iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** II **Month of publication:** February 2024

DOI: https://doi.org/10.22214/ijraset.2024.58325

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue II Feb 2024- Available at www.ijraset.com

A Study of Role and Applicability of Data Mining Techniques in Intrusion Detection Systems

Arun Pandey[1], Dr. Shishir Sharma[2]
Research Scholar, Assistant Prof. Dept. of IT and CS, Dr. C. V. Raman University, Kota, Bilaspur (C.G)

Abstract: ***The swift development of computers technologies changed the way data and information were kept. The risk of this data being exposed to uninvited and unauthorized users arises with this new paradigm of data access. Numerous systems have been created that examine data to look for deviations from a user's or system's typical behavior or look for a known signature in the data. Intrusion Detection Systems (IDS) is the name given to these systems. These systems use a variety of approaches, including machine learning algorithms and statistical methodologies. With the massive rise in the use of network-based services and information sharing on networks, network security has emerged as the fundamental component. The integrity, confidentiality, and availability of computer and network resources are all seriously compromised by intrusion, which also poses a severe risk to network security. Network audit data classification by humans is a costly, time-consuming, and laborious task. An intrusion detection system (IDS) is one tool used to find anomalies and attacks on a network. The network intrusion detection system has made extensive use of data mining techniques to extract valuable information from vast amounts of network data. This work proposes a hybrid model that combines two distinct intrusion detection techniques: anomaly-based and signature-based. The model is separated into two stages. Systems for detecting intrusions make use of audit data produced by network devices, operating systems, and application software. These sources generate enormous databases that contain tens of millions of records. Data mining, which is the process of extracting meaningful patterns from a sizable amount of information, is used to analyze this data. The presented paper deals with the role and the applicability of data mining techniques in designing and developing the IDS Systems.***

Keywords: Data Mining, Intrusion Detection, Clustering, Classification, False Positive

I. INTRODUCTION TO INTRUSION DETECTION SYSTEM

Due to the rapid expansion of networked computer resources in recent years, numerous network-based apps have been created to offer services in a range of sectors, including social media, banking, government, e-commerce, and so on. Unauthorized activity has increased as a result of more machines being networked, both from external and inside threats, such as persons getting unprivileged access for personal gain **[1].** Unauthorized intrusions into computer networks and systems are detected by intrusion detection systems, or IDS. Incidents can involve malware attacks (such worms or viruses), attackers accessing the system without authorization via the Internet, or users acquiring unprivileged root access to the system without authorization. Similar to a network sniffer, an IDS gathers network log data and keeps an eye on a computer system's network activity. An intrusion detection model or approach is used to analyze the gathered network data and look for rule violations. The IDS sounds an alarm to notify the network administrator of any rule violations. Fig. 1 demonstrates a generic model of Intrusion Detection System.

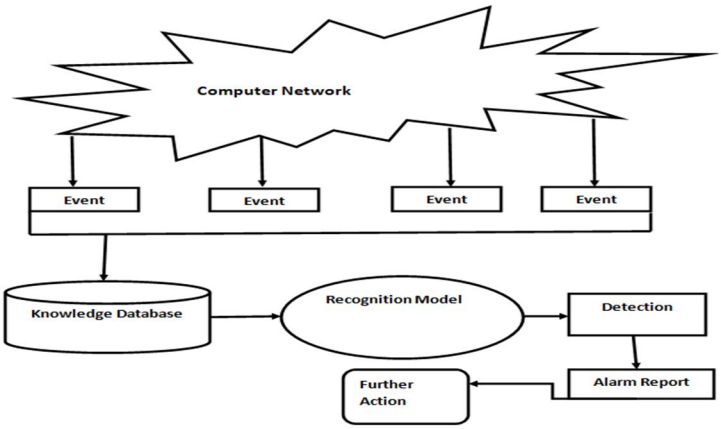


Fig. 1: An IDS Model

Intrusion is another term for malicious online activity. Any behavior that contravenes the network's security rules is considered an incursion **[2]**. In order to fill in the gaps left by firewalls and antivirus programmes, intrusion detection systems (IDS) are hardware and software that are used to detect unauthorized use of, or attacks against, computers or telecommunications networks. Monitoring and analyzing user and system activity, auditing system configuration and vulnerabilities, evaluating the integrity of important system and data files, statistically analyzing activity patterns based on comparison with known attacks, analyzing anomalous activity, and conducting system audits are all capabilities of an intrusion detection system (IDS) **[3]**. One benefit of the IDS is its capacity to record an incursion or threat to an organization, which gives system logs the basis for educating the public about the most recent assault patterns.

