LONTAR KOMPUTER VOL. 15, NO. 1 APRIL 2024 p-ISSN 2088-1541
DOI : 10.24843/LKJITI.2024.v15.i01.p06 e-ISSN 2541-5832
Accredited Sinta 2 by RISTEKDIKTI Decree No. 158/E/KPT/2021

Comparison of Gain Ratio and Chi-Square Feature Selection Methods in Improving SVM Performance on IDS

Ricky Aurelius Nurtanto Diaz[a1,2], I Ketut Gede Darma Putra[b3], Made Sudarma[c4], I Made Sukarsa[b5], Naser Jawas[e6]

[a]Engineering Science Student of Udayana University,
Department of Computer Systems, Faculty of Informatics and Computer,
Institut Teknologi dan Bisnis STIKOM Bali
Renon, Denpasar, Bali, Indonesia
[1] diaz.2391011008@student.unud.ac.id
[2]ricky@stikom-bali.ac.id (Corresponding author)

[b]Department of Information Technology, Faculty of Engineering, Udayana University
Jimbaran Hill, Bali, Indonesia
[3]ikgdarmaputra@unud.ac.id
[5]sukarsa@unud.ac.id

[c]Department of Electrical Engineering, Faculty of Engineering, Udayana University
Jimbaran Hill, Bali, Indonesia
[4]msudarma@unud.ac.id

PhD student, School of Engineering, The University of Warwick
[6]naser.jawas@warwick.ac.uk

Abstract

An intrusion detection system (IDS) is a security technology designed to identify and monitor suspicious activity in a computer network or system and detect potential attacks or security breaches. The importance of accuracy in IDS must be addressed, given that the response to any alert or activity generated by the system must be precise and measurable. However, achieving high accuracy in IDS requires a process that takes work. The complex network environment and the diversity of attacks led to significant challenges in developing IDS. The application of algorithms and optimization techniques needs to be considered to improve the accuracy of IDS. Support vector machine (SVM) is one data mining method with a high accuracy level in classifying network data packet patterns. A feature selection stage is needed for an optimal classification process, which can also be applied to SVM. Feature selection is an essential step in the data preprocessing phase; optimization of data input can improve the performance of the SVM algorithm, so this study compares the performance between feature selection algorithms, namely Information Gain Ratio and Chi-Square, and then classifies IDS data using the SVM algorithm. This outcome implies the importance of selecting the right features to develop an effective IDS.

Keywords: *Feature, Selection, Gain Ratio, Chi-Square, SVM, IDS*

1. Introduction

An intrusion detection system (IDS) is a security technology designed to identify and monitor suspicious activity in a computer network or system and detect potential attacks or security breaches. In network activity, IDS will identify suspicious activity and recognize attacks on the network. When the IDS discovers this attack or activity, the IDS will send reports and notifications to the network administrator.[1] The increase in online threats and attacks shows that developing an Intrusion Detection System is imperative to protect networks and computer systems [2]. IDS is an effective tool for monitoring networks, especially to detect malicious attacks [3]. An IDS can detect network anomaly behavior such as Denial of Service (DoS), Probe, SSH Brute Force (SBF), Brute Force Web (BFW), SQL Injection (SQLI), and other types of attacks. [4]

The importance of accuracy in IDS must be addressed, given that the response to any alert or activity generated by the system must be precise and measurable. The high accuracy ensures that limited resources are not wasted on unnecessary investigations or excessive responses to false alerts. In addition, high accuracy also contributes to a better understanding of potentially emerging attack patterns, which ultimately helps take better precautions and design more resilient defense systems. However, achieving high accuracy in IDS requires a process that takes work. The complex and evolving network environment and the ever-changing diversity of attacks led to significant challenges in developing detection algorithms distinguishing between normal and suspicious activity. The development of data and networks makes data to be processed bigger and included in the Big Data category.

The application of algorithms and optimization techniques needs to be considered to improve the accuracy of IDS. Researchers in recent years using various publicly accessible data sets such as KDDCUP, NSL KDD, Darpa, and other public datasets tried various machine learning-based intrusion detection methods by applying various algorithms, such as the application of PSO [5], SMOTE, and Random Forest [6], including their use in IoT [7]. One algorithm that is also frequently used is the application of Support Vector Machines (SVM), which can learn patterns from training data and apply them to new data to detect suspicious activity.

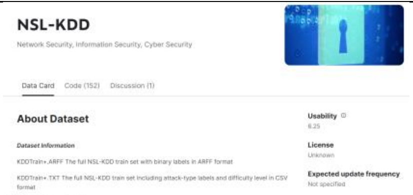
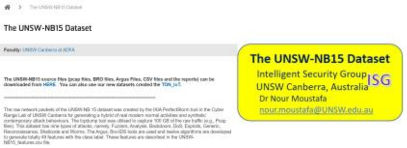
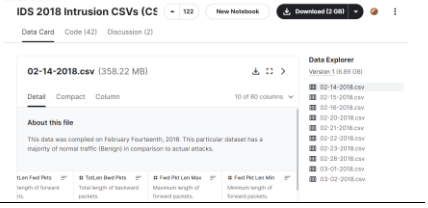
Support vector machine (SVM) is one data mining method with a high accuracy level in classifying data. The previous Comparison of SVM and ANN Classifiers for the COVID-19 Prediction study proves that SVM results have slightly better accuracy than ANN [8]. Judging from the high accuracy, SVM research compared to KNN research shows that the SVM algorithm exhibits higher accuracy when tested with normalization, outperforming the KNN algorithm in normalized and non-normalized conditions. However, the KNN algorithm consistently demonstrates lower accuracy, achieving an SVM accuracy of 84.61% and a KNN accuracy of 64.83% [9]. In another study, SVM achieved high metrics of accuracy, acquisition, precision, and F1 score where this study used intrusion detection domain dataset with 93.75% accuracy on the UNSW-NB15 dataset, 98.92% accurate curation on the CICIDS2017 dataset [10].

