IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 6.252

Volume 2, Issue 2, July 2022

Anomaly Detection for Web Log Data Analysis

Madhur Narwan[1] and Ritu Kadyan[2]

Student, Department of Computer Science Engineering[1]

Assistant Professor, Department of Computer Science Engineering[2]

School of Engineering & Technology, Soldha, Bahadurgarh, Haryana, India

Abstract: *Many methods have been developed to protect web servers against attacks. Anomaly detection methods rely on generic user models and application behaviour, which interpret departures as indications of potentially dangerous behaviour from the established pattern. In this report, we conducted the use of a systematic review of the anomaly detection methods to prevent and identify web assaults; in particular, we utilised Kitchen ham's standard approach for conducting a organized analysis of literature in the computer science area. Logs that record system abnormal states (anomaly logs) can be regarded as outliers, and the improved PCA algorithm has relatively high accuracy in outlier detection methods. Therefore, we use improved algorithm to detect anomalies in the log data. However, there are some problems when using the improved PCA algorithm to detect anomalies, three of which are: excessive vector dimension leads to inefficient kNN algorithm, unlabeled log data cannot support the kNN algorithm, and the imbalance of the number of log data distorts the classification decision of kNN algorithm. In order to solve these three problems, we propose an efficient log anomaly detection method based on an improved PCA algorithm with an automatically labeled sample set. This method first proposes a log parsing method based on N-gram and frequent pattern mining (FPM) method, which reduces the dimension of the log vector converted with Term frequency. Inverse Document Frequency (TF-IDF) technology. Then we use clustering and self-training method to get labeled log data sample set from historical logs automatically. Finally, we improve the PCA algorithm using average weighting technology, which improves the accuracy of the PCA algorithm on unbalanced samples. The method in this article is validated on four log datasets with different types. The maximum recall rate & accuracy achieved for BGL dataset is 100 % & 97.62 % respectively. Similarly maximum F1-score achieved for Spirit dataset is 98.19 %. The accuracy, recall rate and F1-Score for Improved PCA Ensemble technique is 97.62 %, 100 % and 96.55 % for BGL/2 Log Set Data. Similarly, the accuracy, recall rate and F1-Score for Improved PCA Ensemble technique is 97.60 %, 98.79 % and 98.19 % respectively for Spirit/2 log set data.*

Keywords: Frequent Pattern Mining, PCA, KNN Algorithm

I. INTRODUCTION

Due to various their high value, Web servers are gradually becoming targets for assaults as the information technology sector advances. SQL injection and cross-site scripting (XSS) threats have been increasingly common in recent years, which is why Web security has received more attention from academic and industry communities. Anomaly is a term used in internet security research. The analysis of log data is used in web detection. Log files, as crucial recording data, may reveal extensive information at the time of system operation and may be used to trace the majority of assaults. However, log systems create a lot of data, and critical information might be lost in the shuffle. Furthermore, due to the ever-changing nature of assaults and hacking techniques, gathering anomaly data has become increasingly complex, leading to the current problem that manual log file analysis is inadequate to meet log testing standards.

II. PROPOSED METHODOLOGY

We propose a log-put together peculiarity recognition strategy based with respect to further developed kNN with a consequently marked example set. The general structure of the technique is displayed in Fig. 4.1.It essentially comprises of 3 stages: parsing &vectorization of log information, programmed development of information test setting via marks, and peculiarity identification via the further developed K.N.N.

2.1 Log Parsing Method

Logs are simple texts with elements that may vary from one instance to the next. The words "Connection from 10.10.34.12 closed" and "Connection from 10.10.34.13 closed" are considered constant parts in the logs "Connection from 10.10.34.12 closed" and "Connection from 10.10.34.13 closed," for example, because they never change, but the separated parts are known as variability subsections because they are fixing. Although developers specify constant components in source codes, variable portions (such as port numbers and IP addresses) are sometimes dynamically generated and hence unsuitable for anomaly detection. The purpose of log parsing is to extract constants from variable items and generate a well-defined log event (for example, "Connection from * ended"). Both cluster-based and heuristic-based log parsing techniques exist. Clustering-based log parsers determine the distances between logs in the first phase, and clustering algorithms are then used to organize the logs into discrete groups in the second phase. Each cluster generates a template for an event. In heuristic-based approaches, the number of times each word appears in each log position. Common words are chosen and produced as event candidates. Last but not least, decide which candidates will be registered as events. In pre-work, we created and compared four log parsers [24]. We also made an open-source log parsing toolkit available online, which we used to parse raw logs into log events for our research.

