Feature Selection using Chi Square to Improve Attack Detection Classification in IoT Network: Work in Progress

By Zulhipni Reno Saputra Elsi

WORD COUNT 3787
TIME SUBMITTED 03-AUG-2022 11:48PM
PAPER ID 89190412

Feature Selection using Chi Square to Improve Attack Detection Classification in IoT Network: Work in Progress

Zulhipni Reno Saputra Elsi
Faculty of Engineering Universitas Muhammadiyah Palembang/Faculty of Engineering Universitas Sriwijaya
Palembang, Indonesia
zulhipni_renosaputra@um-palembang.ac.id

Deris Stiawan*
Computer Engineering Department, Faculty of Computer Science Universitas Sriwijaya
Palembang, Indonesia
deris@unsri.ac.id

Ahmad Fali Oklilas
Computer Engineering Department, Faculty of Computer Science Universitas Sriwijaya
Palembang, Indonesia
fali@ilkom.unsri.ac.id

Susanto
Faculty of Computer Universitas Bina Insan/Faculty of Engineering Universitas Sriwijaya
Lubuklinggau, Indonesia
susanto@univbinainsan.ac.id

Kurniabudi
Computer Engineering Department Universitas Dinamika Bangsa
Jambi, Indonesia
kurniabudi@unama.ac.id

Yesi Novaria Kunang
Faculty of Computer Science Universitas Bina Darma
Palembang, Indonesia
yesinovariakunang@binadarma.ac.id

Mohd. Yazid Idris
School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia
yazid@utm.my

Rahmat Budiarto
College of Computer Science and IT, AlBaha University
Albaha, Saudi Arabia
rahmat@bu.edu.sa

Abstract— To maintain network security, Intrusion Detection System (IDS) is needed to detect anomaly and attack. Designing proper IDS requires accurate model. This paper proposes a model, which consists of statistical extraction, feature selection, dataset clustering, classification, and performance measurement. Experiments on MQTT-IOT-IDS2020 dataset which contains Normal, scan_A, scan_sU, Sparta and mqtt_bruteforce are conducted. The dataset is statistically extracted using Bidirectional-based features packet header feature with 37 features. Chi square algorithm is selected for performing feature extraction process. 10 relevant and best features are selected and ranked into 5-subsets and 10-subset feature. Three dataset splitting into testing data and training data of 90%:10%, 70%:30% and 50%:50% are created. Binary classification using k-Nearest Neighbor (KNN) and Adaboost algorithms are performed. The experimental results show accuracy level above 99% for all scenarios, with Adaboost algorithm outperforms k-Nearest Neighbor algorithm.

Keywords— Intrusion Detection System, Chi Square, k-Nearest Neighbor, Adaboost, Binary Classification

I. INTRODUCTION

Increasing attacks on network require an IDS to detect anomaly, attack or vulnerabilities in the network [1]. There are many network attacks types, e.g.: brute force attacks [2], DDoS attacks [3], UDP Flood [4]. Visualization is one of the solutions in displaying attacks on the network and to find out anomalies or attacks on devices/sensors by visually managing traffic statistics in the form of graphs, checking device condition and predicting potential problems [5]. With attack visualization, it is easier to recognize and summarize pattern from complex visual images [6].

Thakkar Lohiya [1] use machine learning to design IDS for detecting and classifying attacks. Costantini et al. [7] state that the use of machine learning classifications enables us to analyze interference obtained from electronic devices. Goni et al. [8] discuss machine learning as cyber security tool to deal with network security, data security, endpoint security, identity access security, cloud security, IoT security as well as Fog security.

Data preprocessing is important process before performing IoT traffic data analysis, where the traffic data is processed by feature extraction and feature selection. Feature extraction is usually performed using statistical extraction [9] and packet extraction [10], while feature selection is carried out using principal component analysis (PCA) algorithm [11], Independent Component Analysis (ICA) algorithm [12], Linear Discriminant analysis (LDA) algorithm [12] and T-Shark [13]. Feature selection using feature ranking obtained from Chi square score values provides good performance of classification [14]. Chi square feature values tell the significance of each original feature.

This study proposes pre-processing using statistical extraction features and Chi square algorithm in determining feature rankings based on Chi score values. Then, it uses binary classification methods i.e.: KNN and Adaboost algorithms. The values of the Precision, Recall, F-Measure, and Accuracy of each algorithm are compared to determine the best algorithm.

The paper is structured as follows: Section 2 discusses and reviews related work on IoT anomaly detection issues, and classification performance. Section 3 describes the proposed design and method. Section 4 presents the results and analysis and Section 5 provides conclusions and future work.

