Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

REGULAR

PFARS: Enhancing Throughput and Lifetime of Heterogeneous WSNs through Power-aware Fusion, Aggregation and Routing Scheme

Rahim Khan[1] | Muhammad Zakarya[1] | Zhiyuan Tan[2] | Muhammad Usman[3] | Mian Ahmad Jan*[1] | Mukhtaj Khan[1]

[1]Department of Computer Science, Abdul Wali Khan University Mardan, KPK, Pakistan
[2]School of Computing, Edinburgh Napier University, Edinburgh, United Kingdom
[3]School of Science, Engineering and Information Technology, Federation University, Victoria, Australia

Correspondence
*Mian Ahmad Jan
Email: mianjan@awkum.edu.pk

Abstract

Heterogeneous wireless sensor networks (WSNs) consist of resource-starving nodes that face a challenging task of handling various issues such as data redundancy, data fusion, congestion control, and energy efficiency. In these networks, data fusion algorithms process the raw data generated by a sensor node in an energy-efficient manner to reduce redundancy, improve accuracy, and enhance the network lifetime. In literature, these issues are addressed individually and most of the proposed solutions are either application-specific or too complex that make their implementation unrealistic, specifically, in a resource-constrained environment. In this paper, we propose a novel node level data fusion algorithm for heterogeneous WSNs to detect noisy data and replace them with highly refined data. To minimize the amount of transmitted data, a hybrid data aggregation algorithm is proposed that performs in-network processing while preserving the reliability of gathered data. This combination of data fusion and data aggregation algorithms effectively handle the aforementioned issues by ensuring an efficient utilization of the available resources. Apart from fusion and aggregation, a biased traffic distribution algorithm is introduced that considerably increases the overall lifetime of heterogeneous WSNs. The proposed algorithm performs the tedious task of traffic distribution according to the network's statistics, i.e., the residual energy of neighboring nodes and their importance from a network's connectivity perspective. All our proposed algorithms were tested on a real-time dataset obtained through our deployed heterogeneous WSN in an orange orchard, and also on publicly available benchmark datasets. Experimental results verify that our proposed algorithms outperform the existing approaches in term of various performance metrics such as throughput, lifetime, data accuracy, computational time and delay.

KEYWORDS:
heterogeneous wireless sensor networks, data fusion, data aggregation, nodes vulnerability and routing

1 | INTRODUCTION

Wireless sensor networks (WSNs) consist of sensor nodes, often resource-constrained, which are either deployed randomly or manually in closed proximity of phenomena of interest. These nodes periodically probe the deployed environments after a predefined time interval and transmit their data to a gateway node, i.e., sink, via a wireless transmission mechanism.[1] Due to the limited transmission power of these nodes, direct communication with the gateway is not always possible. Therefore, a multi-hop communication mechanism is used to extend the coverage capacity of a resource-constrained WSN.[2] Although, the multi-hop mechanism increases the coverage capabilities of WSNs, it experiences further challenges in term of message delay, congestion control, power consumption, and reliability. To extend the coverage area with the least possible set of deployed nodes, heterogeneous WSNs (HWSNs) are used in the literature.[3,4] These networks often comprise two types of sensor nodes, i.e., ordinary nodes and cluster heads (CHs). Usually, the cluster heads are more powerful than the ordinary nodes. Typically, the ordinary nodes are deployed in closed proximity of the observed phenomena and their task is to collect data either periodically or upon occurrence of a particular event. The observed data is transmitted to the nearest cluster head either directly or through multi-hop communication. In these networks, the cluster heads are responsible for reliable transmission of ordinary node's data to the sink.[5] Due to dense and random deployment of sensor nodes, redundant and highly correlated data are generated. In particular, the nodes residing in closed proximity generate highly redundant data.[6] As a result, traffic across the network is increased that drastically affect the lifetime of sensor nodes, more precisely the relay nodes. In WSNs, data fusion is a process that is often used to control the network congestion, minimizes redundant data, and replaces outliers/noise false data values with correct data.[7,8] In WSNs, outliers are either generated by a malfunctioning sensor node or through interference by neighbouring nodes.[9] In heterogeneous WSNs, data fusion mechanism is used to fuse the gathered data of multiple embedded sensors before its transmission to a cluster head or sink node. Usually, in heterogeneous WSNs, data fusion mechanism follows a two-level architecture. At level 1, the nodes have limited resources, i.e., low power, short-range wireless communication and scarce computational capabilities. The level 2 nodes, also known as super nodes or cluster heads, have long range transceivers, better processing capabilities, higher data rates and far more better storage reserves.[10] Data fusion approach is used at level 1 nodes to develop a mechanism that controls network congestion by minimizing the redundant packets transmission through sampling or similar techniques. At the same time, this approach substitutes the outliers with relatively accurate data. At level 2 nodes, this approach is used to improve, (a) the quality of service (QoS) locally, e.g., the throughput and end-to-end reliability within a cluster and (b) minimizes the congestion problem globally by discarding redundant data packets.[11]

A tightly coupled concept associated with data fusion, in heterogeneous as well as homogeneous WSNs, is known as data aggregation. This concept is becoming popular in research community of WSNs to refine and summarize the gathered data, usually collected from multiple sensors.[12] Data aggregation is defined as the process of collecting or manipulating raw data from multiple sources, i.e., sensors, and summarizes it to produce refined data that has minimum redundancy and volume. The refined data is eventually transmitted across the network to its destination, i.e., sink. Data aggregation is not only useful in generating refined version of the original data, but also helpful in addressing the congestion problem that arises due to the dense deployment of sensor nodes. In heterogeneous WSNs, aggregation techniques are usually applied at cluster head level, because the data generated by sensor nodes deployed in a closed proximity has a higher value of correlation factor and redundancy. Additionally, in a densely deployed heterogeneous WSN, it is difficult for a cluster head or base station to process such an enormous volume of data, particularly, in its raw form. Therefore, aggregation techniques, either node level or cluster head level, are the ideal solutions in different application areas of heterogeneous WSNs that have a strong emphasis on reliability of the refined data. However, these techniques must be energy-efficient, robust, precise and reliable, i.e., having a highest possible correlation with the original data. In addition to congestion control and redundancy, one of the major challenges is to develop an effective mechanism for prolonging the lifetime of heterogeneous WSNs. A node's lifetime is directly proportional to its sensing and relaying capabilities. These factors are controlled if a node's sampling interval and traffic through it, are minimized.

Apart from data fusion and aggregation techniques, vulnerability of a sensor node, i.e., its importance, is a key factor in heterogeneous WSNs. This factor severely affects the lifetime, throughput and reliability, i.e., the ratio of dropped packets to the generated packets, of a particular network. A sensor node is assumed to be vulnerable or critical if communication ability of a certain portion of a heterogeneous WSN depends, either entirely or partially, on its connectivity and smooth functionality.[13] Therefore, a uniform traffic distribution mechanism needs to consider the vulnerability factor of a sensor node before forwarding packets to it. A particular portion of the network gets disconnected as these nodes begin to deplete their available power rapidly in comparison to ordinary nodes. The idea of a uniform traffic distribution over non-vulnerable paths in heterogeneous WSNs leads to prolonged delay in packets transmission from source node towards the destination that limits its application areas.[14] Therefore, an energy-efficient traffic distribution technique needs to be designed that not only consider the vulnerability of sensor nodes to prolong their lifetime, but at the same time, thoroughly examine the residual energy levels of sensor nodes, particularly neighboring nodes. Moreover, this technique needs to be applicable in different application environments and more importantly minimizes the aforementioned delay factor in WSNs. In the literature, data fusion and aggregation mechanisms are centered on how to efficaciously use the on-board batteries of sensor nodes to enhance the WSNs' lifetime.[15,16,17] These approaches utilize different mechanisms proposed in the literature such as fuzzy set theory, sampling techniques, theory of probability, a hybrid of fuzzy set and probability, and evidence theory proposed by Dempster-Shafer.[18,19,20,21] These approaches are well-suited in WSN's scenario where data redundancy/duplication is a common issue. However, these approaches are based on an unrealistic assumption that a sensor node could always generate accurate data and, therefore, could always work properly. For example, sensor nodes deployed in an outdoor environment are more susceptible than nodes deployed in an indoor environment, because fluctuation in a measured phenomenon of interest may degrade/affect their performance. Therefore, a robust mechanism is needed to resolve these issues in an efficient manner and, at the same time, is based on realistic assumptions about sensor nodes. In heterogeneous WSNs, throughput and lifetime are well-known performance evaluation metrics. These metrics are directly proportional to the on-board battery of a sensor node and need to be utilized in an efficient manner. Usually, in heterogeneous WSNs, the transceiver of a sensor node consumes a significant fraction of the available power, i.e., transmission and reception of packets are controlled by using an energy-efficient routing protocol. These metrics are further improved if these techniques pay a careful attention to the vulnerable nodes in the networks while minimizing the delay factor.

Most of the proposed solutions are either applications-specific or too complex that make their implementation unrealistic, particularly, in constrained-oriented environments. To the best of our knowledge, we are not aware of a single study that solves these tightly coupled problems concurrently and, at the same time, enhances the throughput and lifetime of resource-starving sensor nodes. Any proposed solution needs to be robust, simple and designed according to the nodes' requirements. Besides, the proposed solution needs to be implementable in different platforms with negligible or minor modifications. In this article, a systematic approach based on data fusion mechanism is presented that enhances the accuracy of a sensor's collected data and consistently controls the congestion problem throughout a Heterogeneous WSN. Every sensed value of an embedded sensor is thoroughly examined, prior to further processing or transmission, to differentiate accurate data from the outliers. The proposed mechanism minimizes the energy consumption of a particular sensor by decreasing its ratio of transmitted packets to the sensed one, which ultimately enhances the network lifetime. The main contributions of our work are as follows.

1. A systematic sampling-based aggregation technique is presented to enhance the lifetime of heterogeneous WSNs by minimizing the individual sensor node's transmission ratio.

2. A simplified outlier detection mechanism is adapted to fine-tune the sensor node‘s collected data before its processing.

3. The congestion problem in heterogeneous WSNs is controlled by the removal of duplicate data packets that are usually generated by nodes resides in closed proximity.

4. The traffic is distributed across the network to further enhance its lifetime. The traffic distribution strategy is relaxed for those nodes whom connectivity, i.e., active status, is important to maintain a reliable communication of a particular portion of the heterogeneous WSN.

The rest of the paper is organized as follows. In Section 2, a brief literature review is presented with a strong emphasis on data aggregation and fusion techniques for heterogeneous WSNs. Section 3 describes our proposed scheme for data fusion, aggregation and vulnerability-aware routing. In Section 4, performance evaluations are described in detail and a comparative study of the proposed algorithms with field-proven algorithms on real-time datasets is presented. Finally, concluding remarks and future research directives are discussed in Section 5.

