An Ensemble approach to Big Data Security (Cyber Security)

Manzoor Ahmed Hashmani[1]
High Performance Cloud Computing Center (HPC3)
Center for Research in Data Science (CeRDaS)
Universiti Teknologi PETRONAS
Sri Iskandar, Malaysia

Syed Muslim Jameel[2], Aidarus M. Ibrahim[3],
Maryam Zaffar[4]
Department of Computer and Information Sciences
Universiti Teknologi PETRONAS
Sri Iskandar, Malaysia

Kamran Raza[5]
IQRA University, Karachi, Pakistan

Abstract—**In the past, information safety was centered on event correlation designed for observing and spotting previously identified attacks. Due to the dynamic nature of multidimensional cyber-attacks, these models are no more acceptable. Specifically, these attacks use different strategies and procedures to find their way into and out of an organization. Traditional methods have reached their limit and thus new approaches are needed to find a solution for arising issues and challenges for big data security. To understand the current problem, we critically reviewed the literature related to big data security and the solutions proposed by the scientific community. In this paper, an ensemble approach for big data cybersecurity is proposed. To evaluate our approach, the given benchmark data is fed to three different classifiers namely to a k-nearest neighbor (KNN), support vector machine (SVM), multilayer perceptron (MLP) and the output of the single classifiers were compared to ensemble approach of the three classifiers. The reported results show that the ensemble approach for big data cybersecurity performs better than the single classifiers.**

Keywords—Big data; cyber security; benign, malicious; ensemble approach; Support Vector Machine (SVM); Receiver Operating Characteristic (ROC); Features (F)

I. INTRODUCTION

The progress in the current technology has embarked concerns about the risks to data related having weak security issues such as a virus, malware and compromising systems and services [1]. Lack of all aspects of data security may result compromised data in terms of confidentiality, integrity, and availability of data to outsiders [2]. A lot of efforts have been made to deploy cybersecurity monitoring which significantly worked over the last decade, yet these systems face challenges and issues [3], [1]. For example, Host-based security system and intrusion detection system were proposed to provide protection from the attacks, however, these systems failed to capture the new sophisticated attacks having unknown signatures. Moreover, some commercial systems for monitoring were proposed. These systems include Ganglia, Nagios, and Zabbix. They have provided a quick solution to security problems which have impact system performance. Though, subtle attacks were not detected by these systems [4].

To protect the big data in universal resource locator (URL), different methods for web filtering were deployed. For example, proxy servers are another mitigation approach for brows-able space of the internet [4].

Other methods were proposed for malicious URL detection. The popular method to detect malicious URL is the blacklist method and it is extremely fast and very easy to implement [5]. However, this technique suffers from non-trivial false positive and it is difficult for it to maintain an exhaustive list of malicious URLs [6]. Furthermore, Signature-based detection technique (IDS) is capable to identify the malicious pattern from the already defined type of pattern. However, this technique is not capable to identify new type of malicious attack. Heuristic approach detects the possible malicious pattern through the intelligent guesswork, heuristic approach builds rules (rule of thumb) from the experiences instead of any predetermined formula, although due to lack of rules in most cases, it suffers it accuracy to identify correct malicious pattern [6].