II. RELATED WORK

According to Thaseen and Kumar [16] binary classifier model fails to achieve good attack detection rate and low false alarm level, because not all features captured from network packets contribute to detecting or classifying attacks. Classification algorithms for prediction and analysis are implemented in [17]. Researchers in [18] propose attack classification method in cloud network environments based on machine learning with digital forensic processes. The

researchers use fusion algorithms to detect ICMP attacks, TCPSync attacks, as well as UDP attacks. There are many algorithms that can be used as classifier, including Random Forest algorithm [19], k-Nearest Neighbor [20], Decision Tree [21], Support Vector Machine [22], Naive Bayes [23] and Adaboost [24].

There are 4 performance measurement parameters for classification, i.e.: Precision, Recall, F1 score and support [25] and accuracy level [26].

Confusion matrix introduced in [27] may solve both binary classification and multiclass classification problems.

Kurniabudi et al. [28] introduce 2 frameworks, which are global framework and the preprocessing framework. On the global framework there are five phases: 1) Building topology; 2) Sniffing packet of the sensor; 3) Pre-processing; 4) Learning; 5) Assessment and labeling, while the distributed pre-processing framework consists of 2 stages: 1) Extracting raw data, 2) Reducing data dimension.

Al-Hawawreh et al. [23] evaluate X-IIoTID dataset using several classification algorithms, i.e.: Decision Tree, Naive Bayes, K-Nearest Neighbors, Support Vector Machine (SVM), Logistic Regression, Deep Neural Networks, Gated Recurrent Units. The evaluation uses accuracy value, precision, recall, F1 score as performance metrics. Fleury et al. [29] use machine learning to recognize abnormalities in daily activities in smart homes. The researchers implement PCA algorithm as features extraction technique and use mean values and standard deviations for normalization process. The classification is performed using SVM algorithm with Polynomial and Gaussian kernels.

Alghazzawi et al. [30] recognize and classify distributed denial of service (DDoS) attack using CNN + BiLSTM hybrid deep neural network model on CICDDoS2019 dataset. The highest scoring and most relevant features are selected using ranking-based feature selection technique. This hybrid model can predict attack with accuracy rate of 94.52%, precision 94.74%, recall 92.04%, and F-score 93.44%.

III. DESIGN AND METHOD

This study implements the proposed classification algorithm on a computer with Intel Core i7-11800H CPU @2.30 GHz processor and 16 GB RAM running Ubuntu 20.04 operating system using Python 3.7 with Scikit-learn library for the machine learning feature selection and classification.

A. *Data Resource*

This research uses data sources from the study in [31] generated from MQTT sensor pool and extracted to obtain useful features of attack/interference detection. The dataset consists of five recorded scenarios; normal operation and four attack scenarios. The attacker performs four attacks and each attack was recorded [32]. The type of attack are: 1) Aggressive scanning (Scan A); 2) User Datagram Protocol (UDP) Scanning (Scan_sU); 3) Sparta SSH brute-force (Sparta); 4) MQTT brute-force attack (MQTT_BF). Characteristics of the dataset are listed in Table 1.

TABLE I. DATASET MQTT-IOT-IDS2020 CHARACTERISTIC [31]

No	File Name	File Pcap Size	Total Packet
1.	Normal	192.5 MB	1070577
2.	scan_A	16.2 MB	113940
3.	scan_sU	41.3 MB	255058
4.	Sparta	3,391.1 MB	20688940
5.	mqtt_bruteforce	906.8 MB	10049372

B. *Proposed Model*

In the dataset there are several types of attacks, but in this research all types of attacks are combined, thus, there are only normal data and attack data. The dataset in this study uses the Pcap format, which will be statistically extracted, using Bidirectional-based Features (BF) packet header feature then normalized and labeled.

In the feature selection process, Chi square algorithm is used to obtain more relevant rankings. Before classifying, the dataset is clustered into 2: testing data and training data with composition of 90%;10%, 70%:30% and 50%:50%.

The binary classification process uses supervised algorithms, requiring labeled data and model formation which is based on labeled training data, i.e.: KNN and Adaboost algorithms. The results of this research require validation and classification performance measurement.

Binary confusion matrix is used to validate the model. Table 2 shows the format of binary confusion.

TABLE II. BINARY CONFUSION MATRIX

		Prediction	
		Normal	**Attack**
Actual	**Normal**	True Positive	False Positive
	Attack	False Negative	True Negative

C. *Methodology*

Figure 1 shows the workflow for determining the best machine learning algorithm to be used for IDS on the Internet of Things Smart home network. The workflow of this research is divided into 5 stages, which are Dataset Extraction Stage, Normalization and labeling, the feature selection stage, cluster testing data & training data, classification stage with KNN and Adaboost algorithms and finally the performance measurement stage.

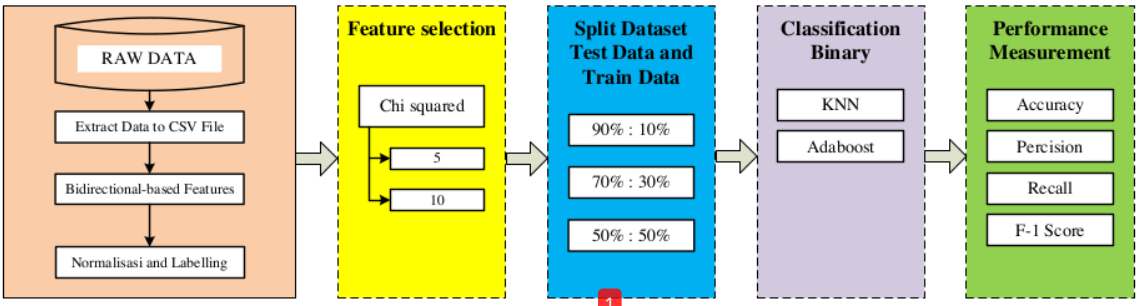


Figure 1. Proposed Method to determine best machine learning algorithm

The experiments are conducted with the following scenarios:

- Experiment 1, Using 5 feature with split dataset 90%:10% and KNN classification
- Experiment 2, Using 10 feature with split dataset dataset 90%:10% and KNN classification.
- Experiment 3, Using 5 feature with split dataset 70%:30% and KNN classification
- Experiment 4, Using 10 feature with split dataset 70%:30% and KNN classification
- Experiment 5, Using 5 feature with split dataset 50%:50% and KNN classification
- Experiment 6, Using 10 feature with split dataset 50%:50% and KNN classification
- Experiment 7, Using 5 feature with split dataset 90%:10% and Adaboost classification
- Experiment 8, Using 10 feature with split dataset 90%:10% and Adaboost classification
- Experiment 9, using 5 feature with split dataset 70%:30% and Adaboost classification
- Experiment 10, using 10 feature with split dataset 70%:30% and Adaboost classification
- Experiment 11, using 5 feature with split dataset 50%:50% and Adaboost classification
- Experiment 12, using 10 feature with split dataset 50%:50% and Adaboost classification

IV. RESULT AND ANALYSIS

Statistical extraction based on the BF packet header feature from the dataset produces 37 features. Then the best and relevant features are selected and ranked from these 37 features based on the value of Chi score. The features that get ranked 1 to 10 are: *bwd_num_bytes, fwd_num_bytes, prt_dst, num_src_flows,bwd_mean_offset, fwd_mean_offset, fwd_num_pkts, bwd_num_pkts, prt_src, fwd_num_psh_flags*, as shown in Table 3.

TABLE III. FEATURE RANKING

Rank	Feature	Chi score vslue
1	bwd_num_bytes	6.974210e+08
2	fwd_num_bytes	5.076926e+08
3	prt_dst	1.695434e+08
4	num_src_flows	9.370210e+07
5	bwd_mean_offset	2.387112e+07
6	fwd_mean_offset	2.386937e+07
7	fwd_num_pkts	6.456904e+06
8	bwd_num_pkts	6.385096e+06
9	prt_src	3.572532e+06
10	fwd_num_psh_flags	2.827373e+06

This research uses subset of 5 features selected from rank 1 to rank 5 and subset of 10 features selected from rank 1 to rank 10.

Performance measurements of each classification algorithm on testing dataset and training dataset are shown in Table 4 for accuracy performance, Table 5 for Recall, Precision and F1-Score performance, Figure 2 for KNN binary confusion matrix with subset 5, Figure 3 for KNN binary confusion matrix with subset 10. Figure 4 for the confusion of Adaboost binary matrix with subset 5 and Figure 5 for the confusion of Adaboost binary matrix with subset 10.

TABLE IV. ACCURACY VALUE COMPARISON

			Accuracy	
Subset	**Data split**	**Percentage**	**KNN**	**Adaboost**
5	Train Data	90%	99.99 %	100 %
		70%	99.99 %	100 %
		50%	99.99 %	100 %
	Test Data	10 %	99.99 %	100 %
		30%	99.99 %	100 %
		50%	99.99 %	100 %
10	Train Data	90%	99.96 %	100 %
		70%	99.95 %	100 %
		50%	99.88 %	100 %
	Test Data	10 %	99.65 %	100 %
		30%	99.97 %	100 %
		50%	99.90 %	100 %

In Figure 2, KNN classifier combined with the use of 5-subset features has 233430 normal data and 12 false alarms for 90% split data. 181557 normal data and 9 false alarms for 70% split data; 129653 normal data and 10 false alarms for 50% split data. Experiments on the training data with split data of 10% has normal data of 25929 and false alarm of 8; split data of 30% has normal data of 77804 and false alarms of 9; and split data of 50% has normal data of 129859 and false alarms of 9.

In Figure 3, KNN classifier combined with the use of 10-subset features has 233355 normal data and 96 false alarms for 90% split data. 181482 normal data and 9 false alarms for

70% split data; 86696 normal data and 124 false alarms for 50% split data. Experiments on the training data with split data of 10% has normal data of 25839 and false alarm of 98; split data of 30% has normal data of 77661 and false alarms of 152; and split data of 50% has normal data of 129559 and false alarms of 130.

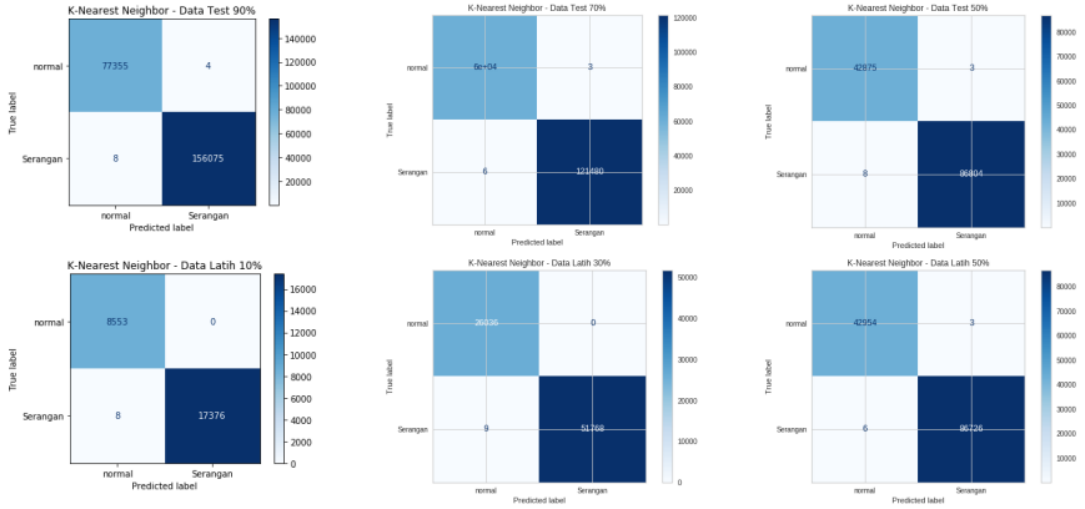
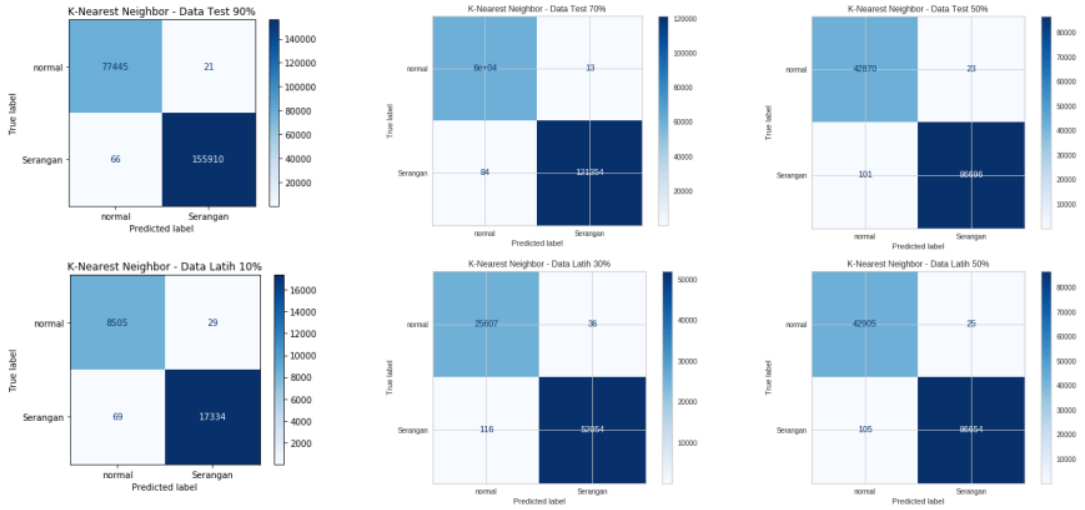


Figure 2. Binary Confussion Matrix KNN with subset 5 Feature



Gambar 3. Konfusi matrik biner KNN dengan subset 10 Feature

In Figure 4, Adaboost classifier combined with the use of 5-subset features has 233442 normal data and 0 false alarms for 90% split data. 181566 normal data and 0 false alarms for 70% split data; 129690 normal data and 0 false alarms for 50% split data. Experiments on the training data with split data of 10% has normal data of 25937 and false alarm of 8; split data of 30% has normal data of 77813 and false alarms of 0; and split data of 50% has normal data of 129689 and false alarms of 0.

In Figure 5, Adaboost classifier combined with the use of 10-subset features has 233442 normal data and 0 false alarms for 90% split data. 181566 normal data and 0 false alarms for 70% split data; 129690 normal data and 0 false alarms for 50% split data. Experiments on the training data with split data of 10% has normal data of 25937 and false alarm of 8; split data of 30% has normal data of 77813 and false alarms of 0; and split data of 50% has normal data of 129699 and false alarms of 0.

