Impacts of wireless sensor networks strategies and topologies on prognostics and health management

Ahmad Farhat, Christophe Guyeux, Abdallah Makhoul, Ali Jaber, Rami Tawil, Abbas Hijazi

Impacts of wireless sensor networks strategies and topologies on prognostics and health management

Ahmad Farhat, Christophe Guyeux, Abdallah Makhoul[a], Ali Jaber, Rami Tawil, Abbas Hijazi[b]

[a]*FEMTO-ST Institute, UMR 6174 CNRS, Université de Bourgogne Franche-Comté, France*
[b]*Department of Computer Science, Lebanese University, Beirut, Lebanon*

Abstract

In this article, we used Wireless Sensor Network (WSN) techniques for monitoring an area under consideration, in order to diagnose its state in real time. What differentiates this type of network from the traditional computer ones is that it is composed by a large number of sensor nodes having very limited and almost nonrenewable energy. A key issue in designing such networks is energy conservation because once a sensor depletes its resources, it will be dropped from the network. This will lead to coverage hole and incomplete data arriving to the sink. Therefore, preserving the energy held by the nodes so that the network keeps running for as long as possible is a very important concern. If we achieve to improve the network lifetime and Quality of Service (QoS). Diagnosing the state of area will be more accurate for a longer time. One of the most important elements to achieve a QoS in WSN is the network coverage which is usually interpreted as how well the network can observe a given area. Obviously, if the coverage decreases over time, the diagnosis quality decreases accordingly. Various coverage strategies are thus proposed by the WSN community, in order to guarantee a certain coverage rate as long as possible, to reach a certain QoS that in turn will impact the diagnosis and prognostic quality. Various other strategies are in common use in WSN like data aggregation and scheduling, to preserve a QoS in wireless sensor networks, as long as possible. We argue that such strategies are not neutral if this network is used for prognostic and health management. Some politics may have a positive impact while other ones may blur the sensed data, like data aggregation or redundancy suppression, leading to erroneous diagnostics and/or prognostics. In this work, we will show and measure the impact of each WSN strategy on the resulting estimation of diagnostics. We emphasized several issues and studied various parameters related to these strategies that have a very important impact on the network, and therefore on data diagnostics over time. To reach this goal, to evaluate both prognostic and health management with the WSN strategies, we have used six diagnostic algorithms.

Keywords: Wireless Sensor Networks, Coverage, Scheduling Mechanisms, Topology, Prognostic and Health Management, Diagnostics, machine learning algorithms.

1. Introduction

Modern industrial plants and areas for military, agricultural, health purposes, are always at risk of failure due to fire, robbery, attack, etc., which will be dangerous and costly since the costs of failure and system downtime are getting pricey. This is why, it is very necessary to evaluate their health and diagnose them at any time, and then plan maintenance activities to avoid disastrous failure results. Prognostic and Health Management (PHM) is a process allowing an advanced system to automatically test the area, diagnose it, isolate the failure, and try predicting the Remaining Useful Life (RUL) for this area before failure occurs [61]. A maintenance scheduling is then determined and the area shutdown is prevented. But if the prediction model and the provided measurements are not accurate, the maintenance activity will not be done on time.

Online measurements of the operating conditions are required for assessing health and diagnosing activity of the area of interest, followed by RUL prediction. A number of sensor nodes usually gather these information. In this study, we consider the case where sensors communicate their information within a Wireless Sensor Network. Unlike the conventional computer networks, WSNs are composed of a large number of sensor nodes with very limited and non-renewable energy. Most of the time, they are deployed to capture the occurrence of possible events in hostile and inaccessible areas [74, 26, 55]. However, there is a classical assumption in PHM that the monitored data are available and complete, which is not always true. Indeed, because of the nature of communication in this network and characteristics of its devices, a WSN is at risk of failure. The accuracy and completeness of data that is going to be captured will be affected by this risk, and consequently PHM will be affected [19, 21, 22]. Therefore to ensure that the data of the monitored area is as accurate as possible, we need to maintain the Quality of Service (QoS) of WSN for the longest possible time.

