International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 05 July 2023 **Revised:** 25 August 2023 **Accepted:** 15 September 2023*

Intelligent FMI-Reduct Ensemble Frame Work for Network Intrusion Detection System (NIDS)

[*1]Ch. Kodanda Ramu, [2]Dr. T. Srinivasa Rao
[*1]Research Scholar, Department of CSE, GITAM (Deemed to be University),
Visakhapatnam, Andhra Pradesh-530045, India.
*Email: 1260312406@gitam.in
[2]Associate Professor, Department of CSE, GITAM (Deemed to be University),
Visakhapatnam, Andhra Pradesh-530045, India.
Email: sthamada@gitam.edu

Abstract: The era of computer networks and information systems includes finance, transport, medicine, and education contains a lot of sensitive and confidential data. With the amount of confidential and sensitive data running over network applications are growing vulnerable to a variety of cyber threats. The manual monitoring of network connections and malicious activities is extremely difficult, leading to an increasing concern for malicious attacks on network-related systems. Network intrusion is an increasing issue in the virtual realm of the internet and computer networks that could harm the network structure in various ways, such as by altering system configurations and parameters. To address this issue, the creation of an efficient Network Intrusion Detection System (NID) that identifies malicious activities within a network has become necessary. The NID must regularly monitor network activities to detect malicious connections and help secure computer networks. The utilization of ML and mining of data approaches has proven to be beneficial in these types of scenarios. In this article, mutual a data-driven Fuzzy-Rough feature selection technique has been suggested to rank important features for the NIDS model to enforce cyber security attacks. The primary goal of the research is to classify potential attacks in high dimensional scenario, handling redundant and irrelevant features using proposed dimensionality reduction technique by combining Fuzzy and Rough set Theory techniques. The classical anomaly intrusion detection approaches that use individual classifiers Such as SVM, Decision Tree, Naive Bayes, k-Nearest Neighbor, and Multi Layer Perceptron are not enough to increase the effectiveness of detecting modern attacks. Hence, the suggested anomaly-based Network Intrusion Detection System named "FMI-Reduct based Ensemble Classifier" has been tested on highly imbalanced benchmark datasets, NSL_KDD and UNSW_NB15datasets of intrusion.

Keywords: Cyber Threats; Network Breach Detection systems; Dimensionality Reduction; Rough Set Theory; Ensemble Classifier.

I. INTRODUCTION

In the digital age, there has been a rapid expansion of internet and network innovations, which has brought about many changes in data sharing among networks and the ubiquity of sensitive information in various sources. This has attracted attention from academia and industry to develop privacy-preserving techniques [1]. Despite the presence of security methods, the unauthorized access to computer systems for the purpose of acquiring confidential information is becoming more prevalent. Maintaining the security of networks is essential for safeguarding information against unauthorized access and to prevent data breaches. Any activity that violates the accessibility, confidentiality, and reliability of data is considered a security threat [2]. In addition to the existing methods of defense, various Privacy Preserving Data Mining (PPDM) techniques have been researched to protect privacy in data, including anonymization, perturbation, cryptography, and normalization-based methods. However, these mitigation strategies are often complicated, take a long time to implement, and have issues such as excessive generalization and suppression. [3]. on the other side, according to the CISCO report, the usage of network traffic in 2017 was 96EB/month and it is expected to reach 278EB/month in 2022 [4]. This passage is talking about the challenges of securing computer networks and web servers. It mentions that the sheer volume of network connections makes manual monitoring difficult and conventional security methods such as firewalls and cryptography are necessary but not enough to protect against today's attacks. The passage suggests that the use of network applications creates a significant danger to the security of web servers, and highlights the need for a more comprehensive approach to secure these systems [5]. The pre-dominant strategy for observing network systems for information infringement/vindictive movement is the utilization of Intrusion Detection System (IDS) [6]. The practice of cyber security revolves around the protection of networks, computers, and data from unauthorized access. However, increasing volume and sophistication of cyber security threats necessitates the need for developing intelligent techniques which can not only anticipate attacks but also essential to identify new attack types It is crucial to detect new and unusual attacks using an anomaly-based method and to find advanced network security solutions to protect against these

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 05 July 2023 **Revised:** 25 August 2023 **Accepted:** 15 September 2023*

threats. It is important to effectively block any harmful requests on the network system. Identifying malicious traffic from normal traffic on a network is efficiently achieved through the use of (NIDS), which automatically monitor network traffic [7]. The Network Intrusion Detection System (NID) plays a Critical role in ensuring the security of network and information systems. A Network Intrusion Detection System (NID) constantly examines network traffic records in real-time to detect any possible security threats by utilizing machine learning algorithms. The anomaly-based methodology of NID is more efficient in recognizing familiar attacks, but it tends to generate a large number of false positives. In contrast, the signature-based methodology could only identify attacks that have been previously defined in its database [8].

SNORT [9] is another approach used in IDS to analyze network traffic and protocol, but it needs more computational time during detector generations. The difficulty of the limitations of both anomaly-based and signature-based approaches to Network Intrusion Detection has led to a challenge for the research community to develop an intelligent framework for NID. ML techniques have been effectively utilized previously for identifying network intrusions., namely synthesized, virtualized, realistic versions have been used as intrusion datasets, such as DARPA'99 [10], KDD_CUP'99 [11], NSL_KDD [12], UNSW_NB15 [13], TUIDS_ Distributed-Denial-of-Service [14], Network Management Protocol _MIB [15]etc. The usage of machine learning models for system intrusion detection presents complex and varied challenges, as the set of data used to train and test typically consist of vast amounts of data, encompassing a range of security incidents. This provides a challenge for researchers to build effective models. The large number of dimensions in intrusion data sets can lead to a number of difficulties that can negatively impact the performance of many conventional classification methods. Also, the other issues like processing high dimensional dataset poses serious challenge in handling noise and redundant data, which could affect the performance of the constitute classifiers during model training. Hence, feature selection (RST) and feature extraction (PCA) techniques were utilized to identify the relevant attributes that contribute towards identifying an attack. Low-dimensional subspace can reduce the model training time, detection time and increases the detection rate also. Given such challenges, the proposed framework utilized Conditional Mutual Information based on fuzzy membership function to select best ranked attributes and rough set theory clearly outlined in methodology section. The task of network intrusion recognition faces many challenges because of complexity and variability of the data involved. Researchers have proposed different machine learning techniques, including Neural Networks [16], Support Vector Machines [17], Decision Trees [18], k-NN[19],, and Naïve Bayes[20], to develop an effective intrusion detection system. However, conventional classifiers often fail to effectually recognize the minority attack types due to the class imbalance problem, where the amount of normal traffic outweighs that of attack traffic. To address this issue, specialized techniques are required to give due importance to the minority class. To mitigate this problem, a selection feature approach based on Fuzzy Logic, Rough-set theory, and Conditional mutual information (CMI) technique was developed to identify the best ranked attributes in the NSL_KDD and UNSW_NB15 datasets.

The subsequent sections of the paper are arranged in the following manner: Section 2 offers a assessment of previous work in this field. Section 3 includes methodology of the proposed FMI-Reduct ensemble framework. Section 4 includes overview of intrusion datasets used for experimentation. Section 5 analyzes the experimental results obtained on intrusion datasets from IG, PCA, RST, IG-PCA, and proposed method using ensemble of classifiers such as C4.5, SVM and k-NN. Section 6 summarizes the features of the proposed technique and suggests for the extension.

II. RELATED WORKS

Anomaly intrusion detection is the subject of extensive investigation and research area due to its potential in detecting new attacks. Since, the implementation of real-world applications has been hindered due to system-complexity and these systems need a large amount of training, testing, and evaluation prior to the deployment. Running IDS over real labeled network traffic patterns with an extensive and comprehensive set of intrusion behaviors is the most idealistic approach for training and testing. Many researchers implemented NIDS system using machine learning algorithms, that monitors the network connections normally Using the content of payload data and statistical attributes (or features) of network activity, such as the following. Prasad et al [21] employed rough-set theory to rank core features of intrusion datasets based on estimated probability. Instance with single or multiple decisions is distinguished using rough-set approximations (i.e., lower, and upper) and its uncertainty is measured using Bayes theorem. The performance of ranked features reduces computational cost, training time, false rate of alarm, and improved rate of detection can be identified in all relevant classifiers.