2 | LITERATURE REVIEW

To enhance the lifetime of WSNs, researchers and scientists have been primarily focused on the challenging task of how effectively the role of cluster heads rotation is applied to achieve their goals.[22] The rotation of cluster heads helps in uniform distribution of the traffic across the network. Various protocols have been proposed for cluster head rotation to enhance network lifetime. Among them,A fuzzy-based mechanism was proposed by Izadi et al[15] to resolve the lifetime issue particularly in WSNs where prior knowledge about sensing data errors is not known. In this scheme, every node is bounded to transmit the computed results, which is based on fuzzy logic controller (FLC), of events. Sensor nodes energy consumption is minimized by enabling these nodes aggregate true values which consequently enhance their lifetime. Low-energy adaptive clustering hierarchy protocol (LEACH) is a well-known protocol that partitions a sensor field into small geographical regions known as clusters.[23] Each cluster has a cluster head node that collects, aggregates and fuses data from member nodes and transmits them to a base station. The protocol operates in rounds and nodes take turn to become cluster heads in subsequent rounds for uniform distribution of energy load. This approach enhances WSN's lifetime but its implementation, for large-scale networks, is not suitable, due to the sensor nodes transceiver's limitation. Various extensions of LEACH protocol have been proposed in literature for large-scale networks. Younis et al.,[24] proposed a hybrid energy efficient distributed (HEED) clustering algorithm where the selection of cluster heads incorporates the intra cluster communication overheads along its residual energy. This algorithm solves the large-scale problem associated with LEACH via multi-hop communication, but, its load balancing mechanism is not efficient as it increases the overall traffic on nodes that reside in the closed proximity of the sink node. An alternative approach, distributed energy efficient clustering (DEEC) algorithm, has further refined the cluster head selection mechanism based on a probabilistic method.[25] The rotation of cluster heads helps in uniform distribution of the traffic across the network. This approach enhances the lifetime of WSNs up to some extent, however, at the same time it generates energy holes, particularly, in WSNs with a single sink module.[26]

To solve this issue, different methods have been proposed in the literature such as assistance approaches, traffic compression, and aggregation.[27,28] The assistance approaches solve this issue by the deployment of nodes having higher power and transmission capabilities. These nodes are deployed in regions where power consumption is relatively higher, i.e., in closed proximity of the gateway module.[29,30] Usually, these nodes work as relay nodes in heterogeneous WSNs and are quite effective in improving the network's lifetime.[31] An alternative approach based on the idea of spreading the traffic uniformly across the network, i.e., dense nodes deployment near the sink vicinity, results in an enhanced overall lifetime of the network, as described in[32,33]. Leu et al.,[34] described the feasibility of pre-deterministic distribution function to enhance WSNs lifetime and evaluated its effectiveness in uniform traffic distribution. The adjustment of transceivers' range plays a vital role in solving the aforementioned problem.[32,35,36] However, the transceivers have their own restrictions on the deployed regions of sensor nodes. Other solutions were proposed in literature to enhance the WSNs lifetime with a mobile gateway module and adoption of multi-hop communication mechanisms.[37] These mechanisms enhance WSNs lifetime but they are not feasible in real environments with large-scale deployment. Moreover, these mechanisms incur a much higher overhead.

An alternative solution, heterogeneous WSNs, has been explored by scientists to address the large-scale deployment issue associated with constraint-oriented sensor nodes. In these networks, the packets of ordinary nodes are transmitted to the gateway by its cluster head. Heterogeneity in WSNs is accomplished by adopting the following models:

1. The heterogeneous nodes, also known as cluster heads, either have a dedicated powered line or powerful replaceable batteries to prolong their lifetime;

2. A higher bandwidth transceiver is integrated with heterogeneous nodes that has a long range and reliable communication capacity;

3. A much powerful processor and memory module is embedded within a heterogeneous node to perform complex operations compared to an ordinary node.

Besides the lifetime optimization algorithms, various mechanisms are presented in literature to filter out any redundant data and outliers using different data fusion or aggregation approaches.[38] Collaborative signal processing in node environment (C-SPINE) is presented where multi-sensor data fusion technique, particularly among collaborative body area networks, is used to create a hybrid data analysis tool i.e., classification, filtering and a timely dependent integration of data[16]. A network's status based clustering and fusion mechanism was presented in[39]. It is based on an unrealistic assumption that the sensor nodes always generate true values. However, this assumption affects the ratio of generated packets to the transmitted ones. A geographic location-based protocol aggregates data according to the position of a node, with embedded global positioning system (GPS).[40] Its energy consumption statistics were not addressed, particularly, in constrained nodes' scenario. A tree-based algorithm was proposed by Xin et al.,[41] to perform aggregation/fusion activity at different levels of a tree. Similarly, a tree-based data aggregation approach was presented by Kuo et al.,[42] to minimize the cost of data transmission. However, a single packet loss at any level of the tree results in information loss for the whole sub-tree.[43,44] A grid-based data aggregation scheme (GBDAS) divides the experimental area into various cells, where each cell has a sensor node, probably the one with the maximum residual energy capable of fusing neighbor nodes data in addition to its own.[45] A Kalman filter-based approach was used to detect and rectify outliers, i.e., noisy data, generated by a malfunctioning sensor node.[46] In[47], the authors described the potentials of truncated bits procedure that drops extra bits of redundant data without losing any information, in resolving the redundancy issue that is tightly coupled with nodes reside in closed proximity. An energy-efficient data fusion mechanism was proposed in[48] to improve the precision of sensor nodes data, which is updated with the local weighted least square estimated average. A reliability and multipath encounter routing (RMER) approach was presented in[49] that minimizes the energy consumption of WSNs by converging the traffic of event monitoring nodes to a single reliable path. However, convergence is a complex process and any failure of a node residing on a reliable path could drastically affect RMER performance in terms of WSNs lifetime and throughput.

In the literature, uniform traffic spreading and multiple path based routing schemes were proposed to enhance the lifetime of constrained-oriented networks.[50] A traffic aware dynamic routing algorithms is presented to exert light-weighted nodes to mitigate the congestion problem by distributing the excessive packets along multiple routes (preferably idle paths) in WSNs. To achieve this goal, a hybrid virtual potential field is created to force the packets to move across idle paths in the network.[51] An efficient data reporting and reliable transmission method, Reliable Energy Balance Traffic Aware greedy Algorithm in multi-sink WSNs (REBTAM), is presented to track objects in multi sink environment of WSNs. Furthermore, this scheme aims to present a balanced energy consumption model for the resource limited networks. For this purpose, under-loaded nodes are utilized to alleviate the congestion problem (preferably with available resource) in WSNs.[52] cluster based route optimization and load balancing protocol (ROL) is presented to guarantee application specific services with a constraint oriented network. Additionally, an optimization tool for a proper load balancing across the network is presented as Nutrient-flow-based Distributed Clustering (NDC).[53] These approaches are highly efficient to enhance the lifetime and throughput of WSNs, but are unable to control prolonged delays that yields due to the transmission of packets over longer paths. Shortest path and least cost paths are the alternative approaches to multiple path schemes, which are used to transmit data with a minimal possible delay. A secure cost aware routing scheme is presented to address networks lifetime optimization and security issues by using energy balance control and probabilistic based random walking scheme. For lifetime optimization and maximum packets delivery ratio, a non-uniform deployment strategy under the same energy and security requirements is adopted. Finally, a quantitative security analysis of their scheme is presented.[54] A routing algorithm, that has the capacity to estimate the distance between non-neighboring nodes specifically in multihop centralized WSNs, is presented. For this purpose, a global table is created and then a recursive function is used to find all possible paths possibly with minimum hop count between source and destination nodes.[55] Both these techniques do not take into account the sensor nodes' vulnerability factor, also known as criticality.[56] The vulnerability factor is defined as importance of a sensor node from the network connectivity perspective. In WSNs, a sensor node is assumed to be critical or vulnerable if the traffic of a particular portion of the network passes through it or that solely depends on its smooth functioning. The lifetime of these nodes is considerably enhanced if any unnecessary traffic is forwarded through other paths or neighboring nodes, if possible. To address this issue, a local cost function based criticality factor was adapted as part of energy-aware routing protocols designed for WSNs. This factor enables the nodes to route the packets over the most reliable paths.[57] This factor was computed using a well-known technique, i.e., logical network abridgment (LNA)[58]. In addition to sensor nodes criticality, the impact factor of routing paths in the networks was also computed and utilized in the traffic distribution strategy. Most of the packets were routed on the paths having a smaller value of computed impact factor, i.e., usually a path that does not contain critical node(s). The idea of local cost function was extended to global cost function that represents a complete optimal path from source to destination, i.e., sensor node to gateway.[59] The global cost function is a summation ($\sum$) of all local cost functions of sensor nodes residing on that path. The routing paths with the smallest values of global cost function are ideal for the transmission of packets from source to destination, because these paths do not have any critical node. An alternative approach was presented in[14] to enhance the lifetime of WSNs through a vulnerability-based routing protocol. Instead of computing a local or global cost function, this approach merges the hop-count and vulnerability factor of nodes to form an efficient routing scheme. In this approach, a packet is forwarded to the neighboring node having a minimal value of criticality factor. In scenarios, where two or more neighbors have a similar criticality factor, then a random selection approach is adapted. Multiple path based and vulnerability-aware routing schemes substantially enhance the lifetime of WSNs, however, these techniques increase the transmission delay.

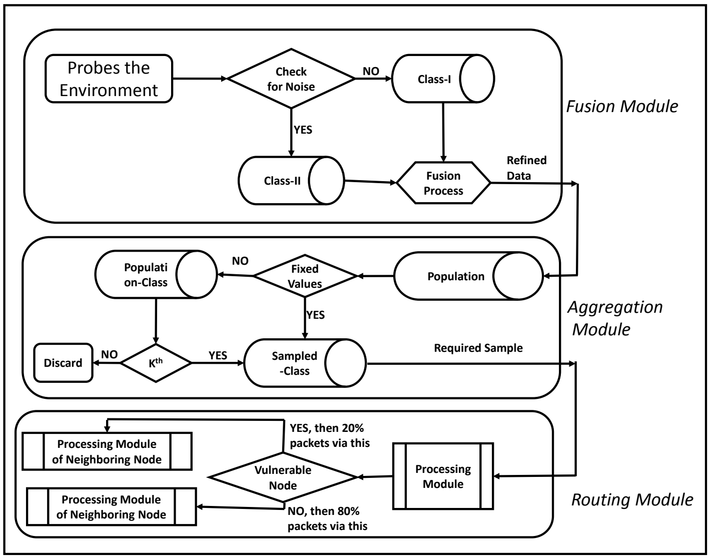

FIGURE 1 Work Flow of the Proposed Techniques on Real Time Data

3 | PROPOSED FUSION, AGGREGATION AND ROUTING IN HWSNS

In this section, we propose several techniques for data fusion, aggregation and vulnerability-based routing as they are critical aspects of heterogeneous WSNs. Fusion and aggregation mechanisms are dedicated to improve the accuracy and precision of the data, collected by the deployed sensors, with strong emphasis on maintaining the overall reliability. A sensor node vulnerability or criticality is defined as the impact of a node on the overall lifetime of a heterogeneous or homogeneous WSN. An efficient utilization of the available power to a vulnerable node leads to longer connectivity of a given network. In order to address these issues, we suggest several algorithms in this section to fuse, aggregate and efficiently transmit real time data as described in Figure. 1. In Section 3.1, we describe a node-level data fusion approach. Section 3.2 describes a hybrid approach to data sampling. Finally, Section 3.3 describes a hybrid shortest path and vulnerability-based routing technique.

3.1 | Node-level Data Fusion in HWSNs

In WSNs, noise generation is tightly linked with the malfunctioning of a sensor that usually happens due to extreme pressure, high temperature, circuit failures and other environmental conditions. In realistic scenarios, it is almost impossible to solve these problems due to remote deployment of the nodes. Therefore, research activities are focused on precise data collection from resource-constrained WSNs without considering the aforementioned challenges. Data fusion is a well-known concept in WSNs, particularly heterogeneous networks, to enhance accuracy of the collected data and is usually performed through in-network processing or in-network data fusion. Data fusion presents a complete view of data collected by sensor nodes, either at ordinary node level or at cluster head (CH) level. One of the most critical issue associated with data fusion is how to increase accuracy of the sensed data with minimum information loss. In this section, a simplified and robust noise detection technique is described where a fusion process is performed at ordinary node level that not only enhances data accuracy, but also saves considerable power by avoiding transmission of the falsified data, i.e. noise. Our proposed approach uses two classes of the same size for temperature data, i.e., class-I for holding accurate data and class-II for noisy data. Initial value of class-I is defined only once at the deployment stage of the nodes. For a temperature sensor, this value is set according to the environmental conditions, i.e., 35^oC to 40^oC, during summer in Pakistan. Once the nodes are fully functional, each temperature sensor probes the environment after a pre-defined interval of time and communicates their readings with microcontroller of the board for testing the accuracy. The microcontroller compares the most recent, i.e., latest, reading of the temperature sensor with previously stored accurate data. If their similarity index is higher, i.e., difference between current value and previously stored value is less than the pre-defined variation or fluctuation range of temperature, then it is stored in class-I. Otherwise, it is stored in class-II with a pointer to its previously stored accurate value and a reserved location in class-I as well. These pointers are extremely useful in replacing noisy data of class-II with their most correlated accurate data of class-I. For class-I, the fluctuation range is 12^{o}C that is determined through the deployment of temperature sensors (three in our deployed testbed) in an open air environment. These three nodes collect their readings over a period of 20 days. Additionally, the fluctuation ratio of temperature in Pakistan during different seasons is thoroughly studied along with valuable suggestions of well-known scientist in agricultural and meteorological sector. The proposed node level data fusion algorithm is presented in Algorithm 1. In Figure. 2, the noise scenarios in real-time dataset collected through our deployed WSN in an orange orchard is shown[60]. In this real testbed, each temperature sensor generates data continuously after a pre-defined time interval, i.e., three seconds in this case. The first value received at time interval 3s is matched with a reference value, i.e., 36^{o}C in this case. Their similarity index is higher as their difference is less than the pre-defined threshold value of 12^{o}C and is stored in class-I. Similarly, the value at time interval 6s, i.e., 34.45^{o}C, is matched with previously stored accurate value, i.e., 34.65^{o}C, and stored in class-I as their difference is lower than the threshold value. A relatively different scenario arises at time interval 12s, where the currently received value is 0^{o}C and its difference from the previously stored accurate value, i.e., 35.33^{o}C, is far greater than the allowed fluctuation range, i.e., 12^{o}C. Therefore, it is identified as noise and stored in class-II. Moreover, an empty location in class-I is reserved at the same timestamp, i.e., 12s. This procedure is applied repeatedly to the upcoming values until the last reading, i.e. 36.99^{o}C, is processed. This is because the sensor nodes are programmed to start the transmission activity after every 20 readings. However, empty locations in class-I for noisy data are replaced with the average of class-I values that are calculated by using Equ. 1.

Algorithm 1 Proposed Node-level data fusion algorithm

```
**INPUT:** Sensed Data values
**OUTPUT:** Return Noise Free Data
Class-I ← **empty**
Class-II ← **empty**
Counter ← 0
Pr-Value ← $37^oC$
Cur-Value ← Most Recent Sensor Reading
**for** every sensed value
  **if** Difference (Cur-Value, Pr-Value) < Threshold-Value        ⊳ Using euclidean or any distance measure
    $Class - I_i$← Cur-Value
    Pr-Value← Cur-Value
  **else**
    $Class - II_j$← Cur-Value        ⊳ Ambiguous value i.e., may be noise of accurate
    i← i + 1 [reserving location in class-I]
  **endif**
  Compute Avg = $\frac{\sum_{i=0}^{n}(Class-I_{val})}{n-\sum_{i=0}^{m}(1)}$
**endfor**
**if** Value (Class-$I_i$) = NULL
  Class-$I_i$ ← Avg        ⊳ In case of noise, average of the previous and currently collected value is computed
**endif**
**Return Noise**
```

$$Avg_{Val} = \frac{\sum_{i=0}^{n}(class - I_{val})}{n - \sum_{i=0}^{m}(1)}. \tag{1}$$

In this equation, $\sum_{i=0}^{m}(1)$ represents the count of noisy data stored in class-II and n is the length or size of class-I, i.e., 20 in this case. The data in class-I is fine-tuned at the corresponding timestamps of 12s, 30s, and 57s, respectively. The graphical representation of with-noise and without-noise dataset is shown in Figure. 2 and Figure. 3, respectively. Furthermore, a similar procedure is repeatedly applied to the collected data of different sensors integrated with a particular board.

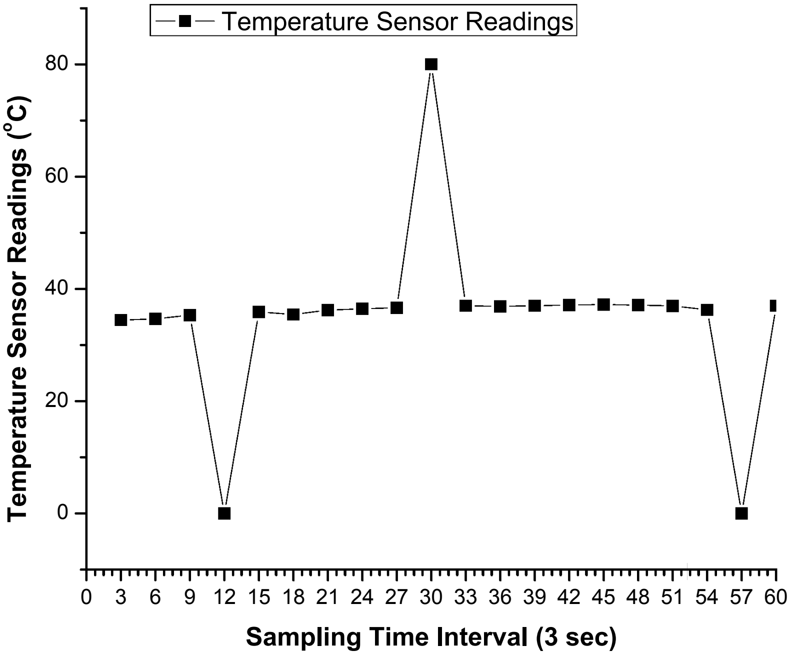

FIGURE 2 Noise scenario in real-time dataset [Outliers represent the noise]

3.2 | Hybrid Sampling Technique in HWSNs

In heterogeneous WSNs, data aggregation or in-network processing is performed by cluster heads and is useful to control redundant data generated by sensor nodes residing in closed proximity. This approach enables ordinary sensing nodes to concentrate on their primary tasks of sensing and reporting the data and let the intended cluster heads perform data processing. Although, this approach resolves the redundancy issue to a greater extent, but at the same time, it underestimates other important issues such as higher congestion rate, transmission of redundant data, in-efficient power utilization etc. Most of these issues are resolved if aggregation mechanisms are localized, i.e., performed by individual source/ordinary nodes, rather than a centralized approach. Moreover, aggregation techniques become more realistic in resource-constrained HWSNs if a two-tier approach is implemented. In this section, an enhanced systematic sampling-based aggregation algorithm is presented that resolves most of the aforementioned issues.

In HWSNs, systematic and simple random sampling techniques are based on the idea of probability, where every unit in a sampling universe has a similar probability value. A sample is a small set of selected values that are based on their assigned probabilities[61]. Each sample is a representation of the entire population or dataset. A modified version of the simple probability based sampling technique that is fine-tuned according to the dynamic nature of WSNs data set is presented to retain maximum information in smaller data set. Initially, the sampling interval is calculated using Equ. 2.

$$K = \frac{N+m}{n}. \tag{2}$$

where, n is the sample size, N is the population size and m is the ratio or size of the fixed selected values in the desired sample. In our scheme, n is 20 and N is 60, respectively. A random starting point is selected and 80% of the data in a sample, that is to be transmitted, is selected randomly according to their K^{th} position in the overall population with zero probability of fixed selected values. The process for the selection of the remaining 20% of the sample data is fixed, i.e., it starts from the initial first value of a sensor and then every value sensed after the pre-defined time interval of 12 seconds. These values are selected based on a fixed interval, i.e., every $12^t h$ value or reading (in this case), i.e., **0** 1 2 3 4 5 6 7 8 9 10 11 **12** 13 14 15 16 17 18 19 20 21 22 23 **24** 25 26 27 28 29 30 31 32 33 34 35 **36**, and so on. The ratio of fixed and random selection of units in a sample leads to a robust sample that mimics the real deployed scenario of HWSNs, i.e., probing the environment after fixed interval of time and

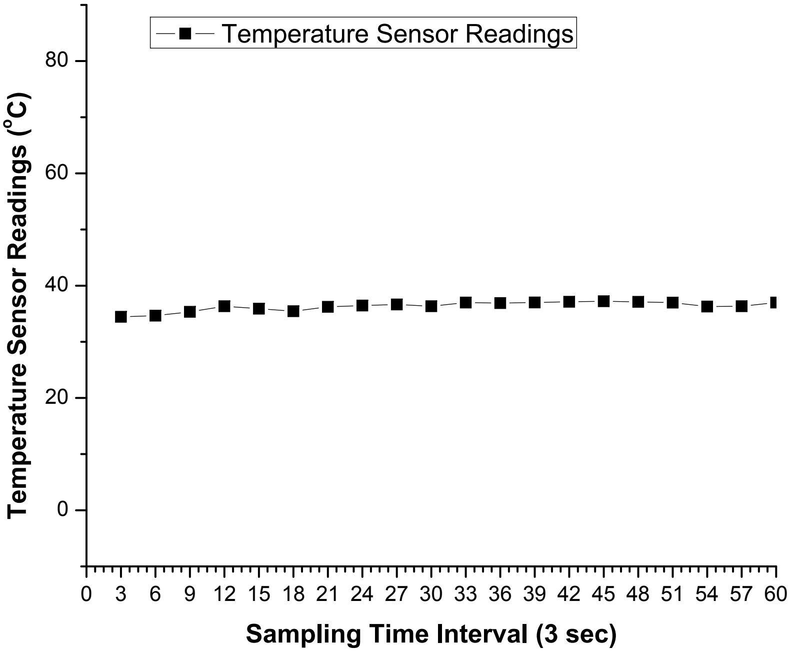

FIGURE 3 Noise removal in real-time dataset

randomness results in an un-biased sample. The population mean (μ) of our proposed sample technique is shown in Equ. 3, that is a representation of the average number of elements in a sample from a particular unit.

