IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A MACHINE LEARNING APPROACH FOR DIGITAL INFORMATION PRESERVING

G.Pavan kumar Reddy
Computer science and Engineering
Madanapalle Institute of Technology and Science, Madanapalle, India

G.Yaswanth kuamr
Computer science and Engineering Madanapalle Institute of Technology and Science, Madanapalle, India

Y.Hari prakash reddy
Computer Science and Engineering, Madanapalle Institute of Technology & Science, Madanapalle, India

G. Jayanth
Computer Science and Engineering Madanapalle Institute of Technology & Science, Madanapalle, India

P Mallikarjuna
Computer Science and Engineering, Madanapalle Institute of Technology & Science, Madanapalle. India

ABSTRACT

In order to ensure a company's Internet security, SIEM (Security Information and Event Management) system is in place to simplify the various preventive technologies and flag alerts for security events. Inspectors (SOC) investigate warnings to determine if this is true or not. However, the number of warnings in general is wrong with the majority and is more than the ability of SCO to handle all awareness. Because of this, malicious possibility. Attacks and compromised hosts may be wrong. Machine learning is a possible approach to improving the wrong positive rate and improving the productivity of SOC analysts. In this article, we create a user-centric engineer learning framework for the Internet Safety Functional Center in the real organizational context. We discuss regular data sources in SOC, their work flow, and how to process this data and create an effective machine learning system. This article is aimed at two groups of readers. The first group is intelligent researchers who have no knowledge of data scientists or computer safety fields but who engineer should develop machine learning systems for machine safety. The second groups of visitors are Internet security practitioners that have deep knowledge and expertise in Cyber Security, but do Machine learning experiences do not exist and I'd like to create one by themselves. At the end of the paper, we use the account as an example to demonstrate full steps from data collection, label creation, feature engineering, machine learning algorithm sample performance evaluations using the computer built in the SOC production of Seyondike.

I. INTRODUCTION

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information. Currently, it is being used for various tasks such as image recognition, speech recognition, email filtering, Facebook auto-tagging, recommender system, and many more.

This machine learning tutorial gives you an introduction to machine learning along with the wide range of machine learning techniques such as Supervised, Unsupervised, and Reinforcement learning. You will learn about regression and classification models, clustering methods, hidden Markov models, and various sequential models .In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability, and we have computers or machines which work on our instructions. But can a machine also learn from experiences or past data like a human does? So here comes the role of Machine Learning. A

Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately. Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predicts the output. Machine learning has changed our way of thinking about the problem. The below fig 1.1 explains the working of machine learning algorithm.

• Machine learning uses data to detect various patterns in a given dataset. • It can learn from past data and improve automatically. • It is a data-driven technology. • Machine learning is much similar to data mining as it also deals with the huge amount of the data.

Feasibility Study Preliminary investigation examines project feasibility; the likelihood the system will be useful to the organization. All systems are feasible if they are given unlimited resources and infinite time. There are aspects in the feasibility study portion of the preliminary investigation: • Technical Feasibility • Operation Feasibility • Economic Feasibility.

Technical Feasibility The technical issue usually raised during the feasibility stage of the investigation includes the following: • Does the necessary technology exist to do what is suggested? • Do the proposed equipment's have the technical capacity to hold the data required to use the new system?

Operation Feasibility: The operational feasibility includes User friendly, reliability, security, portability, availability and maintainability of the software used in the project. Economic Feasibility: Analysis of a project costs and revenue in an effort to determine whether or not it is logical and possible to complete.

At a broad level, machine learning can be classified into three types: 1. Supervised learning 2. Unsupervised learning 1) Supervised Learning Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.

The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not. The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is spam filtering. Supervised learning can be grouped further in two categories of algorithms: ● Classification ● Regression

2) Unsupervised Learning Unsupervised learning is a learning method in which a machine learns without any supervision. The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns. In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data

II. LITERATURE SURVEY

1 INTRODUCTION: Literature survey is something when you look at a literature (publications) on a surface level or an Aerial view. It includes the survey of place people and a publication is in the context of Research. It is a phase where the researcher tries to know what are all the literature related to one area of interest. A literature survey represents a study of previously existing material on the topic of the report. This includes

1. Existing theories about the topic which are accepted universally. 2. Books written on the topic, both generic and specific. 3. Research done in the field usually in the order of oldest to latest. 4. Challenges being faced and ongoing work, if available.

2.EXISTING SYSTEM Most methods to security in the agency have centered on defensive the network infrastructure and not using a or little interest to cease users. As a result, conventional protection capabilities and associated devices, which includes firewalls and intrusion detection and prevention gadgets, deal especially with network stage safety. Although still part of the general security story, such an technique has boundaries in mild of the brand new safety challenges defined in the previous phase. Data Analysis for Network Cyber-Security focuses on tracking and analyzing community traffic statistics, with the purpose of stopping, or quickly identifying, malicious hobby. Risk values had been brought in an information security management system (ISMS) and quantitative assessment changed into performed for particular risk evaluation. The quantitative evaluation confirmed that the proposed countermeasures should lessen risk to a point. Investigation into the value- effectiveness of the proposed countermeasures is an essential future work. It provides users with assault information consisting of the sort of assault, frequency, and goal host t ID and supply host ID.

DISADVANTAGES 1. Firewalls can be difficult to configure correctly. Incorrectly configured firewalls may block users from performing actions on the internet, until the firewall configured correctly. 2. Makes the system slower than before. 3. Firewalls can be difficult to configure correctly. Incorrectly configured firewalls may block users from

performing actions on the Internet, until the firewall configured correctly.

PROPOSED SYSTEM User-centric cyber security allows organizations lessen the danger associated with fast-evolving stopperson realities by way of reinforcing protection toward give up users. User- centric cyber safety isn't always the same as consumer protection. User-centric cyber protection is about answering peoples' wishes in approaches that maintain the integrity of the organization network and its assets. User protection can nearly appear like a matter of protective the community from the person — securing it against vulnerabilities that the person needs to introduce. User-centric safety has the greater value for corporations. Cyber-security systems are real-time and sturdy unbiased structures with high performance necessities. They are used in many software domains, which includes vital infrastructures, which include the country wide power grid, transportation, medical, and defense. These applications require the attainment of stability, overall performance, reliability, performance, and robustness, which require tight integration of computing, communique, and manipulate technological systems

ADVANTAGES 2.3.1 Protects the system against viruses, worms, spyware and other . 2.3.2 Protection against data from theft 2.3.3 Protects the computer from being hacked. 2.3.4 Minimizes computer freezing and crashes. 2.3.5 Gives privacy to use

CYBER ANALYSIS Cyber threat analysis is a process in which the knowledge of internal and external information vulnerabilities pertinent to a particular organization is matched against real -world cyberattacks. With respect to cyber security, this threat-oriented approach to combating cyber-attacks represents a smooth transition from a state of reactive security to a state of proactive one. Moreover, the desired result of a threat assessment is to give best practices on how to maximize the protective instruments with respect to availability, confidentiality and integrity, without turning back to usability and functionality conditions.

If a dataset in your dashboard contains many dataset objects, you can hide specific dataset objects from display in the Datasets panel. For example, if you decide to import a large amount of data from a file, but do not remove every unwanted data column before importing the data into Web, you can hide the unwanted attributes and metrics.

Improve storage efficiency through data reduction techniques and capacity optimization using data reduplication, compression, snapshots and thin provisioning. Data reduction via simply deleting unwanted or unneeded data is the most effective way to reduce a storing's data.

False alarm immunity to prevent customer embarrassment, High detection rate to protect all kinds of goods from theft, Wide-exit coverage offers greater flexibility for entrance/exit layouts, Wide range of attractive designs complement any store décor, Sophisticated digital controller technology for optimum system perform.

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. So you're working on a text classification problem. You're refining your training data, and maybe you've even tried stuff out using Naive Bayes. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper- plane 9 that differentiate the two classes very well (look at the below snapshot).

● SVMs are helpful in text and hypertext categorization, as their application can significantly reduce the need for labeled training instances in both the standard inductive and transudative settings. Some methods for shallow semantic parsing are based on support vector machines. ● Classification of images can also be performed using SVMs. Experimental results show that SVMs achieve significantly higher search accuracy than traditional query refinement schemes after just three to four rounds of relevance feedback. This is also true for image segmentation systems, including those using a modified version SVM that uses the privileged approach as suggested by Vapnik. ● Classification of satellite data like SAR data using supervised SVM. ● Hand-written characters can be recognized using SVM. ● The SVM algorithm has been widely applied in the biological and other services..

They have been used to classify proteins with up to 90% of the compounds classified correctly. Permutation tests based on SVM weights have been suggested as a mechanism for interpretation of SVM models. Support-vector machine weights have also been used to interpret SVM models in the past.

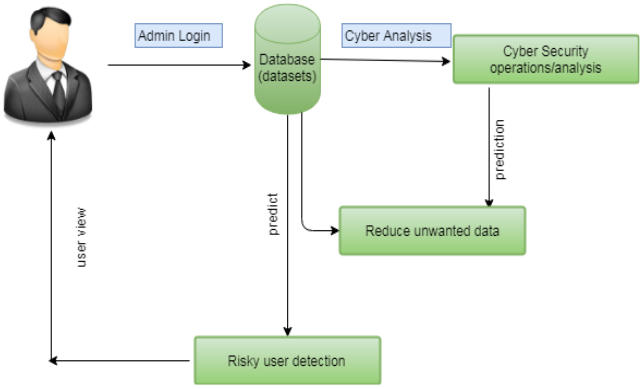
3.ANALYSIS

SOFTWARE REQUIREMENTS: For developing the application the following are the Software Requirements: ● Operating system : Windows 7 & above ● Coding Language : Python. 3.6.2 ● Front-End : Python. 3.6.2 ● Designing : HTML, CSS and JavaScript. ● Data Base : MySQL client 1.3.12 ● Server : Wamp server 2.4

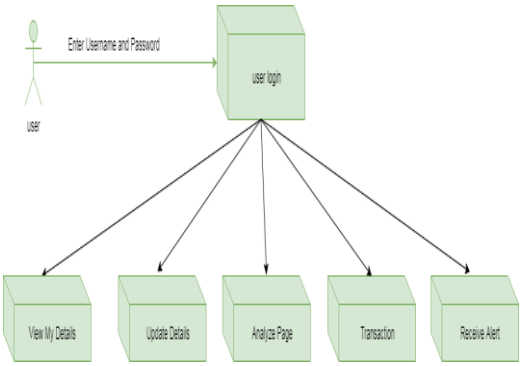
HARDWARE REQUIREMENTS: For developing the application the following are the Hardware Requirements: ● Processor : i3 & above ● RAM : 2GB & above ● Space on Hard Disk : 256GB & above ● Processor : i3 & above ● RAM : 2GB & above ● Space on Hard Disk : 256GB & above.

4.DESIGN

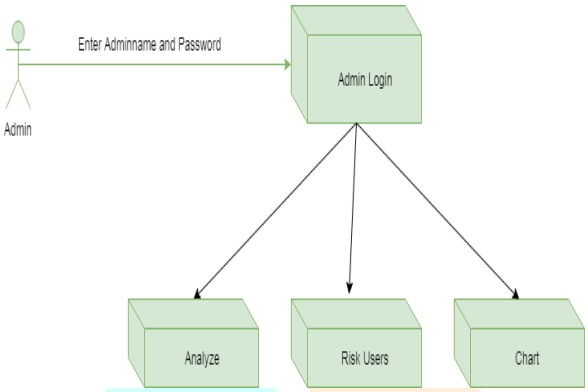
System's design is the process of defining elements of a system like modules, architecture, components and their interfaces and data for a system based on the specified requirements. It is the process of defining, developing and designing systems which satisfies the specific needs and requirements of a business or organization.



USER LOGIN

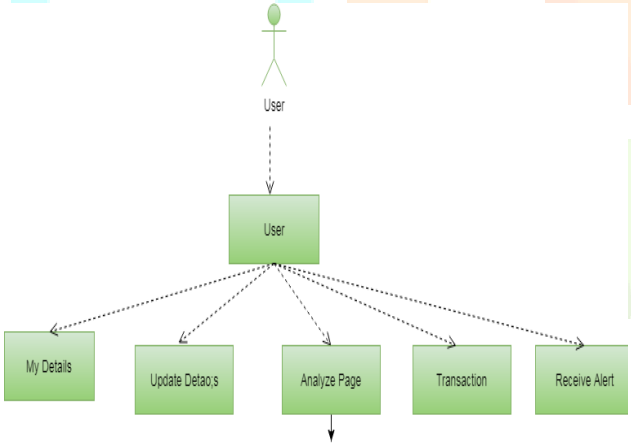


ADMIN LOGIN

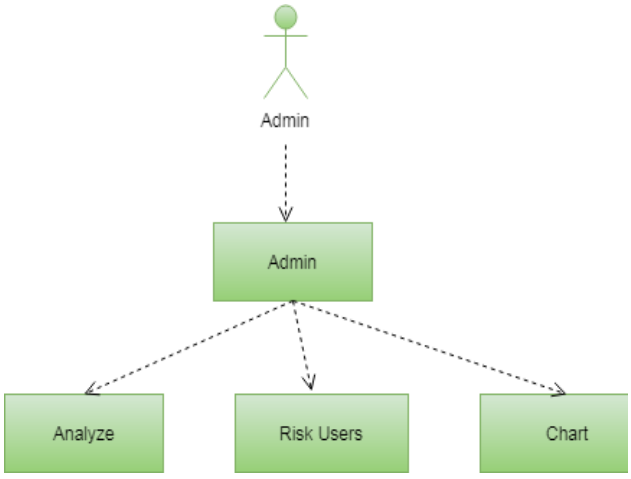


USE CASE DIAGRAMS

a. USER

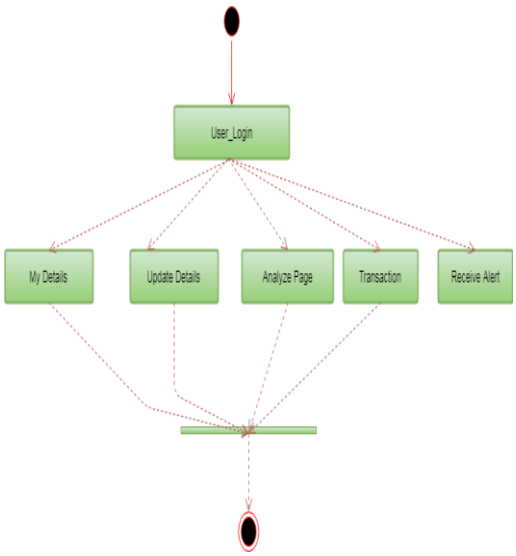


b. ADMIN

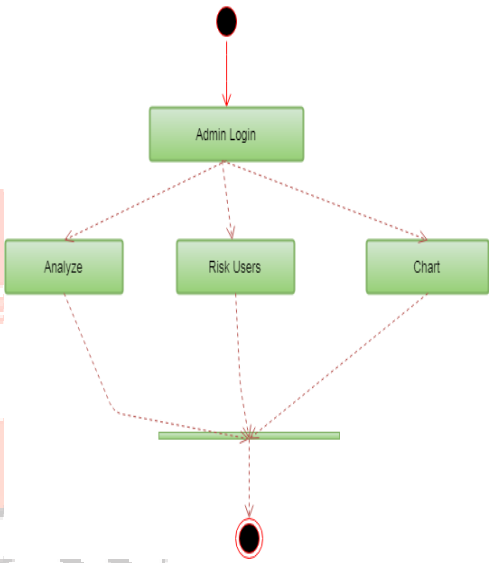


ACTIVITY DIAGRAMS

a. USER



b. ADMIN



5. IMPLEMENTATION AND RESULT

Implementation consist of modules development and their integration as per architecture of our proposed system 5.1 Technologies Used The following technologies are used in our project. 5.1.1 Python Python is a simple, general purpose, high level, and object-oriented programming language. Python is an interpreted scripting language also. Guido Van Rossum is known as the founder of Python programming Python supports multiple programming patterns, including object-oriented, imperative, and functional or procedural programming styles.

Python Features: Python provides lots of features that are listed below. Python is easy to learn and use. It is a developer-friendly and high level programming language. Python language is more expressive and means that it is more understandable and edible. Python is an interpreted language i.e. interpreter executes the code line by line at a time. This makes debugging easy and thus suitable for beginners. Python can run equally on different platforms such as Windows, Linux, Unix and Macintosh etc. So, we can say that Python is a portable language. Python language is freely available at official web address. The source-code is also available. Therefore it is open source. Python supports object oriented language and concepts of classes and objects come into existence. It implies that other languages such as C/C++ can be used to compile the code and thus it can be used further in our python code. Python has a large and broad library and

provides a rich set of modules and functions for rapid application development. Graphical user interfaces can be developed using Python. 10) Integrated It can be easily integrated with languages like C, C++, JAVA etc.

Python Applications: Python is known for its general purpose nature that makes it applicable in almost each domain of software development.

We can use Python to develop web applications. It provides libraries to handle internet protocols such as HTML and XML, JSON, Email processing, request, beautiful Soup, Feed parser etc. It also provides Frameworks such as Django, Pyramid, Flask etc to design and develop web based applications. 2) Python provides a Tk GUI library to develop user interfaces in python based applications. 29some other useful tool kits wx Widgets, Kivy, pyqt that are usable on several platforms. The Kivyis popular for writing multi touch applications. 3) Python is helpful for the software development process. It works as a support language and can be used for build control and management, testing etc. 4) We can use Python to develop console based applications. For example: IPython. 5) Python is awesome to perform multiple tasks and can be used to develop multimedia applications. Some of real applications are: Tim Player etc. 6) To create CAD applications Fandango is a real application which provides full features of CAD. 7) Python can be used to create applications which can be used within an Enterprise or an Organization. Some real time applications are: OpenErp, Tryton, Piccolo etc.

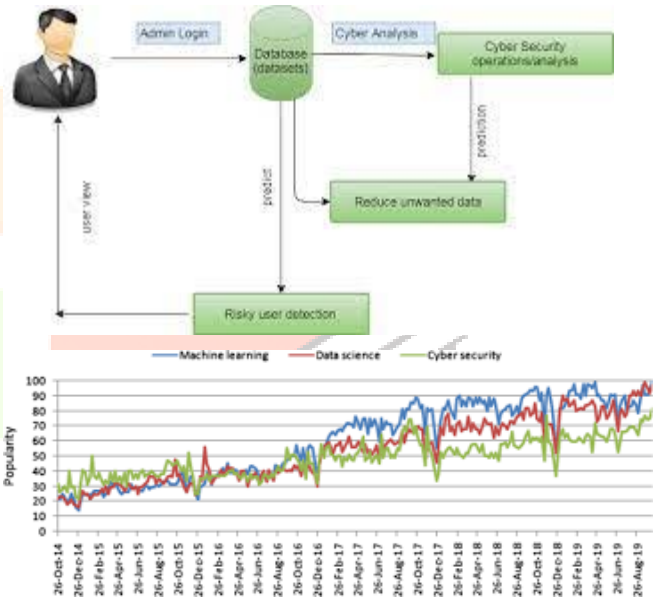
1 Django Django is a web application framework written in Python programming language. It is based on MVT (Model View Template) design pattern. The Django is very demanding due to its rapid development feature. It takes less time to build an application after collecting client requirements. This framework uses a famous tagline: The web framework for perfectionists with deadlines.

Popularity: Django is widely accepted and used by various well-known sites such as: ● Instagram ● Mozilla Features of Django : ● Rapid Development ● Secure ● Scalable ● Fully loaded ● Versatile ● Open Source ● Vast and Supported Community. Django was designed with the intention to make a framework which takes less time to build web applications. The project implementation phase is very time taken but Django creates it rapidly. e Django takes security seriously and helps developers to avoid many common security mistakes, such as SQL injection, cross-site scripting, cross-site request forgery etc. Its user authentication system provides a secure way to manage user accounts and passwords. Django is scalable in nature and has ability to quickly and flexibly switch from small to large scale application project. Django includes various helping task modules and libraries which can be used to handle common Web development tasks. Django takes care of user authentication, content administration, site maps, RSS feeds etc. Django is versatile in nature which allows it to build applications for different different domains. Nowadays, Companies are using Django to build various types of applications like: content management systems, social networks sites or scientific computing platforms etc. Django is an open source web application framework. It is publicly available without cost. It can be downloaded with source code from the public repository. Open source reduces the total cost of the application development.

OUTPUT SCREENS & RESULT ANANLYSIS

TABLE IV. 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*



6. TETING AND VALIDATION

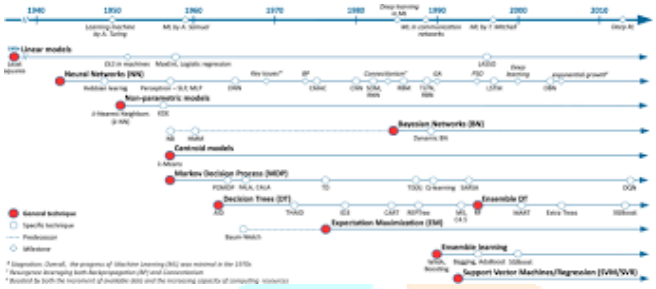
The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product.

unit trying out entails the layout of take a look at cases that validate that the internal application good judgment is functioning nicely, and that application inputs produce legitimate outputs. All selection branches and internal code flow need to be verified. It is the testing of man or woman software program gadgets of the utility. It's miles completed after the finishing touch of an man or woman unit earlier than integration. Functional tests offer systematic demonstrations that features examined are to be had as unique by way of the enterprise and technical necessities, machine documentation, and consumer manuals.

Valid Input : identified classes of valid input must be accepted. Invalid Input : identified classes of invalid input must be rejected. Functions : identified functions must be

exercised. Output : identified classes of application outputs must be exercised.

Interfacing systems or procedures must be invoked. Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. White Box Testing is a checking out wherein the software program tester has understanding of the inner workings, shape and language of the software, or at the least its reason. It is purpose. It is used to test areas that can't be reached from a black container level. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated as a black.



	Hardware	Software	Network
Common attacks	• Hardware Trojan • Illegal clones • Side channel attacks (i.e. snooping hardware signals)	• Software programming bugs (e.g. memory management, user input validation, race conditions, user access privileges, etc.) • Software design bugs • Deployment errors	• Networking protocol attacks • Network monitoring and sniffing
Examples of countermeasures	• Tamper-Resistant Hardware (e.g. TPM) • Trusted Computing Base (TCB) • Hardware watermarking • Hardware obfuscation	• Secure coding practice (e.g. type checking, runtime error, program transformation, etc.) • Code obfuscation • Secure design and development • Formal methods	• Firewall • Intrusion prevention and detection • Virtual Private Network (VPN) • Encryption

7.CONCLUSION

We present a person-centric device mastering device which leverages large information of various security logs, alert records, and analyst insights to the identification of risky consumer. This gadget presents a whole framework and solution to risky user detection for organization security operation middle. We describe briefly how to generate labels from SOC investigation notes, to correlate IP, host, and users to generate consumer- centric features, to choose device mastering algorithms and examine performances, as well as a way to this type of machine gaining knowledge of gadget in an SOC manufacturing surroundings. We also exhibit that the gaining knowledge of gadget is capable of examine extra insights from the records with tremendously unbalanced and constrained labels, in spite of simple machine mastering algorithms. The common carry on pinnacle 20% predictions for multi neural network models is over 5 times higher than current rule-primarily based systems. The complete device learning system is applied in a production environment and completely automatic from information acquisition, daily version refreshing, to actual time scoring, which significantly improve and enhance enterprise risk detection and control. As to the future work, we are able to studies different getting to know algorithms to in addition improve the detection accuracy.

8. REFERENCES

1.The 6 Categories of Critical Log Information", *SANS Technology Institute*, 2020.

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*, 2021.

3. A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection", *IEEE Communications Surveys & Tutorials*, vol. 18.2, pp. 1153-1176, 2019.

4. S. Choudhury and A. Bhowal, "Comparative analysis of machine learning algorithms along with classifiers for network intrusion detection", *Smart Technologies and Management for Computing Communication Controls Energy and Materials (ICSTM)*, 2020.

5. N. Chand et al., "A comparative analysis of SV Mand its stacking with other classification algorithm for intrusion detection", *Advances in Computing Communication & Automation (ICACCA)*, 2019.

6. Alphonse Inbaraj, X., Rao, A.S.: Hybrid agglomeration algorithms for crime pattern analysis. In: 2018 IEEE International Conference on Current Trends toward affiliation Technologies, Coimbatore, India.

7. Dutta, S., Gupta, A.K., Narayan, N.: Identity crime detection exploitation technique. In: 2017 International Conference on methodology Intelligence and Networks. IEEE (2017).

8. Manish Gupta, B. Chandra M.P. Gupta, "Crime technique for Indian police system", Journal of Crime, Vol.2,No.6, 2006.

9. David J. Hand, Heikki Mannila and Padhraic Smyth, "Principles of knowledge mining", MIT Press, 2001.

10. Hsinchun genus, Wingyan Chung, Yi Qin, archangel Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, Homa Atabakhsh, "Crime information Mining: A General Framework and sort of Examples", IEEE notebook computer Society Apr 2004. Cattleman Barnadas, M. (2016). Machine learning applied to crime prediction (Bachelor's thesis, Universitat Politècnica Delaware Catalunya).

11. K. Goeschel, "Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines decision trees and naive Bayes for off-line analysis", *SoutheastCon*, 2016.

12. M. J. Kang and J. W. Kang, "A novel intrusion detection method using deep neural network for in-vehicle network security", *Vehicular Technology Conference*, 2016.