II. WHAT DATA MINING IS

The practice of obtaining knowledge or insights from massive amounts of data through a variety of statistical and computational techniques is known as data mining. The data can be kept in a variety of formats, including databases, data warehouses, and data lakes. It can also be semi-structured, unstructured, or structured. Finding hidden patterns and relationships in the data that may be utilized to generate forecasts or well-informed judgements is the main objective of data mining. This entails analyzing the data using a variety of methods, including anomaly detection, regression analysis, association rule mining, clustering, and classification.

Numerous industries, including marketing, banking, network security, healthcare, and telecommunications, have used data mining extensively. For instance, data mining in marketing can be used to pinpoint target audiences for advertising campaigns, and in healthcare, it can be utilised to pinpoint illness risk factors and create individualized treatment regimens. Similarly, data mining provide an excellent way to detect and protect intrusion in a computer network. IDS based on data mining is capable of effectively locating this user-interested data and forecasting outcomes for future usage. Both the IT industry and general public have shown a tremendous deal of interest in data mining, or knowledge discovery in databases. In order to extract meaningful information from massive amounts of noisy, erratic, and dynamic data, data mining has been used. It is positioned in the centre of the network to collect all incoming packets that are sent across it. After being gathered, the data are sent for pre-processing to eliminate noise and replace any missing or irrelevant features. Subsequently, the preprocessed data undergo analysis and classification based on their severity metrics. If the record is normal, then no further changes are needed; if not, a report is generated with warnings. Alarms are set off based on the data's condition so that the administrator can take proactive measures.

A. Stages involved in Data Mining Process

There are some stages applicable during the data mining. These stages are:

1) *Data Cleaning and Preprocessing:* Preparing the data for analysis through data cleaning and preprocessing is a crucial step in the data mining process. Data cleaning entails removing any superfluous characteristics or features, spotting and fixing outliers, adding missing data, and converting numeric variables from category ones. This entails transforming the data into a format that may be used for analysis as well as eliminating or fixing inaccurate, inconsistent, or incomplete data. Preprocessing also entails dimensionality reduction, feature selection to find significant characteristics, and normalization of the data.
2) *Data Modeling & Evaluation:* The procedure of developing models for machine learning with data and assessing their effectiveness is known as data modeling and evaluation. This entails choosing the right algorithm for the job, adjusting its hyper-parameters to maximize performance, and assessing the algorithm's output using metrics like accuracy or precision. A model is ready for deployment in practical applications once it has undergone training and evaluation. Additionally, abnormalities or outliers in the data can be found via data mining. Applications related to cyber-security and fraud detection will find this very helpful. Once anomalies or outliers have been found, analysts can look into the issue more thoroughly to learn more about it.
3) *Data Exploration and their Visualization:* The process of investigating, evaluating, and displaying data in order to draw conclusions and spot trends is known as data exploration and visualization. This entails taking descriptive statistics, such central tendency, dispersion, and feature correlation, and utilizing them to summarize the data. It also entails visualizing data point distributions and applying clustering or classification algorithms to put comparable data points in one group. By using these techniques, data scientists, data analysts, data engineers, and analytics experts can discover patterns in the data and understand the underlying structure of the information. Additionally useful for communicating and observing the relationships, correlations, and divergences across various datasets are data visualization tools like heat-maps, histograms, bar charts, and scatter plots. Furthermore, by describing information in fewer dimensions, dimensionality reduction techniques like principal component analysis (PCA) can aid in reducing the complexity of datasets. Analysts can determine which machine learning algorithms would be best for their project by first examining and visualizing the data.

4) *Hypothesis Formation:* Now is the moment to search the data for previously unidentified trends, patterns, and clusters. This stage involves the use of grouping, forecasting, and classification algorithms. Appropriate techniques are applied to assess each hypothesis, including loss matrix analysis, bootstrapping, and pass. The most beneficial theories are compiled and subsequently made public.
5) *Deployment:* When data mining is done, the trained models are put to use in a real-world setting. To assure the model's performance, it must be configured for real-time execution and any necessary monitoring measures must be put in place.
6) *Maintenance:* Furthermore, the model might need to be re-trained and re-deployed to production in response to any modifications made to the dataset or model. In conclusion, upkeep is also required to guarantee the model's functionality and maintain it current with modifications to the data or surroundings. Businesses may maintain the accuracy and dependability of their data mining models in production by monitoring these factors.

B. Techniques Applied in Data Mining

Data mining applies certain techniques as per the requirements of system or model. These data mining techniques are:

1) *Classification:* To divide data into preset groups or classes, classification is used. Using a variety of attribute values, this data mining technique determines the class to which an item belongs. The goal is to sort the data into predefined classifications. The most common use of classification is in the prediction of a variable that may possess one of two or more distinct values (spam/not spam; excellent or neutral/negative rating), given one or more input parameters called predictors.
2) *Clustering:* In this kind of data mining, related data points are grouped together into clusters according to specific traits or properties. Finding hidden structures or groupings in data as well as patterns in it are accomplished through the use of clustering. For example, clients who regularly buy particular drinks or food item/s and have consistent taste preferences can be grouped based on sales data. Once these clusters are established, it will be simple to target them with specific message or promotional advertisements. Molecular computation, text mining, online analytics, and medical diagnostics all employ clustering.
3) *Association-Rule Learning:* Finding patterns of correlation between objects in large datasets—like market basket analysis, which identifies the commodities that are frequently purchased together is the goal of this kind of data mining. Learning association rules is used to find if-then patterns among two or more independent variables.
4) *Regression:* Regression is another data mining approach. Regression is used to establish a relationship between variables. Finding the right function to effectively represent the relationship is its goal. Applying a linear function is known as linear regression analysis. Other types of correlations can be taken into account using techniques such as quadratic regression and multiple linear regression.
5) *Anomaly Detection:* To identify outliers, a data mining approach known as anomaly detection is employed (results that depart from the norm). For example, it can recognize unforeseen sales at a store location within a given week of e-commerce data. Among other things, it can be used to detect network intrusions or disturbances and identify credit or debit fraud.
6) *Sequential Pattern Mining:* Sequential pattern mining is a type of data mining that identifies important relationships between events. When we identify a time-ordered sequence that happens at a specific frequency, we can talk about a dependency between events. This method is widely used in medical procedures, DNA-Sequencing, shopping pattern, etc.
7) *ANN Classifier:* An artificial neural network (ANN), sometimes referred to as a "Neural Network", is a process model that may be supported by biological neurons. It consists of a networked assembly of artificial neurons. An input/output unit collection with weights applied to each connection makes up a neural network. Long training times are necessary for neural networks, which makes them better suited for situations where this is feasible. Numerous characteristics are required, such as the network topology or "structure," which are frequently best ascertained empirically. Neural networks have been criticized for their poor interpretability since it is difficult for humans to comprehend the symbolic meaning of the obtained weights. Initially, these features made neural networks less desirable for data mining.
8) *Prediction:* There are two phases involved in both data prediction and data classification. Though the attribute whose values are being anticipated is continuously valued (ordered) instead of category, we do not use the phrase "Class label attribute" for prediction. Creating and applying a model to ascertain the class of an unnamed item or the value or ranges of a specific characteristic that an object is likely to have are two examples of prediction.
9) *Genetic Algorithm:* Genetic algorithms, which are essentially adaptive heuristic algorithms, make up the majority of algorithmic developments. Genetic algorithms are based on natural selection and genetics. These are creative applications of random search that concentrate the search on regions of the solution space that perform better, as shown by past data. They are

widely used to generate outstanding solutions for problems relating to search and optimization. Genetic algorithms mimic the processes of natural selection, meaning that just those species that can adjust to changes in their surroundings will be able to endure, breed, and leave a legacy for future generations.

III. DATA MINING IN IDS

For intrusion detection, a wide range of data mining approaches are available, each with a unique set of benefits. However, it depends largely on the nature or motive of the intrusion detection models that is to be developed.

Finding security flaws in information systems is the aim of intrusion detection. Since intrusion detection keeps an eye on information systems and sounds an alarm when security breaches are found, it is a passive method of security. The misuse of privileges or the exploitation of attacks to take advantage of flaws in software or protocols are two examples of security breaches. The two main categories into which intrusion detection systems are traditionally divided are anomaly detection and misuse detection. The process of misuse detection involves looking for signs or patterns of widely recognized assaults **[4]**. Evidently, that method of detection is limited to known attacks that leave distinctive evidence. In contrast, anomaly detection makes use of a model of typical user or system behaviour and marks notable departures from this model as possibly harmful. The term "user or system profile" refers to this representation of typical user or system behaviour. The capacity of anomaly detection to identify as-yet-unknown assaults is one of its strengths. Furthermore, there are many classifications for intrusion detection systems (IDSs) based on the types of input data they examine. This results in the separation of network-based and host-based intrusion detection systems. IDSs that are based on hosts examine audit sources that are specific to the host, like application logs, system logs, and operating system audit trails. Network-captured network packets are analyzed by network-based intrusion detection systems **[5]**. Data mining is the process of sifting through data to find trends and build connections. The following are some data mining parameters: forecasting, classification, association, sequence analysis and clustering. A data preparation phase is one of the most crucial components of any data mining system. Approximately eighty percent of a typical real-world time is spent on data preparation **[6]**. An attempt at data mining. Inadequate data quality can result in absurd data mining findings that need to be rejected. The selection, assessment, cleaning, enrichment, and modification of the data are all included in data preprocessing. This brain modeling is a methodical approach to creating mechanical solutions. Compared to its more accustomed rivals, this new arrival style to computing also offers a more gradual decline during system overload. An interconnected collection of artificial neurons known as a neural network processes information using a mathematical or computational model that uses a connection approach to computing.

Although a neural network cannot initially be trained with domain knowledge, it can be trained to make decisions by mapping sample pairs of input data into sample output vectors and estimating its weights to approximate each input instance vector to the corresponding output example vector (Hecht-Nielsen, 1988).

IV. APPLICATIONS OF DATA MINING TECHNIQUES IN IDS SYSTEMS

High security controls are necessary with modern network technologies to guarantee secure and reliable information exchange between a client and user. The purpose of an IDS System is to safeguard the system in the event that conventional technologies fail. The process of extracting relevant information from a vast amount of data is known as data mining. Both supervised and unsupervised learning techniques are supported. Since intrusion detection is essentially a data-centric process **[7]**, data mining techniques can help IDS identify anomalous activity, learn from previous incursions, and enhance performance through experience. By enhancing segmentation, it assists in examining the substantial growth in the database and collects only reliable information, enabling organizations to make real-time plans and save time. It can be used for a number of purposes, including identifying suspicious activity, fraud and abuse, terrorist activity, and lying detection in criminal investigations **[8]**.

For the IDS Systems, the following applications of data mining techniques can be discussed.

A. Data Stream Analysis

The term "data stream analysis" refers to continuous data analysis; yet, because data mining requires sophisticated calculations and lengthy processing times, it is primarily applied to static data. The dynamic nature of malicious assaults and breaches makes intrusion detection within the records stream context much more crucial. Furthermore, even though an event appears normal on its own, it may be deemed malevolent if it is seen as a component of a larger set of events. Therefore, it's critical to identify sequential trends, look for outliers, and consider whether sequences of tasks are frequently encountered together **[9].**. Real-time intrusion detection also requires other data mining techniques for finding growing clusters and building dynamic class models in record streams.

B. Devising Innovative IDS Model

The IDS model's data mining technique has a lower false alarm rate and a greater efficiency rate. Data mining techniques are applicable to anomaly-based as well as signature-based detection. Training data is categorized as "normal" or "intrusion" in signature-based detection. Then, a classifier to find acknowledged incursions can be derived. Cost-sensitive modeling, association rule mining, and clarifying algorithms have all been used in research on this location. Completely anomaly-based detection creates models of typical behavior and automatically identifies significant departures from this **[10]**. Statistical techniques, class algorithms, clustering software, and outlier analysis software are examples of methods. The employed solutions must be effective, scalable, and capable of handling excessively large, multidimensional, and heterogeneous amounts of community data.

C. Distributed Data Mining

It is employed to analyze random data, which is naturally dispersed across many databases, making data processing integration challenging. Attacks can originate from a variety of unique locations and target a variety of unique locations. To find those dispersed attacks, community data from several network locations can be examined using distributed data mining techniques.

D. Visualization Tool

These tools are used to display the data as graphs, which makes it easier for the user to interpret the data visually. The aberrant patterns that are found can also be viewed using these tools. These tools could include the ability to view outliers, clusters, relationships, and discriminative patterns. A graphical user interface is actually required for intrusion detection structures in order to enable safety analysts to ask questions about the network data or intrusion detection findings.

V. CHALLENGES

Given that the IDS is a long-standing technology, it will inevitably face certain issues that are incompatible with the contemporary IT landscape. Because it has been around for so long, malevolent actors have developed evasion strategies to fool intrusion detection systems into missing attacks. Some of the major evasion techniques are:

1) *Fragmentation:* This is a simple approach that, in order to remain undetected, breaks the attack payload into numerous packets. Even while tiny packets cannot circumvent an intrusion detection system (IDS), they can be made to require difficult reassembly in order to avoid detection. Adding pauses to the payload while it is sending other parts in the hopes that the IDS would time out is one technique to create fragmentation. Other techniques include sending packets in the wrong order to trick the IDS but not the target host, or transmitting packets so that one fragment replaces data from a prior packet.
2) *Low Bandwidth-Attack:* **T**o avoid IDS detection, attackers can plan an attack that spans many sources and mimics harmless traffic and noise, like that generated by internet scanners, over an extended period of time. By making it difficult for the IDS to match every packet and determine whether this is malicious or benign scanning activity, this strategy works.
3) *Obscurity:* This method of IDS avoidance entails purposeful alteration of protocols to use alternate ports. The intrusion detection system will fail to identify an intrusion if it fails to respond to these protocol violations in the same manner as the target host.

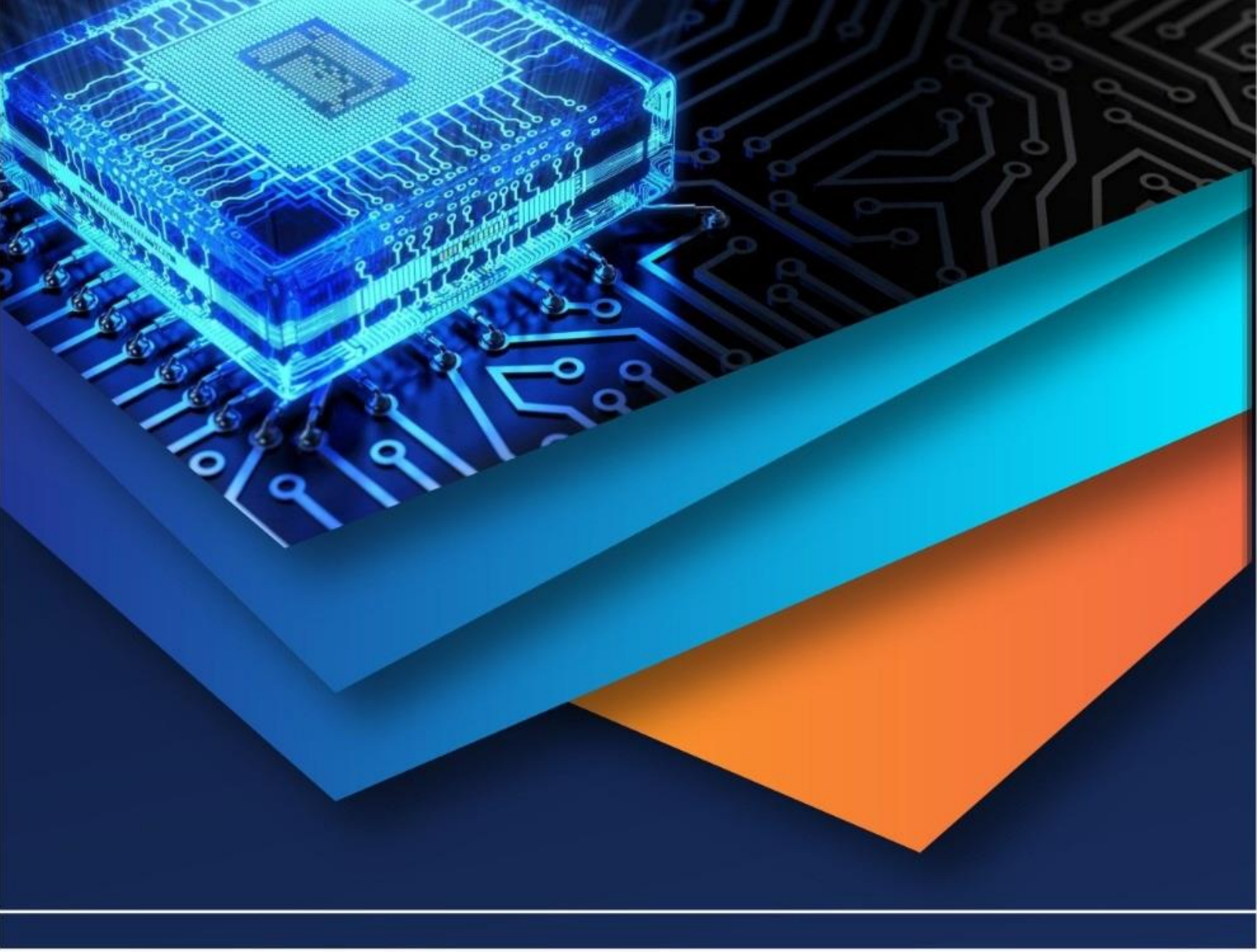
VI. CONCLUSION

A modern computer system's planning involves considering network security. In computer network security, intrusion assault detection is the most crucial problem to solve. Based on the methods used for detection, existing IDS can be categorized into two groups: misuse detection and anomaly detection. The identification of anomalies, misuse, and mis-configuration can be done in a number of ways. Applying data mining techniques greatly increase the capabilities of IDS Systems. The k-means algorithm is a value-sensitive starting method that yields varying clustering outcomes depending on the value of k. An enhanced version of the k-means algorithm is presented in this work in light of its flaws. The data set automatically generates an ideal value through clustering, therefore the k value does not need to be known in advance. After determination, there is no need to modify the clustering centre, and only one scan of the entire data set is required. Both the clustering effect and the enhanced k-means algorithm are now much more efficient.

REFERENCES

[1] Basant Agarwal, Namita Mittal, “Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Techniques”, 2nd International Conference on Communication, Computing & Security, Procedia technology, ScienceDirect, Elsevier Publication, 2012, DOI: 10.1016/j.protcy.2012.10.121
[2] Bischof, H., Leonardis, A., and Selb, A. MDL principle for robust vector quantisation. Pattern Analysis and applications. 2:59-72,1999.
[3] SANS Institute. Understanding Intrusion Detection System. 2001.

[4] K Bala, N Kumar, AK Singh, Performance Evaluation of Advanced Intrusion Detection System, International Journal of Research in Engineering and Technology Volume 7 Issue (10), pp- 10 – 15, 2018.
[5] Shankar Kumar, Dr. Nandeshwar Pd. Singh, Dr. Narendra Kumar."Mechanism, Tools and Techniques to Mitigate Distributed Denial of Service Attacks", Volume 11, Issue I, International Journal for Research in Applied Science and Engineering Technology (IJRASET) Page No: 855-861, ISSN : 2321-965
[6] Curtis A. Carver, Jr., Jeffrey W. Humphries, and Udo W. Pooch,"Adaptation Techniques for Intrusion Detection and Intrusion Response Systems
[7] Eduardo Mosqueira-Rey, Amparo Alonso-Betanzos, Belen Baldonedo Del Rio, and Jesus Lago Pineiro, " A Misuse Detection Agent for Intrusion Detection in a Multi-agent Architecture". Springer-Verlag Berlin Heidelberg 2007
[8] Jiawei, H. and Micheline, K. Data Mining Concepts and techniques, second edition, China Machine Press, pp. 296-303. 2006.
[9] Sabhani, M., and Serpen, G. Why Machine Learning Algorithms Fail in Misuse Detection on KDD Intrusion Detection Dataset. Intelligent Data Analysis, vol 6. (Jne 2004).
[10] Siddiqui, M.K., and Naahid, S.Analysis of KDD CUP 99 Dataset using Clustering based Data Mining. International Journal of Database Theory and Application Vol.6, No. 5. pp.23-24. 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY