*International Journal of Research in Engineering and Science (IJRES)*
*ISSN (Online): 2320-9364, ISSN (Print): 2320-9356*
*www.ijres.org Volume 11 Issue 1 ǁ January 2023 ǁ PP. 520-527*

# A Review of Intrusion Detection Technique Based on Hybrid Ensemble Learning Algorithms

Date of Submission: 17-01-2023 Date of acceptance: 02-02-2023

## I. Introduction:

Every day, different types of new cyber-attacks are discovered, and their sources are becoming more hazardous. As a result, detecting zero-day attacks is a difficult operation that potentially jeopardizes business continuity [1]. Computer attacks are becoming increasingly complex, posing difficulties in accurately detecting the intrusion [2,3]. Network Intrusion detection systems (NIDSs) are meant to monitor computer networks for unusual activities that a regular packet filter would miss Traditional IDSs have a number of flaws, such as the inability to discriminate between new malicious threats; the need for modification; poor accuracy; and a high rate of false alerts. Therefore, machine learning is used to detect new attacks. However, machine learning encounters many challenges because it enhances the computational and time complexity of the task by expanding the search space [4,5]. Numerous studies have been conducted on the use of multiple classifiers instead of single ones and the principle of ensemble learning techniques to ensure high accuracy and a low false alarm rate [6–8]. As a result, ensemble learning can be divided into three categories (i.e., bagging, stacking, and boosting) [9–11]. It is a general meta-approach to machine learning to combine predictions from multiple models to improve predictive performance. Althoughan infinite number of ensembles for any predictive modeling can be created, the subject of ensemble learning is dominated by three methods. The first category of the ensemble is bagging. It is the process of fitting multiple decision trees to different samples of the same dataset and then averaging the results [12]. Alternatively, the stacking method involves fitting many different model types onto the same data and using another model to learn how to best combine the predictions [13]. Boosting involves sequentially adding ensemble members that correct the predictions made by prior models and output a weighted average of the predictions [14]. Recently, researchers applied hybrid principles in feature selection and ensemble methods. Feature selection is a useful approach for intrusion detection systems. This method discovers extremely important features and discards unnecessary ones while causing minimal performance reduction [15–17]. Correlation-based feature selection (CFS) selects strong affinity on similarities that are used as a heuristic evaluation function. The function compares feature vector subsets that are related to the class label but are not associated with one another. The CFS algorithm implies that irrelevant characteristics have a low association with the class and should be removed consequently. Excess traits, on the other hand, should be investigated since they are typically associated with one or more of the other characteristics [18]. The classification algorithms used in ensemble learning frequently mix numerous basic classifiers in some fashion. The proposed work will benefit from this feature of ensemble learning by developing and integrating multiple distinct models, these classifiers are effective at dealing with the same problem and, when combined, produce a predicting output that is more stable and accurate. To begin with, a single classifier may not always be competent to produce the best representation in the hypothesis space. Thus, the use of multiple independent classifiers is necessary to improve prediction performance. Second, a false or inaccurate hypothesis can develop if the training dataset for the learning algorithm is insufficient.

Every day, various types of cyber-attacks are found, and their origins are evolving more dangerous. Identifying zero-day attacks is a complex function that potentially jeopardizes business continuity [1]. Computer hacks are becoming increasingly tricky, posing problems in identifying the intrusion [2,3]. Network Intrusion detection systems (NIDS) are indicated to observe computer networks for unexpected actions that a daily packet filter would miss. Usual IDSs have different flaws, such as the incapability to differentiate between new malicious risks, the need for an update; low precision; and an increased rate of false signals.Moreover, machine learning is utilized to identify new hacks. Therefore, machine learning faces many difficulties because it improves the estimation and time difficulty of the task by extending the search space [4,5]. Multiple analyses have been operated on utilizing multiple classifiers rather than single ones and the rules of ensemble learning approaches to confirm high precision and a low false signal rate [6–8]. As a result, ensemble learning can be split into three classes (i.e., bagging, stacking, and boosting) [9–11]. It is a common meta-strategy for machine

