Research Article

Implementation Of Data Mining And Maching Learing In The Concept Of Cybersecurity To Overcome Cyber Attack

T. Chandrakala[1], S. Nirmala Sugirtha Rajini[2], K.Selvam[3], K. Dharmarajan[4]

[1]Research Scholar,
[2]Professor, Department of Computer Applications, Dr. M.G.R. Educational & Research Institute, Chennai, Tamilnadu, India.
[3]Professor, Department of Computer Applications, Dr. M.G.R. Educational & Research Institute, Chennai, Tamilnadu, India.
[4]Associate Professor, Department of Information Technology, Vels Institute of Science Technology & Advanced Studies, Pallavaram, Chennai, Tamilnadu, India.

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 May 2021

Abstract: This paper examines a high potential routing protocol for cyber security that can be used in WSN based Internet of things applications with heavy traffic loads. DM was the most effective and evolving technologies for extracting previously unknown useful patterns and trends in order to improve an organization's efficiency. All companies are increasingly expanding with the help of data mining capabilities. Data mining aids in the discovery of previously discovered and highly profitable information in large amounts of data. The main goal of database knowledge exploration was to find novel trends in large set of data. It combines a variety of domains, including statics, artificial intelligence and machine learning. DM converts a large data collection into a logical structure and extracts significant information, allowing users to gain insight into the raw data obtained from different Internet of things applications. As a result, the Internet of Things was a network of physical things or object that is fixed with network communication, electronics and sensors to capture and share data. To start the global computation, the BS asks the Cluster heads to compute tasks like help and trust, and then the agent at every CH gives a request to its CMs to start the local computation. This reduces the size and no. of message sent from and to CMs and Cluster heads, as well as between Cluster heads and BS, reducing the amount of energy consumed and extending the network's lifetime.

Keywords: WSN, PDR, Energy Consumption, cyber security, machine learning, massive data, DM, CH, BS.

1. Introduction

The discovery of data "models" is known as data mining. Data dredging is a pejorative term for attempting to retrieve information from data that isn't sponsored by the data.[1] Today, data mining is more akin to machine learning, and most techniques rely on machine learning algorithms to find unusual events concealed within large volumes of data. People and objects are becoming increasingly interconnected as a result of recent developments in communication technology. Since the Internet is accessible, it is possible to link a variety of devices that can communicate and exchange data. Data mining with the IoT was a modern idea that enables users to link different sensors and smart devices to capture real-time data from the world.

The security threats have the most devastating impact on the Software Denied Networking which was the Distributed Denial of Service attack. Distributed Denial of Service can overwhelm the overflow switch or administrator when the network was not protected properly. There are a variety of documentations about the protection of Software Denied Network from Distributed Denial of Service attacks. IDSs are used in the network to aware the controller and detect the packets when a Distributed Denial of Service attack was detected. Machine learning was attracted many researchers to detect Distributed Denial of Service attacks. Therefore, defending the software denied network from problems was active in research area. This research aims to consider the suitable machine learning algorithm to find a Distributed Denial of Service attack in a network area.

DM was an extensive process that it may be proceed to all type of data, more common surveys on the subject can be found in [2], where experts looked at DM and machine learning techniques for analysing medical information. There are several surveys available on each of these techniques, as the ordering of DM techniques in this survey was depend upon the classification, mining pattern and clustering. A common mining pattern over a stream of data, for example, was given. [3] Presents a survey of clustering algorithms for WSNs. The clustering techniques investigated in those papers are solely concerned with network design and management, not with knowledge discovery. In, the author examines traditional classification strategies over data streams in a survey on classification methods over data streams.

Owing to the widespread accessibility of massive quantities of information and the impending require transforming the information into usable knowledge and information, DM has gotten a lot of consideration in the data company and society as an entire in present decades. Consumer retention, research market and fraud detection as well as the discovery of science and production management can all benefit from the information and expertise gained [4]. In the Internet of Things age, where all communicates and interacts with one another, a large amount of data was generated that should be mined properly and analyzed in order to improve IoT functionality. If we can effectively apply data mining methodologies to IoT, this combination will prove to be a game changer in any country's economy. For organizational decision-making, data mining techniques should be combined with IoT. As a result, the primary goal of this paper is to provide a comprehensive overview of a data mining system that has been examined for Internet of things applications. The structure outlined in this research can include a roadmap for investigators interested in using DM to solve IoT applications.

2. Literature survey

S. Haller et al[5] “A world where physical objects are seamlessly integrated into the information network, and where the physical objects can become active participants in business process. Services are available to interact with these ‘smart object’ over the Internet, query their state and any information associated with them, taking into account security and privacy issues.”

Jiang and Gruenwald,[6] examined an estimation of data technique known as CARM that refers the present rule of association between the radars in the recent window sliding. The method was depend on CFI-stream [45], a frequent CARM for the stream of data. To store closed item-sets, it uses a structure of memory data known as the DIU. If a new transmission was received, the tests of algorithm in the CARM and window sliding update the help for the CIS. Instead of generating all possible association rules if CRAM discovers any missing values in sensor readings, it produces the method which has close association with the present readings of sensor. CRAM produces approximate values based on these rules and selected closed item-sets, which include item values that were not included in the original readings. It shows that in the DIU tree, there are currently 4 closed item sets: CBA, BA, DC and C with corresponding supports of 3, 3, 1, and 2 in the right upper corner. These frequently occurring item-sets produce a simple set of rules. This simple rule set can be used to infer all other rules.