At the classification stage, a feature selection stage is needed. Feature selection is an essential step in the data preprocessing phase. This phase entails choosing a subset of pertinent features from a broader set of available features. Some examples of research that apply this technique include the application of IG-R to improve IDS performance [11], research that uses a combination of PSO and CFS for selecting features [12], and the Farmland Fertility Algorithm [13]. The importance of feature selection can be seen from previous studies that prove that optimization algorithms can help increase SVM accuracy by up to 36.2% compared to the SVM process without using feature selection optimization [14]. One feature selection method is the Gain Ratio model, which can improve classification models' accuracy [15]. Other studies indicate that employing the gain ratio method can enhance the efficacy of the Support Vector Machine (SVM) algorithm. This improvement is observed when utilizing features ranging from 100% to 5%, with optimal precision achieved at 50% of the features. However, the highest accuracy and recall are attained when utilizing only 5% of the features. Another research related to feature selection methods is to be able to rank feature sets on microarray datasets. The five feature selection methods include Chi-square, Relief, Gain Ratio, Information Gain, and Symmetrical Uncertainty. Four classification methods were also carried out for the classification stage, which was also carried out ten times, including cross-validation in each classification method. The results of this study, the feature selection method that excels a lot in several Microarray datasets is the Superior Gain Ratio method in the breast, Colon, and Ovarian datasets, with each value of the recognition rate 84.69, 82.25, and 87.91 [16]. The gain ratio is a widely used feature selection method that is useful in improving the accuracy of classification models, assisting in selecting relevant features and reducing complexity.

Another feature selection method that has advantages and is widely used is the Chi-Square (CHI) method. This feature selection method is robust in using statistics developed to measure the relationship between two categorical variables in contingency tables. Features with a strong relationship with the target may be considered for inclusion in the model, while less informative features may be omitted. Related research using the Chi-Square (CHI) method for feature selection [17] states that the Chi-Square method can help optimize the threshold for NeighShrink, which is a denoising algorithm method to reduce additional white Gaussian noise, where the experimental results show that the proposed algorithm is simple and efficient, and provides noise reduction, and can maintain good edges and detail. Another research related to applying Chi-Square combined in Arabic text classification to improve classification performance. This combination significantly enhanced the performance of the Arabic text classification model with a dataset of 5,070 data and Arabic documents classified into six independent classes. The best f-measure obtained for this model is 90.50% when the number of features is as much as 900 [18]. The chi-square method is helpful for feature selection in machine learning and data analysis. This technique helps identify the most informative features by evaluating their relationship with the target variable.

Based on previous research showed that optimization of data input can improve the performance of classification algorithms and saw the ability of SVM in the data classification process, including network data packet patterns, as evidenced by several studies using KDDCUP, NSL KDD, Darpa, and other public datasets, this study compared the performance between feature selection algorithms, namely Information Gain Ratio and Chi-Square. We use these two feature selection algorithms because the data types of all the datasets used have various forms, such as categorical, continuous, and numerical data. Data that the selection has processed feature is also classified using the SVM algorithm. This study uses NSL KDD, UNSW, and CSE CIC datasets IDS2018, and the results of this research comparison will be used as the basis for further research in the classification optimization stage. The details about datasets are shown in Table 1 :

Table 1. Datasets Source

Datasets	Source	Image
NSL KDD	https://www.kaggle.com/datasets/hassan06/nslkdd	 NSL-KDD Network Security, Information Security, Cyber Security Data Card Code (152) Discussion (1) About Dataset *Dataset Information* KDDTrain+.ARFF The full NSL-KDD train set with binary labels in ARFF format KDDTrain+.TXT The full NSL-KDD train set including attack-type labels and difficulty level in CSV format Usability 6.25 License Unknown Expected update frequency Not specified
UNSW	https://research.unsw.edu.au/projects/unsw-nb15-dataset	 The UNSW-NB15 Dataset The UNSW-NB15 Dataset Faculty: UNSW Canberra at ADFA The UNSW-NB15 source files (pcap files, BRO files, Argus Files, CSV files and the reports) can be downloaded from HERE. You can also use our new datasets created the TON_IoT. The raw network packets of the UNSW-NB 15 dataset was created by the IXIA PerfectStorm tool in the Cyber Range Lab of UNSW Canberra for generating a hybrid of real modern normal activities and synthetic contemporary attack behaviours. The tcpdump tool was utilised to capture 100 GB of the raw traffic (e.g., Pcap files). This dataset has nine types of attacks, namely, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms. The Argus, Bro-IDS tools are used and twelve algorithms are developed to generate totally 49 features with the class label. These features are described in the UNSW-NB15_features.csv file. The UNSW-NB15 Dataset Intelligent Security Group ISG UNSW Canberra, Australia Dr Nour Moustafa nour.moustafa@UNSW.edu.au
CSE CIC	https://www.kaggle.com/datasets/solarmainframe/ids-intrusion-csv	 IDS 2018 Intrusion CSVs (CS Data Card Code (42) Discussion (2) 02-14-2018.csv (358.22 MB) Detail Compact Column 10 of 80 columns About this file This data was compiled on February Fourteenth, 2018. This particular dataset has a majority of normal traffic (Benign) in comparison to actual attacks. tLen Fwd Pkts: length of forward ts. # TotLen Bwd Pkts: Total length of backward packets. # Fwd Pkt Len Max: Maximum length of forward packets. # Fwd Pkt Len Min: Minimum length of forward packets. Data Explorer Version 1 (6.89 GB) 02-14-2018.csv 02-15-2018.csv 02-16-2018.csv 02-20-2018.csv 02-21-2018.csv 02-22-2018.csv 02-23-2018.csv 02-28-2018.csv 03-01-2018.csv 03-02-2018.csv

