# Orthogonal variance-based feature selection for intrusion detection systems

Firuz Kamalov Department of Electrical Engineering Canadian University Dubai Dubai, UAE firuz@cud.ac.ae Sherif Moussa Department of Electrical Engineering Canadian University Dubai Dubai, UAE smoussa@cud.ac.ae Ziad El Khatib Department of Electrical Engineering Canadian University Dubai Dubai, UAE ziad.elkhatib@cud.ac.ae

Adel Ben Mnaouer Department of Computer Engineering Canadian University Dubai Dubai, UAE adel@cud.ac.ae

Abstract—In this paper, we apply a fusion machine learning method to construct an automatic intrusion detection system. Concretely, we employ the orthogonal variance decomposition technique to identify the relevant features in network traffic data. The selected features are used to build a deep neural network for intrusion detection. The proposed algorithm achieves 100% detection accuracy in identifying DDoS attacks. The test results indicate a great potential of the proposed method.

*Index Terms*—intrusion detection system, feature selection, network security, variance decomposition, neural network, variance decomposition

# I. INTRODUCTION

The ubiquitous connectivity of the modern world has vielded tremendous gains. Network-based technologies allow to control and synchronize operations of complex systems. Today's networks permeate every facet of daily life making them vulnerable to unwanted intrusions and attacks. The stakes for network protection are as high as ever. It is estimated that malicious network attacks cause billions of dollars in annual damage. Network attacks can also cause physical damage as illustrated by a recent attempt to access a water treatment facility's network in Florida and pump dangerous amounts of sodium hydroxide in the water supply. Therefore, it is imperative to have effective safeguarding mechanisms that can protect networks. The traditional intrusion detection systems (IDS) are manually designed and coded by domain experts. Although this approach served well in the past, the increased variety and volume of malicious attacks has made it more difficult to keep up with the pace. The modern IDS require automation and scalability. There is a need for rapid design and deployment of IDS to provide timely response to the fast evolving network attacks. Our goal is to present a new approach for constructing IDS based on machine learning methods. The proposed method is shown to achieve a 100% detection accuracy making it a promising tool in the fight against malicious attacks.

The recent advances in machine learning techniques and increased computing power have made it an attractive tool for constructing intelligent IDS. Machine learning and artificial intelligence (AI) are already being used in a range of applications against unwanted online intrusions. Random forests and neural networks have been used to build intelligent email filtering systems. AI-based filtering algorithms have the ability to independently learn to distinguish between regular and junk messages. Intelligent systems are used to automatically discover and block suspicious URLs, add connection exceptions, and create new rules. Deep neural networks are also used to detect digital virus signatures. One of the main advantages of AI-based IDS is scalability. Given the tremendous computing power that is available on cloud platforms such as Amazon Web Services, AI-based systems are able to handle any increase in network traffic. AI algorithms do not require human input and can learn and evolve as necessary on their own accord. In fact, AI algorithms benefit from increased traffic as it provides more data for learning. The processing power of AI-based IDS cannot be matched by human experts. Thus, AI offers an attractive avenue for developing automated IDS. It seems increasingly likely that machine learning will play a crucial role in protecting our networks from malicious attacks.

In this paper, we propose a novel machine learning algorithm for detecting malicious attacks. The proposed method is based on a feature selection technique using orthogonal variance decomposition [16]. The algorithm is implemented in two stages. First, the network traffic data is analyzed using the orthogonal variance decomposition and the most relevant features are selected. The features are evaluated based on the total sensitivity index. Second, a deep neural network is constructed based on the selected features to identify malicious traffic in the data. The proposed approach provides a number of benefits:

<sup>© 20</sup>XX IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

- It identifies the key features in network traffic data that can help IT experts better protect the networks.
- It provides an effective AI-based intrusion detection system.

The proposed algorithm can be set up to regularly train itself on new traffic data. As a result, it can remain continuously up-to-date. The test results show that the proposed method achieves a high detection rate.

One of the main challenges in feature selection is feature interactions. Relationships between features can affect their relevance with respect to the target variable. For instance, a feature that is important on its own may lose its relevance when combined with another feature if the two features are correlated. Conversely, a pair features that are unimportant individually can be highly potent when combined into a single subset. To account for all the feature interactions, in theory, every possible subset of the feature set must be considered. Since the total number of subsets is equal to  $2^k$ , where is k is the number of features, it is a computationally infeasible task. Despite various attempts to address the problem of feature interactions it remains an open problem [21]. The orthogonal variance decomposition provides a method for including all the feature interactions under the assumption of feature independence. The variance of the target variable is decomposed according to the features. Consequently, a feature responsible for a higher proportion of the target variance is deemed more relevant. Although feature independence is a seemingly stringent requirement, the method has been shown to perform well even when this condition is not met [10].

Network traffic data is notoriously imbalanced. The majority network signals consists of regular traffic with only a small portion representing malicious attacks. Imbalanced data leads to bias in many machine learning classifiers [19]. Since the main goal of a classifier is to maximize the overall accuracy, it focuses on the majority data at the expense of the minority data. There are exists a number of approaches to combat imbalanced data [8]. The proposed approach is capable of handling imbalanced data by selecting the most relevant features.

Our paper is structured as follows. In Section 2, we present an overview of the existing literature on the subject matter. In Section 3, we describe the details of the proposed approach for constructing AI-based IDS. Section 4 contains the results of the numerical experiment to measure the performance of the method. We conclude with brief remarks in Section 5

#### II. LITERATURE REVIEW

The recent advances in machine learning have propelled their application in a range of domains. In particular, it has been used successfully in IDS. A comparison of popular machine learning techniques for IDS application including SVM, random forest, and extreme learning machine (ELM) was conducted in [1]. The results of numerical experiments showed that ELM outperformed other approaches. Machine learning based IDS have yielded mixed results. Intrusion detection in a cloud-based environment is investigated in [15]. The authors tackle both anomaly detection and categorization of network attacks. The proposed approach employs a combination of linear regression and random forest algorithms. The results reveal 99% detection and 93.6% categorization accuracy. The authors in [12] employ feature selection that is based on a boosting algorithm to reduce the dimensionality of data space. Then a number of machine learning algorithms are applied on the reduced dataset. The results show the effectiveness of the feature selection algorithm in increasing the detection rate up to 90.85%. A multi-faceted approach based on optimizing the model training size was proposed in [7]. The authors are able to reduce the computational time while achieving 99% detection accuracy. In [6], the authors employ a combination of Elman neural network and SVM to develop a method for intrusion information detection. The proposed method achieves detection rate between 87.3-100%

Feature selection is used to reduce the number of variables and obtain a simpler model. It is an important preprocessing step in a number of machine learning applications [11], [20]. In particular, feature selection has been applied in the context of IDS. In [3], the authors propose two separate forward search selection algorithms based on different criteria: linear correlation coefficient and mutual information. Alternatively, in [9], the authors propose to merge mutual information and mutual correlation into a single metric. The resulting feature evaluation method is combined with the decision tree algorithm to predict DDoS attacks. The authors in [2] employ mutual information-based feature selection combined with SVM to design an IDS. One of the main challenges in feature selection is feature interactions. To address this issue, a new method based on orthogonal variance decomposition was proposed in [10]. The proposed methods takes into account all the feature interactions in decomposing the variance of the target variable under the condition of pairwise feature independence.

# III. VARIANCE DECOMPOSITION AND INTRUSION DETECTION

The proposed algorithm consists of two parts: feature selection and deep learning. In the first step, we apply orthogonal variance decomposition to identify the relevant features in network traffic data. The contribution of each feature in the total variance of the target variable can be quantified in terms of the total sensitivity index (TSI). Features are subsequently scored based on the TSI. The second step of the algorithm involves training an artificial neural network on the traffic data using the features selected in the preceding step. The neural network architecture consists of several fully connected layers and a binary output. The network hyperparameters are tuned using cross validation. For convenience, we refer to the proposed algorithm as the total sensitivity neural network (TSNN).

The primary purpose of orthogonal variance decomposition is to decompose the variance of the target variable in terms of the feature variables. Furthermore, variance decomposition takes into account feature interactions. Concretely, the variance of the target variable Y is decomposed as

$$V(Y) = \sum_{i} V_{i} + \sum_{i,j} V_{ij} + \dots + V_{12..k},$$
 (1)

where each term  $V_{i_1i_2..i_s}$  represents the contribution to the variance of Y due to feature interactions in subset  $\{X_{i_1}, X_{i_2}, ..., X_{i_s}\}$ . To obtain the decomposition in Equation 1 suppose that the target variable Y is a function of a set of feature variables  $Y = f(X_1, X_2, ..., X_k)$ . Let the features  $\{X_i\}$ be independently and uniformly distributed over the interval [0, 1]. Then we obtain the following functional decomposition

$$f = f_0 + \sum_i f_i + \sum_{i,j} f_{ij} + \dots + f_{12\dots k},$$
 (2)

where  $f_0 = E[Y], f_i(x) = E[Y|X_i = x] - f_0, f_{ij}(x, y) = E[Y|X_i = x, X_j = y] - f_i(x) - f_j(y) - f_0$  and similarly for higher orders. Note that  $E[f_{i_1i_2..i_s}] = 0$ . To obtain the variance decomposition, we square and integrate the two sides of Eq. (2)

$$\int_{[0,1]^k} f^2 \, d\mathbf{X} = \int_{[0,1]^k} \left( f_0 + \sum_i f_i + \sum_{i,j} f_{ij} + \ldots + f_{12\ldots k} \right)^2 d\mathbf{X},$$

where X is the vector of  $X_i$ 's and  $[0, 1]^k$  is the k-dimensional hypercube. The decomposition in Equation 1 can be used to determine the total contribution of an individual feature to the target variance. This is done by calculating the Total Sensitivity Index (TSI) of a feature given by

$$S_{T_i} = 1 - \frac{\operatorname{Var}(\operatorname{E}[Y|\boldsymbol{X}_{\sim i}])}{V(Y)},\tag{3}$$

where  $X_{\sim i}$  is the vector of all features except  $X_i$ . The details of Equations 1 and 3 can be found in [10], [16], [18].

There exists a number of estimators for  $Var(E[Y|X_{\sim i}])$ . We follow the approach in [13]. Let A and B be a pair of independent sampling matrices. Let j and i denote row and column indexes respectively. Define  $A_B^{(i)}$  to be matrix A, where its *i*th column replaced with the *i*th column of B. Then the variance estimator is given by

$$\operatorname{Var}(\mathbf{E}[Y|\boldsymbol{X}_{\sim i}]) = \frac{1}{n} \sum_{j=1}^{n} f(\boldsymbol{A})_{j} f(\boldsymbol{A}_{\boldsymbol{B}}^{(i)})_{j} - f_{0}^{2} \qquad (4)$$

The details of the algorithm for implementing the final TSI calculations can be found in [10]. Although the utilized feature selection algorithm is more advanced most of the existing methods its implementation and complexity are reasonable. There exist many algorithms for calculating TSI that can be used for implementation.

After selecting the relevant features, we employ an artificial neural network (ANN) to classify the network traffic as malicious or benign. Neural networks have achieved success various domains including image and speech recognition which has prompted their application in IDS. Neural networks have the ability to learn nonlinear hidden patterns in data through several layers of abstraction. As shown in Figure 1, the proposed ANN is constructed with 5 fully connected layers, the size of the ANN was chosen after experimenting with different size architectures. Concretely, experiments with 3 and 4-layer architectures yield slightly lower accuracy than the 5-layer ANN. We use  $L_2$ -regularization with  $\lambda = 0.00001$ to prevent overfitting. The ReLU activation function is used to introduce nonlinearity in the ANN model. The hyperparameters are tuned using cross validation. Binary crossentropy was used as the loss function to train the ANN.



 $\label{eq:linear} \text{Input Layer} \in \mathbb{R}^{10} \qquad \text{Hidden Layer} \in \mathbb{R}^8 \qquad \text{Hidden Layer} \in \mathbb{R}^4 \qquad \text{Hidden Layer} \in \mathbb{R}^2 \qquad \text{Output Layer} \in \mathbb{R}^3$ 

Fig. 1: The ANN architecture used to classify benign and malicious attacks. For illustration purposes, the edge opacity and color are proportional to edge weights. Note that in the actual model the input size is 63.

The TSNN model can be updated on a regular basis to keep up with new threats. We propose a weekly retraining of the model based on newly available data about DDoS attacks. The decomposition step in the TSNN can be performed by an expert or automatically based on a set of predetermined criteria. To deal with different types of attacks the proposed method can be trained on a larger dataset that includes samples various DDoS instances. Alternatively, multiple TSNN models can be trained for different types of DDoS attacks.

## IV. METHODOLOGY AND RESULTS

In this section, we describe the numerical experiments that were conducted to test the performance of the proposed intrusion detection algorithm TSNN. All the computations including the implementation of the feature selection method and the neural network were done in Python using Keras [4] and scikit-learn [14] libraries.

# A. Data

The data in the experiments was obtained from the collaborative project between the Communications Security Establishment and the Canadian Institute for Cybersecurity [17]. It is based on the creation of user profiles containing abstract representations of events and behaviors seen on the network. The dataset is generated by simulating LAN network topology frequently found on the AWS computing platform. For our purposes, we extracted a random sample of size 6,000 from the original dataset. The data is distributed according to 5/1 ratio between benign and DDoS instances. The class distribution represents a realistic scenario where the majority of signals consist of regular network traffic. Each instance of the dataset consists of 63 continuous features and a label. The features consists of various packet and IAT characteristics. The dataset is split into training and test subsets according to 80/20 ratio. The training subset is used for feature selection, ANN training and validation while the test set is used to measure the unbiased performance of the algorithm.

# B. Benchmarks

We benchmark the performance of the proposed TSNN algorithm against two commonly used machine learning algorithms: support vector machines (SVM) and logistic regression (LR). The SVM classifier is a maximum margin classifier. It is designed to find the separating hyperplane with the largest margin between the two target classes. Its main advantage is the ability to handle nonlinear tasks by mapping the data into a higher dimensional representation space using the kernel trick. Logistic regression is a simple linear classifier that maximizes the log-likelihood probability of the sample data. Its main advantage is the low variance which results in less overfitting. The two benchmark classifiers are trained on the full feature set while the TSNN is trained only on a subset of features. Despite the greater amount of information available for training the benchmark algorithms, the TSNN method yields better results.

# C. Results

We test the efficacy of the proposed TSNN algorithm on the dataset described in Section IV-A. First, we apply the algorithm to select the top 10 most relevant features. The selected features are presented in Table I. Among the selected features ACK Flag Coun, Init\_Win\_bytes\_forward, and PSH Flag Count are the most important. The ACK Flag Count parameter belongs to the flag group of attributes. It represents the number of times the ACK flag bit is set to 1 for a given flow of packets sent in forward and backward directions. The flag group of attributes can be used to extract backscattered packets that result from a spoofed denial-of-service attack where the victim responds to a spoofed IP address of another victim used in the DDOS attack instance. The common headers of TCP packets used in flag attributes are SYN+ACK, RST, RST+ACK, and ACK [5]. The min\_seg\_size\_forward and Init\_Win\_bytes\_forward features belong to the flow descriptors group. They are useful for volumetric-flow monitoring and show the quantity of packets in a given direction, minimum, maximum, the mean and standard deviation of the packets as well as other descriptive statistics. These features are used to monitor and detect volume-based DDoS attacks that use a large amount of malicious traffic to bring down a resource.

In the second step of the algorithm, we utilize a deep neural network to classify benign and malicious network signals based on the selected features in Table I. We split the data

TABLE I: The top 10 features selected by the features selection method.

	TSI	feature
1	0.2361	ACK Flag Count
2	0.0847	Init_Win_bytes_forward
3	0.0847	PSH Flag Count
4	0.0713	act_data_pkt_fwd
5	0.0653	Idle Std
6	0.0535	Avg Bwd Segment Size
7	0.0490	Max Packet Length
8	0.0446	Packet Length Variance
9	0.0431	Total Backward Packets
10	0.0416	Subflow Bwd Packets

into training ans test subsets according to 80/20 ratio. The ANN is trained for 200 epochs with batch size 500 using the root mean square propagation optimizer. The summary of the results is presented in Table II. The test results show accuracy of 100% in distinguishing between the benign and DDoS traffic signals. Since the model accuracy is calculated based on the test set - independent of the training set - the chances of overfitting are minimal. Our result compares favorably with similar results in the literature. In addition, we trained and tested the benchmark methods SVM and LR which produced accuracy rate of 98.41% and 99.67% respectively. In addition, the TSNN method achieves perfect performance in precision and recall. Note that the benchmark methods were trained on the full dataset using all 63 features. Despite utilizing fewer features, the TSNN algorithm outperformed the benchmark methods.

TABLE II: Summary of the experiments.

Algorithm	# features	Accuracy rate	Precision	Recall
TSNN	10	100%	100%	100%
SVM	63	98.41%	98.86%	91.05%
LR	63	99.67%	100%	97.89%

One the main issues with the modern machine learning methods is interpretability. Algorithms that are trained on large datasets with a big number of features produce black-box models that are often impossible to humanly comprehend. Therefore, reducing the number of features required to build a model provides a valuable advantage. Models with fewer number of features such as TSNN have better interpretability.

## V. CONCLUSION

The advent of ubiquitous network based technologies has increased the associated vulnerabilities. The need for effective network protection tools has never been greater. In this paper, we propose an AI-based IDS that is capable of distinguishing between regular and DDoS traffic. The proposed method fuses an advanced feature selection technique together with deep learning to produce a simple yet efficient IDS. The proposed algorithm produces 100% accuracy on the tested dataset.

Despite the encouraging results there remains more work to be done to improve the proposed algorithm. Issues such as encrypted payload must be addressed. Other variables such as flow features and packet headers should also be considered. Although additional numerical experiments are required to validate the proposed method, the initial results offer a promising avenue for further research.

#### REFERENCES

- Ahmad, I., Basheri, M., Iqbal, M. J., & Rahim, A. (2018). Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. IEEE access, 6, 33789-33795.
- [2] Ambusaidi, M. A., He, X., Nanda, P., & Tan, Z. (2016). Building an intrusion detection system using a filter-based feature selection algorithm. IEEE transactions on computers, 65(10), 2986-2998.
- [3] Amiri, F., Yousefi, M. R., Lucas, C., Shakery, A., & Yazdani, N. (2011). Mutual information-based feature selection for intrusion detection systems. Journal of Network and Computer Applications, 34(4), 1184-1199.
- [4] Chollet, F., and others (2015). Keras. https://keras.io
- [5] Fachkha, C., Bou-Harb, E., & Debbabi, M. (2015). On the inference and prediction of DDoS campaigns. Wireless Communications and Mobile Computing, 15(6), 1066-1078.
- [6] Fang, W., Tan, X., & Wilbur, D. (2020). Application of intrusion detection technology in network safety based on machine learning. Safety Science, 124, 104604.
- [7] Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020). Multistage optimized machine learning framework for network intrusion detection. IEEE Transactions on Network and Service Management.
- [8] Kamalov, F., & Denisov, D. (2020). Gamma distribution-based sampling for imbalanced data. Knowledge-Based Systems, 207, 106368.
- [9] Kamalov, F., Moussa, S., Zgheib, R., & Mashaal, O. (2020, December). Feature selection for intrusion detection systems. In 2020 13th International Symposium on Computational Intelligence and Design (ISCID) (pp. 265-269). IEEE.
- [10] Kamalov, F. (2018, December). Sensitivity analysis for feature selection. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1466-1470). IEEE.
- [11] Kamalov, F., & Thabtah, F. (2017). A feature selection method based on ranked vector scores of features for classification. Annals of Data Science, 4(4), 483-502.
- [12] Kasongo, S. M., & Sun, Y. (2020). Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset. Journal of Big Data, 7(1), 1-20.
- [13] Homma, T., & Saltelli, A. (1996). Importance measures in global sensitivity analysis of model output, Reliability Engrg. System Safety, 52(1), 1-17.
- [14] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.
- [15] Salman, T., Bhamare, D., Erbad, A., Jain, R., & Samaka, M. (2017, June). Machine learning for anomaly detection and categorization in multi-cloud environments. In 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud) (pp. 97-103). IEEE.
- [16] Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Computer physics communications, 181(2), 259-270.
- [17] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018, January). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In ICISSP (pp. 108-116).
- [18] Sobol', I. Y. M. (2007). Global sensitivity indices for the investigation of nonlinear mathematical models. Matematicheskoe modelirovanie, 19(11), 23-24.
- [19] Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. Information Sciences, 513, 429-441.
- [20] Thabtah, F., Kamalov, F., Hammoud, S., & Shahamiri, S. R. (2020). Least Loss: A simplified filter method for feature selection. Information Sciences, 534, 1-15.
- [21] Zuech, R., & Khoshgoftaar, T. M. (2015). A survey on feature selection for intrusion detection. In Proceedings of the 21st ISSAT International Conference on Reliability and Quality in Design (pp. 150-155).