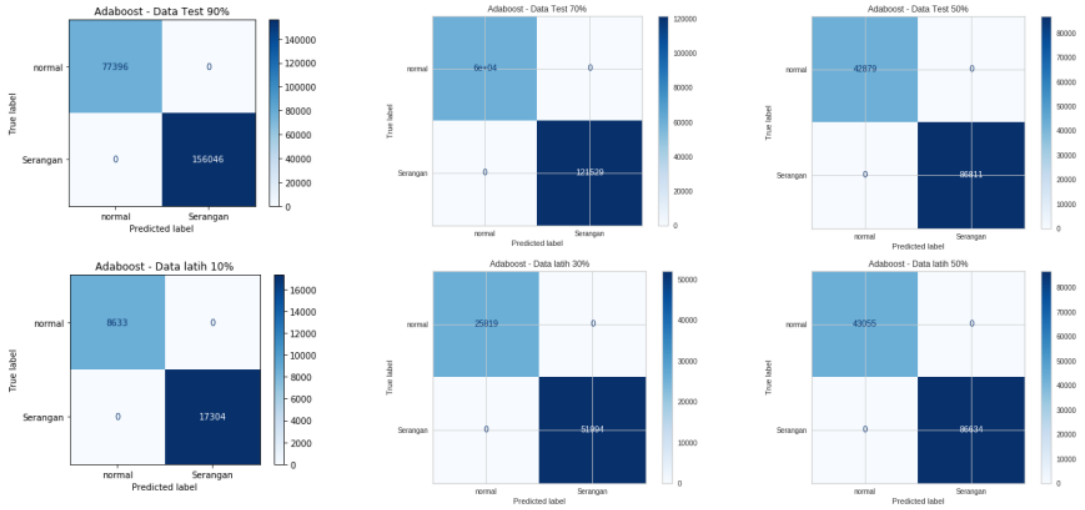


Figure 4. Binary Confussion Matrix Adaboost with subset 5 Feature

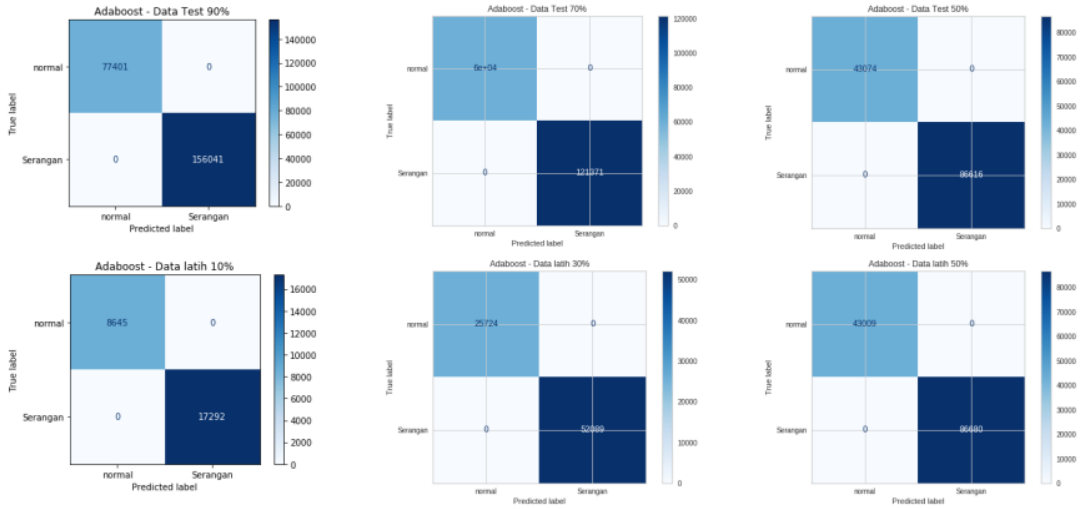


Figure 5. Adaboost Binary Confussion Matrix with subset 10 Feature

TABLE V. COMPARISON OF PRECISION, RECALL, F1 SCORE VALUE

Subset	Data split	Percentage	Precision		Recall		F1 Score	
			KNN	Adaboost	KNN	Adaboost	KNN	Adaboost
5	Testing Data	90%	100 %	100 %	100 %	100 %	100 %	100 %
		70%	100 %	100 %	100 %	100 %	100 %	100 %
		50%	100 %	100 %	100 %	100 %	100 %	100 %
	Training Data	10 %	100 %	100 %	100 %	100 %	100 %	100 %
		30%	100 %	100 %	100 %	100 %	100 %	100 %
		50%	100 %	100 %	100 %	100 %	100 %	100 %
10	Testing Data	90%	100 %	100 %	100 %	100 %	100 %	100 %
		70%	100 %	100 %	100 %	100 %	100 %	100 %
		50%	100 %	100 %	100 %	100 %	100 %	100 %
	Training Data	10 %	100 %	100 %	100 %	100 %	100 %	100 %
		30%	100 %	100 %	100 %	100 %	100 %	100 %
		50%	100 %	100 %	100 %	100 %	100 %	100 %

