SYNERGY ANALYSIS OF ENSEMBLE FEATURE SELECTION ON PERFORMANCE AMELIORATION OF INTRUSION DETECTION SYSTEM

S.Vijayalakshmi and V.Prasanna Venkatesan

Department of Banking Technology, Pondicherry University, Pondicherry, India

ABSTRACT

Unparalleled massive generation of online data by social media platforms, digital banking, networking applications and communication portals have mandated the application of data preprocessing technique in the initial stage for the machine learning models to easily discern patterns/association in the data analysis and classification task. To realize this, effective feature extraction and selection methods have been proposed to simplify the data architecture and relationship between them. This underpins the need for implementing Feature selection in the initial stages of the machine learning pipeline where the decent representation of data becomes available to describe the problem more effectively and clearly. The pruned data generated by these techniques is aimed at effective and timely analysis of the organizational information to decipher any impending threats on the flow of network packets. Collective decisions generated from multiple feature selection techniques surpass the results generated by single feature selection method. This collective ensemble strategy applied in feature selection techniques helps in ameliorating the performance of intrusion detection system inducted in the organizational network. The employment of ensemble design in the feature selection methods holistically improves the IDS performance by enhancing classification efficiency, robustness, stability in accentuating the association between the feature sets with the attack signature (Attack class-oriented feature subset mapping) even when there is disturbance/distortion in the training dataset. This paper thoroughly analyses the efficacy of improving the IDS performance through application of ensemble architecture to feature selection techniques empowered with adoption of DESIRE (Diversity, Equity, Scalability, Inclusivity, Reproducibility (stability) and Enhance Performance) characteristics as highlighted in respective Graphs using NSL-KDD dataset. The diversity generating mechanism instituted in ensemble architecture through data perturbation, function perturbation and hybrid perturbation strategies promises comprehensive coverage of the training set by incorporating cross validation strategies and random sampling techniques

KEYWORDS

Ensemble Feature Selection, Intrusion, Diversity, Equity, Scalability, Inclusivity, Reproducibility (Sensitivity), Performance, Classification Efficiency.

1. INTRODUCTION

Data analysis and preprocessing become a mandatory step in this big data era to arrive at a better understanding of the characteristics of the data and relationship existing between them. This would enable the classification/detection model to easily interpret the underlying fabric of the data composition and discern insightful patterns (attack) in a faster manner [1] [2]. Organizations and business corporates are heavily bombarded with network traffic data comprising both good and bad elements. It becomes imperative for any network security engineer to build a detection system that corners and isolate the infiltrators (intruders) from continuous perpetual of the threat to the entire vicinity. To improve the efficacy of the intrusion detection system (IDS) the dimensionality of the data has to be curtailed to a great extent with the application of feature

selection mechanism where informative/discriminative subset of features are chosen to identify and map attack classes [3]. This summarized feature set has the inherent ability to amplify the classification performance of IDS model and accelerate the detection process. This feature selection (FS) mechanism is also endowed with reduced training time, less memory and storage requirements, high detection accuracy, high computational efficiency etc [4] [5].

High computational cost incurred by intense features would necessitate the intelligent pruning of noisy and outlier data that aids in enhancing the accuracy and preciseness of IDS. A single feature selection strategy would not suffice to address the challenges posed by diverse datasets, as network activity is constantly under flux and the attack landscape is in ad Infinitum mode [6]. A precise and definite representation of feature subset is inadequate to comprehensively recognize the various attacks in the dataset. To introduce the diversity in the selected feature subset and to encompass the features across the breadth and depth of the entire dataset Ensemble Feature Selection (EFS) is recommended. The collective intelligence/decision from multiple base feature learners/selectors aims at reaching out to stable, robust and consistent feature subset [7].

Different base learners have the capacity to imbibe the diverse dataset composition and characteristics, leading to all- encompassing feature subsets. The ensemble architecture incorporated into feature selection is eventually helping to achieve higher(better) detection accuracy of the IDS model [8]. The other benefits accrued through this EFS are to escape from local optima, reduce bias and overfitting, improve generalizability, and have high interpretability. The consistent, robust, efficient, and stable feature subsets extracted using this ensemble paradigm promote performance escalation, high detection accuracy, and reduced friction/tension in the model to produce results in lesser computational time [9].

High variance in ensemble attributes (output) promotes an inclusive culture of accommodating minor classes equally with major classes. Homogeneous and heterogeneous ensemble approach encompassing the data perturbation and function perturbation strategies facilitates the incorporation of changes that can be applied at different levels/strata in Ensemble architecture, viz, dataset level, feature level, base learner method level, combination level, threshold values [10].

The stable and consistent feature subsets generated from this ensemble architecture are aggregated to yield a unique feature subset that strengthens the prediction/induction engine to effectively discriminate the network traffic into genuine and spurious packets. IDS is a network monitoring tool/system manifest in securing the organizational network premises from the clutches of the intruders attempting to thwart the ongoing normal communication in the vicinity [11]. This detection system should be intelligent enough to identify both the zero-day threat and the usual/renowned attack class equally well, for which signature-based, misuse, and anomaly-based detection approaches are endorsed. Application of the EFS paradigm/architecture on the network traffic dataset helps to roll out compact, precise, diverse, stable, robust, and optimal feature subsets that leverage the classification potential of IDS, yielding higher detection accuracy and classification efficiency with a lower false alarm rate. This method also ensures a lower error rate, misclassification, and operational cost of the IDS [12]. This paper has the strength to address pertinent research issues such as

- Does the deployment of ensemble design/architecture in feature selection help to ameliorate the classification efficiency of Intrusion detection system.
- How this feature selector ensemble design aids in comprehensive boosting of diverse traits inherent to IDS meant for encountering modern zero-day threats emanating from the Ultra High Dimensional Database backed with online feature selection for real time processing/scenarios.

The rest of the paper is organized as follows. Section 2 elaborates on the finer details of Feature Selection (FS), Ensemble Feature Selection (EFS), Intrusion Detection System (IDS) and its characteristics. Section 3 discusses the impact of ensemble design on improving the classification efficiency of IDS supplemented with the DESIRE property. Section 4 accentuates the trade-off analysis with and without ensemble architecture integrated with naïve feature selection. Section 5 concludes the paper.

2. BACKGROUND STUDY

2.1. Feature Selection (FS)

Machine learning models confronted with high dimensional data suffers serious performance setback/degradation because of data complexity and quantity culminating to complex models consuming more computational resources and time. It becomes imperative to perform data preprocessing either as feature reduction/selection and extraction to build an efficient model with high discerning ability and improved classification efficiency [13]. The model is as good as the features fed into it. FS is defined as the means to mine distinct, representative and determinant variables from the dataset that has a causal relationship with the target variable ie the attack class (Denial of Service (DoS), Probe, User to Root (U2R), Remote to Local (R2L), Normal) [14].

FS is performed to reduce bias and overfitting issues that plague the classification model, improve generalizability and interpretability of the model (better understanding of underlying fabric ie data composition and interaction), reduce model execution time, decreased training time (time to build the model), reduced inference error and optimum consumption of computational resources [15]. A good selection of the features can influence the learning algorithms to constructively improve the learning speed, generalization capacity, simplicity of the model and a flawless data visualization. Feature subsets with minimal cardinality pave the way for decreased measurement cost with respect to feature collection and misclassification of attack labels leading to a better understanding of the domain. Features can be removed without performance deterioration and no information loss [16]. A good feature selection algorithm should entail an optimal balance between predictive performance and stability.