Tanbeer et al.[7] analysed a tree depend data framework called SP tree to produce a rule of association from wireless sensor networks data along with single scan database. An important concept was to have the bandwidth of all monitoring radars data and build a PT depend on canonical method and finally rearrange the tree in high bandwidth order. The tree of sensor pattern will retain the nodes of detecting radars at the high part of the tree by reorganisation, resulting in high solidity in the structure of tree. After the sensor pattern has been developed, the FP-growth mining technique is used to find the sensor sets that detect frequent events. Experiments are carried out to check that SP-tree outperforms PLT in terms of memory utilisation and runtime. A tree of sensor pattern outperforms PLT in terms of memory and time usage, according to the tests. The aim for this advantage is twofold: first, the PLT needs two database scans, while SP-tree only requires one database scan; 2nd the mining method of sensor pattern was highly potential because of the low bandwidth in the tree model.

Halatchev and Gruenwald [8] examined to find the missing sensor's readings and proposed a centralised approach called DSARM. It employs the rule of association i.e., algorithm for mining to classify radars which record the similar information for many times in a window sliding known as related radars, and then measures information from a radar by data related radars. Because of the radar data nature, it was impossible to apply an algorithm of mining like Apriori straight to sensor data. Because of this reason, the experts developed the DSARM system, which adjusts the Apriori algorithm to the data stream obtained from sensor nodes. Simulation tests on original data obtained by the Department of Transportation in Austin, TX, USA, are used to test this technique for estimating missing values in relevant data sources. To compare DSARM and other methods, performance assessments were carried out.

Umadevi et.al [16] Data mining is the foundation for behavioural analytics, which aims to avoid harm. A probabilistic and prognostic approach is a long-term benefit of machine learning. Machine learning and data mining approaches detect patterns, regularities, and anomalies, allowing for the prevention of cybersecurity breaches. Mercy Beulah et.al [21]The importance of implementing machine learning is defined by a number of factors, including the fact that recognising compromising behaviour takes time, and in many cases, users learn about hacking from a third party. It is essential to automate the review of security protocols and to detect attacks in real time. These facts add to the research's relevance and key goal .

3. Proposed systems

The implicit trust was present always in protocols depend upon cooperation. It involves in IoT routing processes with DM networks. If IoT networks expand, it becomes more danger to threats, necessitating an implementation of powerful protection process.[9] Due to the sensor nodes' limited energy, computing, and storage resources, finding appropriate cryptography for wireless sensor networks is a significant challenge.[10] Reliable minimum Deep Learning Trust Secure Attacker Detection is a new energy-aware routing algorithm that will be proposed for Adhoc networks.[11] DLTSAD addresses four main IoT requirements: energy consumption, reliability, data aggregation, and detection of attackers. DLTSAD was a powerful routing algorithm that identifies paths that use the least amount of total energy for E-E transversal packet while also improving hostile node finding. To apply Elliptic Curve Cryptography technique in IoT, we proposed a cryptography-based security mechanism. Improving the algorithm's decryption and encryption aspects, that provide and exist already for excellent stability.[12]

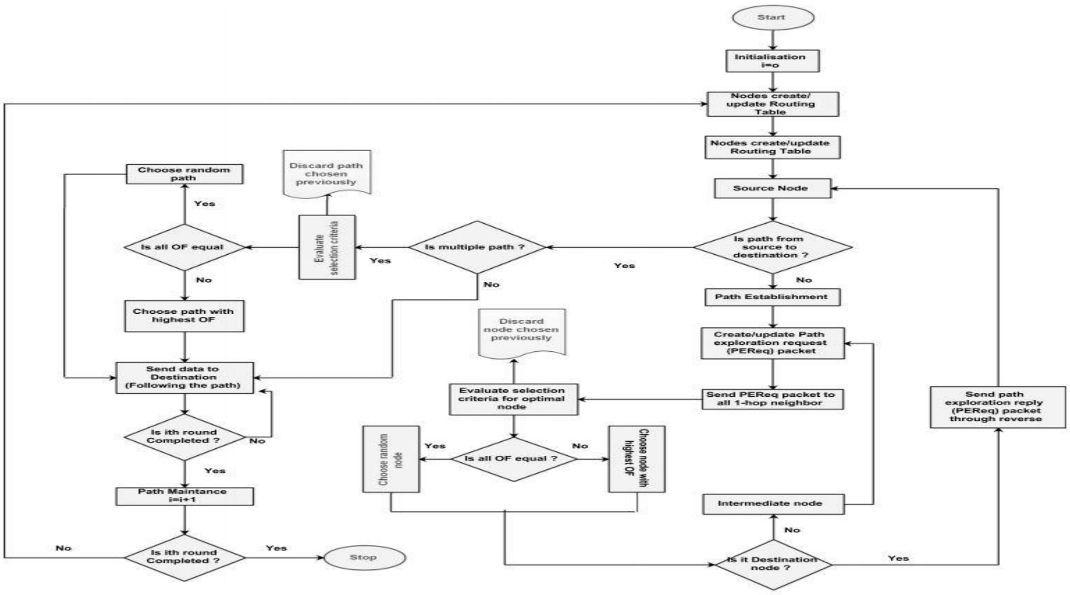

Start
Initialisation i=o
Nodes create/ update Routing Table
Nodes create/update Routing Table
Source Node
Is path from source to destination ?
Yes
No
Path Establishment
Create/update Path exploration request (PEReq) packet
Send PEReq packet to all 1-hop neighbor
Evaluate selection criteria for optimal node
Discard node chosen previously
Is all OF equal ?
Yes
Choose random node
No
Choose node with highest OF
Intermediate node
No
Is it Destination node ?
Yes
Send path exploration reply (PEReq) packet through reverse
Is multiple path ?
Yes
No
Evaluate selection criteria
Discard path chosen previously
Is all OF equal
Yes
Choose random path
No
Choose path with highest OF
Send data to Destination (Following the path)
Is ith round Completed ?
No
Yes
Path Maintance i=i+1
Is ith round Completed ?
No
Yes
Stop