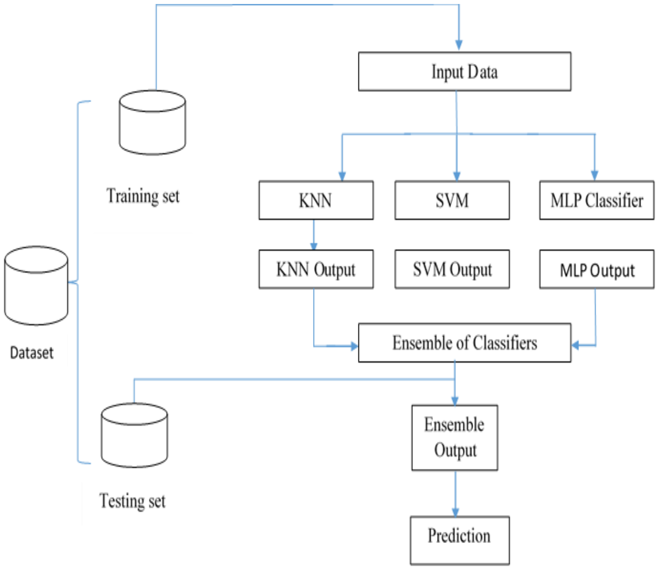

Fig. 1. Proposed Method for Big-Data Cybersecurity.

In this paper, we have proposed an ensemble approach for cybersecurity. In the experimental settings, the data were divided into training and testing sets. The training data were fed to the k-nearest neighbor (KNN), support vector machine, multilayer perceptron (MLP) individually. The output of these single classifiers was combined to form an ensemble approach.

II. Material and Methods

In this section, a benchmark dataset that consists of about 3.2 million features were used. We used data example from already available dataset. The data set is extracted through feature extraction mechanism from a large mail provider (real-time feed supplies 6000-7000 spam and phishing URL per day). The study [7] provides the complete detail of dataset extraction and preparation. The methods used in this study are explained in detail in the following subsections.

A. K-Nearest Neighbors

K nearest neighbors is nonparametric classification that stores available data and classify new data based on how similar they are in terms of distance. In the early 1970's, KNN is considered as one of the most prominent nonparametric techniques in statistical estimation and pattern recognition [8],[9].

B. Support Vector Machine (SVM)

In machine learning, support vector machines supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting).

C. Multilayer Perceptron (MLP)

Multilayer perceptron (MLP) can be designed by connecting the individual perceptron into neural network-based architecture. MLP is recognize as category of feedforward Artificial Neural Network because all input and intermediate layers provide input to their succeeding layers [9].

D. Ensemble Approach

Few recent studies in machine learning domain investigate the comparative analysis among single and ensemble classifiers. Through the variety of experimental results, these studies conclude that in most of the cases the ensemble approaches improve the classification performance over single classifier [10]. However, the effect of ensemble approach for the Big Data security is unknown. These integrated approaches based on diverse classification techniques and could attain the unidentical rate of accurate classified individuals, which ultimately results in more reliable, specific and accurate classification outputs than single classifier approach. A study [11] discuss the core parameters which enhance the performance of ensemble approach over single classifier, this study discusses the statistical, representational and computational interpretation and provides justification for better performance in ensemble classifiers. However, in ensemble classifiers, various critical parameters (for example, sum, product, minimum, maximum, average, Byes, dempster Shafer and decision template) tuning are mandatory for significant improvement in classification results. Fig. 1, defines the proposed ensemble-based approach for big-data cybersecurity.

E. Performance Evaluation

The proposed approach was evaluated using well-known dataset (training and testing. In training part, the parameters of the single classifiers reach to its optimal values (which are near to their target function) from the hypothetical values, and this followed by testing set to validate the performance of the classifiers. This approach reduces the bias and increases the generalization of the reported results. To evaluate and verify the classification performance, a receiver operating characteristics (ROC) were used. A t-test was also applied to see whether benign and malicious URLs are different to validate our results through the statistical analysis of two population mean.

III. Results and Discussion

The comparative analysis among the single and ensemble approaches are shown in, Table 1. The reports result in terms of accuracy of the classification of the two classes (benign versus malicious) are 0.9867 of KNN, 0.9867 of SVM, 0.9833 of MLP and 0.993 of ensemble approach. The classification performance results clearly show that the proposed ensemble approach is slightly higher than the single classifiers.

TABLE I. PERFORMANCE F THE SINGLE AND ENSEMBLE APPROACH

No.	Methods	Accuracy (%)
1	KNN	0.9867
2	SVM	0.9867
3	MLP	0.9833
3	Ensemble approach	0.993

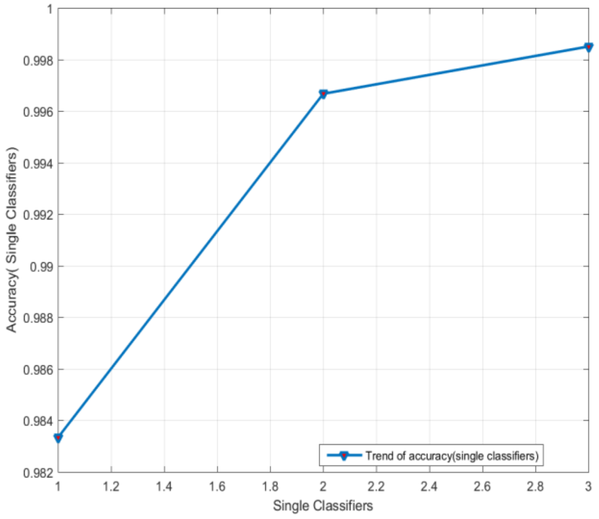

Fig. 2. Classification Performance of Single Classifiers (Benign v/s Malicious).

Fig.2 shows the performance of the single classifiers in terms of accuracy. We have plotted together with the reported accuracy of the single classifiers. The range of the reported accuracy of the single classifies are in between 0.983 to 0.9867. The line graph starts growing up until it reaches 0.09867 and then turns in to constant and this indicates that SVM and KNN are better in classification performance compared to MLP classifier.

Fig.3. shows, results of the methods those are necessary to combine different classifiers and make the ensemble. The x-axis of the figure represents nine different methods from 1 to 9 (as majority voting, maximum, sum, minimum, average, product, Bayes, decision template and Dempster-Shafer) respectively. The y-axis represents the reported accuracy of the different methods. Our reported results show that Dempster-Shafer is superior in ensemble approach compared to other remaining methods.

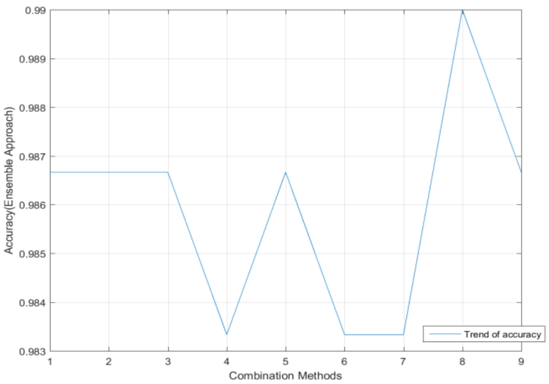

Fig. 3. Results of combination methods in terms of ensemble approach.

Fig. 4. describes the comparison of ROC curve performance of single classifiers and ensemble approach. The reported output shows that ensemble approach is higher than the other methods and this stresses the reported results in Table 1. The same idea was applied to differentiate between benign URLs and malicious URLs. Table 2 indicates the significant difference between benign and malicious features through the mean and standard deviation (having p-value not more than 0.0001). However, all feature (benign and malicious) are not identical.

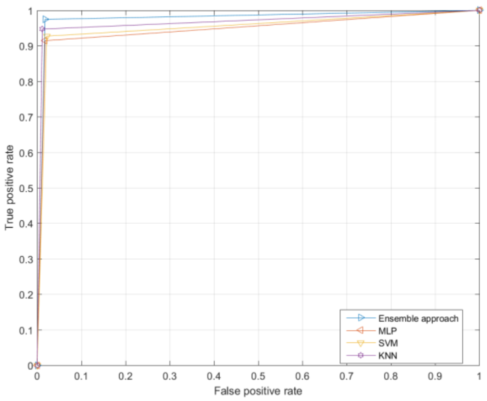

Fig. 4. Comparison of ROC curve performance of single classifiers and ensemble approach.

TABLE II. PERFORMANCE OF THE SINGLE AND ENSEMBLE APPROACH

F.	**Benign**	**Malicious**	**T value**
F. 1	7.087E-02 ± 4.873E-02	6.813E-02 ± 3.334E-02	0.328
F. 2	8.069E-02 ± 3.250E-02	8.855E-02 ± 3.812E-02	1.11
F. 3	0.136 ± 4.670E-02	0.118 ± 4.754E-02	2.00
F. 4	0.455 ± 0.353	0.536 ± 0.347	1.16
F. 5	0.503 ± 0.401	0.539 ± 0.392	0.460
F. 6	0.198 ± 0.250	0.333 ± 0.324	2.33
F. 7	0.123 ± 0.156	0.158 ± 0.152	1.12
F. 8	2.675E-02 ± 4.958E-02	3.006E-02 ± 5.465E-02	0.317
F. 9	1.420E-02 ± 3.466E-02	2.319E-02 ± 4.356E-02	1.14
F 10	3.556E-02 ± 6.019E-02	2.444E-02 $\pm \mathbf{4.072E-02}$	1.08

IV. CONCLUSION

Industrial Revolution IR 4.0 Big Data is considered as an opportunity to provide a more reliable and accurate source for business intelligence [12]. However, the versatile characteristics of Big Data possesses the potential to compromise the reliability and integrity of Big Data (which in result may degrades performance accuracy). Big Data security is considered as one of the serious challenges for researchers. Therefore, in this study, we have proposed a more reliable and accurate ensemble-based approach to classify benign and malicious activities to identify and prevent the possible cyber threat. Our proposed approach is highly accurate and able to classify (between benign versus malicious) an accuracy of 0.993. In future, this study will be further investigated to identify the threat pattern in cybersecurity.

REFERENCES

[1] Y. Ashibani and Q. H. Mahmoud, "Cyber-physical systems security: Analysis, challenges, and solutions," Computers & Security, vol. 68, pp. 81-97, 2017.

[2] C. Everett, "Big data–the future of cyber-security or it's the latest threat?," Computer Fraud & Security, vol. 2015, pp. 14-17, 2015.

[3] A. M. AlMadahkah, "Big Data In computer Cyber Security Systems," International Journal of Computer Science and Network Security (IJCSNS), vol. 16, p. 56, 2016.

[4] M. Mayhew, M. Atighetchi, A. Adler, and R. Greenstadt, "Use of machine learning in big data analytics for insider threat detection," in Military Communications Conference, MILCOM 2015-2015 IEEE, 2015, pp. 915-922.

[5] P. Vinod, R. Jaipur, V. Laxmi, and M. Gaur, "Survey on malware detection methods," in Proceedings of the 3rd Hackers' Workshop on the computer and internet security (IITKHACK'09), 2009, pp. 74-79.

[6] A. Sirageldin, B. B. Baharudin, and L. T. Jung, "Malicious Web Page Detection: A Machine Learning Approach," in Advances in Computer Science and its Applications, ed: Springer, 2014, pp. 217-224.

[7] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Identifying suspicious URLs: an application of large-scale online learning," in Proceedings of the 26th annual international conference on machine learning, 2009, pp. 681-688.

[8] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in Advances in neural information processing systems, 2006, pp. 1473-1480.

[9] L. I. Kuncheva, Combining pattern classifiers: methods and algorithms: John Wiley & Sons, 2004.

[10] T. G. Dietterich, "Ensemble learning," The handbook of brain theory and neural networks, vol. 2, pp. 110-125, 2002.

[11] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[12] S.M. Jameel, M.A. Hashmani, H. Alhussain, and A. Budiman, 2018, June. A Fully Adaptive Image Classification Approach for Industrial Revolution 4.0. In International Conference of Reliable Information and Communication Technology (pp. 311-321). Springer, Cham.