The filter method is based on assigning an importance/relevance score to each feature in the dataset based on the intensity of inclination towards the target class variable**.** This is purely based on the intrinsic characteristics (information) of the data viz. consistency, distance, similarity etc. This is fast, simple, computationally efficient and independent of any classification algorithm but falls short of capturing feature interdependencies [20]. This is deemed to be a correct fit for high dimensional dataset as it can enable quick processing/trimming of data with reduced fuzziness and vagueness. The wrapper method's feature search strategy is influenced by the outcome of the classification algorithm that determines the efficacy of the selected feature subset in amplifying the classification potential [21]. Several rounds of repeated iteration and evaluation of generated feature subsets may increase the computational burden but at the cost of deriving potential, optimal feature subsets highlighting feature dependencies with improved prediction accuracy. The embedded method tries to construct an optimal feature subset as part of the classification process (learning) and is more computational efficient than wrapper but less than filter method [22]. The hybrid method spawns an effective feature subset that judiciously uses the logic of different feature selection techniques in a constructive manner.

Summarized benefits of FS are depicted in Fig.1. In semi-supervised feature selection, the dataset includes labelled and unlabeled data. These algorithms are time consuming and are not advisable for analyzing high dimensional data as they construct graphs in their processing in the form of

similarity matrix such as the Laplacian matrix and these methods are oblivious to the correlation between features in the feature selection process [19]. Supervised methods are effective at resolving classification problems by establishing the association between features and the class label.

Feature Selection (FS) based on supervision strategy	Unsupervised Supervised Semi supervised
Feature selection based on evaluation function	Filter Wrapper Hybrid
Feature selection from a data perspective	Conventional data based Structured data based FS with heterogenous data FS on data streams
Feature selection based on search strategy	Exponential Sequential Metaheuristic Based

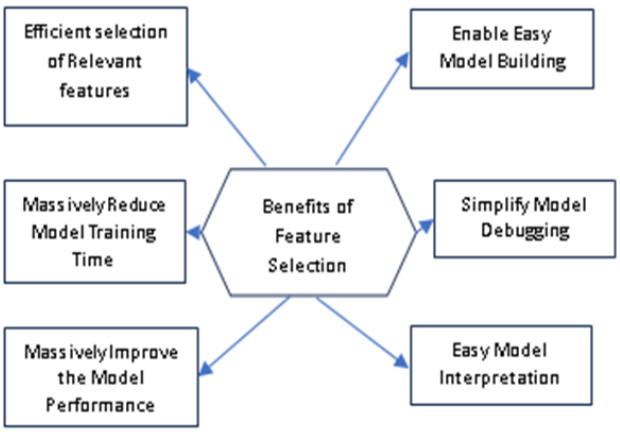
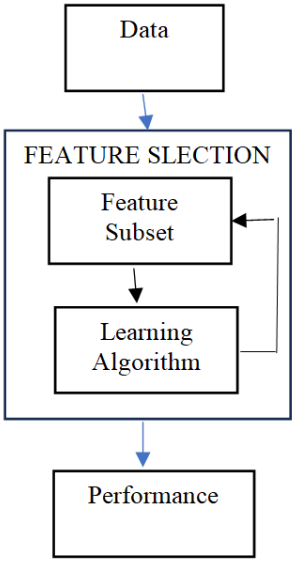
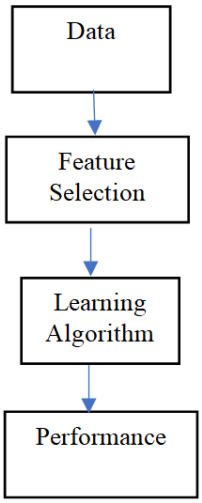


Figure 1. Benefits of Feature Selection

The filter method is based on assigning an importance/relevance score to each feature in the dataset based on the intensity of inclination towards the target class variable**.** This is purely based on the intrinsic characteristics (information) of the data viz. consistency, distance, similarity etc. This is fast, simple, computationally efficient and independent of any classification algorithm but falls short of capturing feature interdependencies [20]. This is deemed to be a correct fit for high dimensional datasets as it can enable quick processing/trimming of data with reduced fuzziness and vagueness. The wrapper method's feature search strategy is influenced by the outcome of the classification algorithm that determines the efficacy of the selected feature subset in amplifying the classification potential [21]. Several rounds of repeated iteration and evaluation of generated feature subsets may increase the computational burden but at the cost of deriving potential, optimal feature subsets highlighting feature dependencies with improved prediction accuracy. The embedded method tries to construct an optimal feature subset as part of the classification process (learning) and is more computationally efficient than the wrapper but less than the filter method [22]. The hybrid method spawns an effective feature subset that judiciously uses the logic of different feature selection techniques in a constructive manner.





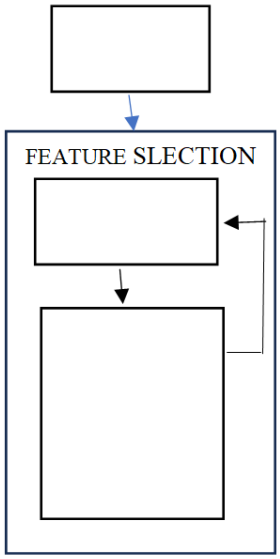


Figure 2. Wrapper FS Method Figure 3. Filter FS Method Figure 4. Embedded FS Method.

Figures 2,3,4 depicts the Wrapper, Filter and Embedded feature selection methods

2.2. Ensemble Feature Selection (EFS)

A feature selector ensemble is defined as several base feature selectors that are effectively used to orchestrate/construct an optimal and stable feature subset from the partitioned training data chunks fed as input and the individual partial decisions are subsequently fused. The main objective of EFS is to identify the best combination of base feature selectors (algorithms) and then the suitable method to aggregate them for effectively arriving at a potential feature subset [23]. Several underlying reasons validate the selection of an ensemble model over an individual model. The uncertainty in the generalization performance of the classifier can be mitigated by adopting an ensemble model that helps to evade the local optima [24]. A single base feature selector can be trained on different instances/subsets of the dataset and produce different orderings of features that could be collectively taken for analysis.

Secondly the voluminous training data can be decomposed into manageable chunks and each smaller partition can be trained with wisely chosen base feature selectors culminating in the adoption of an integration mechanism for fusing the generated partial outputs. By resampling techniques and distributed replication of instances different subspace/population of the datasets can be generated and trained with base feature selectors in an ensemble [25][26]. However to guarantee the improvement of ensemble over single feature selection there are number of factors that need to be considered such as nature and type of base feature selection algorithm (homogenous/heterogenous), diversity generating mechanism such as data perturbation/function perturbation or a hybrid of these two, combination/aggregation method used to merge/fuse the partial/intermediate feature subsets generated by these base selectors and threshold method applied to retain the top relevant and potential features when the ranker methods are used [27].

