TWO LEVEL DATA FUSION MODEL FOR DATA MINIMIZATION AND EVENT DETECTION IN PERIODIC WIRELESS SENSOR NETWORK

Neetu Verma and Dinesh Singh

Assistant Professor, CSE Department, DCRUST, Sonipat, Haryana, India

ABSTRACT

Periodic Wireless Sensor Networks used for many applications such continuous Weather Monitoring, Cattle Monitoring, Water Quality Monitoring and Event Detection Applications. Due to continuous Monitoring, Large Amount of redundant data is transferred over the Network, which causes depletion of energy resources. Data Fusion is used with the Aim of deleting redundant, inconsistent data and provides more accurate data to the Sink. However, redundant data improves the data quality. Therefore, Two Level Data Fusion Model is proposed in this paper to transfer minimized data with the ability of accurately determines the event with minimum Delay. Data Fusion Model employs Cluster based Data Fusion at Two Levels; First Level at Sensor Node (SNs) and another level at Cluster Head (CH)Node. At the first level of Fusion, SNs Send single most common measurement to the CH by using a similarity function. Second Level of Fusion is applied on CH to remove similar multi attribute measurement. Also at CH Node, Multiple Correlation is used to accurately distinguish the abnormal event from the normal event or outliers within minimum delay. Experimental results are performed on real data to validate proposed Data Fusion Model is Better in terms of data transfer over Network, Redundancy, Energy Consumption over the Prefix Filtering Technique (PFF) . This Model is also used to accurately detect early events in case of Emergency.

KEYWORDS

Wireless Sensor Networks, Data fusion, Redundancy, Spatial and Temporal correlation, Event detection, Data minimization.

1. INTRODUCTION

Applications of WSN are air pollution monitoring, forest fire detection, landslide detection, water quality monitoring etc. Such type of applications requires continuous monitoring, thus Periodic Wireless Sensor Network (PWSN) is coming to use [1, 2]. PWSN is the network of tiny sensor nodes (SNs), where SNs periodically collects the sensed data, extract meaningful information by analyzing (aggregation) these data and send this information to the sink node. Since, SNs are placed close to each other's and may produce same or redundant data. Transmission of redundant data causes, wastage of energy resources and bandwidth. However, redundant data increases reliability and accuracy of information. One of the major limitations of WSN is its limited battery life. Continuous transmitting data over the network, consumes more energy resources. To prolong the network life, data reduction techniques are used. Most common data reduction techniques are Data aggregation, Data prediction and Data fusion. Data aggregation is used to aggregate the sensor data and transmitted to the sink node, which reduces communication costs, and save energy [3, 7, 18]. The main limitation is that aggregated data may produce low quality data and proceeds to false positives and negative results. In data prediction, future values of sensors are predicted within certain error bound, without communicating with SN's and forwards the

predicted data to sink node. Predicted data with error bound to also reduce the accuracy. Data fusion, integrates the data and information from multiple sources to produce more accurate and reliable data as compared to data aggregation and prediction. Data fusion can be classified as: data fusion or low level fusion, information fusion or medium level fusion and decision level or high level fusion. Low-level fusion is used to process collected sensor data and produce data that are more accurate or information as an output. Medium- level fusion is used to fuse the important features or information and obtains new accurate information or feature. High- level fusion integrates the input decisions to form a new or better decision. Multi level data fusion uses the above mentioned data fusion level (low, medium and high) at different stages. Sensor fusion is also known as multi-sensor data fusion, which integrates data from different sensors and provides more reliable and accurate result. Multi level data fusion, detect events that occur in a particular area of interest [39-41]. Multi sensor data fusion is also used for wireless monitoring system such as waste water quality monitoring, forest monitoring, greenhouse monitoring etc. [42, 43]. In wireless monitoring, sensor's data are continuously sensed by sensor nodes and forwards to the sink node. Sensor's nodes may produce redundant measurements. Data fusion (at low level) or data aggregation eliminates these redundant data [20, 22-25]. However, data redundancy increases the accuracy and quality of data. There is a tradeoff between accuracy and energy consumption. Also, no research work is found, where efficient wireless monitoring and event detection can be performed by the same data fusion approach. [20, 22, 24, 28, and 29]. Therefore, we proposed a two level Data fusion model for data monitoring and event detection applications. Temporal and spatial correlation, eliminate redundant and inconsistent data and transferred less amount of data over the network. Besides, of using a reduced amount of data, the two level data fusion model identified reliable event with minimum delay. In two level data fusion model, first level of data fusion is applied on SNs, called node level aggregation and second level aggregation is applied on Cluster Head (CH) and termed as CH level aggregation. We compared the node level and the CH level of data fusion model with existing algorithm and found that two level data fusion model provide better results in terms of sending data, aggregated data, event detection accuracy and precision.

2. MATERIALS AND METHODS

Data fusion is a broad area which comprises data aggregation as a sub-process. It focuses on information rather than data and classified by various techniques. Data fusion and data aggregation may be used interchangeably at level 0 of data fusion in many research areas. In WSNs, Sensors are densely deployed due to which same or similar values are frequently occurring. These redundant data are generated due to temporal, spatial and tempo-spatial correlation [14].To reduced communication costs for transferring redundant data various aggregations/fusion approaches are used. Fuzzy-based data fusion approach is able to determine the true values and aggregate it to reduce the burden of base station (BS). It is also used to remove duplicate data, which ultimately reduce energy consumption [22].SVM method is applied on aggregation tree to remove redundant data. Aggregator node finds SNs that have the same data and allow only one sensor node to send the actual data. Various temporal, spatial and tempo-spatial aggregation/fusion approaches has been developed in last year's. Self Based algorithm (SBR) uses temporal correlation to find reduced format of information. SBR is a data compression technique used for compressing multiple streams of sensors that containing historical data. It uses the temporal correlation between multiple measurements on the same sensor to record small details of measurement and compressing multiple time series. The aggregator node and SNs work in a synchronized manner to reduce the amount of message transmitted. Spatial correlation based clustering protocol [19] helps to deactivate some of the nodes that generating redundant information. In this method, CH exploits the spatial correlation for selecting appropriate member nodes to remain active. Sink node defines the correlation region and may vary according to application. CH constructs the cluster according to correlation region

and selects a representative node whose residual energy is higher and which is near to the centre of correlation region. Redundancy Elimination for Accurate Data Aggregation (READA) uses a grouping and compression technique to remove redundancy in the aggregated set before sending to the base station. READA provide an energy efficient way to preserve the accuracy [24].Tempo-spatial data reduction technique exploiting temporal and spatial correlations, temporal presents within same sensor and spatial correlation exists between different sensors. Spatiotemporal Clustering and Compressing Schemes (SCCS)[12]is a lightweight compressing algorithm based on spatio-temporal correlation. In this method, sensor data are stored in a buffer and exploit temporal correlation. When data buffer is full, divide and Conquer algorithm is used to create representative information based on temporal correlation which is sent to the cluster node. Here the spatial correlation is used to create clusters among nodes that sensed the event. But, waiting time is not defined, due to that outdated information may be received at the sink node. Mostly aggregation methods usually result in high delay and reporting outdated data at sink node [6].Waiting time/Sampling rate is an important consideration for fast and efficient aggregation. Dynamic local and global waiting time is introduced at a CH level to decide the most suitable time of aggregation and forwarding the resulting packet to the sink [8]. An Efficient Data Collection Aware of Spatio-Temporal Correlation (EAST) considers spatial and temporal correlation to find out the shortest route for transmitting fast and reliable data towards the sink node [6]. Adaptive sampling rate allows each SNs to adjust its sampling rate, according to dynamics of environmental condition and avoids over sampling and under sampling [9,20]. Adaptive sampling time is better than traditional sampling because it often uses a lower number of samples and provides better efficiency in term of accuracy and confidence. Also, optimized sampling method efficiently reduces energy consumption of network with high data accuracy and good stability [17]. Wavelet-based spatio-temporal [13] data compression algorithm simultaneously looks temporal and spatial correlations among the collected data. It established virtual grid among the sensor nodes and nodes lie inside the grid shows spatial and temporal correlation. To compress the data at the sensor node Wavelet transform algorithm is applied. Since, virtual grid is not based on dynamic region inaccurate information may also produce. EAST algorithm [14] is proposed to achieve higher accuracy by dynamically change region size and threshold error. Two tier tree based aggregation algorithm is proposed to reduce communication cost by using tempo-spatial correlation [3]. Clustering provide better results than tree structure so an optimized in- network aggregation scheme is applied to cluster based routing [15, 16]. In-network aggregation/local aggregation [21] finds redundant data at sensor node. If the current measurement is similar to the existing measurement than delete the current measurement and increase the weight of existing measurement by 1, otherwise add the measurement in the set. At the end of each period, sets (measurements with their weight) are sent to the aggregator node or Cluster Head (CH). CH uses the prefix filtering techniques to find and eliminate redundant pairs of sets and send minimized sets to the sink node. Unordered measurements are subsisting after local aggregation is applied in two tier aggregation. Since, the Jaccard similarity function is used to estimate the similarity between sets of data measures and due to uneven measures, Jaccard extracts less redundant pairs as compared to the distance function(Euclidian and Cosine) and one- way Anova model [5]. The sensor in the underwater wireless sensor network also used two tier approach requires to communicate under the water. SNs periodically eliminate redundant data prior to forwarding its CH. At CH enhanced K-means algorithm is applied to identify nodes generating similar data sets and aggregated these sets and send the result to the sink node [10].

Correlation is also used for detecting events. Common goals of event detection are accurately identified the event with minimum delay. But, sensor data in WSN are unreliable due to noise, malfunctioning of SNs and malicious attacks and produced inconsistent and outlier data. Since such type of factors influences the data quality and results inaccurate event detection. To achieve fast and accurate event detection, these factors need to be distinguished properly and apply some

mechanism to treat/delete these factors if possible [28].Tempo-spatial correlation among sensor nodes is also considered to deal with noisy and outlier data. Statistical, Nearest Neighbor, Clustering, Classification, and Spectral Decomposition approaches are employs to deal with outliers. An efficient real time event detection method is based on spatial-temporal correlation of sensor data. Using this method low frequency error is eliminated by the temporal relation of SNs. Spatial location is used to divide the sensors node into groups and work together to check the failed node. Spatial-temporal correlation [29] is very effective in removing all kinds of erroneous data. WSN is an energy constraint network so each working group is divided in to mini groups and different groups work in different periods in a polling manner to avoid simultaneously working. Many event detection applications have constant time constraints, but using this technique user is allowed to set deadline for the response time of query. Most of the approaches are restricted to detect events in a specific application, therefore a generic state model with neighborhood support (NS) is applied for event detection for generic application. In the generic state model, two level event driven detection model is used to detect an event, sensor node level state model detect abnormal behavior and NS is consider to confirm or refuse the event. After confirmation of the event, network level fusion characterized the event [30]. Threshold level of NS integrates with event is used to separate real alarms from outliers which improves the reliability of event detection. It is a light-weight model and maintains the trade-off between early event detection and reliability. To classify outlier reading as either erroneous data or an actual event, a new clustering based approach combined with nearest neighbor-based [31] is used. In this approach all the sensor data are grouped into clusters. In each cluster, the outlier detection algorithm is applied to determine and labeled cluster as normal cluster or outlier cluster. After this, classify the outlier values as the erroneous value or actual event based on nearest neighborhood algorithm. At the last, trustfulness of each sensor is computed and considers trustful nodes for detecting events. In early event detection technique, spatiotemporal correlations are used to detect an event, which may be exposed to noise and errors. To improve the quality of event detection, a novel approach for abnormal event detection in wireless sensor networks is applied [32]. This approach considers the spatiotemporal correlation between sensor nodes as well as correlations among observed attributes. A dependency model is constructed on observed attributes using Bayesian network to establish dependency model among variables. Using structured learning, dependency structure of abnormal events is obtained. Conditional probability table is maintained by each node using these variables. Attribute correlation confidence is applied to evaluate normal sensor reading or abnormal event pattern. Based on sliding window, time correlation event was detected and space correlation event is detected using the information of a neighbor node. This algorithm improves the accuracy by reducing the impact of interference factors and the rate of false alarm.

On the basis of above discussed techniques, a fact is found that multi-attribute sensor nodes are supportive for efficient environment monitoring and accurate event detection. Pearson coefficient is used to measure the correlation between two variables or two sensors and cannot be applied on more than two variables. So to found a correlation between multiple sensors, multiple- correlation is used. But unfortunately, very limited research work is found where multiple correlation are used for both event detection and monitoring purpose. So we proposed a cluster based fusion algorithm which exploits temporal, spatial correlation and multiple correlation for monitoring and accurate event detection with minimum delay.

