International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 08 | Aug 2019 www.irjet.net p-ISSN: 2395-0072

Indian Water Pollution Monitoring and Forecasting for Anomaly with Fail-Safe Wireless Sensor Networks using Machine learning techniques

V. Geethanjali[1], N.L. Anbarivan[2]

[1]Bachelor of Technology, Peri Institute of Technology, Chennai
[2]Master of Computer Science, VIT, Vellore, India

Abstract – *Water contamination and untreated sewage are the greatest sources of contamination in India. The situation is serious to the point that there is no water body in India that is not contaminated. The single main motivation for water uncleanness in India is urbanization at an uncontrolled rate. In this paper, we put forward a water pollution tracking and prediction system in order to reduce the growing pollution and manage the pollution hotspots over a wider area. Various types of sensors are used to collect data from different locations. The data is modified on various aspects to improve the efficiency of the total system. To ensure the redundancy of the system, a weight based node failure detection and tracking units are also implemented. Collected Data are used to train machine learning algorithms to predict the future pollution rates. This model is optimized using gradient descent on varying datasets and will serve as advanced tool to reduce pollution as well as optimize monitoring of water bodies*.

***Key Words*: AQI, PCA, Water pollution, Edge intelligence, Clustering, TSNE, Linear regression, Anomaly detection.**

1. INTRODUCTION

Water contamination is one of the important issues confronting India at the present time. Specifically, Untreated sewage is the greatest wellspring of such type of contamination in India. There are different sources of contamination. For example, overflow from the rural division and unregulated units from industries. Truth be told, it is said that 80% of the waterbodies in India are exceptionally dirtied. Important water bodies like the Ganga and Yamuna are the most polluted in India.

The single main cause of water contamination in India is urbanization at an uncontrolled rate. This has prompted a few ecological issues in the long haul like scarcity in water supply, accumulation of wastewater. The treatment and transfer of wastewater has also been a noteworthy issue. The zones close to waterways have seen a lot of towns and urban areas and this has added to the developing force to this issue. Uncontrolled urbanization in these territories has additionally prompted age of sewage water. In the urban regions, water is utilized for both modern and local purposes from waterbodies, for example, waterways, lakes, streams, wells, and lakes. 80% of the water that we use for our household reasons for existing is passed out as wastewater. I

The Central Pollution Control Board (CPCB) along with State Pollution Control Boards (SPCBs)/Pollution Control Committees(PPCs) are observing the nature of water bodies at 2500 areas. The area under National Water Quality Monitoring Program (NWQMP) demonstrates that natural contamination is the main reason for water contamination.

2. LITERATURE SURVEY

This work represents the implementation of the two well-known power efficient data gathering and aggregation protocols: PEDAP and PEDAP-PA. Simulations are used to show that both the algorithms perform optimally. The simulations show that keeping all the working nodes together is important. PEDAP-PA performs best among others and when the lifetime of the last node is important, PEDAP is a good alternative. [1]

This paper introduces us a MAC and cross-layer routing approach to QoS assessment in a WSN. The investigation is primarily based on two methods: the best-effort and latency constraint. These approaches can be used for rapid assessment of expected quality of service in the networks as well as finding the time division multiplexing schedules for utilization in network. The final simulation that is done shows that reliability in substantial gain is achieved. [2]

In this paper, data-centric routing is statistically assessed and its performance is compared with traditional end to end routing schemes. The impact of source to destination placement and network density on the energy costs is carefully examined in this paper. The significance of data-centric routing that offers high performance across variety of operational scenarios. [3]

This paper compares various algorithms that make predictions in time series from WSN. A simulation is performed that shows the nature of the data and their entropy deeply influences the performance of the selected algorithm. After the implementation and observation of results, it is concluded that gradually changing data is best for ARMA, and for data with sharp changes, MA is most suited one. [4]

In this paper, a novel technique, DBP, is applied to over 13 million data points from four real world applications. The assessment shows that the technique conquers 99 percent of

the application data and its performance is often better than the other common approaches. [5]

Huge reductions in communication is automatically allowed in this technique. Practical use of DBP includes improving system lifetime from every aspect. The paper is very well explained and the important terms are exquisitely highlighted in detail. [6]

Grouping the sensors into clusters is very well explained in this paper. The technique used here is heterogeneous clustering, which is very energy-efficient. This is done by selecting the cluster head from the cluster with respect to the residual energy of the nodes, transmission range and number of transmissions. The connectivity is considered as a measure of QoS, is ensured by Route identification technique. [7]

This paper analyses the wireless sensor networks that are very important in distributed systems. The paper models and examines the performance value of data aggregation in the network in question. The results show that whatever the sources of cluster, either clustered or random, energy gains can happen with data aggregation. The energy gain is maximum when the number of sources is large and are located relatively close to each other. [8]

This paper analyses the three main phases of fault tolerance and fault detection models at four level of abstractions, namely, hardware, system software, middleware and applications. Four scopes, namely, components of individual node, each node, network and the distributed system also encloses the fault model that is being analyzed.

