

Ensemble learning classifiers hybrid feature selection for enhancing performance of intrusion detection system

Hasanain Ali Al Essa, Wesam S. Bhaya

College of Information Technology, University of Babylon, Hillah, Iraq

Article Info

Article history:

Received Jan 23, 2023

Revised May 4, 2023

Accepted May 30, 2023

Keywords:

Ensemble learning classifier

Feature selection

Intrusion detection system

Multilayer perceptron

XGBoost algorithm

ABSTRACT

Feature selection (FS) plays an important role in the construction of efficient ensemble classifiers; particularly for intrusion detection system (IDS). An IDS is utilized in a network architecture to protect the availability of sensitive information. However, existing IDSs suffer from redundancy, high dimensionality, and high false alarm rate (FAR). Also, lots of models are constructed for outdated datasets, which makes them less flexible to deal with new assaults. Therefore, this paper proposes a new IDS relies on hybrid FS and ensemble classifiers. A hybrid FS approach consists of two techniques, hard-voting and mean. In contrast to recent papers, we use three different FS approaches: extra tree classifier importance as an embedded FS, recursive feature elimination (RFE) as a wrapper FS, and mutual information (MI) as a filter FS. Then, a hard-voting technique has been used to fuse output of these approaches and obtain a reduced subset of features. Since each feature has three weights, a mean technique has been utilized to assign one weight to each feature and obtain an optimal subset of features. The experimental outcomes, utilizing the modern InSDN dataset, confirm that the proposed hybrid FS with ensemble soft voting classifier achieves better results than other ensemble and individual classifiers due to several measures.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Hasanain Ali Al Essa

College of Information Technology, University of Babylon

Hillah, Babil, Iraq

Email: Hasanain@uobabylon.edu.iq

1. INTRODUCTION

The currently, the implementation of electronic technologies in every field and industry operation implies an ongoing positive direction in the growth of global connectivity in terms of the number of connected machines and communications. In this situation, communications networks and systems are always being targeted by intruders. Therefore, inspecting and discovering network attacks play a key role in sustaining critical security functions like availability, confidentiality, privacy, and integrity. For example, it is interesting to note two recent assaults, VPNFilter [1] and distributed denial of service (DDoS) [2], [3], in which thousands of computer devices were hacked, causing huge financial effects in addition to human costs.

In this regard, an intrusion detection system (IDS) is the most extensively utilized defensive line in communication and information technology for inspecting and discovering network attacks, acting as a strong instrument to combat versus various forms of network attacks [4]. An IDS can be classified into network-IDS and host-IDS. A host-IDS uses access Syslog files from end systems. On the other hand, a network-IDS investigates network traffic that passes via the network by using packet filtering [5]. Based on the intrusion detection mechanism, IDS may also be typically categorized as signature-based, anomaly-based, and hybrid [6]. Signature-based compares various kinds of intrusions with a pre-defined of signatures (patterns of

intrusion). One of its drawbacks is that it unable efficiently identify anonymous intrusions because of outdated databases and zero-day assaults. Anomaly-based reveal intrusions through learning anomalous and normal network traffic behaviors and have good detection abilities for anonymous intrusions. However, according to the issues of class imbalance and redundant attributes in IDS datasets, this approach may guide to false positive rate. Machine learning (ML) (i.e., ensemble learning) and statistical approaches are commonly applied for anomaly-based approaches. Hybrid-based integrates anomaly-based and signature-based [7].

Therefore, ensemble learning has been broadly utilized in IDSs due to their ability to learn and detect patterns of intrusion from network traffic via statistical techniques and algorithms [8]. Ensemble learning is a technique that integrates the outcomes of two or more ML classifiers trained separately to produce better performance than individual classifiers. There are various types of integrating, including voting ensemble (i.e., hard voting and soft voting) and stacked model. In this work, we propose ensemble learning classifiers combined by all these types.

Network data traffic contains a lot of irrelevant and redundant attributes or features. Thus, to inspect all attributes, it takes more time (processing cost) and lead to a performance reduction in the classification task. Therefore, it is not suitable to utilize all attributes via the IDS. Consequently, utilizing feature selection (FS) approaches in the preprocessing part have immense potential to enhance the performance of ensemble learning operations when blending with IDS. The benefits of FS involve data reduction, data understanding, reducing processing cost, and determining the amount of storage space required. FS approaches can be classified into filter based, wrapper based, and embedded methods. Many research papers have been implemented based on diverse FS approaches to aid IDS to enhance performance and decrease the rates of false alarms [9]–[11]. In this research, we suggest a new IDS depends on a hybrid FS approach which fuses three different FS approaches that can diminish downsides and inherent biases when employed individually and ensemble learning classifiers. The key contributions of this work are as:

- a. We suggest a new methodology that integrates the advantages of ensemble learning classifiers and FS approaches to obtain an accurate and effective IDS.
- b. In the feature selection stage, we propose a hybrid FS approach contains two techniques, namely, hard-voting and mean. In hard-voting technique three different FS are fused in order to reduce the number of features, then, each feature has three diverse weights due to these methods, mean technique has been applied to assign one weight for each feature and to obtain an optimal subset of features with just 10 features.
- c. Through data preprocessing, to avoid degrading the model's performance and solve the typical unbalanced dataset problem. Therefore, we performed both undersampling on majority classes (normal, DDoS, probe, and DoS) and oversampling (i.e., synthetic minority oversampling technique (SMOTE)) on minority classes (web attack, botnet, and exploitation).
- d. In the classification phase, we propose ensemble learning classifiers based on random forest (RF), extreme gradient boosting (XGBoost), and multilayer perceptron (MLP) combined by hard voting, soft voting, and stacked model in order to enhance outcomes of utilizing only one classifier.
- e. Experimental outcomes, attained depend on the modern InSDN dataset indicate that the proposed hybrid FS with an ensemble soft voting classifier can decrease the number of features from 77 to 10. In addition, it achieves better results compared to other ensemble and individual classifier algorithms due to accuracy, precision, F1 score, and recall with reduced training and testing times, and false alarm rate (FAR) remains at a reasonable level.

