Nanotechnology Perceptions
ISSN 1660-6795
www.nano-*ntp.com*

Clustering Based Anomaly Detection Using Machine Learning

RENUKA KONDABALA [1], RADHA MOGATALA[2], V V NAGENDRA KUMAR[3]

[1,2] ASSISTANT PROFESSORS, DEPARTMENT OF INFORMATION TECHNOLOGY, VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY, BACHUPALLY, HYDERABAD, TELANGANA, INDIA.
[3] ASSISTANT PROFESSORS, DEPARTMENT OF MCA, RAJEEV GANDHI MEMORIAL COLLEGE OF ENGINEERING AND TECHNOLOGY, NANDYAL, ANDHRA PRADESH, INDIA..
**Corresponding Author Email Address: (Renuka_K@Vnrvjiet.In)*

September 5, 2024, submitted
November 12, 2024, accepted
December 8, 2024, published

When cybersecurity is involved, the most crucial thing is immediate and accurate detection of any attempt on illegal intrusion so important data and systems are put at defense and therefore could not be tempered with. Currently, it has an ongoing project that is mainly geared toward implementing the highly effective technique of anomaly detection algorithms. For the most part, this project calls for long and arduous iterations and modification of several parameters. The primary purpose of the task is growing the level of security provided to an organization by using sophisticated tools and technologies. The anomalous behavior, that are possible suspects of breach, can be identified with the help of instances of anomaly detection algorithms, which contribute to a security alert. Such algorithms are used in the process of examining large-scale datasets and network traffic to look for deviations and any anomalies from the norms evident. As we strive to enhance efficiency of our anomaly detection system, we undergo a detailed data preparation procedure. This refers to data collection from heterogeneous sources, cleaning it, and presenting it into a format that chosen algorithms are expecting. The preference of these algorithms is centered around exactness ensuring the parameters are tweaked and optimized to the maximum level of accuracy in reaction to threats. The main objective of the proposed project is to transition from researching and enhancing the anomaly detection algorithms constantly to always meeting the challenges and the variability of the cyber threats. The eventual aim would be to have an extensive and a fresh cybersecurity system which makes the platform lead in the assaults by being able to identify the attempts of access.

Key Words— Anomaly detection, Machine Learning, Clustering Algorithms, Network Security, Cybersecurity, Unsupervised Learning, K-Means, DBSCAN , Isolation Forest , LOF (Local Outlier Factor), NSL-KDD Dataset, IDS-2017, Privacy and Data Security.

I. INTRODUCTION

Search for irregularities figured out as a network stability and keeping factor, as they may be a symptom of threats like cyberattack, failures or wrongdoings in the system. Clustering based anomaly detection is the solution of the most prominent problems that are faced by network

data; it is also another powerful and correct way to find hidden patterns in data. This technique basically is accumulation of data points that creates visually a direct and clear indication of the patterns that are

not usually found. Besides, interruptions and preservation of matrix integrity may be managed implicitly, using clustering algorithms' power. The major focus of this research is, based on clustering patterns, the identification of abnormalities which in turn will promote stability and safeguard the network. This method enables organizations to anticipate cyber threats, weaknesses in the systems and address them through identifying deviations that would otherwise have been spotted via standard system monitoring. It potentially ensures the high level of persistence of the network infrastructure of the whole organization as well. Utilization of clustering for anomaly detection is a core part of network reinforcing defenses in the fight against cyber-attacks penetrating into operational networks and with the main aim of uncharacterized digital attacks.

II. RELATED WORK

Zhang, L., Zhang, W., and Jiang, H [1]
In 2021 an article. The article was titled "A Combination Unsupervised Clustering Approach Approach for Anomaly Identification". IEEE published the article. Within this paper a new approach will be presented fusing the best aspects out of anomaly detection and unsupervised clustering. This method will provide a twelve-step plan which is available to other researchers to take different types of data. The submitted article focuses on increasing accuracy and effectiveness, thereby locating anomalies or outliers that deviate from normal. Such a method makes it possible to identify many types of deviations in a dataset because the algorithms weigh every anomalous pattern present in the data and learn the shape of normal samples and in so doing combine clustering algorithms.

GUO, L., ZHENG, K., AND WANG, Y [2]

Label estimation for network anomaly detection with the aid of clustering techniques: The mentioned paper which is published in Science Direct, highlights the new method which was established by Wang et al. . The paper concentrates on clustering as one of the techniques that can be employed to detect network anomalies. The objective here is to design algorithms for examining network traffic information in an effort to improve the operation of the abnormality spotting system. In this way, which makes spotting anomalies easier than what would have happened if that training data had to be specifically labeled, it helps in estimating labels to data points. The method of clustering involves a grouping of related points within the data that can accurately detect anomalous activities in a network system. This method provides a defense strategy against unauthorized abnormalities on the network systems.

PARK, J., KIM, M., & LEE, S [3]

With the ODC algorithm, an intelligent overpass of network traffic anomalies at a scale never seen before is accomplished." The IEEE publication advances a novel anomaly detection strategy using the methodology of ODC by Lee et al. It is based on a mix of these algorithms: clustering and deep learning. Managing a huge amount of data is the outcome of this method.

The core of ODC Algorithm's value is its capability to deal with enormous volumes of data and scour them for patterns or anomalies by means of using deep learning schemes to enhance detection of network oddities. This technique, through its management of massive traffic and its observation processes thus comes as a new approach to achieving network security.

Natalino, Carlos [4]

Deep Unsupervised Learning for Optical Network Monitoring: As a speech pathologist, one of the essential skills I should highlight in my resume is the ability to detect spectrum anomalies. Carlos Natalino's research work, that has been presented in IEEE, did experimentation on the application of deep unsupervised learning techniques for detecting anomalies in the spectrum of optical networks. According to Winkel Fz., deep learning models can learn the representations of the spectral data by themselves, creating the basis of the early detection of anomalies in the network by not using pre-labelled datasets. This way embraces radiometric methods the most, allowing control of the unhealthy links in the optical networks, where the correct transmission of information may depend on the early discovery of anomalous signals.