A final conclusion is made that a brief survey of the future directions can widely affect the tolerance research in wireless sensor networks. [9] This paper gives an analysis that defines the fault tolerance and the various terms related to it. Various aspects of data constraints such as redundancy and touched-upon fault tolerance has been explored and explained that are used in Wireless Sensor networks. Some of the techniques that has been covered in this paper are redundancy in hardware, NMR and N-version programming software. [10]

3. SYSTEM ARCHITECTURE

The key research methods developed are driven by the various monitoring needs tied to compliance with the national water quality standards, real time public information, and support for atmospheric and health research studies. The main aim is to design, develop and implement a mechanism to identify various contamination issues and assess the level of pollution in relation to the water quality standards as defined by Indian government. Analyzing effluent data values with the predefined thresholds as stated by CPCB and generating alert information depending on the degree of pollution is done.

3.1 Functional modules

- Sensors data Aggregation
- Gateway functionality
- Server analysis (ML, Decision making)
- Network failure optimization

4. METHODOLOGY

We acquired the dataset with various columns of sensor data from various places in India. The average readings of ambient air quality with respect to air quality parameters are collected from pH sensor, Nitratenan n+ nitritenann sensor, fecal coliform sensor, B.O.D. sensor, D.O. sensor, temperature sensor. Data acquired from the source has strident data since few of the stations have been shifted or closed during the period and the corresponding data was marked as NAN or not available. So we have to pre-process the data in order to remove the outliers.

4.1 Datasets

In this dataset we have the pollutant concentration levels occurring on each place. These parameters should be reduced in order to represent the learning and to increase the rate of prediction. We have calculated the water quality index (WQI) for all the available data points. To calculate the WQI, we have to find the individual indexes of each pollutant. Each index of pollutant represents the level of damage caused by the pollutant. Each indexes varies in its own scale.

4.2 Water pollution analysis

From the obtained dataset, various from pollutant concentrations are obtained from PH sensor, NITRATENAN N+ NITRITENANN Sensor, FECAL COLIFORM Sensor, B.O.D. Sensor, D.O. Sensor, TEMPERATURE sensor with respect to the timestamp.

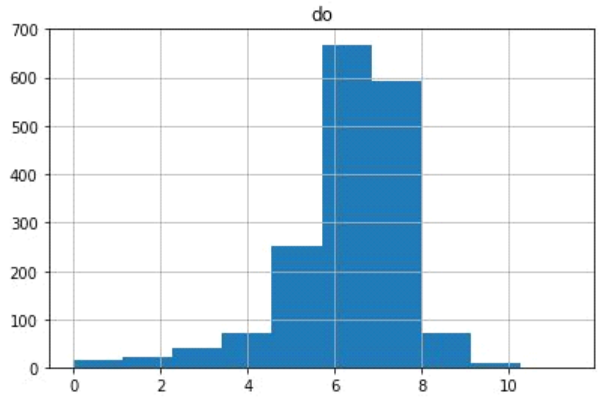

Fig -1: Histogram depicting concentration of D.O. (mg/l)

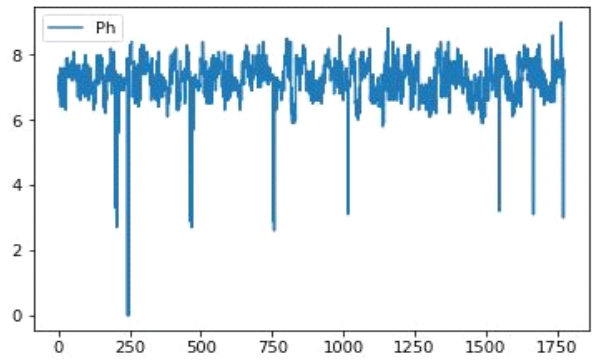

Fig -2: Graph depicting trend of pH

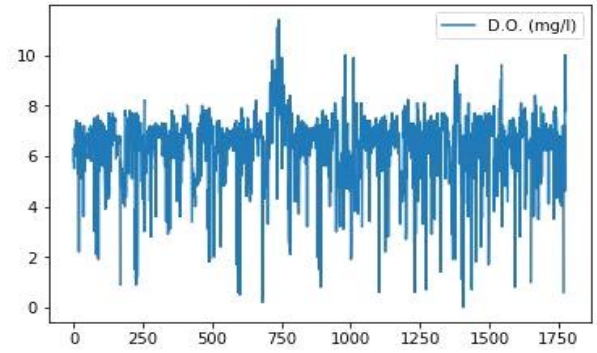

Fig -3: Graph depicting the D.O

These plots show the pollutant concentrations tend to vary a lot depending on the location, season and other affecting factors. These graphs have no increasing or decreasing trends on their measure. So various data cleaning techniques will be implemented to clean the data.

4.3 Outlier analysis

In this problem there are various outliers on the pollution concentration from various sensor readings, so box plot outlier analysis is used to identify and remove the outliers from the Data frame. The box plot consists of various quartiles Q1, Q2, Q3. Q1 and Q3 are the first and third quartile and Q2 is the median in the Data frame. After the Q1 region the points are called smallest non outlier and Q3 region has largest outlier.

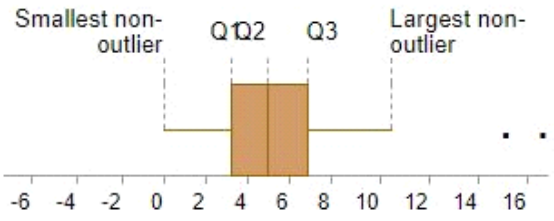

Fig -4: Representation of quartile range

The data points which are away from the Q1 and Q3 regions are classified as outliers. These outliers should be removed so the data will be cleaned, and various machine learning algorithms can be applied.

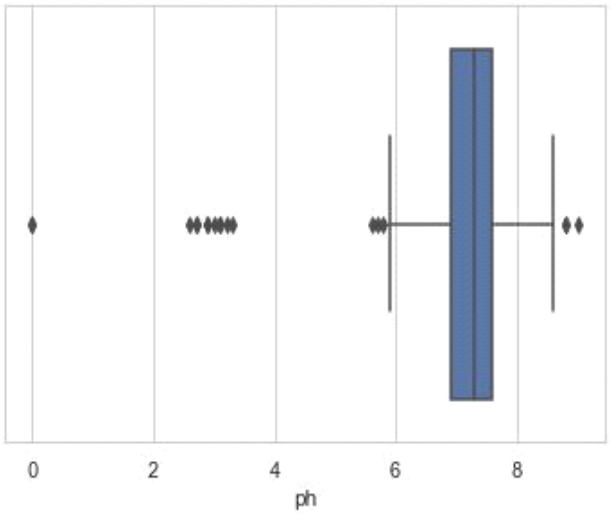

Fig -5: Boxplot analysis for PH values

4.4 Correlation matrix of various features

Correlation matrix is generated for the Data frame to identify the dependency or relationship between the features. Various features like do, ph, co, bod, na, tc data values are taken and correlation matrix is generated. This also helps in feature selection of the chosen data frame.

	do	ph	co	bod	na	tc
do	1	0.051	-0.16	-0.31	-0.21	-0.15
ph	0.051	1	0.093	0.093	0.081	0.02
co	-0.16	0.093	1	0.13	0.058	0.0033
bod	-0.31	0.093	0.13	1	0.15	0.24
na	-0.21	0.081	0.058	0.15	1	-0.0021
tc	-0.15	0.02	0.0033	0.24	-0.0021	1

Fig -6: Correlation matrix of water quality data

We have generated the heat maps for the correlation matrix to identify the level of dependency and relationship among the features.

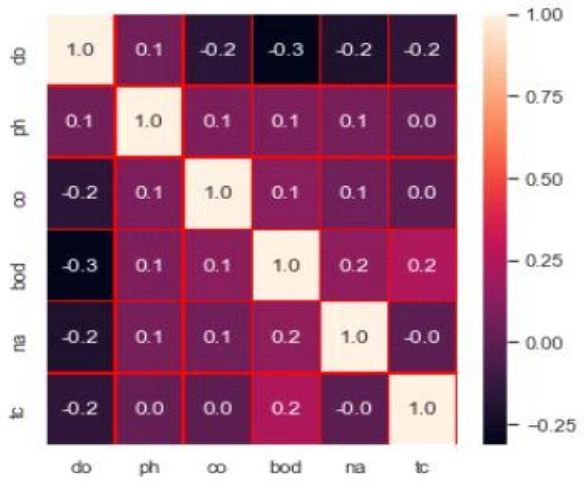

Fig -7: Heat maps of Correlation matrix

4.5 Calculating Water quality index (WQI)

Water quality index (WQI) is a standard rating to depict the overall water quality status that is helpful for the selection of appropriate treatment technique to meet the concerned issues. However, WQI depicts the composite influence of different water quality parameters and communicates water quality information to the public and legislative decision makers. In spite of absence of a globally accepted composite index of water quality. The water quality index of a particular data point is the aggregate of maximum indexed pollutant on that particular area. That pollutants maximum sub index is taken as the air quality index of that particular location. This maximum value of the pollutants is taken as water quality index so as to backtrack the pollutant levels from the water quality index.

Table -1: WQI range

National Sanitation Foundation Water Quality Index (NSFWQI)	
WQI Value	**Rating of Water Quality**
91-100	Excellent water quality
71-90	Good water quality
51-70	Medium water quality
26-50	Bad water quality
0-25	Very bad water quality
Canadian Council of Ministers of the Environment Water Quality Index (CCME WQI)	
95-100	Excellent water quality
80-94	Good water quality
60-79	Fair water quality
45-59	Marginal water quality
0-44	Poor water quality
Oregon Water Quality Index (OWQI)	
90-100	Excellent water quality
85-89	Good water quality
80-84	Fair water quality
60-79	Poor water quality
0-59	Very poor water quality

This method for comparing the water quality of various water sources is based upon nine water quality parameters such as temperature, pH, turbidity, feral coliform, dissolved oxygen, biochemical oxygen demand, total phosphates, nitrates and total solids. The water quality data are recorded and transferred to a weighting curve chart, where a numerical value of Qi is obtained. The mathematical expression for WQI is given by as per Indian government.

$$WQI = \sum_{i=1}^{n} QiWi$$

Where,

= sub-index for ith water quality parameter;

= weight associated with ith water quality parameter;

= number of water quality parameters.

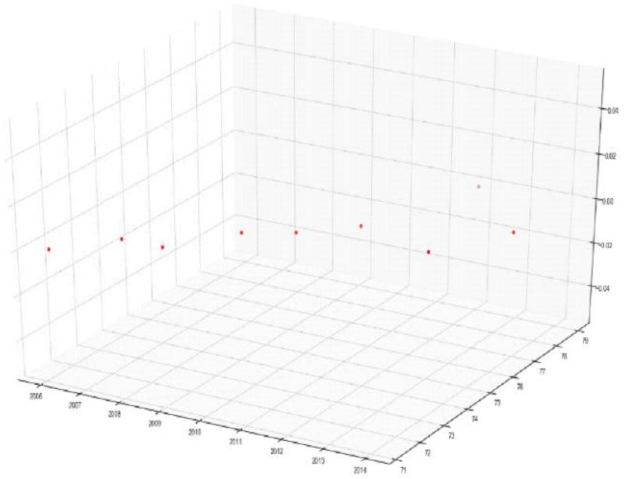

Fig -8: Scatter plot of data points

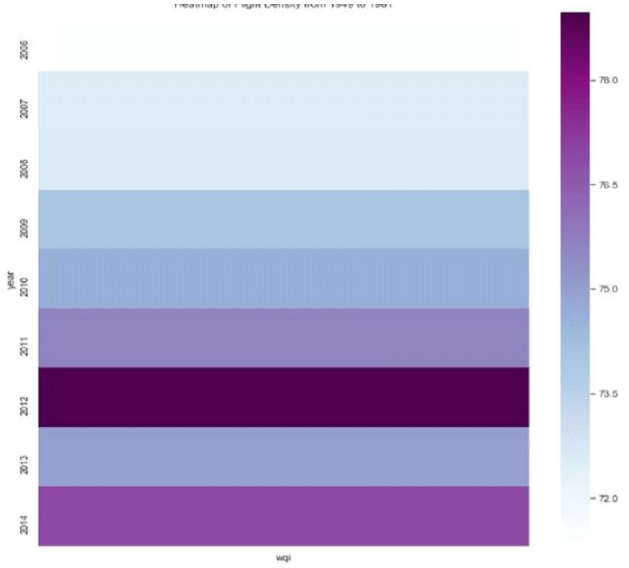

Chart -1: Heat map of WQI vs year

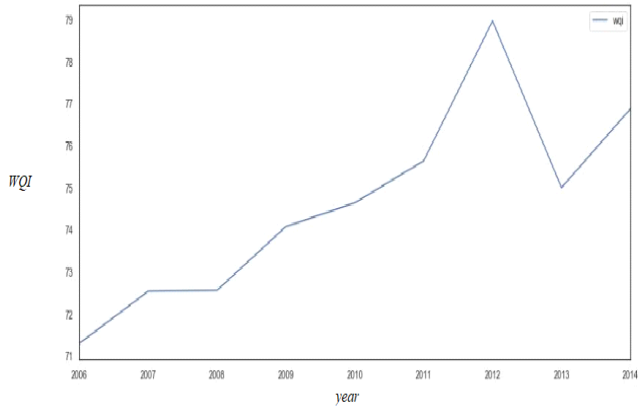

Fig -9: WQI vs year

4.6 Logistic regression with gradient descent

y = B1 + B2 * x ------- 1

Let,

B1 = 0.0

B2 = 0.0

On assigning,

y = 0.0 + 0.0 * x

error = p(j) – y(j)

x=1, y=1

p(j) = 0.0 + 0.0 * 1 ------- 2

p(j) = 0 ------- 3

we calculated the error by

error = 0 – 1

error = -1 ------- 4

We can now use this error rate in our equation for gradient descent to update the its weights. Then we will start with updating the slope intercept first.

B1(t+1) = B2(t) – alpha * error mean

B1(t+1) = B2(t) – alpha * error mean * x

B2(t+1) = 0.0 – 0.01 * -1 * 1

B2(t+1) = 0.01 ------- 5

We have just finished the first iteration of gradient descent and we have updated our weights to B1=0.01 and B2=0.01. This process must be repeated for the remaining 'x' number of instances from our dataset. (x=3000)

4.7 Node failure optimization

Each node should be assigned to a cluster based on distance. Each cluster will have a Cluster Head. The nodes, instead of interacting directly with the sink, should interact with their respective Cluster Heads. The data generated by the sensor nodes should be first received by the Cluster Heads. Cluster Heads will perform the aggregation and bucketing on the data and then send one small aggregated and bucketed data to the sink whenever relevant.

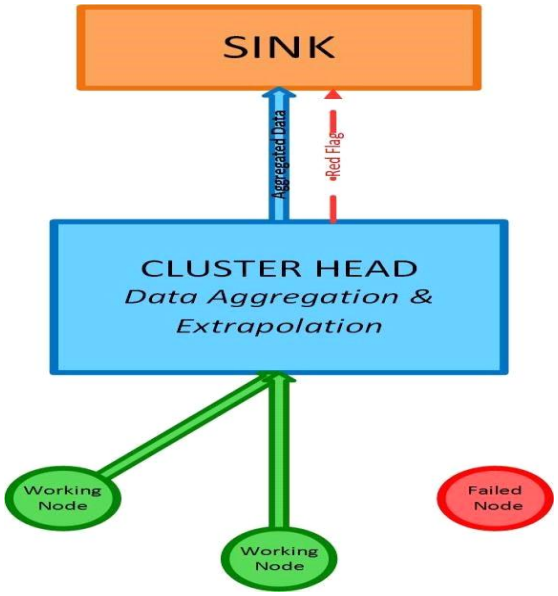

Fig -10: Proposed system architecture

The Cluster Heads are kept in parallel running the node-failure module where each node data is being received should be assigned with a weight and with each received data, the weight should be incremented. The weights should be measured again an ideal weight static variable and if the difference is higher than the threshold, the details should be logged at the Cluster Head database and a red flag should be raised to the sink.

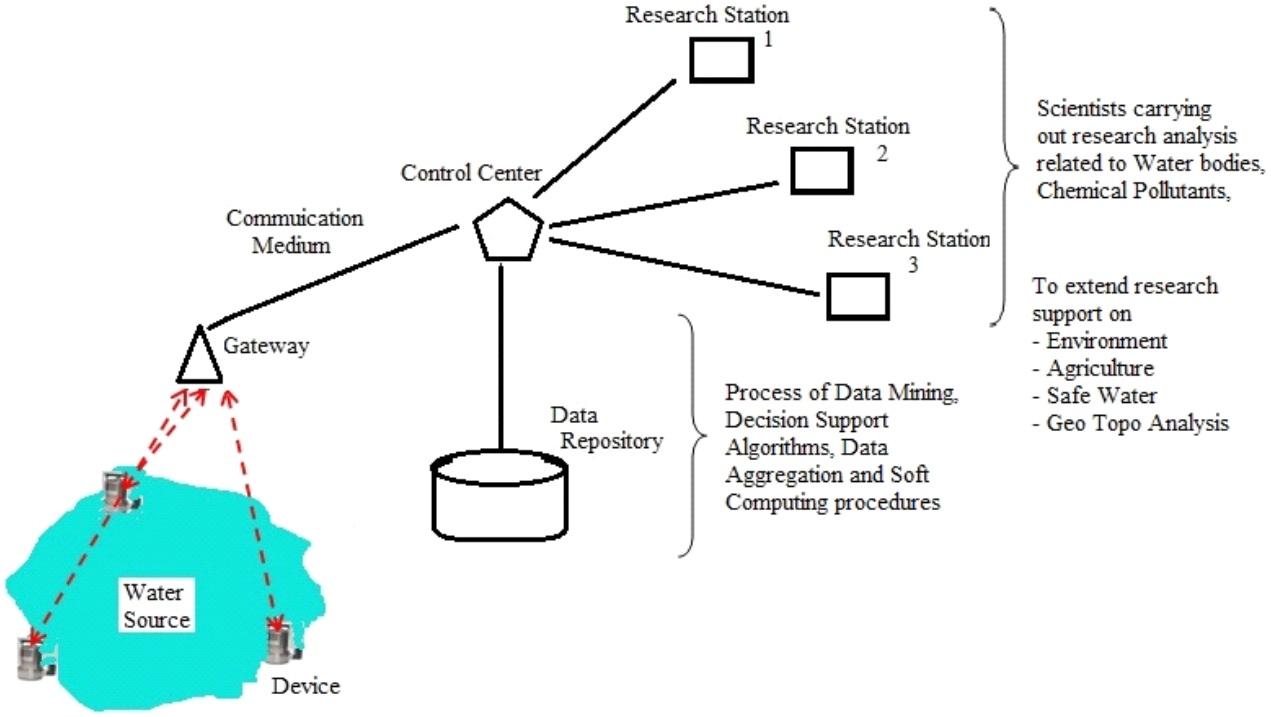

Fig -11. WSN Architecture

4.7.1 Functional modules of simulated environment

- Sensors data Aggregation
- Network failure optimization
- Server analysis (Clustering, Decision making)
- Next Hop Analysis

In the below depiction, gateway is the cluster head (CH) which has multiple sensor nodes, the data are aggregated over the cluster head and low-level analysis are being carried out. Node density can vary among cluster head (CH), these data from the various node and different clusters are formatted and stored in Data Repository. In control center most of the Server analysis such as Clustering, Decision making takes place to find various patterns.

5. SIMULATION ENVIRONMENT AND RESULTS

In this simulation, system architecture uses various sensors to collect the sensor readings / pollutants from various remote water source. So, the detection of failure is the more important aspect to make the system robust.

Various chemical sensors have been deployed to get the pollutant concentrations and data aggregation methods such as boundary value analysis to reduce the data over the network and make the system more efficient over longer run

Algorithm

1. Assign equal weight values to all the sensor nodes. Initialize all the weights to zero (w=0).

2. Increment the weight value when a sensor node transmits data. (Trans ~ w++)

3. Calculate a fitness value (V) based on higher weights of nodes. These nodes form the active node cluster.

4. The cluster head (CH) is selected for the node with the highest weight value (W). CH= Node(W)

5. The next hop (h) can be calculated based on highest weight (W) of neighbor mode which is decided by CH.

This algorithm allows intelligent dynamic route selection based on fitness values as well as the neighbor weight value.

5.1 Node failure detection

The fitness values are calculated according to the timestamp and independent of the values on the sensors. (t ~ fitness value). The failed node is detected by the minimum average of the fitness values (fv) among all the sensors nodes. (min_avg (fv)).

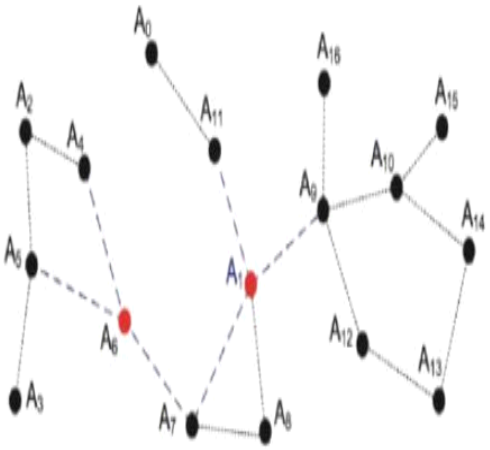

Fig -12: Figure depicting node failure occurring at A6 and A1

Working sensor: - Case study of sensors with cluster head with timestamps in x axis and pollutant concentration in y axis.