2. RELATED WORK

Since it is regarded as one of the most difficult risks in network security, intrusion detection, as a classification issue, has become an extremely prominent research area. However, several solutions have been presented to enhance IDS performance. In this part, we consider some works that full within ML-based IDS, use:

- a. On feature reduction (or selection) approaches

In order to reduce computation time, the approach of feature reduction, which can be utilized as a preprocessing stage in ML techniques, aims to improve the performance of IDSs as well as exclude useless features [12]. In order to acquire an effective and more reliable classifier. Hota and Shrivastava [13] proposed a model that utilized diverse FS approaches to exclude unessential features. The outcomes demonstrate that C4.5 with mutual information (MI) can gain the maximal accuracy for the NSLKDD dataset with just 17 features. Ustebay *et al.* [14], used recursive feature elimination (RFE) with RF for CICI-DS-2017. This dataset contains more than 80 features. They utilized RFE in the experiment to assess the outcomes of choosing 1 to 81 features. The most vital features, Src-port, flow-packets, flow-IAT-Std, and flow-IAT-mean, are selected. Then, MLP for IDS performed with a classification accuracy of 0.89. Because of the

small size of the dataset used to train the model, the performance is inadequate. In order to find important features for network-IDS, Khammassi and Krichen [15] implemented a wrapper-based FS that depends on genetic algorithms and, for classification, logistic regression is used. The outcomes show that their strategy produces detection rates with just 20 and 18 features for the UNSW-NB15 and KDD-Cup99 datasets, respectively. Alazzam *et al.* [10] proposed a new way for FS that depends on pigeon-inspired optimizer. The proposed method binarized continuous variables depends on the cosine similarity measure and is compared with the standard swarm algorithm, which utilizes a sigmoid function. The authors evaluated their method on the UNSW-NB, KDD-CUP99, and NLSKDD datasets. The proposed method outperformed various well-known FS methods owing to false positive rates (FPR), true positive rate (TPR), F1-score, and accuracy.

b. On ensemble learning classifiers

Furthermore, ensemble learning are standard ML techniques that mix various base learner models to minimize FPR and provide more reliable findings than just an individual model. Hsu *et al.* [16] used ensemble classifier technique for network-IDS, using support vector machine, auto encoder models, and RF. Depending on their outcomes, the researchers showed that ensemble classifiers reduce FAR and improve classification accuracy. Jabbar *et al.* [17] suggested an ensemble classifier (cluster-based) for IDS, which utilizes the k-nearest neighbor (KNN) algorithm and alternating decision tree. They showed that the proposed classifier performs better than other existing methods due to detection rate and accuracy. Kumar *et al.* [11], have introduced an ensemble model that relies on chi-squared automatic interaction detection (CHAID), C5, quick unbiased efficient statistical tree algorithms (QUEST), and classification and regression tree (CART) tree-based models. They have utilized the UNSW dataset for the training phase and to evaluate their work versus unseen attacks. The researcher decreases the count of features to only 13 by using MI FS. Then, they utilized decreased features to classify network attacks, which are probe, DoS, exploit, normal, and generic. The proposed model achieved an accuracy of 83.4%. Zhang *et al.* [18] proposed the Relief algorithm and information gain (IG) FS techniques with RF for IDS. On the NSL-KDD dataset, they performed three numbers of experiments. First, the authors investigate the performance utilizing the ReliefF algorithm and IG independently and then check it with their integrated method, ReliefF-IG. The ReliefF-IG method can initially utilize IG to decrease the number of features and then rank the significance by using the ReliefF algorithm, which results in reduced computation complexity and time needed for FS. The outcomes indicate that the ReliefF-IG method can obtain better accuracy than the single ReliefF and IG techniques. Megantara and Ahmad [19] used RFE and mean decrease in impurity (MDI), a hybrid-based FS with the NSL-KDD dataset. To determine the rank of features, MDI is used as a filter approach. Then RFE can therefore reduce the dimension of features via a decision tree classifier. The experimental results show that utilizing decision tree classifier U2R, R2L, probe, and DoS categories obtain 99.4%, 81.3%, 91.2%, and 89.1% accuracy in performance individually.

c. On hybrid methods

Nowadays, various hybrid methods employing both FS and ensemble approaches have been achieved to enhance the performance of the IDS. Kasongo and Sun [20] utilized an ensemble FS based on the Xgboost algorithm to IDS, and performance was evaluated on the UNSW-NB15 dataset utilizing ML approaches. The authors picked 19 out of 42 features due to Xgboost algorithm. The outcomes indicate that applied Xgboost algorithm with decision trees, the detection accuracy was enhanced by 1.9% relative to the benchmark performance utilizing all attributes. To build a model with high accuracy and low FPR, Malik *et al.* [21] proposed a hybrid approach of particle swarm optimization (PSO) and RF. The proposed method improves the accuracy of the model by choosing the most important features for each class. Pham *et al.* [22] proposed a hybrid approach that employs gain ratio (GR) as feature reduction and bagging to integrate tree-based classification models. Bagging models that employ J48 as the classifier model and use 35 features from the NSL-KDD dataset produced the highest performance in experiments. Tama *et al.* [23] proposed a new IDS that depends on integrated FS and two-stage classifier ensembles. The experimental outcomes demonstrate that it performs an important enhancement of the recall measure on the UNSW-NB15 and NSL-KDD datasets.

3. METHOD

3.1. Feature selection

FS is one of the crucial stages in ML techniques and IDSs. The determination of the appropriate FS approach and it's utilized in operations has an impact that will improve the performance of the IDS. It also has the impact of reducing the operational load as it reduces the number of features on the dataset and creates new relationships between features [24]. Therefore, there is no one way or technique for FS [25]. FS approaches can be classified into filter based, wrapper based, and embedded methods. In filter-based approaches, evaluate the significance of the features and the choice of the features (or attributes) depend on the statistics. Wrapper methods, on the other hand, use prediction performance as part of a subset of FS and evaluation operations. While embedding approaches are computationally less costly since they involve an

association between the choice of features and the learning procedure [26]. The technique to be utilized may differ due to the form of the dataset. The major issue in FS is choosing the feature that can effectively recognize between classes. Various FS approaches may be more suitable for various sets of data. Up-to-date intrusion datasets usually include lots of duplicate and useless features. Thus, the first stage in this research is to choose meaningful features and decrease the dimensionality of the used dataset. In this research, a hybrid approach merging by two techniques hard-voting and mean is proposed in order to boost the efficiency of the FS operation and improve the classification accuracy. The key role of this strategy is to assess the redundancy and the importance of the elected subset of features, which is explored in the provided search space in order to the optimum solution.

3.1.1. Mutual information

Having one attribute's information allows you to reduce the uncertainty in the other attribute to a certain extent. In other words, MI is a superior metric to demonstrate the interconnections among attributes X and Y, and it is known as [27]:

$$MI(X, Y) = \sum_{i=1}^n \sum_{j=1}^m P(X_i, Y_j) * \log \left(\frac{P(X_i, Y_j)}{(P(X_i) * P(Y_j))} \right) \quad (1)$$

3.1.2. ExtraTreesClassifier

An ExtraTreesClassifier is an ensemble learning approach that constructs numerous randomized decision trees to effectively model the data. This technique is designed to mitigate the risk of overfitting by introducing randomness into the process of splitting the data. Unlike traditional decision trees that determine split points based on metrics like entropy or Gini impurity [28], the ExtraTreesClassifier makes random splits for all observations in the dataset. By doing so, it encourages diversity among the constituent trees, ultimately contributing to a more robust and less prone-to-overfitting model.

3.1.3. Recursive feature elimination

RFE is a wrapper-based method. RFE starts by recursively eliminating predictors (features) and constructing a model depending on the remaining predictors. It utilizes model performance (i.e., accuracy) to decide which predictors engage the most in order to indicate the target predictor. RFE needs a specified number of predictors to keep, therefore, it is usually not known beforehand how many predictors are optimal [29]. To acquire precise predictors the KNN algorithm is used with RFE FS method in this work.

3.2. Base (single) learner in ensemble learning classifier

The two biggest challenges for each IDS are the FAR and classification accuracy. FAR reflects the number of normal instances detected as attacks (or anomalies), whereas accuracy indicates the number of accurately detected instances. The goal of selecting the base learner in an ensemble learning model is not only to decrease the FAR but also to improve the classification accuracy of the IDS. In this work, RF, XGBoost, and MLP were realized as the base classifier.

3.2.1. Random forest

RF is a machine-learning technique that depends on lots of decision trees. Initially, it specifies how many decision trees are required to be constructed and then utilizes the bootstrap method to randomly pick a group of data for each tree. RF builds its component decision trees in order to reduce the relationship between individual trees. The randomness in the FS operation contributes to the RF performance gains, not the split points in the decision trees of the selected features [30].

3.2.2. Extreme gradient boosting

XGBoost is a popular and effective algorithm. Gradient boosting is a supervised learning approach that combines a set of estimates from many weaker and simpler classifiers to accurately predict a target variable. The XGBoost algorithm achieves well in ML challenges owing to its powerful dealing of distributions, relationships, and a large variety of data types. Moreover, it can deal with a wide range of hyper-parameters that can be fine-tuned. XGBoost can address ranking, regression, and classification issues [28].

3.2.3. Multilayer perceptron

MLP is a neural network, containing an input layer, an output layer, and one or several hidden layers [31] as shown in Figure 1. For classification issues, the number of classes is exactly the number of nodes in the output layer, whereas the amount of features is exactly the number in the input layer. The layers among output and input layers are usually dense (fully-connected) layers and are trained through back-propagation.

When carrying out forward propagation, depending on a transfer function (activation function) from the preceding layer with bias and weight values, the network computes the output of each layer as (2):

$$f(x) = g\left(b^{(2)} + w^{(2)}\left(s\left(b^{(1)} + w^{(1)}x\right)\right)\right) \tag{2}$$

Where $f(x)$ denotes the output matrix, $w(1)$ and $w(2)$ are the weight matrices, $b(1)$ and $b(2)$ are the bias vectors, and g and s are the transfer functions. In our case, we use Relu as the transfer function for hidden layers, which transforms values less than 0 to 0, and the softmax function for the last output layer, which can help specify the best possible prediction.