Xiao Luo, Md. Ahansul Kabir [5]

Unsupervised Learning for $Network Flow based Anomaly Detection in the Deep Learning Delete an example sentence With this IEEE article, Kabir and Luo present the execution of deep and unsupervised learning methods for detecting anomalous networking flows. This technique has the novelty of finding patterns in raw network traffic flow that no other market-based ATI solution has to rely on labeled training data. This method is one of the proofs of how non-supervised learning models can be posed within cybersecurity in order to employ breakthroughs from deep learning to lower network anomaly detection .

Luardi, William Tesardo [6]

The proposed Convolutional Autoencoder with Adversarially Regularized Structure that is the CAA-ARS for network anomaly detection. The author laying out a unique way to form IEEE about using an adversarially regularized convolutional autoencoder is presented by Luardi. This way of boosting the The area of cybersecurity is ever-changing, with the emergence of new advanced threats every day. The model is fully capable of addition of diverse data representations due to its adversarially regularized structure. Furthermore, this feature aids in the most effective detection of network anomalies.

Yating Liu, Quan Yu, Xinyue Shen, Yuantao Gu, and Qingmin Liao [7]

My Bitterness to the Security of the Backbone Network-Unsupervised Anomaly Detection System. This IEEE paper suggests an unsupervised anomaly detection system for backbone networks by making use of a backbone network. On the other hand, Liu et al. introduce the use of unsupervised learning techniques, eliminating the need for known data training, to assist in the identification of anomalies in network traffic. The aim is to enhance the features of the system such as versatility and scalability so that it can satisfy the security requirements at the varying levels of the large networks.

Lec Duc C. [8]

"Unsupervised Ensembles for Insider Threat Anomaly Detection." Lec proposes in his IEEE paper how their utilization of unsupervised ensemble learning can lead to the identification of anomalies that may have been attributed to the insider threats. Differing from conventional methods that employ labeled samples as training data, the methodology merges the models of ensemble learning in order to achieve better effectiveness in detecting such employee behaviors as possible insider threat. Altogether with this approach, it would become easier to improve security strategies against internal risks that can be seen in complex agencies.

Jiewen Mao, Fuke Shen, Dong Jiang, Youngquan Hu, Tongquan Wei. [9] .
CBFS: Introduction The feature selection mechanism in a system network for detecting anomaly based on clustering network. In this IEEE article the researchers Mao et al gave an introduction to their CBFS mechanism which uses clustering algorithm and network anomaly detection. The drivers of this technique instance is selecting notable network data characteristics and integrating them into the anomaly detection system to enhance the efficiency and accuracy of it. This is done through the implication of clustering techniques that are used for identifying and choosing relevant features which, in turn, boosts the sensitivity of security app systems to detect the anomalies in the network domain.

Riyaz Ahamed Ariyaluran Habeeb [10]
The suggested structure, that consist of BroIDS and Flume, Kafka,%Spark streaming, SparkMLlib, Matplot and HBase were evaluated in order to see if this framework manages to achieve its purpose of being effective and fast, especially taking into consideration the accuracy,$ memory consumption, and %execution time. Evaluation will be done to compare the existing techniques of data clustering such as K-means and hierarchical density based algorithms (HDBSCAN), isolation forest, spectral %clustering and agglomerative% clustering. Beyond this, the test cases were performed through the various performance metrics, such as Altered Random Index, and the memory$profile package of the existing methodologies. It was unambiguously proved that the suggested framework meets its objectives and enjoys an accuracy rate as much as 96.51% with the other algorithms.

III. PROPOSED SYSTEM

To make the network more secure against possible cybertraps, the suggested system involves the use of an advanced machine learning architecture in which anomalies are precisely detected in the network data. Defining the usual behavior of a signaling system through machine learning algorithms, such as the K-Means clustering and DBSCAN techniques, represents the first step in the system's methodology. This creates a true picture of normal activity that is not distorted by flawed assumptions. Further suspicion arises. That the model starts with a very fine and completed data cleaning package so all the possible helpful inputs can be steered belongs to procedures. Training is achieved by using large well-defined datasets such as NSL-KDD and IDS-2017, which can be applied to system responses triggered by logs and packet metadata to enable the recognition of both overt and hidden cyber threats. The next step within the system is a participatory approach that the experts employ to identify the abnormal circumstances, flash them back and perform analysis on them as a result of which a reciprocation system is created. During this cyclical process, we optimize the detection models and thereby keep bringing accuracy and reducing false positives in the model. Real-time

monitoring and reporting functionalities of the system, will enable the security personnel to be reactive and informed to incidents quickly. The system will also support a wide range of network setups, configurations, and security protocols, thus enabling it to be applied in different corporate setups. The security and privacy of data is of utmost importance and is achieved through established CCTV system access control and encryption protocols that prevent unauthorized data.One of the most vital parts of the program is the continuous updating and adding new features, which will ensure the algorithms' ability to keep up with novel threat intelligence and cybersecurity studies. This is accomplished by using such means; they not only prevent many kinds of network attacks and weaknesses, but also enhance the company in the long-run to overcome them. The mentioned updates will consider how threat vectors may shift in the future and at the same time will monitor the most recent hacking trends through related information sharing. Hence, as opposed to a combination fast-eradicating or inhibiting, the system serves as a long-term investment in a firm's cybersecurity. Besides the system's defense against present threats through integration of cutting edge research and threat intelligence into operations, it may predict future dangers and always improve on defense capabilities through this new integration process. This can be done by ensuring that the company is not just reactive to cyber threats, but proactive in outdating its resilience from attacks, as it becomes more formidable and more adept at handling these threats as the business thrives.

IV. EXPERIMENTAL ANALYSIS

Dataset
NSL-KDD(42 Attributes)
IDS – 2017(79 Attributes)
NSL-KDD and IDS-2017 data, that are about cyber security and network intrusion detection, are used as well. They constitute the so-called tools which allow determining an intrusion detection system and practicing its development. Below is a summary of every dataset:Below is a summary of every dataset:

NSL-KDD DATASET

NSL-KDD (KDD Cup 99 Dataset Following Feature Selection)
Attributes: The data set in this version has 42 features (attributes), which are derived from the data of network traffic flow. The components of this characteristics is a list of protocols, including connection type, source and destination IP addresses, protocols, service, flags, and more.
A subset of the KDD Cup 1999 dataset referred to as NSL-KDD is exclusively composed of that data set. Real-world data provided by simulators is heavily used in development and testing of intrusion detection systems. The source dataset had this limitation because there were repeated records and feature selection was not applied. NSL-KDD dealt with these shortcomings by eliminating duplicate records and adding the feature selection technique.which NSL-KDD solves by deleting duplicate records and doing feature selection.