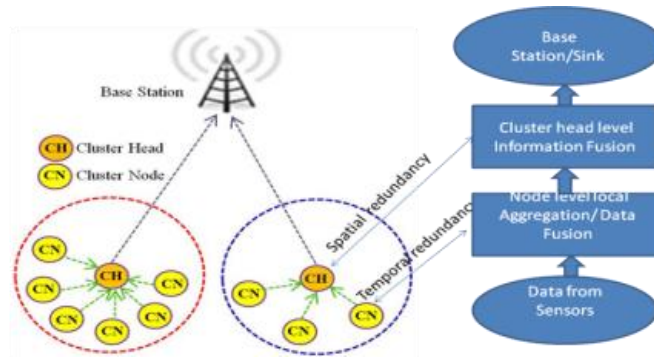## 3. PROPOSED TWO LEVEL DATA FUSION MODEL

Figure 1 Two level data fusion model

In the proposed two level data fusion model, low level data fusion is used at the sensor's node for performing local aggregation and medium level (information) data fusion is used in CH level. K-means clustering algorithm is applied to sensor's location data to form clusters. After clustering the sensor nodes, data fusion methods are applied on node level and the CH level to deliver accurate data to the sink node as shown in figure 1. Sensor node senses the value of temperature, relative humidity, and light intensity to monitor the forest condition and generate an alert if abnormal conditions are arises. At the node level, node of a particular cluster checks the similarities of its sensor value with recorded values, find out the data that is most frequently occurred, and send to the CH [35]. Node level aggregation reduces the amount of data by sending single most similar data, also it remove the local outlier. CH collects the data from cluster nodes and finds spatial redundancy using multi-attribute correlation and multi-attribute Euclidian distance. Multi-attribute Euclidian function reduces the amount of data transfer by removing the redundant data, whereas multiple correlation is considered to detect early and accurate event. At the last, CH sends minimized and accurate data to the sink node.

Multiple correlation coefficient:

A Multiple correlation coefficient is used to determine the influence of two or more variable in a single variable. It determines the maximum degree of linear relationship between two or more independent variables over a single dependent variable.

Suppose, three variables X1, X2 and X3are having N observations. The multiple correlation coefficient of dependent variable X1 on the independent variable X2 and X3 is the simple correlation coefficient between X1 and the combined coefficient of X2 and X3.

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{13}r_{23}r_{12}}{1 - r_{23}^2}} \quad \text{equation 1}$$

Where r12 is the correlation coefficient between variable 1and 2; r13 is the correlation coefficient between variable 1and 3; r23 is the correlation coefficient between variable 2and variable 3.
The value of Multiple correlation coefficient lies between -1 to +1. If the value Multiple correlation coefficient is 1 it shows perfect association between variables, however, if the value is 0 then the dependent variable is uncorrelated with independent variables. The value of Multiple correlation coefficient is greater than or equal to the total correlation coefficient.

Steps of two level data fusion model:

1. Create clusters using k-means clustering algorithm
2.   Choose CH among the cluster node
3. For each cluster node apply node level local aggregation algorithm (level1)
4. Cluster nodes send most similar measurement to the CH
5. CH apply CH level information fusion algorithm to find accurate information(level 2)
6. CH sends the accurate information to the sink node

We apply two level data fusion model for event detection and monitoring the forest fire application. Causes of fire in forest are depending on Climate conditions such as Temperature, Relative Humidity, light intensity, Wind Speed, and direction, Rain activity, duration of dry spell, etc. Considering all these factors are costly and increases the communication cost, therefore considering only those factors which are dominating and finds the relationship between them to determine the actual fire event. It is found that temperature, relative humidity and light intensity are dominating factors and related to each other's [36]. Relative humidity and Light Intensity shows a strong relationship with temperature, Negative correlation exists between temperature and relative humidity and positive correlation between temperature and light intensity. Therefore, these sensors are used for data fusion model to detect early fire event.

3.1 Data fusion at sensor node: Node level aggregation

In PSNs, each period p is divided into t equal time slots, $p=[t_1,t_2….t_p]$.At each slot t_i, every SNs collects new measurement mi(T,RH,LI) and forms a vector of measures as follows: $Mi=[mi(T,RH,LI)_1,mi(T,RH,LI)_2, ….mi(T,RH,LI)_n]$. Sensor Si may collect similar measurement due to temporal correlation. In order to delete the similar values, Si searches the similarities between vectors of measures by using Similarity function. Similarity function is defined as:

$$\text{Similar}(m_i,m_j)= f(x) = \begin{cases} 1, & \text{if } \|m_i - m_j\| \leq \varepsilon \\ 0, & \text{otherwise} \end{cases} \quad \text{equation 2}$$

Where m_i and m_j belongs to M and ε is threshold defined by the application. If two measurements are similar according to similarity function, sensor nodes deletes one measurement and increment its frequency count else, it adds the new measurement in to set and sets the value of frequency count to 1. At the end of t_i period, sensor node selects measurement whose frequency count is higher than others and sends to the CH.

Level1: node level aggregation algorithm

Require: new measure *Mj,* Set of previous measure M(Mi,…Mn).
Output: similar values having high frequency.
1.   For every measure of M in specific period p
2.   If |Mi-Mj|<= δ
3.   remove Mj
4.   F(Mi)= F(Mi)+1
5.   Else
6.   Add to the set M
7.   f(Mj)=1
8.   End for
9.   If   (M == 0) send a message " No value"
10.   Else, send the single measurement to CH whose frequency is higher among others.

3.2 Data Fusion at Level 2: Cluster Head Based Fusion Algorithm

In this section, CH based fusion algorithm is proposed for sending minimized data to sink node and accurately determines event within minimum delay. In cluster based PWSN, CH collects data from their Cluster nodes at a specific time interval.CH collects measurements from their respective cluster nodes and these measurements are same or similar due to the spatial correlation between cluster nodes. Therefore, CH based fusion algorithm is proposed to eliminate redundant data and send the minimized data to sink node. In the proposed algorithm, CH obtains single measurement m(T,RH,LI)from each cluster nodes in a specific time and checks the value of measurement. If the value of T,RH and LI in measurement (m) are below the threshold value (based on application), considered this measurement as a normal measurement and checks its similarity with the measurements in normal set (s1) using multi-attribute Euclidian similarity method. The measurement is similar to the existing measurements if its similarity threshold is less than all the measurements in a normal set. In this case new measurement is deleted otherwise it is added in the normal set. In this way, redundant data are eliminated and CH sends minimized data to sink node.