Different feature selection algorithms applied to the same dataset may generate different feature subsets based on differing evaluation criteria adopted by each method. These generated partial subsets have the power to be better than the others (in their own sense) but not universally/globally accepted as a super performer to boost prediction accuracy [28]. This

necessitates the platform to combine the partial outcomes from the different feature selection algorithms in an attempt to obtain a more robust and optimal feature subset [29]. Fig. 5 illustrates the Homogenous distributed ensemble method where a similar feature selection method is applied iteratively to the distributed training samples across nodes (Data Perturbation). Fig. 6 highlights the Heterogeneous centralized ensemble method where different feature selection methods are applied to the centralized training data samples (same) (Function Perturbation) [30].

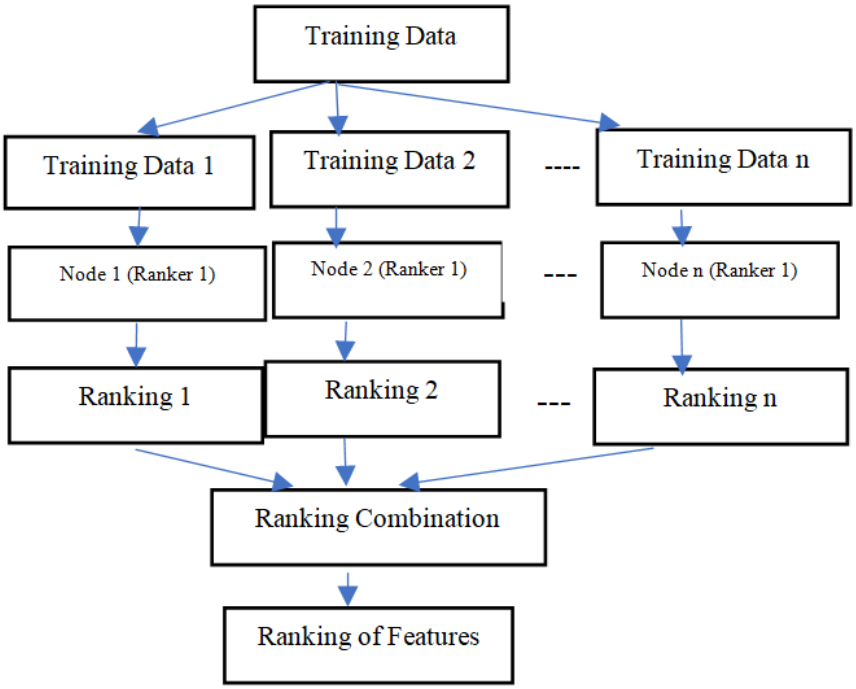


Figure 5. Homogenous Distributed Ensemble Method

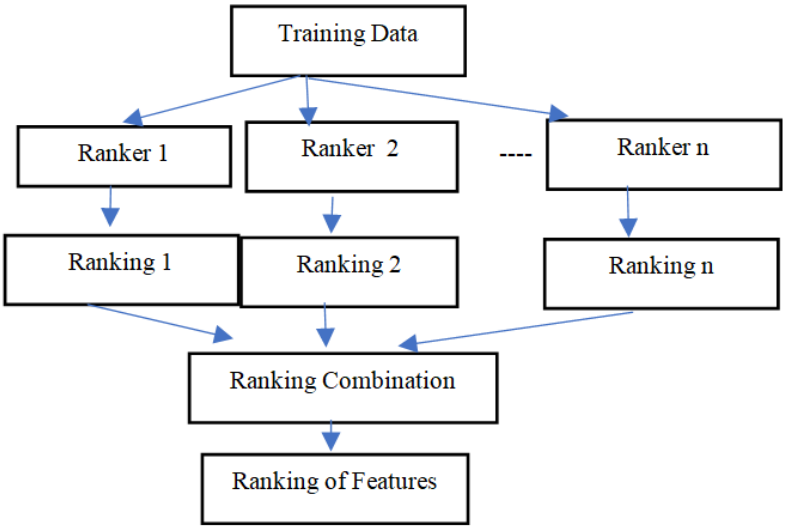


Figure 6. Heterogenous Centralized Ensemble Method

Based on the notion that no best single feature selection technique has been accepted universally so far and the strength of different feature subsets generated from each base feature selector has

the power to discriminate the datasets equally well, ensemble learning has been advocated that combines the advantages of various base learners, overcoming the deficiencies associated with the same [31]. Composing diverse base learners in the ensemble design could guarantee the success of ensemble learning.

EFS consists of two steps viz. in first step, differing base feature selectors are used each generating distinct feature subset that may be distinct label predictions, subset of features or ranking of features and in next step these partial results are integrated and returned as the final ensemble output [32]. Mathematical model for EFS is illustrated by considering a dataset D = (di , . . . , dX), di = (d 1 i , . . . , d Y i) with X instances and Y features. An ensemble of feature selection algorithms (E1, . . . , EN) is applied to D resulting on N feature subsets (F1, . . . , Fn) each one containing S selected features Fn= (fn,1, . . . , Fn,S) [33].

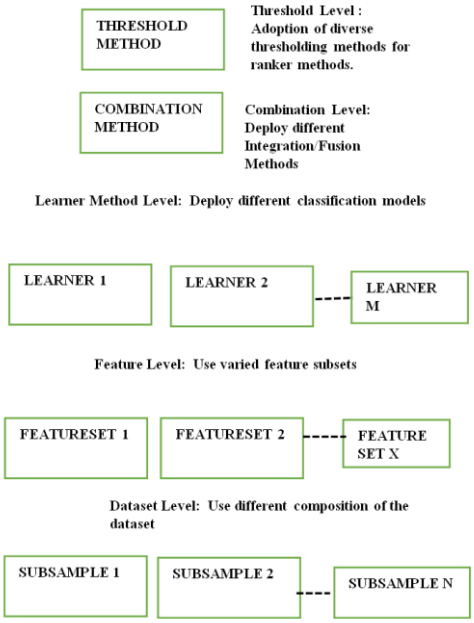


Figure 7. Different Layers in EFS that can be modified.

To design an efficient feature selection ensemble design, several factors have to be considered viz. the cardinality and nature of the individual Feature Selection methods to be used, the size of the different training sets to use, the integration or fusion method to use for amalgamating the partial ensemble feature outputs, the optimal threshold limit and type to be used if the feature selection methods are rankers and the optimal size of the ensemble output [34]. Fig. 7 showcases the different layers in EFS that can be modified/tuned to attain optimal classification performance. The confluence of weak unstable models in an ensemble design might offset the setback/deficiency exhibited in the selection of a single model where the risk of choosing a wrong/flawed base feature selector is annulled [35].

The typical approaches used effectively for ensemble learning are bagging, boosting and stacking particularly adopted for introducing variability and diversity in the ensemble outcomes by sampling the training dataset in such a way that the feature selection algorithm is executed iteratively multiple times over different subsamples [36]. The difference between bagging and boosting ensemble methods rests with the random sampling of the data with and without replacement of the instances. The boosting performs a replacement of the weighted data where the weights assigned are proportional to the misclassification quotient with an intent to mitigate

the classification error. A significant variant of the boosting algorithm viz. The AdaBoost algorithm assigns weight to each data point in the training datasets [37]. A popular ensemble model viz. Random Forest, endorsing the bagging concept where a forest of decision trees is built with each tree representing different random subsets of features [38].