$$\mu = \frac{\sum_{i=0}^{n-m} y_i + \sum_{j=0}^{m} y_j}{n}. \tag{3}$$

where, y_i and y_j are values from each unit in a sample and n describes the sample size. This is an un-biased estimator of the values, i.e., y_i, because every unit has been assigned with equal probability except values that meet the criteria of fixed selection. The standard deviation (σ) of our sampling technique is given by Equ. 4, that describes the measure of variability or spread of the hits, in a given sample with a defined variable.

$$\sigma = \sqrt{\frac{\sum_{i=0}^{n-m}(y_i - \mu)^2 + \sum_{j=0}^{m}(y_j - \mu)^2}{n-1}}. \tag{4}$$

Here, n is the sample size and m represents the number of values whose selection criteria is fixed in the proposed technique, i.e., first reading and then every 12^{th} reading of an ordinary sensor node in heterogeneous WSNs. Its variance is described in Equ. 5, that addresses the average squared distance from the mean.

$$\sigma^2 = \frac{\sum_{i=0}^{n-m}(y_i - \mu)^2 + \sum_{j=0}^{m}(y_j - \mu)^2}{n-1}. \tag{5}$$

where, σ is used to represent an un-biased entity. Additionally, standard deviation (σ) and variance (σ^2) are useful in differentiating accurate data values from outliers.

The proposed sampling-based data aggregation algorithm is described in Algorithm 2. Initially, the sensed data is divided into two classes, i.e., sampled-class and population-class. The former contains fixed values, 5 in this case, that are stored after a fixed time interval. The latter, on the other hand, holds the remaining sensed values, 55 in this case. Fifteen more values of the desired sample data is selected from population-class using a probabilistic approach. Initially, a random value is selected, i.e., value at 8^{th} position in population-class, and is stored in sampled-class with fixed values at the next available location, 6^{th} position in this case. The next selected value of the sample resides in population-class at 11^{th} location, as the value of sampling interval

Algorithm 2 Proposed systematic sampling-based data aggregation algorithm

```
INPUT: Sensed Data values
OUTPUT: Return a Sample Representing the Whole Data Set
Sampled-class ← empty
Population-class ← empty
Counter ← 0
Cur-Pkt ← Most Recent Sensor Reading
for every sensed value                    ▷ This Step is repeated for every sensed value of a sensor node
  if Counter ∈ {0, 12, 24, 36, 48}        ▷ It represents class of fixed values that is 16% in this case
    $Sampled - class_m$ ← Cur-Pkt         ▷ Adding Value to the Sample
    Counter ← Counter + 1
  else
    if Counter = 59   ▷ Completion of Sampling interval i.e., 60 where 20 values (5 fixed + 15 randomly) are transmitted
      $Population - class_n$ ← Cur-Pkt
      Counter ← 0
      Goto Step-19                        ▷ Sampling interval of a node expires i.e., 60 sec
    else
      $Population - class_n$ ← Cur-Pkt
      Counter ← Counter + 1
    endif
  endif
endfor
Compute Sampling Interval $K = (N + m)/n$
Randomly Select Initial Value at (1 to $10^{th}$ location in Population-class)
while m < 20                              ▷ 15 random values are selected from the population class
  $Sampled - class_m$ ← Value at $K^{th}$ position
  m ← m + 1
endwhile
Return: Sample that Represents the Whole Data Set
```

K is equal to three. For the remaining values of the desired sample, the same process is repeatedly applied. The probabilistic selection process is further explained with an example as presented in Table 1. In this table, the data and their locations are kept constant for simplicity and understandability. Otherwise, the data is shuffled and then the next value at the K^{th} position is selected accordingly. In Table 1, the encircled values represent a sample of the population, i.e., 33, 34, 35, 35, 32, 33, 32, 31, 34, 35, 34, 36, 35, 33 and 33. Once it is completed, then a merged sample of values, i.e., a combination of fixed and randomly selected data, is forwarded to the intended cluster head through an optimistic path using the routing technique of Section 3.3.

TABLE 1 Probabilistic Selection of Sample from the Population with value of $K = 3$

1^{st} Position	2^{nd} Position	3^{rd} Position	4^{th} Position	5^{th} Position	6^{th} Position	7^{th} Position	8^{th} Position
30	31	32	33	34	35	34	[✓ 33]
33	[✓ 34]	34	35	[✓ 35]	36	36	[✓ 35]
33	[✓ 32]	32	32	[✓ 33]	33	32	[✓ 32]
30	[✓ 31]	32	33	[✓ 34]	35	34	[✓ 35]
33	[✓ 34]	34	35	[✓ 36]	36	35	[✓ 35]
33	[✓ 33]	32	32	[✓ 33]	33	32	32

The proposed sampling technique is quite useful in heterogeneous WSNs, as it focuses on how to enhance an ordinary node's lifetime rather than the cluster head. In HWSNs, it is the ordinary node that results in disconnection of the networks. In addition to the enhanced nodes' lifetime, the proposed systematic sampling-based data aggregation algorithm plays a vital role in controlling the congestion problem, which is closed linked with dense WSNs. Our proposed algorithm reduces the overall transmission ratio of an ordinary node from 60 to 20, i.e., 66.67% reduction in packet transmission rate. However, our approach is not applicable in applications that have an extremely small or zero transmission delay such as, military surveillance, intensive care, intrusion detection, remote mining etc.

3.3 | Shortest Path and Vulnerability-based Routing

A sensor node's vulnerability, i.e., its importance in term of long-term network connectivity, plays a vital role in prolonging the lifetime of a hetrogeneous WSN. A detailed discussion on the computation of sensor node vulnerability was presented in[14]. The proposed work spreads the traffic across multiple paths using node's vulnerability-aware routing. A node vulnerability is calculated using Equ. 6.

$$V_i = \frac{N_B}{N_A} * \frac{L_B + C}{L_A + C}. \tag{6}$$

where, V_i is the vulnerability of i^{th} node, N_B is the number of nodes before removal of a particular node, N_A is the number of nodes after its removal, L_B is the levels in WSN before removal, L_A represents levels in WSN after removal and C is a constant number. The idea of levels and vulnerability computation is depicted in Figure. 4 and was discussed in our previous work.[14] Here, level 0 represents an actual WSN. In labeling, 0 represents the level of WSN and 1, 2, 3, ..., 11 represent the nodes.

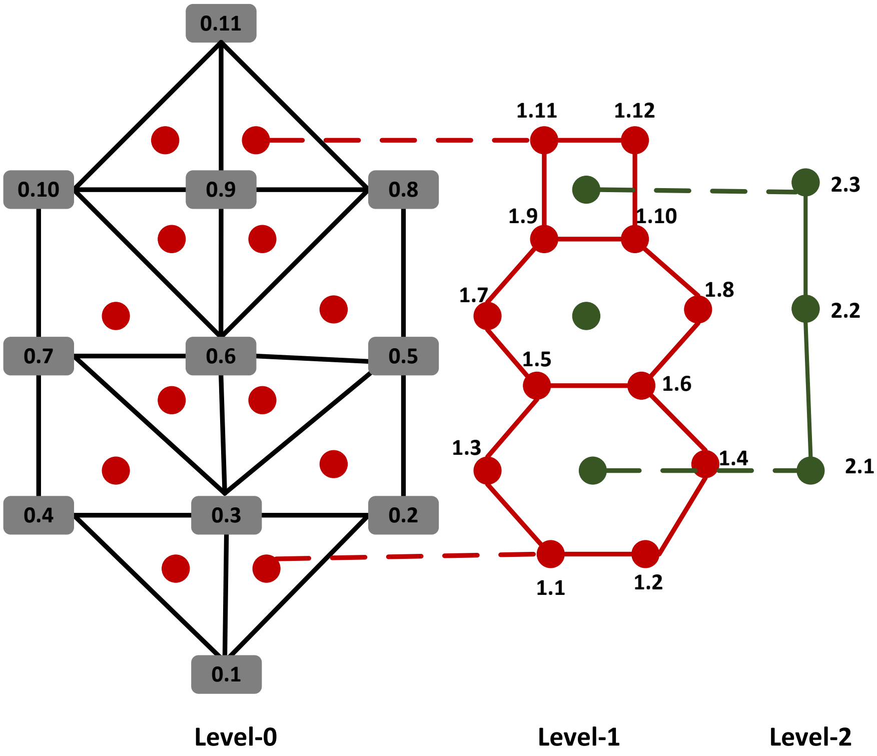

FIGURE 4 Calculation of Levels and Sensor Nodes Vulnerability Using Logical Networks Abridgment (LNA) Technique

The importance of a node from network connectivity and lifetime perspectives is described by its vulnerability value, i.e., V_i. A node's importance is directly proportional to V_i, i.e., higher the value of V_i, greater will be the risk of losing the network connectivity with the demise of that particular node. Vulnerability-aware routing considerably prolongs the lifetime of a heterogeneous WSN by decreasing the traffic on vulnerable paths. However, at the same time, it introduces extra delay in the delivery of packets from source towards the destination, as depicted by Equ. 7.

$$D_{avg} = \sum_{i=0}^{n} T_i(VP) - \sum_{i=0}^{n} T_i(SP). \tag{7}$$

Here, D_{avg} is the average delay produced due to the spreading of traffic across the longest paths in order to increase lifetime of the vulnerable nodes in heterogeneous WSNs. In this equation, $T_i(VP)$ and $T_i(SP)$ represent the times to transmit a packet from source towards the destination through vulnerability-aware routing and shortest path routing techniques, respectively. The vulnerability-aware technique delays every packet, sent from source towards the destination, by different time intervals. The value of average delay ratio for different packets from the same source to a single destination is higher for the longest routing paths in heterogeneous WSNs. Due to packet delay, vulnerability-aware routing protocols are limited to a specific class of traditional WSN application areas. Apart from the packet delay issue, vulnerability-based routing techniques perform exceptionally well in prolonging the lifetime in heterogeneous as well as homogeneous networks.