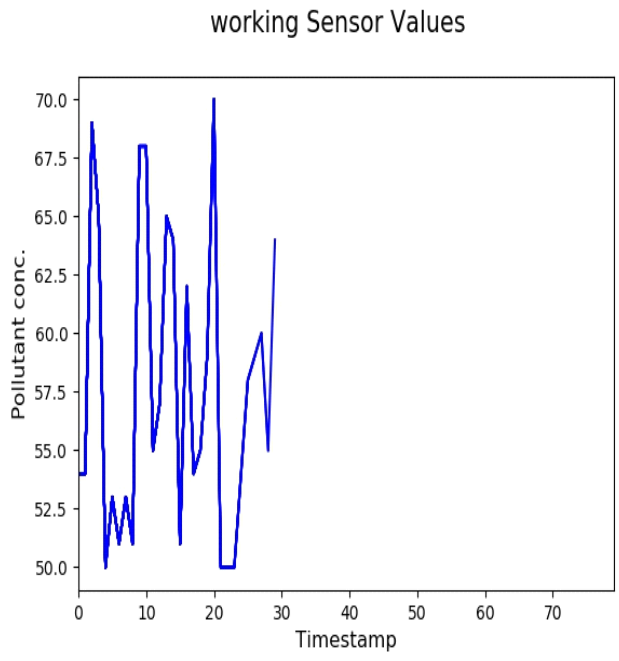

Fig -13: Working sensor data transmission graph

Data aggregation over cluster head (CH): - Pollutant concentration captured with respect to timestamp and corresponding fitness values for a particular sensor in cluster.

Table -2: Data aggregation over cluster head (CH)

timestamp	nitrate_conc.	Fitness_val
0.5	54	1
1.0	54	3
1.5	69	6
2.0	65	10
2.5	50	15
3.0	53	21
3.5	51	28
4.0	53	36
4.5	51	45
5.0	68	55
5.5	68	66
6.0	55	78
6.5	57	91
7.0	65	105
7.5	64	120
8.0	51	136

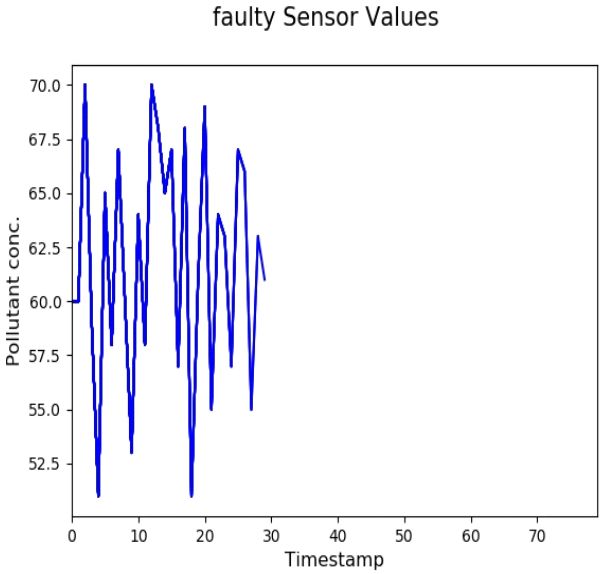

Fig -14: Faulty Sensor data transmission graph

Faulty Sensor: - In the above graph, it is observed some of sensors record no value which we can infer that there is a node failure.

Data aggregation over cluster head (CH): - Pollutant concentration captured with respect to timestamp and corresponding fitness values for a particular sensor in cluster with node failure (some of the fitness value are 0.)

Table -3: Fitness score of sensor nodes

timestamp	D.O._conc.	Fitness_val
0.5	0	0
1.0	95	2
1.5	0	0
2.0	0	0
2.5	88	5
3.0	95	11
3.5	0	0
4.0	93	8
4.5	0	0
5.0	0	0
5.5	0	0
6.0	81	12
6.5	0	0
7.0	0	0
7.5	88	15
8.0	0	0
8.5	0	0
9.0	0	0
9.5	0	0
10.0	82	20

This data is finally taken to analyze for other knowledge extraction patterns such as faulty node clusters, anomalies on clusters as well as on sensor nodes, selecting the next hop (h), etc.

5.2 Cluster node anomaly detection

On the cluster data, various high computational analysis is done to make the system robust and reliable for optimal usage.

Using cluster fitness values, we have clustered the nodes and cluster head according to their behaviors. We have used two principle technique to understand the patterns on large amount of data such as PCA and t-SNE.

5.2.1 Cluster formation: -

Here we are forming clusters on high dimensional view to segregate the desired values, we have spitted the data points into 3 clusters on basis of their behaviors.

1. next hop cluster
2. abnormal cluster
3. failed cluster

Here we are using the fitness value of each sensor nodes to extract the desired patterns

5.2.2 Distributed stochastic neighbor Embedding (t-SNE)

T-distributed stochastic neighbor Embedding (t-SNE) is a widely used machine learning technique that uses non-linear dimensionality reduction technique for data visualization, that makes use of statistical graphs, plots, charts, information graphics, etc. It is an algorithm that is well suited for setting data which has a higher dimension for visualization in a data space that has only two or three dimensions. The algorithm designs a high-dimension object in such a way that the similar plotted points are modelled by nearby points and dissimilar points are portrayed by distant points, embedded with high probability.

$$p_{j|i} = \frac{\exp\left(-|x_i - x_j|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-|x_i - x_k|^2 / 2\sigma_i^2\right)}$$

|xi−xj| be the Euclidean distance between two data points,

|yi−yj| the distance between the map points,

(xi) Gaussian distribution

(σ2i) variance

Then the similarity matrix for the given dataset

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

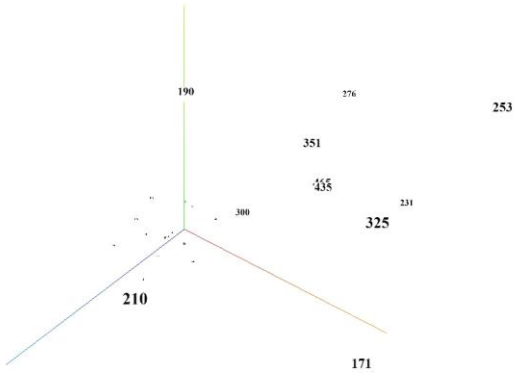

Fig -15: Faulty clusters on Fitness value

In the above figure, the normal node with highest weight value of highest fitness are observed in the left. These sensors can help in choosing the next hop. Similarly, the nodes in the right have lesser weight value which may be formed in failed cluster zone.

5.2.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical method that is used to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. These variables, after the transformation are called principal components. The procedure uses an orthogonal transformation to make the successful transformation. Orthogonal transformations that takes place in two or three-dimensional comprises of stiff rotations, reflections or the combination of the two.

$$Y = W'X$$

Where,

W is the matrix of coefficients that determined by PCA, X is the adjacent data matrix.

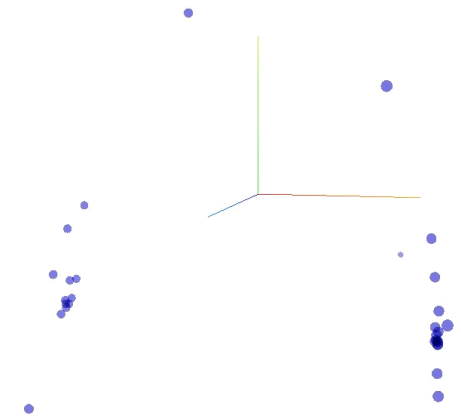

Fig -16: PCA on Fitness Value

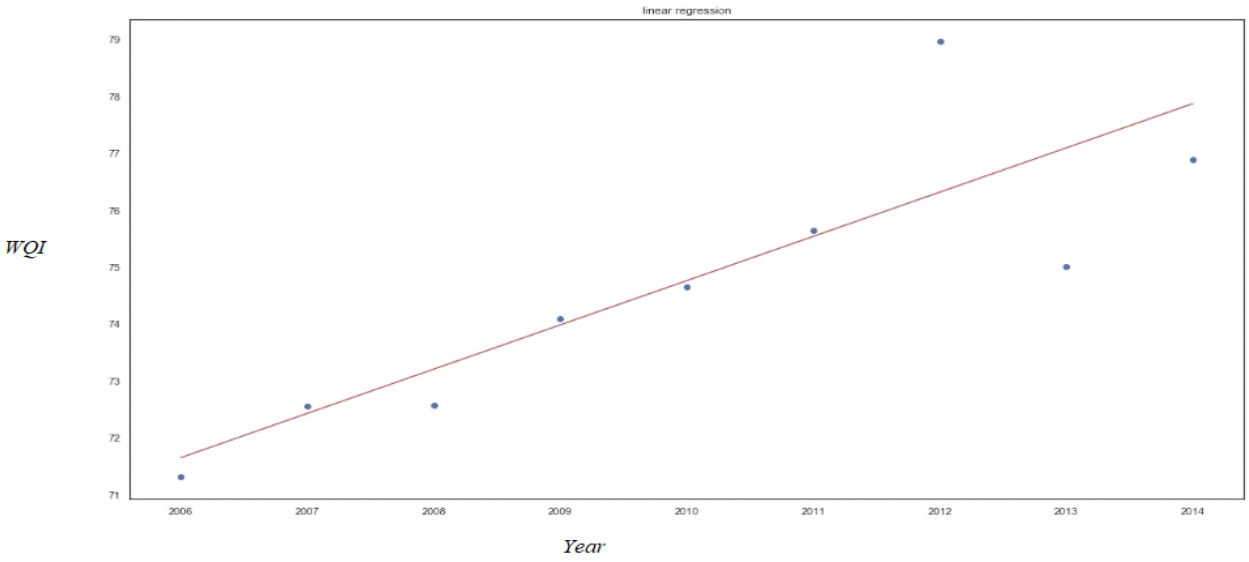

Fig -17: Plotting Logistic regression on WQI data

Table -4: Actual vs predicted data

	year	wqi	Actual	Predicted
0	2006	71.308824	71.308824	71.648936
1	2007	72.549000	72.549000	72.426702
2	2008	72.570943	72.570943	73.204468
3	2009	74.085193	74.085193	73.982234
4	2010	74.648723	74.648723	74.760000
5	2011	75.647013	75.647013	75.537766
6	2012	78.969041	78.969041	76.315532
7	2013	75.009425	75.009425	77.093298
8	2014	76.879588	76.879588	77.871064

5.3 Comparison of wireless sensor routing algorithm

Here, we are comparing the efficiency of similar wireless sensor node failure algorithms such as LEACH and SPIN. We compare the algorithm in terms of their network efficiency, data efficiency and scalability.

5.3.1 Network Efficiency

Here, using ARR algorithm, the data transfer size varies very less with the increase in the cluster heads, this is due to generation of the fitness values over the control center, the data over the network is much lesser than leach and spin.

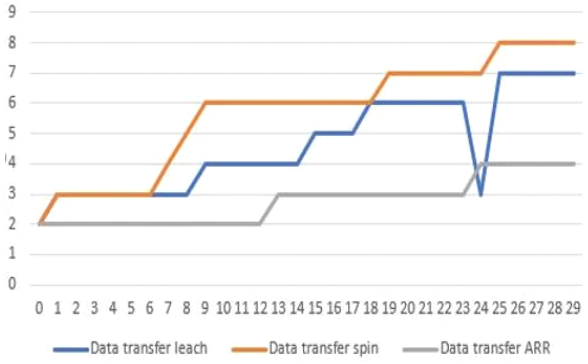

Chart -2: Data transfer vs number of Cluster heads (CH) graph

5.3.2 Data efficiency

Here, the data accumulation over the cluster head is directly proportional to the number of nodes, since the only parameter used to calculate fitness values, is sensor data only. Very less data used for routing and processing than spin and leach.

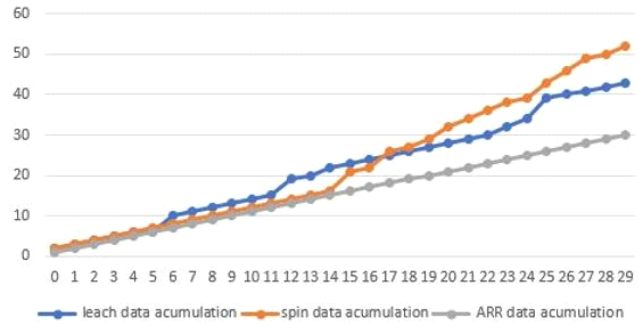

Chart -3: Data accumulation vs number of nodes graph

5.3.3 Scalability:

Here, the performance of various routing algorithms is compared in increasing cluster environment, since ARR have no additional dependencies or querying process it is highly scalable than spin and leach algorithms.

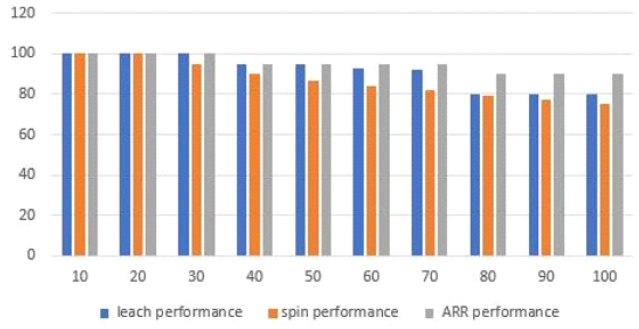

Chart -4: Performance measures of various algorithms

6. CONCLUSION

Using highly advanced machine learning techniques combined with a well modeled architecture we have monitored water contamination in India. This model performs better in terms of data and network efficiency, predicting WQI and architecture. It is also agile and acts as an automatic tracking system to tackle and reduce water pollution.

REFERENCES

[1] Tan, Hüseyin Özgür, and Ibrahim Körpeoğlu. "Power efficient data gathering and aggregation in wireless sensor networks." *ACM Sigmod Record* 32.4 (2003): 66-71.

[2] Dobslaw, Felix, Tingting Zhang, and Mikael Gidlund. "QoS assessment for mission-critical wireless sensor network applications." *2013 IEEE 38th Conference on Local Computer Networks (LCN 2013)*. IEEE, 2013.

[3] Stojkoska, Biljana, and Kliment Mahoski. "Comparison of different data prediction methods for wireless sensor networks." *CIIT, Bitola* (2013).

[4] Kumar, Dilip. "Performance analysis of energy efficient clustering protocols for maximising lifetime of wireless sensor networks." *IET Wireless Sensor Systems* 4.1 (2014): 9-16.

[5] El-Sayed, Hamdy H. "Data Aggregation Energy and Probability Effects on the Performance of EDEEC and MODLEACH Protocol in WSN." *Appl. Math* 12.1 (2018): 171-177.

[6] Gupta, Gaurav, and Mohamed Younis. "Fault-tolerant clustering of wireless sensor networks." *Wireless Communications and Networking, 2003. WCNC 2003. 2003 IEEE*. Vol. 3. IEEE, 2003.

[7] De Souza, Luciana Moreira Sá, Harald Vogt, and Michael Beigl. "A survey on fault tolerance in wireless sensor networks." *Interner Bericht. Fakultät für Informatik, Universität Karlsruhe* (2007).

[8] Iqbal, Muhammad, et al. "Wireless sensor network optimization: multi-objective paradigm." *Sensors* 15.7 (2015): 17572-17620.

[9] Akkaya, Kemal, and Mohamed Younis. "A survey on routing protocols for wireless sensor networks." Ad hoc networks 3.3 (2005): 325-349.

[10] Manjeshwar, Arati, and Dharma P. Agrawal. "TEEN: a routing protocol for enhanced efficiency in wireless sensor networks." null. IEEE, 2001.

BIOGRAPHIES

Geethanjali Vasudevan

N.L.Anbarivan