Outlier Detection and Correction in Wireless Sensor Networks Using a 2-Level Hierarchical Clustering and Gaussian Process Regression Methods

Edward Yellakuor Baagyere, *Member, IAENG*, Regina Esi Turkson, Gideon Evans Wenya, and Iven Aabaah

***Abstract*—In this research paper, we investigated the issue of wireless sensor networks (WSNs) producing anomalous or missing data due to a variety of factors, such as intrusion attacks, node failures, link failures, malicious attacks, and environmental variations. To address this problem, we proposed a novel 2-level hierarchical LEACH protocol approach for electing a referenced sensor with some predefined conditions. This sensor is then tasked with detecting, flagging, and predicting outlier or missing measurements using the Gaussian Process Regression (GPR) technique. We tested the outlier detection, prediction, and correction framework using weather station data. The results demonstrated that the framework is an effective way for detecting, predicting, and correcting outlying sensor measurements and can easily be infused into current WSNs as middleware, as the computational overhead of our approach is low (only simple arithmetic operations are involved) and is purely localized and is therefore scalable to larger WSNs.**

***Index Terms*—wireless sensor networks, leach protocol, gaussian process regression, outlying detection, prediction.**

I. Introduction

WIRELESS Wireless Sensor Networks (WSNs) are a widely used technology designed to monitor and observe physical phenomena in specific areas, delivering accurate and real-time information even in challenging industrial environments characterized by extreme vibrations, noise, humidity, and temperature conditions [1]–[4].

They have a wide range of applications, spanning from healthcare to commercial sectors [5]–[7].

In all these application areas, both homogeneous and heterogeneous sensors are employed to record various quantities such as sound, temperature, humidity, GPS logs, and the occurrence of events. These sensors are either randomly or manually distributed throughout the area, forming a self-organized network. The sensors collect data, process or fuse it, and transmit the information to a sink node via single-hop or multi-hop paths through neighboring nodes. By fusing and transforming the data, the sink node can present the information in a format that is easier for users to comprehend. The transformed data can be utilized in routine decision-making processes to enhance the quality of human life and other physical systems. Consequently, the accuracy of the information gathered by the sensor network is critical to its relevance, as even a minor error could have serious consequences for lives and other systems. This underscores the importance of data integrity as a top priority in the design of Wireless Sensor Networks (WSNs).

However, WSNs are susceptible to various types of failures, including intrusion attacks, node failures, link failures, malicious attacks, environmental variations, limited resource provisioning, and excessive burdens from redundant data processing tasks. When these failures occur, the data generated by the sensors may contain inaccuracies and may not be reliable for making critical decisions. In a densely deployed sensor architecture, data fusion methods are employed to rectify missing information; however, this approach may not be effective if a significant number of sensors are reporting erroneous data.

There is a need to develop a more cost-effective method for predicting missing, corrupt, or outlier sensor values, and to correct these discrepancies before the data is processed for decision-making purposes.

In the detection of outliers or anomalies in WSNs, automated analysis techniques, such as anomaly detection algorithms, are commonly employed to identify and flag data points that deviate significantly from expected patterns or behaviors. These techniques are typically executed on a centralized server or processing unit, which may be situated far from the actual WSN monitoring site. This physical separation can result in increased network latency and reduced reliability, as data must be transmitted from the sensors to the centralized server for analysis.

A more effective approach is to distribute data processing, outlier detection, and data prediction across the nodes. This method will lead to a more efficient use of system resources and reduce the communication overhead associated with processing and transmitting data [8].

The main contributions of the paper are outlined below:

(i) A two-level hierarchical LEACH protocol is proposed for selecting a reference sensor node based on predefined conditions to detect missing sensor measurements.
(ii) A machine learning model is leveraged to predict outliers or identify missing sensor measurements.
(iii) We experimentally demonstrated the feasibility of the

Manuscript received November 22, 2023; revised April 27, 2025.

Edward Yellakuor Baagyere is the Dean of the School of Computing and Information Sciences, C. K. Tedam University of Technology and Applied Sciences, Navrongo, Ghana (e-mail: ebaagyere@cktutas.edu.gh).

Regina Esi Turkson is a Lecturer in the Department of Computer Science and Information Technology, University of Cape Coast, Cape Coast, Ghana (e-mail: rturkson@ucc.edu.gh).

Gideon Evans Wenya is a Lecturer in the School of Advanced Technologies, Engineering Science (SATES), Accra Institute of Technology, Accra-North, Ghana (e-mail: gideonwenya@gmail.com).

Iven Aabaah is an Assistant Lecturer in the Department of Information Systems and Technology, C. K. Tedam University of Technology and Applied Sciences, Navrongo, Ghana (e-mail: iaabaah@cktutas.edu.gh).

proposed localized outlier detection and correction model using WSN data from a weather station.

The rest of the paper is organized as follows: Section II provides comprehensive background information and related work in wireless sensor networks. Section III discusses the conceptual framework for cluster heads. In Section IV, the framework for detecting and flagging outlying sensor measurements is described. Section V, employs a Gaussian Process Regression (GPR) approach to predict outlying sensor measurements. The data used to test the framework and the results are discussed in Section VI. The paper concludes in Section VII.

II. Background on Wireless Sensor Networks and Related Work

This section provides a brief overview of wireless sensor network (WSN) systems, emphasizing the components of WSNs and the functionality of each component. Additionally, it discusses related research on WSNs, specifically focusing on sensor data measurements, the detection of outlier sensors, and the prediction of anomalous sensor measurements.

A. The Structure of a Wireless Sensor Network

A wireless sensor network (WSN) is an embedded system that integrates micro-electronic technology, embedded computing, communication technology, and sensor technology to collaboratively monitor physical or environmental conditions, such as temperature, sound, vibration, pressure, motion, or pollutants across various temporal and spatial domains. These networks can be deployed either randomly or manually in fixed positions within a designated monitoring area. WSNs are self-organizing networks designed primarily for sensing, computing, and communicating data. Figure 1 illustrates a typical architecture of a wireless sensor network, with the various functionalities explained as follows:

The Sink Node

(i) The sink node connects the sensor network to the Internet and other external networks, serving as the gateway to the sensor network when monitoring functions are not in use.

(ii) The sink node converts between protocols and transmits data to the external network.

(iii) The sink node can also function as an extension of the sensor nodes, offering additional energy, memory, and computing resources.

Management Node

(i) The management node, also known as the user node, can be utilized to configure and manage a wireless sensor network (WSN).

(ii) It is used to assign monitoring tasks and to collect monitoring data.

The Power Unit

(i) The energy required to power the sensors for environmental monitoring is sourced from a power unit, which is primarily powered by dry cells. This energy is typically low-cost and delivered promptly.

(ii) Battery power is a critical factor in determining the lifespan and functionality of Wireless Sensor Network (WSN) sensor nodes, as it serves as the primary energy source for their operation. Therefore, effective management and distribution of battery power are essential for maximizing the efficiency and longevity of the WSN system.

The Processor Module

(i) The processor module comprises data processing elements and memory.

(ii) It is responsible for collecting data within a node's monitoring environment, processing it, and storing it in memory.

(iii) The computational power and energy consumption rate of the WSN architecture are determined by the central processing unit embedded in each sensor node.

(iv) A wide variety of microcontrollers, microprocessors, and FPGAs are utilized to enhance the flexibility of CPU implementations

The Communication Module

(i) The communication module consists of a transceiver responsible for both transmitting and receiving data. These two functions are integrated into the same circuitry on a single board.

(ii) The communication module is also capable of receiving commands from the processing unit of each sensor node and subsequently relaying them to other nodes within the network. This is accomplished through communication channels that adhere to a specific network protocol.

Sensing Unit

(i) The sensing unit consists of an analog-to-digital (A/D) converter and multiple sensors.