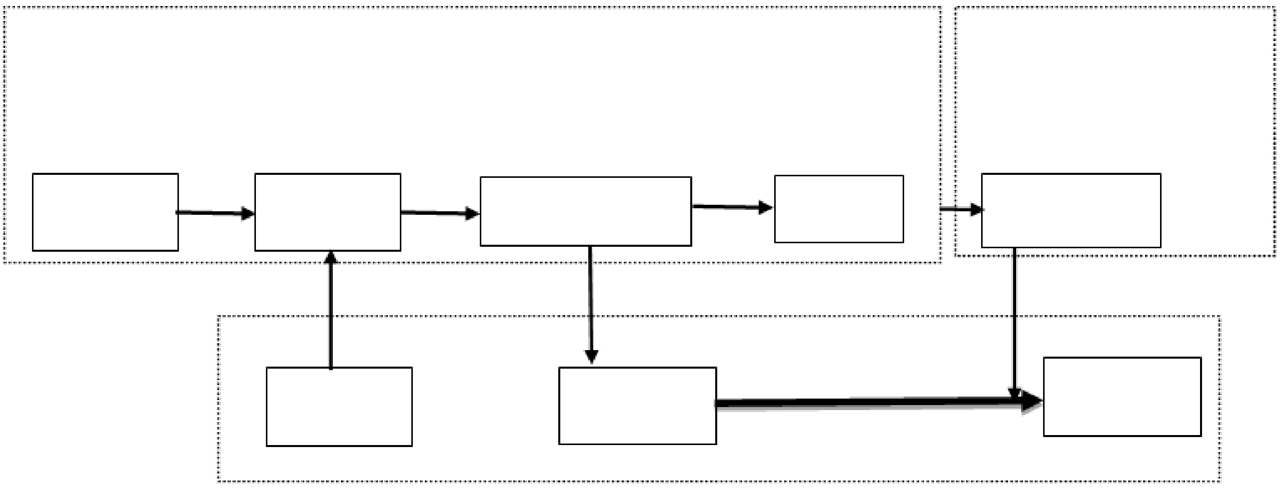

Fig. 2.1 Proposed Technique Framework

2.1.1 Parsing &Vectorization of Log-Data

Our log parsing technique principally incorporates two stages: (1) we propose a strategy dependent on N-gram and FPM to remove invariant part from log information; and (2) we partition logs with a similar invariant part into a gathering and convert each log line into a vector utilizing TF-IDF with word types in its gathering.

2.1.2 Log Vectorization Based on TF-IDF.

We include the word types in each gathering of log information and store them in the word set. For each log line l_j we work out the TF-IDF worth of each word w_i in l_j, as displayed in Equation (1).

$$W(w_i, l_j) = tf(w_i, l_j) * \log\left(\frac{N}{n_{w_i}} + 0.1\right) \dots (2.1)$$

Where $W(w_i, l_j)$ is the heaviness of the word w_i in log line l_j, $tf(w_i, l_j)$ is the word recurrence of word w_i in log line l_j, N is the size of in general log tests, and n_{w_i} is the quantity of log lines containing word w_i.

Log l_j is changed over into vector $[W(w_1, l_j), W(w_2, l_j), \ . \ . \ . \ W(w_N, l_j)]$. We convert each log line into a vector utilizing the above strategy, the vector aspect is the quantity of word types in the word set.

2.2 Anomaly Detection with the Improved PCA

In this segment, we identify abnormalities for tests to be identified, which mostly incorporates three stages: vectorization of log information, improvement of the PCA calculation, and inconsistency discovery. We will portray these three stages exhaustively.

2.2.1 Vectorization of Log Data

For the example to be recognized, we match it to invariants in the invariant layout library. Assuming that it matches, it will be changed over into a log vector as per logs with a similar invariant. Assuming there is no layout to coordinate, the N-gram and FPM-based technique will be utilized to acquire the new invariant format and convert it into a vector as indicated by Section 4.1.

We partition this log line into another gathering, and update the invariant layout library at the equivalent

2.2.2 Improvement of the PCA Algorithm

PCA is a viable element extraction procedure and information portrayal technique. It got a monstrous measure of consideration in design acknowledgment, picture handling and PC vision. PCA produces direct mixes of the first information and expects to observe the best vector space which addresses the circulation of datasets and lessens an enormous of informational collection to a lower aspect for getting successful outcomes. The component space characterized by eigenvector significantly diminishes the element of the first space, which lessens the calculation season of face identification and acknowledgment [24]. The principle objective in the PCA calculation lies in lessening the enormous elements of face information to the elements of the littlest spaces. PCA is considered as a multivariate investigation technique dependent on eigenvector. PCA calculation can be carried out by two fundamental techniques. The first is accomplished by decay of the eigen worth of information covariance framework [25]; then again, the subsequent one is performed by disintegration of a solitary worth of the information lattice. PCA results are communicated as a part or element scores and normalized part score weight. Hence, the came about picture can be communicated as the worth of every eigen faces it was identified with. In the wake of lessening the dimensionality of the dataset, the subsequent log messages are known as the eigen message (or eigenvectors). Through PCA eigen message strategy, every pixel is considered as log messages as a different aspect [26-27].

Contrasted and the customary PCA, the further developed calculation utilizes the mean of each class rather than the particular picture inside the class. Since the normal of each class is a straight blend of inside class pictures, the normal of each class holds an enormous number of varieties of the particular log messages [28]. All in all, the pressure interaction of each picture is more helpful for picture acknowledgment. What's more, one more clear benefit of the further developed PCA is that the preparation time is extraordinarily decreased.

The prominence of PCA comes from three significant properties. To begin with, it is the ideal straight plan for compacting a bunch of high dimensional vectors into a bunch of lower dimensional vectors and afterward reproducing the first set. Second, the model boundaries can be figured straightforwardly from the information - for instance by diagonalizing the example covariance framework. Third, pressure and decompression are simple tasks to perform given the model boundaries - they require just framework augmentation. A multi-dimensional hyper-space is frequently hard to imagine.

The PCA model can be addressed by:

$u_{mx_1} = W_{mx_d} x d_{x_1}$ ... (2.2)

Where u, an m-dimensional vector, is a projection of x - the first d-dimensional information vector(m << d).

PCA is variable-arranged technique, with changes a bunch of connected unique factors into a bunch of uncorrelated factors, called Principal Components (PC). These key parts are straight blends of the first factors. Via completing PCA we trust that a couple of PCs can clarify the majority of the variety in the first information. Along these lines, dimensionality can be decreased with basically no deficiency of data. It is self-evident, that if there should be an occurrence of deduced uncorrelated information, PCA has neither rhyme nor reason. If $x^T = (x_1, \dots \dots \dots . x_p)$indicates a p-dimensional vector of arbitrary factors with anticipated worth μ and covariance framework Σ, then, at that point, we attempt to track down a bunch of new, uncorrelated irregular factors, whose fluctuation diminishes with expanding j = 1, . . . , p. Consequently, for the main head part we search for a direct capacity $\alpha_1 T_X$having greatest fluctuation.