The experiment results in Table 5 shows that Adaboost algorithm outperforms KNN algorithm with accuracy level always greater than 99%. For the Adaboost algorithm, the number of features used and the split dataset does not affect the accuracy value on both testing data and training data, which is 100%. Whereas in KNN algorithm, the number of features used and the split dataset scenarios have impact to accuracy level. In table 4, it is clearly shown that the

difference accuracy in KNN algorithm, with 5-subset feature has accuracy value of 99.99% both for testing data and training data while with 10-subset feature provides accuracy value that varies from 99.65% to 99.97%, both on training dataset as well as testing dataset.

In Table 5 the Precision, Recall and Score F1 value all have 100% values for both KNN and Adaboost algorithms, with subsets 5 and subset 10 on any of split dataset scenarios.

V. Conclusion And Future Works

Based on the experimental results, it can be summarized that the statistical extraction feature based on the BF packet header produces 37 features and 10 best and relevant features are selected from these 37 features using Chi square algorithm for binary classification. Classification using KNN and Adaboost algorithms has high accuracy value above 99.99% both using 5-subset and 10-subset features.

For future work, we plan to use multi-class classification and use other classifiers such as Random Forest algorithm, Decision Tree, Support Vector Machine, Naive Bayes. In addition we also plan to create new dataset from more complex smarthome internet of things network testbed.

References

[1] A. Thakkar and R. Lohiya, "Attack classification using feature selection techniques: a comparative study," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, pp. 1249–1266, 2021, doi: https://doi.org/10.1007/s12652-020-02167-9.

[2] M. K. Kagita, N. Thilakarathne, T. R. Gadekallu, P. K. R. Maddikunta, and S. Singh, "A Review on Cyber Crimes on the Internet of Things." arXiv, 2020, doi: 10.48550/ARXIV.2009.05708.

[3] S. Singh, P. K. Sharma, and J. H. Park, "SH-SecNet: An Enhanced Secure Network Architecture for the Diagnosis of Security Threats in a Smart Home," *Sustainability*, vol. 9, no. 4, pp. 1–19, 2017, doi: https://doi.org/10.3390/su9040513.

[4] Z. Liu *et al.*, "Using Embedded Feature Selection and CNN for Classification on CCD-INID-V1—A New IoT Dataset," *sensors*, vol. 21, pp. 1–37, 2021, doi: https://doi.org/10.3390/s21144834.

[5] D. Stiawan, M. Y. Idris, R. F. Malik, S. Nurmaini, and R. Budiarto, "Anomaly Detection and Monitoring in Internet of Things Communication," 2016, doi: 10.1109/ICITEED.2016.7863271.

[6] N. A. Putri, D. Stiawan, A. Heryanto, T. W. Septian, L. Siregar, and R. Budiarto, "Denial of Service Attack Visualization with Clustering using K-Means Algorithm," in *International Conference on Electrical Engineering and Computer Science (ICECOS) 2017*, 2017, pp. 177–183, doi: 10.1109/ICECOS.2017.8167129.

[7] S. Costantini, G. De Gasperis, and R. Olivieri, "Digital forensics and investigations meet artificial intelligence," *Ann. Math. Artif. Intell.*, vol. 86, no. 1, pp. 193–229, 86AD, doi: 10.1007/s10472-019-09632-y.

[8] I. Goni, J. M. Gumpy, T. U. Maigari, M. Muhammad, and A. Saidu, "Cybersecurity and Cyber Forensics: Machine Learning Approach," *Mach. Learn. Res.*, vol. 5, no. 4, pp. 46–50, 2020, doi: 10.11648/j.mlr.20200504.11.

[9] G. Spanos, K. M. Giannoutakis, K. Votis, and D. Tzovaras, "Combining Statistical and Machine Learning Techniques in IoT Anomaly Detection for Smart Homes," 2019, [Online]. Available: 10.1109/CAMAD.2019.8858490.

[10] I. Cvitic, D. Perakovic, M. Perisa, and B. Gupta, "Ensemble machine learning approach for classifcation of IoT devices in smart home," *Int. J. Mach. Learn. Cybern.*, vol. 12, no. 11, pp. 3179–3202, 2021, doi: 10.1007/s13042-020-01241-0.

[11] P. N. Dawadi, D. J. Cook, M. Schmitter-Edgecombe, and C. Parsey, "Automated assessment of cognitive health using smart home technologies," *echnology Heal. Care*, vol. 21, no. 4, pp. 323–343, 2013, doi: 10.3233/THC-130734.

[12] M. S. Reza and J. Ma, "ICA and PCA integrated feature extraction for classification," in *2016 IEEE 13th International Conference on Signal Processing (ICSP)*, 2016, pp. 1083–1088, doi: 10.1109/ICSP.2016.7877996.

[13] T. Li, Z. Hong, and L. Yu, "Machine Learning-based Intrusion Detection for IoT Devices in Smart Home," in *2020 IEEE 16th International Conference on Control & Automation (ICCA)*, 2020, pp. 277–282, doi: 10.1109/ICCA51439.2020.9264406.

[14] T. D. Diwan *et al.*, "Feature Entropy Estimation (FEE) for Malicious IoT Traffic and Detection Using Machine Learning," *Mob. Inf. Syst.*, no. Distributed Secure Computing for Smart Mobile IoT Networks 2021, pp. 1–13, 2021, doi: https://doi.org/10.1155/2021/8091363.

[15] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, "Exploring feature selection and classification methods for predicting heart disease," *Digit. Heal.*, vol. 6, pp. 1–10, 2020, doi: 10.1177/2055207620914777.

[16] S. Thaseen and C. A. Kumar, "Intrusion Detection Model Using fusion of Chi-square feature selection and multi class SVM," *J. King Saud Univ. - Comput. Inf. Sci.*, 2016, doi: http://dx.doi.org/10.1016/j.jksuci.2015.12.004.

[17] S. K. Bhoi *et al.*, "FireDS-IoT: A Fire Detection System for Smart Home Based on IoT Data Analytics," in *International Conference on Information Technology (ICIT)*, 2018, pp. 161–165, doi: 10.1109/ICIT.2018.0004.

[18] S. Sachdeva. and A. Ali, "Machine learning with digital forensics for attack classification in cloud network environment," *Int. J. Syst. Assur. Eng. Manag.*, vol. 13, pp. 156–165, 2022, doi: https://doi.org/10.1007/s13198-021-01323-4.

[19] B. M. H. A. Allen, A. I. Daood, and W. H, "oster Abstract: Comparison of Classifiers for Prediction of Human Actions in a Smart Home," in *2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI)*, 2018, pp. 287–288, doi: 10.1109/IoTDI.2018.00043.

[20] T. Nugroho, M. Nasrun, and C. Setianingsih, "Smart Lamp Control Based on User Behavior For Two Lamps Using K-Nearest Neighbour," in *2019 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, 2019, pp. 123–128, doi: 10.1109/ICAMIMIA47173.2019.9223423.

[21] F. Alghayadh and D. Debnath, "A Hybrid Intrusion Detection System for Smart Home Security," in *2020 IEEE International Conference on Electro Information Technology (EIT)*, 2020, pp. 319–323, doi: 10.1109/EIT48999.2020.9208296.

[22] V. Jakkula and D. J. Cook, "Detecting Anomalous Sensor Events in Smart Home Data for Enhancing the Living Experience," in *Artificial Intelligence and Smarter Living*, 2011, pp. 33–37,

[Online]. Available: https://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/viewFile/3889/4212.

[23] M. Al-Hawawreh, E. Sitnikova, and N. Aboutorab, "X-IIoTID: A Connectivity- and Device-agnostic Intrusion Dataset for Industrial Internet of Things," *IEEE Internet Things J.*, vol. 9, no. 5, pp. 3962–3977, 2021, doi: 10.1109/JIOT.2021.3102056.

[24] A. R. Javed *et al.*, "Automated cognitive health assessment in smart homes using machine learning," *Sustain. Cities Soc.*, vol. 65, p. 102572, 2021, doi: 10.1016/j.scs.2020.102572.

[25] S. Khare and M. Totaro, "Ensemble Learning for Detecting Attacks and Anomalies in IoT Smart Home," in *2020 3rd International Conference on Data Intelligence and Security (ICDIS)*, 2020, pp. 56–63, doi: 10.1109/ICDIS50059.2020.00014.

[26] P. Singh, N. Singh, K. K. Singh, and A. Singh, "Chapter 5 - Diagnosing of disease using machine learning," in *Machine Learning and the Internet of Medical Things in Healthcare*, K. K. Singh, M. Elhoseny, A. Singh, and A. A. Elngar, Eds. Academic Press, 2021, pp. 89–111.

[27] A. Kulkarni, D. Chong, and F. A. Batarseh, "5 - Foundations of data imbalance and solutions for a data democracy," F. A. Batarseh and R. B. T.-D. D. Yang, Eds. Academic Press, 2020, pp. 83–106.

[28] Kurniabudi, B. Purnama, Sharipuddin, D. Stiawan, Darmawijoyo, and R. Budiarto, "Preprocessing and Framework for Unsupervised Anomaly Detection in IoT: Work on Progress," in *2018 International Conference on Electrical Engineering and Computer Science (ICECOS)*, 2018, pp. 345–350, doi: 10.1109/ICECOS.2018.8605231.

[29] A. Fleury, M. Vacher, and N. Noury, "SVM-Based Multimodal Classification of Activities of Daily Living in Health Smart Homes: Sensors, Algorithms, and First Experimental Results," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 274–283, 2010, doi: 10.1109/TITB.2009.2037317.

[30] D. Alghazzawi, O. Bamasag, H. Ullah, and M. Z. Asghar, "Efficient Detection of DDoS Attacks Using a Hybrid Deep Learning Model with Improved Feature Selection," *Appl. Sci.*, vol. 11, pp. 1–22, 2021, doi: https://doi.org/10.3390/app112411634.

[31] H. Hindy, C. Tachtatzis, R. Atkinson, E. Bayne, and X. Bellekens, "MQTT-IoT-IDS2020: MQTT Internet of Things Intrusion Detection Dataset." IEEE Dataport, 2020, [Online]. Available: http://10.0.82.235/bhxy-ep04.

[32] H. Hindy, E. Bayne, M. Bures, R. Atkinson, C. Tachtatzis, and X. Bellekens, "Machine Learning Based IoT Intrusion Detection System: An MQTT Case Study (MQTT-IoT-IDS2020 Dataset) BT - Selected Papers from the 12th International Networking Conference," 2021, pp. 73–84.

Feature Selection using Chi Square to Improve Attack Detection Classification in IoT Network: Work in Progress

ORIGINALITY REPORT

18%
SIMILARITY INDEX

PRIMARY SOURCES

1	M. Agus Syamsul Arifin, Deris Stiawan, Susanto, Juli Rejito, Mohd. Yazid Idris, Rahmat Budiarto. "Denial of Service Attacks Detection on SCADA Network IEC 60870-5-104 using Machine Learning", 2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2021 Crossref	117 words — 4%
2	link.springer.com Internet	71 words — 3%
3	journal.portalgaruda.org Internet	57 words — 2%
4	Deris Stiawan, Mohd. Yazid Idris, Reza Firsandaya Malik, Siti Nurmaini, Rahmat Budiarto. "Anomaly detection and monitoring in Internet of Things communication", 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), 2016 Crossref	30 words — 1%
5	www.mdpi.com Internet	29 words — 1%
6	arxiv.org Internet	20 words — 1%

7	Muhammad Fachrurrozi, Clara Fin Badillah, Saparudin, Junia Erlina, Erwin, Mardiana, Auzan Lazuardi. "The grouping of facial images using agglomerative hierarchical clustering to improve the CBIR based face recognition system", 2017 International Conference on Data and Software Engineering (ICoDSE), 2017 Crossref	19 words — 1%
8	academic-accelerator.com Internet	19 words — 1%
9	www.ijeat.org Internet	14 words — < 1%
10	www.coursehero.com Internet	13 words — < 1%
11	Keerthilatha M Pai. "Optical Screening of Oral Cancer: Technology for Emerging Markets", 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 08/2007 Crossref	12 words — < 1%
12	Muna Al-Hawawreh, Elena Sitnikova, Neda Aboutorab. "X-IIoTID: A Connectivity-and Device-agnostic Intrusion Dataset for Industrial Internet of Things", IEEE Internet of Things Journal, 2022 Crossref	12 words — < 1%
13	Shaweta Sachdeva, Aleem Ali. "Machine learning with digital forensics for attack classification in cloud network environment", International Journal of System Assurance Engineering and Management, 2021 Crossref	12 words — < 1%
14	Tania Tahmina Jui, Md. Nazmul Hoq, Suryadipta Majumdar, Md. Shohrab Hossain. "Feature	12 words — < 1%

Reduction through Data Preprocessing for Intrusion Detection in IoT Networks", 2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), 2021
Crossref

15 dokumen.pub
Internet
12 words — < 1%

16 artechjournals.com
Internet
10 words — < 1%

17 www.researchgate.net
Internet
9 words — < 1%

18 Kurniabudi, Benni Purnama, Sharipuddin, Deris Stiawan, Darmawijoyo, Rahmat Budiarto. "Preprocessing and Framework for Unsupervised Anomaly Detection in IoT: Work on Progress", 2018 International Conference on Electrical Engineering and Computer Science (ICECOS), 2018
Crossref
8 words — < 1%

19 cot.unhas.ac.id
Internet
8 words — < 1%

20 www.tjprc.org
Internet
8 words — < 1%

21 Daniyal Alghazzawi, Omaima Bamasaq, Hayat Ullah, Muhammad Zubair Asghar. "Efficient Detection of DDoS Attacks Using a Hybrid Deep Learning Model with Improved Feature Selection", Applied Sciences, 2021
Crossref
6 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF