NATURENGS
E-ISSN 2717-8013
https://dergipark.org.tr/en/pub/naturengs

Effective Cyber Attack Detection Based on Augmented Genetic Algorithm with Naive Bayes

Hayriye TANYILDIZ [a,*], Canan BATUR ŞAHİN [a], Özlem BATUR DİNLER [b]

[a] Malatya Turgut Ozal University, Faculty of Engineering and Natural Sciences,Malatya 44210, Türkiye
[b] Siirt University, Faculty of Engineering, Departmant of Computer Engineering, Siirt, 56000, Türkiye

*Corresponding author

ARTICLE INFO

Received	07.12.2023
Accepted	19.12.2023

Doi: 10.46572/naturengs.1401669

ABSTRACT

This study can be considered a vital development in the field of cyber security. Today, the ever-changing and evolving structure of cyber threats constantly challenges defense mechanisms and requires the development of innovative solutions. In this context, the application of the Naive Bayes approach enriched with genetic algorithm offers a significant contribution to existing methodologies in this field. In particular, the use of genetic algorithm in cyber-attack detection optimizes classification processes by determining the most appropriate features from data sets and thus provides a more effective detection mechanism. The integration of the Naive Bayes classifier makes it possible to detect cyber-attacks precisely and quickly based on these selected features. Empirical studies and evaluations have shown that this approach provides superior sensitivity rates and lower false positive rates than traditional techniques, demonstrating its potential to overcome the limitations of existing methods in the field of cybersecurity. These findings can be considered an important step in making cybersecurity strategies more efficient and adaptable, especially considering the constantly evolving and unpredictable nature of cyber threats. The results of this study highlight the importance of developing innovative and effective solutions in the field of cybersecurity and provide a basis for further research in this field.

Keywords: Navie Bayes, Feature Selection, Cyber Attack, Optimisation

1. Introduction

Due to the swift progress of the digital era, cybersecurity has become an essential component of the everyday lives of individuals and companies. Cyber-attacks, characterized by their growing frequency and complexity, are recognized as one of the foremost dangers to information security. The discipline of cybersecurity places great priority on the efficient identification and prevention of these attacks.

Conventional techniques employed presently for identifying cyber-attacks are constrained by the intricate and constantly evolving nature of the threat environment. These methods frequently yield a significant number of incorrect identifications and lack the ability to promptly adjust to novel forms of attacks. The demand for cyber-attack detection systems that are both efficient and adaptable is growing steadily in this particular scenario.

This study explores the utilization of genetic algorithms and Naive Bayes classifier as a novel strategy for detecting cyber-attacks. Genetic algorithms are successful in identifying optimal solutions, yet the Naive Bayes classifier is renowned for its simplicity and computational efficiency. The combination of these two methodologies can enhance the speed and accuracy of cyber-attack detection, particularly when dealing with extensive data sets.

In this study, we investigate how a Naive Bayes approach augmented with a genetic algorithm can provide an effective solution for cyber-attack detection through feature reduction [1]. The performance of the proposed method compared to existing cyber intrusion detection systems will be evaluated through tests and analyses on various data sets.

This study aims to showcase the potential benefits of combining genetic algorithm with Naive Bayes classifier in the domain of cyber intrusion detection [2]. Additionally, it seeks to provide a valuable contribution to the current literature on this topic.

It is possible to summarize some of the research and algorithms on cyber-attack detection in time series data with different data mining algorithms in the literature as follows: Support Vector Machines (SMO-SVM) were assessed using SCADA datasets. The datasets were processed, and the classification performances were

* Corresponding author. e-mail address: hayriyetanyildiz@tedas.gov.tr
ORCID : 0000-0002-6300-9016

improved using principal component analysis for feature selection. As a result of the evaluations on two SCADA datasets, the SMO-SVM algorithm provided the best overall results [3].

However, Zi et al. proposed a method called end-to-end anomaly detection for anomaly detection in twin data to perform real-time anomaly detection quickly and accurately. A multidimensional inverse convolutional network and attention mechanism were applied to the model to search for key features. The study's findings demonstrated that the proposed technique had superior performance in terms of precision and F1 score when compared to state-of-the-art methods [4].

Cai et al. introduced a method for detecting intrusions in industrial control systems using a 1D CWGAN-based approach. The method is based on the idea that using imbalanced data for training deep learning models can greatly decrease detection performance. 1D CWGAN is a network attack pattern generation method that combines 1D CNN and WGAN. Firstly, the problem of low ICS intrusion detection accuracy due to a small number of attack samples was analyzed. This method balances the number of various attack samples in the dataset in terms of data augmentation and aims to increase detection accuracy. The algorithm constructs a framework for modeling network traffic data in two competing networks. It utilizes 1D convolution and 1D inverse convolution based on the temporal characteristics of the network traffic. Additionally, it generates virtual samples that resemble real samples by employing a gradient penalty instead of weight truncation in the Wasserstein Generator Adversarial Networks (WGAN). As a result of validations using large datasets, it was shown that the method improves the classification performance of CNN and BiSRU. For CNN, accuracy increased by 0.75% after data stabilization, while accuracy, recall, and F1 also improved. Compared to BiSRU without data processing, the accuracy of 1D CWGAN-BiSRU increased by 1.34% and accuracy, recall, and F1 increased by 7.2%, 3.46% and 5.29%, respectively [5].

Sahin et al. proposed a novel software vulnerability prediction model utilizing a deep learning approach and Symbiotic Genetics. In this proposed method, a deep Symbiotic-based genetic algorithm model (DNN-Symbiotic GAs) is used to learn the phenotyping of dominant features. Their results showed that the proposed model is good at predicting software vulnerabilities [6].

In their study, Li et al. proposed an unsupervised multivariate anomaly detection method based on Generative Adversarial Anomaly detection achieved high-performance results by simultaneously considering the entire set of variables to capture [7].

Auidibert et al. proposed a fast and stable method called Unsupervised Anomaly Detection for multivariate time series (USAD) based on negatively trained autoencoders. They managed to verify the requirements regarding scalability, stability, robustness, training speed, and high performance in their proposed model [8].

Yang et al. examined the filter-based feature selection technique in detail and comprehensively. In addition, they describe the search algorithms and relevance metrics [9].

2. Materials and Methods

2.1 Datasets

In this study, the WADI dataset from iTrust, Centre for Research in Cyber Security, Singapore University of Technology and Design was used [10].

A water distribution system comprises a vast network of pipelines spanning a wide area, in contrast to a water treatment plant that is often situated in a secure and concentrated place. This water distribution system consists of three main components: treatment, storage, and distribution. The dataset in question focuses on security breaches in the storage part of this system. The WADI testbed used for this purpose is equipped with various features such as chemical dosing systems, booster pumps, valves, various instruments, and analyzers. The purpose of this system is to replicate both cyber-attacks and defenses on Programmable Logic Controllers (PLCs) via network simulations. Additionally, it also replicates the consequences of physical threats such as water breaches and the introduction of dangerous chemicals.

Data collection for the WADI testbed was carried out over a continuous period of 16 days, with 14 days dedicated to recording data under both standard operational and attack scenarios. The installation of this data collection was completed in 2 days. During this period, extensive data was collected, including network traffic as well as information from sensors and actuators.

The WADI dataset, was collected from 123 sensors, and 15 types of attacks were identified in the collected dataset. The duration of each attack ranges from 1.5~ 30 seconds.

2.2 Naive Bayes Classifier

The Naive Bayes classification technique is derived from the theory of Bayes and possesses an easy-to-understand and comprehensible structure. Additionally, it is distinguished by its rapid functionality. This algorithm is based on the assumption that all variables are independent of each other and each has equal importance [11]. This simple approach enables the Naive Bayes algorithm to achieve effective and fast results in data analysis and classification. The algorithm's simplicity and efficiency in handling big datasets are facilitated by the independence and equal significance of variables [12]. This makes it a preferred method, especially for large datasets and real-time applications.

Posterior probability in Naive Bayes is computed assuming feature independence. In the Naive Bayes classifier, the Laplace smoothing technique is used to deal with the zero probability problem. This technique allows the model to be more flexible by setting the probability of combinations of features never observed in the dataset to a small value instead of zero.

2.3 Feature Selection with Genetic Algorithm

Genetic Algorithms operate by employing the principles of natural selection. They embody a heuristic methodology for searching and optimizing, aiming to uncover solutions that approximate the optimal outcome. These algorithms work by iteratively refining and improving the optimal solutions over successive iteration s [13]. Inspired by the process of natural evolution, Genetic Algorithms overcome complex challenges by mimicking this evolutionary mechanism. They continuously develop and refine a pool of potential solutions, often referred to as "individuals", and gradually move towards an optimal or near-optimal solution as the process progresses.

In the field of Machine Learning, the process of training a model involves collecting a significant amount of data to improve its learning ability. Nevertheless, not all of this input is consistently pertinent or valuable to the model, and an abundance of irrelevant data might result in inefficiencies, decelerating the model and potentially yielding erroneous results. Feature selection is an essential stage in this context, aiming to filter the dataset and keep only the data that has a substantial impact on the model's performance, while rejecting redundant information. This approach enhances the efficiency of the model without necessitating any modifications or conversions to the current data. It specifically concentrates on generating a subset of the most pertinent features [14].

The role of feature selection is particularly important in classification systems. By applying this technique, the complexity of classification tasks is reduced, leading to a reduction in the number of computations required. By eliminating irrelevant and chaotic features from the dataset, this approach not only enhances classification accuracy but also decreases training time, the requirement for extensive measurement, and memory consumption. As a result, the classification process becomes more meaningful and simpler and contributes to a more streamlined and effective machine learning model [15].

As can be seen in Figure 1, the raw dataset is utilized to create the feature subset. To determine whether to retain the feature subset extracted from the unprocessed data, various formulas are applied to evaluations. The feature that is selected following the evaluations is appended to the pertinent subset, and the procedure persists until the algorithmic stopping criterion is fulfilled.

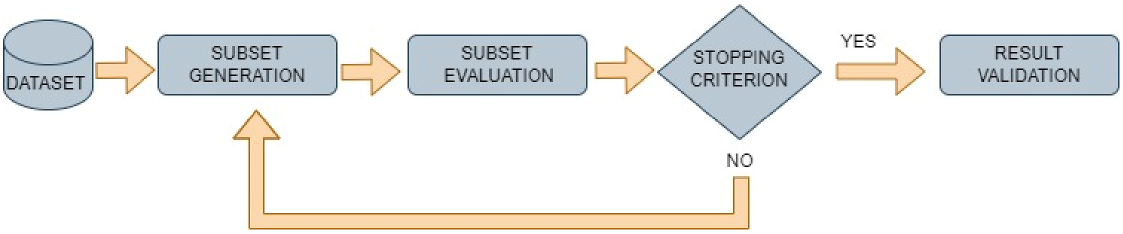


Figure 1. Flowchart of feature selection with genetic algorithm

2.4 Performance Evaluation

Error matrices were obtained for the performance evaluations of the trained models, and the values to be compared were calculated with the help of these error matrices. The error matrix is a 2x2 matrix, and the values in it are as follows:

- True Positive: Comments correctly categorized as positive.
- True Positive: Comments correctly categorized as positive.
- False Positive: Comments incorrectly categorised as positive.
- False Negative: Comments incorrectly categorised as negative.

The formula used to calculate the Accuracy value;

$$\text{Accuracy} = \frac{(\text{TP+TN})}{(\text{TP+FP+TN+FN})} \tag{1}$$

The formula used to calculate the Precision value;

$$\text{Precision} = \frac{\text{TP}}{(\text{TP+FP})} \tag{2}$$

The formula used to calculate the recall value;

$$\text{Recall} = \frac{\text{TP}}{(\text{TP+FN})} \tag{3}$$

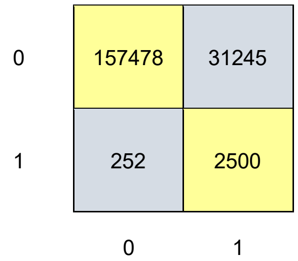
The formula used to calculate the F1-score (F1 value) value;

$$F1 = \frac{2*((\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{4}$$

3. Results and Discussion

The Naive Bayes algorithm is deemed the most suitable classifier in this study for one-step prediction of cyber intrusions in the WADI dataset.

Firstly, data pre-processing techniques were used to maximize the efficiency of the dataset. In the data preprocessing stage, blank data were filled by taking the mean value of the data in that column, and the features in the dataset were subjected to standardization. After the data preprocessing step, the classification process was applied. An examination of the model's confusion matrix revealed that the algorithm accurately predicted 157478 out of the 188723 non-attack values, while erroneously predicting 31245 of them. Out of 2752 attack values, 252 of them were incorrect, and 2500 of them were correct. In the second stage, reclassification was performed with the data obtained by using genetic algorithms for future selection. The parameters selected for the genetic algorithm are given in Table 1.



0	157478	31245
1	252	2500
	0	1

Figure 2. Confusion matrix of Naive Bayes

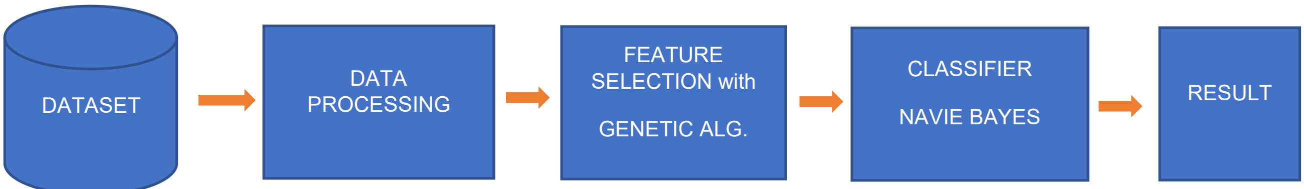


Figure 3. Proposed model workflow

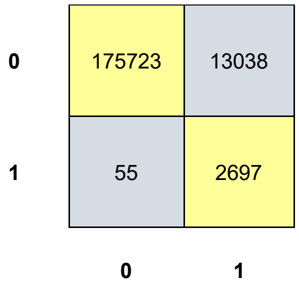
Table 1. Parameters of genetic algorithm

Environment	Population	nGen	cxpb	mutpb
Colab	50	10	0.5	0.2

In the algorithm, the size of the first population is set to 50. This means that the genetic algorithm will initially start with 50 random individuals. The number of generations that the genetic algorithm will work on is determined as 10.

To determine the frequency of crossover and mutation, the crossover probability cxpb was set to 0.5 and the mutation probability to 0.2. Figure 3. shows the workflow used for the proposed model. Following the application of a feature reduction process to the dataset acquired through pre-processing using the genetic algorithm, the dataset was subsequently subjected to a classification process utilizing the Naive Bayes algorithm.

Upon examination of the confusion matrix produced by the model, it is observed that out of the total 188723 non-attack values, this method accurately predicted 175723 while erroneously predicting 13038. Out of 2752 attack values, 55 of them were incorrect and 2697 of them were correct. The performance values of the models run on the Valley dataset without and after feature selection are shown in Table 2.



0	175723	13038
1	55	2697
	0	**1**

Figure 4. Confusion matrix of Naive Bayes+Genetic algorithm feature selection

Table 2. Feature of Naive Bayes after feature selection and before feature selection

	Accuracy	Precision	Recall	F1
Naive bayes without feature selection	0.8355	0.8344	0.9984	0.9091
Naive bayes by selecting feature with genetic algorithm	0.9577	0.9572	0.9998	0.9782

Table3. Comparison of relative studies

	Models	Results
Li et al.[7]	MADGAN	Pre:6.46 Rec:99.99 F1: 0.12
Tian et al. [16]	STADN	Pre:98.49 Rec:45.57 F1: 0.62
Auidibert et al. [8]	USAD	Pre:0.9947 Rec:0.1318 , F1:0.2328
Our Model	Feature Selection(Gen. Alg.)+ Naive Bayes	Acc:0.9572 Pre: 0.95 Rec:0.99 F1:0.97

4. Conclusion

In this study, a comparative analysis between two different implementations of the Naive Bayes classification algorithm: the first is the traditional Naive Bayes approach without any feature selection, and the second is the improved version where feature selection is performed using a genetic algorithm. The effectiveness of both models was evaluated based on key performance indicators such as accuracy, precision, recall, and F1 score.

The standard Naive Bayes model used without the integration of feature selection techniques showed an accuracy rate of 83.55%, a precision rate of 83.44%, a recall rate of 99.84% and an F1 score of 90.91%. These measurements show an overall competent performance of the model and a remarkable strength in the recall capability in particular. However, the sensitivity and F1 score suggest that the model's performance may not be optimally balanced.

On the contrary, the Naive Bayes model, which integrated feature selection through a genetic algorithm, demonstrated a substantial enhancement in performance, attaining an accuracy of 95.77%, precision of 95.72%, recall of 99.98%, and F1 score of 97.82%. These improved results show that the integration of a genetic algorithm for feature selection significantly improves the accuracy and precision of the model. The near-perfect recall and the very high F1 score attest to the model's proficiency in correctly identifying positive examples while effectively reducing false positives.

The aforementioned observations highlight the potential of feature selection utilizing genetic algorithms to substantially enhance the performance of fundamental statistical classification models like Naive Bayes. The study concludes that in scenarios where datasets contain a large number of features that have various effects on classification accuracy, feature selection using genetic algorithms can be a highly advantageous strategy. This study underlines the critical role of model selection and feature engineering in machine learning and demonstrates their significant contribution to optimizing model performance.

References

[1] **Ferriyan,** A. H. Thamrin, K. Takeda, and J. Murai, "Feature selection using genetic algorithm to improve classification in network intrusion detection system," 2017 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), Surabaya, Indonesia, 2017, pp. 46-49,

[2] **Saurabh,** M., Neelam Sharma, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction," Procedia Technology,Volume 4,2012,Pages 119-128, ISSN 2212-0173,

[3] **Alimi, O**., Ouahada, K., Abu Mahfouz, A.M., Rimer, S. & Alimi, K. 2022. Supervised learning-based intrusion detection for SCADA systems. http://hdl.handle.net/10204/12516

[4] **Li,** Z., Duan, M., Xiao, B., & Yang, S. (2023). A Novel Anomaly Detection Method for Digital Twin Data Using Deconvolution Operation With Attention Mechanism. IEEE Transactions on Industrial Informatics, 19, 7278-7286.

[5] **Cai** Z, Du H, Wang H, Zhang J, Si Y, Li P. One-Dimensional Convolutional Wasserstein Generative Adversarial Network-Based Intrusion Detection Method for Industrial Control Systems. Electronics. 2023; 12(22):4653.

[6] **Şahin,** C.B., Dinler, Ö.B. & Abualigah, L. Prediction of software vulnerability based deep symbiotic genetic algorithms: Phenotyping of dominant-features. Appl Intell 51, 8271–8287 (2021).

[7] **Li, D.,** Chen, D., Jin, B., Shi, L., Goh, J., Ng, SK. (2019). MAD-GAN: Multivariate Anomaly Detection for Time Series Data with Generative Adversarial Networks. In: Tetko, I., Kůrková, V., Karpov, P., Theis, F. (eds) Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series. ICANN 2019. Lecture Notes in Computer Science(), vol 11730. Springer, Cham.

[8] **Audibert,** J., Michiardi, P., Guyard, F., Marti, and Maria A. (Zuluaga. 2020). USAD: Unsupervised Anomaly Detection on Multivariate Time Series. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20). Association for Computing Machinery, New York, NY, USA, 3395–3404.

[9] **Lyu,** Y, Yaokai Feng, and Kouichi Sakurai. 2023. "A Survey on Feature Selection Techniques Based on Filtering Methods for Cyber Attack Detection." Information 14, no. 3 (2023): 191.

[10] https://itrust.sutd.edu.sg/itrustlabs_datasets/dataset_info/ (Accessed 3 December 2023)

[11] **Gül, E.** & Kalyoncu, M. (2021). Ağır Vasıta Hava Kompresörü Arıza Durumlarının Naive Bayes Sınıflandırıcısı ile Tahmini. Avrupa Bilim ve Teknoloji Dergisi, (31), 796-800.

[12] **Ron,** K. (2011), Scaling Up the Accuracy of Naive Bayes Classifiers: a DecisionTree Hybrid. Accessed: 24.04.2010, Association For The Advancement Of Artificial Intelligence Website.

[13] **Nabiyev,** V. V., 2016, Yapay Zeka, 5. Baskı, Seçkin Yayınları, Ankara, ISBN: 978-975-02-3727-0.

[14] **Ebren Kara,** Ş. & Şamlı, R. (2021). Genetik Algoritma İle Öznitelik Seçimi Yapılarak Yazılım Projelerinin Maliyet Tahmini. Avrupa Bilim ve Teknoloji Dergisi, (27), 985-994.

[15] **Zahid,** H., Muhammad Nadeem Yousaf, Muhammad Waqas, Muhammad Sulaiman, Ghulam Abbas, Masroor Hussain, Iftekhar Ahmad, Muhammad Hanif, "An effective genetic algorithm-based feature selection method for intrusion detection systems," Computers & Security, Volume 110, 2021, 102448, ISSN 0167-4048,