2.3. Intrusion Detection System (IDS)

Organizations have to safeguard/fend off their own environment and resources from the clutches of perpetrators/cyber criminals who incessantly launch sophisticated threats to shackle the prevailing normalcy in the vicinity. This imposes a strict regimen in the form of Intrusion Detection System (IDS) to be institutionalized by the network security administrators to first prevent the network from falling prey to attack (prevention -proactive) and if so, invoke a reactive/recovery measure to counteract the adversary effect and position(transform) the network as a user friendly [39]. IDS is a network security and monitoring tool designed effectively to capture the adversaries wishing to play spoilsport in the ongoing network communication by shackling the triad security pillars viz. Confidentiality, Integrity and Availability.

IDS can be classified based on their deployment location and the principle adopted for detection methods. Network-based IDS and Host based IDS are the two types that result based on their location of operation. Misuse based, Anomaly based and Hybrid based IDS are the constituent members of IDS classification based on the detection methodology adopted. Network based IDS positions itself in the network perimeter and struggles to safeguard the entire network area and the associated resources from the network miscreants who wishes to exploit the innate vulnerabilities and launch sophisticated attacks that catapult the normal network functioning to a standstill mode [40]. Host based IDS attempts to safeguard an individual node using its log report and activity profiling.

Misuse based IDS cracks for the consistency in attack signature similarity with the network security attack (attack signature) database which is continually updated with the current network dynamics [41]. This is bound to yield a high true positive rate with a deficiency in spotting new attack variants or zero-day threats. Anomaly-based IDS attempts to notify any deviation in the node's/network's behavior as against the network behavior profile progressively constructed over a period of time and flags any discrepancy as a security alarm [42]. Hybrid IDS is a mixture of these both Anomaly based and Misuse based IDS. The attack incidence in the network (reactive) requires the admin staff to be on constant vigil in devising meticulous mining algorithms to decipher and corner the attacker at the point of entry (Early detection) causing less network damage. Fig. 8 depicts the comprehensive IDS framework overview where a timely notification alert is raised on the arrival of a spurious intruder. Fig. 9 exemplifies the higher-level processing of an IDS monitoring and scanning [43].

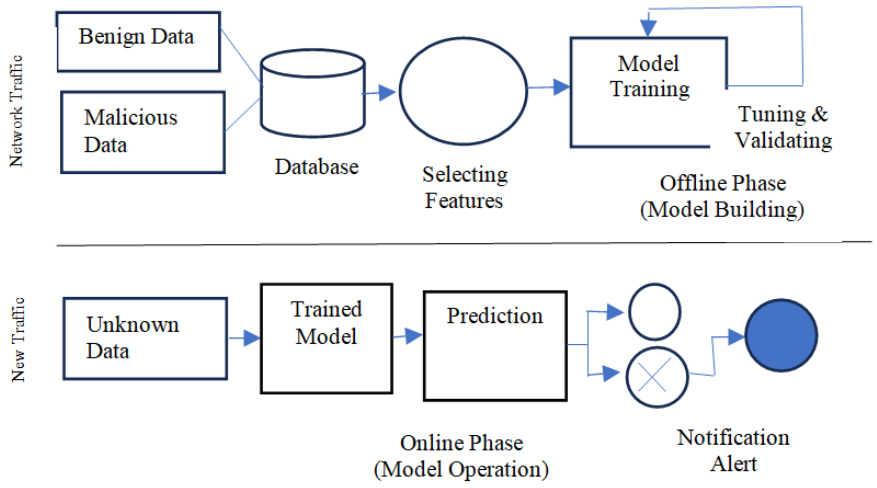


Figure 8. IDS Framework Overview

Mathematical Model for IDS is represented using a specific feature vector y € Y describing basic or specific attributes such as the number of compromised insiders or if a port **scan** was executed. Let I denote a network instance/incident that has been encountered by the organization and it is represented either as I=1 or I=0. A positive intrusion in the vicinity mandates the alarm precursor accordingly i.e. to bear the label as A = 1 or vice versa [44].

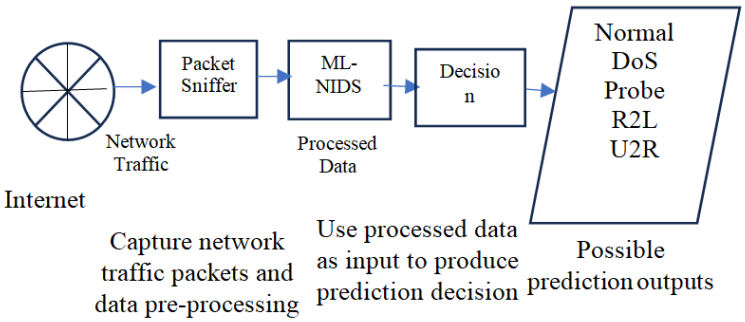


Figure 9. Scheme of Network Intrusion Detection System Deployment

2.3.1. IDS Characteristics

- Less False Positive Rate (FPR)
- Less False Negative Rate (FNR)
- Low computational cost
- Faster
- Scalability
- Accurate results
- High precision and stability
- Speedy detection and early diagnosis
- Computationally efficient
- Reduced Time to detect, Identify and Repair.
- High response time (Reduced execution time)
  High Throughput
- Live/online real time environment/Monitoring

- Lightweight
- Adaptability to evolving threat.
- (Usual/Renowned and Novel attack patterns)
- High detection rate.
- Comprehensive Coverage of diverse attack types.

3. IMPACT OF ENSEMBLE DESIGN ON IMPROVING THE CLASSIFICATION PERFORMANCE OF IDS

Variance preservation in the ensemble design by accommodating diverse base feature selectors helps to achieve improved classification potential with an all-inclusive approach of embracing a 360-degree view of the training dataset. The Diversity, Equity, Scalability, Inclusivity, Reproducibility (stability) and Enhance Performance (DESIRE) characteristics that are inherently endowed with the Ensemble paradigm for Feature selection leverage the potential accrued through successive development and deployment of this model ie Feature selector ensemble.

3.1. Diversity – D

The ensemble design and architecture adopted in feature selection inherently supports the diverse/varied nature that is imperative for any IDS to excel in deciphering unknown and known attacks with ease. A single FS method is not adequate to capture the complete variance/variability that is exhibited by different features as it tends to be biased towards a certain dataset composition eventually leading to overfitting. This may also run the risk of ignoring the other informative features from the dataset that may be pivotal in discriminating the other attack classes definitely. A single Feature subset representation may fail to comprehend well the other looming/impending attacks thus encouraging the adoption of multiple feature selection methods (ensemble) as base learners to yield composite, stable, optimal and robust feature subsets. The collective intelligence/decision/expertise accrued from the diverse base feature selectors aims at easy interpretation of various attacks. This diverse nature also helps to escape from local optima, escape from premature convergence and overfitting. This method also promises to retain useful candidate features by introducing diversity during FS. The weakness/classification error induced by individual feature selection methods is offset with combining the advantages/strengths of these participating feature selection methods.

3.2. Equity – E

The unanimous functioning of the base feature selectors in ensemble design assists in comprehensive coverage of the potential features that help to mine/discern the attacks effectively. This ensemble architecture helps to assign equal weight/proportion to all-encompassing features in a fair way obviating the need for disparagement of other essential features. Ensemble design offers a fair and equal chance for all the features to get selected and helps in detection of minor and major attacks in a class imbalanced scenario. The possibility of minor attacks getting overshadowed by the major attack is ruled out by assigning equitable power to play for all the features.