2. Research Methods

This study used a process of comparing the classification results between feature selection using gain ratio and chi-square. The process starts with acquiring the IDS dataset using the NSL-KDD, UNSW Datasets, and CSE CIC 2018. This research uses normal and abnormal classes, representing normal by 0 and abnormal by 1.

For NSL-KDD data, the data that has been owned is then carried out in the data transformation process by coding the value 'normal' as 0 and other values as 1 in the attack attribute. In the UNSW dataset, the labels (classes) are 0 and 1, where class 0 represents the normal form of network traffic, and 1 represents the attack. On the CSE CIC dataset, we carried out a transformation from the Label (class), namely Benign, to 0, which means normal, and two types of brute force attacks, namely FTP-BruteForce and SSH-BruteForce, to 1, which represents the form of attack.

p-ISSN 2088-1541
e-ISSN 2541-5832

Next, this attribute will be used as a classifier for binary models to identify any attack. The subsequent phase involves implementing feature selection on the existing dataset using the gain ratio and chi-square methods. The feature selection results from these two techniques will cause not all attributes from the original dataset to be used. From the feature selection results, we used a dataset using selected features for training, and the classification process was then tested using the SVM algorithm. The following process compares the final results presented from the confusion matrix process.

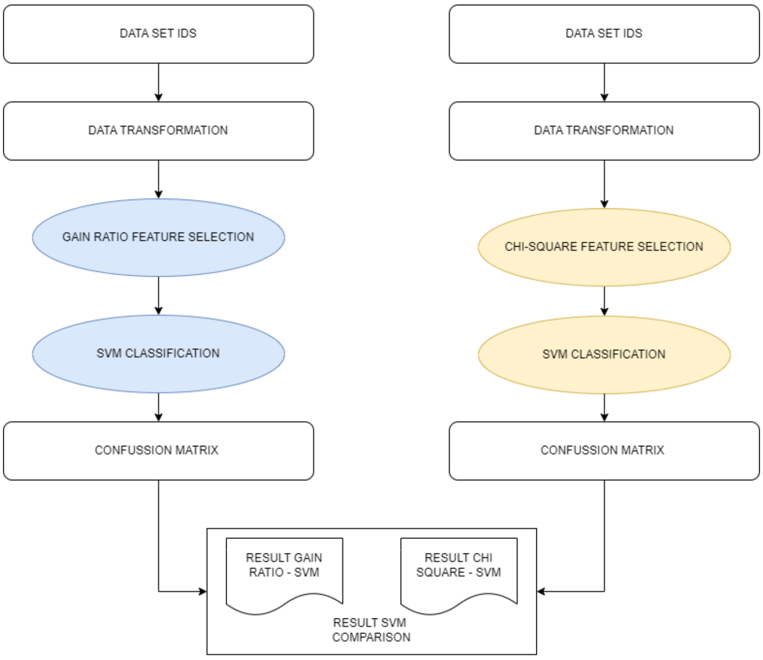


Figure 1. Research Methods

2.1. Data Mining

Data mining is the process of extracting knowledge sourced from large or complex datasets. There are two common techniques in data mining: descriptive and predictive methods [19]. In descriptive approaches, algorithms identify patterns that describe data by examining the relationships among data labels or attributes. Clustering, association rule mining, and sequential pattern discovery represent three model learning methods characterized by their descriptive nature in data mining. Predictive methods use the value of several features to predict a particular value or trait in the future based on data held in the past or present. This technique is also known as the supervised learning method, and some common algorithms include classification, regression, and anomaly detection. [19]

2.2. SVM

Support vector machines (SVM) are a subset of supervised machine learning techniques that construct a binary classification framework for addressing intricate, highly non-linear challenges [20]. Commonly applied in regression and classification challenges, Support Vector Machine (SVM) was conceived by Vapnik as part of the machine learning toolkit. SVM identifies the most effective separator to distinguish between two distinct classes. Additionally, SVM offers a cohesive framework enabling the classification of diverse data through the selected kernel. This is considered one of the advantages of SVM. [21]

2.3. Gain Ratio

The gain ratio is one of the metrics used in classifying data or selecting features in machine learning and data analysis. This metric measures how well a feature or attribute performs at class separation in a dataset. The gain ratio is often used in the context of decision tree algorithms.

Gain Ratio is an increase in Information Gain that attempts to optimize normalized values for a feature in the context of classification. The gain ratio was chosen because of its ability to produce higher accuracy than other filter techniques.[22]. To calculate the Gain Ratio, it is necessary to calculate the Information Gain first, with the process of calculating the gain ratio as follows:

$$\text{Gain Ratio(A)} = \text{Gain(A)/SplitInfo(A)} \tag{1}$$

Gain(A) is the calculation of Information Gain, and SplitInfo(A) is the split of the entropy calculation.

$$\text{Gain(A)} = -\sum_{i=1}^{v}(Pi * log10Pi) - \sum_{j=1}^{v}\left(\frac{Dj}{D}\right) * \left(-\sum_{j=1}^{v}(Pj * log10Pj)\right) \tag{2}$$

$$\text{SplitInfo} = -\sum_{j=1}^{v}\left(\frac{Dj}{D}\right) * \log\frac{Dj}{D} \tag{3}$$

2.4. Chi-Square

A chi-square statistical test tests the difference between theoretical (assumed) and observed distributions. This test is generally used in quantitative research, especially qualitative research, which uses categorized data.

$$Chi - Square\ (\text{tk}, Ci) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{4}$$

Based on the equation above, each feature is given a value for each class, and then the maximum final value is obtained by combining all these values.[18]

2.5. Evaluation Methods

The evaluation method in this study uses the confusion matrix method for precision, Recall, and Accuracy measurements. The confusion matrix compares the model or algorithm's classification results with the actual classification results. The matrix is described in the following table:

Table 2. Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

3. Result and Discussion

The datasets to be used in the classification process in this study are NSL KDD, UNSW, and CSE CIC IDS2018. This dataset contains several network traffic log data labeled as normal or intrusive. We transform the attributes that will become class labels to be classified into Normal and Not Normal.

Subsequently, the preprocessed dataset undergoes the application of the Gain Ratio and Chi-Square feature selection methods. This method ranks features based on their relevance to the target variable (normal or normal nor). Furthermore, the data that has gone through the preprocessing process will proceed to the training phase. In the training process, we used the NSL-KDD Training dataset and UNSW, which provided a special dataset for training. Meanwhile, the training process was carried out for the CSE CIC dataset using 70% of the data as a training source. The last stage is classification with SVM, which evaluates its performance using performance metrics: accuracy, precision, and recall. The process is carried out for all three datasets, and the results are compared.

Before selecting features, the IDS dataset that has been owned will be transformed into a transformation process, especially for attributes that will become labels (classes). The data owned by NSL-KDD has attributes such as Protocol_Type, Src_Bytes, Attack, and Level from a total of

p-ISSN 2088-1541
e-ISSN 2541-5832

43 attributes. The Attack attribute (attribute 42) has several values, including normal, neptune, nmap, spy, etc. The transformation process occurs in this attribute, i.e., encoding the value 'normal' as 0 and the other as 1.