WSN strategies that can be considered are: the topologies [41, 73, 40], coverage [25, 30], deployment strategy [52, 17], scheduling mechanism [71, 30, 63, 64, 31], density [3], security [51, 4], data aggregation [37, 75, 44], packet transfer distance, battery, memory, etc. These strategies have a strong impact on the effectiveness of the network with time, consequently on the quality of the data that will be captured in the monitored area, and finally on PHM. Coverage, for instance, is in a close relation with energy consumption of sensors [77]: the challenge remains in the efficient use of sensors to increase the lifetime of the network while maximizing the coverage. For this reason, before deployment, a prior study of the network is needed which is the only way to ensure that the area is well covered, leading to the reception of accurate data. Although both PHM and WSN strategies have been studied, as far as we know, none of the existing research work has considered the WSN strategies *for* PHM. For instance, in real life application, the monitored area might not be fully covered for several reasons like nodes failure, energy depletion, unadapted initial deployment, scheduling mechanisms, etc. Obviously, this flaw may lead to inaccurate and incomplete data, and if we do not take such issue into consideration while building a PHM process over a WSN, the results provided by diagnostic or prognostic may not be reliable.

In this work, we study the WSN strategies and their relation with prognostic and health management. We focus on the impact of these strategies on the accuracy of the data captured by a wireless sensor network, and its consequences on the health diagnostic of the monitored area. Diagnostic is a very important part in PHM, in which the determination of RUL begins by identifying the current health state of the system. Our objective is to show that usual diagnostic processes that perform well in classical data provided by a well deployed wired network of sensors, may face a dramatic decrease of performances in the case where data are obtained via a WSN, due to the diversity and variation of WSN strategies usual in such networks. For that, we used six machine learning algorithms to diagnose the area state, namely the so-called Support Vector Machines (SVM), Naive Bayes (NB), Random Forests (RF), Gradient Tree Boosting (GTB), Tree-Based Feature Selection (TBFS), and Nearest Neighbors (NN) methods. Then we studied the behavior of these algorithms when particular issues, inherent to the perfectible nature of WSNs, are present. We focused in this study on several important strategies that affect the coverage in WSN, that is, the ones resulting from usual energy consumption of sensors, scheduling mechanisms of sensors, and the problems related to the density and deployment of sensors frequently reported in the wireless sensor network community. In addition, we study the topologies in WSN and its impact on the diagnostics. We used four different types of topology (the most used ones in WSN) which are: distributed, hierarchical, centralized, and decentralized topologies in order to show and study several parameters and issues for each type of topology like density, security, data aggregation, frequency, packet transfer distance, and energy consumption that have an important impact on the diagnostics over time, and therefore on PHM.

This research work presents the issues occurring usually in WSNs strategies that are relevant to consider for diagnostics. The remainder of this article is organized as follows. Section 2 outlines an overview of WSNs and PHM, and we detail the links that can be established between the WSN strategies and the PHM field or research. We simulate and describe the strategies of WSN to show their impact on the quality of data and therefore on the diagnostics. The results of these simulations are given in Section 3. This article ends with a conclusion section, where the contribution is summarized and intended future work is provided.

2. State of the art

2.1. Wireless Sensor Networks

WSNs are event-based systems that rely on the collective effort of several sensor or micro-sensor nodes [5, 7, 6]. This type of network tends to greatly increase the coverage rate for the area that should be monitored and also to increase the accuracy of the information extracted from this area. A sensor node is a tiny device that has the capability to sense new events, compute the sensed values, and communicate information. WSNs are composed by very small sensors with very limited and non renewable energy that can be deployed when monitoring physical and environmental phenomena such as temperature, vibrations, light, humidity, etc [16]. For this energy to be preserved, network throughput has to be low. Another issue is that, as all wireless networks, WSN are not very secure. Both energy limitation and random deployment in hostile and inaccessible areas can cause a node failure (or attack) [70].

Sensor networks may consist of many different types of sensors such as seismic, low sampling rate magnetic, thermal, visual, infrared, acoustic and radar, which are able to monitor a wide variety of ambient conditions [24]. Sensor nodes can be used for continuous sensing, event detection, event ID, location sensing, and local control of actuators. The concept of micro-sensing and wireless connection of these nodes promises many new application in military, environment, health, home, and commercial areas. It is even possible to expand this application classification with more categories such as space exploration, chemical processing, and disaster relief.

Various characteristics are reported for such kind of networks, they are recalled in what follows.

2.1.1. Coverage

Sensor nodes have a short radio range and they collaborate to cover a given surveillance area. The coverage problem arises as: how to ensure that, at any time, any zone in the network is covered by at least one sensor node [65].

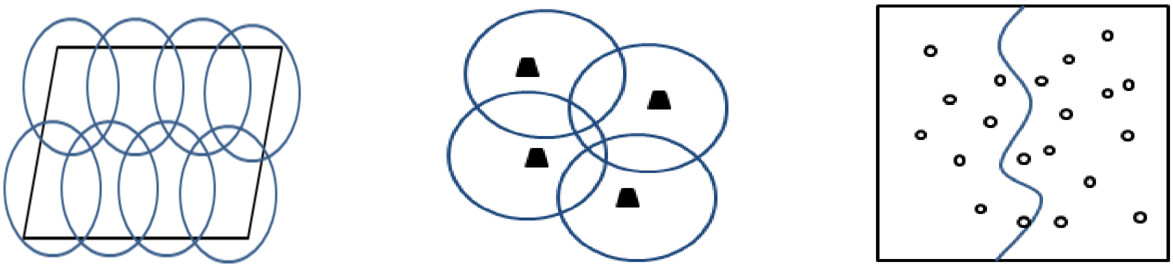
Coverage is in a close relation with energy consumption of sensors. A basic and important function of WSN is to monitor areas or targets for a long period, such as fire monitoring and environment detection. And a critical issue in the WSN applications is the coverage problem because sensors are often deployed in remote or inaccessible environments or are spread in an arbitrary manner [69, 39]. Therefore, the challenge remains in the efficient use of these sensors to increase the lifetime of the network while maximizing the coverage. Indeed, in the WSNs community, coverage is one of the most active areas of research, in which the measurement of how well and for how long the sensors are able to observe the physical space usually defines the coverage problem.

For many years, a lot of works have been dedicated to the coverage-related issues in WSNs since it is a fundamental problem [35, 8, 66, 2, 54, 11]. There is limited energy resource in each sensor node, so this makes energy conserving of the sensors, and prolonging the network lifetime while maximizing the coverage of areas or targets, an important and difficult issue in the applications of WSN. Indeed energy is a very critical resource and must be used very sparingly. Many coverage algorithms were found in recent years, but problems still exist in them [79, 72, 31, 36]. These algorithms are often based on the subject to be covered (area versus discrete points), sensor deployment mechanism (random versus deterministic) as well as other WSN properties (*e.g.*, minimum energy consumption and network connectivity) [43].

Regarding energy preservation problem in WSNs, authors in [77] studied three different approaches, all maintaining the initial coverage QoS. The first approach focuses on optimizing coverage deployment strategy, while the second one consists of planning a scheduling of active sensors that enables other sensors to go into a sleep mode. Finally, the third approach is adjusting the sensing range of sensors for energy conservation. Moreover, recent works show that an efficient density control in high density sensor networks saves significant amounts of energy. Increasing the number of sensors in network makes them closer to each other in the area, which will facilitates the communication between them and reduce the packet transfer distance. This will reduce the energy consumption of sensors, therefore lifetime and coverage rate in WSN will increase. Thus the coverage rate and network lifetime are greatly related to number of sensors in the monitoring area [3, 59, 62, 20]. In order to attain reliability in WSNs (by fault prevention, removal, forecasting, and fault tolerance), sensing coverage and sensing level need to be considered [18]. Based on the researches done before, there are three types of problems related to coverage, which are: area, point, and barrier coverage [15, 43], as shown in Figure 1.

- **Area coverage:** this is the most popular coverage problem in WSNs, which has been widely studied for many years. The main objective of the sensor network is to cover or monitor an area (or region), *i.e.*, each point in the area should be covered, and the network lifetime should be maximized. Figure 1a shows an example of a random deployment of sensors to cover a given parallelogram-shaped area [35, 54].

- **Point coverage:** the objective in this problem is to cover a set of points (targets). It considers how to maximize the network lifetime such that all the objectives in the monitored area are covered. Figure 1b gives an example of monitoring the discrete targets in a WSN, the black nodes form the set of active sensors.



(a) Area coverage.

(b) Point coverage.

(c) Barrier coverage.

Figure 1: Types of coverage problems.

- **Barrier coverage:** barrier coverage problem is detecting the probability of a moving object found when crossing the deployment region of WSN. Its goal is to minimize the probability of undetected penetration through the barrier. Figure 1c shows an example of a general barrier coverage problem where the start and end points of the path are selected from bottom and top boundary lines of the area. The selection of the path depends on the objective.

2.1.2. Topologies

In wireless sensor networks, the connectivity of the network is established via radio transmission between sensors. For two sensors to be able to communicate, they must be within some critical range of each other, as transmission capability is finite. And the packet transfer distance between these sensors in this range depends on the topology used in WSN. A network is connected if any node can communicate with any other node, possibly using intermediate nodes as relays. The variability of this connectivity is due to node failures, introduction of additional nodes, variations in sensor location, which requires the adaptability of underlying network structures and operations. Since sensors may be spread in an arbitrary manner, one of the fundamental issues that arises in sensor networks in addition to coverage is thus the connectivity. In order to ensure connectivity and data accuracy in addition to coverage, WSNs use redundant coverage where multiple sensor nodes cover the same physical location. Therefore, coverage may vary across the network. A solution to save energy in the network rises on finding scheduling mechanisms. The objective of such mechanisms is to activate or deactivate redundant nodes while keeping as much as possible a dense coverage and thus ensuring connectivity.

Another metric to save energy in sensor networks is to reduce the amount of data collected and transmitted via the network. Data gathering in WSNs can be either periodic or event-driven [29]. In periodic applications [45, 46], data is gathered periodically, while in event-driven applications gathering depends on the occurrence of some events. In both cases, the goal of aggregation operations is to reduce energy dissipation by holding packets for as long as possible in intermediate nodes. All packets will be combined together before being forwarded in the network. It is obvious to see that a decrease in energy consumption leads to an increase in the overall delay, and vice versa. A reliable solution would aim at finding an acceptable trade off between energy consumption and delay in WSNs [28, 38]. All of these elements are finally related to the network topologies.

WSNs can be either heterogeneous or homogeneous [42]. In the latter, all nodes have the same role and characteristics. In the former, nodes have different roles: some nodes simply sense and forward information while others aggregate data, manage their area, perform computations, etc. Consequently, some of the nodes can be equipped with higher energy, longer radio range, etc. Several WSN topologies were used in existing monitoring applications, but all of them revolved around four different types (or models) of topologies which are: distributed, hierarchical, centralized, and decentralized ones. They are recalled hereafter.

- **Distributed topology:** in distributed topologies, there is no management of the network by the central node (or a region of it). They consist of a collection of nodes having equal roles. Therefore, no aspect of hierarchy is considered. No prior infrastructure is imposed before the network starts running; each node discovers its surrounding area and decides which node(s) to communicate with. This decision usually relies on the radio range and the transfer distance. Distributed topologies render the network's maintenance an easy task: if a node fails, its neighbors, within their sensing range, will establish new links with other nodes, and the network will continue to work normally.

- **Hierarchical topology:** the organization of sensor nodes can be in several levels, making a hierarchical topology (or a tree topology). Level 0 is represented by the root and there is no level above. From two adjacent levels, sensor nodes are connected in an end to end manner. The hierarchical model can be seen as three different layers: (1) the core layer (the root), which is enhanced for availability and performance, (2) the distribution layer, which implements policies and forwards messages, and (3) the access layer (the leaf nodes) that represents the access point to the network. Scalability is the advantage of Hierarchical WSN. The network is more manageable and the task of isolating and detecting faults is simplified due to the presence of different levels.
- **Centralized topology:** it is one of the easiest topologies to design and implement (also called star topology). All the sensor nodes have a simple task which is sensing new information and forwarding it to a central node where all the data processing will be proceeded with. One of the major problems of this topology is that it presents a single point of failure. The whole network will become paralyzed if a problem occurs at the central node: the data packet cannot be forwarded nor processed when a new event is detected.
- **Decentralized topology:** decentralized topologies are considered as a combination of the distributed and the centralized topologies. The network is divided into regions (or clusters) which are locally managed by a central node (called the Cluster Head CH). This topology offers a reasonable settlement between energy consumption and Quality of Service (QoS). In this type of topology, there is a reduction of congestion problem and the network no longer has a single point of failure.

