ARCHITECTURAL DESIGN STRATEGY ON BIG DATA CYBER SECURITY

A.Rajalakshmi[1],S.Ganesh Kumar[2],B.Amutha[3], Princely Mervin J [4]

1.Assistant Professor,Department of Computer application,SRMIST,Katankulathoor,603203rajalaka@srmist.edu.in,
,2.Associate Professor,school of omputing,SRMIST,Katankulathoor,SRMIST,603203,ganeshk1@srmist.edu.in,
3.Professor/HOD,School of Computing.SRMIST.Katankulathoor,603203,bamutha62@gmail.com
4.Department of CSE -SRMIST SRMIST.Katankulathoor,60320

Abstract- Organizations passes their data through the internet and various networks for the business transactions, and cyber security gives procedures to protect the information and so the systems process it safely and store it for their usage. Since the ratio of cyber threads has been grown very vastly, organizations must be very careful in protecting their official data. National intelligence warned about the cyberattacks and digital spying are the major reason for the nation security, which is also the root cause for many terrorists.Nation security warns that many private and public sector's digital systems are liable to cyberattacks. So, when the world works in digital systems, we need a proper security system to safeguard our data. Nowadays attackers are more in numbers and has the ability to crack the complex systems.This project presents a hyperheuristic framework for SVM configuration optimization. Hyper heuristics is more efficientwhen compared to other methods.

I. INTRODUCTION

Advanced growth in technologies like mobile, social mediaand IOT give a huge digital information. The big data refers the huge digital information today. It consists of both structure and un structed data which is used for quick references. The big data varies from Database as it consists of all kinds of data and many tools available for analyzing the data.

Big data is classified by three characteristics:

1. Volume
2. Varity
3. Velocity

Volume, refers the amount and size of the data stored in day to day basis. The storage plays a vital role for the data storage. **Variety,** refers the different types of data stored in the warehouse such as text, images, audio video etc. **Velocity** defines the rate of data entering into the data store per second.

All the 3V's plays a vital role in the big data environment.Reacher's collect the information from the big data and analyses and work for their various need and it helps to make wise decision using the previous data. Big data has a security threat and the data reliability.

Many attackers intrude the network to get the information for the big data. When the data grows the thread for the data also increases and it need to be addressed. Many companies pay huge sum of amount for their data security and the malware detection is the major concern which need to be noted. The smart contract is stored in the nodes of public Ethereum network which is available in multiple locations and these smart contracts can run in a decentralized fashion and can be stored in a decentralized fashion.

Big Data the worddescribe huge amounts of data that is structured and unstructured.The data arehuge so that it's very complex to process this data using the old database software tools. The latest technologies, devices, and social networking sites, the rapid speed of datasaved by the people is increasing day by day each year. Big data is collection of vast datasets that is difficult and take time to be processed using old computing tools. Big data isn't just a data, instead it's became a whole course, which involves various tools and techniques

II. RELATED WORKS

A. Feature Extraction, Selection and Combination for Malware Family Categorization

- M. Ahmadi et.al (2016) presents that latest malware is meant to be with many characteristics. It causes an infinite rate within the existing number of malwaresidentified. Classification of malware in the live environment with their characteristics is crucial for the many security teams,because day to day they receive enormous malware attacks.It is classified using the signature of the malware family.

Microsoft conducted a malware classification challenge in 2015 which an infinite dataset of near 0.5 terabytes of huge data, and it consists of twenty thousand malware samples. The data set analyzing is inspired by the event set an example of an effective in organizing malware in their family groups. This concept is projected and discussed within this work, where importance has been given to the groups related to the extraction, and selection of a huge set features for the effective representation of malware samples. Features are groups to be with multiple characteristics of malware functionalities. The proposed method got an extremely high accuracy for the Microsoft Malware Challenge dataset.

B. *A Survey on Algorithms for Decision Tree Induction*

- M. P. Basgalupp et.al (2015) states that almost all of the conceptsevolve decision trees as an alternate heuristic to the standard top-down, divide and conquer algorithm. He identified techniques that make use of old algorithms to enhance parts of decision tree classifiers. It provides an overview that's fully focus on traditional algorithms and decision trees and doesn't think about any specific approach. Second, it gives us a classification which indicates the works around decision trees. Finally, it describes applications of traditional algorithms for decision tree in several fields. The work ends by addressing some critical issues and questions that leads to the topic of future research.

C. *Study of the Combination of Meta-Learning with Particle Swarm Algorithms for SVM Parameter Selection*

- R. B. C. Prudencio et.al (2012) presents that Support Vector Machines (SVMs) became the successful algorithm due to its great performance it achieves invarious learning issues. To perform the SVMtechnique it requires some modulations on its model to avoid the trial and error method and this automated parameter selection could be a thanks to cope with this method. The automatedmethod isconsidered an optimizedissue whose goal is makecorrect configuration of parameters which address some learning issues. A study of the mix of Metal earning with Particle Swarm Optimization algorithms to enhance the SVM model, finding for combinations of parameters which increases the success rate of SVM method. Hence, deployeda smaller number of user search points, within the search process, to a suitable solution, would be more cost-effective. The experimentscompare the working of the search algorithms deploying a traditional random initialization and using ML suggestions as an initial population. This method analyzed the learning on convergence of the optimization algorithms, verifying that the mix of PSO techniques with ML derived solutions with higher quality on a group of different problems and different applications.

III. PROPOSED SYSTEM

This project represents a hyper-heuristic framework for SVM configuration. Hyper heuristics is independent and it givesefficient results. The hyper-heuristic framework integrates several components that makes it more efficient than the componentsfor an efficient SVM configuration in big data security. The hyper heuristic framework consists of a dual objective formula of the SVM configuration, in which the accuracy is high and complexity are marked as separate objectives. Theheuristic framework controls both the kernel type and parameters because it has a soft margin parameter. It has the power of decompositionand Paretobased approaches in an adaptive manner to search out an nearest Pareto set of SVM. It has high accuracy when compared to any other mechanism in detecting the malware which keep the big data away from the security thread.

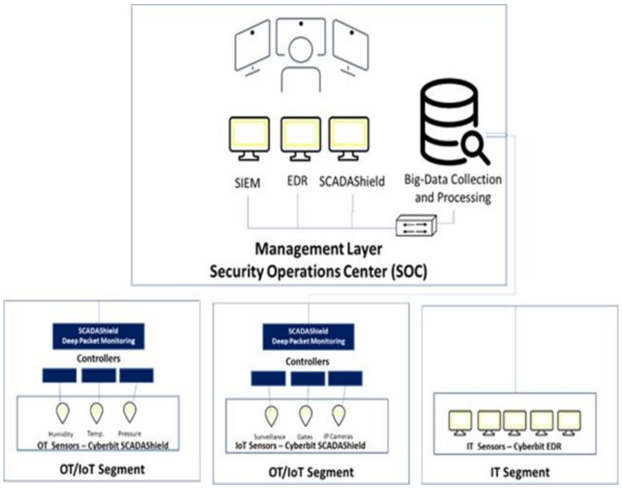
SIEM
EDR
SCADAShield
Big-Data Collection and Processing
Management Layer
Security Operations Center (SOC)
SCADAShield Deep Packet Monitoring
Controllers
Humidity
Temp.
Pressure
OT Sensors – Cyberbit SCADAShield
OT/IoT Segment
SCADAShield Deep Packet Monitoring
Controllers
Surveillance
Gates
IP Cameras
IoT Sensors – Cyberbit SCADAShield
OT/IoT Segment
IT Sensors – Cyberbit EDR
IT Segment

Fig. 1. Architecture

MALWARE AND MALICIOUS BEHAVIOR DETECTION

The method classifies malicious websites by an adaptive support vector machine. To classify malicious websites, they defined the features to represent the essential characteristics of an online page and selected vector machine for learning training data. But this research was focused on only malicious websites. The approach is analogous to our approach using behavior-based features, but our approach can make a behavior model supported malicious and benign behavior by training an oversized amount of information on Big Data platform digital money. Ethereum is employed as an infrastructure to run many decentralized applications thanks to its programmable nature. Ethereum isn't controlled by anyone which implies there's no centralized authority or an organization to regulate Ethereum

HYPER-HEURISTICS

Hyper-heuristic is latest method use to resolve multiple problems. A hyper-heuristic framework gets all the possible inputs and use the best one. The technical people used Hyper Heuristics framework to get which heuristic will be suitable to resolve the problem.

K-MEANS CRYPTOGRAPHIC PUZZLE

A sender contains a packet for transmission. The sender selects a private key in specified length. Sendercreates a puzzle with key and time, where puzzle () represents the generator function, and timeindicates the duration required to solve the puzzle.

The Parameters is measured with time, and it's directly enthusiastic about the measuredcalculation capability of the system, denoted by N and measured in computer operations per second. After generating the puzzle, the user broadcasts and the receiver end, it solves the received puzzle to recover key then computes data that is transferred by the sender.

MD5 SECURITY

The passwordencrypted by the hash algorithm is termed as hashed password. This kind of transmission is usually a subject matter of intrusionto the hackers. The hashed passwords are felt the net as an IP packet. TCP header is the most typical a part of the IP packet. in an TCP header. It has six reserved bits which unused. During this project we proposed a newway to strengthen the hashed passwords using the six reserved bits of a TCP header. In this method we encrypt the hashed password using a private key employing a mathematical relation. The decryption data is carried out in a six-bit TCP header.

IV. **IMPLEMENTATION**

In a system input is raw data that is processed to produce output. The clear and correct output is the major task of any system. While designing the output,developers work to produce the output needed and then consider the needed output designs and report layouts.

Input design- The input design is the bridge between the system andthe user. It links the developing specification and procedures for data preparation and othersteps that is need to put the transaction data for processing. The is data fed to the system through the documents or the printed data. The input data can be directly keyed to the system.The input system directly focuses on the quality and the quantity of the data given. The input data must be free from error and properly validated. The input data is meant to have a security and a proper design.

Input Design must consider the following things:

What data should be given as input?

How it should be arranged or designed?

Input Design isthe process of converting the user data to the system format. It is very important to avoid the errors during the input process and give the proper direction to the management to get correct data from the computer system.

It is achieved by having a user-friendly environment for the info entry to manage huge amount of data. The goal of setting the input data is to makeinput data simple and to make it free from errors. The info entry screen is intended in such how that everyone’s info manipulation performed. It also allows to view the data records. When the input data is entered,the system checks for its validity. Data enters the system through the input interface. Proper data are provided during the initial stages to avoid user’s intervention in the middle. The target of input structure is to make adata layout that's easy to follow.

Output design- Asuper good quality output which meets the needs of the endcustomer and gives the data properly. In output structure it isshown how the information need to be viewed for urgent needs and the textual matter result. It is the foremost critical data to the user. Efficient and good outputresultsdevelopthe relationship to assist user to enhance his decision-making.In modellinga system output, the computer shouldbe processed in a well-arranged manner. Correct result must be created while ensuring that allobjects is meant so user will find the system can find it easily and effectively. When analyzing output data, they must Identify the particular data that's needed to fulfill the necessities.

Select methods for presenting information.

Create the documents, reports which containsdatamanipulated by the computer.

The output sort of a system should accomplish one or more of the subsequent objectives.

Inform data about past status, current information or the scope of work in the future.

Notify the events logs, usages, issues, or warnings in the information

Action need to be triggered

Approve and confirm the action

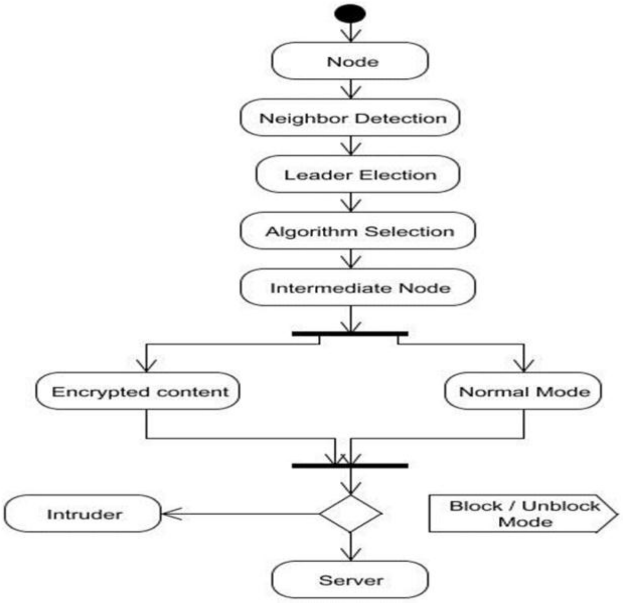

Fig. 2. Activity Diagram

IV. CONCLUSION

In this project, to resolve the matter of reducing the size of training set with the tactic of clustering. The K-means clustering approach is employed to pick out the few most informational samples, which are accustomed build the training data set. Experimental results show that the algorithm reached the aim of reducing the size of the training set, and likely reduced the training and prediction time and also assures the generalization ability of K-SVM algorithm.This can be a severe shortcoming, and it’s going to restrict users from accessing SVMs in real timewhich required processing a huge dataset. This might be useful for the real time system, like security market surveillance and network intrusion detection.

VI. FUTURE WORK

In the future, they are planning to use the work on the huge data set, and analyze uniquely. SVM has been used to reveal some BI problems, but latent period of SVM must improve when applied to the real-time system. In future, KMSVM algorithm should be verified on more real-time BI databases. However, after decrypting error rate rapidly grows increasing the channel error rate. So,to get high system reliability, a vast spread of error patterns must be corrected. A strong efficient code is must which makes the coder and decoder to work easily and alsopossess a high transmission overhead

VII **REFERENCES**

[1] M. Ahmadi, D. Ulyanov, S. Semenov, M. Trofimov, and G. Giacinto, “Feature extraction and integration of malware family classification," in Proc. 6th ACM Conf. Data Appl. Secur. Privacy, 2016, pp. 183_194.

[2] M. P. Basgalupp, R. C. Barros, and V. Podgorica, “Decision-treealgorithms with a multiple objective hyper-heuristic," in Proc. 30thAnnu. ACM Symp. Appl. Comput., 2015, pp. 110_117.

[3] N.-E. Ayat, M. Cheriet, and C. Y. Suen, “Automatic model selection for the optimization of SVM kernels," Pattern Recognit., vol. 38, no. 10, pp. 1733_1745, 2005.

[4] Ben-Hur, “A user's guide to support vector machines," in Data Mining Techniques for the Life Sciences. Methods in Molecular Biology (Methods and Protocols), O. Carugo and F. Eisenhaber, Eds. Vol 609. New York, NY, USA: Humana Press, 2010, pp. 223_239.

[5] C.-C. Chang, “LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 27:1_27:27, 2011.