IMPROVE FAULT DETECTION AND SENSOR ROUTING PROBLEMS WITH NAIVE BAYES CLASSIFIER AND DATA AGGREGATION APPROACH

Abstract

Wireless sensor networks are regularly located in hostile environments to observe changes in environmental factors. Broadening the importance of machine learning to WSN creates wonders and provides credibility to the system. Clustering in wireless sensor networks is a better technique to enhance energy utilization. This chapter uses an ensemble method for clustering and classification to develop machine-learning algorithms for fault detection in wireless sensor networks. We will propose a new routing protocol, Fault tolerance Backup cluster head, and Data Aggregation with Naïve Bayes Classifier to improve fault detection and sensor routing problems. We develop a statistical approach to detect and identify faults in a WSN. The analysis of our model is an accurate and effective fault management framework with minimum energy consumption, delay, and overhead.

Keywords: Fault tolerance, cluster head, Data Aggregation, Naïve Bayes Classifier.

Authors

Dr. Pogula Sreedevi
Associate Professor
Department of CSE
RGM College of Engineering and Technology
Nandyal, Andra Pradesh, India.

Dr. T. Santhi Sri
Professor
Department of CSE
Koneru Lakshmaiah Education Foundation
Andra Pradesh, India.

Dr. E. Poornima
Associate Professor
Department of AIMLE
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India.

Dr. Farooq Sunar Mahammad
Professor
Department of Computer Science and Engineering
Santhiram Engineering College Nandyal
Andra Pradesh, India.

I. INTRODUCTION

The fault tolerance Backup cluster head and Data Aggregation (FTBKCH-DA) algorithm are presented in this chapter. This chapter is organized as follows. Section 1 presents the preamble. Section 2 discusses the various phases of the FTBKCH-DA algorithm in the literature survey. Section 3 illustrates how Machine Learning has been used in Cluster-Head selection and data aggregation techniques. Section 4 studies the performance of the proposed algorithm. Section 1.5 summarizes the chapter.

1. **Preamble:** In other sectors, like data science, where there is a vast amount of data to be analyzed, WSN plays a significant role. The authors employed machine learning methods to categorize this enormous amount of data from several sensors. The predictions from each sensor are then combined using ensemble learning methods. Combining machine learning prediction results reduces the issue of choosing the incorrect models.

   Clustering has a wide range of secondary advantages and associated goals, in addition to promoting network scalability and lowering energy usage through data aggregation. The routing table size kept at each node can be decreased by determining the cluster routing configuration [1].

   Data loss, aggregate mistakes, and calibration faults can cause problems in WSNs. Many people utilize machine learning to find problems with WSNs. The research community has experienced various defects and can be divided into groups based on the felt data [2].

   The suggested technique offers fault tolerance in the two scenarios listed below. 1. Tolerance for faults when CH fails. 2. Tolerance for faults when the route fails. When the CH malfunctions, the network's data transmission is disrupted, and there are significant odds that the data will be lost before reaching its predicted destination. There must be a solution to deal with this since it will cause significant issues when processing critical data on time. Additionally, sending packets through the incorrect relay nodes might be hampered data transmission. Using backup CHs and effective relay node selection, our suggested fault tolerance strategy resolves these problems [3].

   Additionally, sending packets through the incorrect relay nodes might be hampered data transmission. Using backup CHs and effective relay node selection, our suggested fault tolerance strategy resolves these problems [4]. A particular node is chosen and referred to in a network as the cluster head (CH) during the cluster set-up phase, as several small groupings of nodes are partitioned into clusters.

   The reconfiguration of clusters is done during the cluster maintenance phase. Data acquired by CH is pooled from cluster members. The routing method of WSNs based on a network structure involves a variety of categories, including location-based routing, hierarchical routing, and flat routing. Compared to flat routing protocols, hierarchical routing approaches offer apparent communication effectiveness and scalability advantages [5].

Sensing data must be gathered from sensing nodes and transmitted using the sensing network to the BS.

The following are available on the network:

- All CMs can send data directly to the CH because all the nodes are grouped into clusters, each with a CH. Over some hops, the CH sends the data to the BS.
- The Backup CHs are selected by estimating the centrality among the nodes. As part of intra-cluster communication, Aggregator (AG) nodes are deployed for each cluster. These AG nodes can reduce the aggregation overhead in CHs.