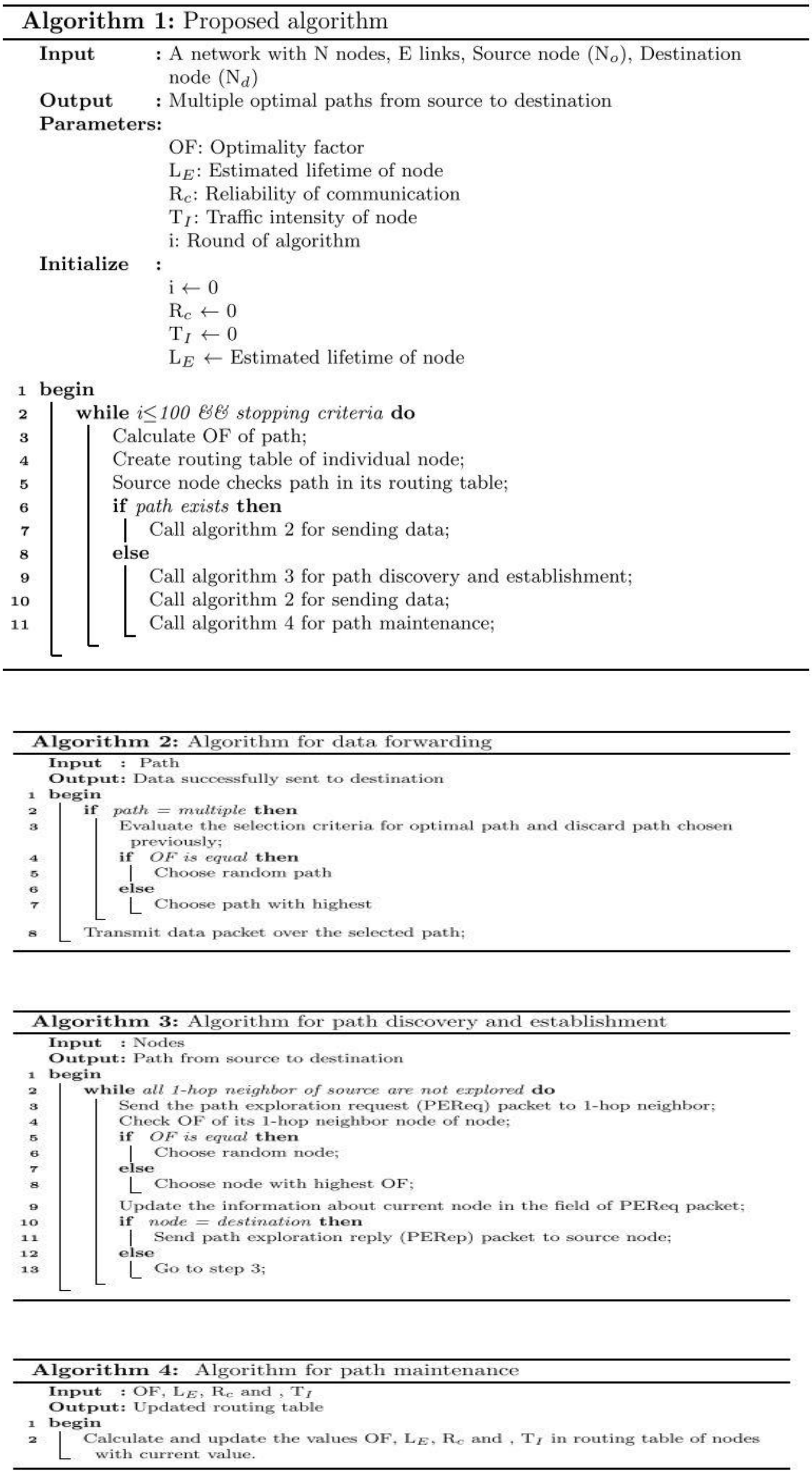
Figure 1: Classification of DM methods for sensor connection.

3.1. Proposed DLTSAD:

To find more secure routes, algorithms consider the reliability of connections. Algorithms determine the most powerful routes.[13] An Algorithm attempts to extend the lifetime of network by identifying paths that include nodes with high potential battery energy. Algorithm also determines with a greater number of protection options.

3.1.1 Algorithm of DLTSAD:

This module was created to propose DLTSAD [DATAGRAM LAYER TRANSPORT SECURITY ASSESSING DOG] focused on collaboration and routing, as well as attacker detection and prevention. In DLTSAD, we go over the strategies and contributions to trust-based protection. This shows how trust depend reasoning would permit each joint to measure the nodes actions and present a trust depend investigation of the DLTSAD protocol through particular trust language.[14]



```
Algorithm 1: Proposed algorithm
Input        : A network with N nodes, E links, Source node ($N_o$), Destination
               node ($N_d$)
Output       : Multiple optimal paths from source to destination
Parameters:
               OF: Optimality factor
               $L_E$: Estimated lifetime of node
               $R_c$: Reliability of communication
               $T_I$: Traffic intensity of node
               i: Round of algorithm
Initialize   :
               $i \leftarrow 0$
               $R_c \leftarrow 0$
               $T_I \leftarrow 0$
               $L_E \leftarrow$ Estimated lifetime of node
1  begin
2      while $i \leq 100$ && stopping criteria do
3          Calculate OF of path;
4          Create routing table of individual node;
5          Source node checks path in its routing table;
6          if path exists then
7              Call algorithm 2 for sending data;
8          else
9              Call algorithm 3 for path discovery and establishment;
10             Call algorithm 2 for sending data;
11             Call algorithm 4 for path maintenance;
```

```
Algorithm 2: Algorithm for data forwarding
Input  : Path
Output: Data successfully sent to destination
1  begin
2      if path = multiple then
3          Evaluate the selection criteria for optimal path and discard path chosen
             previously;
4          if OF is equal then
5              Choose random path
6          else
7              Choose path with highest
8      Transmit data packet over the selected path;
```

```
Algorithm 3: Algorithm for path discovery and establishment
Input  : Nodes
Output: Path from source to destination
1  begin
2      while all 1-hop neighbor of source are not explored do
3          Send the path exploration request (PEReq) packet to 1-hop neighbor;
4          Check OF of its 1-hop neighbor node of node;
5          if OF is equal then
6              Choose random node;
7          else
8              Choose node with highest OF;
9          Update the information about current node in the field of PEReq packet;
10         if node = destination then
11             Send path exploration reply (PERep) packet to source node;
12         else
13             Go to step 3;
```

```
Algorithm 4: Algorithm for path maintenance
Input  : OF, $L_E$, $R_c$ and , $T_I$
Output: Updated routing table
1  begin
2      Calculate and update the values OF, $L_E$, $R_c$ and , $T_I$ in routing table of nodes
         with current value.
```

Figure 2 algorithms of DLSAD

3.2 Framework of data mining for cyber security:

DM was the most effective and evolving technologies for extracting previously unknown useful patterns and trends in order to improve an organization's efficiency. All companies are increasingly expanding with the help of data mining capabilities.[15] Data mining aids in the discovery of previously discovered and highly profitable information in large amounts of data. For eg., by identifying consumers' regular purchasing habits, a company can boost revenue by grouping products that are often purchased together, offering discounts on certain items, or eliminating duplicate items.

The main goal of database knowledge exploration was to find novel trends in large set of data. It combines a variety of domains, including statics, artificial intelligence and machine learning. DM converts a large data collection into a logical structure and extracts significant information, allowing users to gain insight into the raw data obtained from different Internet of things applications.[17] As a result, the Internet of Things was a network of physical things or object that is fixed with network communication, electronics and sensors to capture and share data [18]. The ideal link of DM and Internet of things yields a new cutting-edge technology that will help people from all walks of life. These applications produce a massive amount of disparate data. Since data in Internet of things applications is continuously generated from various sources such as WSN, RFID, and so on.

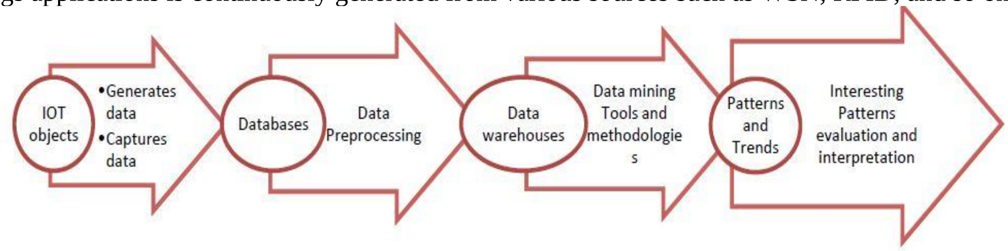


Figure 3: Data mining framework for IoT

3.3. DMA Rules from Wireless Sensor Networks:

A rule of associated could be referred as an allegation, such as A indicates B (A) B, where A and B denote the consequent and antecedent item sets, respectively. That is, database record of transmission containing items from item set A should also contain things from item set B. An appropriate global computation in the implemented method was defining the rule of association in D. [19]It was decomposed and distributed throughout the network, allowing calculation to be done locally and statistical summaries to be obtained and shared. To start the global computation, the BS asks the Cluster heads to compute tasks like help and trust, and then the agent at every CH gives a request to its CMs to start the local computation. This reduces the size and no. of message sent from and to CMs and Cluster heads, as well as between Cluster heads and BS, reducing the amount of energy consumed and extending the network's lifetime.

3.2.1 Trust Computation:

To use the calculated trust value to make a security decision, we must first estimate how much risk is appropriate for each ongoing mission. In other words, for each mission, a trust value threshold (Threshold) must be established. Depending on the security requirements of each ongoing mission, the threshold trust value may be changed. It is simple to determine whether the trustee node satisfies the trust requirement or node by comparing the calculated trust value to the threshold trust value.[20]

T^{d}_{ij}-Trust value I and j nodes, α-positive value, B-negative value

Trust value=20/20+8

=20/28

=0.7

$$\frac{\alpha_{ij}}{\alpha_{ii} + \beta_{ii}} \qquad (1)$$

e Computing Phase:

SCCP will be implemented by all CH that performs the below process;

1. Keeping enable item set and computing the item set of candidate at another stage from the random item sets at the earlier level.
2. Calculating the rate of assistance and trust.

The first part's implementation could be summarised as : an agent on each CH initiates the algorithm's application; this agent will perform the algorithm's major control tasks, such as finding and handling active and item sets of candidate, and communicating with agents on CMs to measure the 2 important measures of law of associations are trust and help. The calculation was decomposed and the agents manage the mechanism as it is replicated iteratively.[20]

3.4 Advantages:

1. Increasing the network's lifespan and achieving a substantial level of security.
2. It increases the performance of network and reduces the entire energy consumption. It also enhances the lifetime of network.
3. PDR and throughput ratio may be increased.
4. Decreased average E-E delay and overhead the routing messages.

3.5 Architecture / Input parameters:

These procedures have 3 parameters to model the optimal factors for selection. They are reliability, lifetime node and the probable traffic intensity.

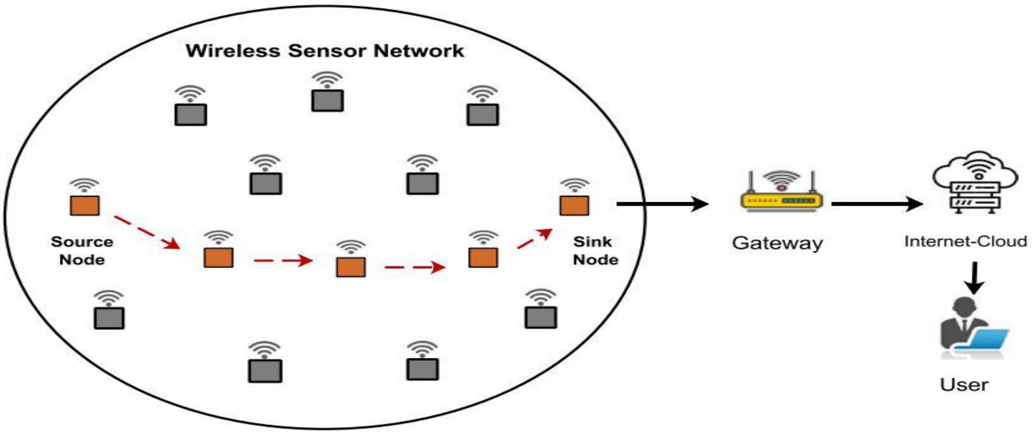


Fig. 4 Architect of WSN

3.6 Procedure Cyber Attack Model:

1. Key generation is a crucial step in which we must produce both a public and a private key.[22] The communication should be decoded by the private key of receiver and encoded by the PK of sender.
2. The PK was produced by the following formula.

   $Q = d * P$
3. ‘d ’ was the arbitrary no. which was chosen inside the range of 1 to n-1. P was the curve point.
4. d was the private key and Q was the PK (public key).

4. RESULTS AND DISCUSSION

4.1 Encryption / Decryption:

In encryption, the message must be represented on the curve. Encrypted document contains deep data implementation. Consider the point ‘M' on the curve ‘E' for ‘m.' Choose ‘k' at random from the list;

$$[1 – (n-1)] \quad (2)$$

Two cipher texts will be produced and it refers as C1 and C2.

$$C1 = k*P \quad (3)$$

$$C2 = M + k*Q \quad (4)$$

C1 and C2 will be sending.

Decryption means get back the message m which was given to customer.

$$M = C2 – d * C1 \quad (5)$$

M was the original message which was send to all.

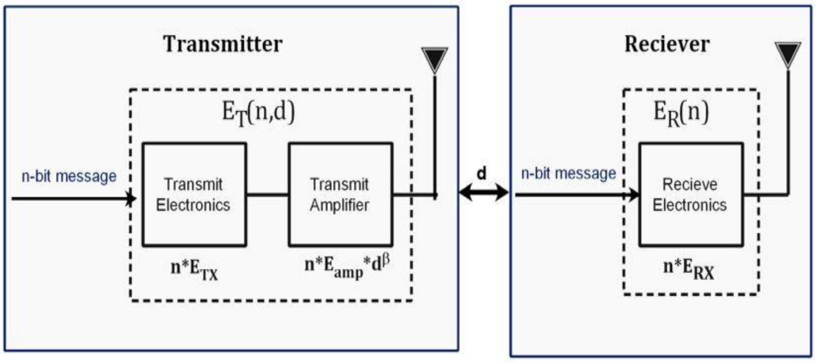


Fig. 6 Proposed protocols

4.2 *Throughput ratio*:

The number of successfully transmitted packets from source to destination per second in a wireless sensor network is known as throughput. The value of a well-designed network should be high, and if it is attacked, the value of throughput will drop dramatically.

Table 1 Simulation Results

Factors	Malicious node	Simulation Time	No. of Nodes	Protocol	Operating Platform	Packet	Simulator
Quantity	1	100000ms	10	AODV	Ubuntu	TCP	NS-2.3

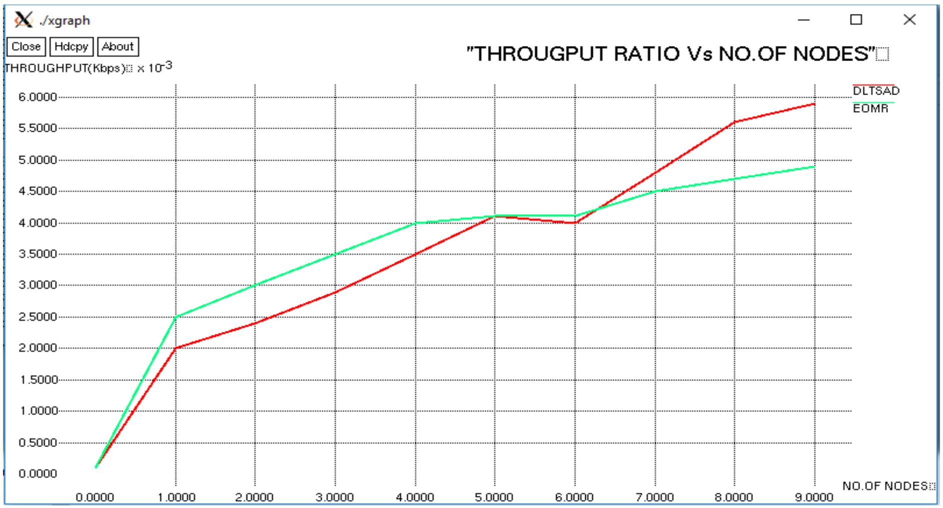


Fig. 7 Throughput ratio

4.3 *PDR:*

PDR refers to the proportion of total packets given to total packets sent from a source node to a destination node in a network. The maximum number of data packets should be sent to the target. As the PDR value rises, the network output rises with it. PDR is determined after a comparison without and with a black hole threat in the network. Packet delivery ratio was identified to be very low during the attack relative to the ratio before the attack, implying that fewer packets enter the sink node.

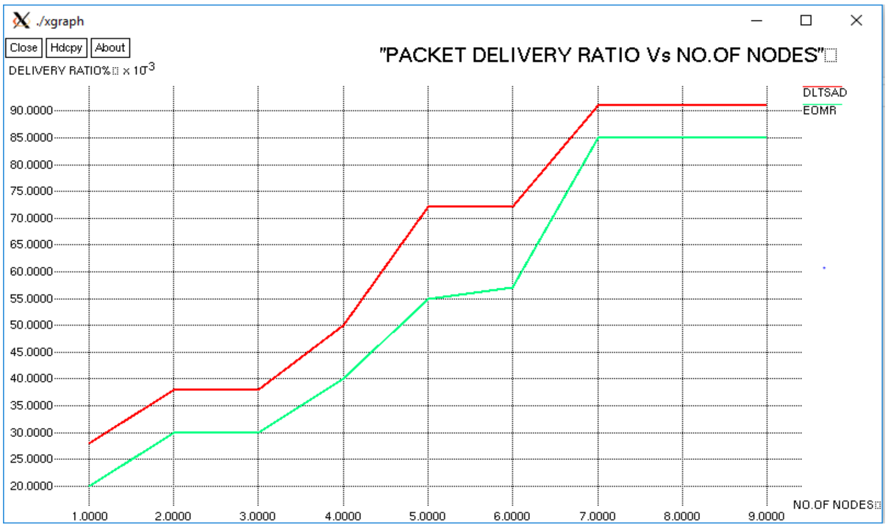


Figure 8 PDR

4.4 *Energy Consumption:*

Energy characterization is important for assessing the needs of an intensive data process which work effectively on mobile gadgets. An experimental analysis of the energy consumption of DM algorithms operating on mobile gadgets is presented in this paper.

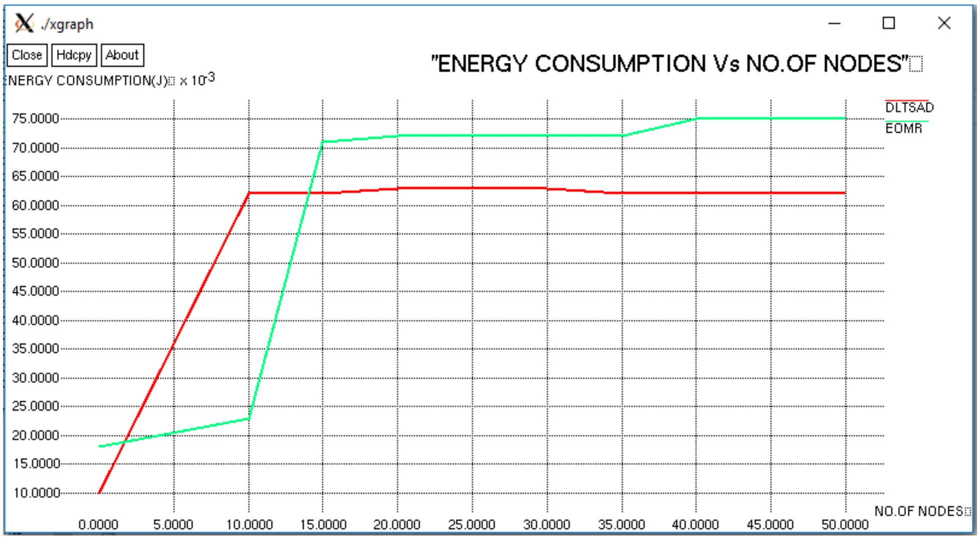


Figure 9 Energy Consumption

4.5 *E-E delay:*

The number of the delays encountered at a series of intermediate nodes on the way to the destination is the packet's end-to-end delay. Each delay is made up of two parts: a fixed transmission and propagation delay and a variable processing and queuing delay at the nodes.