3.3. Scalability – S

The classification efficiency and detection accuracy of any IDS should not falter when faced with large-scale/high dimensional data. The model should naturally evolve and adapt to ever ever-increasing nature of the dataset in a gradient manner without compromising the performance of the system. Scalability is an intrinsic property of the ensemble design thereby promoting the

adoption of IDS effectively to commensurate well with the scaled-up version of dataset. Bagging and boosting, the two ensemble design paradigms have an innate quality of applying parallelly the common feature selection algorithm in parallel to the diverse bootstrap samples of the training data. Vertical and Horizontal distribution of training samples in the form of bootstrap random samples have supported employing judiciously different feature selection algorithms across all data partitions.

The homogeneous and heterogenous ensemble architecture promises to seamlessly address the scalability issue by deriving independent bootstrap samples from the training dataset and employing the same feature selection algorithm across all partitions unanimously (Data perturbation strategy). Function perturbation strategy promises to deploy different feature selection algorithms to the common(unique) bootstrap sample iteratively yielding feature subsets from each partition (dataset) and subsequently aggregating them to yield a final outcome. It is advisable for the IDS to be efficient enough to be proportionate with the scaled-up version of the dataset through repeated vertical and horizontal partition and eventually merging the partial outcomes from these data chunks.

3.4. Inclusivity – I

To tackle the class imbalanced scenario ensemble design/architecture inherently advocates the inclusivity of all-encompassing features equitably to understand and interpret the underlying attack data patterns and interrelationship among them. No prominent single feature selection algorithm exists to tackle all the available datasets comprising varied attack classes. Certain potential discriminant features have the possibility of getting overridden/overshadowed by the other trivial features that are chosen by other algorithms. In an effort to holistically include features spanning (cutting across) the entire dataset and target/identify usual/unusual attacks at the point of entry itself. Assigning equitable weightage to all features ensures a free and fair chance of it being selected judiciously devoid of bias and effectively cornering the zero-day threats/attacks effectively. This strategy also promotes an inclusive culture of accommodating minor attack classes as equally well as major attack classes.

3.5. Reproducibility (Stability) – R

The domain expert has to gain confidence in the selected feature subset for subsequent validation and analysis despite the changes occurring in the training dataset composition. The FS algorithm should remain insensitive to the minimal changes happening in the dataset and produce a stable and consistent feature subset thus ensuring reproducibility without any alarming discrepancy in the final outcome. The stability of the feature subset has a large telling effect on the classification performance of IDS accentuated with aggregation mechanism imposed for combining partial feature subsets generated either from distinct data partitions or from the individual feature selection algorithm. Stability quotient exhibited by this ensemble model improves the credibility of the selected features and helps to gain confidence in the output. The variation in the composition should not have a high impact on the ensemble outcomes, because these outcomes are subsequently used and validated for classification purposes.

A significant variance in the outcome during several rounds/iteration may lower the confidence level of output features. Recomputation/Reconfiguring of feature subset for every trivial change cause to lose the trust and confidence of network security administrators as the resulting accuracy is at stake. The judicious use of aggregation methods viz. union, combination, intersection, mean, median, Robust Rank aggregation, Enhanced Borda Count, weighted voting, Max voting, Confidence and Conflict measure and SVM_Rank has an enduring impact on the stability and credibility of the selected feature subset.

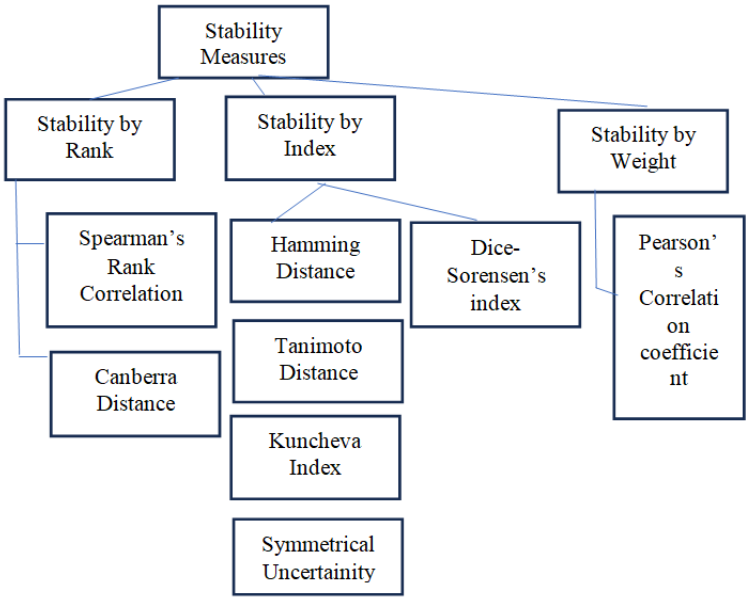


Figure 10. Stability Measures for Features Subset Validation

Jaccard Index and Hamming Distance are the most popular methods used to measure the stability of the selected features that are subset based and the discrepancy in the outcome or consistency miss/loss is attributed to the number of features selected for which the kuncheva consistency index has been proposed. The problem faced by kuncheva stability index demanding the feature subsets to be of similar size is countered by the proposal of new variants of this measure accommodating varying cardinalities. Other popular methods used for computing the similarity among the ordered ranking of features are kendall Tau, Canberra distance and spearman's rank correlation are used.

3.6. Enhances Prediction Performance – E

The ensemble features/output has to be resilient enough to invigorate the classification model to produce predictions with escalated detection rate and minimal false alarm rate. The qualities bestowed in the ensemble attribute/feature subset should offer a competitive edge of enhancing the prediction potential of the IDS model. Guaranteed by the diversity and stability of generated ensemble feature set, there needs to be a check on the quality and relevance of the selected features subset independent of any classifier. The use of an artificial/synthetic data is mandatory to decipher the cluster of pertinent features that are obscured initially in the dataset. The final goal of a feature selection method is to test the efficacy of the selected relevant features on the real dataset.

Various metrics have been proposed to gauge the classification power of the IDS, of which a few have been discussed. Hamming loss measures the extent of misclassification of a feature (selected when it is irrelevant and not selected when it is relevant). The harmonic mean between precision and recall is termed as F1-score and precision is measured by the number of relevant features selected divided by the total number of features selected. Recall is represented as ratio of number of relevant features selected to the total number of relevant features.

An effective solution proposed to transform the ranking of the features returned by the ensemble to a subset of features requires deploying a convenient threshold. To check the authenticity of the ranking of features several techniques exist to ascertain whether the relevant features precede the

irrelevant features. Ranking loss (R) enumerates the number of unconnected features that are better positioned than the relevant features. The minimal set of irrelevant features that are preceding than the relevant ones in the ranking list ensures more scope and space for the latter to exhibit its heightened prominence.

4. TRADE OFF ANALYSIS WITH AND WITHOUT ENSEMBLE DESIGN

Security engineers can make the organization safer through building of effective IDS that adapt to the ever-evolving cyberspace landscape by leveraging the strengths of multiple feature selection techniques. EFS prompts and cues the top priority features for detecting intrusions and this aids the security teams to better perform a decent security profiling for the security analyst to improve their overall security posture. The dynamism instituted in the EFS can empower it to adapt to data distribution changes over time, thereby provisioning the IDS to remain effective against evolving threats and attack vectors. The continuous learning facilitated through synergistic approach established between ensemble approaches with online learning techniques enables the IDS to update its feature selection dynamically as new data arrives incessantly in this big data era.