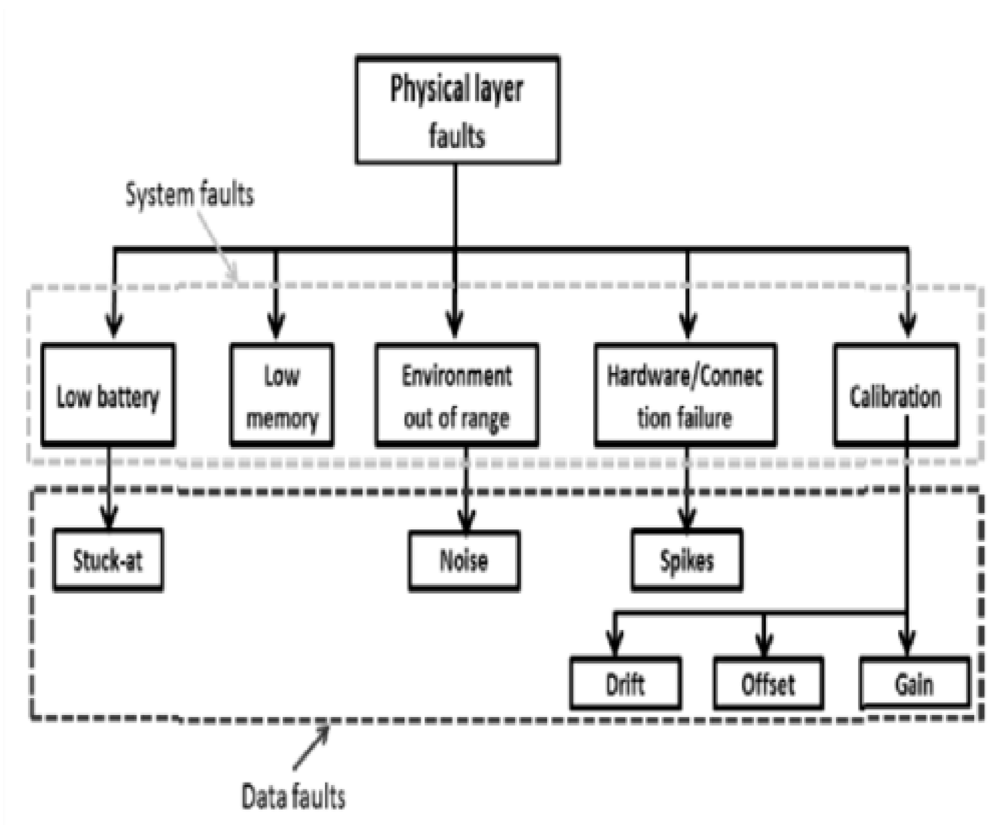

Figure 1: Physical Layer Faults in Wireless Sensor Networks

2. Objectives

- The chapter proposes a fault-tolerant approach for the failure of nodes in the network.
- To improve fault detection, we have implemented the machine learning algorithm, i.e., the Naïve Bayes classification algorithm in Wireless Sensor Networks.
- Based on K-means clustering, clustering algorithms are applied to generate the classifier model.
- The classification model is generated using the Naïve Bayes algorithm.

II. LITERATURE SURVEY

A wireless sensor network (WSN) is a collection of spatially dispersed autonomous sensors that track environmental or physical parameters like temperature, pressure, sound, etc., for data transfer to a destination across a network. The bidirectional nature of current networks enables the capability of sensors to control activities. WSNs have been expanded to include military, industrial, and commercial services and the monitoring and management of machine health [8,9].

The WSNs have a large number of interconnected sensor nodes. Each sensor network comprises various components, including a radio transceiver connected to an external or internal antenna, an electronic circuit containing sensor nodes, an energy supply, and a microprocessor [6].

The placement and arrangement of significant sensor nodes determine the operating power of the WSNs. Monitoring remote environments for patterns in lower frequency data is the most significant technological use for WSNs. When compared to conventional wired systems, the deployment costs would be lower. Every sensing point needs to have a quarter-sized device installed by the installers rather than a variety of wire connections routed through a protective conduit. The main benefit of using WSNs is their capacity for adaptation in dynamic contexts and their cheaper installation costs. Network topologies can be changed via the adaption mechanisms or the network's ability to switch between different operating modes. The main goal of WSNs is to report occurrences of a preset nature or to transmit identifiable data to the BS or sink nodes [7]. Each sensor node serves as a data originator or router in a network. The sensor nodes use a multi-hop or single-hop scheme to send the sensed data to the sink. Because of the limitations of the network, implementing routing protocols is a critical difficulty for WSNs.

Numerous routing strategies are applied to extend the lifetime of sensor networks and establish energy-efficient paths. However, the power of the sensor nodes runs out over effective routes. In sensor networks, some target zones will prevent the observation of events from occurring. The majority of WSN applications are found in the civil and military sectors. The sensor nodes could serve as data sources for detecting and gathering environmental data samples. The restrictions on energy supply, computational power, and wireless link bandwidth are WSNs' drawbacks. Effective data processing based on adopting aggressive energy management techniques is a crucial goal of WSNs [8].

For an extensive and dynamic wireless network, scalability is a problem in a flat network structure. The hierarchy must be considered if WSNs are utilized in applications with higher node density. Utilizing a cluster structure and adding hierarchy to a network increases system capacity and facilitates efficient route management and discovery [9].

1. **Clustering**: The establishment of clusters is crucial to the routing process. Homogeneous or heterogeneous SNs may form the cluster. The reason for clustering is that a direct connection with BS is inappropriate when considering effective energy management because SNs have limited energy resources. Each cluster is typically run by a dominating CH node, which also manages the cluster members (CMs) and does

2. **Routing**: Clusters of WSN use routing as a recommended practice to select the suitable CH and the best communication path, conserving SNs' energy. The various clusters connect via a single-hop or multi-hop during the routing procedure. Without a doubt, direct communication is quicker and more effective than multi-hop communication, but if the deployment area is larger, the effectiveness of direct communication is decreased. Hence, the network must be routed in either an inter-cluster or an intra-cluster environment to maximize the effectiveness of multi-hop communication.

3. **Fault Occurrence**: A system failure that may be verified occurs when the actual service deviates from the anticipated service. Hardware malfunction and software error are the two main reasons why faults happen. Communication faults, link faults, node faults, energy depletion-based faults, physical damage-based faults, intruder attack-based faults, and other fault types are all potential in WSN. There are two ways to fix the WSN flaw. The first is a fault tolerance mechanism in which the required healing procedure must be adopted following the onset of the defect. The second is fault prevention methods, which stop environments where faults arise. The following section describes the proposed methodologies' assumed constraints, network, communication, and fault models.

Constraints

- The nature of SNs is mobility.
- The type of sensor network is homogeneous.
- With a unique SNid, SNs are explicitly identified.
- With all resources, the SSNs are limitless.
- The network is first aware of the SSN deployment topology.
- The V_{pa} is fault-free.

The General Self-Organization Tree-Based Energy Balance (GSTEB) protocol is recommended by authors M. Rangchi and H. Bakhshi as a way to offer load balancing. It is a dynamic tree-based routing system that uses the least amount of energy possible, has little to no packet loss, and also offers data compression to boost performance. But the authors not concentrated at the sink position [10].

In severe conditions or isolated locations where other energy sources are impractical, this offers a possible alternative energy source for powering sensors. On the same circuit board, the dual-band antenna and RF energy harvesting system are simply integrable. To create an antenna for a wideband energy harvesting environment, the difficulty is to increase the bandwidth. There is no information about time division access provided by the authors [11]. In this paper, ATPC, a compact Adaptive Transmission Power Control technique for wireless sensor networks, is introduced. For each of its neighbors, each node in ATPC creates a model that describes the relationship between transmission power and connection quality. In order to dynamically maintain individual link quality over time with this approach, we use a feedback-based transmission power control technique [12].

In this research, a methodical approach to building a collaborative design workflow based on the SOA reference model was provided. The most important step is to pinpoint the services in the model, and this work suggests and implements a number of approaches to produce suitable services for collaborative design. These services could satisfy business requirements, they were in line with information technology, and they could either be mapped to related service components or built at a later point [13,14].

The authors R. Pierdicca, D. Liciotti, et al. provide a data evaluation to identify the ideal sensor configuration based on user testing. The findings of the article show that, in addition to significantly enhancing human contact, embedded localization systems may also be a valuable source for data collection. The goal of this study is to assist retailers and insiders with a variety of tasks, including product development or enhancement, segmentation tactics, and analysis of customer behavior in stores where integrated computer enhances the experience [15].

The authors of this research have used an ensemble method of clustering and classification to deploy machine-learning techniques to detect congestion in wireless sensor networks. The classifier model is created using Expectation- Maximization clustering and Naive Bayes classification based on K-means. The performance of the ensemble approach classifier and the Naive Bayes method alone is used to create the classification model [16].

The convolution neural network (CNN) and Naive Bayes classifier were suggested by the authors to enhance convergence performance and detect node problems. In order to locate and arrange the defects, they next examine convex hull, Naive Bayes, and CNN methods using real-world datasets [17]. The existing paper discusses the classification of documents using supervised learning, with a focus on the concept-based algorithm, and also deals with the hidden patterns in the documents using unsupervised clustering technique and topic-based modeling for the analysis and improvement of systematic arrangement of documents by applying k-means and LDA algorithm [18].

This study also provides an overview of a number of studies that used deep learning and machine learning approaches in a variety of research fields, such as networking, communications, and lossy environments. This survey work's primary goals are to identify potential problems and difficult applications for using various deep learning and machine learning algorithms and strategies in wireless networks, as well as to determine a suitable research path towards the development of a system that can recognize, foretell, and recover from abnormal situations on wireless networks [19].

III. SYSTEM DESIGN

Statistical failure analysis and machine learning methods are used to analyze failure patterns. A machine classifier can indicate the (instantaneous) failure hazard if features have been amassed over time and faults have been identified. However, depending on the gathered sensory data, reliability analysis approaches provide a variety of ways to forecast failure rates and probabilities.ML techniques improve the operation of node clustering and data aggregation as follows:

- Utilize machine learning to extract similarity and dissimilarity (for example, from malfunctioning nodes) in various sensor readings to compress data locally at cluster heads.

- To efficiently choose the cluster head, machine learning methods are used. Appropriate cluster head selection will greatly minimize energy consumption and lengthen the network's lifespan.

1. **Network Model:** There are numerous clusters in the network. The number of clusters in the network is given by N = c_1, c_2, c_3, c_4, c_S. Each cluster comprises the set of nodes C = n_1, n_2, n_3,... n_m, where m is the cluster's total number of nodes. A cluster head, CHC, is kept up by each cluster. The system's nodes are made up of varying amounts of electricity. The BS has positioned a way away from the network. The BS gathers information from the network. The CHs transmit to the BS the data they have gathered from each node n_i, I = 1,2,3,4...m in the cluster c_j, j = 1,2,3,4,5...s. CH and no other node can only send the data to BS.

2. **Task Model:** DAG can be used to express the task model. One CH and a few chosen Vice-Cluster-Heads are in the cluster (VCH). The nodes that are close to the CH are known as the VCHs. The CH distributes the task; the CH separates the task (T) into components (t_1, t_2, t_3,... t_r) and distributes the components (t_i, I = 1,..., r) to VCHs, which then assign the components (t_i, I = 1,..., r) to other nodes, present in the lower levels. The task is distributed so that the nodes at level I each assign themselves a portion of the task, except the node at level 1, and forward the remaining portion to the nodes at level I1. The nodes at level I1 then handle a portion of the task themselves and forward the remaining portion to the nodes at level I2, and so on.

3. **Algorithm**

```
##
For all the nodes n ∈ N
    Divide the network as 'k' clusters
End
    For each node 'n' in cluster 'k' where n ∈ k ∈ N
            CH selection
                    Estimate D_(N−BS)(n)
                    Estimate RE(n)
        Estimate LD(n)
            Calculate CH_wght(n)
                    If CH_wght(n) < CH_wght(n + 1)
                            CH = n
    End for
    BKCH selection
            For each node 'm' in cluster 'k' where m ∈ k & is not CH
                    Estimate centroid (m)
                    If centroid (m) < centroid (m + 1)
                    BKCH = m
            End for
    End BKCH selection
Data transmission phase DCPSO
```

Initialize particles $P_i \quad 1 \leq i \leq N_P$
$For\ i = 1\ to\ N_P$ do
Estimate RE, NL, E_{con}
Calculate fitness $F(P_i)$
$pbest = P_i$
If $F(P_i) < F(Pbest_i)$
$Pbest_i = P_i$
If $F(Pbest_i) < Gbest$
$Gbest = F(P_i)$
End
Compute relay nodes using $Gbest$
End for

4. **Performance of the Algorithm:** For further escalation, the data needs to be combined and transmitted on time to the intended recipients. As a result of frequent retransmissions caused by the existing protocols' lack of fault tolerance and poor forwarder node selection, there is a significant delay between the sensor nodes. The suggested protocol minimizes unneeded elements like excessive data traffic and maximum hops to the destination, which impact data transmission between nodes. The outcome demonstrates that the proposed procedure has a shorter delay in the allotted period than earlier proposed protocols.

   The sensor nodes require energy to operate continuously. Limited energy networks are sensor networks. A longer lifespan is produced via energy use that is optimized. The existing protocols do not provide adequate fault tolerant data aggregation methods that lower communication overhead, which raises the ratio of overall energy consumption. By applying efficient data aggregation techniques, the fault tolerance mechanism in our suggested solution reduces issues like retransmission and the erroneous path and enhances overall energy consumption. The outcome demonstrates that the energy consumption rate is significantly lower than that of traditional techniques.

5. **Naive Bayes Classifier:** Finding sensor values that are odd or aberrant that differ noticeably from the predicted behavior. Fault detection is a critical duty in wireless sensor networks to find broken or damaged sensors that could offer erroneous or unreliable data. By utilizing the probabilistic model to categorize sensor data as normal or defective based on observed attributes, naive Bayes classifiers can be utilized for fault detection.

   The Bayes classifier, which is based on the Bayes theorem, uses conditional probability to solve classification problems.

$$P(Y/X) = P(X/Y)\ P(Y) / P(X)$$

   This is an explanation of the basic Bayes theorem for classification tasks: Predicting the class attribute Y requires treating each attribute, including the class attribute, as a record of the attributes ($X_1, X_2, \ldots, X_d$). The objective of classification is, in particular, to find the value of Y that maximizes $P(Y \mid X_1, X_2, \ldots, X_d)$. Calculating posterior probability is necessary.

Hence, we reach to the result:

$$P(y|x_1,...,x_n) = \{ P(x_1|y)P(x_2|y)...P(x_n|y)P(y)\}/\{P(x_1)P(x_2)...P(x_n)\}$$

Which can be expressed as:

$$P(y|x_1,...,x_n) = \{P(y)\prod_{i=1}^{n}P(x_i|y)\}/\{P(x_1)P(x_2)...P(x_n)\}$$

Now, as the denominator remains constant for a given input, we can remove that term:

$$P(y|x_1,...,x_n)\propto P(y)\prod_{i=1}^{n}P(x_i|y)$$

The cluster parameters table shown below:

Table 1: Cluster Parameters

Parameter	**Value**
Clustering Algorithm	K-Means
Cluster Type	Unequal
Number of Clusters	4-8
Node Deployment	Random way Point

IV. RESULTS AND DISCUSSION

The sensor nodes require energy to operate continuously. Limited energy networks are sensor networks. A longer lifespan is produced via energy use that is optimized. The existing protocols do not provide adequate fault-tolerant data aggregation methods that lower communication overhead, which raises the ratio of overall energy consumption.

By applying efficient data aggregation techniques, the fault tolerance mechanism in our suggested solution reduces issues like retransmission and the erroneous path and enhances overall energy consumption. The outcome demonstrates that the energy consumption rate is significantly lower than that of traditional techniques.

We have compared our proposed method with several existing methods. K-Nearest Neighbor(E-KNN), Fuzzy Optimal Clustering Multiobjetive Parameter (FOC-MOP), Data Aggregation, Energy, Clustering and Scalable Algorithm (DECSA), Adaptive Transmission Power Control (ATPC), Fault Tolerant Optimal Relay with Naïve Bayes (FTOR-NB) algorithms are implemented in terms of Energy Consumption, Overhead and Delay. Figure 2. shows the graphical representation of Energy Consumption.

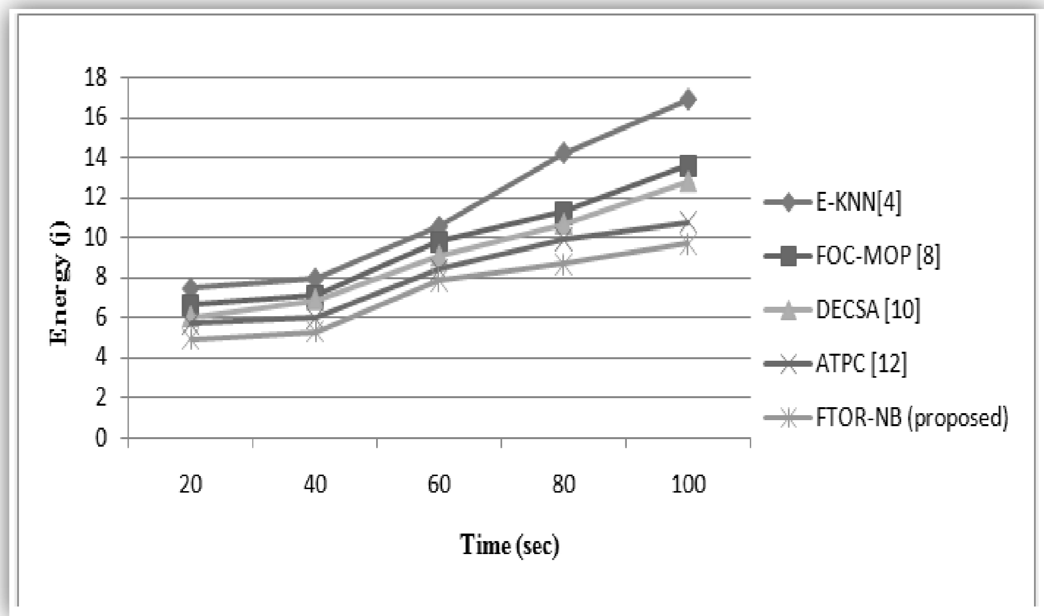

Figure 2: Graphical Representation of Energy Consumption

The number of control packets necessary for the network to function determines the overhead of the network, as shown in Figure 3. The number of retransmissions in the previously suggested operations is fairly high as a result of connection failure and inappropriate relay node selection.

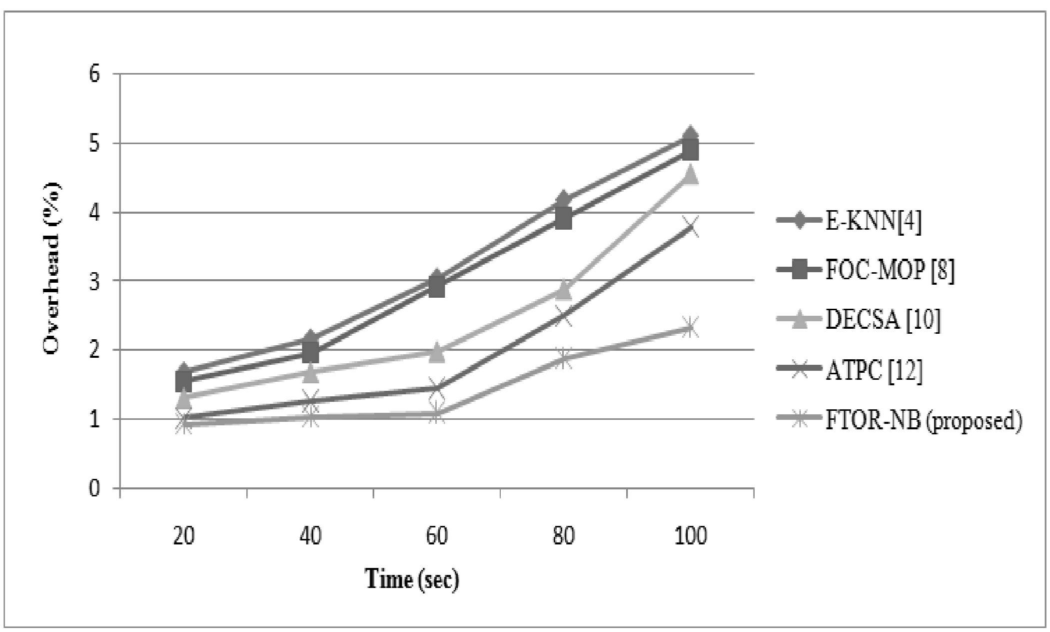

Figure 3: Graphical Representation of Overhead

The suggested protocol minimizes unneeded elements like excessive data traffic and maximum hops to the destination, which have an impact on the transmission of data between nodes. The outcome demonstrates that the proposed procedure has a shorter delay in the allotted period than earlier proposed protocols. The end-to-end delay of the entire network is shown in Figure 4.

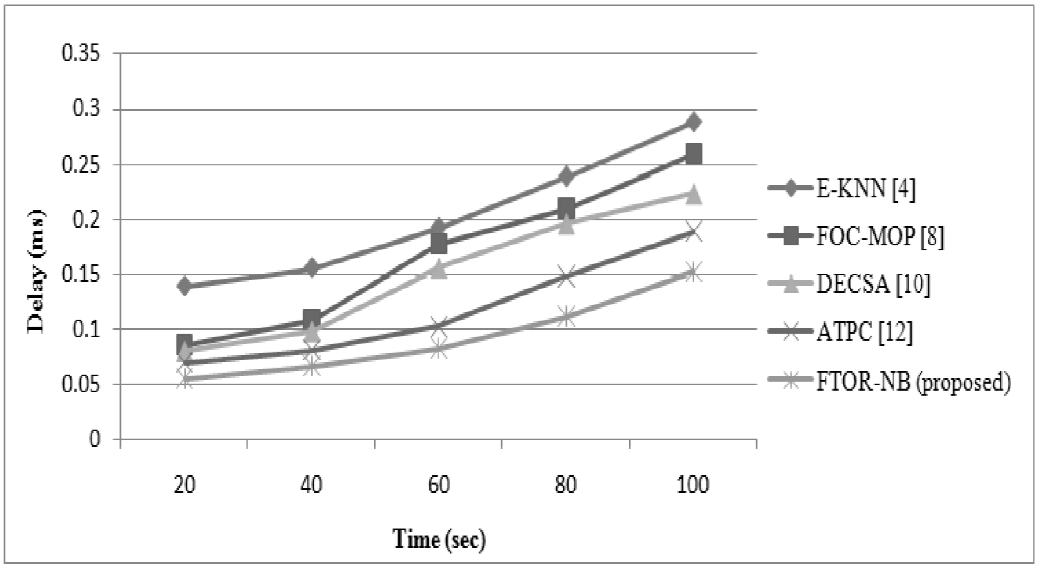

Figure 4: Graphical Representation of Delay

V. CONCLUSION AND FUTURE ENHANCEMENT

A dynamic cluster-based routing protocol method is frequently utilized in wireless sensor networks. Due to the energy depletion of the cluster heads, this strategy commonly uses the cluster reelection process. The energy of the entire sensor network is depleted due to the recurrent cluster head reelection since more advertisements are sent out. This approach is also tested on a heterogeneous wireless sensor network using the multi-objective optimization issue of cluster head selection. To reduce the fitness function, it is best to maximize the sensor node's residual energy and the distance travelled. The performance of the model created to detect the faults, is improved by combining both the Nave Bayes classifier model and the two subsequent approaches. It has been found that using clustering as a pre-processing step helps to minimize the effort required for classification and ensures that labels for comparable items are assigned correctly.

Further studies could look at the merits of alternative clustering methods, the effectiveness with which both maps solve issues in other application areas, as well as the particularities of the challenges encountered there.

REFERENCES

[1] Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications. Available from: https://www.researchgate.net/publication/262452270_Machine_Learning_in_Wireless_Sensor_Networks_Algorithms_Strategies_and_Applications.
[2] Fault-Tolerant Mechanism for Wireless Sensor Network. Available from: https://www.researchgate.net/publication/336086391_FaultTolerant_Mechanism_for_Wireless_Sensor_Network.
Optimal Routing and Clustering Technique for Wireless Sensor Networks. Available from: https://www.researchgate.net/publication/335964886_Optimal_Routing_and_Clustering_Technique_for_Wireless_Sensor_Networks.
[3] Atia Javaid et.al. Machine Learning Algorithms and Fault Detection for Improved Belief Function Based Decision Fusion in Wireless Sensor Networks. Sensors 2019,19, 1334; doi:10.3390/s19061334 www.mdpi.com/journal/sensors
[4] Ritik Argawal; Dattatraya Kalel; M. Harshit; Arun D Domnic; R. Raja Singh, Sensor Fault Detection using Machine Learning Technique for Automobile Drive Applications, 2021 National Power Electronics Conference (NPEC)
[5] Priya, P.I., Muthurajkumar, S. & Daisy, S.S. Data Fault Detection in Wireless Sensor Networks Using Machine Learning Techniques. Wireless Pers Commun 122, 2441–2462 (2022). https://doi.org/10.1007/s11277-021-09001-1
[6] P. Sreedevi, S. Venkateswarlu, An efficient intra-cluster data aggregation and finding the best sink location in WSN using EEC-MA-PSOGA approach, Int. J. Commun. Syst. (2022) e5110. [43]
[7] P. Sreedevi, S. Venkateswarlu, FOC-MOP: Fuzzy optimal clustering based multi-objective parameter route selection for energy efficiency, Wirel. Pers. Commun. (2022) http://dx.doi.org/10.1007/s11277-022-09769-w.
[8] Pogula Sreedevi, Dr. S. Venkateswarlu, A fault tolerant optimal relay node selection algorithm for Wireless Sensor Networks using modified PSO, Pervasive and Mobile Computing
[9] Volume 85, September 2022, 101642
[10] M. Rangchi and H. Bakhshi, "A new energy efficient routing algorithm based on load balancing for wireless sensor networks", International Symposium on Telecommunications (IST'2014), Tehran, pp. 1201-1205, 2014.
[11] Bakkali, J. Pelegri, T. Sogorb, V. Llario and A. Bou-Escriva, "A dual-band antenna for RF energy harvesting systems in wireless sensor networks", Journal of Sensors, pp. 1-8, 2016.
[12] Shan Lin, Fei Miao, Jingbin Zhang,Gang Zhou, Lin Gu,Tian, John A. Stankovic and George J, "ATPC: Adaptive Transmission Power Control for Wireless Sensor Networks", ACM Transactions on Sensor Networks, vol.12, issue 1, pp. 1-8, March 2016.
[13] Wei-Jung Shiang, Hsin Rau and Yu-Hsin Lin, "Service identification of a collaborative design workflow in a dynamically changing environment", Proceedings of the International Conference on Networking, Sensing and Control, pp. 685-690, 2009.
[14] Yang Xiao and Yi Pan, "Capacity and rate adaptation in IEEE 802.11 Wireless LANs", Emerging Wireless LANs, Wireless PANs and Wireless MANs: IEEE 802.11, IEEE 802.15, 802.16 and Wireless Standard Family, Wiley, pp.81-103, 2009.
[15] R. Pierdicca, D. Liciotti, M. Contigiani, E. Frontoni, A. Mancini and P. Zingaretti, "Low cost embedded system for increasing retail environment intelligence" , Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1-6, 2015.
[16] Jayashri B. Madalgi & S. Anupama Kumar Efficiency of Naïve Bayes Technique Based on Clustering to Detect Congestion in Wireless Sensor Network. DOI: 10.1007/978-3-030-34080-3_84.
[17] [17] R. Regin1, S. Suman Rajest & Bhopendra Singh. Fault Detection in Wireless Sensor Network Based on Deep Learning Algorithms. http://dx.doi.org/10.4108/eai.3-5-2021.169578.
[18] Dr. Laxmi Lidiya. S. Suman, Rajest, "Correlative Study and Analysis for Hidden Patterns in Text Analytics Unstructured Data using Supervised and Unsupervised Learning techniques" in International Journal of Cloud Computing, International Journal of Cloud Computing (IJCC), Vol. 9, No. 2/3, 2020.
[19] Pushpender Sarao, Machine Learning and Deep Learning Techniques on Wireless Networks, . ISSN 0974-3154, Volume 12, Number 3 (2019), pp. 311-320.