Instance	Duration	Service	Flag	Src_ bytes	...	Difficulty	Label	Attack
1	0	ftp_data	SF	491	...	20	normal	normal
2	0	private	S0	0	...	19	neptune	neptune
3	25950	private	RSTR	1	...	15	portsweep	portsweep
4	0	ftp_data	SF	334	...	11	warezclient	warezclient
...	...	...	...	...	...	...	...	...
125973	0	ftp_data	SF	151	...	21	normal	normal

Attack Class	Attack Type
DoS	Back, Land, Neptune, Pod, Smurf, Teardrop, Apache2, Updstorm, Processtable, Worm
Probe	Satan, Ipsweep, Nmap, Portsweep, Mscan, Saint
R2L	Guess_Password, Ftp_write, Imap, Phf, Multihop, Warezmaster, Warezclient, Spy, Xlock, Xsnoop, Snmpguess, Snmpgetattack, Httpptunnel, Sendmail, Named
U2R	Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps

IDS-2017 (79 Characteristics):

IDS-2017 (Intrusion Detection System 2017 Dataset) is the name of the dataset.
Attributes: There are 79 representable features (the attributes) in this data set relating to network traffic and system logs. They are, for instance, the network conserving the value of packages, system events, and any features that recognize normal and abnormal behavior.
The determinant of IDS-2017 is to carry out studies and observations about cybersecurity as well as intrusion identification systems. The data set is a very good option for evaluating intrusion detection system performance, as it can generate expert data mainly considering the network traffic and system logs whose conditions can be used to mimic real-world network scenes. Both databases are considered for work in the system while the intrusion detection as well as machine learning models training and testing take place. These datasets serve as grounds for cybersecurity researchers and practitioners to test the validity of intrusion detection algorithms and to discover network anomalies. Using this datasets, strategies to strengthen networking security are also developed.
The effectiveness of the approach is assessed using the ImageNet VID dataset, and results support it. The suggested approach demonstrates notable gains over both static image detectors

and earlier state-of-the-art approaches, particularly for quickly moving objects. Overall, this methodology combines cuboid proposal, short tube detection, and linking techniques to exploit temporal context and improve the accuracy of object detection in videos.

Instance	Destination Port	Flow Duration	Total Fwd Packets	Min Packet Length	...	Label
1	80	38308	1	6	...	BENIGN
2	80	3	2	0	...	DoS Hulk
3	80	5011127	8	0	...	DoS GoldenEye
...	...	...	...	...	...	...
692703	80	99999734	2	0	...	DoS Slowloris

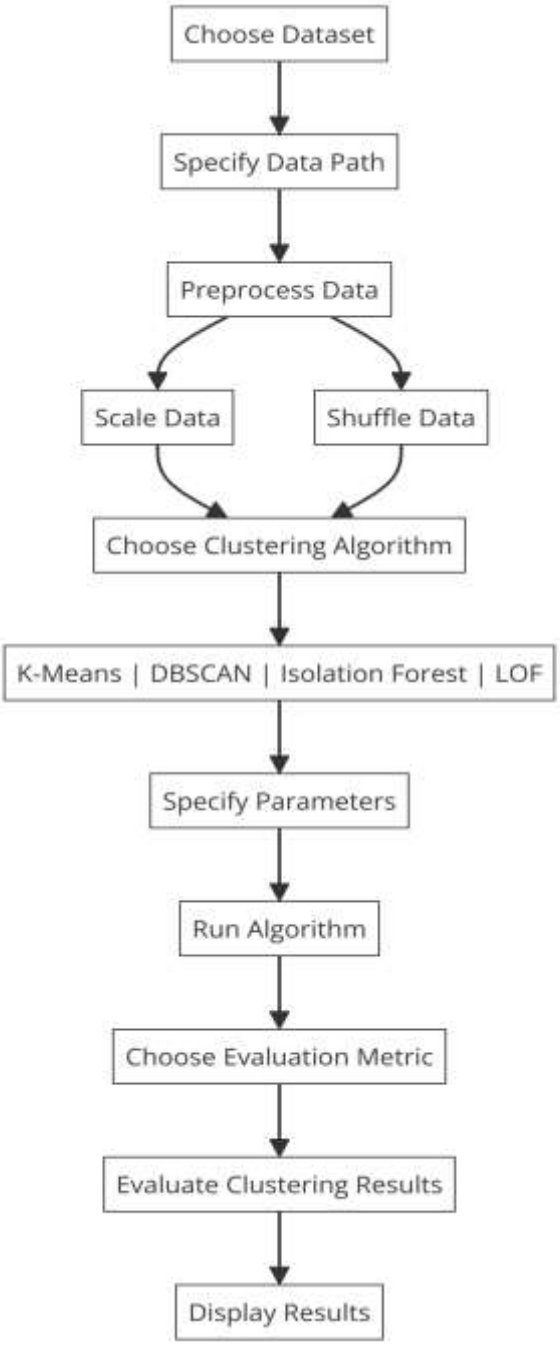
V. ALGORITHMS

As definite and quickly-chained as the field of network security is, one should know how to identificate anomalous activities and possible threats in the network systems. Algorithms of anomaly detection are, therefore, a primary care based system which acts as a front line of defense for automated surveillance to identify the anomalies in the patterns of data and network traffic. Detection algorithms are the foundation for the success of security measures using the proactive approach, as those oddities might show the true danger or vulnerability.

K-Means, DBSCAN, Fire Isolation, and Local Outlier Factor are the algorithms that are most commonly featured in the network security frameworks. The methods each algorithm uses to find outliers and how well-suited it is for various kinds of network data sets set them apart: As a K-Means is a machine learning (ML) algorithm that is known to be the most powerful for partitioning data sets into clusters, which can be mapped to sets of data points in the same concise manner to uncover anomalies present in datasets. Although this does not prevent the algorithm from employing these examples for later training phases, this could be seen as an obstacle. The K-Means Clustering technique truly comes alive when used in a complex network environment since it can locate clusters regardless of shape or size, is noise resistant, and can handle clusters of varying densities. Separation Forest plays a main role in data settings trapped in a high-dimensional area; it is sought to detect anomalies by detecting outliers which are excluded from the normal points. It addresses the variation of the feature density by calculating the local density deviations between a data point and its neighboring values. LOF allows identifying smart anomalies where the data density significantly changes.

Using these algorithms would most likely entail firstly touching bases on path specification and dataset delineation. The second step in data processing is data preprocessing. Data is collected, organized, and scaled to feed the algorithms in the most optimal way. Begin by selecting the appropriate clustering method and then adjust the parameters of that method by taking into consideration the special network data attributes in stage one of the procedure. the metrics criteria are defined for change according to the type of data and the expected result, and a post algorithm execution process are implemented to assess outcomes. The final step is to share the outcomes. This is when the network security vendors will be assessed based on

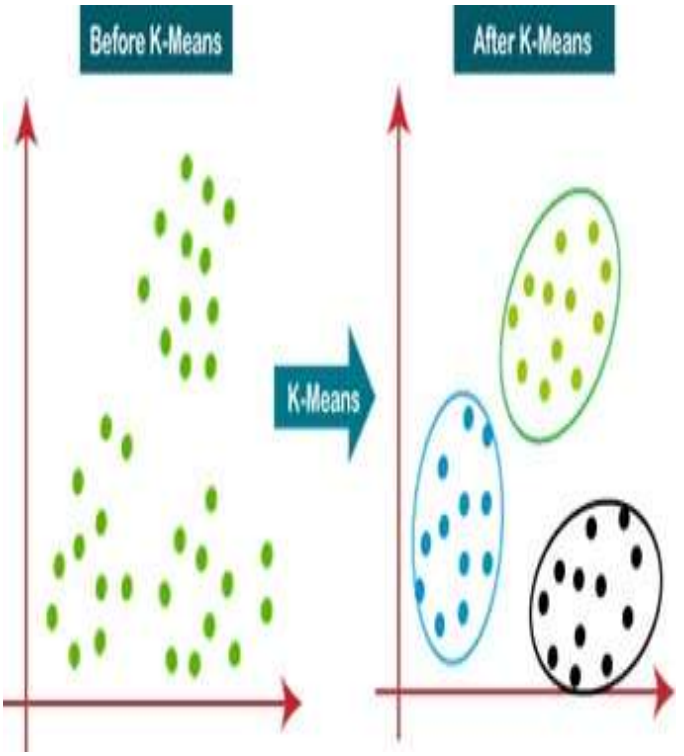
what findings were made and the appropriate response taken to ensure protection against any threats detected.



K Means:

A method for clustering is being actively used to detect anomalies in network data, and it is

called k-Means. It groups data into almost certain relationships or affinities to observe and find patterns among the data elements that fall in clusters aside from the normal randomness. “The data outside known clusters are called anomalies and can be outliers as well”. Concerning these kinds of issues, the K-Means might does not work proper in cases when different clusters are formed in a peculiar way, like in odd shapes, sizes and densities which are typically found in big networked data.



Step 1:Initialization
Suppose we have A(2, 3), B(5, 4), C(9, 6), D(4, 7), E(8, 1), and F(7, 4). Select and initialize two centroids$ at random (k = 2). Assume we select the next two centroids
two centroids$at% random (k = 2). Assume we select the next two centroids:

Centroid 1 (C1) at (2, 3)

Centroid 2 (C2) at (8, 1)

Step 2:Assignment
Determine the distance between each point and each centroid, then allocate the points to the centroid that is closed to them.

Point	**Distance to C1**	**Distance to C2**	**Assignment**
A(2, 3)	0	5	C1
B(5, 4)	3.61	6.32	C1
C(9, 6)	9.22	5.09	C2
D(4, 7)	5.83	5	C1
E(8, 1)	7.07	0	C2
F(7, 4)	6.08	5.09	C2

Step 3: Update
After assigning all of the points, compute the average of points in each cluster to update centroids.

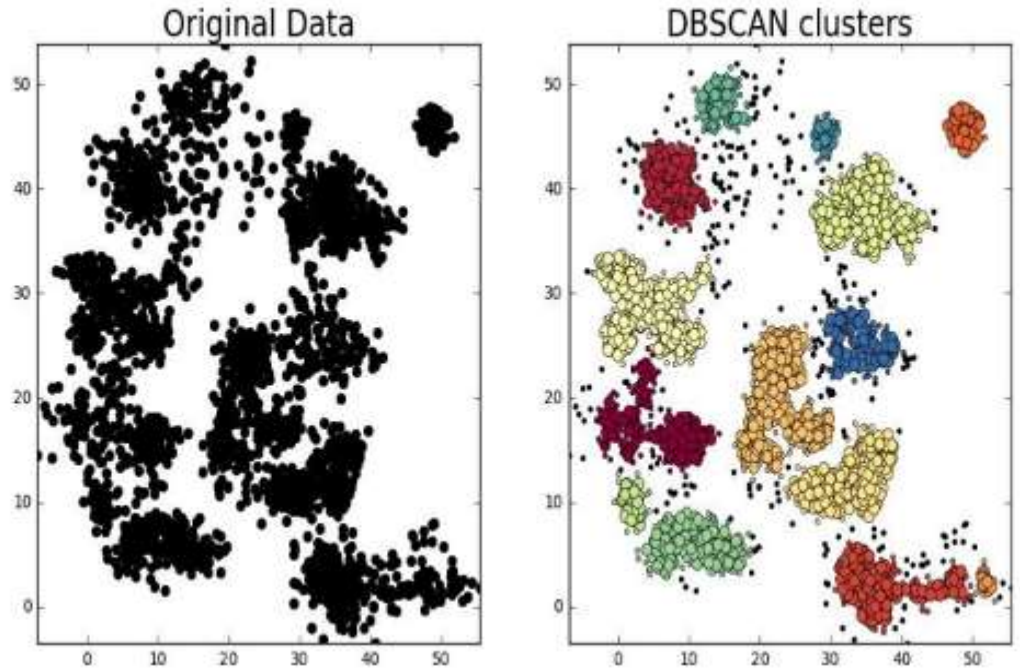
Cluster	**Centroid**
C1	(3.67, 4.67)
C2	(8, 3.67)

Step 4: Convergence Check
Continue steps 2 and 3 until there is no discernible change in the centroids or until the allotted number of iterations is reachedThe method has converged in this instance since the centroids have not changed considerably.

Cluster	Centroid
C1	A,B,D
C2	C,E,F

DBSCAN (Density Based Spatial Clustering of Applications with Noise):

DBSCAN,estimation of density level clustering technique, is employed to identify anomalies. Regular traffic often creates dense clusters, which makes it beneficial for spotting anomalies in network data. On the other hand, anomalies or efforts at intrusion that do not belong in any concentrated area are seen as noise. DBSCAN is a helpful tool for spotting odd patterns and network intrusions since it can discover irregularities in sparse regions.



Assume for the moment that the 2D dataset contains the following$ points:
Data Points: A(2, 3), B(3, 4), C(3, 6), D(4, 5), E(6, 6), F(7, 6), G(8, 5), H(7, 3), I(6,2)

Step 1: Initialization
Set ε (maximum distance) to 1.5 and MinPts to 4.

Step 2: Point Selection
Choose an arbitrary data point, let's start with point A(2, 3).

Step 3: Neighbour Search
Find all the data points within a system ε (1.5) from point A. Calculate the Euclidean distance between A and all other points:
Distance(A,B) = √((2-3)^2 + (3-4)^2) ≈ 1.41
Distance(A,C) = √((2-3)^2 + (3-6)^2) ≈ 3. 16
Distance(A,D) = √((2-4)^2 + (3-5)^2) ≈ 2.83
Distance(A,E) = √((2-6)^2 + (3-6)^2) ≈ 5.00
Distance(A,F) = √((2-7)^2 + (3-6)^2) ≈ 6.40
Distance(A,G) = √((2-8)^2 + (3-5)^2) ≈ 7.81
Distance(A,H) = √((2-7)^2 + (3-3)^2) ≈ 5.10
Distance(A,I) = √((2-6)^2 + (3-2)^2) ≈ 3.16

Step 4: Core Point Identification
Verify if there are more data points in A's ε-neighborhood than or equal to MinPts. In this case, it's not since A only has three neighbors (B, D, and I).

Step 5: Cluster Expansion
Since A is not a core point, it will be marked as noise, and we move on to the next unvisited point.

Step 6: Repeat
Continue this process for each unvisited point until all points are assigned to clusters or marked as noise. In this example, you would visit point B next, and since it has enough neighbors, it would be a core point and start forming a cluster.

Isolation Forest:
The efficacy of this method in high-dimensional datasets is widely recognized, which is crucial when working with substantial volumes of network traffic data. It functions by successfully separating anomalies. Anomalies are defined as data points that can be separated from the rest of the data with fewer partitions. Since anomalies are typically odd occurrences with unique characteristics from regular data, this approach works particularly well with network data.

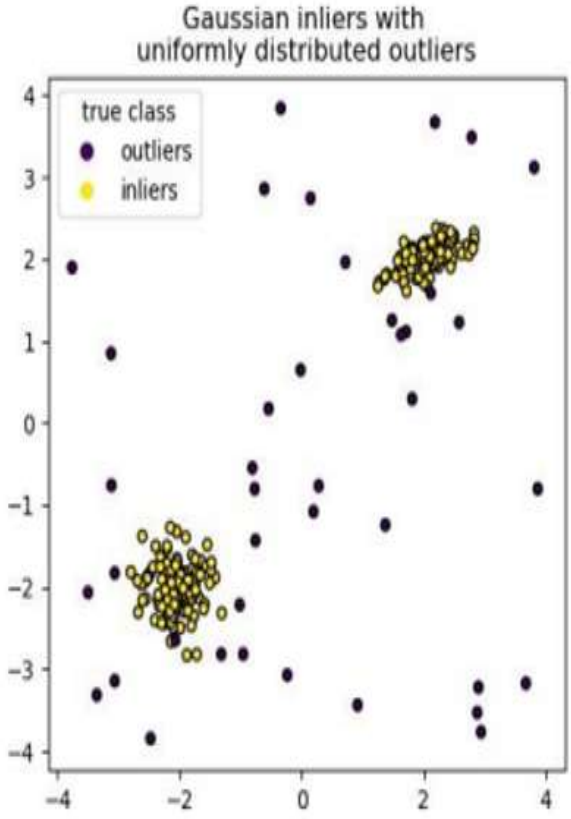


Fig 4.2.3 Isolation Forest(IF)

The formula provides the actual computation of the irregularity result (s) for an example x: The midpoint shortest path of point x over a collection of isolation trees is given by the notation E(h(x)).

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

The formula to approximate the average path length of an misfired search in a Binary Search Tree (BST) is c(n) = 2h(n-1) - (2(n-1)/n), where h(i) is the H.M and can be roughly calculated by ln(i) + 0.5772156649.

n is the no.of external nodes (i.e., the no.of samples used to build the iTrees)

We created a synthetic dataset with 9 points. Most of the points are close to each other except for a few outliers:

[1, 2], [2, 3], [2, 2], [8, 8], [8, 9], [7, 8] (Inliers)

[0, 0], [100, 100], [10, 10] (Outliers)

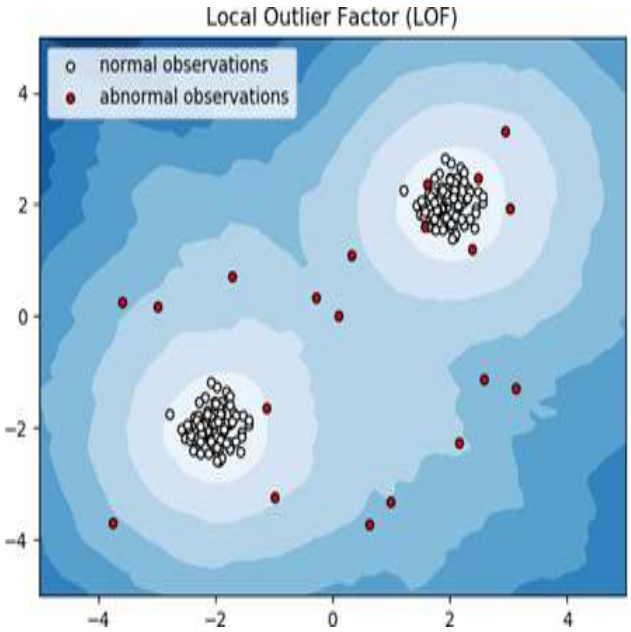
We calculated the anomaly scores and predictions for every data point once the model was fitted. A positive score designates an outlier, while a negative score denotes a data point that is thought to be normal. For inliers, the prediction is 1, while for outliers, it is -1.