Priyadarsini and Anuradha [22] proposed an ensemble model by utilizing two algorithm approaches i.e., Fuzzy-Ensemble Feature Selection (FEFS) and Fusion of Multiple-Classifier (FMC) algorithms. FEFS used to rank best features of KDDCUP'99 intrusion dataset based on Feature Class Distance Function; the results are aggregated by using Fuzzy-Union operation. The experimental results shown in the

proposed system outrages three superior classifiers such as SVM, k-NN and ANNs.

Salo et al [23] presented two-phase anomaly detection architecture by utilizing the concepts of Feature Selection, Feature Extraction and Ensemble Classifier using majority voting scheme. In first phase, to handle unrelated and redundant features in high dimensional or large-scale intrusion datasets ISCX 2012, NSL_KDD, and Kyoto 2006+ features are reduced by using combination of Information-Gain (IG) and Principal Component Analysis (PCA).In second phase, the reduced feature subsets are trained and tested using ensemble classifier such as SVM, k-NN and MLP exhibits prominent performance in-terms of evaluation measures. Comparisons are made with other advanced techniques, evaluating metrics such as precision, Rate of detection, and false rate.

Senthilnayaki et al [24] reduced the dimensions of NSL-KDD intrusion dataset attributes by using the concept of Rough- Set Theory (RST), neighborhood RST and Fuzzy Set Theory (FST) to identify/remove the redundancy among attribute values. The features obtained from proposed feature selection algorithm named Maximum-Dependence Maximum-Significance Algorithm with modified k-NN classifier are evaluated to efficiently discriminate the attack categories such as DOS, Probe, U2R and R2L, the results proved to be robust in improving accuracy and decrease False Alarm Rate effectively.

Hakim et al [25] analyzed the impact of NSL_KDD attributes by choosing various Filter and Wrapper based methods to implement IDS. The experimental results were compared on feature subsets (Gain Ration, Information Gain, Chi-Squared & Relief) with popular classifiers Naive **B**ayes, J48, Random Forest, and KNN show the performance of feature subsets considerably influences the speed of training and testing time in terms milliseconds (ms), significantly affects the accuracy on different classifiers.

Rodda and Erothi [26] proposed a rough-set based ensemble framework based on Quick Reduct algorithm, the proposed algorithm is evaluated on multiple reduct subspaces with homogeneous or heterogeneous containing active/passive classifiers such as C4.5, Naïve Bayes and SVM to train & test KDD'99 and dataset of NSL_KDD and its performance is measured in-terms evaluation measures (accuracy, false alarm rate, precision, detection rate, g-mean and f-score). The experimental outcomes demonstrate that the suggested Rough Set Theory (RST) method delivers a higher rate of detection and a lower rate of false alarms compared to other leading ensemble methods such as the Random Subspace Method (RSM) and Bagging.

Raza and Qamar [27] addressed the limitations of attribute dependency approaches for feature selection; the authors proposed a novel heuristic based dependency-calculation method by utilizing intermediate two-dimensional grid data structure to capture the measures of each individual attribute dependency of unique or non-unique class categories based on the change of attribute values. The proposed method on 6 multivariate datasets evaluated in two phases by calculating the cardinality of equivalence class structure of decision attribute, show the improvement of accuracy by reducing execution time and memory consumption as well attracted over positive region-based method.

Luo et al [28] proposed dominance-based rough-set theory approach to address inconsistencies of multi-class classification problems and change of attribute values at run-time across different-stages of granulations. The experimental results of incremental and non-incremental approaches w. r. t computational time are evaluated on three popular standard datasets such as User Knowledge Modeling, Car Evaluation & Turkiye Student Evaluation; the incremental approach on artificial taxonomy attribute values outperforms non-incremental approach.

Popoola and Adewumi [29] presented a novel framework for IDS using Discretized Differential Equation (DDE) to select optimal feature subset of NSL-KDD intrusion dataset with C4.5 Machine Learning algorithm. The identified DOS attacks such as Neptune, Smurf and Satan reduced misclassification error rate and false positives, increased accuracy, precision and recall measures.

Ambusaidi et al [30] introduced a new feature selection technique, i.e. mutual-information based feature selection approach to tackle both linear and non-linear depended features of KDD'99, NSL_KDD and Kyoto 2006+ intrusion datasets. The intrusion datasets are evaluated on proposed Least Square Support Vector Machine based Intrusion Detection model (LSSVM-IDS), the method show higher accuracy, lower false alarm rate and computational time are compared with other state-of-the-art methods.

Kumar and Batth [31] proposed modified Naïve Bayes classifier to classify normal from abnormal patterns of NSL-KDD intrusion dataset. The problem of bias and over-fitting can be estimated by using three popular feature selection or attribute selection methods such as correlation-based, information gain and gain ratio obtained improved results on proposed classifier in detecting minority class attack categories (U2R & R2L), also reduced false alarm rate when compared to decision tree classifiers i.e., J48 & REPTree.

El-Alfy and Alshammari [32] to reduce model computational time under big flow, the authors presented a scalable learning scheme using genetic algorithm and new approaches to address attribute subset selection using rough-set theory to approximate the core (reduct) which has similar capability of discerning full set of original attributes. The performance of proposed approaches is evaluated on four cyber-security intrusion datasets, namely Spam-base, NSL_KDD, Kyoto 2006+ and CDMC2012, show lower

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *05 July 2023* ***Revised:*** *25 August 2023* ***Accepted:*** *15 September 2023*

running time (in seconds) and the impact of reduct attributes in both Sequential and Genetic Algorithm approaches.

Hasan et al [33] the overall performance of the intrusion detection model was improved by removing irrelevant and repetitive instances/features using RRE-KDD approach to address biased features and records of KDDCUP'99 intrusion dataset. The RRE-KDD method show better accuracy with higher execution time on Random Forest classifier. The selected subset features are evaluated using EM and K-means are compared with varying number of clusters.

Alhaj et al [34] presented a two-tier feature extraction method to select most relevant features of DARPA2000 intrusion dataset. The first phase includes the ranking of features using information-gain entropy in decreasing order. Second phase includes some additional features that are enhancing the discriminative/correlation from the initially ranked attributes to classify class labels.

Elhag et al [35] utilized genetic fuzzy systems to build pair-wise learning framework for intrusion detection. The proposed approach using fuzzy sets and divide-and-conquer learning model achieved a significant improvement in detecting both majority classes (Dos & Probe attacks) and minority classes (U2R & R2L attacks) of KDD_CUP'99 intrusion dataset.

Wang et al [36] employed filter and wrapper-based approaches to reduce features from 41 to 10 in high dimensional network intrusion dataset (KDD_CUP'99 and DDos).The authors utilized Information Gain (Filter method) to find relevance between independent and dependent attributes to obtain important features based on rank. Wrapper based approach use searching method to select optimal feature subset to evaluate C4.5 and Bayesian Network (BN). The accuracy and detection performance thus improved because of a fewer subset of features during the intrusion detection.

Keerthi et al. [37] carried out experiments on benchmark KDD_CUP'99 and UNB_ISCX Datasets related to network intrusions. The authors utilized Feature extraction method using Principal Component Analysis (PCA) for dimensionality reduction. The accuracy obtained from 10 PCA's was compared with classifiers random forest (99.7%) and C 4.5 (98.8%).

Moustafa and Slay [38] examined the feature characteristics of UNSW_NB15 and KDD_CUP'99 intrusion datasets, the author's utilized Association Rule Mining (ARM) algorithm to select relevant features in to measure their efficiency. The experiments were compared with existing classifiers, the outcomes proved the efficiency of proposed technique to assess the complexity in terms of accurateness and False Alarm Rate. The experiments were proved that the original features of KDD_CUP'99 are less efficient compared to attributes of the UNSW_NB15 data set. However, comparing UNSW_NB15 and KDD_CUP'99, the accuracy of KDD'99 dataset is healthier than the dataset of UNSW_NB 15, and the false alarm rate of the KDD'99 dataset is lower than the dataset of UNSW_NB 15.

Pascoal et al [39] presented a novel and automatic anomaly detection learning scheme for selecting relevant features (mutual information metric) and outlier detection (Principal Component Analysis) from the operational point of view under real traffic conditions. The efficacy of the proposed approach is assessed under two real time network scenarios i.e., Business and Residential customer's profiles. The comparative results show six different types of detector engines under different scenarios like with and with not having selection of feature, and outlier detection reduced False Positive Rate.