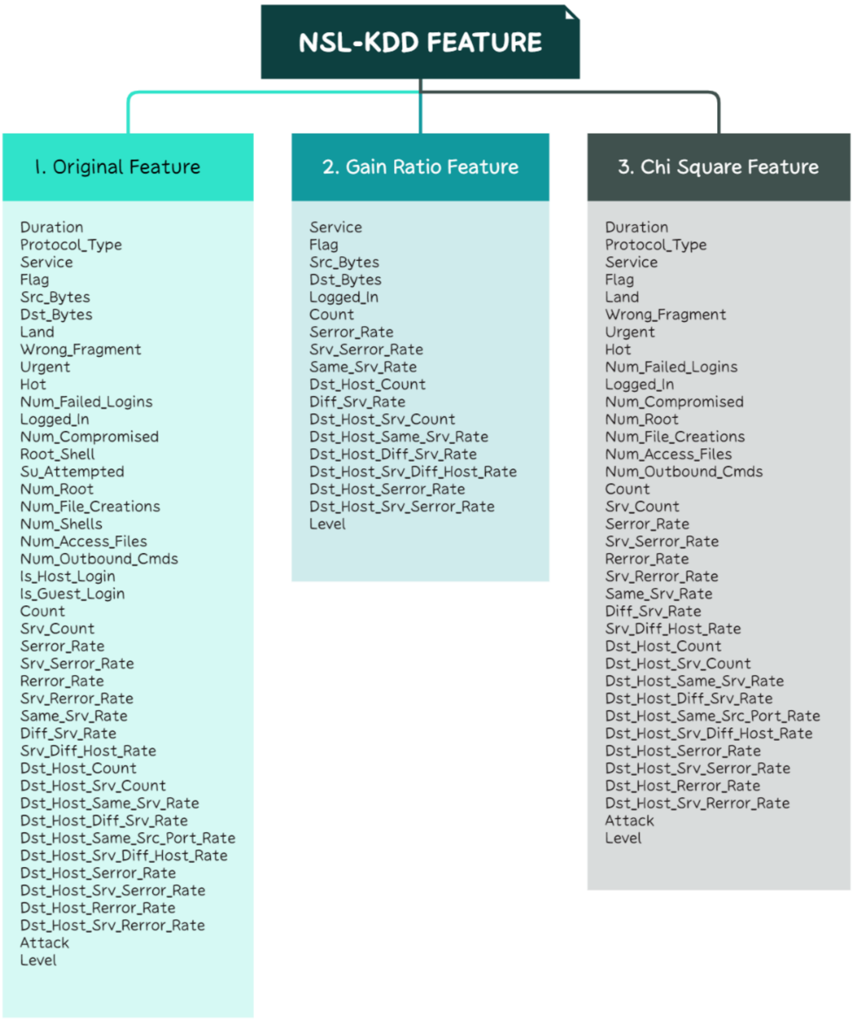


Figure 2. NSL KDD Features

At the feature selection stage, we compared the selection results of two feature selection methods, namely Gain Ratio and Chi-Square. Several attributes are obtained from the feature selection results of the Gain Ratio method, such as Service, Flag, Logged_In, Count, Dst_Host_Srv_Count, and Dst_Host_Count. From the feature selection results using the gain ratio, 19 attributes, including labels, will be used in the next stage. There are 24 features removed from the use of the gain ratio method. In the Chi-Square selection process, 36 features were used, and only seven features were ignored by this method. The attributes obtained from the Chi-Square method feature extraction results include Duration, Protocol_Type, Service, Flag, Land, and 31 other features.

In the next stage, we implement the feature selection results from the NSL KDD dataset with SVM using the dot kernel. The following are the results of SVM performance with the selection of Gain Ratio features:

Table 3. Gain ratio and SVM Performance

	Predicted Not Normal	Predicted Normal	Class Recall
Actual Not Normal	7583	5250	59,09%
Actual Normal	304	9407	96,87%
Class Precision	96,15%	64,18%	

These results obtained accuracy in data testing using a feature selection model, and the gain ratio was 75.36%. Furthermore, for SVM performance results using Chi-Square feature selection are as follows:

Table 4. Chi-square and SVM Performance

	Predicted Not Normal	Predicted Normal	Class Recall
Actual Not Normal	51286	7344	87,47%
Actual Normal	2835	64508	95,79%
Class Precision	94,76%	89,78%	

We used the UNSW and CSE CIC datasets IDS2018 in another experiment while utilizing the previously mentioned feature extraction methods. For the UNSW dataset, which includes 49 features and labels, the Gain Ratio feature extraction stage resulted in 12 optimal parts, whereas the Chi-Square method identified 42 features for use. The detailed UNSW dataset features are shown in Figure 3.

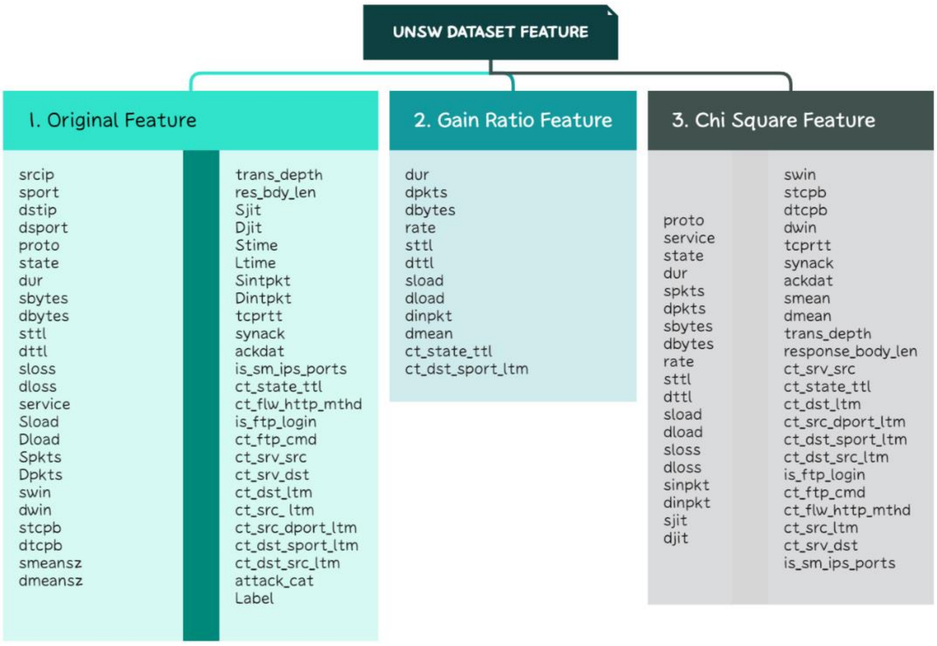


Figure 3. UNSW Dataset Features

Regarding the CSE CIC dataset IDS2018, which encompasses 80 elements, including labels, the Gain Ratio feature extraction stage produced the best 14 features. In contrast, the Chi-Square feature extraction method identified the best six features, as shown in Figure 4.

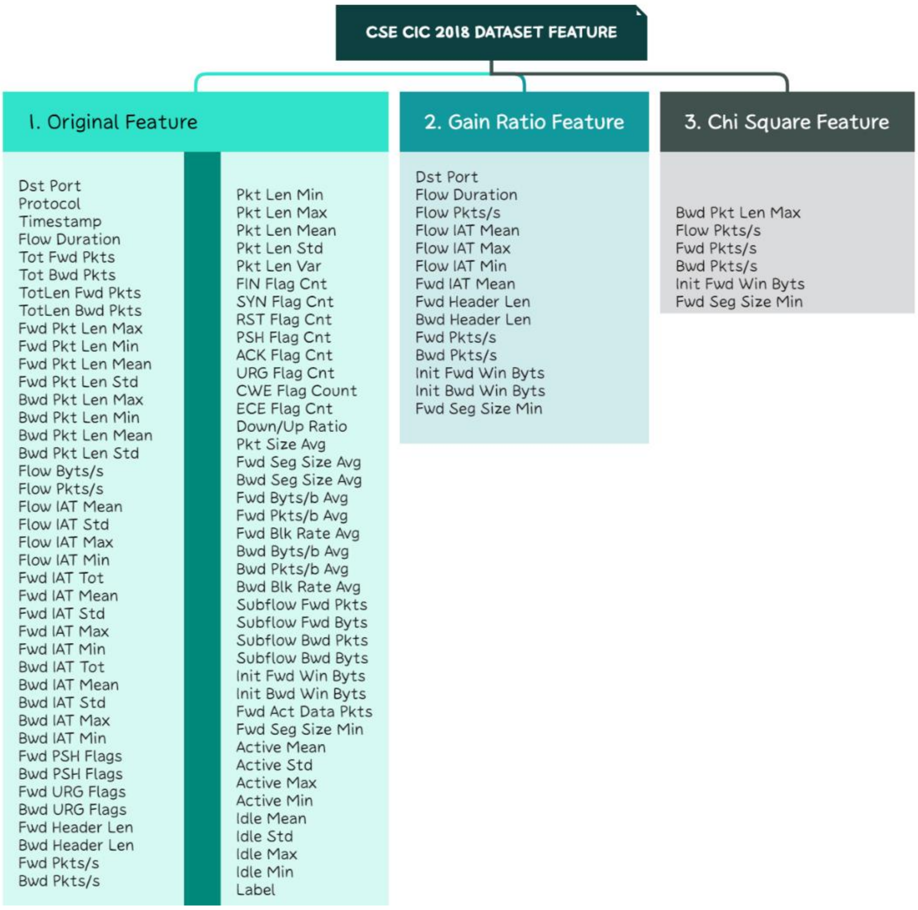


Figure 4. CSE CIC 2918 Dataset Features

The results of the feature extraction process from the two datasets above are then processed using SVM with dot kernels. The following are the performance results of SVM with the Gain Ratio feature selection for the UNSW and CSE CIC IDS2018 datasets:

Table 5. SVM+Gain Ratio Dataset UNSW

	Predicted Not Normal	Predicted Normal	Class Recall
Actual Not Normal	45184	148	99,67%
Actual Normal	20103	16897	45,67%
Class Precision	69,21%	99,13%	

Table 6. SVM+Gain Ratio Dataset CSE CIC

	Predicted Not Normal	Predicted Normal	Class Recall
Actual Not Normal	8570	2858	74,99%
Actual Normal	588	19441	97,06%
Class Precision	93,58%	87,18%	

From these results, accuracy was obtained in testing data testing using the feature selection model, with a gain ratio of 75.40% for the UNSW dataset and an accuracy of 89.05% for the CSE CIC dataset. Furthermore, for SVM performance results using Chi-Square feature selection are as follows:

Table 7. SVM+Chi Square Dataset UNSW

	Predicted Not Normal	Predicted Normal	Class Recall
Actual Not Normal	45027	305	99,33%

Actual Normal	16574	20426	55,21%
Class Precision	73,09%	98,53%	

Table 8. SVM+Chi Square Dataset CSE CIC

	Predicted Not Normal	Predicted Normal	Class Recall
Actual Not Normal	8649	2779	75,68%
Actual Normal	5289	14740	73,59%
Class Precision	62,05%	84,14%	

These results showed that SVM performance using test data with a feature selection model from Chi-Square resulted in an accuracy of 79.50% for the UNSW dataset and 74.35% for the CSE CIC dataset. Figure 5 shows a comparison of the accuracy results of the selection of Gain Ratio and Chi-Square features for all datasets used:

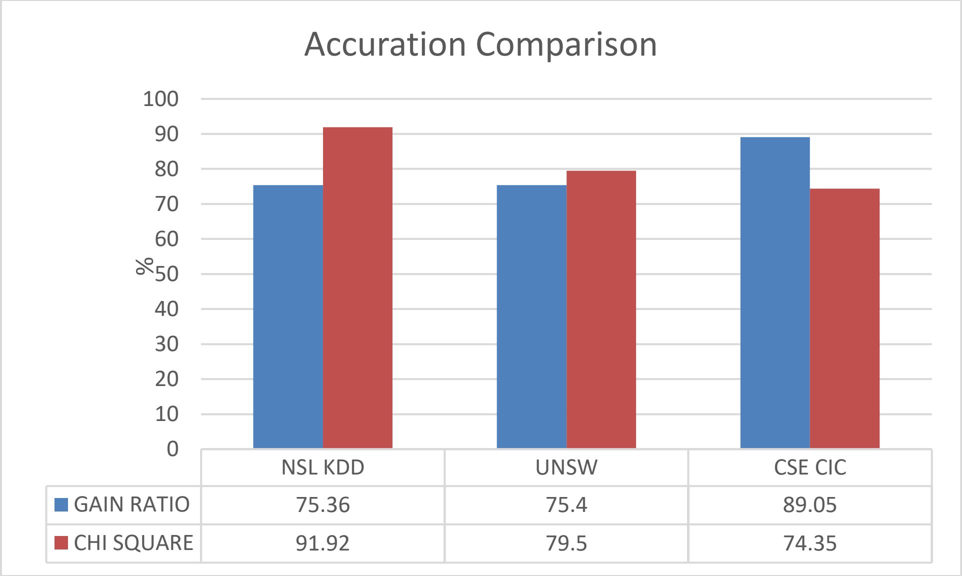


Figure 5. Accurate comparison

Analyzing the accuracy outcomes shown in Figure 5, it becomes evident that choosing the Chi-Square feature selection yields advantages over the Gain Ratio in experiments conducted on two datasets, NSL KDD and UNSW. Conversely, the choice of Gain Ratio features proves superior in tests involving the CSE CIC dataset. When compared with the difference in accuracy levels in the three datasets, results were obtained for the superiority of Chi-Square on the NSL KDD dataset with a percentage of 18%. Here is a table of accuracy differences for the three datasets used:

Table 9. Accuracy Difference

Dataset	Gain Ratio	Chi-Square	Difference
NSL KDD	75,36	91,92	18%
UNSW	75,4	79,5	5%
CSE CIC	89,05	74,35	17%

A comparison of precision and recall results of both feature selection techniques on all datasets can be seen in Figure 6:

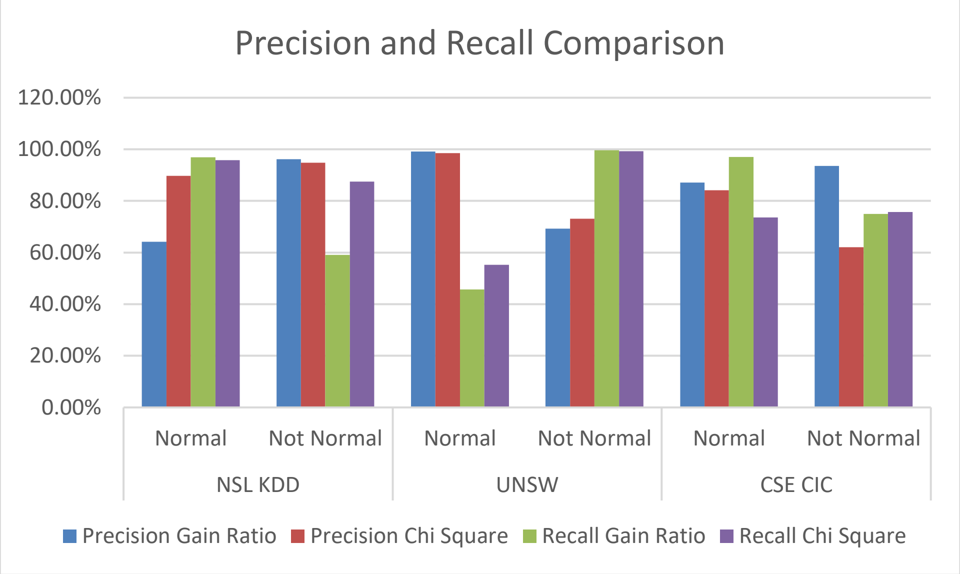


Figure 6. Precision and recall comparison

The experimental results shown in Figure 6 show that both feature selection methods improve SVM performance compared to using the entire feature. The experimental findings from NSL KDD, UNSW, and CSE CIC datasets indicate that SVM accuracy improves when employing Chi-Square feature selection compared to Gain Ratio feature selection.

4. Conclusion

This study has compared the selection methods of Gain Ratio and Chi-Square features in the context of IDS to improve SVM performance. Results show that both methods can improve SVM performance in detecting intrusions. From the comparison results, the Gain Ratio value is smaller than the Chi-Square for testing two datasets, so the Chi-Square method is more recommended in feature selection because it provides slightly better results in terms of accuracy. Based on the accuracy results, the two selection features can work optimally if the features they have are not too few, or in other words, they have enough features to be the basis for the classification process. This can be seen from the accuracy results when Gain Ratio and SVM are used for the CSE CIC dataset, which produces an accuracy of 89.05%. The dataset features obtained by the Gain Ratio are more than those provided by Chi-Square and are sufficient for classification. The same thing can be seen when Chi-Square and SVM produce higher accuracy than Gain Ratio and SVM for the NSL KDD and UNSW datasets. This outcome implies the importance of selecting the right features to develop an effective IDS.

References

[1] L. Yang and A. Shami, “IDS-ML: An open source code for Intrusion Detection System development using Machine Learning[Formula presented],” *Software Impacts*, vol. 14, Nov. 2022, doi: 10.1016/j.simpa.2022.100446.

[2] M. A. Hossain and M. S. Islam, “Ensuring network security with a robust intrusion detection system using ensemble-based machine learning,” *Array*, p. 100306, Sep. 2023, doi: 10.1016/j.array.2023.100306.

[3] Z. Yang *et al.*, “A systematic literature review of methods and datasets for anomaly-based network intrusion detection,” *Computers and Security*, vol. 116. Elsevier Ltd, May 01, 2022. doi: 10.1016/j.cose.2022.102675.

[4] B. M. Serinelli, A. Collen, and N. A. Nijdam, "On the analysis of open source datasets: Validating IDS implementation for well-known and zero-day attack detection," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 192–199. doi: 10.1016/j.procs.2021.07.024.

[5] N. Kunhare, R. Tiwari, and J. Dhar, "Particle swarm optimization and feature selection for an intrusion detection system," *Sādhanā*, vol. 45, 2020, doi: 10.1007/s12046-020-1308-5S.

[6] R. Alshamy, M. Ghurab, S. Othman, and F. Alshami, “Intrusion Detection Model for Imbalanced Dataset Using SMOTE and Random Forest Algorithm,” in *Communications in Computer and Information Science*, Springer Science and Business Media Deutschland GmbH, 2021, pp. 361–378. doi: 10.1007/978-981-16-8059-5_22.

[7] D. Musleh, M. Alotaibi, F. Alhaidari, A. Rahman, and R. M. Mohammad, “Intrusion Detection System Using Feature Extraction with Machine Learning Algorithms in IoT,” *Journal of Sensor and Actuator Networks*, vol. 12, no. 2, Apr. 2023, doi: 10.3390/jsan12020029.

[8] D. N. Avianty, Prof. I. G. P. S. Wijaya, and F. Bimantoro, “The Comparison of SVM and ANN Classifier for COVID-19 Prediction,” *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 13, no. 2, p. 128, Aug. 2022, doi: 10.24843/lkjiti.2022.v13.i02.p06.

[9] D. A. Anggoro, “Comparison of Accuracy Level of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) Algorithms in Predicting Heart Disease,” *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, pp. 1689–1694, May 2020, doi: 10.30534/ijeter/2020/32852020.

[10] J. Gu and S. Lu, “An effective intrusion detection approach using SVM with naïve Bayes feature embedding,” *Comput Secur*, vol. 103, Apr. 2021, doi: 10.1016/j.cose.2020.102158.

[11] Y. K. Saheed and F. E. Hamza-Usman, “Feature Selection with IG-R for Improving Performance of Intrusion Detection System,” 2020.

[12] T. Ahmad and M. N. Aziz, “Data preprocessing and feature selection for machine learning intrusion detection systems,” *ICIC Express Letters*, vol. 13, no. 2, pp. 93–101, 2019, doi: 10.24507/icicel.13.02.93.

[13] T. S. Naseri and F. S. Gharehchopogh, “A Feature Selection Based on the Farmland Fertility Algorithm for Improved Intrusion Detection Systems,” *Journal of Network and Systems Management*, vol. 30, no. 3, Jul. 2022, doi: 10.1007/s10922-022-09653-9.

[14] A. F. Indriani and M. A. Muslim, “SVM Optimization Based on PSO and AdaBoost to Increasing Accuracy of CKD Diagnosis,” *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, p. 119, Aug. 2019, doi: 10.24843/lkjiti.2019.v10.i02.p06.

[15] S. J. Pasha and E. S. Mohamed, “Advanced hybrid ensemble gain ratio feature selection model using machine learning for enhanced disease risk prediction,” *Informatics in Medicine Unlocked*, vol. 32, Jan. 2022, doi: 10.1016/j.imu.2022.101064.

[16] N. D. Cilia, C. De Stefano, F. Fontanella, S. Raimondo, and A. S. di Freca, “An experimental comparison of feature-selection and classification methods for microarray datasets,” *Information (Switzerland)*, vol. 10, no. 3, 2019, doi: 10.3390/info10030109.

[17] C. J. Zhang, X. Y. Huang, and M. C. Fang, “MRI denoising by NeighShrink based on chi-square unbiased risk estimation,” *Artificial Intelligence in Medicine*, vol. 97, pp. 131–142, Jun. 2019, doi: 10.1016/j.artmed.2018.12.001.

[18] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, “Feature selection using an improved Chi-square for Arabic text classification,” *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 2, pp. 225–231, Feb. 2020, doi: 10.1016/j.jksuci.2018.05.010.

[19] J. H. Joloudari, H. Saadatfar, A. Dehzangi, and S. Shamshirband, “Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection,” *Informatics in Medicine Unlocked*, vol. 17, Jan. 2019, doi: 10.1016/j.imu.2019.100255.

[20] C. Ioannou, V. Vassiliou, and by Ieee, “Network Attack Classification in IoT Using Support Vector Machines,” 2021, doi: 10.3390/jsan.

[21] S. İlkin, T. H. Gençtürk, F. Kaya Gülağız, H. Özcan, M. A. Altuncu, and S. Şahin, “hybSVM: Bacterial colony optimization algorithm based SVM for malignant melanoma detection,” *Engineering Science and Technology, an International Journal*, vol. 24, no. 5, pp. 1059–1071, Oct. 2021, doi: 10.1016/j.jestch.2021.02.002.

[22] P. Nimbalkar and D. Kshirsagar, “Feature selection for intrusion detection system in Internet-of-Things (IoT),” *ICT Express*, vol. 7, no. 2, pp. 177–181, Jun. 2021, doi: 10.1016/j.icte.2021.04.012.