Next, we look for $\alpha_2 T_X$, uncorrelated with $\alpha_1 T_X$&greatest change, and so on Besides, α_j, j = 1, . . . , p is scaled to meet the imperative α_j^T, $j\alpha^j$ = 1. T The deviation of the PCs prompts the outcome, that the vectors of coefficientsα_1. . ……….$\alpha_p$for every PC are the eigenvectors of Σ relating toλ_1, . . . , λ_p eigen values, with $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_p$. As referenced previously, we trust that m << p PCs will represent a large portion of the difference in x. If the p × p framework of eigenvectors is signified by A = (α_1. . ……….α_p), the vector z of head parts can be composed as **z** = **A**^{T} x

2.2.3 Principal Components Analysis Based Anomaly Detection

Common datasets for interruption recognition are normally extremely huge and multidimensional. With the development of high velocity organizations and conveyed network based information concentrated applications putting away, handling, communicating, imagining and understanding the information is turning out to be more perplexing and costly. To handle the issue of high dimensional datasets, analysts have fostered a dimensionality decrease strategy known as Principal Component Analysis (PCA). In numerical terms, PCA is a method where n connected irregular factors are changed into $d \leq n$ uncorrelated factors. The uncorrelated factors are straight mixes of the first factors and can be utilized to communicate the information in a diminished structure. Normally, the primary head part of the change is the straight blend of the first factors with the biggest fluctuation. At the end of the day, the main head part is the projection on the course wherein the change of the projection is augmented. The second head part is the direct mix of the first factors with the second biggest fluctuation and symmetrical to the primary head part, etc. In numerous informational indexes, the initial a few head parts contribute the majority of the fluctuation in the first informational index, so the rest can be ignored with insignificant loss of the change for aspect decrease of the dataset. PCA has been broadly utilized in the space of picture pressure, design acknowledgment and interruption identification.

They estimated the distance of every perception from the focal point of the information for inconsistency identification. The distance is processed dependent on the amount of squares of the normalized head part scores.

2.3 Performance Parameter

This part discusses the mostly often used metrics for evaluating the various practices described in the evaluated literature.

2.3.1 Accuracy (ACC) is the clearly recognized payload ratio divided by total generated payloads.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

2.3.2 False Alert Rate (FAR) : The False Alert Rate (FAR) is the likelihood of a falsed alarming being raised. Whenever the trued magnitude is -ve, a positive result will be given

$$.FAR = \frac{FP}{FP+TN}$$

2.3.3 True-Negative or Rate/Specificity: T.N.R/Specificity is a metric that indicates the ratio of false -ve that are genuinely identified.

$$TNR = \frac{TN}{FP + TN}$$

2.3.4 TPR, Recalling Sensitivity & Detection Rate: The T.P.R, also known as Reminder, Sensitivity, or Detection Rate (DR), is a metric that indicates the ratio of true +ve that are accurately identification.

$$TPR = \frac{TP}{TP + FN}$$

2.3.5 Precision, or Positive-Predictive-Value (P.P.V): It is the proportion of malicious payloads accurately detecting to the total no. of malicious payloads.

$$PPV = \frac{TP}{TP + FP}$$

2.3.6 False-Negative-Rate: It is the percentage of positive test results linked with a test, or the conditional likelihood of a negative test result given the presence of the disease.

$$FNR = \frac{FN}{FN + TP}$$

2.3.7 F1-Score: It is a testing accuracy metric that considers dual accuracy and recall. The weighting H.M of 2 performance measurement, P & R, is used to calculate the F1-Score[176–178].

$$F1 - Score = \frac{1}{\alpha.\frac{1}{p} + (1-\alpha).\frac{1}{R}}$$

2.3.8 Classification error: The no. of samples incorrectly-classified (FP+FN) is referred to as "classification error," and it is computed using the formula:

$$CE = \frac{f}{n}.100$$

2.3.9 Matthews Correlation Coefficient

M.C.C is a proportion of the nature of twofold orders (two-class)[17]. M.C.C is a connection coefficient that offers a benefit bet^{n} - 1 and +1 for noticed & anticipated double groupings. A coefficient of +1 means a perfect forecast, a coefficient of 0 signifies no improvement over irregular expectation, and a coefficient of - 1 indicates complete conflict among forecast and perception. It's exactly the same thing as the phi-coefficient.

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

2.3.10 Area under Curve

A collector working trademark bend, or R.O.C bend, is a chart that showing how well a grouping models performs across all arrangement edges by plotting the T.P.R against the F.P.R. The A.U.C esteem is a 2-D area underneath the whole R.O.C bend that runs from 0 to 1 (model via 100% mistaken forecasts) and mirrors a model's capacity to recognize classes.

III. Simulation and Result Analysis

All The performance parameter of web log datasets is shown in the table 3.1In this table, basically four datasets BGL, Liberty, Spirit & thunderbird are used. The maximum recall rate & accuracy achieved for BGL dataset is 100 % & 97.62 % respectively. Similarly maximum F1-score achieved for Spirit dataset is 98.19 % .

Table 3.1: Performance Parameter of Web Log Datasets

Sr. No.	Dataset	Recall Rate	F 1-Score	Accuracy
1	BGL	**100 %**	96.55 %	**97.62 %**
2	Liberty	96.29 %	94.52 %	92.83 %
3	Spirit	98.79 %	**98.19 %**	97.60 %
4	Thunderbird	97.53 %	96.34 %	95.17 %

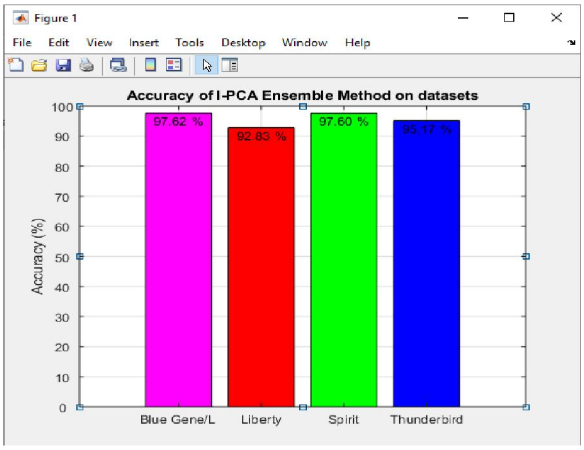

Fig 3.1 Computation of Accuracy of Improved PCA Ensemble Method on different Datasets

Computation of Accuracy of Improved PCA Ensemble Method on different Datasets is represents in fig 3.1 The accuracy of BGL, Liberty, Spirit & thunderbird is 97.62 %, 92.83 %, 97.80 % and 96.17 % respectively. The accuracy is enhanced by improved PCA technique.

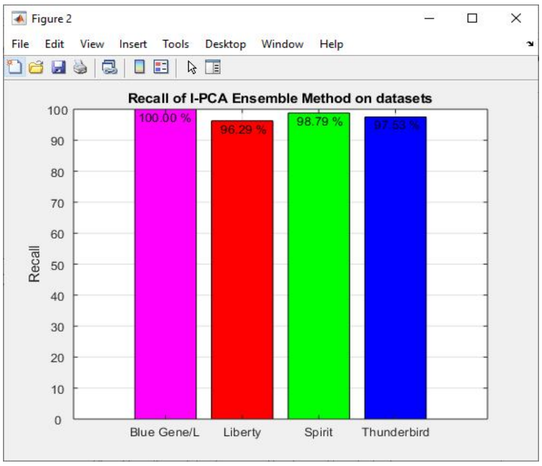

Fig 3.2 Computation of Recall Rate of Improved PCA Ensemble Method on different Datasets

Computation of Recall Rate of Improved PCA Ensemble Method on different Datasets is represents in fig 3.2. The Recall Rate of BGL, Liberty, Spirit & thunderbird is 100 %, 96.29 %, 97.79 % and 97.53 % respectively. The recall rate is enhanced by improved PCA technique.

Computation of F1-Score of Improved PCA Ensemble Method on different Datasets is represents in fig 4. The F-Score of BGL, Liberty, Spirit & thunderbird is 96.55 %, 94.52 %, 96.19 % and 96.34 % respectively. The F1-Score is enhanced by improved PCA technique.

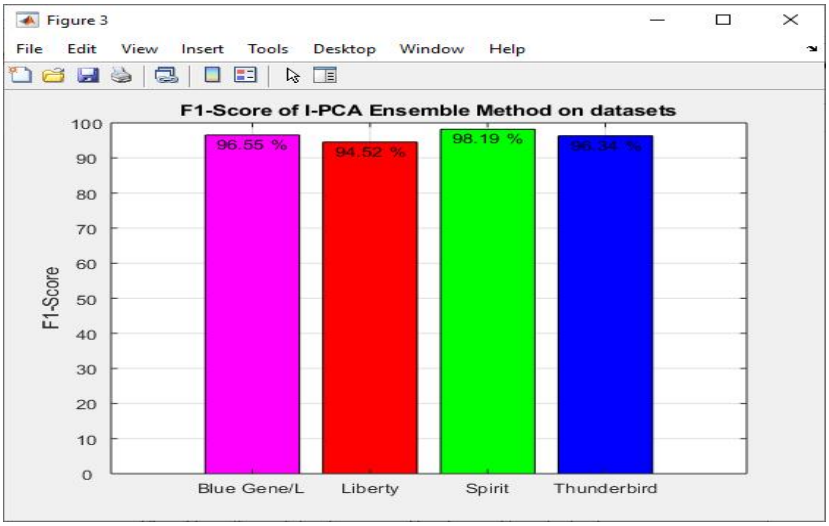

Fig 3.3 Computation of F1-Score of Improved PCA Ensemble Method on different Datasets

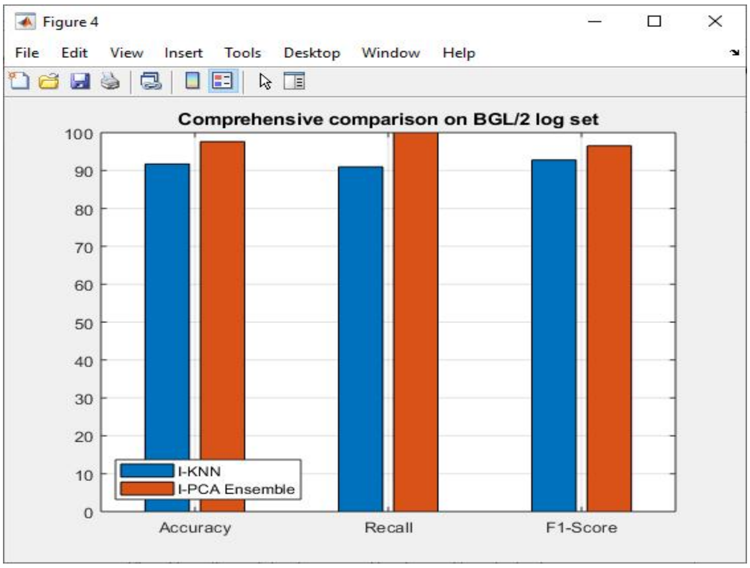

Fig 3.4 Comprehensive comparative analysis of I-KNN & I-PCA Ensemble on BGL/2 Log set

Comparative Analysis of I-KNN & I-PCA Ensemble on BGL/2 Log set is represents in fig 3.4. The accuracy, recall rate and F1-Score for Improved KNN is 91.73 %, 90.97 % and 92.784 % respectively. Similarly, the accuracy, recall rate and F1-Score for Improved PCA Ensemble technique are 97.62 %, 100 % and 96.55 % respectively. It represent that the performance parameter of BGL/2 Log set is enhance by improved PCA Technique.

Comparative Analysis of I-KNN & I-PCA Ensemble on Spirit/2 Log set is represents in fig. The accuracy, recall rate and F1-Score for Improved KNN are 91.97 %, 99.15 % and 96.13 % respectively. Similarly, the accuracy, recall rate and F1-Score for Improved PCA Ensemble technique is 97.60 %, 98.79 % and 98.19 % respectively. It represent that the performance parameter of Spirit/2 Log set is enhance by improved PCA Technique.

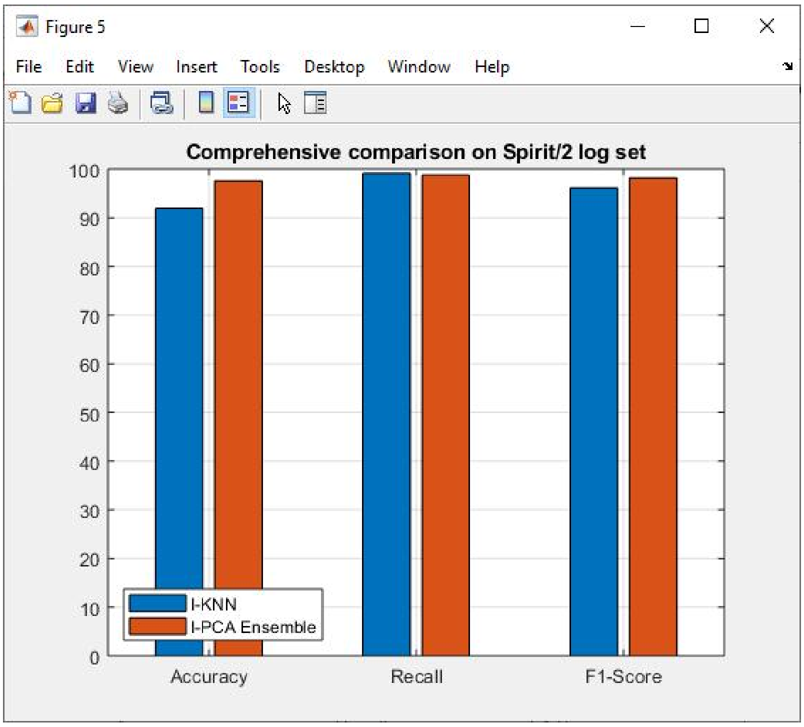

Fig 3.5 Comprehensive comparative analysis of I-KNN & I-PCA Ensemble on Spirit/2 Log set

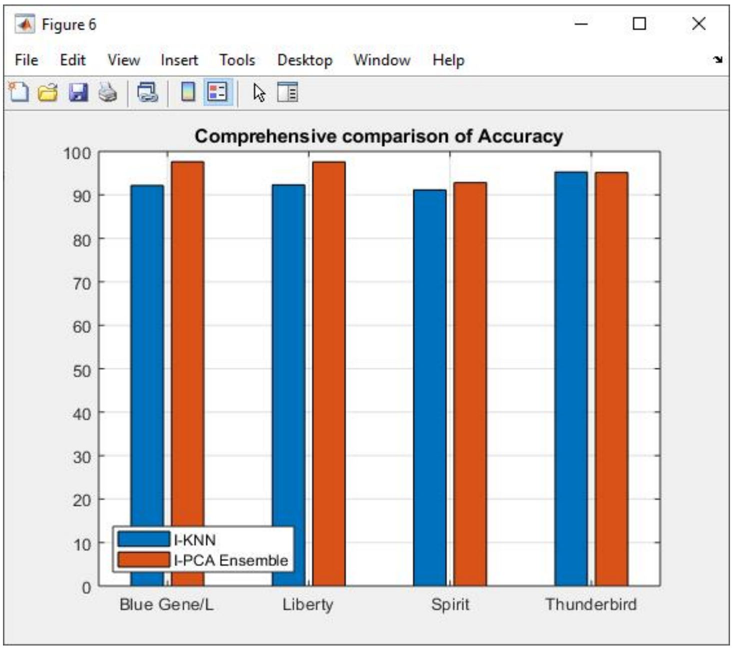

Fig 3.6 Comparative Analysis of I-KNN & I-PCA Ensemble for different Datasets.

Comparative Analysis of I-KNN & I-PCA Ensemble for different Datasets is represents in fig3.6. The accuracy for BGL, liberty, Spirit & thunderbird datasets is 92.15 %, 92.32 %, 91.14 % and 95.27 % respectively. Similarly, the accuracy for BGL, liberty, Spirit& thunderbird datasets is 97.62 %, 92.83 %, 97.60 % and 95.17 % respectively. The maximum accuracy achieved with improved PCA Ensemble algorithm for BGL, Liberty and Spirit is 97.62 % , 92.83 % and 97.60 % respectively. Similarly, the maximum accuracy achieved with improved KNN algorithm for thunderbird is 95.27 % .

Table 3.2 Comparative analysis of I-KNN & I-PCA Ensemble on BGL/2 Log set

Sr. No.	Parameter	I-KNN	I-PCA Ensemble
1	Accuracy	91.73 %	**97.62 %**
2	Recall Rate	90.97 %	**100 %**
3	F-1 Score	92.784 %	**96.55 %**

Table 3.3 Comparative analysis of I-KNN & I-PCA Ensemble on Spirt/2 Log set

Sr. No.	Parameter	I-KNN	I-PCA Ensemble
1	Accuracy	91.97 %	**97.60%**
2	Recall Rate	**99.15 %**	98.79 %
3	F-1 Score	96.13 %	**98.19 %**

Table 3.4 Comparative Analysis of I-KNN & I-PCA Ensemble for different Datasets

Sr. No.	Data Set	I-KNN	I-PCA Ensemble
1	BGL	92.15 %	**97.62%**
2	Liberty	92.32 %	**92.83 %**
3	Spirit	91.14 %	**97.60 %**
4.	Thunderbird	**95.27 %**	95.17 %

Table 3.3 and Table 3.4 represents comparative analysis of I-KNN & I-PCA Ensemble on BGL/2 Log set and Spirt/2 Log set. In case of BGL/2 Log Set, the accuracy, recall rate and F1-Score for Improved KNN is 91.73 %, 90.97 % and 92.784 % respectively. Similarly, the accuracy, recall rate and F1-Score for Improved PCA Ensemble technique is 97.62 %, 100 % and 96.55 % respectively. In case of Spirit/2 Log Set, the accuracy, recall rate and F1-Score for Improved KNN is 91.97 %, 99.15 % and 96.13 % respectively. Similarly, the accuracy, recall rate and F1-Score for Improved PCA Ensemble technique is 97.60 %, 98.79 % and 98.19 % respectively. Comparative Analysis of I-KNN & I-PCA Ensemble for different Datasets is represents in table 6. The accuracy for BGL, liberty, Spirit & thunderbird

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 2, July 2022

Impact Factor: 6.252

datasets is 92.15 %, 92.32 %, 91.14 % and 95.27 % respectively. Similarly, the accuracy for BGL, liberty, Spirit& thunderbird datasets is 97.62 %, 92.83 %, 97.60 % and 95.17 % respectively. The maximum accuracy achieved with improved PCA Ensemble algorithm for BGL, Liberty and Spirit is 97.62 %, 92.83 % and 97.60 % respectively. Similarly, the maximum accuracy achieved with improved KNN algorithm for thunderbird is 95.27 %.

IV. CONCLUSION

The Comparative Analysis of I-KNN & I-PCA Ensemble for different Datasets is represents in fig 5.6. The accuracy for BGL, liberty, Spirit & thunderbird datasets is 92.15 %, 92.32 %, 91.14 % and 95.27 % respectively. Similarly, the accuracy for BGL, liberty, and Spirit & thunderbird datasets is 97.62 %, 92.83 %, 97.60 % and 95.17 % respectively. The maximum accuracy achieved with improved PCA Ensemble algorithm for BGL, Liberty and Spirit is 97.62 % , 92.83 % and 97.60 % respectively. Similarly, the maximum accuracy achieved with improved KNN algorithm for thunderbird is 95.27 % .The maximum recall rate & accuracy achieved for BGL dataset is 100 % & 97.62 % respectively. Similarly maximum F1-score achieved for Spirit dataset is 98.19 %. The accuracy, recall rate and F1-Score for Improved PCA Ensemble technique is 97.62 %, 100 % and 96.55 % for BGL/2 Log Set Data. Similarly, the accuracy, recall rate and F1-Score for Improved PCA Ensemble technique is 97.60 %, 98.79 % and 98.19 % respectively for Spirit/2 log set data.

REFERENCES

[1]. Thang, T.M.; Nguyen, K.V. FDDA: A Framework For Fast Detecting Source Attack In Web Application DDoS Attack. In Proceedings of the Eighth International Symposium on Information and Communication Technology, NhaTrang, Vietnam, 7–8 December 2017; Association for Computing Machinery: New York, NY, USA, 2017; SoICT 2017; pp. 278–285.

[2]. Tripathi, N.; Hubballi, N. Slow Rate Denial of Service Attacks against HTTP/2 and Detection. Comput. Secur. 2018, 72, 255–272.

[3]. Najafabadi, M.M.; Khoshgoftaar, T.M.; Calvert, C.; Kemp, C. User Behavior Anomaly Detection for Application Layer DDoS Attacks. In Proceedings of the 2017 IEEE International Conference on Information Reuse and Integration (IRI), San Diego, CA, USA, 4–6 August 2017; pp. 154–161.

[4]. Zolotukhin, M.; Hämäläinen, T.; Kokkonen, T.; Siltanen, J. Increasing web service availability by detecting application-layer DDoS attacks in encrypted traffic. In Proceedings of the 2016 23rd International Conference on Telecommunications (ICT), Thessaloniki, Greece, 16–18 May 2016; pp. 1–6.

[5]. Shirani, P.; Azgomi, M.A.; Alrabaee, S. A method for intrusion detection in web services based on time series. In Proceedings of the 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE), Halifax, NS, Canada, 3–6 May 2015; pp. 836–841.

[6]. Tripathi, N.; Hubballi, N.; Singh, Y. How SecureareWeb Servers? An Empirical Study of Slow HTTP DoS Attacks and Detection. In Proceedings of the 2016 11th International Conference on Availability, Reliability and Security (ARES), Salzburg, Austria, 31 August–2 September 2016; pp. 454–463.

[7]. Wang, C.; Miu, T.T.N.; Luo, X.;Wang, J. SkyShield: A Sketch-Based Defense System Against Application Layer DDoS Attacks. IEEE Trans. Inf. Forensics Secur. 2018, 13, 559–573.

[8]. Wang, Y.; Liu, L.; Si, C.; Sun, B. A novel approach for countering application layer DDoS attacks. In Proceedings of the 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 25–26 March 2017; pp. 1814–1817.

[9]. Xie, Y.; Tang, S. Online Anomaly Detection Based on Web Usage Mining. In Proceedings of the 2012 IEEE 26th International Parallel and Distributed Processing SymposiumWorkshops PhD Forum, Shanghai, China, 21–25 May 2012; pp. 1177–1182.

[10]. Lin, H.; Cao, S.; Wu, J.; Cao, Z.; Wang, F. Identifying Application-Layer DDoS Attacks Based on Request Rhythm Matrices. IEEE Access 2019, 7, 164480–164491.

[11]. Xiao, R.; Su, J.; Du, X.; Jiang, J.; Lin, X.; Lin, L. SFAD: Toward effective anomaly detection based on session feature similarity. Knowl.-Based Syst. 2019, 165, 149–156.

[12]. Kozik, R.; Chora´s, M.; Hołubowicz,W. Evolutionary-based packets classification for anomaly detection in web layer. Secur. Commun. Netw. 2016, 9, 2901–2910.

[13]. Wang, L.; Cao, S.; Wan, L.; Wang, F. Web Anomaly Detection Based on Frequent Closed Episode Rules. In Proceedings of the 2017 IEEE Trustcom/BigDataSE/ICESS, Sydney, NSW, Australia, 1–4 August 2017; pp. 967–972.

[14]. Yuan, G.; Li, B.; Yao, Y.; Zhang, S. A deep learning enabled subspace spectral ensemble clustering approach for web anomaly detection. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 3896–3903.

[15]. Bronte, R.; Shahriar, H.; Haddad, H. Information Theoretic Anomaly Detection Framework for Web Application. In Proceedings of the 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC), Atlanta, GA, USA, 10–14 June 2016; Volume 2, pp. 394–399. [CrossRef]

[16]. Luo, Y.; Cheng, S.; Liu, C.; Jiang, F. PU Learning in Payload-based Web Anomaly Detection. In Proceedings of the 2018 Third International Conference on Security of Smart Cities, Industrial Control System and Communications (SSIC), Shanghai, China, 18–19 October 2018; pp. 1–5. [CrossRef]

[17]. Ren, X.; Hu, Y.; Kuang, W.; Souleymanou, M.B. A Web Attack Detection Technology Based on Bag of Words and Hidden Markov Model. In Proceedings of the 2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), Chengdu, China, 9–12 October 2018; pp. 526–531.

[18]. Kozik, R.; Chora´s, M.; Hołubowicz,W. HardeningWeb Applications against SQL Injection Attacks Using Anomaly Detection Approach. In Image Processing & Communications Challenges 6; Chora´s, R.S., Ed.; Springer International Publishing: Cham, Switzerland, 2015; pp. 285–292.

[19]. Maggi, F.; Robertson,W.; Kruegel, C.; Vigna,G. Protecting aMoving Target: Addressing Web Application Concept Drift. In Recent Advances in Intrusion Detection; Kirda, E., Jha, S., Balzarotti, D., Eds.; Springer:Berlin/Heidelberg, Germany, 2009; pp. 21–40.

[20]. Valeur, F.; Vigna, G.; Kruegel, C.; Kirda, E. An Anomaly-Driven Reverse Proxy for Web Applications. In Proceedings of the 2006 ACM Symposium on Applied Computing, Dijon, France, 23–27 April 2006; Association for Computing Machinery: New York, NY, USA, 2006; pp. 361–368.

[21]. Guangmin, L. Modeling Unknown Web Attacks in Network Anomaly Detection. In Proceedings of the 2008 Third International Conference on Convergence and Hybrid Information Technology, Busan, Korea, 11–13 November 2008; Volume 2, pp. 112–116.

[22]. Yu, S.; Guo, S.; Stojmenovic, I. Fool Me If You Can: Mimicking Attacks and Anti-Attacks in Cyberspace. IEEE Trans. Comput. 2015, 64, 139–151.

[23]. Sakib, M.N.; Huang, C. Using anomaly detection based techniques to detect HTTP-based botnet C traffic. In Proceedings of the 2016 IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia, 22–27 May 2016; pp. 1–6.

[24]. Medvet, E.; Bartoli, A. On the Effects of Learning Set Corruption in Anomaly-Based Detection of Web Defacements. In Detection of Intrusions and Malware, and Vulnerability Assessment; Hämmerli, M.B., Sommer, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 60–78.

[25]. Davanzo, G.; Medvet, E.; Bartoli, A. Anomaly detection techniques for a web defacement monitoring service. Expert Syst. Appl. 2011, 38, 12521–12530.

[26]. Juvonen, A.; Sipola, T.; Hämäläinen, T. Online anomaly detection using dimensionality reduction techniques for HTTP log analysis. Comput. Netw. 2015, 91, 46–56.

[27]. Wang,W.; Guyet, T.; Quiniou, R.; Cordier, M.O.; Masseglia, F.; Zhang, X. Autonomic Intrusion Detection. Know.-Based Syst. 2014, 70, 103–117.

[28]. Vartouni, A.M.; Kashi, S.S.; Teshnehlab, M. An anomaly detection method to detect web attacks using Stacked Auto-Encoder. In Proceedings of the 2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), Kerman, Iran, 28 February–2 March 2018; pp. 131–134.

[29]. Zolotukhin, M.; Hämäläinen, T.; Kokkonen, T.; Siltanen, J. Analysis of HTTP requests for anomaly detection of web attacks. In Proceedings of the 2014 World Ubiquitous Science Congress: 2014 IEEE 12th International Conference on Dependable, Autonomic and Secure Computing, DASC 2014, Dalian, China, 24–27 August 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 406–411.

[30]. Asselin, E.; Aguilar-Melchor, C.; Jakllari, G. Anomaly detection for web server log reduction: A simple yet efficient crawling based approach. In Proceedings of the 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, USA, 17–19 October 2016; pp. 586–590.

[31]. Zhang, S.; Li, B.; Li, J.; Zhang, M.; Chen, Y. A Novel Anomaly Detection Approach for Mitigating Web-Based Attacks Against Clouds. In Proceedings of the 2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing (CSCloud), New York, NY, USA, 3–5 November 2015; IEEE Computer Society: Piscataway, NJ, USA, 2015; pp. 289–294.

[32]. Zhang, M.; Lu, S.; Xu, B. An Anomaly Detection Method Based on Multi-models to Detect Web Attacks. In Proceedings of the 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 9–10 December 2017; Volume 2, pp. 404–409.

[33]. Parhizkar, E.; Abadi, M. OC-WAD: A one-class classifier ensemble approach for anomaly detection in web traffic. In Proceedings of the 2015 23rd Iranian Conference on Electrical Engineering, Tehran, Iran, 10–14 May 2015; pp. 631–636.

[34]. Kozik, R.; Choras, M. Adapting an Ensemble of One-Class Classifiers for aWeb-Layer Anomaly Detection System. In Proceedings of the 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 3PGCIC, Krakow, Poland, 4–6 November 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 724–729.

[35]. Cao, Q.; Qiao, Y.; Lyu, Z. Machine learning to detect anomalies in web log analysis. In Proceedings of the 2017 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, 13–16 December 2017; pp. 519–523.

[36]. Yu, J.; Tao, D.; Lin, Z. A hybrid web log based intrusion detection model. In Proceedings of the 2016 4th IEEE International Conference on Cloud Computing and Intelligence Systems, CCIS 2016, Beijing, China, 17–19 August 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 356–360.

[37]. Threepak, T.;Watcharapupong, A. Web attack detection using entropy-based analysis. In Proceedings of the International Conference on Information Networking, Phuket, Thailand, 10–12 February 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 244–247.

[38]. Swarnkar, M.; Hubballi, N. Rangegram: A novel payload based anomaly detection technique against web traffic. In Proceedings of the 2015 IEEE International Conference on Advanced Networks and Telecommuncations Systems (ANTS), Kolkata, India, 15–18 Decemnber 2015; pp. 1–6.

[39]. Xu, H.; Tao, L.; Lin, W.; Wu, Y.; Liu, J.; Wang, C. A model for website anomaly detection based on log analysis. In Proceedings of the 2014 IEEE 3rd International Conference on Cloud Computing andIntelligence Systems, Shenzhen, China, 27–29 November 2014; pp. 604–608.

[40]. Park, S.; Kim, M.; Lee, S. Anomaly Detection for HTTP Using Convolutional Autoencoders. IEEE Access 2018, 6, 70884–70901.

[41]. Chora´s, M.; Kozik, R. Machine learning techniques applied to detect cyberattacks on web applications. Log. J. IGPL 2014, 23, 45–56.

[42]. Tharshini, M.; Ragavinodini, M.; Senthilkumar, R. Access Log Anomaly Detection. In Proceedings of the 2017 Ninth International Conference on Advanced Computing (ICoAC), Chennai, India, 14–16 December 2017; pp. 375–381.

[43]. Kozik, R.; Chora´s, M.; Hołubowicz,W. Packets tokenization methods for web layer cyber security. Log. J. IGPL 2016, 25, 103–113.

[44]. Kamarudin, M.H.; Maple, C.; Watson, T.; Safa, N.S. A LogitBoost-Based Algorithm for Detecting Known and UnknownWeb Attacks. IEEE Access 2017, 5, 26190–26200.

[45]. Yu, Y.; Liu, G.; Yan, H.; Li, H.; Guan, H. Attention-Based Bi-LSTM Model for Anomalous HTTP Traffic Detection. In Proceedings of the 2018 15th International Conference on Service Systems and Service Management (ICSSSM), Hangzhou, China, 21–22 July 2018; pp. 1–6.

[46]. Nguyen, X.N.; Nguyen, D.T.; Vu, L.H. POCAD: A novel pay load-based one-class classifier for anomaly detection. In Proceedings of the 2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS), Danang, Vietnam, 14–16 September 2016; pp.

74–79.

[47]. Lu, L.; Zhu, X.; Zhang, X.; Liu, J.; Bhuiyan, M.Z.A.; Cui, G. One Intrusion Detection Method Based On Uniformed Conditional Dynamic Mutual Information. In Proceedings of the 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), New York, NY, USA , 1–3 August 2018; pp. 1236–1241.

[48]. Moustafa, N.; Misra, G.; Slay, J. Generalized Outlier Gaussian Mixture technique based on Automated Association Features for Simulating and DetectingWeb Application Attacks. IEEE Trans. Sustain. Comput. 2018, 1.

[49]. Alrawashdeh, K.; Purdy, C. Fast Activation Function Approach for Deep Learning Based Online Anomaly Intrusion Detection. In Proceedings of the 2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS), Omaha, NE, USA, 3–5 May 2018; pp. 5–13.

[50]. Kaur, R.; Bansal, M. Multidimensional attacks classification based on genetic algorithm and SVM. In Proceedings of the 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, 14–16 October 2016; pp. 561–565.

[51]. Angiulli, F.; Argento, L.; Furfaro, A. Exploiting N-Gram Location for Intrusion Detection. In Proceedings of the 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI), Vietrisul Mare, Italy, 9–11 November 2015; pp. 1093–1098.

[52]. Hiremagalore, S.; Barbará, D.; Fleck, D.; Powell, W.; Stavrou, A. transAD: An Anomaly Detection Network Intrusion Sensor for the Web. In Information Security; Chow, S.S.M., Camenisch, J., Hui, L.C.K., Yiu, S.M., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 477–489.

[53]. Favaretto, M.; Spolaor, R.; Conti, M.; Ferrante, M. You Surf so Strange Today: Anomaly Detection in Web Services via HMM and CTMC. In Green, Pervasive, and Cloud Computing; Au, M.H.A., Castiglione, A., Choo, K.K.R.;,Palmieri, F., Li, K.C., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 426–440.

[54]. Kozik, R.; Chorás, M. The http content segmentation method combined with adaboost classifier for web-layer anomaly detection system. Adv. Intell. Syst. Comput. 2017, 527, 555–563.

[55]. Kozik, R.; Chora´s, M.; Holubowicz, W.; Renk, R. Extreme Learning Machines for Web Layer Anomaly Detection. In Image Processing and Communications Challenges 8; Chora´s, R.S., Ed.; Springer International Publishing: Cham, Switzerland, 2017; pp. 226–233.