Data Point	**Anomaly Score**	**Prediction**
[1, 2]	0.079	1
[2, 3]	0.069	1
[2, 2]	0.098	1
[8, 8]	0.115	1
[8, 9]	0.093	1
[7, 8]	0.084	1
[0, 0]	-0.039	-1
[100, 100]	-0.303	-1

[10, 10]	0.006	1

The data points [0, 0] and [100, 100] have been correctly identified as outliers. The point [10, 10], although it is farther away from the cluster, was not considered as much of an outlier compared to the others, which is reflected in its lower but still positive anomaly score

Local Outlier Factor:
When anomalies have different local density properties but are otherwise identical to normal data, LOF can be useful. It implements the density estimation method that is based on the evaluation of the neighborhood of every data vector relative to its neighbors. The spatial density functions called Local Outlier factors are nanoparticles called contrast agents help in the picturization of such nanoscale anomalies inside the tissue like clots for a trend to be considered significant, it is required that points arrayed in very different categories should be present local densities as abnormal. These are the events that stand apart from the overall design of the underlying constant. This could suggest that someone potentially within is trying to penetrate the system or that any other irregular activity is occurring there are sold their goods by means of online platforms.



A strategy called regional departure rate (LOF) is used as well understand how much the value of a datapoint differs from its neighbors in the local area, detection of the outliers or exceptional points in directly is because of the use of anomaly detection techniques, while in the second case identification of contextual information is supported by the use of sentiment analysis. Every data point is evaluated and given a weight by LOF in such a way that a higher than average score symbolizes a higher weight and increased likelihood that is outside the norm.An example of calculating LOF is illustrated by the 2-dimensional dataset below for the basic data.

Assuming a dataset as 2D and having 5 data points to simply point out how LOF fences out a normal distribution or an entire group of observations from the otherwise normal points. The dataset includes an obvious outlier to help illustrate the calculation more effectively:

Data Points: A(1, 1), B(2, 2), C(3, 3), D(4, 4), E(10, 10)

We aim to calculate the LOF score for each data point, particularly focusing on point E(10, 10) to show how it is identified as an outlier.

1.Select a Data Point:
We start with data point E.

2.Calculate the k-Nearest Neighbors:
Let's choose k = 2 for simplicity.
Calculate the distances between point E and all other data points:
Distance(E, A) = √((10-1)^2 + (10-1)^2) ≈ 12.73
Distance(E, B) = √((10-2)^2 + (10-2)^2) ≈ 11.31
Distance(E, C) = √((10-3)^2 + (10-3)^2) ≈ 9.90
Distance(E, D) = √((10-4)^2 + (10-4)^2) ≈ 8.49

The two nearest neighbors of E are C and D.

3. Calculate the Reachability Distance:
For k=2, we consider the distances to C and D, which are the two nearest neighbors.
Reach-Dist(E, C) = max(Distance(E, C),
Distance(E, D)) = max(9.90, 8.49) = 9.90
Reach-Dist(E, D) = max(Distance(E, D),
Distance(E, C)) max(8.49, 9.90) = 9.90

4. Calculate the Local Reachability Density (lrd):

lrd(E) = 1 / [average Reach-Dist(E, N) / k]
= 1 / [(9.90 + 9.90) / 2] = 1 / 9.90 ≈ 0.101

5. Determine the Score for LOF:

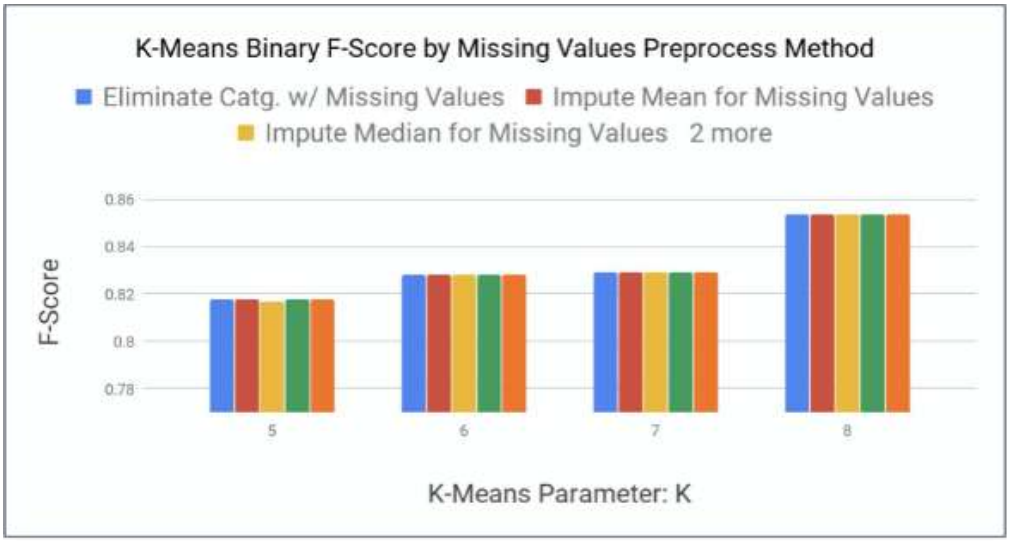
To find the LOF of E, we need the lrd of its neighbors (C and D). Let's assume after calculations similar to above, we find lrd(C) ≈ 0.333 and lrd(D) ≈ 0.333 (for illustration).

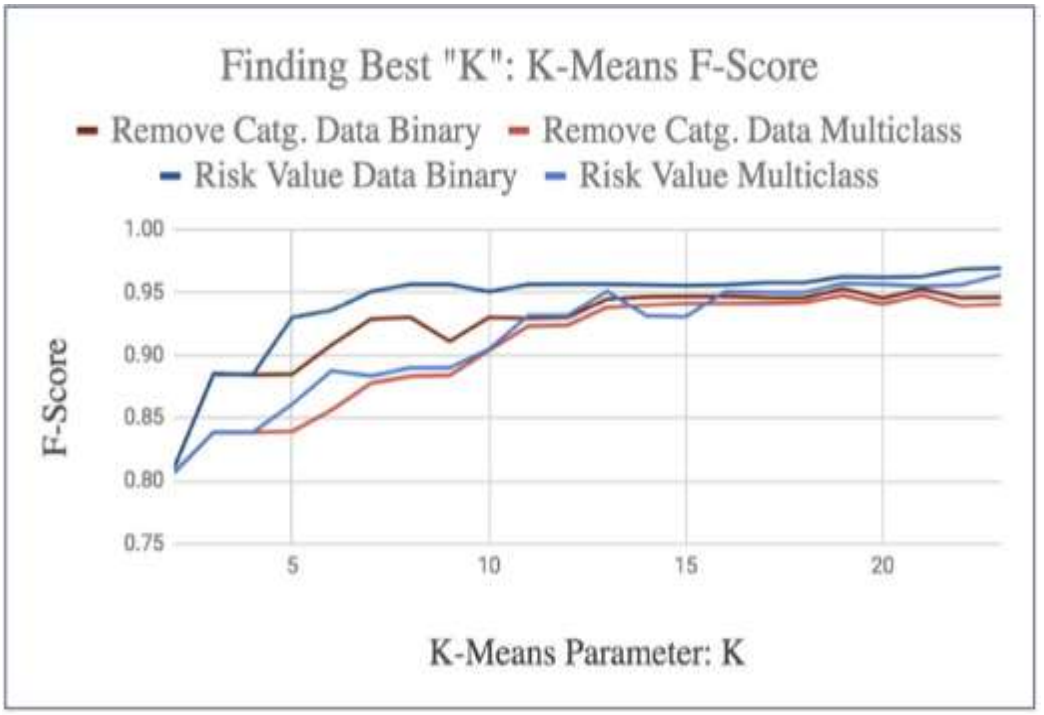
LOF(E) = [(lrd(C) + lrd(D)) / k] / lrd(E)
= [(0.333 + 0.333) / 2] / 0.101 ≈ 3.30

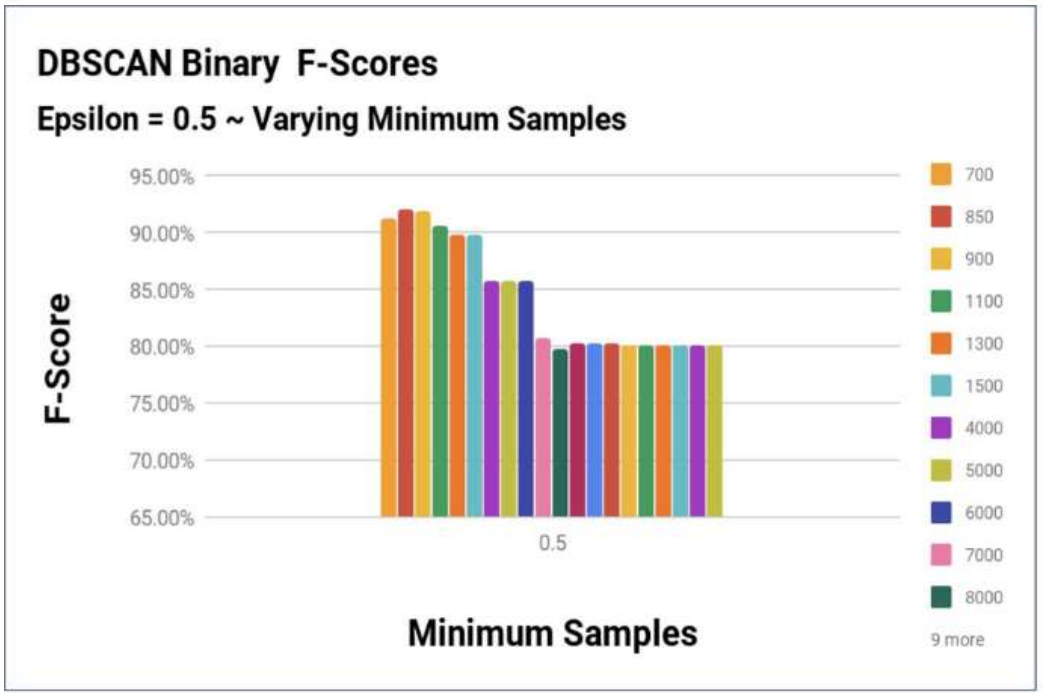
Because LOF(E) is significantly greater than 1, it suggests that E is an outlier relative to its neighbors. In comparison, the LOF scores for points A, B, C, and D would be closer to 1, indicating they are not outliers. This method effectively highlights how point E, with its high LOF score, deviates significantly from the pattern established by the other points, identifying it as an outlier in the dataset.