learning to mix estimations from various methods to enhance estimated performance. Although an unlimited number of ensembles for any estimated modeling can be developed, the subject of ensemble learning is overwhelmed by three models. The first category of the costume is bagging. It is the methodology of fitting various decision trees to different examples of the same dataset and then averaging the impacts [12]. Alternatively, the stacking technique involves working for various approach classes onto the same data and utilizing another approach to learn how to best mix the forecasts [13]. Boosting involves consecutively adding ensemble members that update the predictions made by primary models and result in a valued average of the projections [14]. Currently, researchers use hybrid guides in feature selection and ensemble learning. Feature selection is a helpful method for intrusion detection methods. This approach discovers important features and discards unwanted ones while causing minimal implementation decrease [15–17]. Correlation-based feature selection (CFS) selects strong relations on similarities utilized as a heuristic evaluation procedure. The procedure compares feature vector subsets connected to the class label but not correlated with one another. The CFS algorithm exposes irrelevant factors with a low relationship with the class and should be deleted. Extra features should be analyzed since they are generally linked with one or more other factors [18]. The classification algorithms utilized in ensemble learning continually combine multiple basic classifiers. The suggested work will satisfy this element of ensemble learning by creating and combining different distinct models; these classifiers are valid at negotiating the same issue and, when mixed, produce an estimated more durable and perfect result.First, a single classifier may not always be qualified to create the best presentation in the theory space. Thus, utilizing multiple independent classifiers is essential to enhance forecast execution. Second, a false or incorrect theory can create if the training dataset for the learning algorithm is inadequate.

## II. Literature Review:

https://assets.researchsquare.com/files/rs-591679/v1_covered.pdf?c=1631871021

In the 1970s, the need for security systems is felt more than ever due to the increasing speed, efficiency, number of computers. In 1977 and 1978, the International Standard Organization held a meeting between governments and inspection bodies of Electronic Data Processing that the outcome that meeting was to prepare a report on the status of security, inspection, and control of systems at that time. At the same time, the US Department of Energy began very detailed studies on the inspection and the security of computer systems due to concerns about the security of its systems. This study was carried out by a person named James P. Anderson. The Report presented by Anderson in 1980, can be introduced as the main core of the concepts of intrusion detection [19]. Raman Singh et al. [42] presented a proposed system based on Extreme Learning Machine. This machine solves the problem of the neural network in terms of speed. This system was used aimed to reduce computational memory and time using creating a profile of network traffic, and as well as two alpha and beta profiles were used. Alpha and beta profiles can reduce the effect of unaligned data. The beta profile can reduce the size of the experimental data set, while its features are maintained in practice, and the alpha profile is used to reduce the effect of discovery time. Folino et al. [11] used an intrusion detection system based on ensemble classification, aimed to increase group accuracy. The ensemble structure of the NIDS makes possible the detection of sophisticated attacks and alarms in a proper manner, and the advantages of using this ensemble classifier include reducing error variance and bias, and it is appropriate for unbalanced classification. The proposed method works well to identify attacks and minimize the alarms but needs to be improved for specific attacks. Aburomman et al. [1] presented a new method based on the support vector machine, K-Nearest Neighbor, and particle algorithm, and the weighted majority algorithm classifier for the intrusion detection system. Six support vector machine classifiers and six K-Nearest Neighbor classifiers with different values have been used in this method. Then, WMA was used as a classifier combination. The Local uni-modal sampling (LUS) algorithm was used to select high-quality parameters. The proposed method has used LUS-WMA that has better accuracy than a method that uses the WMA classifier, but the performance of WMA alone is better than the proposed method. Gautam et al. [14] used the proposed algorithm based on information theory and entropy, in which this algorithm obtained the entropy after the classi-fication of features, and classification is based on bias and features. The results show that the rate of detection and accuracy of the proposed algorithm is better than the Fast Feature Reduction in Intrusion Detection Data sets (FFRIDD) and Multi-Level Dimensionality Reduction Methods (MLDRM) selection algorithm. A hybrid semisupervised learning technique was introduced using the Active smart vector learning machine (ASVM) and Fuzzy C-Means (FCM) in the design of an intrusion detection system that has an excellent performance. This system is considered as a binary classification and hence works faster than multi-classifiers [24]. Li et al. [27] presented a new hybrid method based on the density peaks clustering and k nearest neighbors in order to increase the accuracy rate that DPNN was used to train, and kNN was used for classification. Finally, the proposed DPNN method has better accuracy than the support vector machine, and there are many other methods in the field of machine learning. Vinayakumar et al. presented a proposed hybrid intrusion detection system (Scale-Hybrid-IDS-Alert Network) based on a high level of a scalable framework on a hardware server that the capability to classify unpredictable cyber-attacks, monitor network, and host-level event. The framework distributed based on a deep learning

model with the DNNs method used for analyzing big data in real-time and optimal network parameters and network typologies for DNNs. Based on the tests obtained, the performance of the DNNs is higher than that of the classical method [46]. El-Sappagh et al. [39] depicted different classification methods of data mining for true detection false alarm and high accuracy. Based on many methods of data mining on KKD CUP99 that disclose all attack classes with high accuracy. In this paper, the best accuracy gain in the multilayer perceptron method by 92%, and the best training time in the rule-based model is 4 seconds. Elmasry et al. [9] proposed a method using an ensemble weighted majority algorithm to increase accuracy a feature selection method to decrease the number of features for detection attacks. This method increase accuracy of detection by 10, false-positive rate, reduces to 0.05%. Zhang et al. [51] proposed class imbalance processing technology for IDS data set, which combines Synthetic Minority Over-Sampling Technique (SMOTE) and undersampling for clustering based on Gaussian Mixture Model (GMM). The advantage of their novel method verified using the UNSW-NB15 and CICIDS2017 data sets. This model shows an effective solution to imbalanced data in an intrusion detection system.

In the 1970s, the necessity for safety procedures is felt more than ever due to the growing speed, ability, and different computers. In 1977 and 1978, the ISO bore a meeting between governments and Electronic Data Processing inspection bodies. The result of that conference was to read a statement on the status of safety, assessment, and management techniques at that time. The US Department of Energy began the very exhaustive analysis of the evaluation at the same time as the security of computer procedures due to worries about the safety of its systems. James P. Anderson was taken out for this research. Anderson delivered the report in 1980. The details suggested the central core of the ideas of intrusion detection [19]. Established on Extreme Learning Machine, an offered system was presented by Raman Singh et al. [42]. This method solves the issue of the neural network in terms of acceleration. This approach was utilized to reduce predicted remembrance and time by creating a network traffic profile, and two alpha and beta profiles were applied. Alpha and beta profiles can decrease the effect of unaligned information. The beta profile can decrease the size of the testing data set. Its elements are maintained in exercise simultaneously, and the alpha profile is utilized to reduce the impact of discovery time. Folino et al. [11] applied an intrusion detection process to grow group precision based on chorus classification. The chorus structure of the NIDS creates identifying sophisticated attacks and alarms potential. The advantages of utilizing this ensemble classifier include decreasing mistake conflict and bias, which is suitable for unstable classification. The proposed system works well to identify hacks and underestimate the alarms but needs development for personal attacks. Aburomman et al. [1] showed a new approach based on the support vector machine, K-Nearest Neighbor, the weighted majority algorithm classifier, and particle algorithm for intrusion detection. Six K-Nearest Neighbor classifiers and six support vector machine classifiers with different values have been used in this method. The Local uni-modal sampling (LUS) algorithm was utilized to choose high-quality parameters. The offered process has operated LUS-WMA, which has better accuracy than a technique that employs the WMA classifier. Regardless, the implementation of WMA isolated is better than the offered procedure. Gautam et al. [14] utilized the algorithm based on information theory and entropy. This algorithm achieved entropy after ranking elements, and type is established on bias and factors. The results show that the proposed algorithm's rate of identification and accuracy is better than the FFRIDD and MLDRM picking algorithms.

A compound semi supervised understanding process was oriented using the ASVM and FCM in organizing an intrusion detection technique with a superb interpretation. This process is believed to be a binary sort and works faster than multi-classifiers [24]. Li et al. [27] proposed a new hybrid process based on the density ridges clustering and k nearest neighbors to expand the precision rate that DPNN was used for training, and KNN was used for sorting. The offered DPNN technique has better accuracy than the support vector device, and many other device learning styles exist. Vijayakumar et al. introduced a hybrid intrusion detection technique (Scale-Hybrid-IDS-Alert Network) established on a highly scalable framework on a hardware server that can manage incidental cyber-attacks and monitor network and host-level affairs. The framework is based on a deep learning approach with the DNNs method used for studying extensive data in real-time and optimal network parameters and web typologies for DNNs. Established on the tests achieved, the implementation of the DNNs is higher than that of the classical method [46]. El-Sappagh et al. [39] defined various data mining category systems for accurately detecting wrong alarms and high precision. Based on many forms of data mining on KKD CUP99 that highly accurately disclose all attack classes. In this article, the best accuracy in the multilayer perceptron process by 92%, and the best workout time in the rule-based model is 4 seconds. Elmasry et al. [9] offered a project utilizing an ensemble weighted plurality algorithm to improve accuracy and a feature selection strategy to decrease the number of segments for detection attacks. This process increases the accuracy of detection by 10, false-positive rate decreases to 0.05%. Zhang et al. [51] suggested class inequality processing technology for the IDS data set, which integrates the Synthetic Minority Over Sampling Technique (SMOTE) and under selection for clustering founded on Gaussian Mixture Model (GMM). The benefit of their novel technique was confirmed using the UNSW-NB15 and CICIDS2017 data sets. This model solves imbalanced data in an intrusion detection procedure.

## III. Methodology

**https://www.researchgate.net/profile/Krishna-Veni-5/publication/348161131_Efficient_feature_selection_and_classification_through_ensemble_method_for_network_intrusion_detection_on_cloud_computing/links/613738942b40ec7d8bed9801/Efficient-feature-selection-and-classification-through-ensemble-method-for-network-intrusion-detection-on-cloud-computing.pdf**

This proposed work is categorized into two phases, the first phase, focus on ensemble-based feature selection to detect and benchmark the key features for intrusion detection. The second phase focuses on an ensemble-based classification model for intrusion detection using a majority voting technique. The methodology utilized in this proposed work by implementing an ensemble technique.

This proposed work is organized into two stages. The first stage focuses on ensemble-based component selection to detect and benchmark the crucial features for intrusion detection. The second stage focuses on an ensemble-based sort model for intrusion detection utilizing a plurality voting method. The methodology applied in this proposed work is by executing an ensemble method.

### 3.1 Ensemble-based feature selection

We proposed an effective feature selection methodology dependent on the ensemble technique. In which, we proposed a univariate ensemble filter feature selection method (UEFFS), It uses five-different filter feature selection techniques to get the optimal feature subsets from the given intrusion datasets. Moreover, we use the combination rule, in which the optimal feature subsets are aggregated to obtain the final feature subsets. The proposed ensemblebased feature selection framework comprises six functional units. That includes (1) Intrusion datasets, (2) univariate ensemble-based filter feature selection (UEFFS), (3) model construction, (4) performance evaluation analysis, (5) comparative analysis, and (6) statistical analysis. These functional units are used to control, manage, and implement the core idea of the proposed feature selection methodology for IDS. Figure 1 illustrated our proposed framework for constructing the UEFFS method for intrusion detection.

We proposed an effective feature selection methodology dependent on the ensemble technique. In which, we proposed a univariate ensemble filter feature selection method (UEFFS), It uses five-different filter feature selection techniques to get the optimal feature subsets from the given intrusion datasets. Moreover, we use the combination rule, in which the optimal feature subsets are aggregated to obtain the final feature subsets. The proposed ensemblebased feature selection framework comprises six functional units. That includes (1) Intrusion datasets, (2) univariate ensemble-based filter feature selection (UEFFS), (3) model construction, (4) performance evaluation analysis, (5) comparative analysis, and (6) statistical analysis. These functional units are used to control, manage, and implement the core idea of the proposed feature selection methodology for IDS. Figure 1 illustrated our proposed framework for constructing the UEFFS method for intrusion detection.

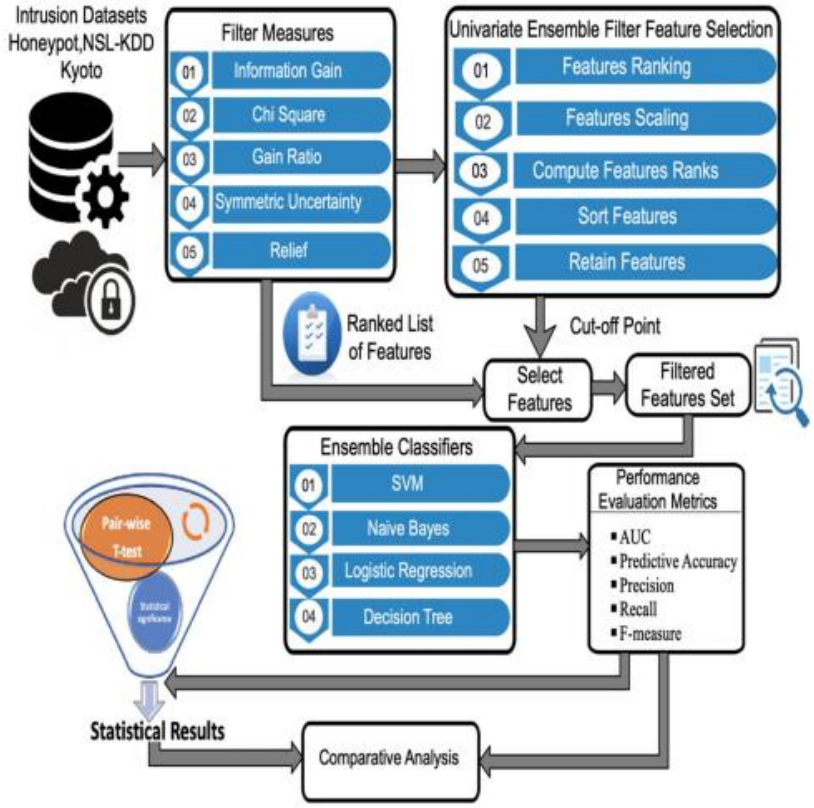


**Fig. 1 Overall framework of the UEFFS method**

### 3.2 Information gain (IG)

Information-Gain (IG) is one of the top most widely used feature selection methods based on mutual information. IG is simple and computationally efficient. It quantifies the data between the jth feature fj and the class labels C, i.e., the amount of information in bits about the class prediction, in the presence of that feature and knowing the corresponding class distribution.

### 3.3 Gain-ratio

The Gain ratio filter feature selection method is measured to be one of the discrepancy measures that provides regularised notch to improve the Info Gain method score. The split information value can be applied to measure the ratio, the formula for split information as follows [21]: SplittinfoAð Þ¼ D Xm j¼1 Dj

---

j j D

---

log2 Dj D ð4Þ where the structure of v partitions represents the Split Info. The gain ratio formula as follows: Gain Ratio ð Þ¼ A InfoGainð Þ A =SplitInfoðÞ ð A 5Þ

### 3.4 Chi squared

The Chi squared FS is the most widespread statistical measure feature selection technique, which processes the relationship between two variables. It may assist to assess the independence of a feature from its class. The Formula for Chi squared FS defined as follows [22]: v2 ¼ X ijOijEij2 .Eij ð6Þ where i and j are two variables and O denotes Observed value and E denotes expected value and v2represents value of Chi squared.524

### 3.7 Proposed UEFFS method

The first step of the proposed algorithm is to compute features from the following three intrusion datasets namely,Honeypot, NSL-KDD, Kyoto. In step two, the five-univariate filter-based measures were used for ranking each feature in an intrusion dataset. This process has been represented in step 2 to step 3 of the algorithm. Then, all the computed ranks were scaled using the first filter measure. The proposed approach

evaluates the significance of the features by their association to the class and classifies independent features corresponding to their degrees of weights. Features with the highest weights or ranks are then selected for inducing classification. To equalize the impact of dissimilar scales, the proposed approach changes the values to the identical scale (i.e., range from 0 and 1). Features with the uppermost weights or ranks are then selected to rank 1, the existing approach has been selected rank 0 to a feature with the uppermost weights or ranks. Following this, the scale ranks are ordered in the ascending order and combine them. Finally, the proposed algorithm calculates a mean to find the ranks and significances of each feature. This process was executed in the following key step repeated for the continuing (n - 1) measures which also represented in step 4 to step 19. After that, rank aggregations were executed in step 25, In this evaluation process, we used an empirically proved existing combined method for combining the five filter measures [24]. Furthermore, the feature score, weight, and priority were computed, as presented in steps 32, 33 and 34.

**4.18 Ensemble Classifier**

Ensemble classifiers stock the projections of numerous base approaches [***]. Broadly practical and theoretical proof shows that model mixture raises predictive precision [76],[77]. Ensemble learners make the base systems in a dependent or independent way. For instance, the bagging algorithm Originated different base models from bootstrap samples of the primary data [78]. On the other hand, boosting algorithms increase an ensemble in a dependent trend. They Repetitively add a base approach that is qualified to ignore the mistakes of the existing ensemble [79]. The literature offers other additions to bagging and boosting [80].

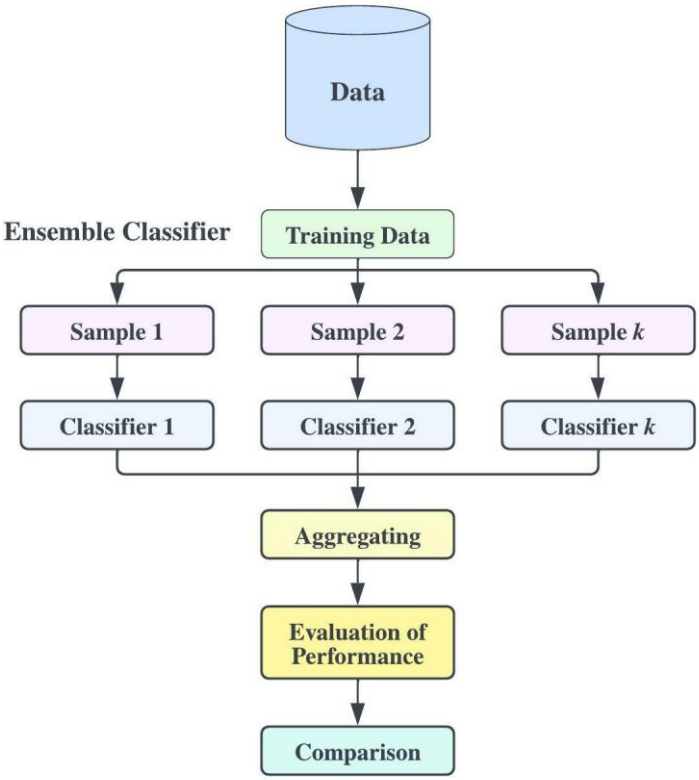


**Figure 3: Workflow of single v. ensemble classifiers [***]**

**5.3 Classification performance results**

The methodology utilized in this proposed work to detect and classify network attacks by applying ensemble voting methods. The proposed ensemble-based predictive model consists of three stages that include: • Collecting representative network traffic data from different intrusion datasets. • By using ensemble-based feature selection methodology for selecting most of the relevant features during the process of building predictive models, in which we used the UEFFS method.

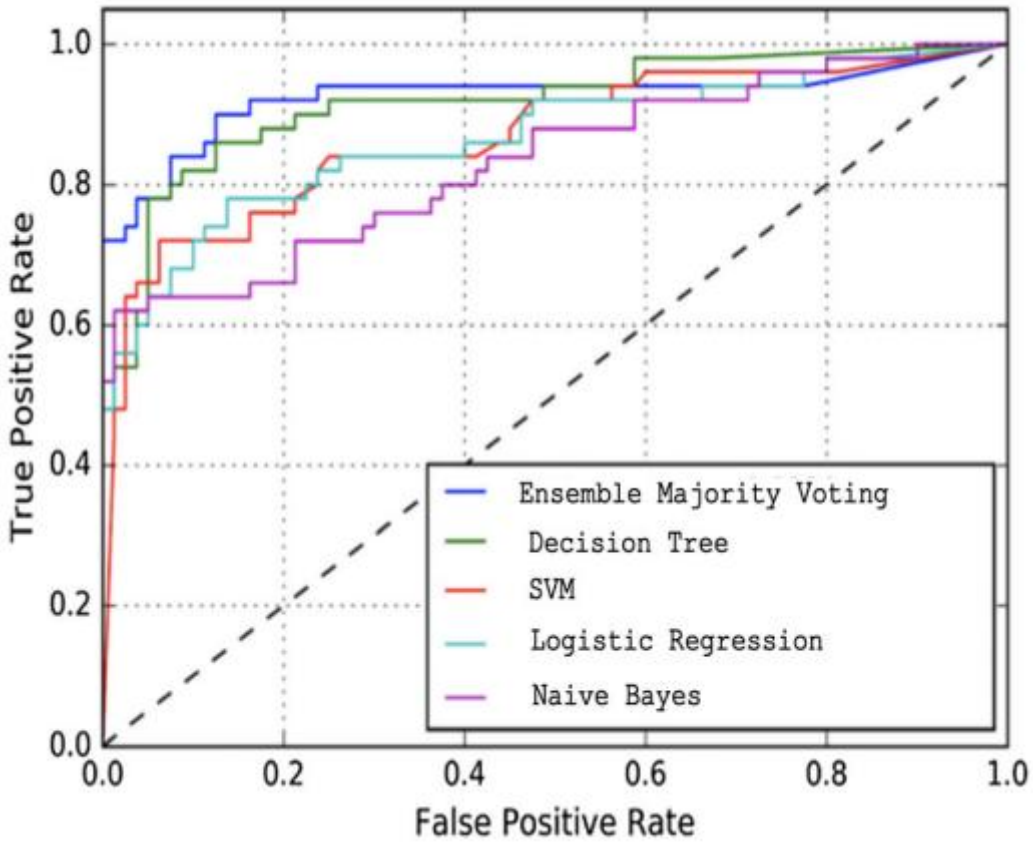


**Figure1: Fig. 1** ROC curve

Construct the predictive model to classify the known threats using ensemble classifiers and evaluate the performance and detection abilities. To build the model, the first step to identify the source of data, which needs to get collected, as well as features from the intrusion datasets. The second step is to select most of the relevant features, in which we used univariate ensemble-based filter feature selection techniques. The third step showcases the existing supervised learning algorithms for classification being used on the intrusion dataset. Then, classify the known threats using ensemble classifiers, evaluate the performance and detection abilities. After building a model, we evaluate its performance on the test data. ROC-Curve plots with two parameters:'' FPR (''False Positive Rate'') and ''TPR (''True Positive Rate''), also, AUC defines how the model is efficient in discriminating between classes and signifies the measures of distinguishable classes. Figure 9 visualizes the ROC curve, which illustrates that our proposed ensemble model achieved a better predictive value than the other four classifiers.

6 Conclusion An efficient intrusion detection system was designed and developed for the cloud environment by using ensemble method feature selection and classification. Our proposed system relies on univariate ensemble feature selection technique, this approach used for the selection of valuable reduced feature set from given intrusion datasets, in which simplicity and speed of five univariate filter methods were used for contributing features towards intrusion detection, while the ensemble classifiers that can competently fuse the single classifiers to produce a robust classifier and which may classify the network attacks. The basic goal of the ensemble method used for increasing the predictive accuracy than any of the single classifiers. In which, we employed a majority voting ensemble technique to discriminate the network traffic as attack data and normal data. Moreover, we evaluated this proposed work on three different intrusion datasets, namely Honeypot real-time datasets-KDD and Kyoto and measures the classification performance by applying ''area under the receiver operating characteristic curves'' AU-ROC metrics, TPR, FPR, precision, recall, DR, FAR across various classifiers to distinguish the network traffic as attack data and normal data. Hence, we proved that this proposed method achieved a strong considerable amount the performance improvement in the accuracy and robustness of various classification tasks, this contributing to FS and key steps in the intrusion detection system. Future works would incorporate multivariate measures to study the irrelevant feature selection. Another Future work to the base models for the ensemble method would be to replace it with a deep neural network model in

the selection process. Acknowledgements This paper and research work behind it would not have been possible without the support of our institution. We thankful to our esteemed institution SRM IST. Many thanks to the editor and reviewers for their concern and valuable comments for improving our manuscript. We also thank the referees for their useful suggestions. The author KS designed and developed the model for feature selection, classification, and developed a real-time environment for data collection using honeypot in the AWS, and also wrote the manuscript. The co-authors SM analysed and interpreted the experiments. S.S modified the final manuscript and answered most queries raised by reviewers. P.S edited the original version of the manuscript.

## References:

[1]. Krishnaveni, S., Sivamohan, S., Sridhar, S.S. *et al.* Efficient feature selection and classification through ensemble method for network intrusion detection on cloud computing. *Cluster Comput* **24**, 1761–1779 (2021). https://doi.org/10.1007/s10586-020-03222-y

[2]. Md Haris Uddin Sharif and Shamim Uddin Ahmed (2022) "Efficient Cyber Intrusion Detection Technique Based On An Ensemble Classifier" Journal of Theoretical and Applied Information Technology 100(16).