A hybrid protocol, i.e., a composite of vulnerability and shortest path based routing techniques, is required to address the issues associated with both protocols and incorporates their best features. As a result, we have extended this approach by integrating the shortest path algorithm with vulnerability-aware routing protocol to form a robust hybrid routing mechanism. The proposed mechanism prolongs the lifetime of heterogeneous WSNs through node vulnerability and reduces delay in packets transmission by utilizing the shortest path, whenever possible. Instead of a single path as in the case of shortest path algorithm or multiple paths as in the case of uniform traffic spreading techniques, sensor nodes hold the vulnerability and hop-count value of two neighboring nodes in our scheme that have the shortest path to the base station or cluster head. In a realistic scenario, every sensor node stores two optimum paths. The shortest path, i.e., path-1, and the second shortest path, i.e., path-2, are composed of three and four hops to a given cluster head, respectively. The optimum paths are computed using Equ. 8.

$$P_{opt_i} = hop_{count} * max(\frac{E_r}{E_s}). \tag{8}$$

where, P_{opt_i} represents the optimal path from a sensor node to its cluster head in heterogeneous WSNs and hop_{count} is the number of hops to reach the given cluster head. Every node in heterogeneous WSNs hold two such paths with minimum value of hop-count, i.e., path-1 with the smallest value and path-2 with second smallest value. E_r and E_s represent the residual energy and starting energy of a node, respectively. The traffic is distributed uniformly by providing equal and similar probabilities, i.e. half, to these paths and the neighboring nodes consume approximately similar battery power after pre-defined time intervals. The uniform distribution is possible only if vulnerable nodes (N_i) do not resides on these paths. However, if one of the optimal paths contain one or more N_i, then the traffic distribution strategy is biased i.e., minimizing the frequency of packet transmission on this path. The path selection model, with required biased distribution function, is represented by Equ. 9.

$$P_{best_i} = P_{opt_i} - \sum_{i=0}^{n} w_i(N_v). \tag{9}$$

where, P_{best_i} is the node having the minimum value of N_v and maximum residual energy, i.e., an ideal scenario for the successful and reliable transmission of packet. In heterogeneous WSNs, high power nodes are considered more reliable. In this equation, w_i is the weightage assigned to the vulnerable nodes and its value ranges from 30% to 80% and may vary based on the application requirements. In case, when both the neighboring nodes are vulnerable, a uniform distribution strategy is more suitable and more realistic than a probabilistic strategy. Similarly, when both the neighboring nodes are not vulnerable, uniform packet distribution is preferred. Our proposed hybrid technique not only prolong the lifetime by using a consistent traffic distribution scheme among neighboring nodes with higher residual energy and similar class of vulnerability values, but it also avoids the paths with vulnerable nodes. At the same time, it also addresses the transmission delay problem associated with vulnerability-based routing technique by integrating the benefits of the shortest path algorithm with vulnerability-aware routing.

To understand the proposed technique working phenomena, a portion of a heterogeneous WSN is taken as an example in Figure. 5. In this figure, the nodes are labeled with their vulnerability values. These values are calculated only once after the deployment. Assume that an ordinary node has the capacity to transmit 501 packets with its available on-board battery power and the heterogeneous WSN's lifetime model is tightly coupled with disconnection of the first node. Consider an event occurs in the sensing range of nodes A, B and C. These nodes sense the event continuously after pre-defined time interval of one second, and are eager to transmit it to the cluster head as long as it remains in their area, i.e. 500 sec. The shortest paths of these nodes have a common node F that rapidly depletes its available power due to a higher traffic load, i.e., it has forwarded approximately 166 packets of all source nodes. The source nodes have enough residual energy and are able to transmit more packets, i.e., 335, but the network is disconnected as node F depleted its available power. Each of these source nodes use two optimum paths instead of one shortest path, and further enforce a biased traffic distribution strategy, i.e. minimum packets on vulnerable path. Biased strategy is used in those situations where one of the routing path contains vulnerable nodes. The traffic spreading strategy is make biased by assigning more weightage to the ordinary path (nodes), in this case 80%, and slightly minimum weightage to the vulnerable path, i.e., 20%. This strategy enhances the network lifetime with an acceptable range of overall packet delay and enables every source node to transmit their desired packets, 500 in this case, successfully to the cluster head. The vulnerable node F still has the residual energy of transmitting approximately 71 packets.

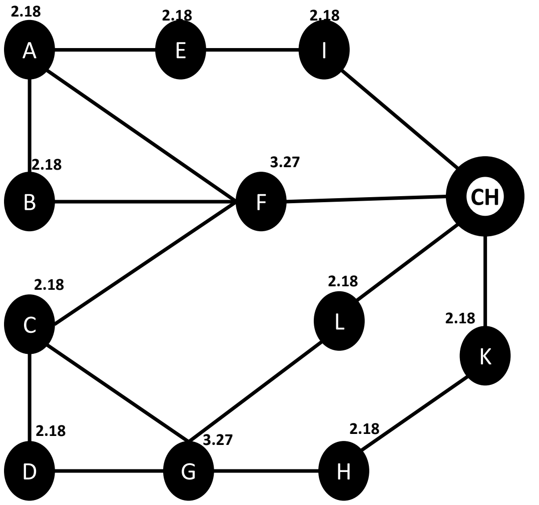

FIGURE 5 A heterogeneous WSN with computed nodes vulnerability values

In case when both the paths do not have vulnerable nodes then a uniform traffic distribution strategy is adopted. In Figure. 5, a node G has two optimal paths to the cluster head, i.e., through L and HK, and both have no vulnerable nodes. Therefore, half the packets are transmitted through L and half are routed through HK. Similarly, if both the paths contain vulnerable nodes, then a uniform distribution strategy is preferred using a biased distribution strategy. This scenario is described in Figure. 5 by node C whose optimal paths towards the cluster head are passed through F and G, both vulnerable nodes. The performance of the proposed technique in term of lifetime is exceptionally well in comparison to its counterpart algorithm, i.e., the shortest path algorithm, with a small compromise on the packet delay that arises when the packets flow on marginally longer paths.

4 | PERFORMANCE EVALUATION

In this section, we evaluate the overall performance of the proposed algorithms by comparing them with field-proven techniques such as, (a) the shortest path algorithm (b) outliers detection (OD) (c) vulnerability-based routing (d) pattern anomaly value (PAV) (e) MPAV and (f) rare pattern drift detector (RPDD) algorithms. These algorithms were implemented in OMNET++[62], an open source discrete events simulation environment. We run the experiments using similar topologies, transmission power, on-board batteries, nodes, and a single-gateway in heterogeneous WSNs. In our simulation set-up, we modeled a heterogeneous

TABLE 2 Simulation Values.

Parameters	values
WSN deployed area	500m × 500m
Sensor node (ordinary)	95, 190, 285, 380
Cluster heads (CH)	5, 10, 15, 20
Base station	1
Initial energy (E_s)	1,150, 2,300, 6,600, 13,000 mAh
Residual energy (E_r)	$E_s - E_{consumed}$
Transceiver energy (T_i)	5,026 mA
Transmission range (T_r)	100m
Initial hop count (HC) of CH (T_r)	0
Initial HC of an ordinary node (T_r)	∞
Maximum distance between nodes	100m

WSN with n randomly distributed sensor nodes comprised of 95% ordinary nodes and 5% cluster head nodes in an area with dimension 500m × 500m. Various characteristics and parameters of the experiments are shown in Table 2[63].

Every cluster is assumed to have equal number of ordinary nodes, as described in[64]. The ordinary nodes either communicate: (a) directly with the cluster heads or (b) through other nodes if the cluster heads do not resides in their transmission range. Additionally, the power consumed by a particular node to transmit a packet to its cluster head and to its neighboring nodes is assumed to be similar to a wasp-mote agriculture pro-board developed by Libelium[1], i.e., 5,026 mAh - the initial energy of an ordinary board. In our simulation setup, this is equivalent to the actual power provided by an on-board battery of a wasp mote agricultural board, i.e., 1150, 2300, 6600 and 1300 mAh.

We use several metrics to evaluate the performance of our proposed techniques. These metrics include energy consumption, throughput, computational time, accuracy and nodes drop ratio. The energy consumption is computed in term of network lifetime, throughput in term of number of packets received at the sink node and computational time in term of algorithm's complexity.

4.1 | Results and Discussion

In this section, we explain the results obtained for our three strategies as discussed in Section 3. To keep the contents easy to follow, the discussion is divided into three sections.

4.1.1 | The Hybrid Routing Scheme

In heterogeneous WSNs, the sensor nodes rely on their on-board batteries that are not always rechargeable. Therefore, energy-efficient algorithms need to be designed to enhance the lifetime of these networks. The network lifetime is an important parameter, in both heterogeneous and homogeneous WSNs, that is used to evaluate the performance of an algorithm in real environmental conditions. The network lifetime is defined as the time when the very first node exhausts its overall power or when a node consumes its battery. The proposed hybrid algorithm, described in Section 3.3, has a far better lifetime than its competitors, i.e., the shortest path algorithms and vulnerability-based routing techniques as shown in Figure. 6. The load balancing criteria for the proposed algorithm is that 20% of the generated packets are transmitted over a vulnerable optimal path and 80% packets over a non-vulnerable path. Additionally, in scenarios, where both optimal paths are either vulnerable or

[1] http://www.libelium.com/products/waspmote/

non-vulnerable, then a uniform traffic distribution strategy, i.e., 50%, is adopted. The proposed technique enhances heterogeneous WSNs lifetime from 20% to 50% over its competitive algorithms as shown in Figure. 6. Four different on-board battery levels are used in simulation setup to test the performance of these algorithms, because a wasp-mote agricultural board has different power options as described in start of this section.

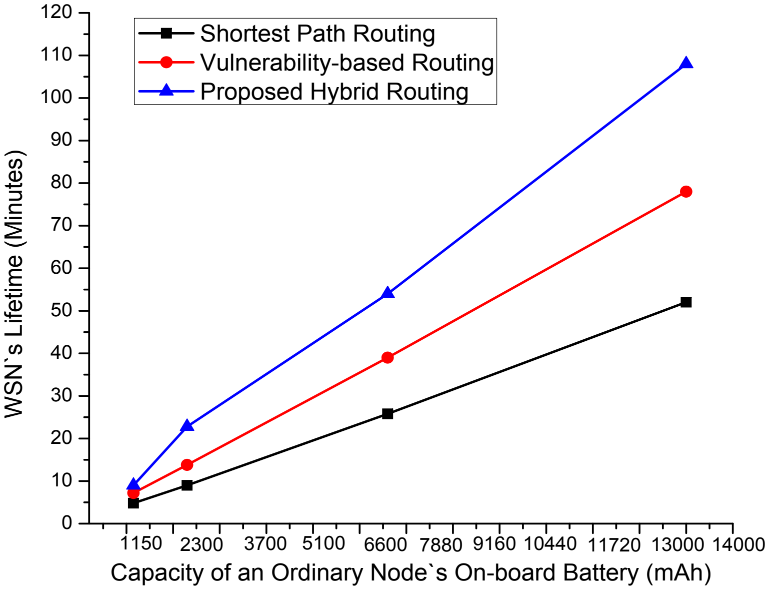

FIGURE 6 Lifetime comparison of the proposed hybrid algorithm, with traffic distribution ratio of 20% traffic load on an optimal path and 80% load on a non-vulnerable optimal path, with the shortest path and vulnerability based routing algorithms [higher values are "best"]

The performance of proposed algorithm is tested in different realistic scenarios by assigning different weightage to the optimal and vulnerable paths in a heterogeneous WSN. A uniform traffic distribution strategy for the proposed algorithm, i.e., 50% load on vulnerable optimal path and 50% on non-vulnerable optimal path, ensures to prolong the network lifetime from 12% – 27% against its rivals algorithms, i.e., the shortest path and vulnerability-based routing techniques, as shown in Figure. 7. Moreover, the proposed scheme is well-suited in different application areas of WSNs by simply tuning its weightage factor, accordingly. In WSN's application areas where packets delivery within a shortest stipulated time is preferred over the network's lifetime, then a 100% weightage to the shortest optimal path needs to be assigned in the proposed scheme. In lieu scenarios, the weightage factor assigned to the vulnerable and non-vulnerable optimal paths is adjusted accordingly to keep the network operational for its maximum possible duration. We observed that our approach becomes similar to the shortest path algorithm if the traffic distribution strategy is completely biased, i.e., 100% load on the most optimal path.

In addition to the network lifetime, throughput is another important evaluation factor that is defined as the total number of packets received at the gateway or destination. If an algorithm prolongs WSNs lifetime up to its maximum possible limit, but its throughput is very low, then such an algorithm is of no use. Therefore, an efficient algorithm needs to prolong WSNs lifetime and at the same time, needs to provide maximum throughput. In this scenario, the proposed scheme outperforms its rival algorithms as shown in Figure. 8. The proposed algorithm reports a maximum throughput by assigning a lower weightage to the vulnerable path, i.e.,20%, and a higher weightage to the non-vulnerable optimal path, i.e. 80%. The Shortest path algorithm generates a maximum throughput only when the source nodes do not have any common node on their routes, a scenario that is not possible particularly in randomly deployed WSNs. Similarly, in the vulnerability-based technique, throughput of the network is inversely proportional to the vulnerable nodes. A network with a smallest or no vulnerable nodes has the highest throughput, whereas,

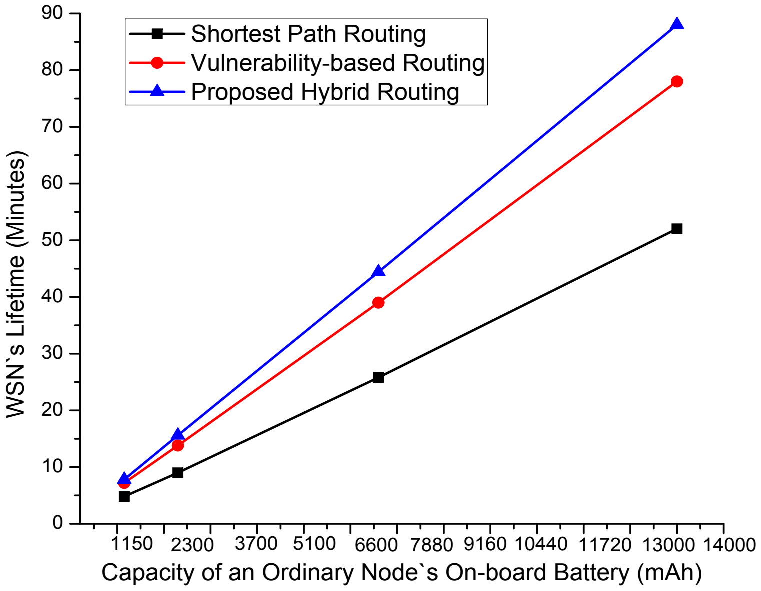

FIGURE 7 Lifetime comparison of the proposed hybrid algorithm, with traffic distribution ratio of 50% traffic load on an optimal path and 50% load on a non-vulnerable optimal path, with the shortest path and vulnerability based routing algorithms [higher values are “best"]

a network with the highest vulnerable nodes has a lowest throughput. Our proposed algorithm covers both these scenarios by assigning a different weightage factor to the vulnerable and non-vulnerable optimal paths. A realistic simulation scenario of the proposed algorithm’s uniform traffic distribution strategy is depicted in Figure. 9, that shows the stability of our scheme. Additionally, the proposed approach is tested on different weightage factors and we observed that its performance, in terms of throughput, is compromised only if the weightage factor is fully biased, i.e., either 100% on a non-vulnerable optimal path or on a vulnerable optimal path.

4.1.2 | Node-level Data Fusion Technique

In heterogeneous WSNs, ordinary sensor nodes have limited resources, i.e., on-board battery, processing power and communication. Due to these limitation, the data sensed by these sensors are highly susceptible to noise or outliers that usually occur due to the malfunctioning of an embedded sensor or interference with another node’s data. Reporting noisy data to a sink, gateway or cluster head not only consume the limited resources but also affects the accuracy of a real-time system. Therefore, a mechanism is needed to detect noisy data prior to its transmission.

For real-time data, the worst case complexity of the proposed fusion algorithm is O(1), whereas pattern anomaly value (PAV), MPAV, and rare pattern drift detector (RPDD) algorithms have complexities of $O(n^2)$, $O(n)$ and $O(n^2 + n)$ respectively. Similarly in case of static data set, the proposed fusion scheme worst case complexity is $O(n)$, where n represents size of the dataset. The performance of our proposed noise detection algorithm, i.e., time to refine the raw data, is depicted in Figure. 10. This figure clearly shows that our proposed technique outperforms the existing algorithms, i.e., outliers detection (OD), PAV, MPAV and RPDD. This is because these algorithms are designed for general-purpose systems, whereas, our proposed algorithm is specially designed for sensor nodes of heterogeneous WSNs. The accuracy ratio of these algorithms is presented in Figure. 11, that shows the performance of our proposed technique against the existing schemes. The RPDD algorithm has a slightly higher accuracy in comparison to our technique, however, its complexity is much higher, as shown in Figure. 10. The higher complexity forbids the implementation of RPDD on an ordinary node of a heterogeneous WSNs. The main issue associated

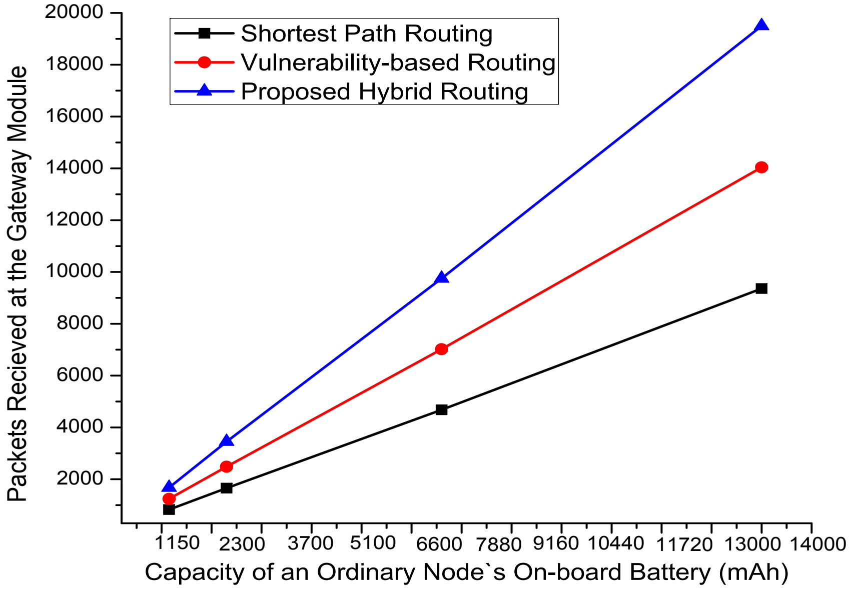

FIGURE 8 Throughput of the proposed hybrid algorithm, with traffic distribution ratio of 20% traffic load on a vulnerable optimal path and 80% load on a non-vulnerable optimal path, with the shortest path and vulnerability based routing algorithms [higher values are "best"]

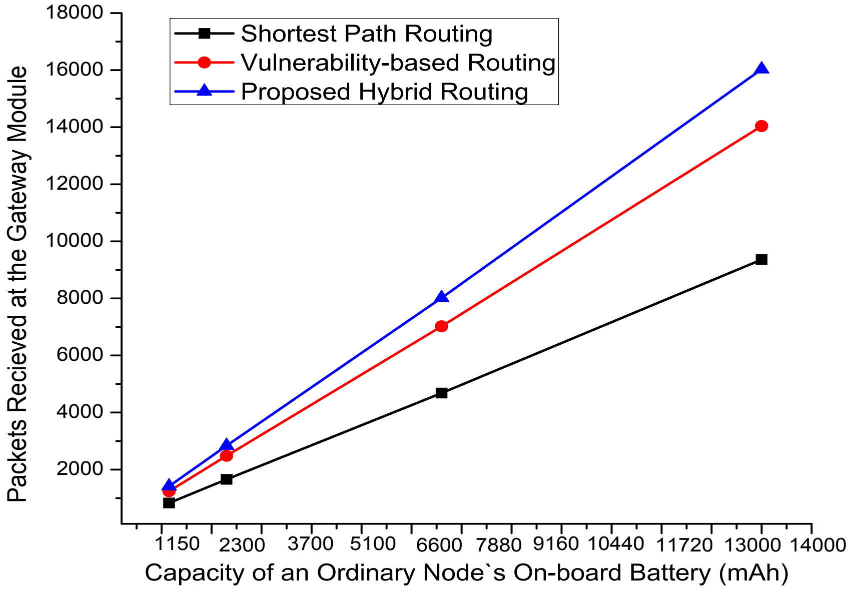

FIGURE 9 Throughput of the proposed hybrid algorithm, with uniform traffic distribution strategy i.e. 50% load on both vulnerable and non-vulnerable optimal path, with the shortest path and vulnerability based routing algorithms [higher values are "best"]

with OD algorithm is its susceptibility to multivalued noise that arises due to its dependence on local average. Our proposed node level data fusion algorithm solves this issue with the application of a global average rather than a local average. These algorithms were tested on a real-time dataset obtained from our deployed WSN, as described in Rahim et al.[60] .

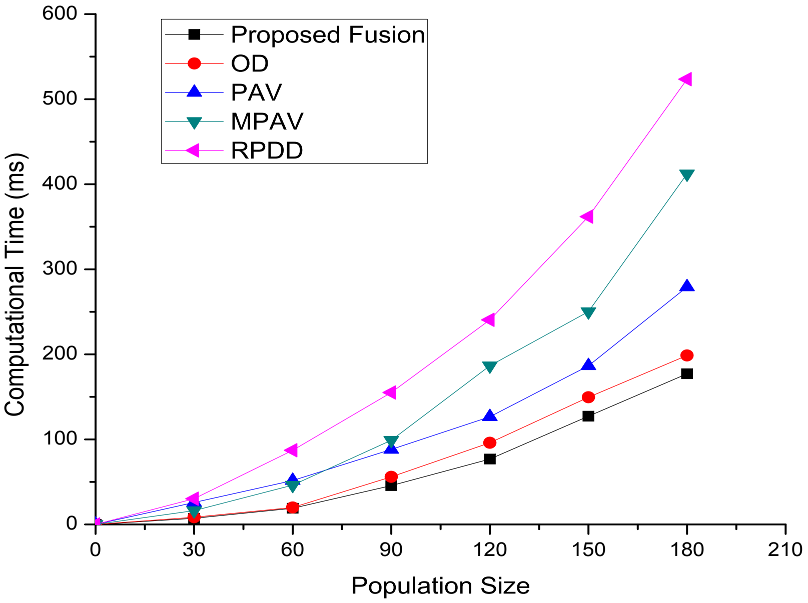

FIGURE 10 Running time comparison of the proposed node level data fusion algorithm with “OD", “PAV", “MPAV" and “RPDD" algorithms over different size of the populations that is 30 – 180 packets [lower values are “best"]

4.1.3 | The Hybrid Sampling Technique

Redundancy is one of the major issues associated with a densely deployed heterogeneous WSN and has accomplished considerable attention from the research community. Data aggregation is common tool that is used to minimize the redundant data, generated by nodes resides in closed proximity, both locally and globally. Data aggregation is directly proportional to the lifetime of heterogeneous WSNs, increasing its ratio ensues in prolonged networks lifetime. The proposed hybrid sampling algorithm aggregates the collected data of an ordinary node in heterogeneous WSNs and forwards a sample, a subset that is used to represent the entire population, to its destination i.e. CH or gateway. A comparative study of how the network's lifetime is enhanced by embedding the proposed hybrid sampling algorithm with routing techniques, as described in Section 4.1.1, is presented in Figure. 12. We observed that an embedded node level data aggregation technique enhances the lifetime of a heterogeneous WSN from 40.0 – 66.67%. Data aggregation is an effective technique that is used to reduce the transmission ratio of a particular node in a network and hence improving its lifetime.

Additionally, the proposed aggregation technique is helpful in controlling the congestion problem, that is closely linked with the dense deployment of sensor nodes, by reducing the packets transmission ratio. This ratio is controlled, if an ordinary node drops a certain proportion of its collected data and forwards or transmits a sample of it. The drop out packets ratio of different nodes, in proposed data aggregation technique, is shown in Figure. 13, where 64 and 0 are the highest and lowest drop-out ratio of ordinary nodes respectively. The difference in nodes drop-out ratio is due to the fact, that a random selection criteria was set for the source nodes in heterogeneous WSNs. Moreover, the worst case complexity of the proposed hybrid sampling technique

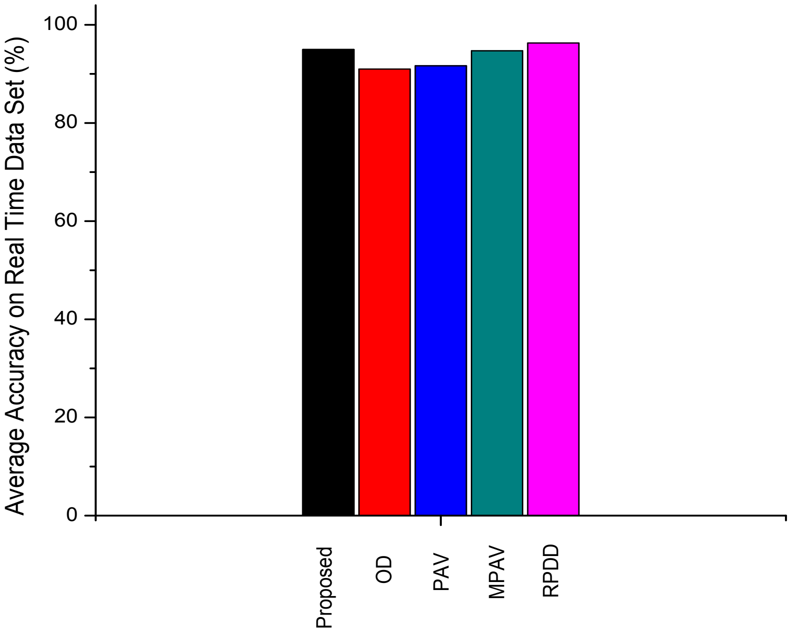

FIGURE 11 Accuracy comparison of the proposed node level data fusion algorithm with “OD", “PAV", “MPAV" and “RPDD" algorithms over different size of the populations [higher values are “best"]

is $O(n + m)$, where n represents size of the data set or collected data by the deployed sensor nodes and m is the randomly selected initial values as described in subsection 4.1.1.

5 | CONCLUSION AND FUTURE WORK

In this paper, a simplified data fusion algorithm was presented to resolve noise issue that is generated either by a malfunctioning sensor node or by harsh environmental conditions at ordinary nodes. These nodes are usually at level-1 in heterogeneous WSNs. Moreover, an energy-efficient hybrid data aggregation algorithm that is a hybrid of fixed and random sampling techniques, was introduced that resolves numerous challenges faced by ordinary nodes of heterogeneous WSNs. Our evaluation suggest that the proposed data aggregation algorithm not only prolonged WSNs’ lifetime, but at the same time, controlled the congestion problem by minimizing the ratio of transmitted packets. The simulation results verified the outstanding performance, specifically in terms of accuracy and lifetime of heterogeneous WSNs, of the proposed algorithms against its closest rivals techniques. Apart from data fusion and aggregation, a hybrid routing algorithm, the shortest path and vulnerability-based routing were presented that considerably enhanced heterogeneous WSN’s lifetime and were verified by the simulation results. In the future, we plan to investigate methods and techniques for dynamic gateways and cluster heads in heterogeneous WSNs with strong emphasis on data fusion and aggregation. We also plan to further improve the precision and accuracy of our aforementioned algorithms, i.e., node-level data fusion. Additionally, two-level data fusion and aggregation at the ordinary node and cluster head is also an interesting future research directive.

References

1. Kobo HI, Abu-Mahfouz AM, Hancke GP. A Survey on Software-Defined Wireless Sensor Networks: Challenges and Design Requirements.. *IEEE Access* 2017; 5(1): 1872–1899.

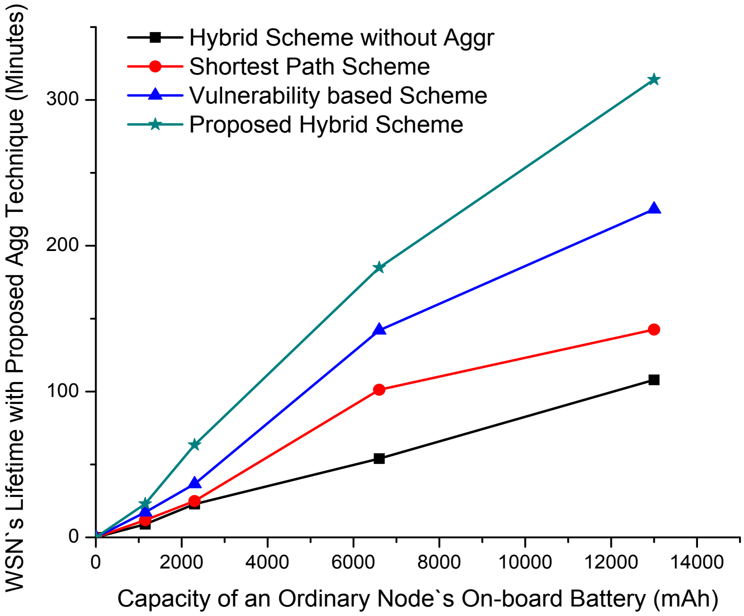

FIGURE 12 Lifetime comparison of heterogeneous WSNs by adopting routing schemes with both embedded data aggregation facility i.e. the proposed hybrid sampling technique and without it [higher values are “best"]

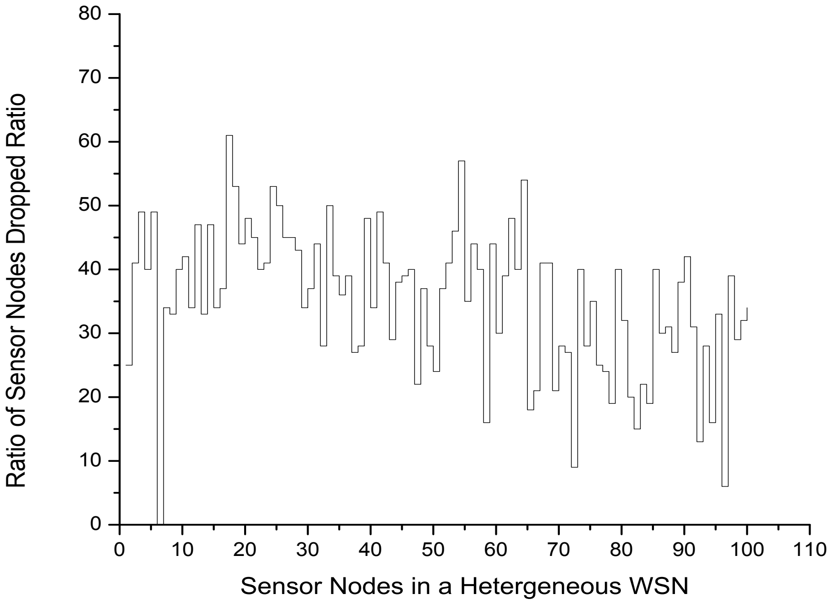

FIGURE 13 Packets dropped by sensor nodes in heterogeneous WSNs to enhance its lifetime

2. Gavalas D, Venetis IE, Konstantopoulos C, Pantziou G. Mobile agent itinerary planning for WSN data fusion: considering multiple sinks and heterogeneous networks. *International Journal of Communication Systems* 2017; 30(8): e3184.

3. Naranjo PGV, Shojafar M, Mostafaei H, Pooranian Z, Baccarelli E. P-SEP: A prolong stable election routing algorithm for energy-limited heterogeneous fog-supported wireless sensor networks. *The Journal of Supercomputing* 2017; 73(2): 733–755.

4. Wang Z, Cao Q, Qi H, Chen H, Wang Q. Cost-effective barrier coverage formation in heterogeneous wireless sensor networks. *Ad Hoc Networks* 2017; 64: 65–79.

5. Lin CC, Chen YC, Chen JL, Deng DJ, Wang SB, Jhong SY. Lifetime enhancement of dynamic heterogeneous wireless sensor networks with energy-harvesting sensors. *Mobile Networks and Applications* 2017; 22(5): 931–942.

6. Zhong H, Shao L, Cui J, Xu Y. An efficient and secure recoverable data aggregation scheme for heterogeneous wireless sensor networks. *Journal of Parallel and Distributed Computing* 2018; 111: 1–12.

7. Hahmann S, Miksch J, Resch B, Lauer J, Zipf A. Routing through open spaces–a performance comparison of algorithms. *Geo-spatial Information Science* 2018; 21(3): 247–256.

8. Li W, Fu Z. Unmanned aerial vehicle positioning based on multi-sensor information fusion. *Geo-spatial Information Science* 2018; 21(4): 302–310.

9. Gilbert EPK, Kaliaperumal B, Rajsingh EB, Lydia M. Trust based data prediction, aggregation and reconstruction using compressed sensing for clustered wireless sensor networks. *Computers & Electrical Engineering* 2018.

10. Plata-Chaves J, Bertrand A, Moonen M, Theodoridis S, Zoubir AM. Heterogeneous and Multitask Wireless Sensor Networks-Algorithms, Applications, and Challenges.. *J. Sel. Topics Signal Processing* 2017; 11(3): 450–465.

11. Jan MA, Jan SRU, Alam M, Akhunzada A, Rahman IU. A comprehensive analysis of congestion control protocols in wireless sensor networks. *Mobile networks and applications* 2018; 23(3): 456–468.

12. Jan SRU, Jan MA, Khan R, Ullah H, Alam M, Usman M. An Energy-Efficient and Congestion Control Data-Driven Approach for Cluster-Based Sensor Network. *Mobile Networks and Applications* 2018: 1–11.

13. Ahlawat P, Dave M. A hybrid approach for path vulnerability matrix on random key predistribution for wireless sensor networks. *Wireless Personal Communications* 2017; 94(4): 3327–3353.

14. Khan R, Khan SN, Ahmad M, Muhammad T. Increasing network lifetime and data transfer through node vulnerability aware routing in Wireless Sensor Networks. In: IEEE. ; 2010: 1–5.

15. Izadi D, Abawajy JH, Ghanavati S, Herawan T. A data fusion method in wireless sensor networks. *Sensors* 2015; 15(2): 2964–2979.

16. Fortino G, Galzarano S, Gravina R, Li W. A framework for collaborative computing and multi-sensor data fusion in body sensor networks. *Information Fusion* 2015; 22: 50–70.

17. Illiano VP, Lupu EC. Detecting malicious data injections in wireless sensor networks: A survey. *ACM Computing Surveys (CSUR)* 2015; 48(2): 24.

18. Qin J, Fu W, Gao H, Zheng WX. Distributed k-means algorithm and fuzzy c-means algorithm for sensor networks based on multiagent consensus theory. *IEEE transactions on cybernetics* 2017; 47(3): 772–783.

19. Yuan K, Xiao F, Fei L, Kang B, Deng Y. Modeling sensor reliability in fault diagnosis based on evidence theory. *Sensors* 2016; 16(1): 113.

20. Zheng Y, Cao N, Wimalajeewa T, Varshney PK. Compressive sensing based probabilistic sensor management for target tracking in wireless sensor networks. *IEEE Transactions on Signal Processing* 2015; 63(22): 6049–6060.

21. Tirkolaee E, Hosseinabadi A, Soltani M, Sangaiah A, Wang J. A hybrid genetic algorithm for multi-trip green capacitated arc routing problem in the scope of urban services. *Sustainability* 2018; 10(5): 1366.

22. Wang J, Gao Y, Liu W, Sangaiah AK, Kim HJ. An improved routing schema with special clustering using PSO algorithm for heterogeneous wireless sensor network. *Sensors* 2019; 19(3): 671.

23. Heinzelman WB, Chandrakasan AP, Balakrishnan H. An application-specific protocol architecture for wireless microsensor networks. *IEEE Transactions on wireless communications* 2002; 1(4): 660–670.

24. Younis O, Fahmy S. HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. *IEEE Transactions on mobile computing* 2004; 3(4): 366–379.

25. Qing L, Zhu Q, Wang M. Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks. *Computer communications* 2006; 29(12): 2230–2237.

26. Wang J, Gao Y, Liu W, Wu W, Lim SJ. An asynchronous clustering and mobile data gathering schema based on timer mechanism in wireless sensor networks. *Comput. Mater. Contin* 2019; 58: 711–725.

27. Song C, Liu M, Cao J, Zheng Y, Gong H, Chen G. Maximizing network lifetime based on transmission range adjustment in wireless sensor networks. *Computer Communications* 2009; 32(11): 1316–1325.

28. Wang J, Gao Y, Yin X, Li F, Kim HJ. An enhanced PEGASIS algorithm with mobile sink support for wireless sensor networks. *Wireless Communications and Mobile Computing* 2018; 2018.

29. Ammari HM, Das SK. Promoting heterogeneity, mobility, and energy-aware voronoi diagram in wireless sensor networks. *IEEE Transactions on Parallel and Distributed Systems* 2008; 19(7): 995–1008.

30. Nudurupati DP, Singh RK. Enhancing Coverage Ratio using Mobility in Heterogeneous Wireless Sensor Network. *Procedia Technology* 2013; 10: 538–545.

31. Wang J, Ju C, Gao Y, Sangaiah AK, Kim Gj. A PSO based energy efficient coverage control algorithm for wireless sensor networks. *Comput. Mater. Contin* 2018; 56: 433–446.

32. Ma G, Tao Z. A nonuniform sensor distribution strategy for avoiding energy holes in wireless sensor networks. *International Journal of Distributed Sensor Networks* 2013; 9(7): 564386.

33. Baranidharan B, Santhi B. DUCF: Distributed load balancing Unequal Clustering in wireless sensor networks using Fuzzy approach. *Applied Soft Computing* 2016; 40: 495–506.

34. Leu JS, Chiang TH, Yu MC, Su KW. Energy efficient clustering scheme for prolonging the lifetime of wireless sensor network with isolated nodes. *IEEE communications letters* 2015; 19(2): 259–262.

35. Demertzis A, Oikonomou K. Avoiding energy holes in wireless sensor networks with non-uniform energy distribution. In: IEEE. ; 2014: 138–143.

36. Lian J, Naik K, Agnew GB. Data capacity improvement of wireless sensor networks using non-uniform sensor distribution. *International Journal of Distributed Sensor Networks* 2006; 2(2): 121–145.

37. Luo J, Hubaux JP. Joint mobility and routing for lifetime elongation in wireless sensor networks. In: . 3. IEEE. ; 2005: 1735–1746.

38. Wang J, Gao Y, Liu W, Sangaiah AK, Kim HJ. An intelligent data gathering schema with data fusion supported for mobile sink in wireless sensor networks. *International Journal of Distributed Sensor Networks* 2019; 15(3): 1550147719839581.

39. Jung WS, Lim KW, Ko YB, Park SJ. Efficient clustering-based data aggregation techniques for wireless sensor networks. *Wireless Networks* 2011; 17(5): 1387–1400.

40. Koutsonikolas D, Das SM, Hu YC, Stojmenovic I. Hierarchical geographic multicast routing for wireless sensor networks. *Wireless networks* 2010; 16(2): 449–466.

41. Xin G, Fei-qi D. Complete ternary tree-based data aggregation routing algorithm for wireless sensor networks. In: IEEE. ; 2013: 578–581.

42. Kuo TW, Tsai MJ. On the construction of data aggregation tree with minimum energy cost in wireless sensor networks: NP-completeness and approximation algorithms. In: IEEE. ; 2012: 2591–2595.

43. Pan JS, Kong L, Sung TW, Tsai PW, Snášel V. α-Fraction First Strategy for Hierarchical Model in Wireless Sensor Networks. *Journal of Internet Technology* 2018; 19(6): 1717–1726.

44. Arora VK, Sharma V, Sachdeva M. A distributed, multi-hop, adaptive, tree-based energy-balanced routing approach. *International Journal of Communication Systems*: e3949.

45. Wang NC, Chiang YK, Hsieh CH, Chen YL. Grid-based data aggregation for wireless sensor networks. *Journal of Advances in Computer Networks* 2013; 1(4).

46. Wang Q, Liao H, Wang K, Sang Y. A variable weight based fuzzy data fusion algorithm for WSN. In: Springer. ; 2011: 490–502.

47. Pradhan S, Sinha E, Sharma K. Data Fusion by Truncation in Wireless Sensor Network. In: Springer. 2018 (pp. 544–551).

48. Xiao L, Boyd S, Lall S. A scheme for robust distributed sensor fusion based on average consensus. In: IEEE. ; 2005: 63–70.

49. Dong M, Ota K, Liu A. RMER: Reliable and energy-efficient data collection for large-scale wireless sensor networks. *IEEE Internet of Things Journal* 2016; 3(4): 511–519.

50. Zin SM, Anuar NB, Kiah MLM, Ahmedy I. Survey of secure multipath routing protocols for WSNs. *Journal of Network and Computer Applications* 2015; 55: 123–153.

51. Ren F, He T, Das SK, Lin C. Traffic-aware dynamic routing to alleviate congestion in wireless sensor networks. *IEEE Transactions on Parallel and Distributed Systems* 2011; 22(9): 1585–1599.

52. El-Fouly FH, Ramadan RA, Mahmoud MI, Dessouky MI. REBTAM: Reliable energy balance traffic aware data reporting algorithm for object tracking in multi-sink wireless sensor networks. *Wireless Networks* 2018; 24(3): 735–753.

53. Hammoudeh M, Newman R. Adaptive routing in wireless sensor networks: QoS optimisation for enhanced application performance. *Information Fusion* 2015; 22: 3–15.

54. Tang D, Li T, Ren J, Wu J. Cost-aware secure routing (CASER) protocol design for wireless sensor networks. *IEEE Transactions on Parallel and Distributed Systems* 2015; 26(4): 960–973.

55. Cota-Ruiz J, Rivas-Perea P, Sifuentes E, Gonzalez-Landaeta R. A recursive shortest path routing algorithm with application for wireless sensor network localization. *IEEE Sensors Journal* 2016; 16(11): 4631–4637.

56. Khan R. An efficient load balancing and performance optimization scheme for constraint oriented networks. *Simulation Modelling Practice and Theory* 2019; 96: 101930.

57. Arvanitis T, Constantinou C, Stepanenko A, Sun Y, Liu B, Baughan K. Network visualisation and analysis tool based on logical network abridgment. In: IEEE. ; 2005: 106–112.

58. Ahmed G, Khalid Z, Khan NM, Vigneras P. Energy efficient and vulnerability aware routing in wireless sensor networks. In: ; 2008: 1-4

59. Khan NM, Ali I, Khalid Z, Ahmed G, Ramer R, Kavokin AA. Quasi centralized clustering approach for an energy-efficient and vulnerability-aware routing in wireless sensor networks. In: ACM. ; 2008: 67–72.

60. Khan R, Ali I, Zakarya M, Ahmad M, Imran M, Shoaib M. Technology-Assisted Decision Support System for Efficient Water Utilization: A Real-Time Testbed for Irrigation Using Wireless Sensor Networks. *IEEE Access* 2018.

61. Johnson JL. Design of experiments and progressively sequenced regression are combined to achieve minimum data sample size. *International Journal of Hydromechatronics* 2018; 1(3): 308–331.

62. Varga A. *OMNeT++*: 35–59; Berlin, Heidelberg: Springer Berlin Heidelberg . 2010

63. Jan MA, Usman M, He X, Rehman AU. SAMS: A seamless and authorized multimedia streaming framework for WMSN-based IoMT. *IEEE Internet of Things Journal* 2018; 6(2): 1576–1583.

64. Shin H, Moh S, Chung I, Kang M. Equal-size clustering for irregularly deployed wireless sensor networks. *Wireless Personal Communications* 2015; 82(2): 995–1012.

□