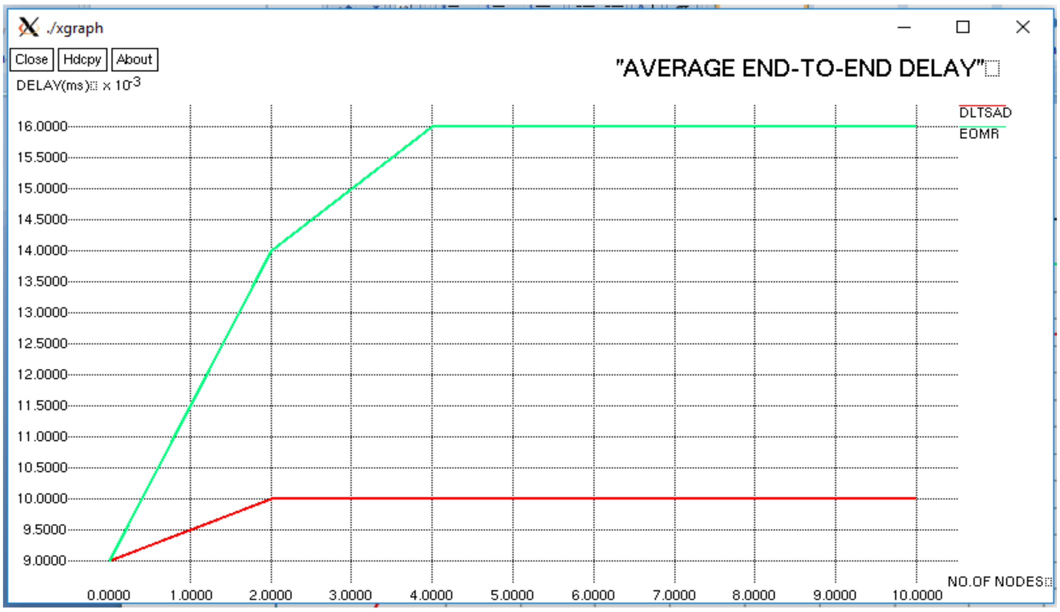


Fig. 10 E-E delay

V. OUTPUT

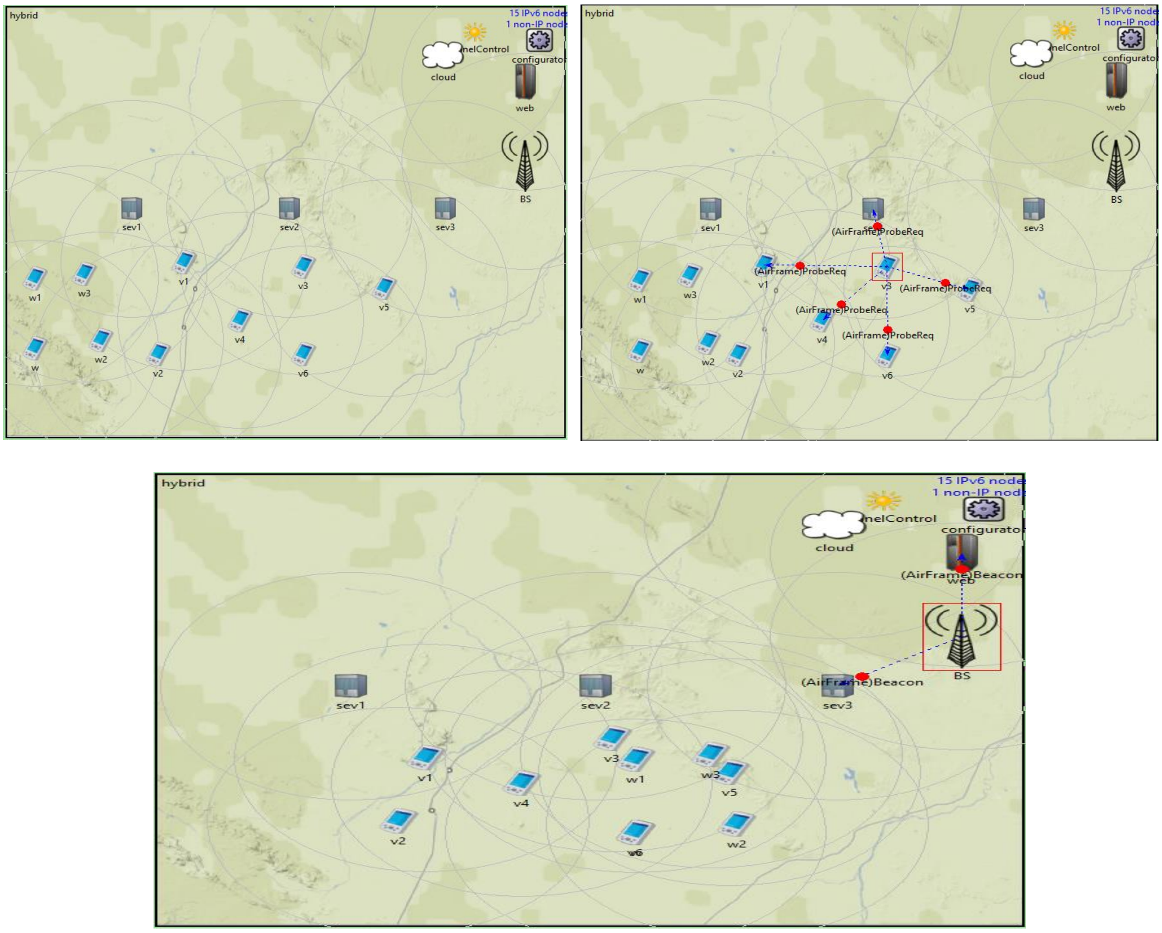


Fig. 11 outputs

5. Conclusion

The growing demand for DM techniques in the field of wireless sensor networks prompted the growth of a slew of algorithm. These algorithms addresses specific problems associated to the form and implementation of

WSNs. This paper examines a high potential routing protocol for cyber security that can be used in WSN based Internet of things applications with heavy traffic loads. DM was the most effective and evolving technologies for extracting previously unknown useful patterns and trends in order to improve an organization's efficiency. All companies are increasingly expanding with the help of data mining capabilities. Data mining aids in the discovery of previously discovered and highly profitable information in large amounts of data. The main goal of database knowledge exploration was to find novel trends in large set of data. It combines a variety of domains, including statics, artificial intelligence and machine learning. DM converts a large data collection into a logical structure and extracts significant information, allowing users to gain insight into the raw data obtained from different Internet of things applications.

Reference

1. A. Rozyyev, H. Hasbullah, and F. Subhan, "Indoor child tracking in wireless sensor network using fuzzy logic technique," Research Journal of Information Technology, vol. 3, no. 2, pp. 81– 92, 2011.
2. R. Szewczyk, E. Osterweil, J. Polastre, M. Hamilton, A. Mainwaring, and D. Estrin, "Habitat monitoring with sensor networks," Communications of the ACM, vol. 47, no. 6, pp. 34–40, 2004.
3. S. H. Chauhdary, A. K. Bashir, S. C. Shah, and M. S. Park, "EOATR: energy efficient object tracking by auto adjusting transmission range in wireless sensor network," Journal of Applied Sciences, vol. 9, no. 24, pp. 4247–4252, 2009.
4. P. K. Biswas and S. Phoha, "Self-organizing sensor networks forintegrated target surveillance," IEEE Transactions on Computers, vol. 55, no. 8, pp. 1033–1047, 2006.
5. L. T. Lee and C. W. Chen, "Synchronizing sensor networks with pulse coupled and cluster based approaches," Information Technology Journal, vol. 7, no. 5, pp. 737–745, 2008.
6. Jiang and Gruenwald, S. A. Aljunid, B. Ahmad, A. Yahya, R. Kamaruddin, andM. S. Salim, "Wireless sensor actor network based on fuzzy inference system for greenhouse climate control," Journal of Applied Sciences, vol. 11, no. 17, pp. 3104–3116, 2011.
7. Tanbeer, Gruenwald "Monitoring forest cover changes using remote sensing and GIS: a global prospective," Research Journal of Environmental Sciences, vol. 5, pp. 105–123, 2011.
8. Halatchev, B.Mukherjee, and D. Ghosal, "Wireless sensor network survey," ComputerNetworks, vol. 52,no. 12,pp. 2292–2330, 2008.
9. T. Arampatzis, J. Lygeros, and S. Manesis, "A survey of applications of wireless sensors and wireless sensor networks," in Proceedings of the 20th IEEE International Symposium on Intelligent Control (ISIC '05), pp. 719–724, June 2005.
10. Y.-C. Tseng, M.-S. Pan, and Y.-Y. Tsai, "Wireless sensor networks for emergency navigation," Computer, vol. 39, no. 7, pp. 55–62, 2006.
11. T. Yairi, Y. Kato, and K. Hori, "Fault detection by mining association rules fromhouse-keeping data," in Proceedings of the 6th International Symposium on Artificial Intelligence, Robotics and Automation in Space, pp. 18–21, 2001.
12. O. Horovitz, S. Krishnaswamy, and M. M. Gaber, "A fuzzy approach for interpretation of ubiquitous data streamclustering and its application in road safety," Intelligent Data Analysis, vol. 11, no. 1, pp. 89–108, 2007.
13. J. Gama, P. P. Rodrigues, and L. Lopes, "Clustering distributed sensor data streams using local processing and reduced communication," Intelligent Data Analysis, vol. 15, no. 1, pp. 3–28, 2011.
14. Z. A. Aghbari, I. Kamel, and T. Awad, "On clustering large number of data streams," Intelligent Data Analysis, vol. 16, no. 1, pp. 69–91, 2012.
15. A. Boukerche and S. Samarah, "An efficient data extraction mechanism for mining association rules from wireless sensor networks," in Proceedings of the IEEE International Conference on Communications (ICC '07), pp. 3936–3941, June 2007.
16. S.Umadevi, S. NirmalaSugirtha Rajini,A. Punitha & Viji Vinod, (2020), " Dimensionality Reduction in Machine Learning Technique using Principal Component Analysis", Test Engineering and Management, Jauary - February 2020 ISSN: 0193 - 4120 Page No. 14546 - 14552 .
17. S.Umadevi, S. Nirmala Sugirtha Rajini, A. Punitha & Viji Vinod(2020), "Performance Evaluation Of Machine Learning Algorithms In Dimensionality Reduction", International Journal of Advanced Science and Technology, Vol. 29, No. 9s, pp. 3845-3853
18. M.Anita Priscilla Mary , M.S.Josephine , V.Jeyabalaraja & S.Nirmala Sugirtha Rajini(2020), "Identification and Performance valuation for Effective Utilization of Electrical Energy Resource using K Means Clustering Algorithm", International Journal of Advanced Science and Technology, Vol. 29, No. 9s, (2020), pp.55-62.

19. S. Uma Devi & S. Nirmala Sugirtha Rajini (2019), " Detection of Traffic Violation Crime Using Data Mining Algorithms", Jour of Adv Research in Dynamical & Control Systems, Vol. 11, No.9, pp. 982-987.
20. Mercy Beulah, E,Nirmala Sugirtha Rajini,S & Rajkumar, N (2016), "Application Of Data Mining In Healthcare: A Survey", Asian Journal of Microbiology, Biotechnology & Environmental Sciences, vol.18, no. 4, pp. 999-1001, ISSN-0972-3005.
21. Mercy Beulah, E , Nirmala Sugirtha Rajini, S & Raj Kumar, N(2015)," Data Mining and Business Intelligence applications in Shipping Industry" , International journal of applied environment sciences(IJAES), vol. 10, no.1, pp.87-91, ISSN0973-6077 (Scopus Indexed).
22. Kiruthika , C & Nirmala Sugirtha Rajini,S(2014), "An Ill-identified Classification to Predict Cardiac Disease Using Data Clustering", International Journal of Data Mining Techniques and Applications, vol. 03,pp. 321-324, ISSN:2278-2419.