III. OVERVIEW OF METHODS

Despite research initiatives for the development of ensemble classifier confronted with some major issues include obtaining a quality training data points, selecting diverse sub-spaces, important features, extra computational time for training and aggregating multiple base-classifiers in the field of Intrusion detection that are still remain un-tackled. The state of ensemble art methods in the field of intrusion detection needs quality training data; hence, ensemble-classifiers reduce miss-classification errors and better accuracy in comparison with individual classifiers. Constructing models using a single classifier to handle all types of major security breaches is inadequate and results in low detection rates and high false positive rates. Diversity, in terms of varying errors produced by a group of base classifiers on subsets selected using advanced techniques such as filter and wrapper methods can enhance performance. Various techniques have been proposed to measure diversity in ensemble classifiers [40]. The proposed framework with ensemble classifier for intrusion detection involves multiple classifiers such as C 4.5, SVM and k-NN are utilized to aggregate their outcomes to receive more accurate results. Combining multiple outcomes using majority voting scheme in the ensemble model have benefit to improve detection accuracy and low FAR over a single base classifier.

A. Feature Selection

Handling redundant and irrelevant attributes in high-dimensional intrusion datasets (DARPA'99, KDDCUP'99, NSL_KDD, UNSW-NB15 etc.,) has caused everlasting challenge for network-anomaly detection. Feature subset selection (FSS) and Feature Extraction (FE) is considered as one of the most important data pre-processing step in data-mining/machine-learning areas, especially to build classification model for accurate prediction. Intrusion dataset with FSS and FE techniques helps to provide an optimal informative low-dimensional attribute subspace that are needed to model the network data from distinguishing normal from abnormal/anomaly data. Eliminating such attributes from

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *05 July 2023* ***Revised:*** *25 August 2023* ***Accepted:*** *15 September 2023*

intrusion dataset with spectral information improves the process of classification and it aids traditional classifiers to make precise decisions throughout attack identification time. This results in improving model accuracy and to reduce model complexity/processing time during classifier building phase. The proposed framework for NIDS utilized Fuzzy Mutual Information (FMI) to estimate correlation between dataset features and Rough Set Theory (RST) that analytically selects the optimal attributes for classification, subsequently the selected attributes is used in the training and testing phase.

1) Conditional Mutal Information

Mutual Information (MI) can be employed for Feature Selection (FSS), evaluating the relationship between two variables in terms of target/decision variables to quantify uncertainty for classification tasks. However, MI calculated across the entire dataset may not accurately reflect the correlation between features. To address this limitation, this study first reevaluates MI on identified instances, and then presents a novel FSS method based on Conditional Mutual Information (CMI).MI is mainly used to describe how much information is shared between random variables i.e., MI with higher indicates more relevance between variables, $MI = 0$ indicates totally unrelated with each other. For example, given two random variables X and Y , and there $MI(X;Y)$ is defined as Eq.1.

$$MI(X;Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} p(x_i,y_j)\, ln\left[\frac{p(x_i,y_j)}{p(x_i)p(x_j)}\right] \quad (1)$$

In another way we can find mutual information of two random variables X and Y by using

$$MI(X;Y) = H_e(X) + H_e(Y) - H_e(X,Y)$$

Where $H_e(X), H_e(Y)$ and $H_e(X,Y)$ are given in Eq.2, Eq.3 and Eq.4

$$H_e(X) = -\sum_{i=1}^{n} p(x_i)\ln p(x_i) \quad (2)$$

$$H_e(Y) = -\sum_{i=1}^{n} p(y_i)\ln p(y_i) \quad (3)$$

$$H_e(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{n} p(x_i,y_j)\ln\left[p(x_i,y_j)\right] \quad (4)$$

Similarly, CMI is used to quantify the amount of shared information between two variables X and Y when additional variables are known (Z). Specifically, when the value of a random variable Z is given, CMI is defined as Eq.5

$$CMI(X;Y/Z) = H_e(X/Z) - H_e(X/Y,Z) \quad (5)$$

Where $H_e(X/Z), H_e(X/Y,Z)$ are given in Eq.6, Eq.7

$$H_e(X/Z) = -\sum_{i=1}^{n}\sum_{j=1}^{n} p(x_i,z_j)\ln\left[p(x_i/z_j)\right] \quad (6)$$

$$H_e(X/Y,Z) = -\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} p(x_i,y_j,z_k)\ln\left[p(z_j,x_i/y_j)\right] \quad (7)$$

This definition shows that the (X) variable brings information about the other variable(Y) that is not already present in Z, and the larger the value of $CMI(X;Y/Z)$, the more information it contains. In accordance with the definitions, the following equation applies

$$CMI(X;Y/Z) = MI(X,Z;Y) - MI(Z;Y) \quad (8)$$

In other words, $CMI(X;Y/Z)$ represents the additional amount of information that the variable can provide with respect to the given variables (Z and Y), which is crucial and especially beneficial in the context of feature selection

2) Fuzzy Logic

As we know, the most traditional ML approaches (such as K-NN, SVM, NN, Linear Regression etc.,) can process only numerical inputs. The network packet typically includes heterogeneous data can be noisy, incomplete, redundant, and inconsistent. Therefore, the data pre-processing stage is considered a major challenge in the field of data mining, particularly when dealing with diverse datasets. It is important to transform raw data and features with different scales degrade the classification/model performance. To address these issues, the proposed architecture employs Fuzzy Logic, which is a computing approach based on the degree quantified by Function of membership (MF). The Function of membership is a curve which maps every data point in the space of input to a value membership ranging from 0 to 1. Fuzzy Logic has been proven to be an effective means for decision-making and has been utilized in the design of a fuzzy classification system, let D be a dataset containing attribute set $A = \{a_1, a_2, ..., a_n\}$ with X_i data points are categorized by a function of membership $f_A(X_i)$. An instance X_i transformed to linguistic values (LOW, MEDIUM, and HIGH) as stated in Eq.9 .The function of triangular membership used in this research deal with inexact facts to replaces the crisp boundaries with membership degree functions. The utilization of fuzzy logic in this proposed architecture allows for more flexible and nuanced boundaries to be established for each attribute, as opposed to the traditional crisp boundaries. This makes it particularly useful for datasets with numeric (discrete or continuous) values Using fuzzy logic, a numerical value can be assigned to an attribute to express how much that value belongs to that attribute, instead of having clear cut-off values. This approach allows for more flexible and nuanced classifications. Each attributes then represented as fuzzy form i.e., low, medium and high. This method uses Chandra and Varghese [41] approach to measure the degree of membership of certain value of an entity to a particular class that does not have sharply defined boundaries.

$$f_A(x:\min,\mathrm{mid},\max) = \begin{cases} 0, & x < \min \\ \frac{x-\min}{\alpha_1-\min}, & \min \le x \le \alpha_1 \\ \frac{x-\alpha_1}{\alpha_2-\alpha_1}, & \alpha_1 < x \le \alpha_2 \\ \frac{x-\alpha_2}{\max-\alpha_2}, & \alpha_2 < x \le \max \\ 0, & x > \max \end{cases} \quad (9)$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *05 July 2023* ***Revised:*** *25 August 2023* ***Accepted:*** *15 September 2023*

B. Rough Set Theory

The selected features using Filter Approach i.e., conditional mutual information (MI) leads to high bias towards features with large range of possible values. Hence, FMI subspace or subset (S^{FMI}) will be exposed for further reduction by applying Wrapper Approach i.e., Rough set Theory Pawlak [42] to choose minimal subset called reduct. Typically, a reduced attributes set has the same ability to classify objects as the full set of original attributes.

A decision-system S^{FMI} from the universe U to a given target class (C_k) can be represented as $S^{FMI}=(U,A,V,f)$, where U is universal set or finite set of non-empty objects i.e., $U=\{x_1,x_2,x_3,\ldots,x_n\}$, A is non-empty set of conditional attributes (C) and Decision attribute (D) i.e., $A=\{a_1,a_2,a_3,..,a_n\}$ set of nonempty attribute values represented as $V=\bigcup_{a\in A}V_a$, V_a is called the value set of attribute set A. $\forall a\in A$, there is a corresponding function $f_A:U\to V_a$ .The problem of approximating uncertainty between C and D using information granules are induced by dominance-relation D_R for a given $R\in C$ and $x\in U$ can be defined as Eq.10 and Eq.11.

Set of nonempty objects dominating x, also called as R dominating-set w.r.t x.

$$D^{+}_{R}=\{y\in U: yD_Rx\} \tag{10}$$

Set of nonempty objects dominating x, also called as R dominating-set w.r.t y.

$$D^{-}_{R}(x)=\{y\in U: xD_Ry\} \tag{11}$$

For a given, $Q\in A$, the associated equivalence-relation or indecernability could be demarcated as Eq.12

$$IND(Q)=\{(x,y)\in U\otimes U\mid \forall a\in Q, f_a(x)=f_a(y)\} \tag{12}$$

The partition or subspace of U generated by IND (Q), considered as set of indiscernible-classes, it is denoted as $\frac{U}{Q}$. if $(x,y)\in IND(Q)$ then x, y are in-discernible by attribute from Q ., The equivalence-classes (EC) of the $IND(Q)$ relation are denoted by $[x]_Q$ .Let $R\subseteq U$., then P_{lower} and P_{upper} approximation, then $\left[P_*(X),P^*(X)\right]$ of set X can be represented as Eq.13 and Eq.14:

$$R(X)=\left\{X\in U\mid [x]_Q\subseteq X\right\} \tag{13}$$

$$P^*(X)=\left\{X\in U\mid [x]_Q\cap X\neq\phi\right\} \tag{14}$$

Let $P,Q\subseteq C$, then there is an equivalence-relations over U, then the +ve, -ve and boundary regions with some ambiguity are indicated as

$\left[\mathrm{POS}_P(Q),\mathrm{NEG}_P(Q)\text{ and }\mathrm{BOUD}_P(Q)\right]$ Can be defined as Eq.15, Eq.16 and Eq.17:

$$\mathrm{POS}_P(Q)=\bigcup_{x\in\frac{U}{Q}}R(X) \tag{15}$$

$$\mathrm{NEG}_P(Q)=U-\bigcup_{x\in\frac{U}{Q}}P^*(X) \tag{16}$$

$$\mathrm{BOUD}_P(Q)=\bigcup_{x\in\frac{U}{Q}}P^*(X)-\bigcup_{x\in\frac{U}{Q}}P_*(X) \tag{17}$$

The $\mathrm{POS}_P(Q)$ of the partition $\frac{U}{Q}$ w.r.t P of all objects in U can be certainly classified in the partition $\frac{U}{Q}$ by means of P. Q–depends on P in a degree $"m"(0\le m\le n)$ denoted by $P\underset{m}{\Rightarrow}Q$

$$m=\gamma_{P(Q)}=\frac{|POS_P(Q)|}{|U|} \tag{18}$$

Where P is a set of C, Q is D and $\gamma_{P(Q)}$ is the quality of classification. If $m=1$., Q - depends on P; if $(0\le m\le 1)$., Q– depends partially on P: and if $m=0$ then Q does not depend on P.

The goal of FS is to remove redundant features, so that the reduced subset has similar capability of classifying objects as the complete-set of conditional attribute. The set of all possible-reducts is represented as Eq.19.

$$\mathrm{Reduct.}(S)=\left\{R\subset S\mid \gamma_{R(S)}=\gamma_{S(S)},\forall B\subset R,\gamma_B(D)\neq\gamma_S(D)\right\} \tag{19}$$

A dataset/database may have more than one reducts, the optimal reducts is represented as Eq. 20:

$$\mathrm{Reduct.}(S)_{\min}=\left\{R\in \mathrm{Reduct.}\mid \forall R'\in \mathrm{Reduct.},|R|\le|R'|\right\} \tag{20}$$

C. Proposed Method

The aim of the proposed FMI-Reduct based Ensemble classifier is to provide an effectual network intrusion detection system with low FAR and high DR. The general working of proposed architecture includes following phases: Pre-processing, Feature Selection, Classification & Validation.

Pre-processing step is necessary in our proposed architecture, since intrusion datasets include categorical, continuous and binary values with different scales, which can degrade the performance of individual classifiers involved in intrusion detection. Initially the input dataset (S) with categorical or nominal features are converted to numeric using popular One-Hot-Encoding method available in python libraries. As stated before, scaling attributes (shown in Alg.1) into a normalized range (i.e., -1.0 to 1.0), improve the intrusion detection performance in-terms of all evaluation measures.

Feature selection phase was performed to rank important features of intrusion datasets. The overlapping boundaries of each attribute is analyzed (using Eq.9), map features to {Low, Medium, High}. The mutual information w.r.t decision attributes are then ranked based on correlation among independent attributes. The model trained on selected features show biased outputs in detecting one type of intrusion from other type. Hence, the features selected from FMI stage (shown in Alg.2,Eq.(8)) was proposed to further reduce by applying rough-set theory, a mathematical approach to identify

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *05 July 2023* ***Revised:*** *25 August 2023* ***Accepted:*** *15 September 2023*

indiscernible relations between FMI selected attributes, the upper, lower and boundary approximation as shown in Eq.15 through Eq.17. The proposed system utilized Quick Reduct Algorithm (QRA, shown in Alg.3) to select minimal FMI Subsets are determined by comparing the equivalence relationships produced by a group of ranked attributes referred to as "reduct" (Eq-20).Which has the same ability to classify the full set of objects in the FMI conditional attribute set. The dependency and classification induced by the set of FMI conditional attribute (C) and decision attribute (D) is represented as $Y_c(D)$.

Classification phase (presented in Alg.4), once the reduced dimensionality subspace is obtained from proposed FMI-reduct are then provided as input to the ensemble classifier with the combination of active and lazy base classifiers viz C 4.5, SVM and k-NN. Since, ensemble models are more powerful than individual classifiers to improve the predictive performance in-terms of accuracy. Training network intrusion detection models using heterogeneous or homogeneous classifiers on the FMI-reduct of the identical domain creates diversity between the base classifiers. By combining these diverse classifiers with use of a decision system, the result of the ensemble classifier is improved and the overall rate of rate is reduced. Each of the base classifiers is potentially trained on a subspace of the FMI-reduct... Hence, the proposed system utilized ensemble classifiers with the aim to reduce FAR and to enhance the overall NIDS accuracy.

Validation phase (shown in Alg.5), the labels of class of the test set or instances are predicted by combining the decisions made by the constituent base classifiers using a majority voting scheme.

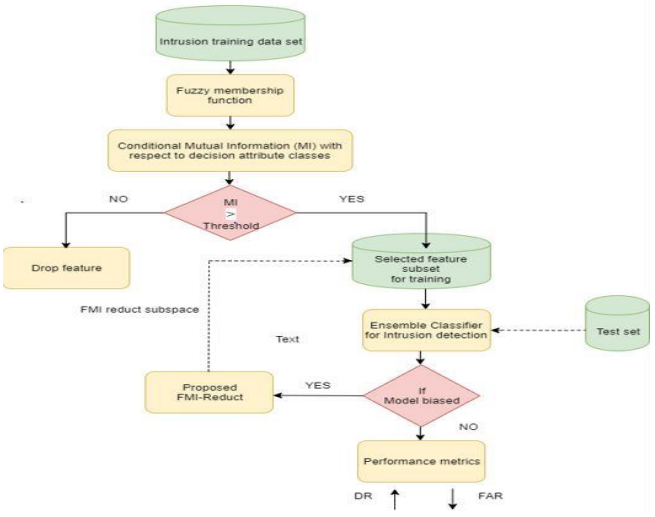

Figure 1: Architecture of FMI-Reduct Ensemble Framework for NIDS

3) Basic Notations

Notation: Input data set $S = \langle U, Q, V, F \rangle$

U_i – Dataset objects from a given input data set $U = \{u_1, u_2, ..., u_n\}$.

Q_i – Set of attributes of a given input data set $Q = \{q_1, q_2, ..., q_n\}$ including conditional Attribute (C) and decision attribute (D). i.e. $Q = (C \cup D)$.

V_{Q_i} – Value of i^{th} attribute in j^{th} position.

$f \rightarrow$ Mapping function, that is $f : U \times V \rightarrow V$

$Q_i^T \rightarrow$ Set of all features transformed from Q_i i.e., in our case (Alg.1)

$Q_i' \rightarrow$ Set of features selected from Q_i i.e., in our case FMI subset (Alg.2)

$Q_i'' \rightarrow$ Set of all features selected from Q_i' i.e., in our case FMI-reduct subset (Alg.3)

$C^b(U) \rightarrow$ Constituent base classifiers used in Alg.4

$y^b(U) \rightarrow$ Ensemble classifier model shown in Alg.4

$y_i \rightarrow$ Predicted class for a given test sample used in Alg.5

4) Algorithms

Algorithm 1: Pre-processing phase
1. Read original dataset U with U_i objects 2. Replacing nominal features Q_i using One-Hot-Encoder 3. Procedure (U) for each Q_i in Q for each j in Q_{ij} $Q_{ij} = \left[u_j * Q_{ij} + c + \mathrm{E}_{err} \right]$ end for end for 4. return $\left(Q_i^T \right)$

Algorithm 2: Fuzzy Mutual Information (FMI)
1. Input data set U with Q_i^T 2. Procedure FMI (U) for each Q_i in Q_i^T Calculate probability correlation matrix using fuzzy values {low, medium, high} w.r.t to D values, show in Eq.(9) Calculate CMI using fuzzy values show in Eq.(8) Append to rank list, select attributes with high correlation (Q_i') end for return (Q_i')

Algorithm 3: Quick Reduct Algorithm for FMI-Reduct Subset
1. Input dataset U with Q_i'

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *05 July 2023* ***Revised:*** *25 August 2023* ***Accepted:*** *15 September 2023*

2. Procedure QRA (U)

$R \leftarrow \{\phi\}$

while $Y_R(\mathrm{D}) \neq Y_C(\mathrm{D})$

$S_i \leftarrow R$

$\forall u \in (C - R)$

If $Y_{R\cup\{u\}}(D) > Y_S(D)$

$S_i \leftarrow R \cup \{U\}$

end if

$R \leftarrow S_i$

end while

3. reduct $= \bigcap_{i=1}^{n} S_i$

4. return (reduct)

Algorithm 4: Ensemble Classification Model

1. Input dataset U with FMI- reduct subspace containg U_i training objects and Q_i features.

2. Construct ensemble classifier model $y^b(U)$ for U

3. for each $b_i = 1:$ to n

$y^b(U_i) = c^b(U_i)$ // training objects

end for

4. return $\left(y^b(U_i)\right)$

Algorithm 5: Validation/Testing phase

1. for each U_i in U

$y_i = y^b(U_i)$ // involves ensemble model to classify U_i

end for

2. $y_i = \mathrm{argmax}_{y\in\{c_1,c_2,...,c_n\}}\left\{c^b(U) : b = 1,2,...n\right\}$

3. return (y_i) // final prediction

4. Evaluate performance measures

IV. DATASET DESCRIPTION

D. KDD_CUP'99

KDD_CUP'99 [43] is a part of data collected from Defence Advanced Research Projects Agency (DARPA), popularly called as DARPA'98[44] Dataset. It was created by MIT_Lincoln-Laboratory (MLL) to evaluate network systems for intrusion detection (ID). As of today most researchers Considered KDD_CUP'99 dataset for assessment of IDS and popularly used for AB- IDS (anomaly-based detection). This data set includes 4 major categories of attacks such as (DOS), Probe, User to Root (U2R), and Remote to Local (R2L), the features of this dataset is categorised into three different groups basic features, traffic features (containing details of Host and Service) and content Features. The Authors in [44] statistically analysed this dataset, the experimental results showed two major issues in the dataset records, which highly influences the performance of existing Machine learning Classifiers of anomaly-detection approaches. To solve inherent dataset problems, the authors proposed modified data set version (NSL_KDD) contains selected network connections of the whole dataset of KDD'99 discussed in section 5.2. The limitations of this dataset clearly outlined in [44]

E. NSL_KDD

Since 1999, many researchers have utilized the KDD_CUP'99 dataset for misuse-based and anomaly-based detection approaches. The dataset contains approximately 4 million records gathered from the dataset of DARPA'98, which is comprised of 4GB of tcp__dump connection data. The NSL-KDD dataset is a refined version of the KDD'99 intrusion dataset, which eliminates duplicate instances to avoid biased classification results, NSL_KDD [45] has number of versions, which are publicly available, and this passage is discussing a dataset used for evaluating the performance of a system for detecting network attacks. The dataset includes 42 features and records of both normal network traffic patterns and 22 different kinds of attacks. The data records consist of 41 features which are divided into 4 categories and include both numerical and categorical data. The distribution of different types of attacks, such as Dos, probe, unauthorized access from a remote system (U2R), and unauthorized access to a local machine (R2L), is shown in a figure(3) and a table(1).

Table 1: Class wise details of KDD data set Attributes

Attribute column number	Attribute type	Attribute column names
1 to 9	Basic Features (B)	Duration, protocol_type, service, src_byte, dst_bytes, flag, land, wrong_fragment, urgent
10 to 22	Content Features(C)	Hot, num_failed_logins, logged_in, num_compromised, root_shell, su_attempted, num_root, num_file_creations, num_shells, num_access_files, num_outbound_cmds, s_hot_login, s_guest_login
23 to 31	Traffic Features (T)	Count, serror_rate, rerror_rate, same_srv_rate, diff_srv_rate, srv_count, srv_serror_rate, srv_rerror_rate, srv_diff_host_rate
32 to 41	Host Features(H)	dst_host_count, dst_host_srv_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_rerror_rate, dst_host_srv_rerror_rate
42	Class variable	class

Table 2: Summary of Intrusion Dataset Features

Attr. NOs	Features	Description
1-9	Basic Features (B)	Provide statistical information of the individual TCP/IP (packet) connections
10-22	Content Features(C)	Represent the details(like domain knowledge) of the packet.
23-31	Traffic Features (T)	Represent the network traffic information and these features were computed using a time window for every two seconds.
32-41	Host Features(T)	Represent the details of the host, which are designated to assess attacks, and were computed using a window for every time interval for more connections.
42	Class variable	The four classes of attacks are Denial of Service (DoS), Probing, User to Root (U2R), and Remote to Local (R2L) attacks. Each of these attack classes is characterized by a unique set of features and behavior patterns, which makes it important to develop a multi-class intrusion detection system that can accurately detect and classify these different types of attacks: Dos,Probe,U2R & R2L

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *05 July 2023* ***Revised:*** *25 August 2023* ***Accepted:*** *15 September 2023*

Table 3 : Summary of Attacks

SNO	Attack	Description
1	Denial of Service (DOS)	Attacker tries to prevent legitimate user from using a service
2	Probe	Attacker tries to gain information about the target host
3	Remote to Local (R2L)	The attacker in this scenario does not have a legitimate account on the victim machine, but they still attempt to gain unauthorized access
4	User to Root (U2R)	Attacker has local access to victim machine and tries to gain super user privileges.

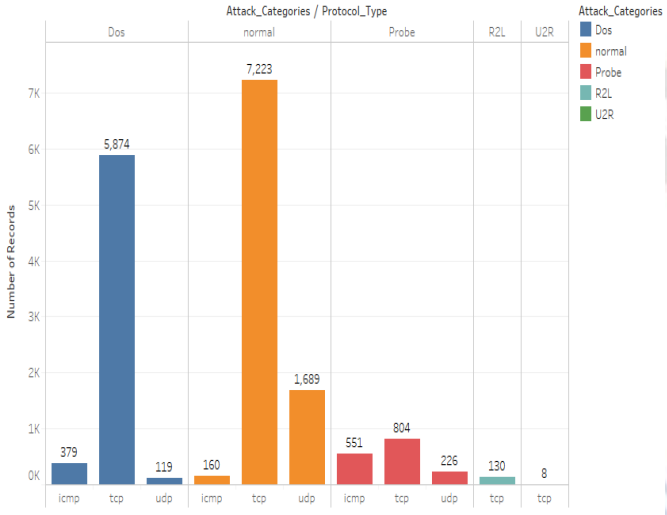

Figure 2 : Summary of Class Distribution in Overall NSL_KDD dataset

F. UNSW- NB15

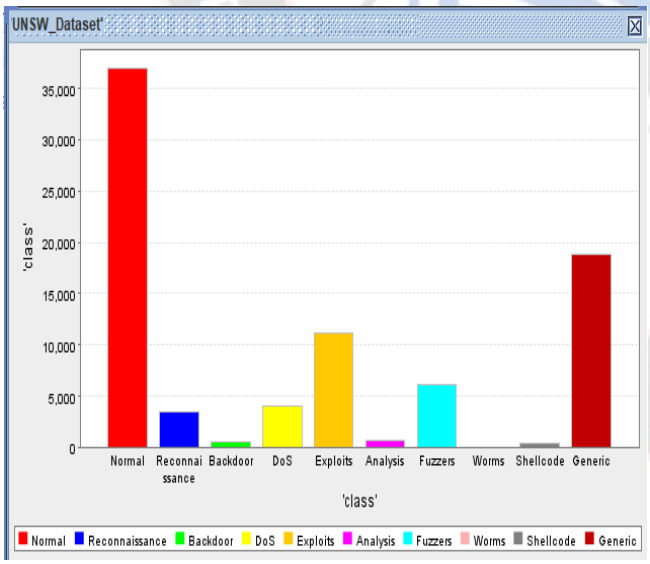

Figure 3: Distribution of UNSW_NB15 Dataset

UNSW_NB15 is a synthetic and complex dataset as it is generated from the modern IXIA test bed configuration tool [46] containing some modern attacks and The NSL-KDD dataset is an enhanced version of the KDD'99 intrusion dataset, developed by the Australian Centre for Cyber Security (ACCS), it can be used for reliable modern IDS evaluation. This dataset was used to create fusion of modern network scenarios (with 9 attacks) updated frequently from CVE dictionary i.e., Well-documented security risks. The IXIA testbed simulations conducted for two scenarios, first one: to generate 1 attack/sec until it should capture 50 GBs, second configuration is to make 10 attacks/sec for another 50 GB spcap files. The Extraction of features from pcap files to csv file includes 49 features with major categories such as flow(5), basic(13),content (8), time(9),connection(7), and additional General(5) features.The description of each feature is provided in Table 4, and the distribution of number of samples of each category is provided in Fig.5.

Table 4: Classwise details of UNSW-NB15 data set attributes

Attribute column number	Attribute type	Attribute column names
1 to 9	Basic Features (B)	Duration, protocol_type, service, src_byte, dst_bytes, flag, land, wrong_fragment, urgent
10 to 22	Content Features(C)	Hot, num_failed_logins, logged_in, num_compromised, root_shell, su_attempted, num_root, num_file_creations, num_shells, num_access_files, num_outbound_cmds, s_hot_login, s_guest_login
23 to 31	Traffic Features (T)	Count, serror_rate, rerror_rate, same_srv_rate, diff_srv_rate, srv_count, srv_serror_rate, srv_rerror_rate, srv_diff_host_rate
32 to 41	Host Features(H)	dst_host_count, dst_host_srv_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_rerror_rate, dst_host_srv_rerror_rate
42	Class variable	class

V. EXPERIMENTAL RESULT

The state-of-the-art and proposed methods were experimented on a PC with Intel Core i5-2400,3.10 GHZ CPU and 4-GB of Random access memory running on 32 bits OS. The effectiveness of the proposed approach is measured on 2 intrusion datasets i.e NSL_KDD and UNSW_NB15. These datasets contain attacks with unequal class distribution and it may degrade the performance of intrusion detection. To mitigate the impact of skewed class distributions and redundancy, the dataset splitted into train (75%) and test (25%), table-7 shows important features selected for proposed framework with Filter (FMI) and Wrapper based (FMI-reduct) methods.

Table 5: FMI-Reduct method feature subsets

Data set	Algorithm	Phase Algorithm	Subset
NSL KDD	Alg-2	FMI	1,5,6,7,9,10,12,13,16,17,22,29, 33,34,40,41
	Alg-3	FMI-Redut	5,6,7,12,13,24,33,40,41

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *05 July 2023* ***Revised:*** *25 August 2023* ***Accepted:*** *15 September 2023*

UNSW-NB15	Alg-2	FMI	6,9,10,11,14,15,16,20,21,23,34,35,36,37,41,42,45
	Alg-3	FMI-Redut	6,1,11,15,16,20,23,37,41,42,45

The reduced dimensions shown in table 5 are sufficient for detecting potential intrusions. The FMI subset for NSL_KDD reduced from 42 to 16 attributes, they are selected based on fuzzy conditional mutual information w.r.t class labels indicating greater than 70% threshold value. The features from FMI subset are further reduced from 16 to 9 based on attribute dependency using classical RST shown in table 7. The QRA (Alg. 3) generated 8 reduct subsets and the final core/ reduct subset is used as an input to the proposed ensemble framework shown in Fig.2, Similarly for UNSW_NB15 dataset.

It is clearly observed from Fig.6, the performance of individual base classifiers viz. C 4.5, SVM, k-NN and proposed ensemble NIDS framework outrages individual classifiers on FMI-reduct subspace for intrusion detection compared to other state-of-the-art feature selection methods. During experiments we have observed, the change in any SVM kernel function for proposed framework obtained higher accuracy than individual classifiers. The accuracy and overall execution time shown in Fig.6 almost acceptable compared to other feature selection methods. The accuracy and execution time results shown for different feature subsets are chosen based on following settings i.e IG subset with top 10 attributes, PCA's with average of top5, RST with single core/reduct subset after performing intersection operation between all generated reducts, IG+PCA subset similar to authors experimented by the author Salo. The base classifiers with different odd (minimum 3) combinations were experimented using existing classifiers such as NaiveBayes, C 4.5, and SVM. k-Nearest Neighbors. But combination with Naïve Bayes classifier degrades the performance of other classifiers in ensemble. Our experiments indicated best results for the combination C 4.5, SVM and k-NN by changing SVM parameters, others with default parameters. The performance of ensemble classifiers with different combinations increased execution time (with full features), reduced accuracy due to bias problem, increased false alarm rate for intrusion detection. To mitigate the above issues, this paper experimented on different existing feature selection algorithms along with proposed feature selection i.e FMI-reduct.

In order to study the impact of ensemble classifiers between majority and minority class detection for considered FMI-reduct are clearly examined using the confusion matrix shown in Table 6.For "normal" connection, it is observed that IG+PCA and FMI-reduct subspaces reduced FAR compared to IG, PCA and RST. The detection of "Dos" (majority) attack is higher in proposed with slight difference compared to RST, whereas similar detection observed in PCA and IG+PCA. The similar behavior observed in the case of "Probe" (majority) and "R2L" (minority) attack. Finally, the proposed ensemble NIDS framework detected minority class i.e. "U2R" on IG+PCA and FMI-reduct subspaces.

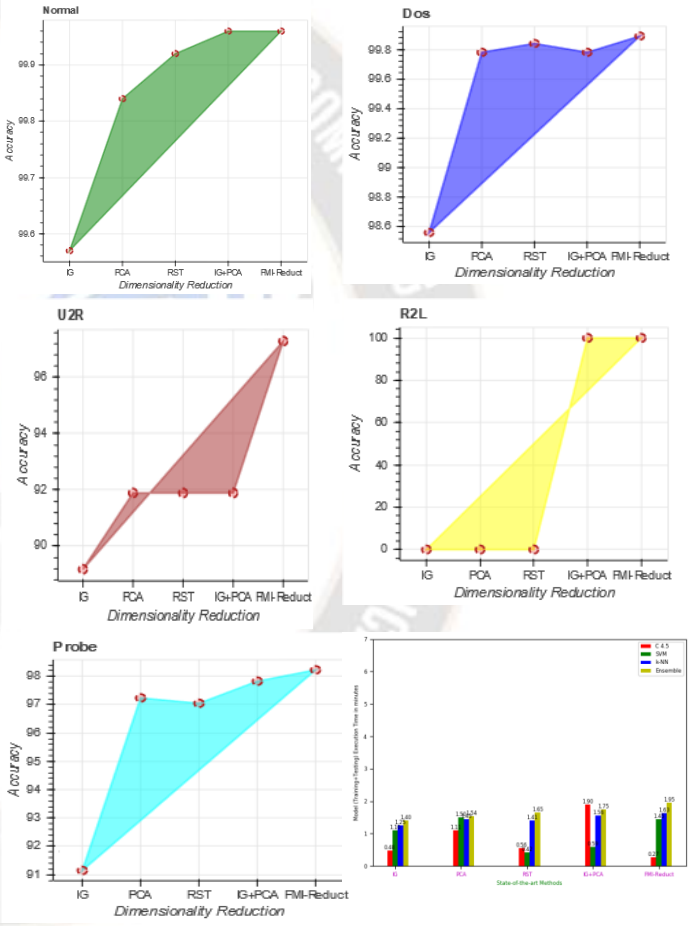

Figure 5: FMI-Reduct Ensemble classifiers Accuracy and Execution Time for NSL_KDD dataset

Table 6: NSL_KDD confusion matrix analysis for FMI-Reduct Ensemble classifier for NIDS

Methods	Act/pre	Normal	DOS	Probe	R2L	U2R	TP	TN	FP	FN	Total
IG	Normal	2557	6	3	2	0	0	2557	11	0	2568
PCA		2564	0	3	1	0	0	2564	4	0	2568
RST		2566	1	0	1	0	0	2566	2	0	2568
IG+PCA		2567	1	0	0	0	0	2567	1	0	2568
Proposed		2767	1	0	0	0	0	2567	1	0	2568
IG	DOS	14	1858	13	0	0	1858	0	13	14	1885
PCA		2	1881	2	0	0	1881	0	2	2	1885

RST		2	1882	0	1	0	1882	0	1	2	1885
IG+PCA		2	1881	1	1	0	1881	0	2	2	1885
Proposed		2	1883	0	0	0	1882	0	1	2	1885
IG	Probe	17	27	464	1	0	464	0	28	17	509
PCA		5	9	495	0	0	495	0	9	5	509
RST		5	4	494	2	4	494	0	10	5	509
IG+PCA		2	6	498	2	1	498	0	9	2	509
Proposed		2	4	500	3	0	500	0	7	2	509
IG	R2L	4	0	0	33	0	33	0	0	4	37
PCA		3	0	0	34	0	34	0	0	3	37
RST		3	0	0	34	0	34	0	0	3	37
IG+PCA		3	0	0	34	0	34	0	0	3	37
Proposed		1	0	0	36	0	36	0	0	1	37
IG	U2R	1	0	0	0	0	0	0	0	1	1
PCA		1	0	0	0	0	0	0	0	1	1
RST		0	0	0	1	0	0	0	1	0	1
IG+PCA		0	0	0	0	1	1	0	0	0	1
Proposed		0	0	0	0	1	1	0	0	0	1

Table 7: Classwise analysis of NSL_KDD dataset

Attacks	Methods	TP	FP	FN	Recall	Precision	F- score	G-mean
Dos	IG	1858	13	14	99.25	99.30	99.27	99.27
	PCA	1881	2	2	99.89	99.89	99.89	99.89
	RST	1882	1	2	99.89	99.94	99.91	99.91
	IG+PCA	1881	2	2	99.89	99.89	99.89	99.89
	Proposed	1883	2	0	100	99.89	99.94	99.94
Probe	IG	464	28	17	95.84	94.30	95.06	95.06
	PCA	495	9	5	99.00	98.21	98.60	98.60
	RST	494	10	5	98.99	98.01	98.49	98.49
	IG+PCA	498	9	2	99.60	98.22	98.90	98.90
	Proposed	500	7	2	99.60	98.61	99.10	99.10
U2R	IG	0	0	1	0	0	0	0
	PCA	0	0	1	0	0	0	0
	RST	0	1	0	0	0	0	0
	IG+PCA	1	0	0	100	100	100	100
	Proposed	1	0	0	100	100	100	100
R2L	IG	33	0	4	89.18	100	94.28	94.43
	PCA	34	0	3	91.89	100	95.77	95.85
	RST	34	0	3	91.89	100	95.77	95.85
	IG+PCA	34	0	3	91.89	100	95.77	95.85
	Proposed	36	0	1	97.29	100	98.62	98.63

The relevance of FMI-reduct ensemble NIDS classifier is evaluated on two extreme measures such as Precision and Recall shown in Table 7. The TP, FP and FN indicates the best result with slight difference with "IG+PCA", zero FN's indicate 100% recall compared to other methods. The recall parameter with 100% indicates the classifiers are best in detecting/ classifying intrusion connections. The precision indicates the total number of positive intrusion connections recognized from actual positives such as Dos, Probe, U2R and R2L. The FP values for IG subset fails to identify actual positive intrusions, increasing FN values for recall measures indicating abnormal connection entered into network systems. Whereas RST, PCA and IG+PCA almost similar to FMI-reduct. The decrease in F-Score and G-mean values indicates the feature set degrades the performance of minority class detection for a given subspace. The Enhancement of F-Score and G-mean for FMI-reduct set indicates that the attributes are good at detecting minority class i.e R2L and U2R, most sensitive to the performance of ensemble classifiers.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *05 July 2023* ***Revised:*** *25 August 2023* ***Accepted:*** *15 September 2023*

Table-8:FAR analysis for NSL_KDD data set

Methods	TN	FP
IG	2557	11
PCA	2564	4
RST	2566	2
IG+PCA	2567	1
Proposed	2567	1

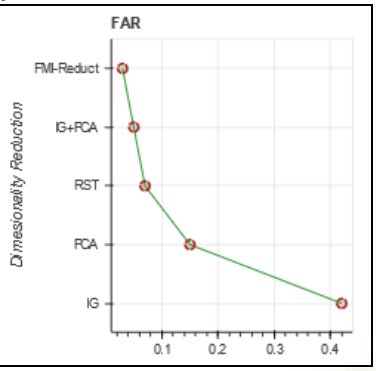

Table 8 present the improvement of reducing FAR showing 0.03, it is clearly observed that IG+PCA and FMI-reduct correctly classified 2571 normal connections out of 2568, whereas IG method leads to high FAR i.e 0.42, PCA and RST obtained 0.15 and 0.07 respectively. It is observed clearly from the figure shown in table 10, showing lesser FAR for proposed FMI-reduct subset.

Accuracy is another important measure used in machine learning, A measurement that indicates the proportion of correct predictions to the total number of instances., in our experiment the datasets include positive (Attacks) and negative (Normal) connections, the proposed ensemble NIDS classifier show highest accuracy compared to all considered state-of-the-art methods for both NSL_KDD and UNSW_NB15 shown in fig 7 and Table 9 respectively.

Table 9: Evaluation measure analysis for UNSW_NB15 dataset using different feature subsets on FMI-Reduct Ensemble framework for NIDS

Methods	ACC	DR	Precision	FAR	F-SCore	G-Mean	AUC
IG	0.7685	0.4102	0.5839	0.1036	0.4446	0.4699	0.6533
PCA	0.7689	0.4125	0.5852	0.1040	0.4461	0.4716	0.6542
RST	0.9329	0.7980	0.7570	0.0508	0.7705	0.7740	0.8736
IG+PCA	0.9519	0.8274	0.7764	0.0406	0.7961	0.7989	0.8934
Proposed	0.9569	0.8483	0.7869	0.0416	0.8133	0.8154	0.9033

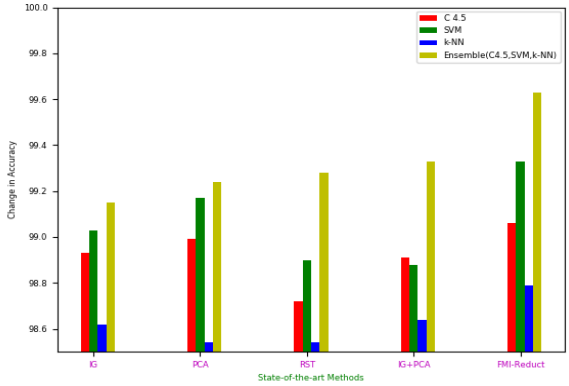

Figure 6: The change in NSL_KDD accuracy measure for different feature subsets on FMI-Reduct with homogeneous and hetrogeneous classifiers

VI. CONCLUSION AND FUTURE WORK

The detailed analysis has been carried-out in the field of IDS (Intrusion Detection) by considering popular benchmark intrusion datasets. The effectiveness of proposed FMI-reduct ensemble NIDS framework have been compared due to the necessity of feature selection or dimensionality reduction methods in recent studies for classification models in intrusion detection. The motive of proposed NIDS framework is to handle major challenges such as to select optimum feature subset, extra computational time for identifying quality training data points using RST, aggregating multiple base-classifiers and reducing mis-classification error, to improve accuracy compared to individual base classifiers is the major task for intrusion detection that are still remain un-tackled. The experimental results shown in Figures and Tables, the proposed NIDS architecture delivers to show best results for detecting intrusions such as Dos, Probe, U2R and R2L attacks of NSL_KDD, similar performance have been found for UNSW_NB15 dataset. However the proposed novel approach continuous to show promising results for intrusion detection, but still its suitability could be tested for other real time intrusion datasets such as TUIDS DDos, SNMP_MIB etc.,by giving due importance to multi-class imbalance should be considered in the future work.

REFERENCES

[1] S. Rodda, U.S. Erothi, "Class imbalance problem in the network intrusion detection systems," In 2016 international conference on electrical, electronics, and optimization techniques (ICEEOT), Ieee, pp. 2685-2688, 2016.

[2] L. Xu, C. Jiang, J. Wang, J. Yuan, Y. Ren, “Information security in big data: privacy and data mining,” Ieee Access, vol.2, 1149-1176.

[3] K. Mandala, K. SobhanBabu, U.S. Rao Erothi, “Least Square Privacy Preserving Technique for Intrusion Detection System,” 2014.

[4] B.B. Prakash, K. Yeswanth, M.S. Srinivas, S. Balaji, Y.C. Sekhar, A.K. Nair, "An Integrated Approach to Network Intrusion Detection and Prevention,” In Inventive Communication and Computational Technologies:

Proceedings of ICICCT, pp. 43-51, 2020. Springer Singapore.