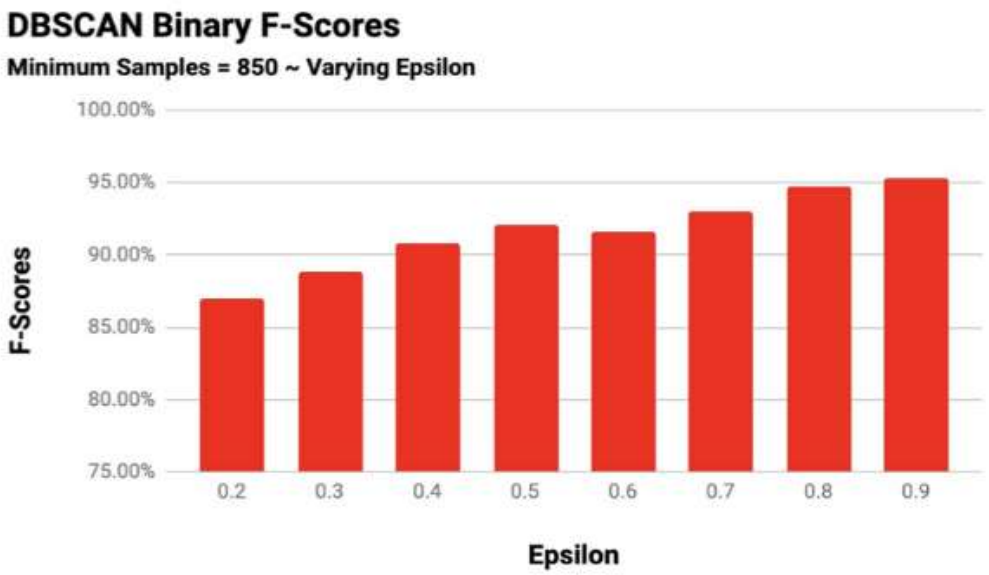
VI RESULT AND ANALYSIS

S.NO	ALGORITHM	BINARY F-SCORE	MULTICLASS F-SCORE	RUN TIME IN SECONDS
1	K-MEANS	95.63%	88.98%	0.0000188
2	DBSCAN	96.19%	92.09%	214.2
3	LOF	54.19%	53.46%	259.8
4	Isolation Forest	63.66%	55.79%	24.3




Finding Best "K": K-Means F-Score
Remove Catg. Data Binary
Remove Catg. Data Multiclass
Risk Value Data Binary
Risk Value Multiclass
F-Score
1.00
0.95
0.90
0.85
0.80
0.75
5
10
15
20
K-Means Parameter: K


DBSCAN Binary F-Scores
Epsilon = 0.5 ~ Varying Minimum Samples
F-Score
95.00%
90.00%
85.00%
80.00%
75.00%
70.00%
65.00%
0.5
Minimum Samples
700
850
900
1100
1300
1500
4000
5000
6000
7000
8000
9 more



VII CONCLUSION

This method builds a whole framework for discovering uncommon evidences of threats in the environment, by the means of the combination of the abilities to identify patterns of the Anomaly Detection algorithms, IF and LOF , with the data-separation capabilities of DBSCAN and K-Means, the Clustering algorithms. Extramission of traffic patterns baselines is done by K-Means function of feature and it clearly distinguishes the trend that is farthest to the baseline in DBSCAN. The IF algorithm in this situation is able to identify outliers fast and cheap before the system generally performs for higher quality. The local traffic variation and the network activity of the environment around are the variables, which are calculated using the LOF algorithm in order to detect the anomalies. It seems obvious that the algorithm having and using these algorithms offers a strong defensive system which could enhance network security by distinguishing and terminating potential threats.

Data set	Algo rithm	Para meter(s)	Binary F-Score	Multi class F-Score	Run Time (seconds)
NSL-KDD	K-Means	K = 8	95.63%	88.98%	0.0000188
NSL-KDD	DB SCAN	Esp = 0.8 ~ MS = 650	96.19%	92.09%	214.2

NSL-KDD	LOF	Contaminatio n = 0.3	54.19%	53.46%	259.8
NSL-KDD	ISO Forest	Contaminatio n = 0.25	63.66%	55.79%	24.3

VIII FUTURE SCOPE

The synergistic effect of the hybrid models and the deep learning techniques is really an exciting development in the cybersecurity field. However, the most significant impact of these approaches in cybersecurity is the improvement in ADS accuracy. The distinctive ability deep learning models have of recognizing even minute details of network behavior that may hint at a security attack is because they are built to deduce complex patterns from large masses of data. Utilizing the two approaches together with the normal machine learning methods gives rise to a hybrid system which overcomes the weaknesses of the latter and incorporates the advantages of both models. This mixture can serve as the sole factor determining the level of accuracy in the detection of anomalies with minimized possibilities of false positives and quick response to the real threats.

Hence, along with the distribution and utilization of cloud and IoT technologies, every day vulnerabilities of the field are broad, too. It is not a luxury or supplementary but indispensable only way for these technologies to ensure anomaly detection. Data collection algorithms can be configured to examine suspicious behavior that may imply a cybersecurity threat by analyzing information from the myriad of networked devices, as well as from cloud-based services. The timely detection is illustrating when anti-natural disaster actions possess the likelihood of being employed as the attack develops.

IX REFERENCES

[1]. Zhang, L., Zhang, W., & Jiang, H. (2021). A Hybrid Unsupervised Clustering-Based Anomaly Detection Method. IEEE.

[2] .Wang, Y., Guo, L., & Zheng, K. (2020). Clustering label estimation for network anomaly detection.Science

[3] .Lee, S., Kim, M.,& Park, J. (2021). Intelligent Anomaly Detection for Large Network Traffic With Optimized Deep Clustering Algorithm. IEEE.

[4] .Carlos Natalino. (2021). Spectrum Anomaly Detection for Optical Network Monitoring Using Deep Unsupervised Learning. IEEE.

[5] Md.Ahansul Kabir,Xiao Luo. (2020). Unsupervised Learning for Network Flow based Anomaly Detection in the Era of Deep Learning. IEEE.

[6] WILLIAM TESARDO LUARDI. (2023). ADVERSARIALLY REGULARIZED CONVOLUTIONAL AUTOENCODER FOR NETWORK ANOMALY DETECTION. IEEE.

[7] .YATING LIU, YUANTAO GU XINYUE SHEN, QINGMIN LIAO , QUAN YU. (2023). AN UNSUPERVISED ANOMALY DETECTION SYSTEM FOR NETWORK SECURITY IN BACKBONE NETWORK. IEEE.

[8]. DUC C. LE. (2021). ANOMALY DETECTION FOR INSIDER THREATS USING UNSUPERVISED ENSEMBLES. IEEE.

[9] .JIEWEN MAO,YOUNGQUAN HU,DONG JIANG,TONGQUAN WEI,FUKE SHEN. (2020). CBFS: A CLUSTERING-BASED FEATURE SELECTION MECHANISM FOR NETWORK ANOMALY DETECTION. IEEE

[10] .Lijuan Wang, Jun Shen, Fang Dong (2021). A Hybrid Unsupervised Clustering-Based Anomaly Detection Method.

[11] Riyaz Ahamed Ariyaluran Habeeb,Fariza Hanum Nasaruddin,Abdullah Gani Clustering-based real-time anomaly detection—A breakthrough in big data technologies.