances its performance [2], [3].

Hence, feature selection is also added with data mining in IDS. Researchers set their focus on the ensemble, as it produces better results than single methods [4], [5]. A data mining algorithm which produces good results in one application may not be suitable for another one. Also, nowadays, problems are very complex in nature.

To solve them, multiple skills or experts or algorithms are required. Ensemble solves this by combining multiple base classifiers and forming a strong classifier. When an IDS identifies disruptive behaviour on the network, it must sound an alarm. Different types of network activity generate alarms from different IDSs. In the majority of cases, the following are the most common trigger mechanisms: Detection of anomalies and Detection of misuse

## 2. RELATED WORKS

Our research is focused on enhancing the efficiency of IDS. Related to this, research work has been done based on the following research focus:

### 2.1 RF1 - CLASSIFICATION ALGORITHMS USED IN IDS

Network packets move around the network. These packets may be normal or malicious. By analyzing these packets, intrusions are identified. The Intrusion Detection System (IDS) does this well by incorporating DM. Data mining algorithms are classified into four categories, such as classification, clustering, regression, and association rules. Classification algorithms are supervised, machine learning algorithms.

From our previous work, Overview of Data Mining in Intrusion Detection Systems, it is identified that, the following classification algorithms are widely used in IDS. The Random Forest (RF) algorithm is highly accurate but takes more computation time. The Decision Tree (DT) performs well with good accuracy, but suffers due to overfitting and needs pruning. The K-Nearest Neighbor (KNN) algorithm is simple to understand and implement, but requires more storage and the selection of k-value is a difficult one.

The Artificial Neural Network (ANN) is highly reliable and stable, but more complex in structure; takes a long time to train, and its false positive rate is high. Genetic Algorithm (GA) efficiency is better, but it is more complex. With small data sets, the Support Vector Machine (SVM) performs well. It produces a good detection rate and a lower false positive rate. But it requires more memory and the selection of the kernel is a tedious one. It's training and testing speed is slow. It is easy to construct and produces good accuracy, but time complexity is greater.

### 2.2 RF2 - DIFFERENT FEATURE SELECTION METHODS USED IN CLASSIFICATION

Data and features are inseparable. The number of features also increases the dimensionality of the data. To handle high dimensional data, relevant and non-redundant features are

required. Feature selection selects optimal features which are more informative and useful. These features reduce the complexity of the model and enhance the performance of the model [6]-[8]. Different feature selection methods are available, namely filter, wrapper, embedded, and hybrid. All these methods have their own merits and limitations. Researchers apply these methods according to their requirements.

From our previous work, Methodical Survey on IDS with Feature Selection, it is observed that the filter method is widely used, because of its characteristics. Filter methods are classifier independent and fast. They are both cost-effective and time-efficient. According to performance, it is less. Wrapper methods are complex and classifier dependent. They are slow. But they perform well and produce a good feature subset. A hybrid method has been introduced by combining the strength of the filter and the wrapper to overcome the drawbacks of these two. Its performance is better than a filter and faster than a wrapper. Researchers who try to attain both accuracy and speed in their work, select a hybrid method. The Embedded method does both learning and feature selection simultaneously. It is not suitable for large data sets.

### 2.3 RF3 - DIFFERENT ENSEMBLE CLASSIFIERS WITH FEATURE SELECTION USED IN IDS AND THEIR CHALLENGES

Nowadays, any problem in any subject requires the collaboration of several experts or abilities to resolve. Problems are very complex in nature. Hence, in the research field, the ensemble method is highly preferable to a single method. In an ensemble, instead of relying on a single method, different multiple methods are combined and a strong one is formed. In classification as well, ensemble classifier produces good results. Hence, more researchers set their focus on ensemble with feature selection. In ensemble, researchers have tried different combinations of classifiers and proved their results [9]-[13]. While producing good results, it also faces certain challenges. In an ensemble, different base classifiers are combined and a strong classifier is formed. During training and testing, all base classifiers are trained and tested. During the testing phase, all base classifiers are treated equally and individually participate in voting. It leads to more computation effort and time.

### 2.4 RF4 - PERFORMANCE METRICS USED IN IDS MODELS

There are different performance metrics that are available to evaluate the performance of an IDS model [14], [15]. Accuracy, Detection rate, False Positive Rate, True Positive Rate, Testing time, ROC Curve, Precision, Recall, F1-Score, Stability, Cost, Root Mean Squared Error are some of them. Different combinations of them have been applied by researchers to prove their work. Generally, performance metrics such as prediction accuracy, false positive rate, and testing time are the common metrics used by many researchers. They did not succeed in achieving good results in all three of these metrics.

## 3. RESEARCH MOTIVATIONS

There is an unbelievable growth of technologies. It produces a huge volume of high dimensional data. The dimensionality of data is related to the number of features. Managing these large numbers of features is very challenging for an Intrusion Detection System. While analyzing data, feature selection is a wonderful task. It speeds up the process. Filter, Wrapper, Hybrid and Embedded are different methods of feature selection. These methods, with their merits and demerits, select optimal features from the entire set of data.

The hybrid method attains better accuracy than the filter and faster than the wrapper by combining the strengths of the filter and the wrapper. The complexity of the hybrid method requires further research and needs a new combination of filters and wrappers. Complex problems require an ensemble method rather than a single one, as it performs well. But ensemble methods also face certain challenges. All these motivate us to design and develop an efficient framework for an IDS with enhanced accuracy, a low false positive rate, and less testing time.

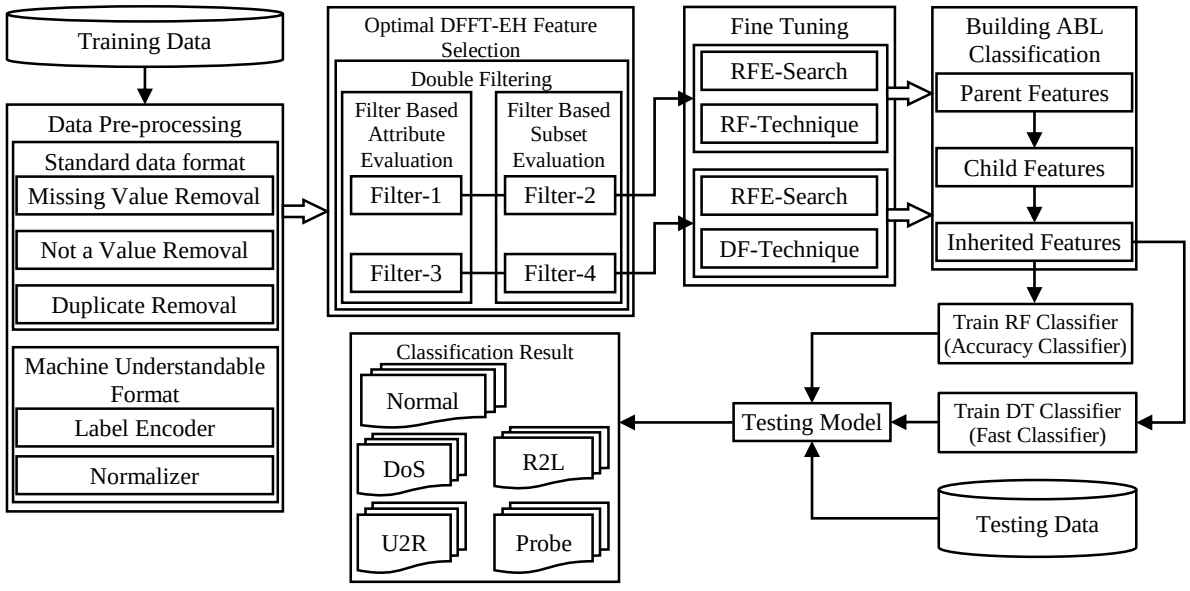


Fig.1. JAA-IDS for Intrusion Detection System

## 4. JAA-IDS - PROPOSED WORK

From this research focus, it is observed that there is a need to design an efficient framework for intrusion detection systems using ensemble classifiers with feature selection. It led to the design of JAA-IDS. In this proposed work, a new combination of filters and wrappers, Double Filtered Fine Tuning Ensemble Hybrid (DFFT-EH), has been introduced. Using double filters, relevant and non-redundant features are selected.

From double filtered features, the best subset is obtained through the wrapper method. Instead of relying on a particular method, the ensemble hybrid provides a chance to use different filter and wrapper methods. DFFT-EH selects optimal features which enhance the performance of the classifier. In ensemble, all base classifiers are treated equally and their uniqueness is not identified and applied. In ensemble classifiers, computation effort and computation time are greater.

To overcome these drawbacks, the ABL Classifier has been introduced in the proposed work. This ABL Classifier has been constructed with a Random Forest (RF) and a Decision Tree (DT). Among these two classifiers, Random Forest performs well and produces better accuracy than Decision Tree. Decision Tree is faster but less accurate than Random Forest.

By identifying their uniqueness, DT is considered a Fast Classifier (FC) and RF is considered an Accuracy Classifier (AC). Therefore, packets are first passed to the fast classifier. It first classifies them. If any uncertain case arises, the same packet is passed without any human intervention to the accuracy classifier and it classifies them. This ABL Classifier produces good accuracy and takes less computation time. Its false positive rate is also lower. Hence, the proposed work succeeds in achieving good results in three performance metrics such as accuracy, False Positive Rate, and Testing Time. The proposed work is depicted in Fig.1.

**Algorithm 1: Auto Bi-Level Classification**

**Input**: *X_test*, *Model*1, *Model*2, *N*

// *X_test* is test data

// *Model*1 is Decision Tree classifier

// *Model*2 is Random Forest classifier

// *N* = len(*X_test*)

// $\partial$ - Threshold

// *CS* - Confidence Score

**Output**: *ans*1

// *ans*1 is an array having classification result

```
Begin
  for i in range(0,N)do
  {
    X = Get network connection record (i) from X_test
    Call procedure Bilevel_prediction(X)
  }
End
Procedure Bilevel_prediction (X) do
Begin
  result1= Model1.predict_proba(X)
  CS = Max(result1)
  predictedClass = CS
  if(CS>∂)
    insert predictedClass in ans1
  else
    result2 = Model2.predict_proba(X)
    CS = Max(result2)
    predictedClass = CS
    insert predictedClass in ans1
    print(ans1)
  End
End Procedure
```

In Auto Bi-Level classification algorithm, each network connection record will fall into five classes, such as Normal, DoS, Probe, U2R, and R2L. A probability score of these classes for each record is obtained using model1. Maximum of these score is considered as Confidence Score (CS). If CS is greater than threshold value, then that class is the predicted class of that particular record. Otherwise, for strong confirmation, same record is passed to model2. Model2 predict the class of that record.

## 5. IMPLEMENTATION AND RESULTS

In phase-I, DFFT-EH produces optimal features subset. In phase-II, ABL classifier is constructed using optimal features subset and RF and DT classifiers. The Proposed work has been implemented in Python 3.7.10 programming language using NSL KDD dataset. The proposed work performs well with good accuracy of 99.20% and False Positive Rate of 0.06%. It achieves less classification time of 1.97s. The Table.1 represents the performance of proposed work.

Table.1. Performance of ABL-Classifier with DFFT-EH Feature Selection

| Classes | Values |
|---|---|
| Accuracy | 99.20 % |
| False Positive Ratio (FPR) | 0.0006% |
| True Positive Ratio (TPR) | 0.9989% |
| Training Time | 14.6971 secs |
| Testing Time | 10.701 secs |
| Precision | 0.99% |
| Recall | 0.99% |
| F1-Score | 0.99% |

## 6. PERFORMANCE ANALYSIS

The proposed work does both binary classification and multiclass classification. In binary classification, it identifies whether a network connection record is normal or an attack. In multiclass classification, it not only identifies the attack, but also identifies the class of attack. ABL-Classification Report for Five

Classes in terms of precision, recall, and F1-Score is given in Fig.2.

A number of performance metrics are available to evaluate the performance of a model. Generally, accuracy, FPR, and testing time are used. The ABL classifier produces good results in all three metrics. The outcome of the ABL classifier is compared with standard algorithms and is given in Table.2.

The results prove that the proposed work produces good results in terms of accuracy, FPR and testing time while other standard algorithms are not successful in producing good results in all these three metrics.

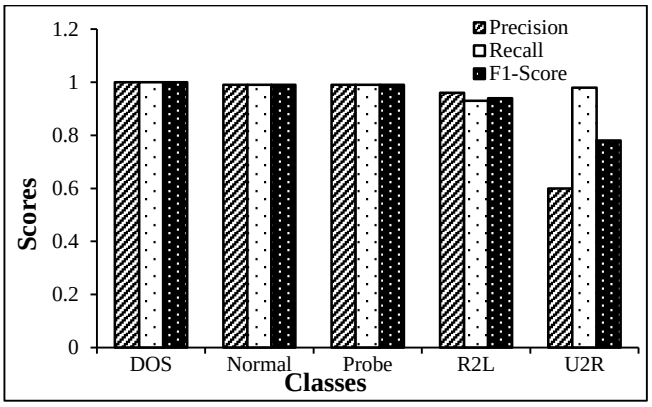


Fig.2. ABL-Classification Report for five classes

Table.2. Comparison with other Standard Algorithms in terms of Accuracy, FPR and testing time

| Classifier | Accuracy | FPR | TPR | AUC | Testing Time (s) |
|---|---|---|---|---|---|
| **RF** | 99.33 | 0.0006 | 0.9986 | 0.9989 | 13.2952 |
| **DT** | 98.95 | 0.0012 | 0.9989 | 0.9988 | 0.0956 |
| **Voting Classifier** | 98.95 | 0.0012 | 0.9989 | 0.9989 | 13.3084 |
| **ABL Classifier (Proposed)** | 99.20 | 0.0006 | 0.9989 | 0.9991 | 1.9701 |

## 7. CONCLUSION

This article proposes a novel framework called JAA-IDS for intrusion detection system to provide better performance in terms of accuracy, false positive rate and testing time. This proposed work has two main phases. In the first phase, optimal features are selected by the DFFT-EH feature selection method. The ABL Classifier is built in the second phase using Random Forest and Decision Tree, as well as an optimal feature subset, to provide better results.

## REFERENCES

[1] Tarfa Hamed, Rozita Dara and Stefan C. Kremer, "An Accurate, Fast Embedded Feature Selection for SVMs", *Proceedings of 13th International Conference on Machine Learning and Applications*, pp. 135-140, 2014.

[2] Z. Xue-qin, G. Chun-Hua and L. Jia-jun, "Intrusion Detection System Based on Feature Selection and Support Vector Machine", *Proceedings of International Conference on Communications and Networking*, pp. 1-5, 2006.

[3] A. Shadi, M. Ljawarneh and Muneer Bani Yasin, "Anomaly-Based Intrusion Detection System through Feature Selection Analysis and Building Hybrid Efficient Model", *Journal of Computational Science*, Vol. 23, No. 1, pp. 1-10, 2017.

[4] T. Karthikeyan and K. Praghash, "Improved Authentication in Secured Multicast Wireless Sensor Network (MWSN) using Opposition Frog Leaping Algorithm to Resist Man-in-Middle Attack", *Wireless Personal Communications*, Vol. 113, pp. 1-17, 2021.

[5] T. Karthikeyan and K. Praghash, "Data Privacy Preservation and Trade-off Balance Between Privacy and Utility Using Deep Adaptive Clustering and Elliptic Curve Digital Signature Algorithm", *Wireless Personal Communications*, Vol. 116, pp. 1-16, 2021.

[6] R. Manikandan and M. Ramkumar, "Design of Autonomous Production using Deep Neural Network for Complex Job", *Materials Today: Proceedings*, Vol. 4, pp. 1-12, 2021.

[7] Z. Xue-Qin, G. Chun-Hua and L. Jia-Jun, "Intrusion Detection System Based on Feature Selection and Support Vector Machine", *Proceedings of 1st International Conference on Communications and Networking*, pp. 1-5, 2006.

[8] 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.

[9] Suleman Khan, Joseph H. Anajemba, Mohit Mittal, Mamdouh Alenezi and Mamoun Alazab, "The Use of Ensemble Models for Multiple Class and Binary Class Classification for Improving Intrusion Detection Systems", *Sensors*, Vol. 20, No. 9, pp. 255-265, 2020.

[10] 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 Australasian Multiconference on Computer Science Week*, pp. 1-6, 2018.

[11] Fadi Salo, Ali Bou Nassif and Aleksander Essex, "Dimensionality Reduction with IG-PCA and Ensemble Classifier for Network Intrusion Detection", *Computer Networks*, Vol. 148, pp. 164-175, 2019.

[12] 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.

[13] Yuyang Zhou, Guang Cheng, Shanqing Jiang and Mian Dai, "An Efficient Intrusion Detection System Based on Feature Selection and Ensemble Classifier", *Journal of Lates Files*, Vol. 14, No. 8, pp. 1-13, 2019.

[14] Wei-Chao Lin, Shih-Wen Ke and Chih-Fong Tsai, "CANN: An Intrusion Detection System based on Combining Cluster Centers and Nearest Neighbors", *Knowledge Based Systems*, Vol. 78, pp. 13-21. 2015.

[15] W. Du, Z. Cao, T. Song, Y. Li and Y. Liang, "A Feature Selection Method based on Multiple Kernel Learning with Expression Profiles of Different Types", *BioData Mining*, Vol. 10, No. 1, pp. 1-16, 2017.