(ii) The A/D converter is used to convert an analog signal into a digital signal. As a result, the input is an analog signal received from the sensor, and the output is a digital signal. This digital signal is then sent to the microcontroller embedded within the sensor node for further processing.

(iii) Each sensor node in a Wireless Sensor Network (WSN) can possess multiple sensing capabilities simultaneously, depending on the application. For instance, acoustic sensors, resonant temperature sensors, and magnetic field sensors can detect a variety of physical phenomena, including temperature, sound, pressure, and gravity.

(iv) Sensing nodes also have the capability to function as routers, forwarding data to neighboring nodes within the WSN architecture.

The current WSN architecture can be modified to effectively identify outlying neighboring sensors for each node, facilitating the prediction of outlying sensor measurements. We propose a localized outlier detection scheme that employs a 2-level hierarchical LEACH protocol and the Gaussian Process Regression method for the prediction process.

B. Related Work

In this section, we survey related works in wireless sensor networks, with a primary focus on outlier detection and correction.

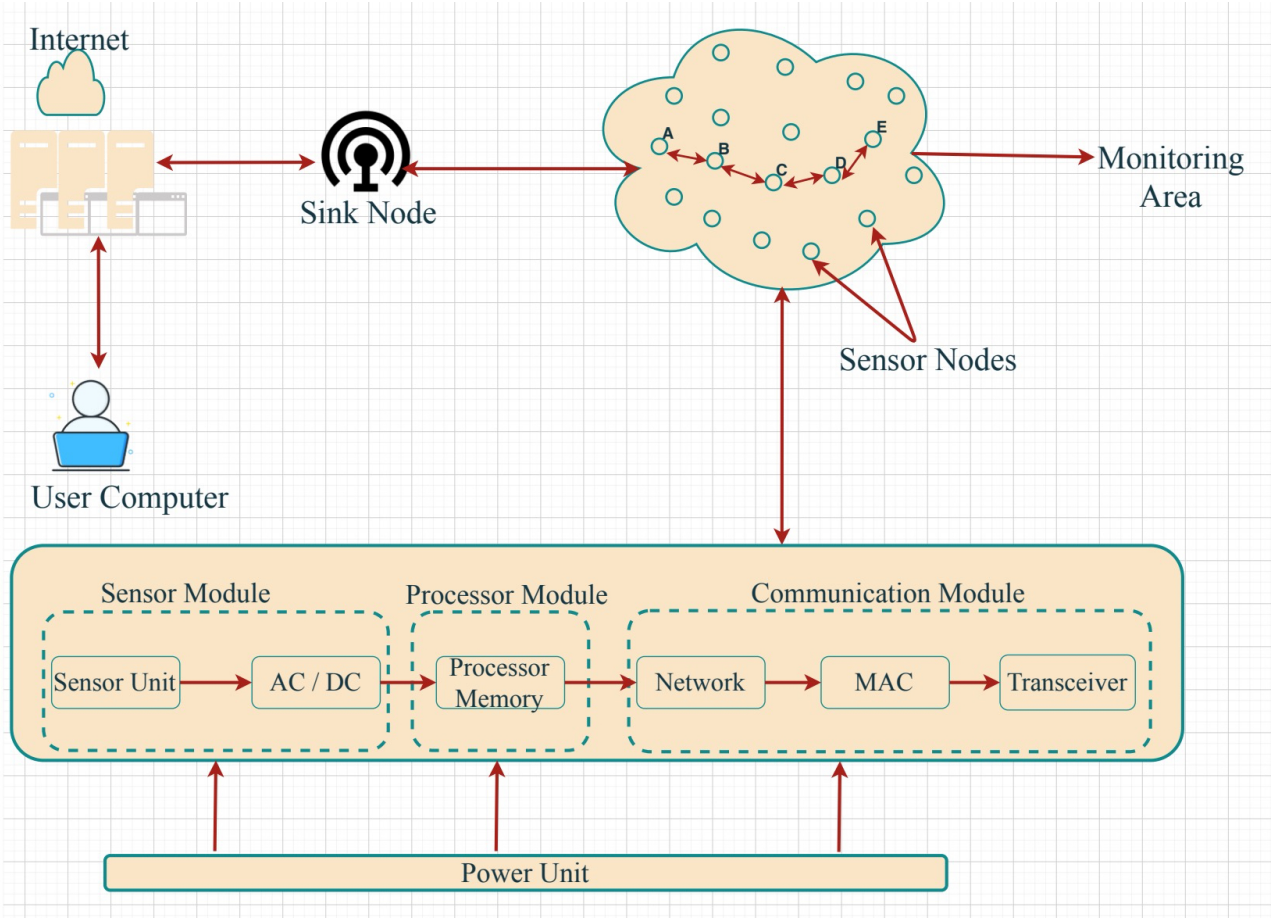

Fig. 1: Model of a Sensor Network

Outlier detection in sensor measurements has garnered significant research attention across various domains [9]–[16]. However, there have been few attempts to correct these outlying sensor measurements using appropriate and computationally efficient methods. The authors in [17] proposed a multi-sensor information filtering framework that enables a desired number of sensors with the most accurate measurements to participate in information exchange while discarding sensors with inaccurate and outlier measurements. Shantala and Vijayakumar [18] introduced an outlier detection scheme for identifying events and attacks in wireless sensor networks.

Wang Feng *et al.* [19] designed a middleware framework for processing heterogeneous information from various devices in the context of the Internet of Things (IoT) using a service-oriented architecture. This data processing middleware architecture is employed for both theoretical analysis and experimental verification, utilizing environmental monitoring sensor data. The authors reported that the architecture demonstrates superior adaptability to multi-sensor and multi-stream application scenarios within the IoT framework, while also enhancing the overall value of IoT by improving the utilization of heterogeneous data. Socoró *et al.* [10] ooutlined a method for detecting anomalous sensor measurements in health monitoring systems by analyzing physiological data from medical sensors. The research aimed to effectively distinguish false alarms from true alarms.

In [9],], the authors proposed a segmentation algorithm utilizing a ”one-class support vector machine approach” to identify anomalies in turbomachines. Wu *et al.* [20] developed localized scalable algorithms for detecting outlying sensors and events. Their algorithm is capable of clearly detecting event boundaries and identifying outlying sensors. The authors in [21] described an online anomaly detection system that employs an ensemble of classifiers, which can be executed on embedded systems such as wireless sensors. Their work considers both single and multi-dimensional input classifiers for predicting errors. The authors further tested the framework using both synthetic and real-world data, reporting that the use of ensemble classifiers significantly enhances the overall detection of anomalies The authors in [22]proposed an approach to predict critical nodes in an opportunistic sensor network based on a multiple attribute decision-making process. They employed the “TOPSIS algorithm” to identify the ferry node with the maximum comprehensive contribution, which is deemed a critical node. Sundararajan and Arumugam [23] proposed “an intrusion detection algorithm to mitigate sinkhole attacks on the LEACH protocol in WSNs”. The implemented Intrusion Detection System (IDS) utilized the number of packets transmitted and received to compute the intrusion ratio, which indicates normal or malicious activity. The IDS agent alerts the network to cease data transmission whenever a sinkhole attack is detected. The authors in [24] proposed a support vector machine approach based on an improved particle swarm optimization technique to predict dynamic errors in sensor networks. The root mean square error and mean absolute percentage error were used for model evaluation in terms of

prediction accuracy and precision.

A K-means and neural network approach for detecting outliers in social network analysis is proposed by [14]. The authors in [25] introduced a multivariate spatial and temporal correlation method aimed at enhancing prediction accuracy and reducing data for WSN. Simulations were conducted using simple linear regression and multiple linear regression functions to evaluate the proposed method. They confirmed that prediction accuracy is lower when simple linear regression is employed, while multiple linear regression yielded the most accurate predictions. Additionally, a congestion and traffic path prediction model is proposed by [26] to forecast and minimize sensor data congestion. The authors evaluated the effectiveness of the model through simulations conducted with NS-2 and MATLAB®. They employed a network grid representation method that preserved the fine-scale structure of a transportation network, as described in [27]. This method was utilized to predict road traffic by converting static images derived from network-wide traffic speed into a "Spatiotemporal recurrent convolutional network (SRCN)". They demonstrated the effectiveness of the model using real-world data and reported that the SRCNs outperformed "other deep learning-based algorithms in both short-term and long-term traffic prediction".

Our work is distinct and innovative compared to the approaches employed by other researchers. We propose a scheme that incorporates the traditional LEACH protocol [28], [29] to elect a reference sensor at level-1 and a cluster head sensor at level-2 during each round of the protocol. The level-1 cluster head, referred to as the reference sensor, is a dedicated sensor responsible for predicting and correcting outlying or missing sensor measurements using simple arithmetic operations integrated into the sensors as middleware. A sensor is selected as the reference sensor based on its residual energy and data variability index during each round of the protocol. Subsequently, the Gaussian Process Regression (GPR) method is utilized to predict missing sensor values and to correct any missing or outlying sensor measurements.

III. The Cluster Head Conceptual Framework

Sensor nodes are typically organized into disjoint sets known as clusters. This grouping of sensors into clusters, based on specific criteria, primarily aims to enhance network stability through the efficient utilization of resources. Additionally, it serves as an energy-saving mechanism for the WSN architecture [28], [29]. Furthermore, clustering techniques not only facilitate efficient data collection and transmission but also optimize resource usage and enhance overall network performance.

Clustering routing techniques utilize the data aggregation and information processing capabilities of cluster heads to minimize the volume of data transmitted across the network. By aggregating and processing data at the cluster head level, this approach decreases the amount of data that individual nodes must transmit, thereby conserving energy and extending the network's lifespan. This method is particularly advantageous in resource-constrained WSNs where nodes have limited battery power and energy conservation is essential for prolonged network operation [30]–[33].

Clustering routing techniques leverage the data aggregation and information processing capabilities of cluster heads to reduce the amount of data transmitted throughout the network. By aggregating and processing data at the cluster head level, this approach reduces the amount of data that needs to be transmitted by individual nodes, which in turn helps conserve energy and extend the network lifetime. This method is especially beneficial in resource-constrained WSNs where nodes have limited battery power, and energy conservation is crucial for prolonged network operation [30]–[33].

The LEACH (Low Energy Adaptive Clustering Hierarchy) protocol introduced significant advancements in clustering techniques by considering both the minimum transmission energy and the overall energy consumption within the network. LEACH improved upon previous clustering protocols by addressing not only the minimum transmission energy but also the total energy consumption throughout the network. This approach helps ensure that no single node becomes overly burdened by managing a disproportionate share of the network's communication demands.

In this paper, we utilize a cluster head conceptual framework based on the LEACH routing protocol to group sensors according to their *location, elevation, and homogeneity*. This approach aims to detect and correct outlier sensor measurements.

A. The LEACH Protocol

The LEACH protocol is an energy-efficient, cluster-based routing framework that incorporates media access and application-specific data aggregation. This configuration results in improved performance regarding system lifetime, latency, and application-perceived quality [29], [34].

In the LEACH protocol, all nodes, except for the cluster heads, transmit data to the head of their respective clusters. The cluster heads then perform data aggregation, compression, and onward transmission of the pre-processed data to the base station. During each round, each node employs a probability-based approach to determine whether it will become a cluster head for that round. A fundamental assumption of the LEACH protocol is that every node in the network possesses sufficient energy to transmit data directly to the base station or to the nearest cluster head. However, the protocol acknowledges that continuously utilizing this full energy capacity may not be the most efficient strategy for extending the network's lifespan. To address this, the protocol implements a random rotation of cluster heads, which helps to distribute energy consumption among various nodes. This approach prevents the depletion of a single node's energy by evenly distributing energy usage across the network.

The LEACH protocol employs a round-based approach, with each round consisting of two phases: cluster head selection and data transmission. During the cluster head selection phase, each node has a probability of $1/P$ of being selected as a cluster head for only one round. Once selected, a node cannot be a cluster head again for P rounds. Any node that was not a cluster head in the previous round selects the nearest cluster head and joins that cluster. The cluster head then creates a schedule for each node in its cluster to transmit its data.

Equation 1 mathematically expresses the LEACH protocol as described above.

$$T(n) = \begin{cases} \frac{p}{1-P*[r \mod \frac{1}{P}]} & \text{if } n \in G \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

B. The 2-level Hierarchical Clustering Algorithm

The 2 -level hierarchical clustering algorithm is an extension of the LEACH protocol, allowing for multiple clustering heads (CH) at different levels [28].

Assuming there are h_i levels in the clustering hierarchy, then h_1 is designated as the level-1 cluster head, representing the lowest level, while h_2 is the highest level. Consequently, each cluster network features 2-level hierarchical cluster heads: the *1-level cluster head* is responsible for sensor data comparison, prediction, and correction, while the *2-level cluster head* manages data fusion, aggregation, and the onward transmission of processed data to the base station (sink).

Algorithm 1 illustrates the level-1 cluster head selection process. The time and space complexity of the algorithm are O(N^2) and O(N), respectively.

Algorithm 1: Level-1 cluster head selection Procedure

Data: Sensor Nodes
Result: Level-1 cluster head selected
for *Each sensor node* **do**
  Compute Sensor Data Variability Index ζ using Equation 2;
for *Each sensor node* **do**
  Choose a random node number R;
  Compute a threshold value $T(n)$ using Equation 1;
if $R < T(n)$ *And* $\zeta > \theta_i$ **then**
  Sensor node is selected as a Level-1 Cluster head;
  Sensor node broadcast it status to the WSN.
else
  Sensor node select itself as a normal node;
  Decide which cluster to join in this round based on its current residual energy and distance to cluster head.

Then, the level-1 CHs serve as reference sensors for calculating the distance di_s, which is defined in Equation 3 in Section IV-B. These d_i values are then calculated and standardized using Equation 4. Sensors with high y_i values are flagged as outliers, and their corresponding values are predicted using a GP regression method. The corrected and fused data are subsequently communicated to the level-2 CHs which ultimately relay the aggregated data or estimates based on this aggregated data to the processing center.

The level-2 cluster heads are selected during a second round using the standard LEACH protocol selection process, as illustrated in Equation 1. Due to the high correlation between data signals from nearby nodes, we can implement a clustering infrastructure to predict corrupted or missing sensor data. This approach enables local processing of all data within a cluster, thereby minimizing the volume of data that must be transmitted to the end user. Additionally, data aggregation techniques further decrease the amount of data that needs to be sent by combining multiple data signals into a more compact set of information that retains the same informational content as the original signals.

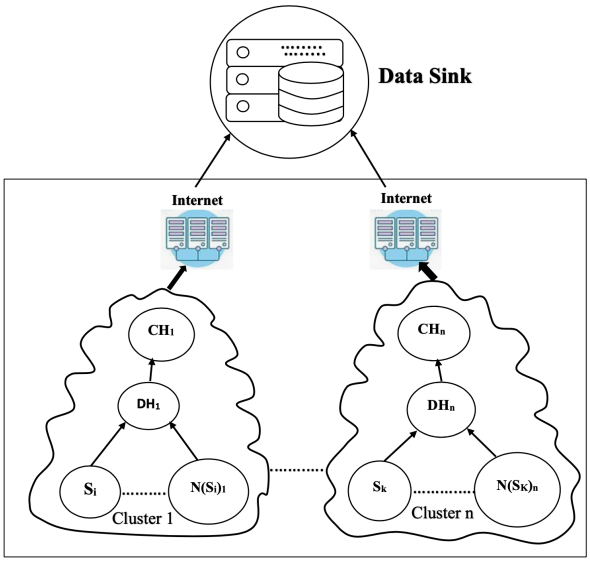

Fig. 2: The Hierarchical Clustering Framework

C. Node Cluster Selection Process

Sensors are grouped into clusters based on predefined criteria, including distance, homogeneity, and elevation. Consequently, sensors within a particular cluster exhibit similar measurements. Thus, the measured values obtained from homogeneous data should be very similar within the same time frame. The proposed conceptual framework of the 2-level hierarchical clustering system is illustrated in 2.

D. Assumption on Nodes within each Cluster

The following assumptions were made when determining the cluster heads:

(i) Nodes can communicate with one another using the MAC layer protocol embedded within the sensors to coordinate neighboring broadcasts, ensuring that no collisions occur.
(ii) The WNS is homogeneous, meaning that the nodes are measuring the same or similar sensor data within a specific cluster at a particular time or spatial domain.
(iii) All level-2 nodes can communicate with the Base Station (BS) while maintaining sufficient energy.
(iv) That all level-1 nodes have enough residual energy to serve as a reference sensor for comparative analysis and flagging outlying sensors
(v) That data receive from these sensors are subject to the same problems
(vi) That the sensors are not separated by more than a few tens of kilometres and hence can be expected to experience reasonably similar weather conditions. As such, their readings are likely to display a strong degree of correlation
(vii) That nodes can use different power level for communication.

The advantage of using the cluster head conceptual framework are outlined below:

(i) The computational overhead is very low because one sensor (the data head) is involved in data aggregation, comparison, and prediction during each round.
(ii) The computational overhead is low because only the relevant sensor values are transmitted to the sink node or base station.
(iii) This reduces the number of transmissions in the network.
(iv) Data aggregation of cluster heads from their cluster members also reduces duplicate transmission and enhances the network lifetime owing to energy efficiency

E. Advantages of performing sensor data correction close to the data site

(i) It allows data reduction while providing information when unexpected behavior occurs.
(ii) Only relevant data/information are sent to the sink, leading to efficient use of resources in terms of bandwidth owing to less usage of communication channels
(iii) Nodes with corrupt or missing data can be corrected using their neighbors' sensor data.
(iv) The cost of communication for transporting 'raw' data to the sink is high.

IV. Localized Outlying Sensor Detection Framework

This section defines the neighborhood of a given sensor and outlines the procedure for detecting an outlying sensor.

Sensor Network Neighbourhood Defined

Let $\Phi(\mathcal{R}, S_i)$ be the Euclidean distance $\mathcal{R}$ around sensor S_i and $\mathcal{N}(S_i)$ be the neighbors of sensor S_i at a distance $\mathcal{R}$ from Sensor S_i. Thus, the neighborsof S_i are all sensors $\in |\mathcal{N}(S_i)|$ within that Euclidean space.

Outlying sensor defined

A sensor whose reading is an outlier (described as an outlying sensor) when the observation of its readings deviates significantly from that of other neighboring sensor readings [35], [36] because of either environmental factors or system defects within the WSN. In WSNs, sensor nodes are assigned to monitor the behavior of the physical world; thus, the sensed data must have a pattern that represents the true behavior of the sensed data.

However, more often than not, data obtained from WSNs are normally unreliable because they are affected by noise and measurement errors. To generate redundant data, low-cost, low-quality sensors are normally distributed over a given region. These sensors normally have low energy (battery life), memory, computational capacity, and communication bandwidth. These limitations make the data generated by sensor nodes prone to errors and can, therefore, be defined as outlying sensors, that is, their measured values deviate greatly from the true values. Sensor data can also be vulnerable to malicious attacks, such as denial of service (DoS) attacks, black holes, and eavesdropping, in which sensor network data are deliberately altered by an adversary. All these factors (environmental and system defects) lead to the generation of unreliable sensor data, which affects the accuracy of the raw data and consequently that of the processed or aggregate data to be used for decision-making. Because events that occur in the physical environment or world cannot be accurately detected using inaccurate and incomplete data, it is important to have a high level of assurance on the reliability and accuracy of sensor data before using the data for decision-making.

Types of Outlying Sensor Readings:

(i) Point outlier: A sensor data reading that is different from all other readings within a defined time frame.
(ii) Contextual outlier: A sensor reading that is anomalous in a specific context or neighborhood, such as a particular longitude, latitude, date, or time within a month.
(iii) Collective outlier: A collection of sensor readings that are anomalous with respect to the entire dataset or a processed sensor data set.

In this work, we used the collective outlying sensor reading approach in our analysis.

A. The Rate of Variability in Sensor Data Set

The rate of variability in a sensor data set is used to measure the integrity of the sensor data. Data from neighboring sensors are assumed to have approximately the same information and should present the same rate of variability. Otherwise, this could indicate that at least one of the data sources is corrupted by noise. Therefore, sensor readings with a lower rate of variability are considered to have higher integrity and are preferred over those with a high rate of variability. The rate of variability index, ζ is defined as [21]:

$$\zeta = \frac{std(x)}{mean(x)}, \tag{2}$$

where $std(x)$ is the standard deviation and x is a vector of sensor measurement. Figure 3 showed the rate of variability index for the temperature measurements for the month of May obtained for the Chimet sensor measurements website [1]. It is observed that some measured values have very high of ζ, signifying the possibility of noise in their measurements.

B. Procedure for Outlying Sensor Data Detection

Let S_i denote a referenced sensor whose status is determined stochastically using the LEACH routing protocol. At each round, S_i is selected based on its residual energy and data variability index ζ as defined in Equation 2 (Section IV-A). Sensors with high residual energy and a low variability index ζ in each of the n clusters are preferred and elected as level-1 cluster heads.

A large ζ value (due to a large standard deviation) is indicative of corrupted sensor data, and thus, that sensor is not qualified to be a level-1 cluster head.

Let $\mathcal{N}(S_i)$ be the neighbors of sensor S_i named $S_i\alpha_1, S_i\alpha_2, S_i\alpha_3, \ldots, S_i\alpha_n$ within the same cluster before S_i was chosen as a level-1 cluster head.

[1] http://www.chimet.co.uk/

The difference, d_i, between the measurement of the reference sensor S_i and the median of the set $\{\alpha_1^{(i)}, \alpha_2^{(i)}, \alpha_3^{(i)}, \ldots, \alpha_n^{(i)}\}$, is defined as follows:

$$d_i = x_i - med\ \alpha_i^{(i)} \tag{3}$$

The set $\mathcal{D} = \{d_1, d_2, d_3, \ldots, d_n\}$ contains values obtained using Equation 3. If d_i is extreme in the set $\mathcal{D}$, then sensor $S_i\alpha_i$ is an outlying sensor whose measurements are likely wrong due to factors such as intrusion attacks, node failures, link failures, malicious attacks, environmental variations, resource constraints, or problems with measuring instruments. The values in the set $\mathcal{D}$ can be standardized by the following procedure:

$$\gamma_i = \frac{d_i - mean(d_i)}{std(d_i)} \tag{4}$$

A decision for flagging a sensor as outlying is defined as follows: *If $|\gamma_i| \geq \theta$ or sensor data measurements fall outside the 95% confidence level*, then the sensor is flagged as outlying. $\theta > 1$ is a preselected value. In this work, θ is defined as follows:

$$\theta_i = \frac{min(\gamma_i) + max(\gamma_i)}{2} \tag{5}$$

This approach enables the identification of outlying sensors and can contribute to the overall reliability and security of the WSNs.

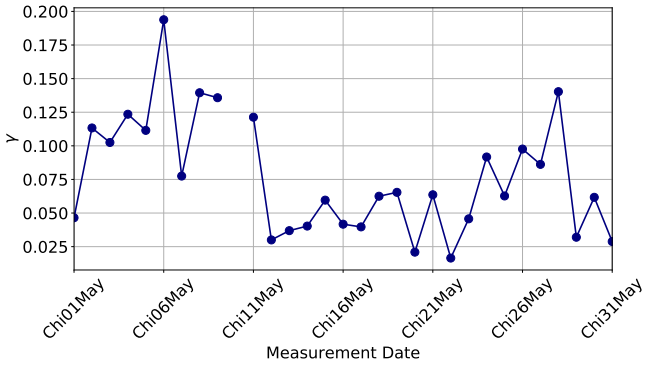

Fig. 3: Sensor Measurement Variability Index for the Month of May, 2017 (The referenced Month in the work)

The algorithm for detecting and correcting outlying sensor measurements is shown in Algorithm 2. The time and space complexity of the algorithm are O(N^2) and O(N), respectively.

Algorithm 2: Outlier Sensor Measurements Detection and Correction

Data: Sensor Measurements
Result: Outlying Sensors Detected and Corrected
while *Threshold Condition $T(n)$ and Sensor Data Variability Index $\zeta >$ predefined value θ_i* **do**
  for *Each Neighbouring Sensor Measurement $\alpha_i^{(i)}$* **do**
    Calculate d_i using Equation 3;
    Standardizes d_i values using Equation 4 ;
  if *Equation 4 value $|\gamma_i| \geq \theta_i$ or sensor data measurements fall outside the 95% confidence level* **then**
    Flag a neighbouring sensor node measurement as outlying;
    Predict the outlying sensor measurements using GPR method
    Correct the outlying sensor measurements;
  else
    Transmit sensor measurements to level-2 cluster heads for data aggregation and onward processing
  Transmit sensor measurements to level-2 cluster heads for data aggregation and onward processing

V. Predicting Outlying Sensor Measurements

Sensor data that is corrupted due to environmental factors and system failures can be predicted, and these predicted values can be used to replace the corrupted values, preserving the data structure more effectively than simply discarding the outliers.

Gaussian Processes (GPs) are state-of-the-art tools for modeling many machine learning tasks [37], [38]. They are a powerful modeling framework that incorporates kernels and Bayesian inference, offering several benefits over traditional machine learning frameworks such as linear models and support vector machines

A. Gaussian Process

A Gaussian Process (GP) is formally defined as a collection of a finite number of random variables indexed by time or space (a stochastic process), such that every finite linear combination of these variables follows a multivariate normal distribution [37]. Formally, a function f is a $\mathcal{GP}$ if a finite set of values $f(x_1), \ldots, f(x_n)$ follow a multivariate normal distribution, where the inputs $\{x_n\}_{n=1}^N$ correspond to arbitrary sized domain vectors. GPs can be viewed as a probabilistic generalization of the notion of a multivariate function, where the function and its derivatives are governed by a Gaussian distribution. This property allows GPs to capture the inherent uncertainty and noise present in real-world data, making them well-suited for modeling many machine learning tasks.

B. Gaussian Process Regression

Regression is a supervised machine learning problem that aims to predict the relationship between an attribute set x and the response variable y indexed either by time or space, which can be continuous or discrete. Both linear and non-linear models have been used in regression analysis. Rather than assuming a defined function $f(x)$, Gaussian Process Regression (GPR) can represent $f(x)$ blindly but rigorously, letting the data define the underlying relationship. GPR is a supervised learning method that provides both predictions and the associated confidence levels. GPR modeling can handle both linear and nonlinear functions as base functions, making it less parametric and more nonparametric in nature. GPR provides a flexible and natural mechanism for selecting between simple and complex models for various inputs and applications.

Given a data set $\mathcal{D}$ of n observations, $\mathcal{D} = (x_i, y_i) \mid i = 1, \ldots, n$, where $x_i \in \mathbb{R}^d$ describes the attributes and y_i is the scalar output of the i-th data point, GPR finds a function f that associates the x_i to y_i using appropriate $\mathcal{GP}$ parameters. $\mathcal{GP}s$ are completely defined by specifying two functions, the mean $m(x)$ and covariance $cov(x, x')$ functions [37], [38]. Hence, $\mathcal{GP}$ can be formally written as:

$$f(x) \sim GP(m(x), cov(x, x')) \tag{6}$$

The covariance function is one of the most important factors in Gaussian Process ($\mathcal{GP}$) implementation. It is often based on the Euclidean distance between two attribute sets x and x', as defined in Equation 7 [37], [38]:.

$$cov(x, x') = k(\|x - x'\|), \tag{7}$$

for a decreasing function k.

The mean function $m(x)$ and covariance function $cov(x, x')$ are typically predefined prior to $\mathcal{GP}$ implementation. These prior properties can be determined from sample data [37].

One of the most commonly used covariance functions is the squared exponential (SE) covariance function defined in Equation 8.

$$cov(x, x') = \sigma_f^2 \exp(-\frac{1}{2\ell^2}\|x - x'\|^2) \tag{8}$$

However, a novel approach that folds noise into the covariance function, as shown in Equation 9, can be used to increase flexibility and adaptability. In this work, the SE covariance function defined in Equation 9 was used to implement GPR for predicting outlying or missing sensor-measured values.

$$cov(x, x') = \sigma_f^2 \exp(-\frac{1}{2\ell^2}\|x - x'\|^2) + \sigma_n^2 \delta(x, x'), \tag{9}$$

The function describes the covariance between x and x' and the parameter ℓ controls the "influence range", meaning how much the data points influence each other. A larger ℓ means a greater influence range. The Kronecker delta function δ indicates whether two data points are the same or different. The maximum allowable covariance is defined as σ_f^2. If $x \approxeq x'$, then $cov(x, x')$ approaches this maximum, meaning $f(x)$ is nearly perfectly correlated with $f(x')$.

Therefore, for the function to be smooth, neighbours must be alike. The influence of separation will depend on the length parameter, ℓ, providing much flexibility in the definition of the SE covariance function.

VI. Data, Implementation and Discussion

In order to test the localized outlying detection framework outlined in this paper, we used WSN data from a weather station in England locate at four (4) different locations shown at the following websites [39]–[42] Each sensor was capable of taking readings of several variables associated with local weather conditions, such as wind speed, tide height, and air temperature. A new reading for these variables is recorded every minute and stored, while every five minutes, the reading is transmitted and uploaded to the internet. The files from the websites, therefore, provide daily measurements in a single file that covers a 24-hour period at 5-minute intervals. In the cluster formation process, we assumed that

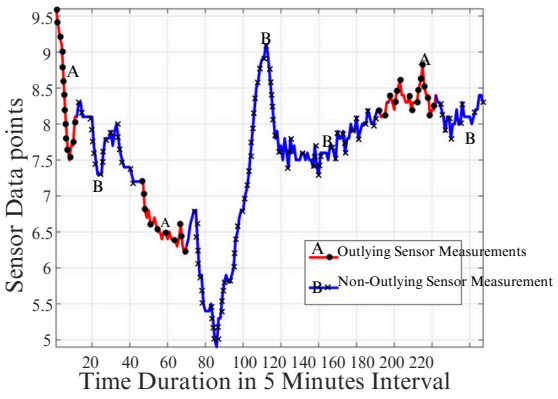

Fig. 4: Outlying Sensor measurement of Air Temperature (°C)

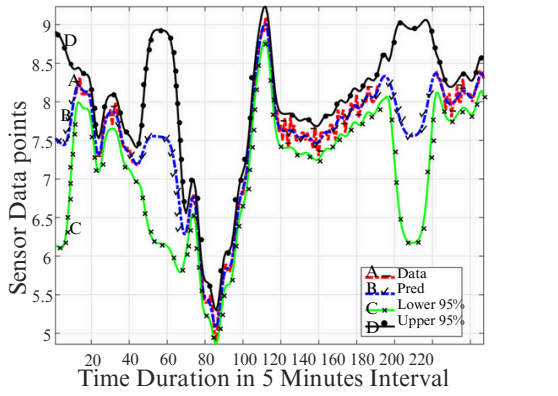

Fig. 5: Predicted outlying Air Temperature(°C)

sensor measurements within a month were a cluster, and each of the daily measurements came from individual sensors. And thus, in the ideal situation, these sensor measurements should have similar values within the one-month duration. The same scenario is applied to the remaining 3 sensor locations. The sensor measurement data are accessed from May 1–May 31, May 10th sensor measurement is not available, as shown in Figure 3. Using the sensor measurements from the http://www.chimet.co.uk weather station as a baseline for the studies, thus for a month May with 30 days of sensor measurements data, it implies that each cluster will have at most 30 sensors and the entire monitoring area will then have about 120 sensors in total.

In using the 2-level hierarchical clustering framework outlined in Section III-B,two of these sensors are elected as Level 1 and level-2 cluster heads for outlying detection, correction, and data aggregation and transmission, respectively, based on their residual energy and their rate of variability. The level-1 cluster head is the reference sensor selected based on its high residual energy and low rate of data variability, and each of the remaining 28 sensor measurements within each cluster was compared with the reference sensor's measurement. During the simulation process, it was realized that for sensor measurements from the **Chimet** sensor websitehttp://www.chimet.co.uk, the May 22 measurement has a low variability indexγ as shown in Figure 3 aand May 22, which is taken to be a sensor node, is assumed to have the highest residual energy during the LEACH protocol routing

and is therefore used as the reference sensor.

The referenced sensor in each cluster, as shown in Figure 2, was used for comparison with the neighboring sensor measurements. Sensor measurements with $|\gamma_i| \geq \theta$ *($\theta > 1$) or sensor data measurements that fall outside the 95%* confidence level are classified as outlying, and their values at that time stamp are replaced with those of the referenced sensor measurement at the same time stamp. In our simulation, the value of θ is given by Equation 5 in Section IV-B.

Five sensor measurements were used to evaluate the proposed model, as discussed below.

(i) Sensor Measurement of Air Temperature: Figure 4 shows the case of an outlying sensor measurement of air temperature(°(C)) obtained from the http://www.chimet.co.uk where a sensor measurement is compared with the referenced sensor measurement, x_i.
The outlying measurements are discarded (as shown in the red line) and consequently, these discarded values were predicted using the GPR method, as shown in Figure 5.

(ii) Sensor Measurement of Wind Direction: Figures 6 show a similar case for sensor measurements for the wind direction obtained from http://www.chimet.co.uk. The outlying sensor measurements shown in red were discarded, and their predicted values were obtained using GPR. The result are shown in Figure 7.

(iii) Sensor Measurement of Water Depth: Figures 8 also show the case for sensor measurements for the Water depth obtained from http://www.chimet.co.uk.The Sensor measurements with $|\gamma_i| \geq \theta$ *($\theta > 1$) or sensor data measurements that fall outside the 95%* confidence level are classified as outlying and their values at that time stamp are replaced with those of the referenced sensor measurement at that same time stamp as shown in Figure 9.

(iv) Sensor Measurement of Wind Speed: Figures 10 show a similar case for sensor measurements for the wind speed obtained from http://www.chimet.co.uk. Similarly, our model was able to detect the outlying sensor measurements, flag the sensor at that time instance as an outlying sensor, and trigger the GPR method to predict the outlying sensor measurements.

(v) Sensor Measurement of Maximum Gust: Figures 12 shows the case for a sensor measurements for Maximum Gust obtained from http://www.chimet.co.uk. Similarly, the outlying sensor measurements were detected and flagged, and the GPR method was used to predict them, as shown in Figure 13.

The results demonstrated that the proposed framework will be an effective method for detecting, predicting and correcting outlying sensor measurements when adopted as a middleware in the WSN architecture

VII. Conclusion

Sensors are important tools, and their application domains are wide and relevant. When deployed to monitor a given area (forming a network), they provide sensitive data for making quality decisions that affect almost every aspect of our lives, ranging from health and commercial to military applications. However, sensor networks and their delivery systems are vulnerable to several types of failure and disruption. This can include intrusion attacks, node failures, link failures, malicious attacks, environmental variations, limited resource provisioning, and overburdening from redundant data processing tasks. Additionally, failures can occur because of problems with the measuring instruments. As these failures occur, the data produced by the sensors contains inaccurate information and may not be reliable for making critical decisions. Data fusion methods from densely deployed sensors is an attempt, though, to compensate for these errors, but this method may not also be able to provide correct information if a majority of the sensors are reporting erroneous data. Therefore, there is a need to provide an efficient and scalable approach for handling outlying, missing, or anomalous sensor measurements. We proposed a 2-level hierarchical LEACH protocol approach to elect a referenced sensor with predefined conditions for detecting, flagging, and predicting missing, outlying, and anomalous sensor measurements. We then leveraged Gaussian Process Regression (GPR) as an effective and efficient machine learning method for predicting outlying or missing sensor measurements. We tested the outlying detection, prediction and correction framework by using weather station data. The results demonstrated that the framework is an effective way to detect, predict, and correct outlying sensor measurements.

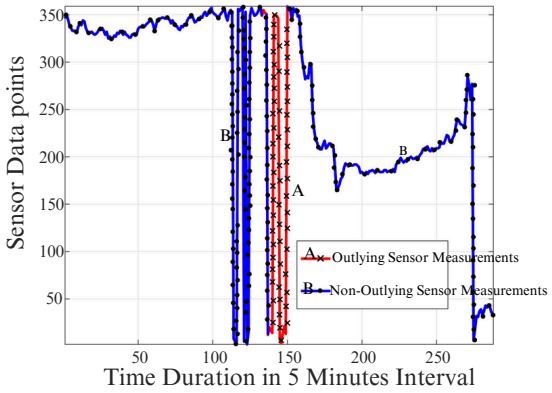

Fig. 6: Outlying Sensor measurement of Wind Direction (°C)

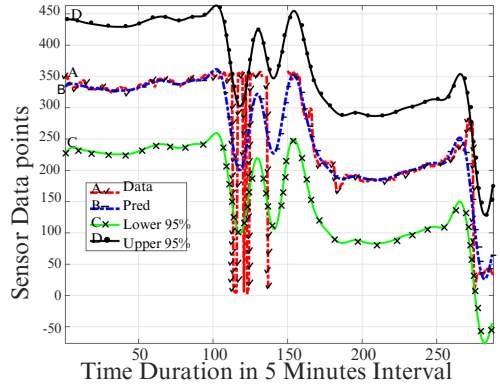

Fig. 7: Predicted Outlying Sensor Measurement of Wind Direction (°C)

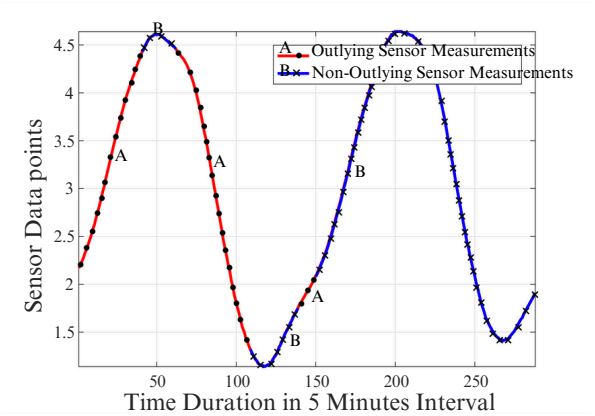

Fig. 8: Outlying Sensor measurement of Water Depth (m)

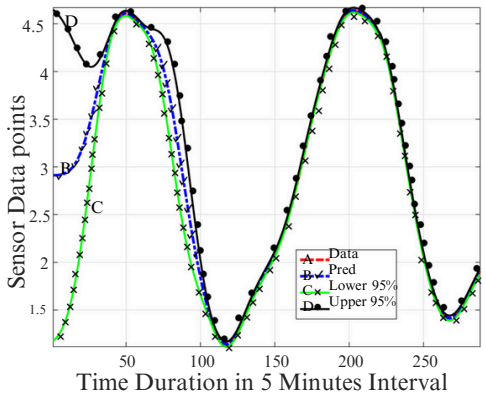

Fig. 9: Predicted Outlying Sensor Measurement of Water Depth (m)

As a future research direction, we shall use fuzzy logic and deep learning together with novel data fusion approaches for the detection and prediction of outlying sensors.

REFERENCES

[1] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *computer networks*, vol. 52, no. 12, pp. 2292–2330, 2008.

[2] H. Alemdar and C. Ersoy, "Wireless sensor networks for healthcare: A survey," *computer networks*, vol. 54, no. 15, pp. 2688–2710, 2010.

[3] C. F. García-Hernández, P. H. Ibarguengoytia-Gonzalez, J. García-Hernández, and J. A. Pérez-Díaz, "Wireless sensor networks and applications: a survey," *IJCSNS International Journal of Computer Science and Network Security*, vol. 7, no. 3, pp. 264–273, 2007.

[4] S. Yassine, "A routing protocol for the wireless body area sensor network (wbasn)," *IAENG International Journal of Computer Science*, vol. 49, no. 2, pp. 279–285, 2022.

[5] P. R. Bhaskaran, J. D. Rathnam, S. Koilmani, and K. Subramanian, "Multiresonant frequency piezoelectric energy harvesters integrated with high sensitivity piezoelectric accelerometer for bridge health monitoring applications," *Smart Materials Research*, vol. 2017, 2017.

[6] C.-S. Tu, C.-H. Chang, S.-C. Chang, C.-S. Lee, and C.-T. Chang, "A decision for predicting successful extubation of patients in intensive care unit," *Internationa journal of BioMed Researchl*, vol. 2018, p. 11, 2018. [Online]. Available: https://doi.org/10.1155/2018/6820975

[7] Y. A. Bangash, Y. E. Al-Salhi *et al.*, "Security issues and challenges in wireless sensor networks: A survey." *IAENG International Journal of Computer Science*, vol. 44, no. 2, pp. 135–149, 2017.

[8] R. Vidhyapriya and P. Vanathi, "Energy efficient adaptive multipath routing for wireless sensor networks." *IAENG International Journal of Computer Science*, vol. 34, no. 1, pp. 56–64, 2007.

[9] L. Martí, N. Sanchez-Pi, J. M. Molina, and A. C. B. Garcia, "Anomaly detection based on sensor data in petroleum industry applications," *Journal of Sensors*, vol. 15, no. 2, pp. 2774–2797, 2015. [Online]. Available: http://www.mdpi.com/1424-8220/15/2/2774

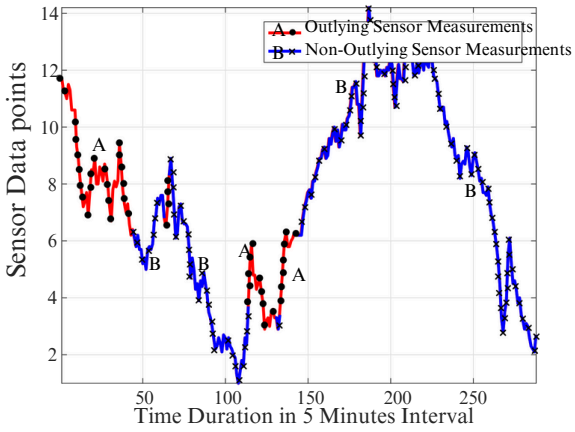

Fig. 10: Outlying Sensor Measurements of Wind Speed (knots)

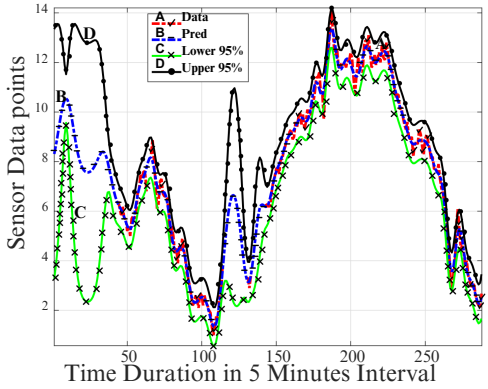

Fig. 11: Predicted Sensor Measurements of Wind Speed (knots)

[10] S. A. Haque, M. Rahman, and S. M. Aziz, "Sensor anomaly detection in wireless sensor networks for healthcare," *Sensors*, vol. 15, no. 4, pp. 8764–8786, 2015. [Online]. Available: http://www.mdpi.com/1424-8220/15/4/8764

[11] J. C. Socoró, F. Alías, and R. M. Alsina-Pagès, "An anomalous noise events detector for dynamic road traffic noise mapping in real-life urban and suburban environments," *Sensors*, vol. 17, no. 10, 2017. [Online]. Available: http://www.mdpi.com/1424-8220/17/10/2323

[12] D. Wang, J. Wan, M. Wang, and Q. Zhang, "An mef-based localization algorithm against outliers in wireless sensor networks," *Sensors*, vol. 16, no. 1041, 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/7/1041

[13] G. Chen, Y. Zhang, R. Huang, F. Guo, and G. Zhang, "Failure mechanism of rock bridge based on acoustic emission technique," *Sensors*, vol. 2015, p. 11, 2015. [Online]. Available: http://dx.doi.org/10.1155/2015/964730

[14] P. Kaur, "Outlier detection using kmeans and fuzzy min max neural network in network data," in *2016 8th International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE, 2016, pp. 693–696.

[15] A. Abid, A. Kachouri, and A. Mahfoudhi, "Outlier detection for wireless sensor networks using density-based clustering approach," *Journal of IET Wireless Sensor Systems*, 2017.

[16] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE communications surveys & tutorials*, vol. 12, no. 2, pp. 159–170, 2010.

[17] V. P. Bhuvana, C. Preissl, A. M. Tonello, and M. Huemer, "Multi-sensor information filtering with information based sensor selection and outlier rejection," *IEEE Journal of Sensors*, 2018.

[18] S. D. Patil and B. Vijayakumar, "An outlier detection scheme for wireless sensor networks," in *2016 5th International Conference on Wireless Networks and Embedded Systems (WECON)*. IEEE, 2016, pp. 1–6.

[19] F. Wang, L. Hu, J. Zhou, K. Zhao *et al.*, "A data processing middleware

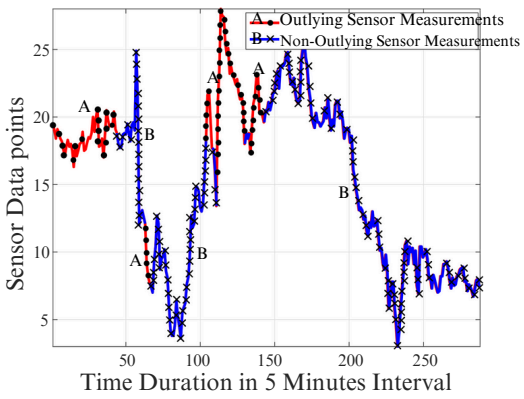

Fig. 12: Outlying Sensor Measurements of Maximum Gust (knots)

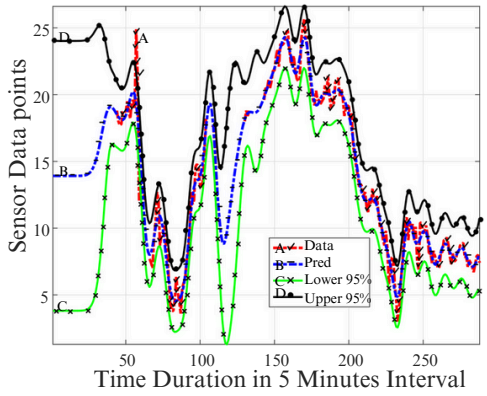

Fig. 13: Predicted Sensor Measurements of Maximum Gust (knots)

based on soa for the internet of things," *Sensors*, vol. 2015, 2015.