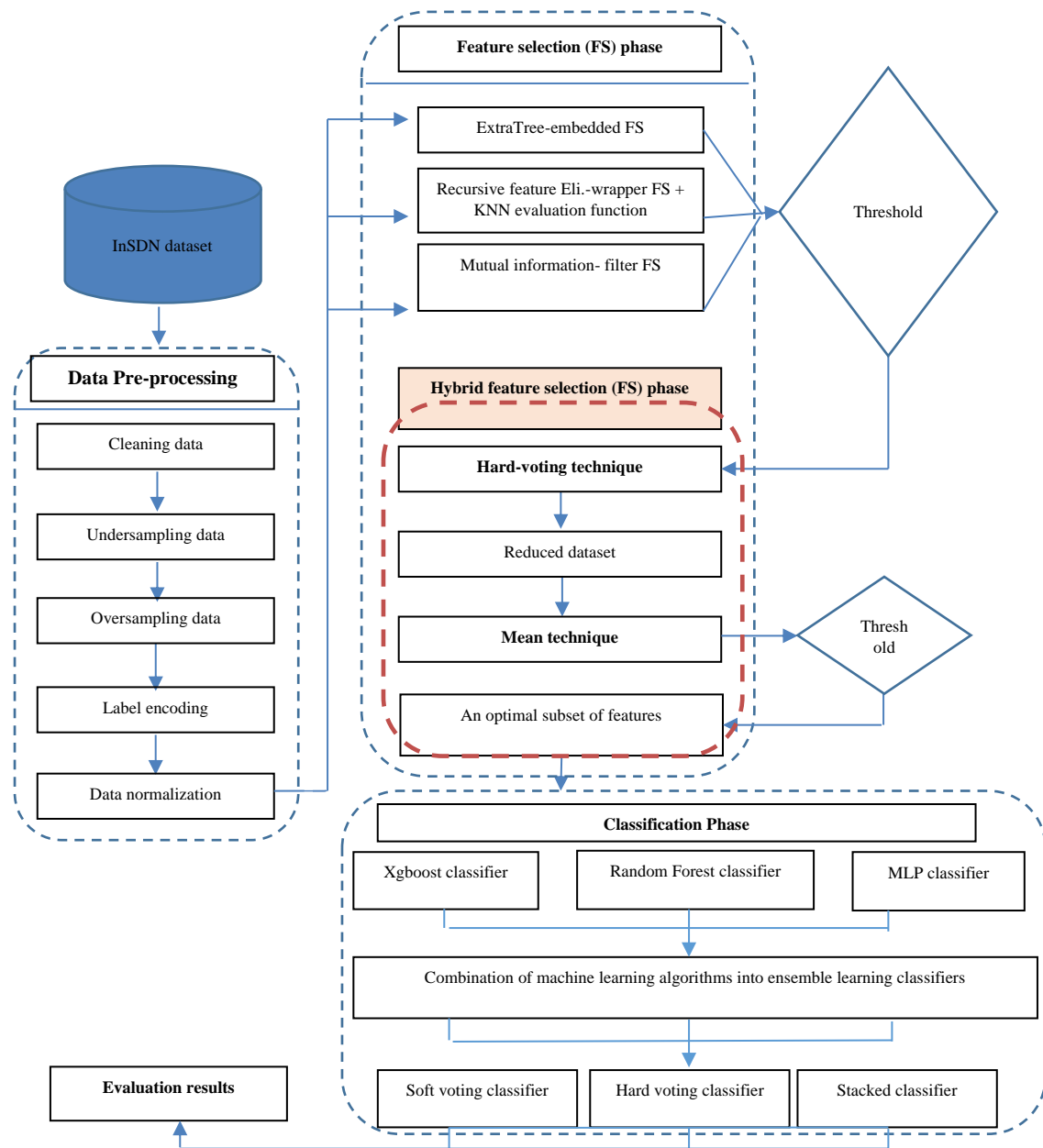


Figure 1. The proposed hybrid FS approach and ensemble learning classifiers

3.3. Proposed method

To improve the recognition capability of IDS and protect the network or service providers from assault, we suggest ensemble learning classifiers and a hybrid-based FS approach. Throughout the trials, we

divide the dataset into training and test sets to verify the efficiency of the models and identify normal data and diverse kinds of assaults (or attacks). Figure 1 shows the proposed hybrid FS and ensemble learning classifiers, which contains of the next four major stages:

3.3.1. Data pre-processing

The initial stage is to convert raw or initial data into an appropriate format for inspection by implementing cleaning data, undersampling data, oversampling data, label transforming (or encoding), and standardization of the InSDN dataset. The data preprocessing steps are defined in detail in section 4.2.

3.3.2. Feature selection

It is possible to assist the classification task in reaching its ultimate goal by selecting the appropriate feature and number of features. In addition to overwhelm the issue of the curse of dimensionality. Feature reduction phase is an important stage of model classification that could be carried out using a knowledge field or different intrinsic processes. In this work, the feature reduction stage is divided into two steps. In the first step, we trained separately three diverse FS approaches (extra tree classifier importance, RFE, and MI). Then, a hybrid approach is proposed to decrease the dimensions of the dataset and pick the most meaningful features for diverse kinds of attacks. The hybrid FS approach explains in detail in section 4.3.

3.3.3. Classification phase

In this stage, we employ the acquired reduced subset of features, obtained using the mean technique. These features are utilized to train three individual base models: XGBoost, RF, and MLP classifiers. Subsequently, ensemble learning classifiers, including hard voting, soft voting, and stacked models, are constructed based on these base models. This comprehensive approach aims to enhance the classification accuracy of the IDS. For more detailed, the classification phase is shown in section 4.4.

3.3.4. Evaluation results

Thanks to the ensemble learning classifiers, we can effectively identify and categorize various types of attacks as well as normal network traffic. These approaches achieve low FAR and high classification accuracies, providing robust defense mechanisms for our system. The evaluation results are demonstrated in detail in section 4.5.

4. RESULTS AND DISCUSSION

4.1. Dataset description

One of the widespread issues for ML IDSs is the unavailability of the datasets. Illegal issues and privacy are the primary reason for the absence of datasets in the IDS area. The network traffics include sensitive information, where the visibility of such information might disclose company and clients' secrets. To solve the preceding gap, several authors are generating their information to avoid any sensitive issues. In this research paper, we assess the proposed ensemble learning classifiers with hybrid FS utilizing the newly released InSDN dataset [32]. The InSDN dataset contains recent popular attack types like DDoS, Probe, DoS, Botnet, password-guessing, web, and exploitation. Moreover, the normal network traffic in the dataset includes common applications like hypertext transfer protocol (HTTP), hypertext transfer protocol secure (HTTPS), electronic mail (Email), distribute database system (DNS), secure shell (SSH), and file transfer protocol (FTP). The InSDN dataset contains 361,317 observations for attacks and normal traffic, wherein 292,893 for attack observations and 68,424 for normal observations. Table 1 shows how these data observations are distributed.

Table 1. The proportion of classes in the InSDN-dataset

Labels	Instances	%
Legitimate	68.424	18.93
Exploitation (U2R)	17	0.0047
DoS	69.044	19.105
Probe	98.129	27.15
Password-Guessing	1405	0.388
BOTNET	164	0.0453
Web-Attack	192	0.053
DDoS	123.942	34.3
Sum	361.317	--

4.2. Data pre-processing

In order to train the model, it is necessary to prepare the dataset because the data can be duplicated, noisy, inconsistent, and incomplete. The InSDN dataset is known in a flow_based form, with over 80 features obtained using the CICFlowMeter tool [33]. Thus, in this research paper, the pre-processing stage contains cleaning data, under sampling data, oversampling data, label encoding, and data normalization.

- Cleaning data: the dataset includes socket information like Src-IP, Dst-IP, flow-ID, and so on. All these features are eliminated to overcome the over-fitting issue, and in addition, the socket features may vary from one network to another.
- Undersampling data: different portions in dataset classes lead to dataset imbalance, which is a serious issue that affects ensemble learning and degrades the model's performance. Therefore, we eliminate instances picked randomly from normal, DDoS, probe, and DoS classes.
- Oversampling data: from the statistics of the InSDN dataset as shown in Table 1, we can observe that the number of observations (examples) for web attack, botnet, and exploitation are small, and most ML approaches will ignore them, causing bad performance. Therefore, we used the SMOTE technique to duplicate the instances in minority classes. The new statistics of the InSDN dataset after performing undersampling and oversampling are depicted in Table 2.