Ensemble Feature selection helps to amplify the discriminative power of IDS by combining the power of different feature selection methods and generating a robust, stable and optimal feature subset that helps the ML classifier in IDS to better understand the underlying dataset composition and effectively discern the malicious and normal traffic with reduced false alarm rate. The minimal redundancy and maximum relevance to the target class can be achieved with reduced unique and informative features averting the clutter of the dataset leading to confusion in the model. Variety in feature selection techniques helps to develop a more generalized model that performs well across different datasets reducing the risk of overfitting.

EFS aims to eliminate unrelated, noisy and outlier features that confuse the IDS model in arriving at an optimal classification outcome and increased interpretability. This reduction in noise can lead to clearer patterns in the data. A well-chosen subset of features can significantly boost the predictive power of a machine learning model used in IDS culminating in better detection rates for both familiar and unfamiliar attacks. A consistent improvement in precision and recall significantly escalates the F1 score which is vital for evaluating the performance of IDS. EFS leads to better generalization when applied to new, unseen data by reducing the risk of overfitting where the model overlearns the noise and outliers giving less attention to the underlying patterns. EFS incorporates cross-validation techniques ensuring that the nominated features perform well across diverse sections/segment of the training data. A small feature set can lead to simpler and faster models that are easier to interpret and maintain that is beneficial for security analysts. With fewer features to process the training time for ML models can be significantly reduced in real time scenarios.

5. EXPERIMENTAL RESULTS AND DISCUSSION

5.1. Background and Methodology

The effective functioning of IDS is very crucial in latency sensitive environments (network monitoring, scanning and probing) where real time quick affirmative decisions have to be made to keep the adversaries at bay. This solution has to be endured with accelerated feature selection process empowered with minimal time for choosing highly informative representative features and ascertaining the judicious use of these features in both ensemble feature selection and individual/baseline feature selection method. To address these issues, multiple feature selection

methods viz. ReliefF, Mutual Information, Corelation Coefficient, Anova, Chi square methods were employed in both ensemble mode and in solo (individual) mode. The performance metrics derived from this method helped to gauge the effectiveness of the proposed model in establishing improved classification accuracy along with enhanced area coverage in Area Under Precision Recall Curve (AUPRC)

5.2. Precision-Recall Curve Analysis

To further validate the model's performance, Precision-Recall (PR) curves were generated for both the Ensemble Feature Selection (using 5 distinct Feature selection methods) and a Baseline Single Filter Method. The experiments were conducted under identical conditions using the NSL-KDD dataset.

5.2.1. Experimental Setup

Classifier: Random Forest with 10-fold cross-validation
Metrics: Precision, Recall, F1-Score, Area Under PR Curve (AUPRC)
Tools: Scikit-learn, Matplotlib

5.2.2. Observations

The PR curve for the ensemble method showed consistently higher precision across all recall values. AUPRC improved from 0.82 (baseline) to 0.91 (ensemble). The ensemble approach was more robust against class imbalance, especially in detecting minority classes such as U2R and R2L attacks

5.2.3. Interpretation

The higher AUPRC confirms that ensemble-based feature selection enhances not only accuracy but also precision and recall trade-offs—critical in real-time IDS, where false positives and false negatives carry high costs

5.3. Performance Trend with Varying Feature Subsets

A line graph was plotted to compare classification performance (accuracy) with varying numbers of selected features for both Ensemble and Baseline feature selection methods

5.3.1. Experimental Setup

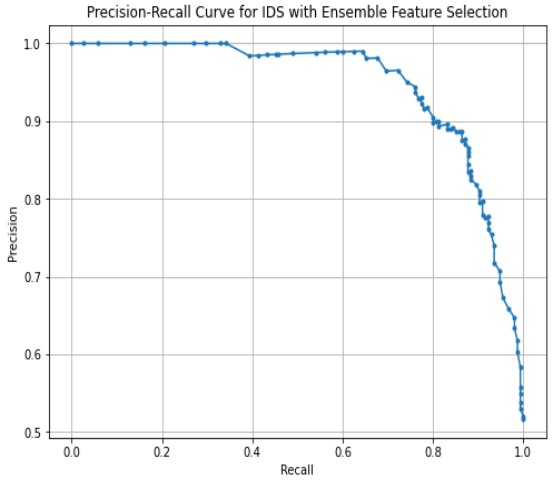
Dataset: NSL-KDD
Classifier: Random Forest
Feature Subset Sizes: 5, 10, 15, 20, 25, 30, 35
Evaluation: 10-fold cross-validation
Metric: Classification Accuracy

5.3.2. Observations

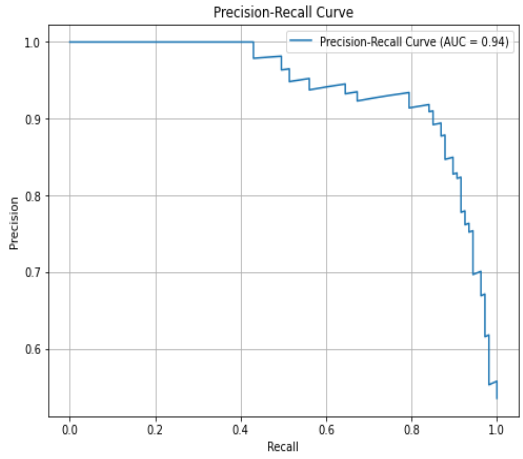
Ensemble Method showed a steep accuracy increase with 5 to 15 features, reaching a plateau beyond 20 features. Baseline Method showed gradual improvement but consistently underperformed the ensemble across all subset sizes. Maximum gain of 6–8% in classification accuracy was observed in the 10–20 feature range when using ensemble selection

5.3.3. Interpretation

The ensemble method identified a compact yet informative feature subset faster than the baseline method with reduced training time and improved generalization. The plateau suggests that additional features beyond a threshold add little or no value and may even introduce noise. The supremacy of the results obtained by deploying ensemble feature selection over the baseline (single) feature selection methods has been depicted using Precision-Recall (PR) curve and a line graph plotted with detection rate (accuracy) against the number of features selected. The area under the curve (AUC) in the PR curve gives an idea about the model's performance. Greater area coverage in the PR curve illustrates better model's performance and a higher precision and recall value ensures fewer false positives and false negatives. The ideal point on the curve is the top right corner where both precision and recall are maximized. Graph 1 and 2 depicts the PR curve for both the Ensemble feature selection and Baseline (Single) Feature Selection Method.



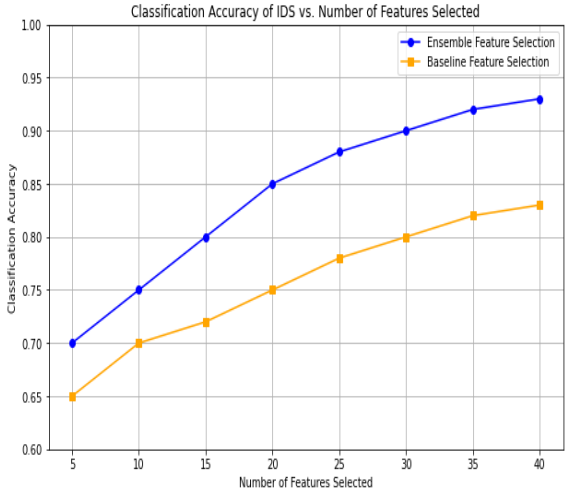
Graph 1: Precision Recall Curve for Ensemble Feature Selection



Graph 2: Precision Recall Curve for Baseline Feature Selection

The line graph demonstrates that with the deployment of ensemble feature selection an optimal number of features is selected that helps to improvise the functioning of IDS effectively over the baseline method where a single feature selection method is used. With the selection of most optimal informative and representative features through EFS aids in enhancing the

classification accuracy of the IDS as shown in the graph than its counterpart. Graph 3 shows the line graph highlighting the performance improvement in Classification of IDS with varying number of optimal features selected for both Ensemble and Baseline feature selection methods.



Graph 3: Classification accuracy of IDS vs Number of features selected for both Ensemble and Baseline Methods

4. CONCLUSION

This paper advocates the application of ensemble architecture in feature selection techniques to enhance the classification potential of IDS with improved data quality and reduced data dimensionality. The need for historical data (background) necessitates the pruning/trimming of the training dataset to contain only principal/informative variables that aid to discern the attack packets in the inward network traffic in to various attack categories/classes. Incorporation of ensemble paradigm complemented with DESIRE (Diversity, Equity, Scalability, Inclusivity, Reproducibility (stability), Enhance Performance) characteristic promotes constructively the selection of informative features from the dataset. This compact representative features spawned from EFS helps to ameliorate the network monitoring efficiency, classification accuracy, robustness and preciseness of the NIDS in efficient assignment of spurious network packets in to respective attack classes. The adequacy and superiority of this ensemble feature are selection technique is analyzed with NSL-KDD dataset creating significant/remarkable results in terms of Accuracy, Precision, Recall and F1 measure as witnessed in the above illustrated graphs. Class specific, Cost-sensitive and Privacy preserving Feature selection augmented with distributed ensemble strategy for multi class IDS forms the scope for foreseeable enhancement.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] Hindy H. Brosset D, Bayne E, Amar Seeam, Christos Tachtatzis, Robert Atkinson, and Xavier Bellekens. 2018. “A Taxonomy and Survey of Intrusion Detection System Design Techniques, Network Threats and Datasets”. 1, 1 (June 2018), 35 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

[2] Torabi M, Udzir N.I, Abdullah M.T, Yaakob R, "A Review on Feature Selection and Ensemble Techniques for Intrusion Detection System", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 5, 2021

[3] Alshamy R, Akcayol M.A, "Intrusion Detection Model using Machine Learning Algorithms on NSl-KDD Dataset", International Journal of Computer Networks & Communications (IJCNC) vol 16, No 6, November 2024, DOI: 10.5121/ijcnc.2024.16605

[4] Umar M.A, Chen Z, Shuaib K, Liu Y, "Effects of feature selection and normalization on network intrusion detection, Data Science and Management, 2024, ISSN 2666-7649, https://doi.org/10.1016/j.dsm.2024.08.001.

[5] Lin Y, Ren X, Wang S, " Ensemble Feature Selection based on Multiple Metrics and Improved Aggregation Strategies", ISCER '24: Proceedings of the 2024 3rd International Symposium on Control Engineering and Robotics, Pages 99 – 103, https://doi.org/10.1145/3679409.3679429

[6] Haro A, Cerruela G, Pedrajasa N, "Ensembles of feature selectors for dealing with class-imbalanced datasets: A proposal and comparative study", Elsevier, Information Sciences, 2020, Spain.

[7] Zhang Y, Zhang H, Zhang B, "An Effective Ensemble Automatic Feature Selection Method for Network Intrusion Detection. Information 2022, 13, 314. https:// doi.org/10.3390/info13070314

[8] Can Q.T, Nguyen T.D, Pham M.B, Nguyen T, AnTran T.H, Dinh T.M, "An Innovative Hybrid Model for Effective DDoS Attack Detection in Software Defined Networks", International Journal of Computer Networks & Communications (IJCNC) vol 16, No 6, November 2024, DOI: 10.5121/ijcnc.2024.16607

[9] Alkasassbeh M, Baddar A.H, "Intrusion Detection Systems: A State-of-the-Art Taxonomy and Survey", *Arab J Sci Eng* **48**, 10021–10064 (2023). https://doi.org/10.1007/s13369-022-07412-1

[10] Albulayhi K, Haija A.A, Alsuhibany S.A, Jillepalli A.A, Ashrafuzzaman M, Sheldon F.T, "IoT Intrusion Detection Using Machine Learning with a Novel High Performing Feature Selection Method". *Appl. Sci.* **2022**, *12*, 5015. https://doi.org/10.3390/app12105015

[11] Haq N F, Onik A R, and Shah F M, "An ensemble framework of anomaly detection using hybridized feature selection approach (HFSA)," *2015 SAI Intelligent Systems Conference (IntelliSys)*, London, UK, 2015, pp. 989-995, doi: 10.1109/IntelliSys.2015.7361264.

[12] A. Khanna et al. (eds.)," Ensemble Feature Selection Method Based on Recently Developed Nature-Inspired Algorithms", International Conference on Innovative Computing and Communications, Advances in Intelligent Systems and Computing 1087, https://doi.org/10.1007/978-981-15-1286-5_39

[13] Chimphlee, W., & Chimphlee, S. (2023). Intrusion Detection System (IDS) Development Using Tree-Based Machine Learning Algorithms. International Journal of Computer Networks & Communications (IJCNC), 15(4). Academy and Industry Research Collaboration Center (AIRCC).

[14] Soheili M, Amir Haeri M.A, "Distributed Ensemble Feature Selection Framework for High-Dimensional and High-Skewed Imbalanced Big Dataset," *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, Orlando, FL, USA, 2021, pp. 1-8, doi: 10.1109/SSCI50451.2021.9659937.

[15] Jaw E, Wang X, "Feature Selection and Ensemble-Based Intrusion Detection System: An Efficient and Comprehensive Approach", *Symmetry* **2021**, *13*, 1764. https://doi.org/10.3390/sym13101764

[16] Songma1 S, Netharn W, Lorpunmanee S, "Extending Network Intrusion Detection with Enhanced Particle Swarm Optimization Techniques", International Journal of Computer Networks & Communications (IJCNC) vol 16, No 4, November 2024, DOI: 10.5121/ijcnc.2024.16404

[17] Ren Y, Zhang L, and Suganthan P. N, "Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]," in *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 41-53, Feb. 2016, doi: 10.1109/MCI.2015.2471235.

[18] Pes B, Dessì N, Angioni M, "Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data, Information Fusion, Volume 35, 2017, Pages 132-147, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2016.10.001.

[19] Brahim A.B, Limam M, "Robust ensemble feature selection for high dimensional data sets," *2013 International Conference on High Performance Computing & Simulation (HPCS)*, Helsinki, Finland, 2013, pp. 151-157, doi: 10.1109/HPCSim.2013.6641406.