[5] A. Razzaq, A. Hur, H.F. Ahmad, M. Masood, "Cyber security: Threats, reasons, challenges, methodologies and state of the art solutions for industrial applications," In 2013 IEEE Eleventh International Symposium on Autonomous Decentralized Systems (ISADS), pp. 1-6, 2013. IEEE.

[6] A. Shrivastava, M. Baghel, H. Gupta, "A review of intrusion detection technique by soft computing and data mining approach," International Journal of Advanced Computer Research, , vol. 3, no. 3, pp. 224, 2013.

[7] M.A. Khan, Y. Kim, "Deep learning-based hybrid intelligent intrusion detection system," Comput. Mater. Contin, vol. 68, pp. 671-687, 2021.

[8] D.G. Bhatti, P.V. Virparia, "Soft computing-based intrusion detection system with reduced false positive rate," Design and Analysis of Security Protocol for Communication, pp.109-139, 2020.

[9] F. Erlacher, F. Dressler, "On high-speed flow-based intrusion detection using snort-compatible signatures," IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 1, pp.495-506, 2020.

[10] R. Lippmann, J.W. Haines, D.J. Fried, J. Korba, K. Das, "The 1999 DARPA off-line intrusion detection evaluation," Computer networks, vol. 34, no. 4, pp.579-595, 2000.

[11] K.D.D. Cup, "Dataset," available at the following website http://kdd. ics. uci. Edu/databases/kddcup99/kddcup99. Html, vol. 72, pp. 15, 1999.

[12] M. Tavallaee, E. Bagheri, W. Lu, A.A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," In 2009 IEEE symposium on computational intelligence for security and defense applications, IEEE, pp. 1-6, 2009.

[13] N. Moustafa, J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," In 2015 military communications and information systems conference (MilCIS), pp. 1-6, 2015. IEEE.

[14] P. Gogoi, D.K. Bhattacharyya, B. Borah, J.K. Kalita, "MLH-IDS: a multi-level hybrid intrusion detection method", The Computer Journal, vol. 57, no. 4, pp. 602-623, 2014.

[15] A. Manna, M. Alkasassbeh, "Detecting network anomalies using machine learning and SNMP-MIB dataset with IP group," In 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), IEEE, pp. 1-5, 2019.

[16] S. Rodda, "Network intrusion detection systems using neural networks", In Information Systems Design and Intelligent Applications: Proceedings of Fourth International Conference INDIA 2017, Springer Singapore, pp. 903-908, 2018.

[17] S.A. Mulay, P.R. Devale, G.V. Garje, "Intrusion detection system using support vector machine and decision tree," International journal of computer applications, vol. 3, no. 3, pp. 40-43, 2010.

[18] V. Kosamkar, S. Sangita, "Improved Intrusion detection system using C4. 5 decision tree and support vector machine," PhD diss., Doctoral dissertation, Mumbai University, 2013.

[19] W. Li, P. Yi, Y. Wu, L. Pan, J. Li, "A new Intrusion Detection System based on KNN classification algorithm in wireless sensor network," Journal of Electrical and Computer Engineering, http://dx.doi.org/10.1155/2014/240217, pp. 1-7, 2014.

[20] M. Panda, M.R. Patra, "Network Intrusion Detection using Naïve Bayes," International Journal of Computer Science and Network Security, Vol. 7, no.12, pp. 258-263, 2007.

[21] M. Prasad, S. Tripathi, K. Dahal, "An efficient feature selection based Bayesian and Rough set approach for intrusion detection," Applied Soft Computing, vol. 87, pp. 105980, 2020.

[22] P.I. Priyadarsini, G. Anuradha, "A novel ensemble modeling for intrusion detection system," International Journal of Electrical and Computer Engineering, vol. 10, no. 2, pp. 1963, 2020.

[23] F. Salo, A.B. Nassif, A. Essex, "Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection," Computer Networks, vol.148, pp.164-175, 2019.

[24] B. Senthilnayaki, K. Venkatalakshmi, A. Kannan, "Intrusion Detection System using Fuzzy Rough Set Feature Selection and Modified KNN Classifier," International Arab Journal of Information Technology, vol. vol.16, no.4, pp. 746-753, 2019.

[25] L. Hakim, R. Fatma, "Influence Analysis of Feature Selection to Network Intrusion Detection System Performance Using NSL-KDD Dataset," In 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), IEEE. pp. 217-220, 2019.

[26] S. Rodda, U.S. Erothi, "A Roughset Based Ensemble Framework for Network Intrusion Detection System," International Journal of Rough Sets and Data Analysis (IJRSDA), vol. 5, no. 3, pp. 71-88, 2018.

[27] M.S. Raza, U. Qamar, "Feature selection using rough set-based direct dependency calculation by avoiding the positive region," International Journal of Approximate Reasoning, vol. 92, pp. 175-197, 2018.

[28] C. Luo, T. Li, H. Chen, H. Fujita, Z. Yi, "Incremental rough set approach for hierarchical multicriteria classification," Information Sciences, vol. 429, pp. 72-87, 2018.

[29] E. Popoola, A.O. Adewumi, "Efficient Feature Selection Technique for Network Intrusion Detection System Using Discrete Differential Evolution and Decision," IJ Network Securityvol. vol.19, no.5, pp. 660-669, 2017.

[30] M.A. Ambusaidi, X. He, P. Nanda, Z. Tan, "Building an intrusion detection system using a filter-based feature selection algorithm," IEEE transactions on computers, vol. 65, no.10. pp. 2986-2998, 2016.

[31] K. Kumar, J.S. Batth, "Network intrusion detection with feature selection techniques using machine-learning algorithms," International Journal of Computer Applications, vol. 150, no. 12, 2016.

[32] E.S. El-Alfy, M.A. Alshammari, "Towards scalable rough set based attribute subset selection for intrusion detection

using parallel genetic algorithm in MapReduce," Simulation Modelling Practice and Theory, vol. 64, pp. 18-29, 2016.

[33] M.A. Hasan, M. Nasser, S. Ahmad, K.I. Molla, "Feature selection for intrusion detection using random forest," Journal of information security, vol. 7, no. 3, pp. 129-140, 2016.

[34] T.A. Alhaj, M.M. Siraj, A. Zainal, H.T. Elshoush, F. Elhaj, "Feature selection using information gain for improved structural-based alert correlation," Pl oS one vol. 11, no. 11, pp. e0166017, 2016.

[35] S. Elhag, A. Fernández, A. Bawakid, S. Alshomrani, F. Herrera, "On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on intrusion detection systems," Expert Systems with Applications, vol. 42, no. 1, pp. 193-202, 2015.

[36] W. Wang, Y. He, J. Liu, S. Gombault, "Constructing important features from massive network traffic for lightweight intrusion detection," IET Information Security, vol. 9, no. 6, pp. 374-379, 2015.

[37] M.A. Aziz, A.A. Ewees, A.E. Hassanien, "Multi-objective whale optimization algorithm for content-based image retrieval," Multimedia tools and applications, vol. 77, pp. 26135-26172, 2018.

[38] N. Moustafa, J. Slay, "The significant features of the UNSW-NB15 and the KDD99 data sets for network intrusion detection systems," In 2015 4th international workshop on building analysis datasets and gathering experience returns for security (BADGERS),. IEEE, pp. 25-31, 2015.

[39] C. Pascoal, M.R. De Oliveira, R. Valadas, P. Filzmoser, P. Salvador, A. Pacheco, "Robust feature selection and robust PCA for internet traffic anomaly detection," In 2012 Proceedings Ieee Infocom, IEEE, pp. 1755-1763, 2015.

[40] L.I. Kuncheva, C.J.Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," Machine learning, vol. 51, no. 2, pp. 181, 2003.

[41] B. Chandra, P.P. Varghese, "Fuzzifying Gini Index based decision trees", Expert Systems with Applications," vol. 36, no. 4, pp. 8549-8559, 2009.

[42] Z. Pawlak, "Rough sets and intelligent data analysis," Information sciences, vol.147, no. 1-4, pp.1-12, 2002.

[43] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, October 1999.

[44] DARPA 1998.Available on: http://www.ll.mit.edu/r-d/datasets/1998-darpa-intrusion-detection-evaluation-dataset.

[45]NSL-KDD dataset, Available on : https://www.unb.ca/cic/datasets/nsl.html.

[46] S. Meftah, T. Rachidi, N. Assem, "Network based intrusion detection using the UNSW-NB15 dataset," International Journal of Computing and Digital Systems, vol. 8, no.5, pp. 478-487, 2019.