Table 2. The new statistics of classes in the InSDN-dataset

Labels	New instances	%
Legitimate	35114	30.52
Exploitation (U2R)	701	0.609
DoS	26313	22.87
Probe	25235	21.93
Password-Guessing	1951	1.69
BOTNET	701	0.609
Web-Attack	1257	1.092
DDoS	23767	20.65
Sum	115039	--

- Label encoding: the used dataset includes continuous, binary, and symbolic values. For example, the attribute 'protocol' in the InSDN dataset contains symbolic worth for instance: "udp", "icmp", and "tcp". Because various models receive just numerical inputs, the conversion step is regarded as crucial and has an important effect on IDS classification accuracy. In this research, we substitute each specific worth with an integer number to deal with the non-numerical features.
- Normalization: various ranges between features can degrade ensemble learning classifiers, for instance, a feature that takes on a high integer value, such as 'Flow Duration' can dominate the classification performance. Thus, we utilized a fast and simple normalization method named min-max method [34] using (3) to map the feature values into 0–1 range.

$$\underline{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3)$$

Wherein x_{max} and x_{min} denote the maximum and minimum values of feature x .

4.3. Feature reduction or selection phase

It is crucial to select the most meaningful features in order to achieve a low FAR, high detection rate, and low computational time. In this research, feature reduction (FR) is divided into two steps. First, three FR approaches (shown in section 4) were tested separately on the InSDN dataset. Due to the threshold value, each of them produces a different subset of features, wherein each feature that does not improve classifier performance is removed. Then, a hybrid FR method is proposed to exclude irrelevant and inconsistent features and provide a reduced dataset (an optimal subset of features), and utilizing that reduced dataset, ensemble learning classifiers can provide good results in various types of classification issues. The hybrid method contains two techniques. First, a hard-voting technique is used to reduce the subset of features from 77 to 15, where each feature that has just one vote will be removed, for instance, 'Fwd Pkts/s' feature is selected from only a MI method. On the other hand, each feature that has two or three votes will be selected. Second, the mean technique is used to combine the outputs of three FR approaches and assign a single weight for each feature because each feature has three weights from the previous step. This technique reduced the number of features from 15 to 10.

4.4. Classification phase

When an optimal subset of features is picked through the proposed hybrid FS approach, it will be fed into the classification phase (training stage), where two steps are employed. In the first step, three classifier models (i.e., RF, XGBoost algorithm, and MLP) are trained individually. In the second step, in order to achieve better classification accuracy, the classifier models commonly integrate numerous base classifiers in some manner. These models are efficient to handle the same issue and jointly perform a predicting outcome with constancy and improved accuracy by constructing numerous separate classifiers and merging them.

The aims for utilizing ensemble learning classifiers to enhance the efficiency are computational reason, generalization problem, and statistical reason. Initially, maybe an individual classifier is not sufficient to achieve the optimal generalization in the hypothesis area, thus, it is required to merge individual models to boost the model's performance. Secondly, when the original dataset is not adequate to train the classifier, the outcome may guide to a false or weak hypothesis. Finally, to produce an appropriate hypothesis, a separate classifier might consume a large amount of computational time, in which case the technique could be more likely to encounter issues.

Hard voting, soft voting, and stacking classifiers are most common in ensemble learning. They typically achieve better outcomes in classification tasks and are extensively used to construct many ensemble learning models. Furthermore, ensemble approaches have been demonstrated to enhance accuracy in several scenarios, such as IDS. Ensemble learning classifiers provide techniques for security experts to inspect similarities to previously known harmful or normal samples. Among tree-based methods, RF and XGBoost have been extensively utilized in the area of anomaly detection owing to their simple parameters and high efficiency, in addition to MLP are chosen to build the ensemble for multi classification IDS in this research.

4.5. Analysis of results

The proposed approach is assessed depending on its ability to classify network-traffic data into a valid kind. To assess the efficiency of the hard voting, soft voting, and stacking models, six metrics were used to assess the performance of the final classification for the proposed approach. The mathematical computations of the used evaluation criteria are clarified in [35]. Before the proposed hybrid approach for FS, we preprocessed the InSDN dataset samples to overcome overfitting features. Then, we implemented extra tree classifier importance (ETCI), RFE, and MI methods on the training set to attain the significant ranking of 77 features as shown in Table 3.

Table 3. The details of selected important features from InSDN dataset

FS techniques	Feature names
Embedded methods-ExtraTreeClassifier importance-features	Protocol, Init Bwd Win Byts, SYN Flag Cnt, FIN Flag Cnt, ACK Flag Cnt, Down/Up Ratio, Bwd Pkts/s, Flow Pkts/s, Bwd Header Len, Pkt Len Std, Bwd Seg Size Avg, Pkt Len Max, Bwd Pkt Len Mean, Bwd Pkt Len Max, Subflow Bwd Pkts, Fwd Header Len
Wrapper-based approaches-RFE-features	Protocol, Flow Duration, Tot Fwd Pkts, Fwd Pkt Len Min, Fwd Pkt Len Max, Fwd Pkt Len Mean, Fwd Pkt Len Std, Flow Pkts/s, Bwd Pkt Len Max, Bwd Pkt Len Min, Bwd Pkt Len Mean, Flow Byts/s, Flow IAT Mean, Fwd IAT Max, Fwd IAT Tot, Fwd IAT Mean, Bwd IAT Tot, Bwd IAT Mean, Bwd IAT Max, Bwd IAT Min, Fwd Header Len, Bwd Header Len
Filter-based methods-MI-features	Bwd Header Len, Init Bwd Win Byts, Fwd Header Len, Bwd IAT Tot, Bwd IAT Mean, Bwd IAT Max, Protocol, Bwd Pkts/s, Flow Pkts/s, Flow IAT Mean, Tot Fwd Pkts, Subflow Fwd Pkts, Flow Duration, Subflow Bwd Pkts Tot Bwd Pkts
Hard-voting technique (reduced subset of features)	Bwd Header Len, Bwd IAT Max, Bwd IAT Mean, Bwd IAT Tot, Bwd Pkt Len Max, Bwd Pkt Len Mean, Bwd Pkts/s, Flow duration, Flow IAT Mean, Flow Pkts/s, Fwd Header Len, Init Bwd Win Byts, Protocol, Subflow Bwd Pkts, Tot Fwd Pkts
Mean technique (an optimal subset of features)	Bwd Header Len, Protocol, Fwd Header Len, Bwd Pkts/s, Bwd IAT Tot, Bwd IAT Mean, Bwd IAT Max, Flow Pkts/s, Flow IAT Mean, Init Bwd Win Byts

There are some features that have the least significant ranking in these methods, which may reduce the model's performance. We select 0.01 and 0.7 as the thresholds for ETCI and MI FS approaches, respectively, to exclude insignificant features, while the unwanted features are excluded recursively owing to the scoring model (i.e., accuracy) utilizing the RFE method on KNN algorithm. Thus, in ETCI features with a significant value greater than or equal to 0.01 were kept, whereas RFE chose the features using the KNN algorithm ($k=3$), and the selected features (29 features) had an accuracy of 0.98 were kept in RFE while in MI features with a significant value greater than or equal to 0.7 were kept. After eliminating insignificant features from three methods, respectively, three subsets of features were attained. 15 features were kept by ETCI, 22 features were kept by RFE, and 15 features were kept by MI.

Two hybrid techniques were used to obtain their union group: hard-voting and mean. In hard-voting, 15 features were kept based on hard opinion (or majority voting), which means that the feature is selected if and only if it has two or three votes. Then, using the mean technique, because each feature has three weights, take their sum and divide it by three to uniform these weights. After that, we pick 0.11 as a threshold for mean technique and obtained the optimal subset of features that includes just 10 features. After implementing the proposed hybrid-based feature reduction approach on InSDN dataset as shown in Algorithm 1, 10 significant features were finally chosen which can be further used in the classification phase. Table 3 depicts the details of an optimal subset of features for InSDN dataset. Finally, to greatly enhance the model performance of IDS, hard voting classifier, soft voting classifier, and stacked model are proposed.

Algorithm 1. The proposed hybrid FS approach

Input: list1-ExtraaTreeImp-sorted, list2-RFE-sorted, list3-mutualInf-sorted

Output: Opt-sorted-list //an optimal subset of features

Begin

Hard-voting-dic=0, mean-dic=0

For each feature

If that feature is found in at least two lists then

// **hard-voting technique**

Save the feature in a hard-voting-dic. // **the output of this step is a reduced subset of features.**

End

For each feature in a hard-voting-dic

Acquire that feature and get all its three weights from list1-ExtraaTreeImp-sorted, list2-RFE-sorted, and list3-mutualInf-sorted and then compute a mean value and save it in a mean-dic.

End

For each feature in the mean-dic

If the weight of that feature exceeds a threshold value

// Mean technique

Save it in an Opt-sorted-list

End

Sort the opt-sorted-list from most weighted to least weighted.

Return an Opt-sorted-list // This list contains an optimal subset of features, which is then used to train the proposed models.

4.5.1. Comparison performance between our proposed FS approach and with no FS

To assess the efficacy of the proposed IDS approach, we compared the results among no FS and the suggested hybrid FS to identify normal instances and attacks. Gratitude to the chosen of important features (an optimal subset of features) via the proposed hybrid FS approach, the average worths of these measures, like accuracy, detection rate, precision, F1-score, and FAR, have improved significantly. Table 4 sums up the classification performance depend on the InSDN-dataset, which contains the outcomes of the single learner classifier and ensemble learning models. It is shown that the ensemble learning models are not better enough in many measures with no FS approaches.

Table 4. Comparison performance for individual and ensemble classifiers with no FS approaches

Classifier	Accuracy	Precision	Recall	F1-Score	FAR
RF	0.948	0.939	0.941	0.939	0.05
XGBoost	0.938	0.936	0.938	0.937	0.06
MLP	0.934	0.933	0.934	0.933	0.07
Ensemble-Hard-voting	0.976	0.975	0.974	0.975	0.05
Ensemble-Soft-voting	0.977	0.976	0.977	0.976	0.04
Ensemble-Stacked model	0.976	0.975	0.976	0.975	0.06

On the other hand, Table 4 indicates that the proposed hybrid FS with all ensemble learning classifiers achieves good results on the InSDN dataset. In particular, ensemble soft voting model yields the best and the highest accuracy of 0.999, precision or positive predictive value of 0.998, sensitivity or recall of 0.998, F1 score of 0.998 which is a good measure to utilize for the performance of each classifier, in the multi-classification task, and the minimum FAR of 0.02 relies on the InSDN dataset. Also, every single classifier utilizing an optimal subset of features achieves higher detection accuracy, precision, recall, F-score, and lower FAR compared to the single classifier with full features as shown in Tables 4 and 5, which significantly confirms the importance of the proposed hybrid FS approach.

Table 6 shows the building (or training) and testing times for all single and ensemble classifiers on the InSDN dataset. According to the number of selected features, the proposed hybrid FS approach with all

single and ensemble sharply decreases the time overhead. It is also shown that the ensemble stacked model and single classifiers are faster than the ensemble soft voting classifier, which is neglected due to improved performance. The aim here is to come to terms between accuracy and speed-up to gain the best outcome possible. Overall, the outcomes illustrate that the number of selected features impacts the time required to construct and test all classifiers. At last, from Tables 4-6 the prediction outcomes of the individual and ensemble classifiers utilizing the optimal subset of features are all greater than the classifiers utilizing all features in terms of all metrics. Thus, the proposed hybrid FS approach with all proposed ensemble learning classifiers can be crucial in distinguishing between normal and attack instances.

Table 5. Comparison performance for individual and ensemble classifiers by utilizing our proposed hybrid FS approach

Classifier	Accuracy	Precision	Recall	F1 score	FAR
RF	0.985	0.984	0.985	0.984	0.03
XGBoost	0.978	0.967	0.978	0.974	0.05
MLP	0.980	0.977	0.971	0.973	0.04
Ensemble-Hard voting	0.998	0.997	0.998	0.997	0.03
Ensemble-Soft voting	0.999	0.998	0.998	0.998	0.02
Ensemble-Stacked model	0.997	0.996	0.995	0.996	0.02

Table 6. Summary of building time and test time comparison

Classifier	With no FS approach		With our proposed hybrid FS approach	
	Building or training time (sec)	Testing time (sec)	Building or training time (sec)	Testing time (sec)
RF	32.325	1.465	16.161	1.461
XGBoost	306.562	1.812	128.211	1.610
MLP	174.729	0.672	41.327	0.276
Ensemble-Hard voting	466.914	3.991	237.494	4.314
Ensemble-Soft voting	462.483	4.936	232.502	4.294
Ensemble-Stacked model	342.98	3.633	233.905	3.627

