Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 5, pp. 1021–1031
DOI : 10.47974/JSMS-1303

An intrusion detection system for industrial IoT using chi-square feature selection

Ved S. Bilaskar *
Shyam V. Aradhye †
Snehal S. Shinde §
Deepak D. Kshirsagar ‡
Pushparaj R. Nimbalkar @
Department of Computer Science and Engineering
School of Computational Sciences
COEP Technological University
Pune 411005
Maharashtra
India

Abstract

The expansion of the Industrial Internet of Things (IIoT) has led to advancements in various industries, but that has also exposed infrastructure which is critical to increasing cyber threats. In response, this paper addresses the important role of feature selection in improving the performance of Intrusion Detection Systems (IDS). Utilizing the Edge-IIoTset, this paper introduces a model which includes data pre-processing, feature selection using chi-square method, intrusion detection using machine learning and performance analysis. The experimentation reveals that the initial implementation of the PART model, using all features, yields accuracy of 97.5162% but shows a high model build-up time of 61.23 seconds. Further refinement through the chi-square feature selection method identifies a subset of the top ranked 20 features based on chi-sqauare score, achieving an enhanced accuracy of 97.5202% with a significantly reduced model build-up time of 21.54 seconds. Comparative analysis with other feature selection methods establishes the superiority of the chi-square method for the Edge-IIoT dataset.

Subject Classification: *68T99.*

Keywords: *Industrial Internet of Things, Cybersecurity, Intrusion detection system, Feature selection, Machine learning.*

* *E-mail:* **vedbilaskar21@gmail.com (Corresponding Author)**
† *E-mail:* **aradhyeshyam22@gmail.com**
§ *E-mail:* **shindesnehal0211@gmail.com**
‡ *E-mail:* **ddk.comp@coeptech.ac.in**
@ *E-mail:* **prn22.comp@coeptech.ac.in**

1. Introduction

The Internet of Things (IoT) has changed the way devices interact, connecting the physical world with the digital realm. This network of devices, sensors, and systems which is interconnected has revolutionized various sectors, including transportation, agriculture [1], industrial processes [2,3], and healthcare [4]. According to a report [5], active IoT devices till 2022 is approximately 11.57 billion. It is projected to increase to 25.44 billion till 2030. As IoT continues to grow, its industrial kind, known as IIoT, stands out for its impact on smart manufacturing. The use of IIoT has revolutionized industrial processes, offering connectivity and efficiency. IIoT seamlessly integrates physical devices, sensors, and industrial systems, giving rise to a new era of connectivity and efficiency in manufacturing processes. However, this interconnected landscape is not safe from cyber threats, as the growing interdependence of devices and networks introduces vulnerabilities that malicious actors can easily exploit. With these challenges in mind, the implementation of robust cybersecurity measures, particularly Intrusion Detection Systems (IDS), becomes important to safeguard the integrity, availability, and confidentiality of industrial infrastructure.

According to a statistical report [6], in February 2023, senior executives of UK-based organizations participated in a survey, and 20 percent of them predicted a notable rise in cyberattacks targeting IIoT. To effectively counter against these cyber threats, it is of utmost importance to employ advanced attack detection mechanisms accustomed to the unique characteristics of IIoT ecosystems. IDS is the first line of defense by identifying and responding to anomalous activities within the network. However, the sheer volume and complexity of data generated in IIoT environments pose a challenge in accurately detecting malicious behavior. As a result, accurate and effective feature selection (FS) algorithms are required to improve IDS performance by locating the most appropriate and discriminative characteristics within massive data sets related to IIoT network traffic. Network intrusion detection systems (NIDS) [7], make it possible to recognize and stop online threats. Workstations can be protected from multiple grid breaches by using NIDS to identify and address network activity.

This research paper seeks to address the critical importance of FS in the context of IDS for IIoT. This research explores the significance of FS in increasing the accuracy and decreasing model build-up time of intrusion

detection, ultimately contributing to the resilience of IIoT environments against cyber threats.

The remainder of this paper is structured as follows: Literature Review conducted is discussed in Section 2. Proposal of the model being used in the experimentation is provided in Section 3. In Section 4, the actual experiment is described in detail along with an analysis of the outcomes. The suggested model is contrasted with other models used in related publications in Section 5. Conclusion is stated in Section 6.

2. Related Work

In [7], a deep feed forward neural network (DFFNN) combined with a genetic search method and rules-based algorithm is proposed for feature extraction. When tested on various samples from NSL-KDD and UNSW-NB15 datasets, it achieved 98% detection rate with FAR of 0.9% and detection rate of 98.9% with FAR 1% respectively.

The system introduced in [8] used Chi-square (Chi) and a modified Decision Tree (MDT) for intrusion detection in SCADA networks and Industrial IoT. The system selected five top-ranked features of Chi and was tested on ICS-SCADA, WUSTL-IIoT-2018, CICDS2018, and NSL-KDD. The MDT achieved higher accuracy of 99.98% on the WUSTL-IIoT-2018 dataset compared to others.

The proposed framework in [9] for cyberattacks in IIoT used an ensemble FS method. Mutual Information (MI), PCC, ANOVA, and Chi were combined for ensemble FS, and two reduced sets were created by performing union and intersection operation between the features. The sets were then tested on the Edge-IIoT-2022 dataset with RF, DT, CTB, and XGB. The framework achieved higher accuracy of 99.61 with XGB using 37 features (union) and 97.84 with XGB using 22 features (intersection).

IDS for IIoT was proposed in [10] and was evaluated on the UNSW-NB15. It implemented a Genetic Algorithm combined with the RF model for FS. It generated feature vectors Vb, including 10 feature vectors for binary classification, and Vm, including 7 feature vectors for multiclass classification. For binary classification, Logistic Regression (LR) was used as a baseline model, and XGB, ET, RF, and DT were applied. Similarly, for Vm, Naive Bayes (NB) was the baseline model, and the same models were applied. The GA-RF achieved accuracy of 87.61% using 16 features for binary and 77.64% using 17 features for multiclass classification.

The work [11] solely focused on applying the maximum correlation minimum redundancy (MRMR) algorithm with the support vector

machine (SVM) on the UNSW-NB15 dataset, which has 42 features in total, and the MSU dataset, which has 26 features. It then compared the experimental results on both datasets to draw their conclusion that the combination or independence of attributes has a great impact on the classification of attacks. On UNSW-NB15, a maximum detection accuracy of 69.96% was observed when 19 features were used. On MSU, a detection accuracy of 95.67% was achieved with 21 features used.

In [12], a system was proposed which utilized the WUSTL-IIOT-2021 dataset. Recursive feature elimination (RFE) was applied to reduce the 41 features of the dataset to 11. RFE relies on a simple recursive method to terminate the feature which has the minimal importance while maintaining the performance of the model. DT, RF, LR, GNB along with 10-fold cross-validation is used. The reduced dataset results in increased efficiency and maintains the accuracy of 99.97% for DT with a prediction time of 0.1517 μs.

The research [13] proposes an IDS using RF model. Pearson's Correlation Coefficient (PCC) was applied. The model was validated on the WUSTL-IIOT-2021 and BoT-IoT. Matthew's correlation coefficient (MCC) was used to study the impact of their model on the Bot-IoT dataset which is imbalanced. It showed 99% accuracy for BoT-IoT and 99.27% accuracy for the other dataset.

3. Proposed Model

The architecture of the proposed model is shown in Figure 1.

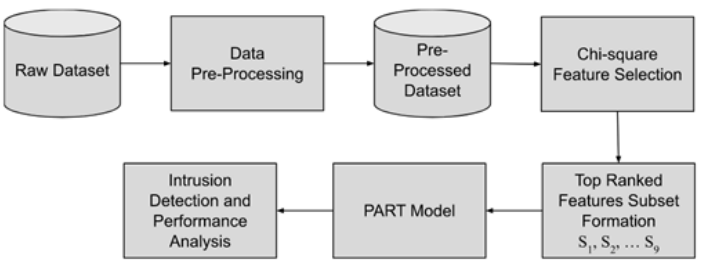

Figure 1

Proposed model architecture for IDS in IIoT

3.1 *Data Pre-Processing*

Preprocessing the raw dataset for Machine Learning involves removal of unnecessary features, handling missing values and duplicates,

converting values to appropriate forms, encoding nominal and ordinal features, and optionally replacing values. This ensures data is suitable for ML models, addressing issues like bias, noise, and accuracy. The target column may also need conversion based on ML model requirements. After preprocessing a pre-processed dataset is obtained, this is used for further experimentation.

3.2 *Chi-square Feature Selection*

Feature selection (FS) is a significant step in ML, aiming to improve model performance, interpretability, and computational efficiency. It aims to retain relevant attributes while discarding redundant ones. Unnecessary features can lead to issues such as the curse of dimensionality, information abundance but knowledge shortage [14], decreased accuracy, and increased time and space complexity. Based on the features evaluated using FS, it's important to select a certain number of features. The aim is to reduce their quantity while ensuring accuracy remains high and Model Build-up Time (MBT) decreases. This is accomplished by utilizing techniques like top-ranking. These methods help streamline the FS process and improve the model performance. In this study, Chi is used.

3.3 *PART Model*

This step is fundamental in developing an IDS. ML models serve as the core detectors, classifying or predicting new traffic as either legitimate or malicious based on previous training data. To assess the efficiency and reliability of the selected features in FS, they undergo testing using ML models. The experiment makes use of the PART ML model.

3.4 *Intrusion Detection and Performance Analysis*

The analysis, employing ML models with FS, focuses on evaluating how FS techniques affect the dataset. Important factors to consider during this assessment include increasing or at least maintaining model accuracy, decreasing the time required for detection, and selecting a low number of features. Other important parameters like Precision, F1-score, Recall, False Alarm Rate (FAR), are calculated.

Based on these factors, a model is proposed to select a FS technique. The aim is to enhance accuracy, reduce detection time, and utilize fewer features. This framework aims to improve IDS performance by leveraging the benefits of FS in the ML analysis process.

4. Experimentation and Result Analysis

The proposed model was implemented on a machine with macOS installed, featuring a 1.8 GHz Dual-Core Intel Core i5 processor and 8 GB of 1600 MHz DDR3 memory. Weka 3.8.6, Python 3, and Google Colab were the tools used.

4.1 *Dataset - Edge-IIoTset [15]*

The organized testbed utilized for data collection comprised seven layers like Blockchain, Edge computing, IoT & IIoT perception layer, etc. Various types of IoT devices were employed, such as digital sensors, ultrasonic sensors, and pH sensor meters. Multiple sources were analyzed to have flow features from network packets, like logs, alerts and network traffic. Out of the 1176 features found, 61 features having high correlations are proposed. The dataset contains 14 attacks categorized into five major categories - DoS/DDoS, Injection, Man In The Middle (MITM), Information Gathering and Malware, in addition to normal traffic

4.2 *Data Pre-Processing*

The raw data consisted of 157800 entries with 61 features and two label columns. Firstly, unnecessary features like ip.dst_host, frame.time, ip.src_host, http.file_data, icmp.transmit_timestamp, etc as specified in [15] were eliminated after which 46 features were obtained. Next missing values were checked, and any duplicate entries were removed. The format of values was standardized, with hexadecimal numbers converted to their appropriate form for features expecting integer values. Additionally, a few entries consisting of data in inappropriate form were dropped. Nominal features were one-hot encoding, while ordinal features were label encoded. Lastly, the target column was converted from ordinal to nominal format to align with the requirements of the Weka 3.8.6 tool used for analysis. Finally, 152028 entries having 48 features and two label columns, for binary and multiclass classification were obtained.

4.3 *Chi-square Feature Selection and PART Model*

The pseudocode for the proposed feature selection model is as shown in Figure 2. Initially, the dataset containing all the 48 features S was tested using the PART model [16], known for its effectiveness with the specific dataset being used. Key parameters such as Accuracy and Model Build-up Time (MBT) were noted, whereas parameters such as Precision, F1-score,

Recall , FAR were calculated. The goal was to minimum features while maintaining or improving accuracy and reducing MBT. Next, the Chi, coupled with a ranker algorithm, was applied to the dataset. Relevant feature subsets containing top ranked features based on Chi score, in multiple of five (S_1, S_2, … S_9) were formed. PART was applied on each of these subsets and their accuracy and MBT were recorded as given in Table 1.

Pseudocode: Proposed Feature Selection model

Input: Redundant feature set S = {F_1, F_2, …., F_{48}}

Output: Relevant feature set S_R = {f_1, f_2, …., f_r}

1. Begin
2. To measure accuracy and MBT of S using PART, i.e., Accuracy(S) and MBT(S)
3. Apply Chi on S
4. Formation of Relevant feature subsets.
   S_1= Top Ranked 5 features based Chi score
   S_2= Top Ranked 10 features based Chi score
   :
   S_9 = Top Ranked 45 features based Chi score
5. for S_i in { S_1, S_2, … S_9 } :
6. Measure accuracy and MBT of S_i using PART, i.e., Accuracy(S_i) and MBT(S_i)
7. Select S_i with min features && Accuracy(S_i) ≥ Accuracy(S) && MBT(S_i) < MBT(S)
8. **En**d

Figure 2

Pseudocode for the proposed feature selection model

4.4 *Performance Analysis*

When using PART with all features, an accuracy of 97.5162% was achieved, but the MBT was high at 63.73 seconds. To reduce MBT and potentially improve accuracy, further experimentation was conducted. After application of PART on feature subsets S_1, S_2, … S_9, subset S_4 with top 20 features achieved the highest accuracy of 97.5202% and a reduced MBT of 21.54 seconds. Its comparison with other subsets is as shown in Table 1.

Table 1

Comparison of PART on various top ranked feature subsets

Feature Subset	Accuracy (%)	Precision (%)	F1 score (%)	Recall (%)	FAR	MBT (seconds)
S_1	96.6789	99.83255717	88.3143935	79.1791168	0.00025013288310	4.5
S_2	97.5156	99.91647423	91.5026209	84.3957503	0.00013288309410	12.57
S_3	97.5136	99.93118364	91.4941494	84.3708499	0.00010943313640	17.03
S_4	97.5202	99.97541306	91.5151242	84.3700000	0.00003908326298	21.54
S_5	97.5195	99.98032755	91.5123005	84.3666998	0.00003126661039	27.49
S_6	97.5162	99.96557490	91.5012379	84.3583997	0.00005471656818	39.19
S_7	97.5162	99.96557490	91.5012379	84.3583997	0.00005471656818	44.15
S_8	97.5162	99.97540580	91.5004727	84.3500996	0.00003908326298	49.73
S_9	97.5162	99.99015942	91.4993246	84.3376494	0.00001563330519	56.58
All features	97.5162	99.99015942	91.4993246	84.3376494	0.00001563330519	61.23

5. Comparison

To compare the efficacy of Chi, other filter-based feature selection methods used in traditional IDSs like Symmetrical Uncertainty used in [17], Correlation [9], Information Gain [18], and Gain Ratio [18], coupled with the ranker algorithm, were employed. Their top ranked 20 features were tested using PART, as Chi gave best results with top 20 features, with the results as shown in Table 2. It was observed that Chi achieved the highest accuracy of 97.5202% among all methods. Correlation had a shorter MBT of 14.04 seconds compared to Chi's 21.54 seconds, but it achieved a lower accuracy of 96.6704%. Therefore, it can be concluded that Chi is the best-suited FS method, with its top-ranked 20 features, for the Edge-IIoT dataset.

Table 2

Comparison of Chi to other FS methods

Feature Selection Method	Accuracy (%)	Precision (%)	F1 score (%)	Recall (%)	FAR	MBT (seconds)
Symmetrical Uncertainty	97.5103	99.92625731	91.4823232	84.3542496	0.00011724978900	24.26
Correlation	96.6704	99.91084540	88.2726346	79.0629150	0.00013288309410	14.04

Contd...

Chi Square	97.5202	99.97541306	91.5151242	84.3700000	0.00003908326298	21.54
Information Gain	97.5097	99.91153052	91.4810314	84.3625498	0.00014069974670	21.72
Gain Ratio	97.5143	99.95574133	91.4946771	84.3542496	0.00007034987337	24.55

6. Conclusion

This paper proposes a filter-based feature selection for the Edge-IIoT dataset. The pipeline includes preprocessing, feature selection and ML. The original dataset, with all features, achieved an accuracy of 97.5162% and a model build-up time of 63.73 seconds when tested using the PART. Hence, the chi-square, coupled with the ranker algorithm, was utilized. Features were selected using a top-ranked approach, with subset sizes in multiples of five. After analysis, the subset with the top ranked 20 features demonstrated the highest accuracy of 97.5202% with a model build-up time of 21.54 seconds. This was compared with other traditional feature selection methods such as Symmetrical Uncertainty, Correlation, Information Gain, and Gain Ratio, using their top-ranked 20 features. The chi-square method outperformed all of them, proving to be the best feature selection method for the Edge-IIoT dataset.

In future, wrapper, embedded, and evolutionary methods will be studied, and feature selection will be employed for intrusion detection in IIoT.

References

[1] Sharma, Pankaj, et al. “Sustainable Agriculture: An IoT based case-study.” *TARU Journal of Sustainable Technologies and Computing*, vol. 1, no. 1, pp. 45-52 (2019). DOI: 10.47974/TJSTC.014.2019.v01i01.

[2] Sharma, Raj Kumar, and Jailia, Manisha. “Machine Learning and IoT-based Garbage Detection System for Smart Cities.” *Journal of Information and Optimization Sciences*, vol. 44, no. 3, pp. 393–406 (2023). DOI: 10.47974/JIOS-1349.

[3] Katyare, Poonam, et al. “Real-time Data Modeling for Forecasting Fuel Consumption of Construction Equipment Using Integral Approach of IoT and ML Techniques.” *Journal of Information and Opti-*

mization Sciences, vol. 44, no. 3, pp. 427–437 (2023). DOI: 10.47974/JIOS-1363.

[4] Andrews, Leo John Baptist, et al. "Application of Internet of Things - A Smart Healthcare Concept for Botswana." *TARU Journal of Sustainable Technologies and Computing*, vol. 1, no. 2, pp. 93-103 (2019). DOI: 10.47974/2019.TJSTC.004.

[5] Vailshery, L.S. "Number of IoT Connected Devices Worldwide 2019-2030." Statista, 2023, www.statista.com/statistics/1183457/iot-connected-devices-worldwide/.

[6] Petrosyan, Ani. "Share of Companies in the United Kingdom (UK) Expecting Significant Increase in Cybercrime within a Year as of February 2023, by Type of Attack." Statista, 2023, www.statista.com/statistics/1426517/cyberattacks-uk-companies-expected-increase-within-a-year-by-type/.

[7] Potnurwar, Archana V., et al. "Deep Learning-Based Rule-Based Feature Selection for Intrusion Detection in Industrial Internet of Things Networks." (2023).

[8] Ahakonye, Love Allen Chijioke, et al. "SCADA Intrusion Detection Scheme Exploiting the Fusion of Modified Decision Tree and Chi-square Feature Selection." (2023).

[9] Arya, Laxmi, and Govind P. Gupta. "Ensemble Filter-based Feature Selection Model for Cyber Attack Detection in Industrial Internet of Things." (2023).

[10] Kasongo, Sydney Mambwe. "An Advanced Intrusion Detection System for IIoT Based on GA and Tree Based Algorithms." (2021).

[11] Zhang, Xueying, et al. "Research on Feature Selection for Cyber Attack Detection in Industrial Internet of Things."

[12] Alani, Mohammed M. "An Explainable Efficient Flow-based Industrial IoT Intrusion Detection System." (2023).

[13] Mohy-eddine, Mouaad, et al. "An Effective Intrusion Detection Approach Based on Ensemble Learning for IIoT Edge Computing." (2022).

[14] Li, J.-Q., et al. "Industrial Internet: A Survey on the Enabling Technologies, Applications, and Challenges." *IEEE Communications Surveys & Tutorials,* vol. 19, no. 3, pp. 1504–1526 (2017).

[15] Ferrag, Mohamed Amine, et al. "Edge-IIoTset: A New Comprehensive Realistic Cyber Security Dataset of IoT and IIoT Applications for Centralized and Federated Learning." TechRxiv (2022), https://doi.org/10.36227/techrxiv.18857336.v1.

[16] Al Nuaimi, Taraf, et al. "A Comparative Evaluation of Intrusion Detection Systems on the Edge-IIoT-2022 Dataset." (2023).

[17] Altameem, Ayman, et al. "P-ROCK: a sustainable clustering algorithm for large categorical datasets." *Intell. Autom. Soft Comput* 35.1 : 553-566 (2023).

[18] K, Albulayhi, et al. "IoT Intrusion Detection Using Machine Learning with a Novel High Performing Feature Selection Method." *Applied Sciences*, vol. 12, no. 10 (2022), https://doi.org/10.3390/app12105015.

Received May, 2024