[20] W. Wu, X. Cheng, M. Ding, K. Xing, F. Liu, and P. Deng, "Localized outlying and boundary data detection in sensor networks," *IEEE Journal of transactions on knowledge and data engineering*, vol. 19, no. 8, pp. 1145–1157, 2007.

[21] H. H. Bosman, G. Iacca, H. J. "W ́ortche", and A. Liotta, "Online fusion of incremental learning for wireless sensor networks," in *2014 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2014, pp. 525–532.

[22] Q. Chen, L. Liu, Z. Yang, and K. Guo, "Prediction approach of critical node based on multiple attribute decision making for opportunistic sensor networks," *Sensors*, vol. 2016, 2016.

[23] R. K. Sundararajan and U. Arumugam, "Intrusion detection algorithm for mitigating sinkhole attack on leach protocol in wireless sensor networks," *Sensors*, vol. 2015, p. 12, 2015. [Online]. Available: http://dx.doi.org/10.1155/2015/203814

[24] M. Jiang, L. Jiang, D. Jiang, F. Li, and H. Song, "A sensor dynamic measurement error prediction model based on napso-svm," *Sensors*, vol. 18, no. 233, 2018. [Online]. Available: http://www.mdpi.com/1424-8220/18/1/233

[25] C. Carvalho, D. G. Gomes, N. Agoulmine, and J. N. de Souza, "Improving prediction accuracy for wsn data reduction by applying multivariate spatio-temporal correlation," *Sensors*, vol. 11, no. 11, pp. 10 010–10 037, 2011. [Online]. Available: http://www.mdpi.com/1424-8220/11/11/10010

[26] G.-W. Lee, S.-Y. Lee, and E.-N. Huh, "Congestion prediction modeling for quality of service improvement in wireless sensor networks," *Sensors*, vol. 14, no. 5, pp. 7857–7880, 2014. [Online]. Available: http://www.mdpi.com/1424-8220/14/5/7857

[27] H. Yu, Z. Wu, S. Wang, Y. Wang, and X. Ma, "Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks," *Sensors*, vol. 17, no. 1501, 2017. [Online]. Available: http://www.mdpi.com/1424-8220/17/7/1501

[28] S. Bandyopadhyay and E. J. Coyle, "An energy efficient hierarchical clustering algorithm for wireless sensor networks," in *Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies INFOCOM*, vol. 3. IEEE, 2003, pp. 1713–1723.

[29] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proceedings of the 33rd annual Hawaii international conference on System sciences*. IEEE, 2000, pp. 3005 – 3014.

[30] J. Wang, Z. Zhang, F. Xia, W. Yuan, and S. Lee, "An energy efficient stable election-based routing algorithm for wireless sensor networks," *Sensors*, vol. 13, no. 11, pp. 14 301–14 320, 2013. [Online]. Available: http://www.mdpi.com/1424-8220/13/11/14301

[31] T. M. Rahayu, S.-G. Lee, and H.-J. Lee, "A secure routing protocol for wireless sensor networks considering secure data aggregation," *Journal of Sensors*, vol. 15, no. 7, pp. 15 127–15 158, 2015. [Online]. Available: http://www.mdpi.com/1424-8220/15/7/15127

[32] A. E. Tümer and M. Gündüz, "Energy-efficient and fast data gathering protocols for indoor wireless sensor networks," *Sensors*, vol. 10, no. 9, pp. 8054–8069, 2010. [Online]. Available: http://www.mdpi.com/1424-8220/10/9/8054

[33] O. O. Ogundile and A. S. Alfa, "A survey on an energy-efficient and energy-balanced routing protocol for wireless sensor networks," *Sensors*, vol. 17, no. 1084, 2017. [Online]. Available: http://www.mdpi.com/1424-8220/17/5/1084

[34] W. B. Heinzelman, A. P. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Transactions on wireless communications*, vol. 1, no. 4, pp. 660–670, 2002.

[35] V. Barnett, T. Lewis *et al.*, *Outliers in statistical data*. Wiley New York, 1994, vol. 3, no. 1.

[36] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.

[37] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, 2006, vol. 1.

[38] C. K. Williams, "Prediction with gaussian processes: From linear regression to linear prediction and beyond," in *Learning in graphical models*. Springer, 1998, pp. 599–621.

[39] Cambermet Limited, "Cambermet." [Online]. Available: https://www.cambermet.co.uk/

[40] Bramblemet Limited, "Bramblemet." [Online]. Available: https://www.bramblemet.co.uk/

[41] Sotonmet Limited, "Sotonmet." [Online]. Available: https://www.sotonmet.co.uk/

[42] Chimet Limited, "Chimet." [Online]. Available: https://www.chimet.co.uk/

EDWARD YELLAKUOR BAAGYERE (M'16) received the B.Sc. degree (Hons.) in computer science from the University for Development Studies (UDS), Tamale, Ghana, in 2006, the M.Phil. degree in Computer Engineering from the Kwame Nkrumah University of Science and Technology (KNUST), Kumasi, Ghana, in 2011, and the DEng. degree in Computer Science and Technology from the University of Electronic Science and Technology of China, in 2016.

He is currently an Associate Professor with the Department of Computer Science, and the Dean of the School of Computing and Information Sciences at the C. K. Tedam University of Technology and Applied Sciences (CKT-UTAS), Navrongo, Ghana.

Professor Baagyere current research interests include Deep/Machine Learning and Application, Mobile Sensor Networks, Cryptography and Application, Social Networks, and Internet of Things.

REGINA ESI TURKSON obtained her PhD in computer science and technology at the University of Electronic Science and Technology of China (UESTC), Chengdu, China. She is currently a Lecturer/Researcher at the Department of Computer Science and Information Technology at the University of Cape Coast (UCC), Ghana. Her research interest includes Artificial Intelligence, Machine Learning, Neural Networks and Computer Security and Cryptography.

GIDEON EVANS WENYA is a research fellow and a lecturer in the School of Advanced Technologies, Engineering and Science (SATES), Accra Institute of Technology, Ghana. He received his B. Ed. degree in Mathematics Education from the University of Education, Winneba, Ghana, in 2010 , and the M.Eng degree in Electronic Science and Technology at the University of Electronic Science and Technology of China (UESTC) Chengdu, Institute of Fundamental Frontier Sciences(IFFS), China, in 2015. His current research interest include nano-fabrication of metal nanoparticles in thin films , Nano-materials for Energy, in particular their processing, properties, structure, performance in thin film materials (dielectrics and composites).

IVEN AABAAH obtained his BSc and MSc Degrees in computer science from the Kwame Nkrumah University of Science and Technology (KNUST), in 2015, and 2019 respectively. He is currently a PhD candidate at the C. K. Tedam University of Technology and Applied Sciences (CKT-UTAS), Ghana. He is an Assistant Lecturer with the Department of Information Systems and Technology at CKT-UTAS. His research areas include Wireless Networks, Ad-Hoc Wireless Networks, Wireless Network Security, and Distributed Data Management.