Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 218 (2023) 318–327

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

# Hybrid Meta-Heuristic based Feature Selection Mechanism for Cyber-Attack Detection in IoT-enabled Networks

Arun Kumar Dey[a], Govind P. Gupta[a], Satya Prakash Sahu[a]

[a]Department of Information Technology, National Institttue of Technology, Raipur 492010, C. G., India

## Abstract

Today's technologically advanced connected world is mostly reliant on the Internet of Things (*IoT*)-enabled smart gadgets and easy connectivity. These smart gadgets are more susceptible to malicious practices found in network traffic, which is one of the biggest challenges in cyber security domain. As a result, many systems and end-users are adversely affected by this practice. However, intrusion detection systems (*IDS*) are often applied to guard against cyber-attacks. Since, *IDS* plays a key role in detecting and preventing cyber-attacks in IoT-enabled networks, but design of an efficient and fast IDS system for cyber-attack detection is still a challenging research issue. Moreover, *IDS* datasets contain multiple features and to design an efficient and fast IDS, feature selection (*FS*) is an essential mechanism to remove the irrelevant and redundant features from large *IDS* datasets. Thus, this paper has proposed a hybrid feature selection scheme in which statistical test-based filter approaches such as Chi-Square $(\chi^2)$, Pearson's Correlation Coefficient (*PCC*), and Mutual Information (*MI*) are combined with a Non-Dominated Sorting Genetic Algorithm (*NSGA-II*)-based metaheuristic approach for optimization of features. In the proposed scheme, filter-based methods are employed to rank the features for guided population initialization in *NSGA-II* for faster convergence towards a solution. Performance evaluation of the proposed scheme is evaluated using the ToN-IoT dataset in terms of number of selected features and accuracy. Experimental outcomes are compared with some latest state-of-art techniques. Result analysis confirms the superior performance of the proposed scheme with minimum number of optimized features (only 13 out of 43 features) and maximum accuracy (99.48%).

*Keywords*: Meta-Heuristic; Cyber Security; Internet of Things (IoT); Intrusion Detection; Feature Selection

## 1. Introduction

Today's world requires easy connectivity and smart devices that are associated to the Internet of things (*IoT*). *IoT*

* Corresponding author. Tel.: +91-9891952480
*E-mail address:* gpgupta.it@nitrr.ac.in

10.1016/j.procs.2023.01.014

is an embedded system having sensing abilities [1]. These smart devices are more susceptible to cyber-attacks. As an example, businesses are commonly targeted by Distributed Denial of Service (*DDoS*) attacks [2] that can be used for hacking multiple systems simultaneously. Penta Security estimates that cybercrime damages will exceed *USD* 10 trillion by 2025 [2]. Although, the advent of technologies in the domain of cyber-security, cyber-attacks are often protected with intrusion detection systems (*IDS*). As *IDS* are one of the most productive tools existing in the literature [3]. *IDS* is a software program designed to monitor the malicious activities of the network traffic [4]. *IDS* is mostly classified into two groups [5] (i) signature-based (*sIDS*) and (ii) anomaly-based (*aIDS*) [5]. In *sIDS*, pattern of bytes in network packets are deeply inspected, but it is not capable to identify the new attacks. While *aIDS* is well suited to identify known as well unknown attacks. An efficient *IDS* is necessary for enhancing the security of *IoT* networks. Thus, this paper focus on design of an efficient *IDS* based on hybrid feature selection approach.

Intrusion detection datasets contain big volumes of data that has a high degree of dimension (features). So, FS is crucial to overcoming the "curse of dimensionality" [6]. *FS* is the process of selecting the most pertinent features to constructing a robust framework [7]. *FS* technique is classified into three common forms viz. filter-based method, wrapper-based method, and embedded-based methods [8], [9]. In the filter-based method, every feature is compared with the target features, which are then ranked using statistical techniques [10]. While, in the wrapper method, wrapping more than one features to create subset of features, then checks with the target feature to find which subset will provide higher accuracy. However, wrapper method is computationally expensive than filter method [10][11]. In the embedded-based method, features are selected based on learning process of a particular machine learning technique [9]. Metaheuristic-based techniques are mostly used in optimization of features that are relevant for prediction in *IDS*, especially because of their high degree of accuracy [12]. In this context, evolutionary algorithms (*EA*) based multi-objective optimization methods are highly valuable since it is more efficient than single-objective optimization methods [13]. A multi-objective optimization hurdle can be solved by selecting the most optimal solution from the Pareto Front (*PF*), also known as trade-off solutions. When using multi-objective *EA*-based methods, the entire *PF* is calculated in one simulation run, after evaluating the population of solutions [13].

In this research work, a hybrid feature selection mechanism is proposed to get optimized subset of feature for cyber-attack detection. This paper makes the following contributions:

1. A hybrid metaheuristic-based *FS* mechanism is proposed using statistical test-based filter methods and *NSGA-II*-based metaheuristic technique.
2. In the proposed scheme, an initial population in *NSGA-II* is guided by three feature ranking-based filter methods for faster convergence towards a solution.
3. Performance of the proposed scheme is evaluated in terms of number of selected features and accuracy using the benchmark IDS dataset such as ToN-IoT.

The other section of the article is arranged as follows: related works based on *FS* are outlined in Section 2. Preliminary studies of statistical test-based filter methods and metaheuristic technique is discussed in Section 3. The proposed work for metaheuristic-based hybrid *FS* is presented in Section 4. Performance evaluation and Result analysis are discussed in section 5. This paper is concluded in Section 6.

## 2. Related Work

To ensure network security, researchers are continuously focusing on improving the cyber-attack detection frameworks for timely detection of new attacks. In this study, we have reviewed some important literature related to cyber-attack detection and *IDS,* as summarized in Table 1.

In recent years, several hybrid methods using ensemble feature selection and classification have been developed to upgrade the work of *IDS*. Hajisalem *et al.* [5] developed a hybrid technique called ABC-AFS for intrusion detection. Furthermore, Fuzzy C-Means clustering is used to identify different subsets of the dataset after a correlation-based FS selects a relevant feature with Weka (data mining tool). Simulation results of the ABC-AFS model shows that accuracy (99.0%) and optimized feature subset (6 out of 41 features) for NSL-KDD. Limitation of this work is that dataset is old and does not contain advanced IoT-based attacks. Kumar *et al.* [14] proposed a framework which combines the result of the gain ratio, correlation coefficient, and random forest mean decrease accuracy as an ensemble technique for feature selection. Optimized features set obtained by filter methods. Further, optimized feature set is classified by Random Forest (*RF*), *KNN*, and XGBoost respectively to get different classification results. Experimental result shows that XGBoost outperforms the random forest and *KNN* in terms of 99.34% of accuracy and relevant feature subset is 18 out of 41 features using NSL-KDD. For DS2OS, accuracy is 99.43% and selected feature subset is 6 out of 12 features. The shortcomings of this approach are reduced features set can be further optimized by meta heuristic approach and modern *IDS* dataset is not considered in this study.

In 2020, Nazir *et al.* [3] introduced another way to choose attributes for cyber-attack detection in network traffic. The proposed procedure presents a new feature selection using meta heuristic approach named Tabu Search- Random Forest (*TS-RF*) that used to solve combinatorial problem. The proposed method decreases the time complexity as well as the number of features from 43 to only 16 features for UNSW-NB-15 dataset and achieved better accuracy (83.12%). Class imbalance problem of UNSW-NB 15 dataset is not handled in this technique. In 2020, Roopak *et al.* [15] introduced a hybrid technique to find out DDoS attack. In this model deep learning technique and multi-objective optimization technique such as *NSGA-II* are merged to design *IDS*. Features are optimized by *NSGA-II* then CNN and LSTM techniques are applied to detect intrusions. The model is evaluated by *IDS* dataset CISIDS2017 and achieves accuracy of 99.03%. The limitation of this work is that number of selected features are not discussed in this work.

Alazzam *et al.* [4], have introduced a *FS* technique is handled by cosine similarity. The technique known as Cosine_PIO is developed to obtain optimized subset of features. Cosine_PIO produces accuracy (96.0%) and optimized feature subset (7 out of 41 features) for KDDCup99 dataset. For UNSW-NB15 dataset, accuracy (91.7%) and selected feature subset (5 out of 49 features). Accuracy (88.3%) and feature subset (5 out of 41 featues) using NSL-KDD dataset. The shortcoming of Cosine_PIO is that imbalanced class problem in dataset is not addressed. SaiSindhuTheja *et al.* [16], developed a new approach based on the Oppositional Crow Search Algorithm (*OCSA*) for selecting features for *IDS*. *OCSA* is a population-based technique that uses Recurrent Neural Network (*RNN*) for classification. With this algorithm, it is able to achieve the maximum amount of diversity in the produced solution. This experiment considered binary classification and the result showed that the model achieved accuracy (94.12%) and optimised feature subsets (10 out of 41) using KDDCup99. It has the disadvantage that data imbalanced problem is not handled.

Another hybrid technique called TP2SF used blockchain-IPFS combined Fog–Cloud for privacy preserving in *IoT* network is proposed by Kumar *et al.* [17]. TP2SF is designed in three components such as trustworthy component, privacy component, and IDS component. TP2SF is evaluated by ToN-IoT dataset to protect *IoT* based smart cities. Gad *et al.* [18] anticipated a machine learning based *IDS* to detect intrusion in vehicular ad-hoc networks. Author used Chi-Square method for feature selection and SMOTE technique to handle data imbalance problem. The model is evaluated by ToN-IoT dataset and achieves highest accuracy (99.10%) using XG-Boost technique. In 2022, Bacha et al. [19] proposed KPCA-KELM techniques to detect anomalies in N-BaIoT dataset. KPCA is applied to get best features and KLEM is used for classification. Rresult showed that the model achieved accuracy (99.40%). Limitation of this work is that number of selected features are not discussed by the authors.

Table 1: Summary of related work based on feature selection techniques for *IDS* datasets.

| References | Year | Technique | Hybrid | Datasets | Selected Features | Classifier | Accuracy (%) |
|---|---|---|---|---|---|---|---|
| Hajisalem *et al.* [5] | 2018 | ABC-AFS | ✓ | NSL-KDD | 06/41 | CART | 99.00 |
| Kumar *et al.* [14] | 2019 | ICADS | ✓ | NSL-KDD | 18/41 | RF, XG-Boost, KNN | 99.34 |
| Nazir *et al.* [3] | 2020 | TS-RF | ✓ | UNSW-NB15 | 16/43 | RF | 83.12 |
| Roopak *et al.* [15] | 2020 | DL+NSAG-II | ✓ | CISIDS2017 | NA | CNN, LSTM | 99.03 |
| Alazzam *et al.* [4] | 2020 | Cosine_PIO | ✕ | KDDCup99 | 07/41 | Decision Tree | 96.00 |
| SaiSindhuTheja *et al.* [16] | 2020 | OCSA | ✓ | KDDCup99 | 10/41 | RNN | 94.12 |
| Kumar *et al.* [17] | 2021 | TP2SF | ✓ | ToN-IoT | 19/43 | XG-Boost | 98.84 |
| Gad *et al.* [18] | 2021 | Chi2- SMOTE | ✕ | ToN-IoT | 20/43 | XG-Boost | 99.10 |
| Bacha *et al.* [19] | 2022 | KPCA-KLEM | ✕ | N-BaIoT | NA | KLEM | 99.40 |

**Abbreviations**: NA denotes that no information is available.

## 3. Preliminary Concepts

This section provides a preliminary explanation of the ranking-based filters methods such as Chi-Square Test, Pearson's Correlational Coefficient (*PCC*), and Mutual Information (*MI*). In addition, a brief description of the *NSGA-II-based* meta-heuristic technique are discussed.

### *3.1 Chi- Square ($\chi^2$)*

$\chi^2$ test is applied for testing of hypothesis on population of variance. In $\chi^2$, the features are ranked according to their level of independence amid the all-independent features '$I_{f_i}$' and the target feature '$D_{f_j}$' and comparing the Chi distribution to a one-degree-of-freedom distribution [20]. Eq. (1) defines the computation of $\chi^2$.

$$\chi^2\left(I_{f_i}, D_{f_j}\right) = \frac{r*\,(p*q-s*t)^2}{(p+t)*(p+q)*(t+q)*(s+q)} \tag{1}$$

Here $p$ is frequency of $I_{f_i}$ and $D_{f_j}$ in the dataset, $t$ is frequency of $I_{f_i}$ appearing without $D_{f_j}$, $s$ is frequency of $D_{f_j}$ appearing without $I_{f_i}$ and $q$ is frequency of neither $D_{f_j}$ nor $I_{f_i}$ appearing with each other in the dataset, $r$ represent total count of features, $i = 1,2, \ldots\ldots, 43$ features, and $j = 0, 1$ (target class). The higher value of $\chi^2\left(I_{f_i}, D_{f_j}\right)$ indicates higher importance feature [13].

### *3.2 Pearson's Correlational Coefficient (PCC)*

The Pearson's Correlation Coefficient ($PCC$ ) measures the relationship between independent features, say $I_f = \{f_1, f_2, f, \cdots\cdots\ \cdots\cdots, f_{n-1}\}$ and dependent or target feature, say $D_f = \{f_n\}$. It computes the linear relationship between $I_f$ and $D_f$, lies between $-1$ to $+1$ [14]. If $I_f$ and $D_f$ are dependent then the value of $PCC = \pm 1$, and if $I_f$ and $D_f$ are independent then $PCC = 0$. $PCC\ (I_f, D_f)$ is calculated in Eq. (2).

$$PCC(I_{f,}D_f) = \frac{cov(I_f, D_f)}{\sigma_{I_f}\sigma_{D_f}} \tag{2}$$

where $cov(I_f, D_f)$ defines the covariance between $I_f$ and $D_f$, $\sigma_{I_f}$ ,$\sigma_{D_f}$ are the standard deviation of $I_f$ and $D_f$ respectively. A lesser value of $PCC(I_{f,}D_f)$ indicates that $I_f$ and $D_f$ are strongly correlated [13].

### *3.3 Mutual Information (MI)*

*MI* is a measure of "mutual dependence" between two arbitrary features such as independent feature and dependent feature [10]. In machine learning, *MI* measures the amount of information that a feature provides to a correct prediction on a target variable. In mathematical terms, the *MI* between independent and dependent variable is computed by Eq. (3).

$$I(x; y) = e(x) - e(x|y) \tag{3}$$

Here $I(x; y)$ is the mutual information between independent feature $(x)$ and dependent feature $(y)$, $e(x)$ is entropy of $x$ and $e(x|y)$ represents the conditional entropy of $x$ for given $y$.

### *3.4 Non-dominated Sorting Genetic Algorithm -II (NSGA-II)*

*NSGA-II* is a metaheuristic technique that belongs to the evolutionary algorithms (*EA*) class that uses non-dominated sorting as well as distance of crowding to improve multi-objective performance [21]. Moreover, it is also ideal as a feature selection tool [22]. Genetic manipulator of this algorithm are selection, crossover, and mutation. First, in the selection process, two parents are selected, and a new population is created by using crossover and mutation. Second, in a crossover operation, two parents exchange their genetic information for offspring [23]. Then, the one-point crossover method is applied to derive a child chromosome who has a high crossover probability ($p_{crossover}$). There is an arbitrary point where two parents are fragmented in half. New children will inherit the first half of their inheritable factor from the first biological parent and the other half from the next biological parent. After that, using the mutation operator, the child is more likely to escape from the local optimum with mutation probability ($p_{mutation}$) [24]. Further, an individual's gene is altered with a one-bit flip mutation technique. After that, the population is then arranged based on two aspects: non-dominated rank and density of solution (crowding distance) [11]. First, an initial rank index is assigned to each individual, based on its level of non-domination. Second, crowding distance operator is used between two chromosomes, if they have equal pareto front that is same non-domination rank. As a result, populations are sorted from top to bottom without losing reasonable solutions. Then, to create a new generation, the best solutions are moved to a mating pool. Each step is repeated until a stopping criterion is met. It returns non-dominated pareto front solutions that approximate the best. A major advantage of *NSGA-II* is its low mathematical complexity of $O(mn^2)$ where $m$ is total count of objectives and $n$ is population size [21].

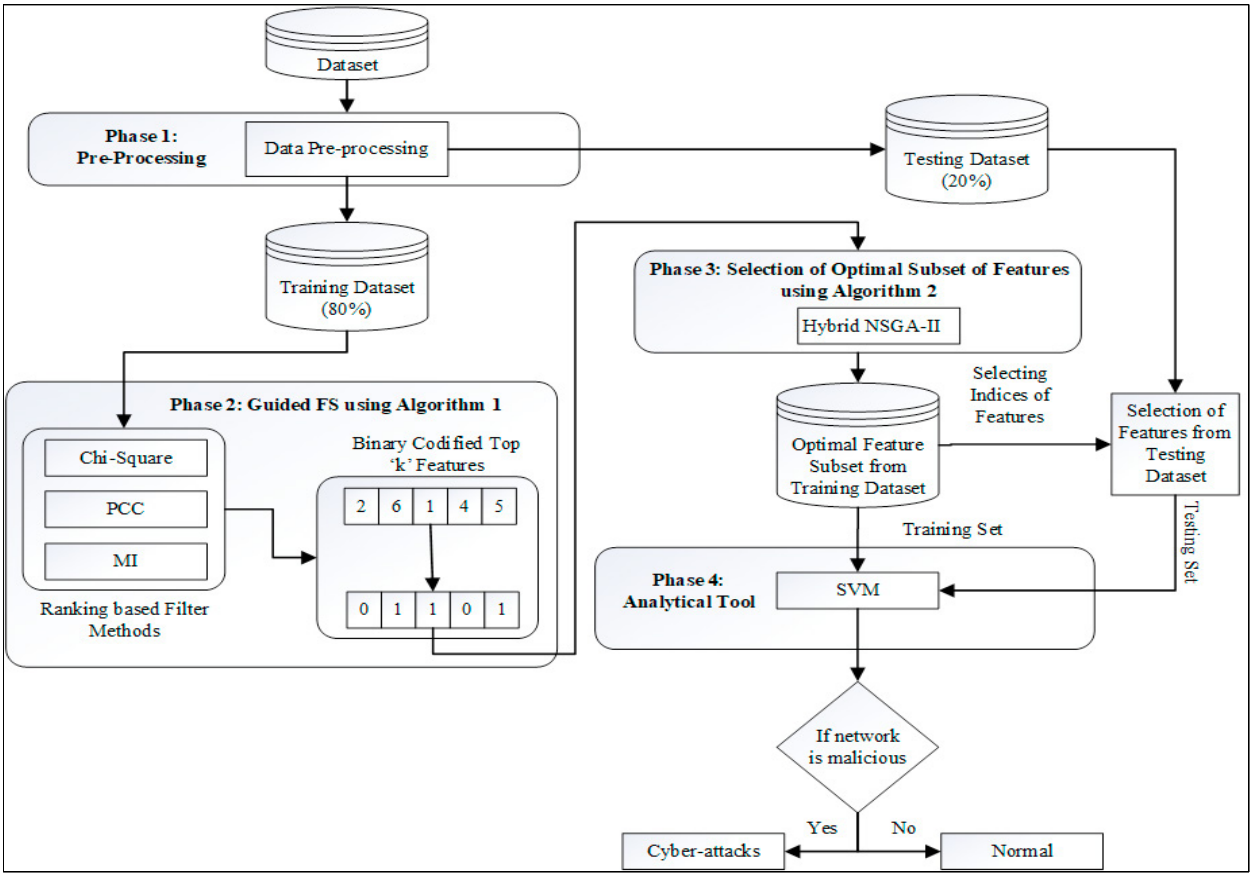

Dataset
Phase 1: Pre-Processing
Data Pre-processing
Testing Dataset (20%)
Training Dataset (80%)
Phase 3: Selection of Optimal Subset of Features using Algorithm 2
Hybrid NSGA-II
Selecting Indices of Features
Optimal Feature Subset from Training Dataset
Selection of Features from Testing Dataset
Phase 2: Guided FS using Algorithm 1
Chi-Square
PCC
MI
Ranking based Filter Methods
Binary Codified Top 'k' Features
2 6 1 4 5
0 1 1 0 1
Training Set
Testing Set
Phase 4: Analytical Tool
SVM
If network is malicious
Yes
No
Cyber-attacks
Normal

Fig. 1. The overall workflow of the proposed scheme for the detection of cyber-attacks.

## 4. Proposed Hybrid Scheme for *FS* in Cyber-attack Detection

In this section, a hybrid scheme for *FS* for designing of an efficient cyber-attack detection model has been proposed wherein a guided initial population sample for the *NSGA-II* process is generated using ranking based filter methods such as Chi-Square, *PCC*, and *MI*. Fig. 1 shows the overall workflow of the proposed scheme for cyber-attack detection. The model consists of four phases. As a first phase, complete *IDS* dataset is preprocessed. In the second phase, three filter methods are applied to rank top ten features in the preprocess dataset. In the third phase, optimal subset of features is selected using hybrid *NSGA-II* algorithm. In the fourth phase, the optimal feature subset feed to base machine learning classifier such as *SVM* for classification to prediction of cyber-attacks and normal activities. Details of each step of the proposed scheme is defined as follows:

### *4.1 Phase 1: Data pre-processing*

*IDS* datasets have different types of attributes, including both categorical and symbolic data. Therefore, label and data transformation are crucial to predict the vital attributes of the dataset. Moreover, a binary class is also created for the target class based on its value. There are two types of records: normal and malicious. Normal records are denoted by 1 while malicious records are denoted by 0. Also, it is essential to normalize the data to avoid bias resulting from high value features in the dataset [25]. Dataset is rescaled by Min-max normalization method to scale the values of each feature to 0-1, as defined in Eq. (4) [22].

$$f_{normalized} = \frac{f - f_{minimum}}{f_{maximum} - f_{minimum}} \tag{4}$$

Here $f$ is the feature to be scaled down, $f_{minimum}$ denotes the minimum value, and $f_{maximum}$ denotes the maximum value for a specific feature present in the dataset.

### *4.2 Phase 2: Guided feature selection using ranking-based filter methods*

*FS* is a technique for selecting the optimum subset of features in accordance with a certain standard [14]. In this work, feature ranking-based filter methods such as Chi-Square, *PCC*, *MI* are used to produce ranked features from the pre-processed dataset, as discussed in section 4.1. Next, top ten ranked attributes are taken and then convert those into binary codified chromosomes of the population i.e., represented by 1, and not selected feature are represented by 0, as defined in Algorithm 1. Since, this process will provide some guidance for population initialization in multi-objective optimization technique such as NSGA-II. The reason to apply guided population initialization using ranked-based filter methods, as wrapper methods are more expensive for population initialization [13].

### *4.3 Phase 3: Hybrid NSGA-II for selection of optimal subset of features*

In this section, hybrid multi-objective optimization technique, called Hybrid *NSGA-II* is applied to get optimal subset of features. In this hybrid model, initial hybrid population is used, as defined in Algorithm 2, line #1 to provide some guidance for initial population initialization. For this experiment, Table 2 shows the NSGA-II algorithm parameters. Moreover, in this work two objectives such as (i) maximum classification accuracy, and (ii) minimum number of attributes are optimized. Additionally, binary individuals are used to identify whether an attribute is picked or not. A classifier's accuracy is taken as a fitness function in the train dataset with the picked attributes using classifier such as Support Vector Machines.

### *4.4 Phase 4: Support Vector Machine (SVM) used for cyber-attack detection*

Machine learning techniques such as SVM are used for classification analysis. During the training phase, the *SVM* model aims to select the hyperplane that maximize the smallest distance between each category [26]. Hyperplane is a kind of a boundary in between two class, which decides the new data belongs to cyber-attack class or normal class. In practice, these types of choices lead to better decisions. The hyperplane is selected based on the data points, which are very close to opponent class called support vector. Its greatest advantage is that overlearning and high dimension issues are overcome with *SVM* [27]. In this experiment hyper-parameter is tuned for *SVM*, as shown in Table 3.

Table 2 Parameter setting for Hybrid NSGA-II.

| Algorithm | Parameters |
|---|---|
| *NSGA-II* | $p_{crossover=} = 0.8$ , $p_{mutation} = 0.2$ , $pop_{size} = 100$ , $generation = 100$ , $initial_{Type} = Top\ 10\ guided\ and\ 90\ random$ |

**Algorithm 1**: Guided Population Initialization using Rank-based Filter Methods

**Input**: Set of actual features ꜰ $\in \{f_1, f_2, f, \cdots\cdots \cdots\cdots, f_n\}$

**Output**: Top ‘k’ ranked features converted into binary codified chromosome ($R_f$)

```
Initialize $n$-dimension feature space ꜰ $\in \{f_1, f_2, f_3, \cdots\cdots \cdots\cdots, f_n\}$
First ranked features subset $C_f$ using Chi-Square Test
Second ranked feature subset $P_f$ using $PCC$
Third ranked feature subset $M_f$ using $MI$
$R_f$ = Top ‘k’ ranked features using $C_f$, $P_f$, and $M_f$ are converted into binary chromosome
return $R_f$
```

**Algorithm 2**: Hybrid NSGA-II for Selection of Optimal Subset of Features

**Input**: Pre-processed dataset, $pop_{size}$ , $generation$ , $p_{crossover}$ , $p_{mutation}$ , $initial_{Type}$

**Output**: Optimal subset of features

```
$InitialHybridPopulation$ ←Top ‘k’ guided population initialized using Algorithm1 and rest of the population initialized randomly
    in $n$-dimension feature space ꜰ $\in \{f_1, f_2, f_3, \cdots\cdots \cdots\cdots, f_n\}$
Evaluate multi-objective fitness values of each chromosome in $InitialHybridPopulation$
while $iter_{num} < generation$ do
    foreach $chromosome \in InitialHybridPopulation$ do
        $parent_1 \leftarrow chromosome$
        $parent_2 \leftarrow rnadom\_choose(InitialHybridPopulation)$
        $rnd \leftarrow random()$
        if $rnd < p_{crossover}$ then
            $descendant \leftarrow crossover(parent_1, parent_2)$
        else
            if $rnd < p_{mutation} + p_{crossover}$ then
                $desendant \leftarrow mutation(parent_1)$
            end
        end
        if $descendant \notin DescendantPopulation$ then
            include $descendant$ to $DescendantPopulation$
        end
    end
    Evaluate multi-objective fitness values of each chromosome in $DescendantPopulation$
    Merge $InitialHybridPopulation$ & $DescendantPopulation$
    Compute non dominance sorting
    Choose chromosome for next generation
    Increase $iter_{num}$
end
return $InitialHybridPopulation$
```

## 5. Experimental Results and Discussion

This section shows performance evaluation of the proposed scheme using a benchmark *IDS* dataset such as ToN-IoT. The proposed scheme is implemented in Python 3.9 with SVC package on a Windows 11 PC with Intel(R) Core (TM) i7 CPU, with 16GB RAM and 500GB SSD. Performance metrics such as accuracy and number of selected features; are used to assess the performance of the proposed scheme. Result of the proposed scheme is compared with original NSGA-II and some existing techniques, which are evaluated by ToN-IoT dataset.

Table 3 Parameter setting for SVM.

| Parameter | Kernel | Gamma | C | Probability |
|---|---|---|---|---|
| Value | rbf | scale | 5000 | True |

### *5.1 Description of the ToN-IoT Dataset*

At the Cyber Range and IoT Labs at the School of Engineering and Information Technology, UNSW Canberra @ the Australian Defense Force Academy (ADFA) collected data from a large-scale and realistic network [28].

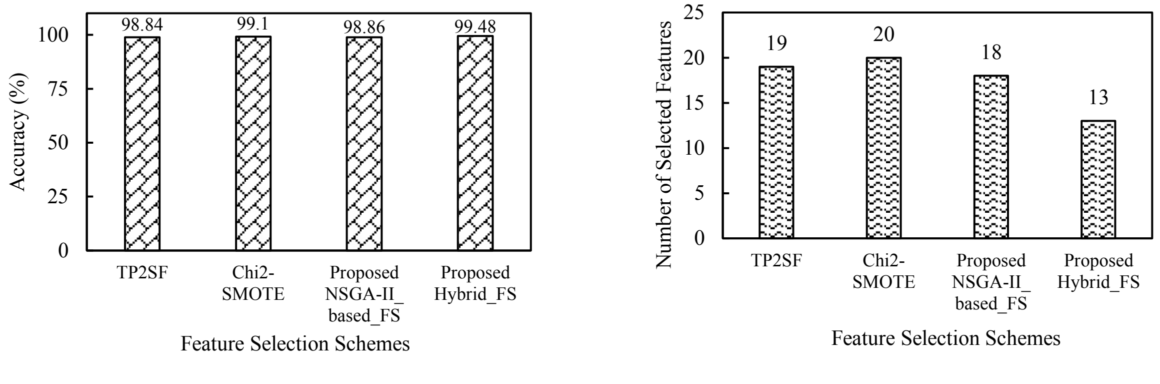
Telemetry data from IoT services, network traffic from IoT networks, and operating system logs are included in this dataset [29]. In this work, publicly available train-test network traffic flows dataset is utilized for the experiment. ToN-IoT covers different cyber-attacks, such as: ransomware, DoS, and DDoS. Dataset having original 44 features, which is extracted by Bro-IDS, also known as Zeek. Dataset having 796,380 (3.56%) normal flows and 21,542,641 (96.44%) malicious flows, that is, 22,339,021 records in total. Table 4 shows various feature name with their index number.

Table 4 ToN-IoT features with their index number and name

| Feature # | Feature name | Feature # | Feature name | Feature # | Feature name |
|---|---|---|---|---|---|
| $f_0$ | ts | $f_{15}$ | dst_ip_bytes | $f_{30}$ | http_trans_depth |
| $f_1$ | src_ip | $f_{16}$ | dns_query | $f_{31}$ | http_method |
| $f_2$ | src_port | $f_{17}$ | dns_qclass | $f_{32}$ | http_uri |
| $f_3$ | dst_ip | $f_{18}$ | dns_qtype | $f_{33}$ | http_version |
| $f_4$ | dst_port | $f_{19}$ | dns_rcode | $f_{34}$ | http_request_body_len |
| $f_5$ | proto | $f_{20}$ | dns_AA | $f_{35}$ | http_response_body_len |
| $f_6$ | service | $f_{21}$ | dns_RD | $f_{36}$ | http_status_code |
| $f_7$ | duration | $f_{22}$ | dns_RA | $f_{37}$ | http_user_agent |
| $f_8$ | src_bytes | $f_{23}$ | dns_rejected | $f_{38}$ | http_orig_mime_types |
| $f_9$ | dst_bytes | $f_{24}$ | ssl_version | $f_{39}$ | http_resp_mime_types |
| $f_{10}$ | conn_state | $f_{25}$ | ssl_cipher | $f_{40}$ | weird_name |
| $f_{11}$ | missed_bytes | $f_{26}$ | ssl_resumed | $f_{41}$ | weird_addl |
| $f_{12}$ | src_pkts | $f_{27}$ | ssl_established | $f_{42}$ | weird_notice |
| $f_{13}$ | src_ip_bytes | $f_{28}$ | ssl_subject | $f_{43}$ | type |
| $f_{14}$ | dst_pkts | $f_{29}$ | ssl_issuer | | |

### *5.2 Evaluation of Proposed Model*

As discussed in sections 4.1, different pre-processing steps have been applied to simplify datasets for machine learning techniques. The proposed hybrid method is based on ranking-based filter methods and *NSGA-II*. It is observed that ToN-IoT has only one solution in their pareto front. In the solution, proposed model achieves accuracy with 99.48% and the optimized feature sets are having 13 features. Table 5 shows the optimized features subset and accuracies acquired by ToN-IoT datasets.



(a) (b)

Fig. 2. (a) Classification accuracy using ToN-IoT dataset; (b) Number of selected features using ToN-IoT dataset.

### *5.3 Result comparison of proposed model using ToN-IoT dataset*

According to section 4.2, the goal of *FS* is to choose the features that contain the most information in the dataset. It has been reported that, many techniques for this subject have been used over the years that are considered only legacy methods. For ToN-IoT dataset, results have been compared with original NSGA-II and two existing methods: Kumar *et al.* [17], and Gad *et al.* [18] in terms of total count of selected features, classification accuracy, as demonstrates in Table 6. Fig. 2 (a) and 2 (b) shows the graphical representation of the result for accuracy and selected features for the three examined techniques and each bar represents the score for ToN-IoT dataset. As the Fig. 2 (a) and 2 (b) shows that proposed model achieved the highest accuracy (99.48%) and minimum number of selected features (13 out of 43features) respectively. Therefore, it can be decided that proposed model shows good results, as compared to the other existing techniques.

Table 5. Statistics of Pareto optimal solutions for the proposed model on the ToN-IoT dataset.

| Dataset | Accuracy | Index of optimized feature set | Selected Features |
|---|---|---|---|
| ToN-IoT | 99.48% | $\{f_0, f_1, f_3, f_9, f_{13}, f_{19}, f_{21}, f_{23}, f_{24}, f_{28}, f_{36}, f_{37}, f_{39}\}$ | 13 |

Table 6. Comparison of the selected features and accuracy obtained by the proposed model with existing techniques for ToN-IoT dataset.

| Reference | FS Technique | Classifier | Selected Features | Accuracy (%) |
|---|---|---|---|---|
| Kumar *et al.* [17] | *TP2SF* | XG-Boost | 19 | 98.84 |
| Gad *et al.* [18] | *Chi$^2$-SMOTE* | XG-Boost | 20 | 99.10 |
| Proposed NSGA-II_based_FS Scheme | *NSGA-II* | SVM | 18 | 98.86 |
| Proposed Hybrid_FS Scheme | Filter + *NSGA-II* | SVM | 13 | 99.48 |

## 6. Conclusion

This paper proposed a hybrid feature selection scheme by combining the filter methods and the *NSGA-II-based meta-heuristic* technique to minimize the number of features and maximize the classification accuracy for the *IoT* networks. In first phase, model uses data pre-processing steps. In second phase, filter methods such as Chi-square, *PCC*, and *MI* are applied to rank the top '*k*' feature. Then these top '*k*' features are converted into binary encoded individuals for guided population initialization. In the third phase, to make a hybrid model, ranked-based filter methods are used as defined in phase 2 for guided population initialization in *NSGA-II* algorithm. In the fourth phase, machine learning technique such as *SVM* is used for the classification. To evaluate the performance of the model benchmark IDS dataset namely ToN-IoT is used. Results are compared with original NSGA-II and state-of-the-art cyber-attacks detection techniques with similar experimental scenarios. Overall, the proposed model outperformed in terms of accuracy (99.48%) and minimum optimized number of features for ToN-IoT dataset from 43 to only 13. In future, we plan extend to proposed model with other real time network dataset using others metaheuristic techniques.

## Acknowledgements

This work was supported by the MATRICS Project funded by the Science and Engineering Research Board (SERB), India, under the Project No: MTR/2019/001285.

## References

[1] E. Oriwoh and M. Conrad, "'Things' in the Internet of Things: Towards a Definition," *Int. J. Internet Things*, vol. 4, no. 1, pp. 1–5, 2015.

[2] PentaSecurity, "Ways to Deal with Cyber Risks in 2022," 2022. https://www.pentasecurity.com/blog/ways-to-deal-with-cyber-risks-in-2022/.

[3] A. Nazir and R. A. Khan, "A novel combinatorial optimization based feature selection method for network intrusion detection," *Comput. Secur.*, vol. 102, no. 10, p. 102164, Mar. 2021, doi: 10.1016/j.cose.2020.102164.

[4] H. Alazzam, A. Sharieh, and K. E. Sabri, "A feature selection algorithm for intrusion detection system based on Pigeon Inspired Optimizer," *Expert Syst. Appl.*, vol. 148, 2020, doi: 10.1016/j.eswa.2020.113249.

[5] V. Hajisalem and S. Babaie, "A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection," *Comput. Networks*, vol. 136, pp. 37–50, 2018, doi: 10.1016/j.comnet.2018.02.028.

[6] B. Bauer and M. Kohler, "On deep learning as a remedy for the curse of dimensionality in nonparametric regression," *Ann. Stat.*, vol. 47, no. 4, pp. 2261–2285, 2019, doi: 10.1214/18-AOS1747.

[7] H. M. Huan Liu, *Feature Selection for Knowledge Discovery and Data Mining*. Springer International Publishing, 1998.

[8] P. Agrawal, H. F. Abutarboush, T. Ganesh, and A. W. Mohamed, "Metaheuristic algorithms on feature selection: A survey of one decade of research (2009-2019)," *IEEE Access*, vol. 9, pp. 26766–26791, 2021, doi: 10.1109/ACCESS.2021.3056407.

[9] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," *Egypt. Informatics J.*, vol. 19, no. 3, pp. 179–189, 2018, doi: 10.1016/j.eij.2018.03.002.

[10] S. Sen, S. Saha, S. Chatterjee, S. Mirjalili, and R. Sarkar, "A bi-stage feature selection approach for COVID-19 prediction using chest CT images," *Appl. Intell.*, vol. 51, no. 12, pp. 8985–9000, Dec. 2021, doi: 10.1007/s10489-021-02292-8.

[11] Z. Alizadeh Afrouzy, M. M. Paydar, S. H. Nasseri, and I. Mahdavi, "A meta-heuristic approach supported by NSGA-II for the design and plan of supply chain networks considering new product development," *J. Ind. Eng. Int.*, vol. 14, no. 1, pp. 95–109, 2018, doi: 10.1007/s40092-017-0209-7.

[12] A. Al Shorman, H. Faris, and I. Aljarah, "Unsupervised intelligent system based on one class support vector machine and Grey Wolf optimization for IoT botnet detection," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 7, pp. 2809–2825, 2020, doi: 10.1007/s12652-019-01387-y.

[13] R. Kundu and R. Mallipeddi, "HFMOEA: A hybrid framework for multi-objective feature selection," *J. Comput. Des. Eng.*, vol. 9, no. 3, pp. 949–965, 2022, doi: 10.1093/jcde/qwac040.

[14] P. Kumar, G. P. Gupta, and R. Tripathi, "Toward Design of an Intelligent Cyber Attack Detection System using Hybrid Feature Reduced Approach for IoT Networks," *Arab. J. Sci. Eng.*, vol. 46, no. 4, pp. 3749–3778, 2021, doi: 10.1007/s13369-020-05181-3.

[15] M. Roopak, G. Y. Tian, and J. Chambers, "An Intrusion Detection System Against DDoS Attacks in IoT Networks," *2020 10th Annu. Comput. Commun. Work. Conf. CCWC 2020*, pp. 562–567, 2020, doi: 10.1109/CCWC47524.2020.9031206.

[16] R. SaiSindhuTheja and G. K. Shyam, "An efficient metaheuristic algorithm based feature selection and recurrent neural network for DoS attack detection in cloud computing environment," *Appl. Soft Comput.*, vol. 100, p. 106997, 2021, doi: 10.1016/j.asoc.2020.106997.

[17] P. Kumar, G. P. Gupta, and R. Tripathi, "TP2SF: A Trustworthy Privacy-Preserving Secured Framework for sustainable smart cities by leveraging blockchain and machine learning," *J. Syst. Archit.*, vol. 115, no. July 2020, p. 101954, 2021, doi: 10.1016/j.sysarc.2020.101954.

[18] A. R. Gad, A. A. Nashat, and T. M. Barkat, "Intrusion Detection System Using Machine Learning for Vehicular Ad Hoc Networks Based on ToN-IoT Dataset," *IEEE Access*, vol. 9, pp. 142206–142217, 2021, doi: 10.1109/ACCESS.2021.3120626.

[19] S. Bacha, A. Aljuhani, K. Ben Abdellafou, O. Taouali, N. Liouane, and M. Alazab, "Anomaly-based intrusion detection system in IoT using kernel extreme learning machine," *J. Ambient Intell. Humaniz. Comput.*, no. Aljuhani 2021, 2022, doi: 10.1007/s12652-022-03887-w.

[20] I. S. Thaseen, C. A. Kumar, and A. Ahmad, "Integrated Intrusion Detection Model Using Chi-Square Feature Selection and Ensemble of Classifiers," *Arab. J. Sci. Eng.*, vol. 44, no. 4, pp. 3357–3368, 2019, doi: 10.1007/s13369-018-3507-5.

[21] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, 2002, doi: 10.1109/4235.996017.

[22] A. Da Li, Z. He, and Y. Zhang, "Bi-objective variable selection for key quality characteristics selection based on a modified NSGA-II and the ideal point method," *Comput. Ind.*, vol. 82, pp. 95–103, 2016, doi: 10.1016/j.compind.2016.05.008.

[23] S. Casale, A. Russo, and S. Serrano, "Multistyle classification of speech under stress using feature subset selection based on genetic algorithms," *Speech Commun.*, vol. 49, no. 10–11, pp. 801–810, 2007, doi: 10.1016/j.specom.2007.04.012.

[24] S. Yildirim, Y. Kaya, and F. Kılıç, "A modified feature selection method based on metaheuristic algorithms for speech emotion recognition," *Appl. Acoust.*, vol. 173, p. 107721, 2021, doi: 10.1016/j.apacoust.2020.107721.

[25] S. K. Sahu, S. Sarangi, and S. K. Jena, "A detail analysis on intrusion detection datasets," *Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC 2014*, pp. 1348–1353, 2014, doi: 10.1109/IAdCC.2014.6779523.

[26] 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 Med. Unlocked*, vol. 17, no. October, p. 100255, 2019, doi: 10.1016/j.imu.2019.100255.

[27] Z. Tao, L. Huiling, W. Wenwen, and Y. Xia, "GA-SVM based feature selection and parameter optimization in hospitalization expense modeling," *Appl. Soft Comput. J.*, vol. 75, pp. 323–332, 2019, doi: 10.1016/j.asoc.2018.11.001.

[28] Nour Moustafa, "ToN_IoT datasets." IEEE Dataport, 2019, [Online]. Available: doi: https://dx.doi.org/10.21227/fesz-dm97.

[29] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and Adna N Anwar, "TON-IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," *IEEE Access*, vol. 8, pp. 165130–165150, 2020, doi: 10.1109/ACCESS.2020.3022862.