International Journal of Research in Engineering, Science and Management
Volume 7, Issue 4, April 2024
https://www.ijresm.com | ISSN (Online): 2581-5792

Machine Learning Based User Transaction Alert Using Decentralized Approach-SOS

S. Salini[1], R. Manoj[2*], M. S. Jayanth[3], M. Sanjay[4]
[1]Assistant Professor, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, India
[2,3,4]B.Tech. Student, Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, India

***Abstract*: To guarantee cyber security of an endeavor, regularly SIEM (Security Information and Event Management) framework is in put to normalize security occasions from diverse preventive advances and hail cautions. Examiners in the security operation center (SOC) explore the cautions to choose if it is really pernicious or not. In any case, for the most part the number of alarms is overpowering with lion's share of them being untrue positive and surpassing the SOC's capacity to handle all cautions. Since of this, potential malevolent assaults and compromised has may be missed. Machine learning is a practical approach to diminish the wrong positive rate and move forward the efficiency of SOC examiners. In this paper, we create a client- centric machine learning system for the cyber security operation center in genuine undertaking environment. We examine the normal information sources in SOC, their work stream, and how to use and prepare these information sets to construct an successful machine learning framework. The paper is focused on towards two bunches of perusers. The to begin with bunch is information researchers or machine learning analysts who do not have cyber security space information but need to construct machine learning frameworks for security operations center. The moment bunch of groups of onlookers are those cyber security professionals who have profound information and mastery in cyber security, but do not have machine learning encounters and wish to construct one by themselves. All through the paper, we utilize the framework we built in the Symantec SOC generation environment as a case to illustrate the total steps from information collection, name creation, include building, machine learning calculation choice, show execution assessments, to hazard score generation.**

***Keywords*: machine learning, user transaction, alert system, decentralized approach, anomaly detection, fraud detection, privacy-preserving techniques, distributed ledger technology, privacy-enhancing technologies.**

1. Introduction

Cyber security episodes will cause noteworthy budgetary and notoriety impacts on venture. In arrange to distinguish noxious exercises, the SIEM (Security Information and Event Management) framework is built in companies or government. The framework relates occasion logs from endpoint, firewalls, IDS/IPS (Interruption Detection/Prevention Framework), DLP (Data Loss Protection), DNS (Domain Name System), DHCP (Dynamic Host Configuration Protocol), Windows/Unix security occasions, VPN logs etc. The security occasions can be assembled into diverse categories [1]. The logs have terabytes of information each day.

From the security occasion logs, SOC (Security Operation Center) group creates so-called utilize cases with a pre-determined seriousness based on the analysts' encounters. They are ordinarily run the show based connecting one or more pointers from distinctive logs. These rules can be network/host based or time/frequency based.

If any pre-defined utilize case is activated, SIEM framework will produce an caution in genuine time. SOC examiners will at that point explore the cautions to choose whether the client related to the alarm is unsafe (a genuine positive) or not (untrue positive). If they discover the cautions to be suspicious from the examination, SOC investigators will make OTRS (Open-Source Ticket Request System) tickets. After starting examination, certain OTRS tickets will be heightened to level 2 examination framework (e.g., Co3 Framework) as extreme security occurrences for encourage examination and remediation by Occurrence Reaction Team.

However, SIEM regularly produces a part of the alarms, but with an exceptionally tall untrue positive rate. The number of alarms per day can be hundreds of thousands, much more than the capacity for the SOC to explore all of them. Since of this, SOC may select to examine as it were the alarms with tall seriousness or stifle the same sort of alarms.

This may possibly miss a few extreme assaults. Subsequently, a more shrewdly and programmed framework is required to distinguish unsafe users.

The machine learning framework sits in the center of SOC work stream, joins diverse occasion logs, SIEM cautions and SOC examination comes about and creates comprehensive client chance score for security operation center. Instep of straightforwardly burrowing into huge sum of SIEM cautions and attempting to discover needle in a bundle, SOC investigators can utilize the chance scores from machine learning framework to prioritize their examinations, beginning from the clients with most elevated dangers. This will enormously move forward their productivity, optimize their work line administration, and eventually improve the enterprise's security.

Specifically, our approach develops a system of client centric machine learning framework to assess client hazard based on alarm data. This approach can give security examiner a

*Corresponding author: manojcbmjesus@gmail.com

comprehensive hazard score of a client and security examiner can center on those clients with tall hazard scores.

2. Literature Survey

A Novel Intrusion Detection Method Using Deep Neural Network for In-Vehicle Network Security

Description: In this paper, we propose a novel interruption location method utilizing a profound neural organize (DNN). In the proposed method, in-vehicle arrange bundles traded between electronic control units (ECU) are prepared to extricate moo- dimensional highlights and utilized for separating ordinary and hacking bundles. The highlights perform in tall productive and moo complexity since they are created specifically from a bitstream over the arrange. The proposed method screens an trading parcel in the vehicular arrange whereas the highlight are prepared off-line, and gives a real-time reaction to the assault with a altogether tall discovery proportion in our tests.

Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection.

Description: Interruption location is one of the challenging issues experienced by the cutting edge organize security industry. A organize has to be ceaselessly observed for identifying arrangement infringement or suspicious activity. So an interruption discovery framework needs to be created which can screen organize for any hurtful exercises and create comes about to the administration specialist. Information mining can play a gigantic part in the improvement of a framework which can identify arrange interruption. Information mining is a method through which critical data can be extricated from colossal information storehouses. In arrange to spot interruption, the activity made in the arrange can be broadly categorized into taking after two categories- ordinary and bizarre. In our proposed paper, a few classification methods and machine learning calculations have been considered to categorize the arrange activity. Out of the classification procedures, we have found nine reasonable classifiers like Bayes Net, Calculated, IBK, J48, Portion, JRip, Arbitrary Tree, Irregular Woodland and REP Tree. Out of the a few machine learning calculations, we have worked on Boosting, Sacking and Mixing (Stacking) and compared their correctnesses as well. The comparison of these calculations has been performed utilizing WEKA device and recorded underneath concurring to certain execution measurements. Recreation of these classification models has been performed utilizing 10-fold cross approval. NSL-KDD based information set has been utilized for this recreation in WEKA.

A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection.

Description: Arrange assaults have ended up more unavoidable in the cyber world. There are different assaults such as refusal of benefit, checking, benefit acceleration that is expanding day by day driving towards the prerequisite of a more strong and versatile security procedures. Peculiarity location is the fundamental center of our paper. Support Vector Machine (SVM) is one of the great classification calculation connected extraordinarily for interruption discovery. In any case, its execution can be essentially moved forward when it is connected in integration with other classifiers. In this paper, we have performed a comparative examination of SVM classifier's execution when it is stacked with other classifiers like Bayes Net, Ada Boost, Calculated, IBK, J48, Arbitrary Woodland, JRip, One R and Straightforward Cart. Multi-Classifier calculation have way better classification control when compared to a single classifier calculation extraordinarily for identifying moo recurrence assaults such as figure secret word, rootkits, spyware etc. Our preparatory investigation over NSL-KDD'99 dataset appears that stacking of SVM and Irregular Timberland gives the best execution with precision of around 97.50% which clearly way better than SVM (91.81%).

A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection.

Description: This overview paper portrays a centered writing study of machine learning (ML) and data mining (DM) strategies for cyber analytics in bolster of interruption location. Brief instructional exercise depictions of each ML/DM strategy are given. Based on the number of citations or the pertinence of an developing strategy, papers speaking to each strategy were distinguished, studied, and summarized. Since information are so critical in ML/DM approaches, a few well-known cyber information sets utilized in ML/DM are portrayed. The complexity of ML/DM calculations is tended to, talk of challenges for utilizing ML/DM for cyber security is displayed, and a few proposals on when to utilize a given strategy are given.

3. Problem Statement

In the quickly advancing scene of computerized exchanges, guaranteeing the security and judgment of client exchanges is fundamental. In any case, existing centralized exchange checking frameworks are helpless to single focuses of disappointment, information breaches, and protection concerns. To address these challenges, there is a basic require for a decentralized approach to client exchange caution frameworks that leverages machine learning methods whereas joining an SOS (Spare Our Souls) highlight for crisis situations.

The issue at hand is to plan and execute a machine learning-based client exchange alarm framework that works in a decentralized way, utilizing blockchain or conveyed record innovation. This framework ought to be able of recognizing bizarre or suspicious exchanges in real-time and creating alarms to inform clients of potential dangers. Furthermore, it ought to coordinated an SOS usefulness that permits clients to trigger crisis alarms in cases of false or unauthorized exchanges, subsequently improving client security and security

A. Existing System

Most approaches to security in the venture have centered on securing the arrange foundation with no or small consideration to conclusion clients. As a result, conventional security capacities and related gadgets, such as firewalls and interruption location and anticipation gadgets, bargain basically with arrange level assurance. In spite of the fact that still portion of the by and large security story, such an approach has

restrictions in light of the unused security challenges depicted in the past section.

Data Examination for Arrange Cyber-Security centers on checking and analyzing arrange activity information, with the purposeful of anticipating, or rapidly distinguishing, malevolent movement. Chance values were presented in an information security management system (ISMS) and quantitative assessment was conducted for nitty gritty hazard appraisal. The quantitative assessment appeared that the proposed countermeasures may decrease chance to a few degree. Examination into the cost-effectiveness of the proposed countermeasures is an imperative future work. It gives clients with assault data such as the sort of assault, recurrence, and target have ID and source have ID. Ten et al. proposed a cyber-security system of the SCADA framework as a basic foundation utilizing real-time checking, irregularity location, and affect examination with an assault tree-based technique, and relief strategies.

Disadvantages:

- Firewalls can be difficult to configure correctly.
- Incorrectly configured firewalls may block users from performing actions on the Internet, until the firewall configured correctly.
- Makes the system slower than before.
- Need to keep updating the new software in order to keep security up to date.
- Could be costly for average user.

4. Methodology

Issue Definition and Scope Identification:

- Clearly characterize the targets of the exchange alarm system.
- Identify the sorts of exchanges to screen and the potential dangers or inconsistencies to detect.
- Information Collection and Preprocessing:
- Collect exchange information from decentralized sources such as blockchain systems or disseminated ledgers.
- Preprocess the information to handle lost values, normalize numerical highlights, and encode categorical variables.
- Ensure information protection and security measures are input, particularly in decentralized environments.

Include Engineering:

- Extract pertinent highlights from exchange information that can be utilized to prepare machine learning models.
- Consider highlights such as exchange sum, timestamp, sender/receiver addresses, exchange sort, etc.
- Explore procedures like dimensionality lessening to handle high-dimensional information efficiently.

Show Choice and Training:

- Choose fitting machine learning calculations for inconsistency location, such as Separation Woodland, One-Class SVM, or autoencoders.
- Train the chosen models utilizing chronicled exchange information whereas guaranteeing a decentralized preparing approach, conceivably utilizing unified learning or homomorphic encryption.
- Fine-tune hyperparameters to optimize show execution and generalization.

Alarm Era and Decentralized Choice Making:

- Implement components for creating alarms when bizarre exchanges are detected.
- Design a decentralized decision-making prepare for affirming the authenticity of alarms, including agreement instruments or savvy contracts.
- Integrate an SOS include that permits clients to trigger crisis alarms in case of suspicious or false transactions.

Arrangement and Integration:

- Deploy the machine learning demonstrate and alarm framework in a decentralized environment, leveraging blockchain or disseminated record technology.
- Integrate the framework with existing exchange stages or wallets to give real-time observing and alarming functionalities.
- Ensure compatibility with distinctive blockchain conventions and interoperability with different decentralized applications (DApps).

Assessment and Execution Monitoring:

- Evaluate the execution of the framework utilizing measurements such as exactness, review, and F1-score.
- Continuously screen the system's execution and adjust the models as essential to progress precision and diminish untrue positives/negatives.
- Conduct normal security reviews to distinguish and relieve potential vulnerabilities.

A. Proposed System

We create a client centric machine learning system for the cyber security operation center in genuine venture environment. We talk about the commonplace information sources in SOC, their work stream, and how to use and handle these information sets to construct a viable machine learning framework. The paper is focused on towards two bunches of perusers. The to begin with gather is information researchers or machine learning analysts who do not have cyber security space information but need to construct machine learning frameworks for security operations center. The moment bunch of gatherings of people are those cyber security professionals who have profound information and mastery in cyber security, but do not have machine learning encounters and wish to construct one by themselves. All through the paper, we utilize the framework we built in the Symantec SOC generation environment as an illustration to illustrate the total steps from information collection, name creation, include building, machine learning calculation determination, show execution assessments, to chance score generation.

Advantages:

- Protects system against viruses, worms, spyware and other.

- Protection against data from theft.
- Protects the computer from being hacked.
- Minimizes computer freezing and crashes.
- Gives privacy to users.
- Securing the user-aware network edge.

5. System Implementation

A. Machine Learning Algorithms

In our framework, we attempted a few machine learning calculations [3]-[7], counting Multi-layer Neural Organize (MNN) with two covered up layers, Random Forest (RF) with 100 Ginisplit trees, Support Vector Machine (SVM) with outspread premise work bit and Logistic Regression (LR). In our hone, we discover that Multi-layer Neural Organize and Arbitrary Woodland work beautiful well for our issue. A few approval comes about from these models will be appeared afterward.

B. Model Performance Measures

As a common hone, modeling information ought to be arbitrarily part into preparing and testing sets and diverse models ought to be assessed on test holdout information. Other than AUC, we too characterize two measures of demonstrate goodness in Equations (1) to (2) below:

$$\text{Model Detection Rate} = \frac{\textit{Number of Risky Hosts in Certain Predictions}}{\textit{Total Number of Risky Hosts}} \times 100\% \quad (1)$$

$$\text{Model Lift} = \frac{\textit{Proportion of Risky Hosts in Certain Predictions}}{\textit{Overall Proportion of Risky Hosts}} \quad (2)$$

In differentiate to the AUC that assesses show on the entirety test information, discovery rate and lift reflect how great the demonstrate is in finding hazardous clients among distinctive parcels of forecasts. To calculate these two measurements, the comes about are to begin with sorted by the demonstrate scores (in our case, the likelihood of a client being hazardous) in plummeting arrange. Discovery rate measures the adequacy of a classification show as the proportion between the comes about gotten with and without the demonstrate. For illustration, assume there are 60 unsafe clients in test information, from beat 10% of the forecasts, the demonstrate captures 30 hazardous clients, the location rate is rise to 30/60=50%. Lift measures how numerous times it is superior to utilize a show in differentiate to not utilizing a demonstrate. Utilizing the same case over, if the test information has 5,000 clients, the lift is rise to (30/500)/(60/5000)=5. Higher lift suggests superior execution from a show on certain forecasts.

C. Model Validation Results

To approve the adequacy of our machine learning framework, we take one month of demonstrate running comes about and calculate the execution measures. We part the information arbitrarily into preparing (75% of the tests) and testing (remaining 25%) sets. The table underneath records distinctive models' AUC on test information. With normal AUC esteem over 0.80, Multi-layer Neural Arrange and Arbitrary Woodland accomplishes fulfilling precision.

Table 1
Model AUC on test data

	MNN	RF	SVM	LR
MEAN	**0.807**	**0.829**	0.775	0.754
STANDARD ERROR	0.006	0.004	0.016	0.008

Table 2
Model detection rates on top 5%~20% predictions

Top % Predictions	MNN	RF	SVM	LR
5%	31.67%	25.00%	20.00%	31.67%
10%	58.33%	43.33%	46.67%	50.00%
15%	70.00%	70.00%	70.00%	68.33%
20%	78.33%	**80.00%**	80.00%	76.67%

Table 3
Model lifts on top 5%~20% predictions

Top % of Predictions	MNN	RF	SVM	LR
5%	**6.82**	5.30	4.19	6.82
10%	6.25	4.55	4.92	5.30
15%	4.92	4.92	4.92	4.80
20%	4.09	4.19	4.19	4.00
Average	***5.52***	*4.74*	*4.56*	*5.23*

Table 3 records the location rates for distinctive models on best 5% to 20% expectations separately. It is promising that Irregular Timberland is able to identify 80% of the genuine hazardous cases with as it were 20% most noteworthy predictions.

Finally, we assess the demonstrate lift too on beat 5% to 20% forecasts as recorded in Table 3. For best 5% expectations, Multilayer Neural Organize accomplishes lift esteem of 6.82, meaning that it is nearly 7 times way better than current rule-based framework. If we see at the normal lifts on beat 5% to 20% expectations, Multilayer Neural Arrange is the most noteworthy with normal lift over 5.5 as recorded on the final.

D. Model Implementations and Active Learning

As of now the machine learning framework has been executed in a genuine venture generation. The highlights and names are being overhauled day by day from chronicled information. At that point the machine learning show is revived and conveyed to the scoring motor day by day to make beyond any doubt it captures the most recent designs from the information. After that, the chance scores are produced in genuine time when modern cautions are activated, so SOC investigators can take activity right absent for tall chance clients. At long last, SOC analysts' notes will be collected and bolstered back into verifiable information for future demonstrate refinement. The entire prepare has been streamlined consequently from information integration to score era. The framework too effectively learns modern bits of knowledge produced from analysts' examinations.

6. System Architecture

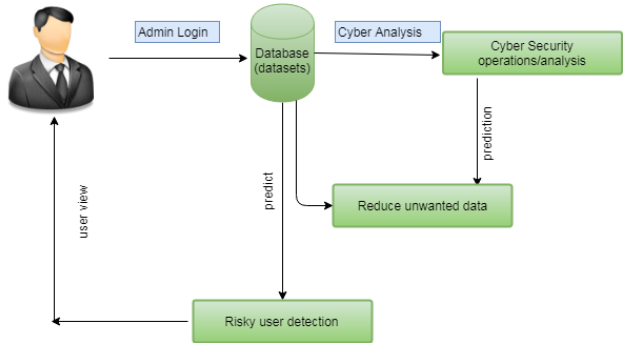

Fig. 1. System architecture

Components:

Admin Login: This grants access to the system for administrative tasks.

Database (datasets): This centralized repository stores user data and transaction information.

Cyber Analysis likely refers to the process of analyzing user activity for signs of suspicious or malicious behavior.

Cyber Security operations/analysis: This might involve investigating and responding to security threats.

Reduce unwanted data: This suggests a process to filter out irrelevant or unnecessary data from the user activity logs.

Risky user detection: This could be a system that identifies users exhibiting suspicious behavior.

User view: This indicates a section of the system where users can interact with it, possibly to view their transaction history or receive alerts.

Predict: This might refer to a machine learning model that can predict future risky user behavior based on historical data.

Prediction: The output of the machine learning model, indicating the likelihood of a user engaging in risky transactions.

Decentralized Approach:

While the exact details of the decentralized approach aren't clear from the text, it likely refers to a system where the data and processing logic are distributed across multiple nodes rather than being stored in a central server. This can offer several advantages, such as improved security, scalability, and fault tolerance.

In a decentralized system, there might be multiple copies of the user data and transaction logs stored across different nodes. This makes it more difficult for attackers to tamper with the data, as they would need to compromise multiple nodes. Additionally, if one node fails, the system can still function using the remaining nodes.

Overall Process:

1. User transactions are logged in the central database.
2. The unwanted data is filtered out from the transaction logs.
3. The filtered data is then analyzed for signs of risky user behavior. This might involve machine learning models that can identify patterns in user activity that are indicative of fraud or other malicious activity.
4. If a user is flagged as risky, an alert is generated and sent to the appropriate authorities.
5. Users might also be able to view their transaction history and receive alerts through a user interface.

Benefits of Decentralized Approach:

Security: Decentralized systems can be more secure than traditional centralized systems, as there is no single point of failure that attackers can target.

Scalability: Decentralized systems can be easily scaled to accommodate more users and transactions.

Fault tolerance: If one node in a decentralized system fails, the system can still function using the remaining nodes.

III. Result and Discussion

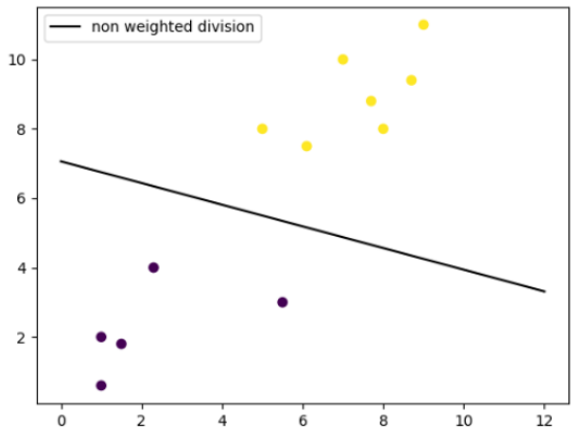

Fig. 2. Linear SVM

The decision boundary doesn't have to be a line. It's also referred to as a hyperplane because you can find the decision boundary with any number of features, not just two.

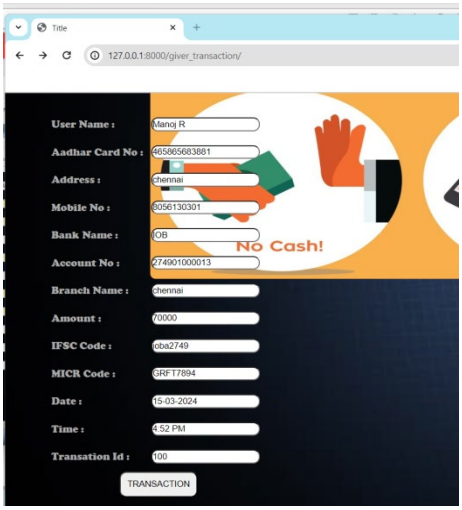

Fig. 3. Transaction of user and admin

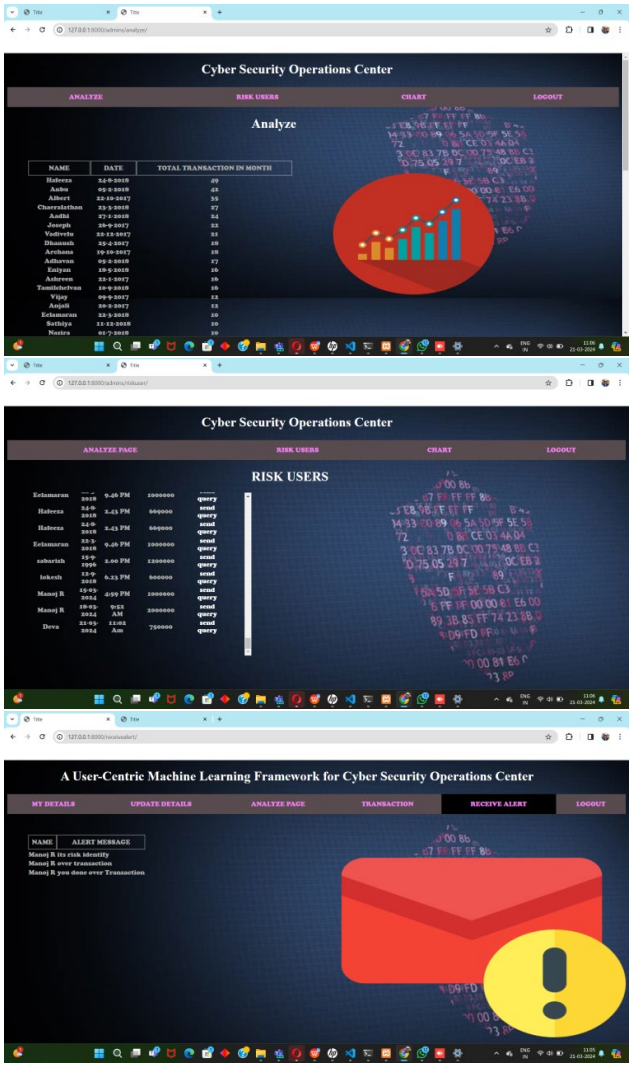

Fig. 4.

7. Conclusion

In this paper, we show a user-centric machine learning framework which leverages enormous information of different security logs, alarm data, and examiner bits of knowledge to the recognizable proof of unsafe client. This framework gives a total system and arrangement to hazardous client discovery for undertaking security operation center. We depict briefly how to produce names from SOC examination notes, to connect IP, have, and clients to produce user-centric highlights, to select machine learning calculations and assess exhibitions, as well as how to such a machine learning framework in SOC generation environment. We too illustrate that the learning framework is able to learn more bits of knowledge from the information with exceedingly unequal and constrained names, indeed with basic machine learning calculations. The normal lift on best 20% expectations for multi neural arrange show is over 5 times superior than current rule-based framework. The entire machine learning framework is actualized in generation environment and completely computerized from information securing, day by day demonstrate reviving, to genuine time scoring, which enormously move forward SOC analyst's proficiency and upgrade undertaking chance discovery and administration.

References

[1] SANS Technology Institute. The 6 Categories of Critical Log Information´ 2013.

[2] X. Li and B. Liu. "Learning to classify text using positive and unlabeled data", Proceedings of the 18th international joint conference on Artificial intelligence, 2003.

[3] A. L. Buczak and E. Guven. ³A survey of data mining and machine learning methods for cyber security intrusion detection´, IEEE Communications Surveys & Tutorials 18(2):1153-1176, 2015.

[4] S. Choudhury and A. Bhowal. ³Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection´, Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015.

[5] N. Chand et al. ³A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection´, Advances in Computing, Communication, & Automation (ICACCA), 2016.

[6] K. Goeschel. ³Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and naive Bayes for off-line analysis´, Southeast Con, 2016.

[7] M. J. Kang and J. W. Kang. ³A novel intrusion detection method using deep neural network for in-vehicle network security´, Vehicular Technology Conference, 2016.