[20] Abdullah M, Balamash A, Alshannaq A, Almabdy S, "Enhanced intrusion detection system using feature selection method and ensemble learning algorithms", International Journal of Computer Science and Information Security (IJCSIS), Vol. 16, No. 2, February 2018, pp. 48-55

[21] Preethi D, Khare N, "EFS-LSTM (Ensemble-Based Feature Selection With LSTM) Classifier for Intrusion Detection System", International Journal of e-Collaboration Volume 16 • Issue 4 • October-December 2020

[22] Akhiat Y, Touchanti K, Zinedine A, Chahhou M, "IDS-EFS: Ensemble feature selection-based method for intrusion detection system", Multimedia Tools and Applications (2024) 83:12917–12937 https://doi.org/10.1007/s11042-023-15977-8

[23] Singh K.J, and Tanmay D, "Efficient Classification of DDoS Attacks Using an Ensemble Feature Selection Algorithm", J. Intell. Syst. 2017, https://doi.org/10.1515/jisys-2017-0472 Received September 14, 2017.

[24] Chanu U.S, JohnsonSingh K, Chanu Y.J, "An ensemble method for feature selection and an integrated approach for mitigation of distributed denial of service attacks", Concurrency ComputatPractExper. 2022; JohnWiley&Sons,Ltd., https://doi.org/10.1002/cpe.6919.

[25] Roopak M, Tian G.Y, Chambers J, "Multi-objective-based feature selection for DDoS attack detection in IoT networks", IET, The Institution of Engineering and Technology Netw., 2020, Vol. 9 Iss. 3, pp. 120-127

[26] Abdurohman, M., Prasetiawan, D., & Yulianto, F. A. (2023). DDoS Attacks Detection using Dynamic Entropy in Software-Defined Network Practical Environment. International Journal of Computer Networks & Communications (IJCNC), 15(3), 115–???. https://doi.org/10.21512/comtech.v8i4.3902

[27] Galar M, Fernandez A, Barrenechea E, Bustince H, and Herrera F, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches", IEEE Transactions on Systems, Man, and Cybernetics — Applications and Reviews, VOL.42,NO.4,JULY2012

[28] Vijayalakshmi S, Venkatesan V.P, "A Survey on Application of Metaheuristics Techniques for Ensemble Feature Selection (EFS)", Proceedings of the International Conference on Automation, Computing and Renewable Systems (ICACRS 2022) IEEE Xplore Part Number: CFP22CB5-ART: ISBN: 978-1-6654-6084-2

[29] Bolón-Canedo V, Sánchez-Maroño N, Betanzos A.A, "Recent advances and emerging challenges of feature selection in the context of Big Data", Knowledge-Based Systems (2015), doi: http:// dx.doi.org/10.1016/j.knosys.2015.05.014

[30] Touzene A, Farsi A.A, Zeidi N.A, "High Performance NMF Based Intrusion Detection System for Big Data IoT Traffic", International Journal of Computer Networks & Communications (IJCNC) vol 16, No 2, November 2024, DOI: 10.5121/ijcnc.2024.16203.

[31] Liu Z, et al., "A class-oriented feature selection approach for multi-class imbalanced network traffic datasets based on local and global metrics fusion", Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2015.05.089i.

[32] Wenhao H, Li H, Li J, "Ensemble Feature Selection for Improving Intrusion Detection Classification Accuracy", AICS 2019, Association for Computing Machinery. ACM ISBN 978-1-4503-7150-6/19/07, https://doi.org/10.1145/3349341.3349364.

[33] Yusof A.R, Selamat A, Hamdan H, Abdullah M.T, "Adaptive Feature Selection for Denial of Services (DoS) Attack", 2017 IEEE Conference on Application, Information and Network Security (AINS).

[34] Seijo-Pardo B, Bolón-Canedo V, Alonso-Betanzos A, "Testing Different Ensemble Configurations for Feature Selection", Neural Process Lett DOI 10.1007/s11063-017-9619-1, Springer Science+Business Media New York 2017

[35] Bolón-Canedo V, Alonso-Betanzos A. Ensembles for feature selection: A review and future trends, Elsevier, [J]. Information Fusion, 2019, 52: 1-12.

[36] Canuto A.M.P, Abreu M.C.C, Oliveira L.D, Xavier J.C, Santos A.M, "Investigating the influence of the choice of the ensemble members in accuracy and diversity of selection-based and fusion-based methods for ensembles", Elsevier, Pattern Recognition Letters, 2007, doi:10.1016/j.patrec.2006.09.001

[37] Nguyen, H. S., & Ha, T. D. (2023). A lightweight method for detecting cyber attacks in high-traffic large networks based on clustering techniques. International Journal of Computer Networks & Communications (IJCNC), 15(1).

[38] Fahada C.A, Taria Z, Khalila I, Almalawia A, Zomayab A.Y, "An optimal and stable feature selection approach for traffic classification based on multi-criterion fusion", Elsevier Future Generation Computer Systems 36 (2014) 156–169, http://dx.doi.org/10.1016/j.future.2013.09.015

[39] Wang H, Khoshgoftaar M, Napolitano A, "A Comparative Study of Ensemble Feature Selection Techniques for Software Defect Prediction", 2010 Ninth International Conference on Machine Learning, IEEE Computer Society, DOI 10.1109/ICMLA.2010.27

[40] Seijo-Pardo B, Bolon-Canedo V, Porto-Diaz I, and Alonso-Betanzos A, "Ensemble Feature Selection for Rankings of Features", IWANN 2015, Part II, Springer, LNCS 9095, pp. 29–42, 2015. DOI: 10.1007/978-3-319-19222-2 3

[41] Vijayalakshmi S, Venkatesan V.P, "Ameliorated/Accelerated Intrusion Detection System (AIDS) Using Multiattribute Foveat Analysis with Recurrent Neural Network Augmented by Behavior Pattern Profile (BPP)", Dogo Rangsang Research Journal, UGC Care Group I Journal, ISSN: 2347-7180, Vol-10, Issue-07, No. 16, July 2020.

[42] Huynh, T.T, Nguyen H.T, "Effective Multi-Stage Training Model for Edge Computing Devices in Intrusion Detection", International Journal of Computer Networks & Communications (IJCNC) vol 16, No 1, November 2024, DOI: 10.5121/ijcnc.2024.16102.

[43] Al-Akhras, M., Alawairdhi, M., Alkoudari, A., & Atawneh, S. (2020). Using Machine Learning to Build a Classification Model for IoT Networks to Detect Attack Signatures. International Journal of Computer Networks & Communications (IJCNC), 12(6), 99–116. https://doi.org/10.5121/ijcnc.2020.12607

[44] Tran Hoang Hai, L. H. Hoang, & E. Huh. (2020). Network Anomaly Detection based on Late Fusion of Several Machine Learning Algorithms. International Journal of Computer Networks & Communications (IJCNC), 12(6), https://doi.org/10.5121/ijcnc.2020.12608

AUTHORS

Mrs. S. Vijayalakshmi M.C.A., M.Phil. graduate currently pursuing Ph.D. in Dept. of Banking Technology, Pondicherry University. Her research interest includes Artificial Intelligence, Cyber security, Deep Learning and applications of DL models in security engineering mainly on domains such as Intrusion/Anomaly Detection System. I have 12 years of teaching and research experience and have scholarly publications in international repute conferences and erudite blind peer reviewed journals

Dr. V. Prasanna Venkatesan, Professor, Dept. of Banking Technology, Pondicherry University has research thrust on domains like software architecture, service-oriented architecture, Business Intelligence, Smart Banking, Banking Technology. Eleven scholars have successfully earned their Ph.D degree under his able guidance and support. He has 27 years of teaching and research experience to his credit. He has meticulously completed one project and has 127 publications in peer-reviewed international conferences and journals.