If the value of measurement (T or RH or LI) exceeds threshold, add this measurement in suspicious set (s2). Measurements presents in suspicious set are considered as either outlier or actual event. Multiple correlation is used to correctly identify between actual event or outlier. In generic state model [30], spatial correlation with neighborhood hood support is used to detect events. In this method, normal state is occurred when values of each variable/sensors are within its threshold value and suspicious state is present when values of any of the sensors exceed the threshold value. When a suspicious state is occurring, NS is used to determine the abnormal state (s3) or outliers (s4). If $NS(S_{suspicious}/S_n)$ is greater than or equal to neighbor threshold, abnormal state is occurring or if NS is less than neighbor threshold, outlier state is occurring. The main limitation of this method is that it considers the measurement lies in a suspicious set in determining the abnormal event or outlier. But in actual scenario, when an event is occurring, the measurement of nearby sensors varies accordingly due to spatial correlation and may or may not exceed the threshold value and must be considered to determine the event. For example, when a fire occurs in one place and it is about to start than reading of sensor near to that place exceeding the threshold values and placed in suspicious set. Suppose in a cluster, there are 10 sensors and its NS threshold is 0.5. Out of 10, 3 sensors are near to the event and placed in suspicious set. On the basis of earlier method, a fire event is not raised until it reaches NS threshold value. But, in a real scenario, if the actual fire event is occurred than the sensor nodes those are near to suspicious sensor are reflected accordingly and correlate with the suspicious sensor. These sensors may also be placed in suspicious set and used for event detection. Therefore we apply multiple correlation to find out all the measurement which are significantly varies with event. If number of measurement is greater than or equal to NS threshold than actual event is detected else outlier is detected. Figure 2 shows the event detection method in proposed algorithm. In this way this algorithm is used to detect accurate and early event detection.
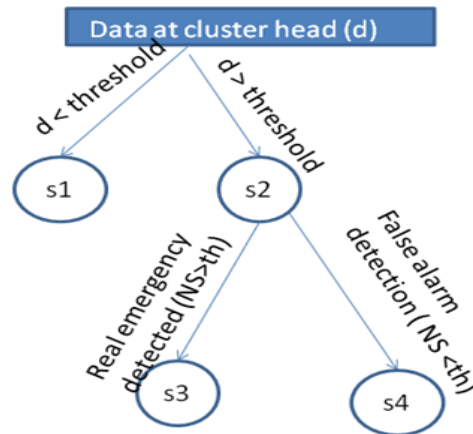
Figure 2: Event detection at CH

Event detection method used at CH level is represented as follows:

- S1set denotes normal state:

Normal state is occurred, when value of m type of sensor (Z_i) in measurement are less than threshold value.

$$Z_i \in S_1 \gg Z_1 \wedge Z_2 \wedge \ldots Z_m < TH_i,$$

The measurement is added in S1(normal set) set when measurement is in normal state.

- S2 set denotes Suspicious state:

Suspicious state is occurred when values of some or all sensors in measurement exceed its threshold.

$$Z_i \in S_2 \gg Z_1 \vee Z_2 \vee {}_{\ldots} Z_m < TH_i,$$

The measurement is added in set S2 (suspicious set) when suspicious state is occurred, after that CH checks the correlation of suspicious measurements with normal measurement and includes those measurement in suspicious set whose correlation coefficient is greater than 0.75 because when an event is occur at a node it is correlated with its neighboring node. Neighborhood support (NS) is used to determine the abnormal measurement or outlier and determined by ratio of number of suspicious measurement to the total number of measurement.

- S3 denotes Abnormal measurement:

If NS ≥ th (threshold) it indicates that an actual event is detected. When NS confirm abnormal measurement, it will send to sink node immediately.

- S4 denote outlier:

If value of NS is less than threshold then measurements are considered as outlier.

Level 2: CH based fusion

Require: New measurement m(T,RH,LI), Euclidian distance threshold(T_d), Multiple correlation coefficient (R)

Output : S is the resultant set and classify an actual event or outliers in abnormal condition.

1. For new every measure in CH at specific time interval
2. Check if (m_i(T,RH,LI)< threshold) [normal condition]

// Check its similarity with existing measurement using multi attribute Euclidean similarity method

3. $d = \sqrt{(m_i - m_j)^2}$ // calculate its similarity index
4. If (d > T_d) // Not similar
5. Add measurement M_i in set S,
6. Else
7. Rejected
8. Send reduced set S to the sink node
9. If (m_i(T,RH,LI) ≥ threshold [suspicious condition]
10. Add measurement in suspicious set

//Check the event using Multiple correlation applied on suspicious measurement and all measurement in normal set

11. $R_{T,RH,LI} = \sqrt{\dfrac{r_{TRH}^2 + r_{TLI}^2 - 2r_{TLI}r_{RHLI}r_{TRH}}{1-r_{RHLI}^2}}$
12. If (R>.5)
13. Add measurement in suspicious set
14. If (NH > th)
15. Real emergency detected
16. Else
17. False Alarm reported

4. RESULTS AND DISCUSSION

Intel Berkeley Research Lab [34]. Deployment of sensors is shown in figure 3. Mica2Dot sensors are used to collect temperature, relative humidity, light intensity and voltage sensor values in timestamp topology. Temperature is measured in degree Celsius, Relative humidity is ranging from 0-100% and light intensity is expressed in Lux. From the dataset it is observed that relative humidity is negatively correlated and light intensity is positively correlated with temperature. We take only one day sensor data for evaluating the results. Periodic network with Cluster-based network topology is used to implement the proposed algorithm, where the SNs are divided into several clusters using k-means clustering algorithm. Each cluster has a CH which is responsible for collecting the data from sensor node and forwarding this data to the sink node in a periodic manner. In this paper CH is defined using an appropriate clustering algorithm. There is a single hop communication between sensor node and their appropriate CH or CH and sink. Each sensor node periodically collects sensor data and applied Local aggregation algorithm [35] to remove redundant or outlier data by using temporal correlation. After local aggregation sensor node sends minimized data to their respective CH. CH collects data from their sensor node and employ CH based fusion algorithm to forward minimized data to sink node for monitoring and early event detection. The objective of proposed algorithm is to successfully forward minimized and accurate data to sink node for energy conservation. Another objective of this algorithm is to detect actual/accurate event with minimum delay or detect accurate event as early as possible.
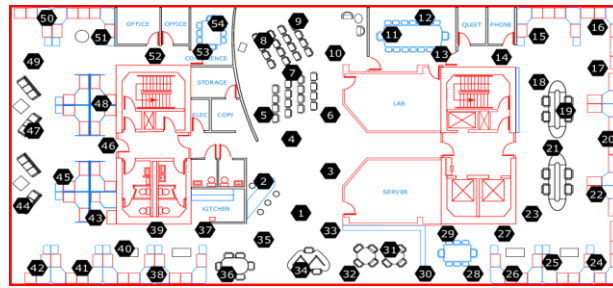
Figure 3: Sensors deployment in Intel laboratory

Various parameters are used to evaluate the performance of proposed model: similarity threshold (δ), time period (t), distance threshold t_d , multiple correlation factor(R) .The multiple correlation factor is fixed to 0.75. Jaccard threshold (t_j) is used when compare the CH based fusion algorithm with PFF method [21]. PFF algorithm is used only for data monitoring, so we also compare our CH based fusion algorithm with an existing algorithm (SM/NS TED) [30]. We demonstrate the performance of two level data fusion model for data monitoring and event detection for Periodic WSN as below:

A Proposed data fusion model is designed to monitor the sensor data efficiently by minimizing data transfer rate between sensor node and the sink node. The Metrics used to evaluate the performance of the data fusion model for data monitoring are:

1. **Data redundancy**: Redundancy represents a repetition of same or similar data in WSN measurements. This section shows the data redundancy at node level and CH level.

   a) **Data redundancy at node level** : node level aggregation is applied on the sensor node to find similar data or redundant data. At node level redundancy is influenced by similarity threshold (δ) and time period (t) and represented in figure 4.a and figure 4.b. It has been observed that when both δ and t increases, percentage of redundancy is increasing with same values in both algorithms because the same similarity function is used to find redundancy.
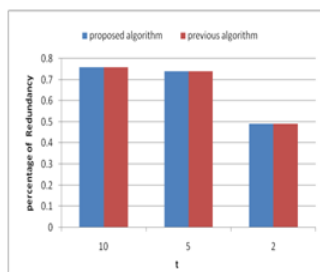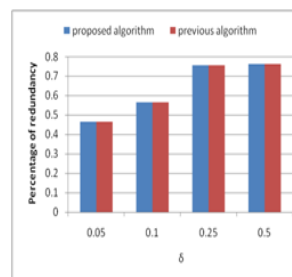

Figure 4.a   Figure 4.b

Figure4 Data redundancy at node level

   b) **Data redundancy at CH level**

   After receiving measurements from sensor node CH node apply second level fusion algorithm, i.e CH based fusion algorithm to check the similarities of measurement between sensor nodes. In our CH based fusion algorithm CH receives single most common measurement from sensor node and finds the similarities between measurements of other sensor nodes whereas in PFF it searches similarity between sets of measurements.

Therefore, our algorithm finds more redundancy as compared to PFF as shown in figure 5. Redundancy depends on three factors: δ, t_d, t. In figure 5.a t, t_d is fixed and varied the value of δ, in figure 5.b t, δ is fixed and t_d is varied and in figure 5.c t_d, δ is fixed and t increases. On the following observation, it is to be concluded that the proposed algorithm found 10 to 25 % (in case of t_d) and 43 % (in case t) more redundant measurements as compared to PFF, because Euclidian function found more similar measurement as compared Jaccard function which is applied on sets of measurement in PFF. It is also found that PFF searches more redundant sets when value of δ increased as PFF is dependent on δ.
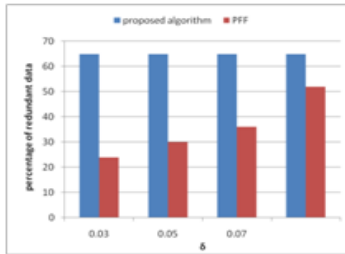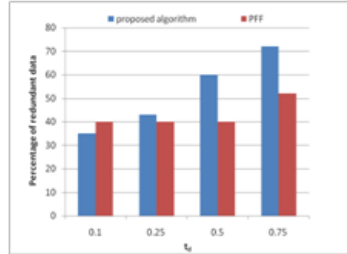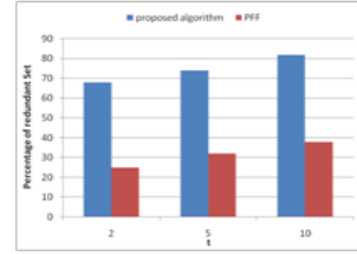
| Figure 5.a | Figure 5.b | Figure 5.c |

Figure:5 Percentage of data sent to Sink node

2. **Aggregated data**: Aggregated data is the amount of data which is transferred from sensor node to sink node or intermediate node after aggregation. This section shows the amount of aggregated data after node level aggregation and CH level.

a) **Aggregated data at node level:** At the node level sensor node find the similar measurement and sends single most similar measurement at each time period. Amount of data aggregated is depending on similarity threshold (δ), time period (t) and it increases when t and δ increases as shown in figure 6. The figure 6.a and figure 6.b show the comparison of the proposed algorithm with previous algorithm and without aggregation method. From the above observation it is found that at each period, 100 percent data is transferred when aggregation method is not used, whereas, in our approach 2 to 18 % (in case of t) and 8 to 45% (in case of δ) less data is transferred as compared to existing algorithm when δ and t rise.
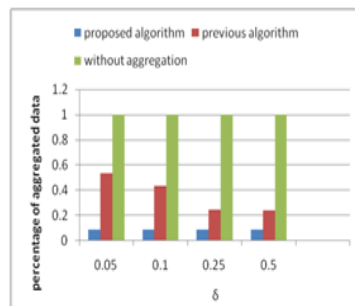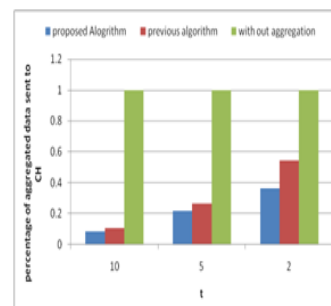

| Figure: 6.a | Figure 6.b |

Figure 6: Percentage of aggregated data

b) **Aggregated data at CH level:** CH node applies CH based data fusion algorithm on data collected from sensor nodes and removes the redundant measurement using Euclidian similarity method. CH then forwards reduced or aggregated data to the sink node. Since more measurement are reduced when applying the Euclidian method and transferred less

amount of measurement as compared to PFF and shown in figure 7. Figure 7.a to 7.c shows that proposed algorithm transferred 18% to 50% less measurement than PFF.
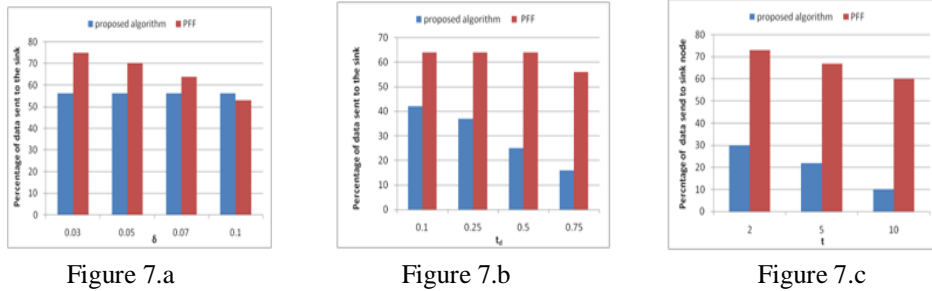
c)

Figure 7.a   Figure 7.b   Figure 7.c

Figure 7: Percentage of data sent to sink

3. **Data Accuracy:** The Accuracy of a network is estimated by percentage of loss measure when measurements are transferred from sensor node to sink node; in other words, it is evaluated as measurements taken by source node and did not reach at sink node. It is also defined as aggregation error or loss measure. Here we consider loss measure at node level and the CH level to evaluate overall loss measures.

a) **Loss measure at node level:** This section represents the impact of data accuracy on similarity threshold (δ) and time period (t). At sensor node, it is estimated by dividing the measurement of sensor node whose value or similar value did not receive at the CH over the total number of sensor measurements. Figure 8.a and figure8.b represents the percentage of loss measure when t and δ increased and concluded that loss measure increases, when t increases whereas it remains constant (90 %) in our case and increases (45 % to 75 %) in previous algorithm if δ increase. This is because proposed algorithm sends single measurement at each period of time where as in previous algorithm only similar data is eliminated and transfer remaining data to CH.
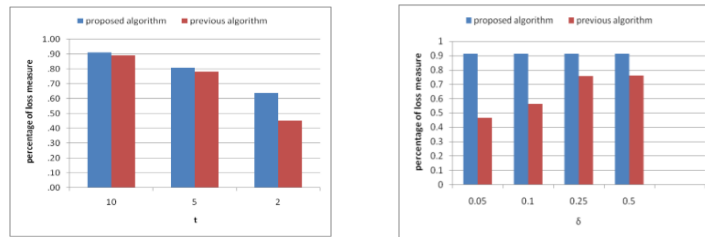

Figure: 8.a   Figure 8.b

Figure 8 Percentage of loss measure

b) **Loss measure at CH level:** loss measure at CH level is calculated by dividing the number of measurements at CH node whose value or similar values, not arrived at the sink node over the total number of measurements at CH node. Figure 9.a, figure 9.b, figure 9.c show the impact of δ, t_d and t on data accuracy and observed that percentage of loss measure increase when t and t_d increase. PFF is based on δ thus loss measure increases (2.5% to 10%) when similarity threshold increase. It is also found on the basis of observations that measure loss at CH in our algorithm is 1.5 to 23 % more than the PFF approach. The reason of decreasing data accuracy in proposed algorithm is to transferring less measurement as compared to PFF.
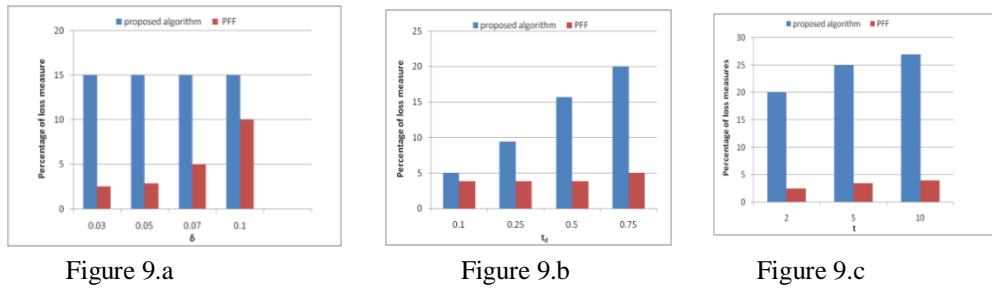
| Figure 9.a | Figure 9.b | Figure 9.c |

Figure 9: Percentage of loss measures

4. **Energy consumption:** Energy consumption of WSN is mainly depends on data sent and received. In this section we study the total amount of energy used to transfer sensor data towards sink node. To find energy consumption in proposed model we focus only on energy consumption by communication (transmission/reception).

a) **Energy consumption at node level:** Energy consumption at node level is determined by the amount of data communicated by sensor node to the CH after applying node level aggregation. Figure 10.a and figure 10.b show the energy consumption comparison of our approach with previous algorithm when δ and t varies. Energy consumption in proposed algorithm is reduced to 20% to 70% compared to previous algorithm when δ varies.
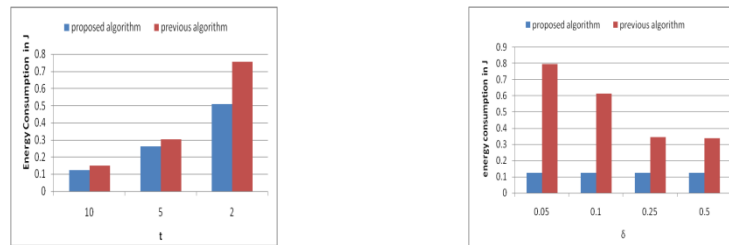

| Figure10 .a | Figure 10.b |

Figure 10: Energy Consumption at sensor node

b) Energy consumption at CH level: Energy consumption at CH level depending on the amount of data sent towards sink node. A lesser amount of amount of energy is required to transfer the reduced amount of data over the network. Figure11.a to figure 11.c show the comparsion of energy consumption at CH when proposed and PFF algorithm is used. It is to be concluded on the basis of the observation that the proposed algorithm consume approximately 30 % less amount of energy when t_d and t increases, however in case of δ it is reduced by 8 to 60 %.

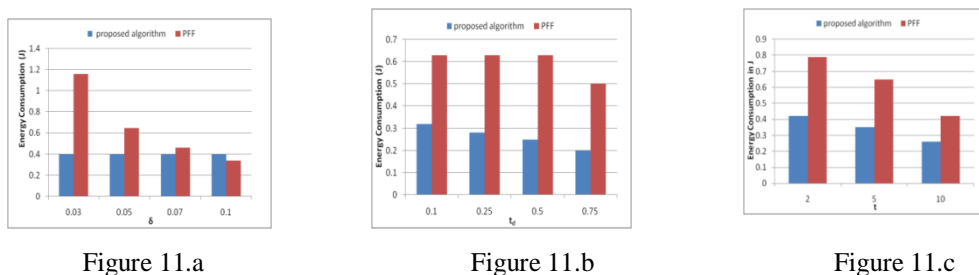| Figure 11.a | Figure 11.b | Figure 11.c |

Figure 11: Energy Consumption

Following metrics are used to evaluate the performance of the data fusion model for event detection: Event detection is defined as discovering considerable occurrence of data pattern which is unusual with normal patterns. Event detection is used to find out the particular event at time t   Performance of data fusion model is analyzed by considering following metrics: accuracy, precision, robustness, delay. To simulate the data fusion model for fire event in forest, we consider normal measurements are considered as "normal event" or abnormal measurement represents "fire event". We insert the some data manually to the real data sets which represent fire event or outliers. We also compare our model with SM/NS TED [30] for event detection.

1. **Accuracy:** Accuracy is represented as total number of correctly reported events (TP+TN) divided by the total number of events. Accuracy = TP+TN/ total event detected. Figure 12 show that accuracy decreases when the outlier rate varied from 5 to 30 percent. Outlier indicates a temporary error or sensor malfunctioning, therefore false negative and false positive events arise when outlier increase. A specified amount of increment is observed in accuracy if proposed algorithm is compared with SM/NS TED.
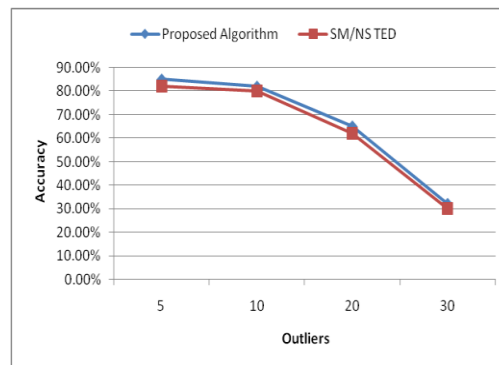

Figure 12: Accuracy when outliers increases

2. **Reliability:** Precision and recall are used to measure the reliability of fire event detection. Precision is measured as division of true positive event by true positive and false positive events and recall is considered as a division of True positive event by the true positive and false negative events. For the ideal situations value of precision and recall is near to 1. To calculate precision we varied the time duration from 2 to 10 minutes as shown in figure 13. It has been observed that data fusion model provide better precision than SM/NS TED because more measurement are included in suspicious set to conform about event in the same time duration.
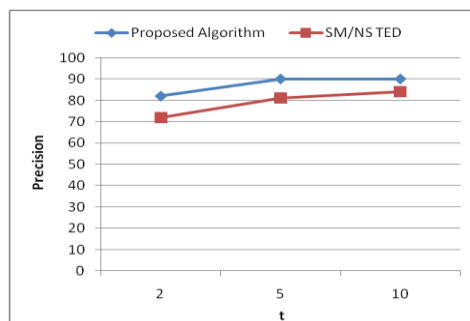

Figure13: Precision varies with time

3. **Delay Detection time:** Delay detection time is required to measures the time at which particular event is occurring. Figure 14 represents a delay time used by the data fusion model and SM/NS TED. Data fusion model detect accurate event with minimum delay as compared to SM/NS TED. This because to detect an event data fusion model considers the neighboring measurement that are correlated with events, whereas SM/NS waits until some measurements exceeds the threshold value.
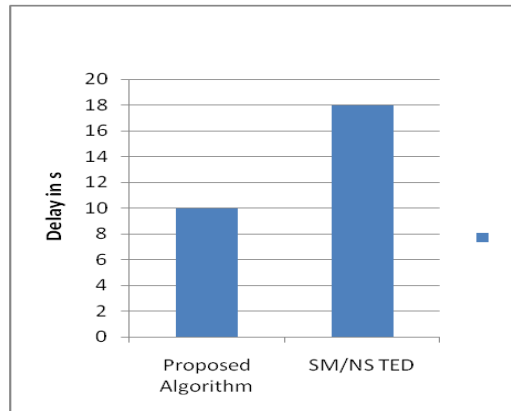

Figure 14: Delay(s)

4. CONCLUSION

Data fusion is applicable in WSNs for transferring minimized and accurate data. This paper presents a two level data fusion model for data minimization and event detection in the Periodic Two level Data fusion model uses multi-attribute sensors for data minimization as well as event detection, whereas previously illustrated techniques are limited to one attribute and accomplish either data minimization or event detection functions. The proposed data fusion model utilizes both spatial and temporal correlation for transferring minimized data as well as finding accurate event to the sink node.

Two level fusion model represents, the first level of fusion at sensor node and second level at CH level. First level fusion model or low level data fusion uses temporal correlation to transferred single more accurate information to CH by remove redundant and outlier data.CH node identifies incoming data from the different sensor node and reduced the data transfer rate by merging the redundant data. CH based data fusion also categorized incoming events as normal event, abnormal event and outliers.

The simulation results show that in a normal condition, proposed a data fusion model transferred up to 50% less data to sink node as compared to previous algorithm which saves the network life. In an abnormal condition, proposed model used multiple correlation to find accurate event with minimum delay and provides better results in term of accuracy and reliability over the existing algorithms.

To simulate the results, we are using three attribute, more attributes may be used to enhance the performance of the model. Sampling time and similarity threshold confer the impact of data transferring rate of WSN's, hence an optimized sampling data and similarity threshold needs to be analyzed.

REFERENCES

[1] Kezhong Liu et. al.,(2015) " Spatiotemporal Correlation Based Fault-Tolerant Event Detection in Wireless Sensor Networks", International Journal of Distributed Sensor Networks ,Volume 11(10)

[2] Kalman Tornai et. al., (2016) " Monitoring Scheme for Event and Danger Detection inWireless Sensor Networks", AHSWN 30(1-2), pp.145-161.

[3] Jacques M. Bahi et. al, (2012) "A Two Tiers Data Aggregation Scheme for Periodic Sensor Networks" Ad Hoc & Sensor Wireless Networks, 21(1), pp.1–24

[4] Amany Abu Safia , Zaher Al Aghbari and Ibrahim Kamel (2019) , "Distributed Environmental Event Monitoring using Mobile Wireless Sensor Network", Procedia Computer Science 155, pp 335–342.

[5] Hassan Harb et.al.(2017) "Comparison of Different Data Aggregation Techniques in Distributed Sensor Networks" , IEEE Access,vol.5pp pp. 4250-4263.

[6] Leandro A. Villas et.al, 2013 "An energy-aware spatio-temporal correlation mechanism to perform efficient data collection in wireless sensor networks," Computer Communications, vol: 36 , pp 1054–1066

[7] Prakashgoud Patil and Umakant Kulkarni, 2013 " SVM based Data redundancy Elimination for Data aggregation in Wireless Sensor Network", International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE

[8] Imane Horiya Brahmi et.al, (2015) "A Spatial Correlation Aware Scheme for Efficient Data Aggregation in Wireless Sensor Networks" Proceedings of the 40th IEEE Local Computer Networks Conference Workshops,pp.847-854

[9] Abdallah Makhoul et. al. , (2015)"An adaptive scheme for data collection and aggregation in periodic sensor networks", Int. J. Sensor Networks, Vol-18(1/2),pp 62-74.

[10] Hassan Harb et. al, 2015 "An Enhanced K-means and ANOVA-based Clustering Approach for Similarity Aggregation in Underwater Wireless Sensor Networks", IEEE Sensors Journal, vol. 15(10).

[11] Carlos Carvalho et. al., 2011,"Improving Prediction Accuracy for WSN Data Reduction by Applying Multivariate Spatio-Temporal Correlation", Sensors 2011, 10010-10037

[12] N.D. Pham, T.D. Le, K. Park, H. Choo, 2010, "Sccs: spatiotemporal clustering and compressing schemes for efficient data collection applications in WSNs ", Int. J. Commun. Syst. 23 (2010) 1311–1333.

[13] Dubois S., Péteri R., Ménard M. (2009) A Comparison of Wavelet Based Spatio-temporal Decomposition Methods for Dynamic Texture Recognition. In: Araujo H., Mendonça A.M., Pinho A.J., Torres M.I. (eds) Pattern Recognition and Image Analysis. IbPRIA 2009. Lecture Notes in Computer Science, vol 5524. Springer, Berlin, Heidelberg

[14] Leandro A. Villas et.al, 2013 , " An energy-aware spatio-temporal correlation mechanism to perform efficient data collection in wireless sensor networks ",Computer Communications 36 ,pp- 1054–1066

[15] Jacques Bahi, Abdallah Makhoul and Maguy Medlej,2012, "An Optimized In-Network Aggregation Scheme for Data Collection in Periodic Sensor Networks", 11-th Int.Conf. on Ad Hoc Networks and Wireless, Jan 2012, Serbia. pp.153-166.

[16] Yao Lu,"Benefits of data aggregation on energy consumption in wireless sensor networks", IET Commun., Vol. 11 (8), pp. 1216-1223

[17] Xinglin Piao, (2014)"Correlated Spatio-Temporal Data Collection in Wireless Sensor Networks Based on Low Rank Matrix Approximation and Optimized Node Sampling", sensors 2014, 23137-23158.

[18] Sukhchandan Randhawa and Sushma Jain , 2017 "Data Aggregation in Wireless Sensor Networks: Previous Research, Current Status and Future Directions", Wireless Personal Communications,97(4),1-71.

[19] Ghalib A. Shah and Muslim Bozyigit, "Exploiting Energy-aware Spatial Correlation in Wireless Sensor Networks",2nd international conference on communication system software and middleware,IEEE Xplore.

[20] Prasanna Sridhar, Asad M. Madni, and Mo Jamshidi, 2006 "Hierarchical Data Aggregation in Spatially Correlated Distributed Sensor Networks", World Automation Congress (WAC), July 24-26, Budapest, Hungary.

[21] Elena Fasoloy et. al., 2007, "In-network Aggregation Techniques for Wireless Sensor Networks: A Survey, IEEE wireless communication, vol 14(2), pp 70-87

[22] Davood Izadi, 2015"A Data Fusion Method in Wireless Sensor Networks, Sensors 2015, 2964-2979

[23] Alia Ghaddar et. al., 2010,"Algorithm for data similarity measurements to reduce data redundancy in wireless sensor networks",IEEE international symposium on A world of wireless, mobile, multimedia networks

[24] Kavi Khedo et.al., 2010 "READA: Redundancy Elimination for Accurate Data Aggregation in Wireless Sensor Networks", Wireless Sensor Network, 2, 300-308

[25] M.C. Vuran, O.B. Akan, I.F. Akyildiz, 2004, " Spatio-temporal correlation: theory and applications for wireless sensor networks ", Comput. Networks, 45, 245–259.

[26] A. Deligiannakis, Y. Kotidis, N. Roussopoulos, 2004, " Compressing historical information in sensor networks",Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, SIGMOD'04, ACM, NewYork, NY, USA, 2004, pp. 527–538.

[27] Li and Wang, 2013 "Automatic ARIMA modeling-based data aggregation scheme in wireless sensor networks", EURASIP Journal on Wireless Communications and Networking, Research paper pp-81-105.

[28] Yang Zhang, Nirvana Meratnia, And Paul Havinga, 2010, "Outlier Detection Techniques For Wireless Sensor Networks: A Survey, IEEE Communications Surveys & Tutorials, Vol. 12, No. 2, Second Quarter 2010

[29] Fangfang Li and Zhibo Feng, 2011,"An Efficient Real-time Event Detection Approach Based on Temporal-Spatial Correlations in Wireless Sensor Networks", IEEE, 2011 International Conference on Computer Science and Network Technology,1245-1249

[30] Yanning Yang et. al., 2012 "A generic state model with neighbourhood support from wireless sensor networks for emergency event detection", Int. J. Emergency Management, Vol. 8, No. 2, 2012

[31] Asmaa Fawzy, 2013,"Outliers detection and classification in wireless sensor networks", Egyptian Informatics Journal

[32] Mengdi Wang, Anrong Xue, and Huanhuan Xia , 2017, "Abnormal Event Detection in Wireless Sensor Networks Based on Multiattribute Correlation", Hindawi Journal of Electrical and Computer Engineering Vol 2017, 1- 8

[33] Carlos Carvalho et.al ,2011 "Improving Prediction Accuracy for WSN Data Reduction by Applying Multivariate Spatio-Temporal Correlation" Sensors (Basel) 2011; 11(11): 10010–10037.

[34] Http://Db.Csail.Mit.Edu/Labdata/Labdata.Html

[35] Neetu Verma, Dinesh Singh, 2019," Local Aggregation Scheme for Data Collection in Periodic Sensor Network" International Journal of Engineering and Advanced Technology (IJEAT) ,Volume-9 (2), 2249 – 8958

[36] Neetu verma and Dinesh Singh, 2020,"Analysis of cost-effective sensors: Data Fusion approach used for Forest Fire Application", Material science procedding,24, 2283-2289