5. CONCLUSION

The objective of this work is to present the significance of hybridization FS and ensemble learning classifiers in order to enhance performance of the IDS. We propose a new IDS approach, which depends on hybrid FS and ensemble learning classifiers to reduce the high dimension and deal with unbalanced data networks with low FAR as well as improve accuracy and F1-score metrics. A proposed hybrid FS approach depends on fusion of three different FS approaches using hard-voting and mean techniques. First, we get 15 relevant features due to the hard-voting technique via extra tree classifier importance, RFE, and MI. So, according to these three methods, each feature has three diverse weights; therefore, a mean technique has been used, to assign one weight to each feature and obtain an optimal subset of features with just 10 features. Then, the ensemble learning classifiers depend on RF, XGBoost, and MLP combined by hard voting, soft voting, and stacked model are introduced to build the prediction model. Finally, the suggested IDS is validated by utilizing an up-to-date InSDN dataset.

Experimental outcomes indicated that all the proposed ensemble learning classifiers had low FAR and high accuracy, and the best classifier due to accuracy, precision, recall, F1 score, and FAR is the ensemble (RF+XGBoost+MLP) combined by soft voting method with the subset of features (10 f) obtained via our proposed hybrid FS approach. Accuracy of this classifier is 0.999, precision or positive predictive value is 0.998, sensitivity or recall is 0.998, F1 score is 0.998, which is a good measure to utilize for the performance of each classifier, in the multi-classification task, and the minimum FAR is 0.02. Since ensemble soft voting achieved higher accuracy, higher F1 score, lower FAR, and decrease significantly the training (or building) time from 462.483 s to 232.502 s, it is clear that ensemble soft voting is very effective in detecting and classifying normal and attacks type. Also, when compared to utilizing full features (77 f), that is, without FS approach, it illustrates improving performance on diverse measures. Ensemble soft voting classifier had the best and highest performance, so we can show that the proposed hybrid FS with ensemble soft voting classifier is the most efficient and accurate one compared to other single or ensemble classifiers trained in this work.

In data preprocessing stage, we solved the imbalanced dataset issue to avoid degrading the model's performance. Therefore, we performed both undersampling on majority classes (normal, DDoS, probe, and DoS) and oversampling (i.e., SMOTE technique) on minority classes (web attack, botnet, and exploitation).

In the future, we plan to implement the proposed FS approach over other IDS datasets with different ML models.

ACKNOWLEDGEMENTS

We would like to thank the University of Babylon, Hillah, Iraq.




REFERENCES

- [1] L. Marinos and M. Lourenço, *Threat Landscape Report 2018 - 15 Top Cyberthreats and Trends*, no. January. Greece: ENISA, 2018, doi: 10.2824/622757.
- [2] "ENISA Threat Landscape 2020 - Distributed denial of service", Sep. 2020. Available: <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2020-distributed-denial-of-service>. [Accessed: Sep. 28, 2023]
- [3] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, and P. Faruki, "Network Intrusion Detection for IoT Security Based on Learning Techniques," *IEEE Communications Surveys and Tutorials*, vol. 21, no. 3, pp. 2671–2701, 2019, doi: 10.1109/COMST.2019.2896380.
- [4] S. Latha and S. J. Prakash, "A survey on network attacks and intrusion detection systems," in *2017 4th International Conference on Advanced Computing and Communication Systems, ICACCS 2017*, 2017, pp. 1–7, doi: 10.1109/ICACCS.2017.8014614.
- [5] M. Liu, Z. Xue, X. Xu, C. Zhong, and J. Chen, "Host-Based Intrusion Detection System with System Calls," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–36, Sep. 2019, doi: 10.1145/3214304.
- [6] R. Singh, H. Kumar, R. K. Singla, and R. R. Ketti, "Internet attacks and intrusion detection system," *Online Information Review*, vol. 41, no. 2, pp. 171–184, Apr. 2017, doi: 10.1108/OIR-12-2015-0394.
- [7] H. T. Elshoush and I. M. Osman, "Alert correlation in collaborative intelligent intrusion detection systems—A survey," *Applied Soft Computing*, vol. 11, no. 7, pp. 4349–4365, Oct. 2011, doi: 10.1016/j.asoc.2010.12.004.
- [8] A. Drewek-Ossowicka, M. Pietrolaj, and J. Rumiński, "A survey of neural networks usage for intrusion detection systems," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 1, pp. 497–514, Jan. 2021, doi: 10.1007/s12652-020-02014-x.
- [9] L. Li, Y. Yu, S. Bai, J. Cheng, and X. Chen, "Towards Effective Network Intrusion Detection: A Hybrid Model Integrating Gini Index and GBDT with PSO," *Journal of Sensors*, vol. 2018, pp. 1–9, 2018, doi: 10.1155/2018/1578314.
- [10] H. Alazzam, A. Sharieh, and K. E. Sabri, "A feature selection algorithm for intrusion detection system based on Pigeon Inspired Optimizer," *Expert Systems with Applications*, vol. 148, pp. 1–16, Jun. 2020, doi: 10.1016/j.eswa.2020.113249.
- [11] V. Kumar, D. Sinha, A. K. Das, S. C. Pandey, and R. T. Goswami, "An integrated rule based intrusion detection system: analysis on UNSW-NB15 data set and the real time online dataset," *Cluster Computing*, vol. 23, no. 2, pp. 1397–1418, Jun. 2020, doi: 10.1007/s10586-019-03008-x.
- [12] S. Maza and M. Touahria, "Feature Selection Algorithms in Intrusion Detection System: A Survey," *KSII Transactions on Internet and Information Systems*, vol. 12, no. 10, pp. 5079–5099, Oct. 2018, doi: 10.3837/tiis.2018.10.024.
- [13] H. S. Hota and A. K. Shrivastava, "Decision Tree Techniques Applied on NSL-KDD Data and Its Comparison with Various Feature Selection Techniques," 2014, pp. 205–211, doi: 10.1007/978-3-319-07353-8_24.
- [14] S. Ustebay, Z. Turgut, and M. A. Aydin, "Intrusion Detection System with Recursive Feature Elimination by Using Random Forest and Deep Learning Classifier," *International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism, IBIGDELFT 2018 - Proceedings*, pp. 71–76, 2019, doi: 10.1109/IBIGDELFT.2018.8625318.
- [15] C. Khammassi and S. Krichen, "A GA-LR wrapper approach for feature selection in network intrusion detection," *Computers and Security*, vol. 70, pp. 255–277, 2017, doi: 10.1016/j.cose.2017.06.005.
- [16] Y.-F. Hsu, Z. He, Y. Tarutani, and M. Matsuoka, "Toward an Online Network Intrusion Detection System Based on Ensemble Learning," in *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*, Jul. 2019, pp. 174–178, doi: 10.1109/CLOUD.2019.00037.
- [17] M. A. Jabbar, R. Aluvalu, and S. S. S. Reddy, "Cluster Based Ensemble Classification for Intrusion Detection System," in *Proceedings of the 9th International Conference on Machine Learning and Computing*, Feb. 2017, pp. 253–257, doi: 10.1145/3055635.3056595.
- [18] Y. Zhang, X. Ren, and J. Zhang, "Intrusion detection method based on information gain and ReliefF feature selection," in *2019 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2019, pp. 1–5, doi: 10.1109/IJCNN.2019.8851756.
- [19] A. A. Megantara and T. Ahmad, "Feature Importance Ranking for Increasing Performance of Intrusion Detection System," in *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*, Sep. 2020, pp. 37–42, doi: 10.1109/IC2IE50715.2020.9274570.
- [20] S. M. Kasongo and Y. Sun, "Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset," *Journal of Big Data*, vol. 7, no. 105, pp. 1–20, Dec. 2020, doi: 10.1186/s40537-020-00379-6.
- [21] A. J. Malik, W. Shahzad, and F. A. Khan, "Network intrusion detection using hybrid binary PSO and random forests algorithm," *Security and Communication Networks*, vol. 8, no. 16, pp. 2646–2660, Nov. 2015, doi: 10.1002/sec.508.
- [22] N. T. Pham, E. Foo, S. Suriadi, H. Jeffrey, and H. F. M. Lahza, "Improving performance of intrusion detection system using ensemble methods and feature selection," in *Proceedings of the Australasian Computer Science Week Multiconference*, Jan. 2018, pp. 1–6, doi: 10.1145/3167918.3167951.
- [23] B. A. Tama, M. Comuzzi, and K.-H. Rhee, "TSE-IDS: A Two-Stage Classifier Ensemble for Intelligent Anomaly-Based Intrusion Detection System," *IEEE Access*, vol. 7, pp. 94497–94507, 2019, doi: 10.1109/ACCESS.2019.2928048.
- [24] P. Dhal and C. Azad, "A comprehensive survey on feature selection in the various fields of machine learning," *Applied Intelligence*, vol. 52, no. 4, pp. 4543–4581, Mar. 2022, doi: 10.1007/s10489-021-02550-9.
- [25] V. Hajisalem and S. Babaie, "A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection," *Computer Networks*, vol. 136, pp. 37–50, May 2018, doi: 10.1016/j.comnet.2018.02.028.
- [26] A. Destrero, S. Mosci, C. De Mol, A. Verri, and F. Odone, "Feature selection for high-dimensional data," *Computational Management Science*, vol. 6, no. 1, pp. 25–40, 2009, doi: 10.1007/s10287-008-0070-7.
- [27] X. Su, L. Li, F. Shi, and H. Qian, "Research on the Fusion of Dependent Evidence Based on Mutual Information," *IEEE Access*, vol. 6, pp. 71839–71845, 2018, doi: 10.1109/ACCESS.2018.2882545.
- [28] D. R. Patil and T. M. Pattewar, "Majority Voting and Feature Selection Based Network Intrusion Detection System," *EAI*




- Endorsed Transactions on Scalable Information Systems*, vol. 22, no. 6, 2022, doi: 10.4108/eai.4-4-2022.173780.
- [29] K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sensors and Actuators, B: Chemical*, vol. 212, pp. 353–363, 2015, doi: 10.1016/j.snb.2015.02.025.
- [30] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/BF00058655.
- [31] M. T. Camacho Olmedo, M. Paegelow, J. F. Mas, and F. Escobar, "Geomatic Approaches for Modeling Land Change Scenarios. An Introduction," Switzerland: Springer International Publishing AG 2018, 2018, pp. 1–8, doi: 10.1007/978-3-319-60801-3_1.
- [32] M. S. Elsayed, N.-A. Le-Khac, and A. D. Jurcut, "InSDN: A Novel SDN Intrusion Dataset," *IEEE Access*, vol. 8, pp. 165263–165284, 2020, doi: 10.1109/ACCESS.2020.3022633.
- [33] I. Ahmad, S. Namal, M. Ylianttila, and A. Gurtov, "Security in Software Defined Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2317–2346, 2015, doi: 10.1109/COMST.2015.2474118.
- [34] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Data Preprocessing for Supervised Learning," *International Journal of Computer Science*, vol. 60, no. 1–2, pp. 111–117, 2011.
- [35] S. Elhag, A. Fernández, A. Altalhi, S. Alshomrani, and F. Herrera, "A multi-objective evolutionary fuzzy system to obtain a broad and accurate set of solutions in intrusion detection systems," *Soft Computing*, vol. 23, no. 4, pp. 1321–1336, Feb. 2019, doi: 10.1007/s00500-017-2856-4.

BIOGRAPHIES OF AUTHORS



Hasanain Ali Al Essa    received his bachelor degree in Computer Science from the University of Babylon in 2007, Iraq. He received his Master Degree in Computer and Information Science from Tula State University in 2014, Russia. Currently, he is a Lecturer at Department of Computer Science, Babil, Iraq. His research interests include machine learning, artificial intelligence, deep learning, computer security, and computer network. He can be contacted at email: hasanain@uobabylon.edu.iq.



Wesam S. Bhaya    is currently working as Professor at University of Babylon, Iraq. He has completed his Ph.D. in Computers and Informatics from Informatics Institute for Postgraduate Studies, Iraq. His main area of interest focuses on computer virus and antivirus, information warfare, computer security, computer network, computer architecture, operating systems, assembly programming language, C programming language, delphi programming language, mobile programming, internet system installation, network devices installation, computer maintenance, social networks, and steganography, where he is the author/co-author of over 30 research publications. He can be contacted at email: wesambhaya@uobabylon.edu.iq.