2.2. Prognostic and health management

Maintenance is an important activity in industrial field. It is either performed to restore a machine/component, or to prevent it from breaking down. It aims at increasing system availability, readiness, and enhancing safety. Through time, different strategies have evolved in order to bring maintenance to its current state: condition-based and predictive maintenance. This evolution was caused by the increasing demand of reliability in industry. PHM is a tool to predict the RUL of engineering assets and is the key process of condition-based and predictive maintenance. Nowadays, industrial machines are required to avoid shutdowns while offering safety and reliability [53]. Research in PHM field has gained and was given a great deal of attention. Prognostic models are developed in an attempt to predict the RUL of machinery (or monitored area) before failure takes place. If the prediction model and the provided measurements are not accurate, the maintenance activity will possibly be performed either too soon or too late.

Corrective maintenance is the first form of maintenance. In this strategy, actions are only taken when the system breaks and can no longer perform the intended tasks; in fact, sudden shutdowns cost money and time in addition to client's trust and safety. Maintenance became a periodic activity for solving these problems. Domain experts depend on their knowledge and the observation of upcoming events to set time intervals so that the components are inspected and replaced if needed. Preventive maintenance (often called periodic) is performed regardless of the machine's condition which is considered the main drawback. But sometimes the machine can be in a healthy state so the maintenance will be unnecessary and will cost extra and avoidable fees. But even with periodic maintenance and inspections, random failures still occur. For that, in the early nineties, Condition Based Maintenance (CBM) was proposed and developed [32].

CBM is based on real-time observations. It is an on-line approach that assesses machine's health through condition measurements. CBM aims to increase the system reliability and availability, as any maintenance strategy, while reducing maintenance costs. This particular strategy has benefits which include avoiding unnecessary maintenance tasks and costs, as well as not interrupting the normal machine operations [32]. CBM decreases the number of maintenance operations and causes the influence of human error to be reduced. Predictive maintenance (PM) is a new maintenance that has recently emerged. Based on the current condition, it predicts the system health in the future and defines the needed maintenance activities accordingly. Extra tasks are required for shifting from traditional maintenance strategies to CBM and PM. These tasks encompass data analysis and modeling, system surveillance, and decision making support system. This scientific approach is called Prognostics and Health Management (PHM). PHM is the core activity of CBM and PM. The steps of PHM are: data acquisition, data processing, health assessment, diagnostics, prognostics, and decision making support [34], this is done following the steps described in Figure 2.

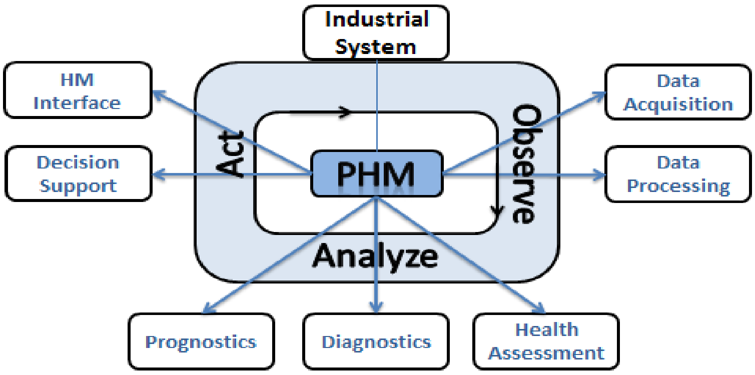


Figure 2: The process of PHM.

Obviously, Prognostics and Health Management requires information about the targeted physical assets. The data is acquired by means of sensors which are placed on/around the critical components. The stored data will later on be fed as inputs to the developed algorithm for health assessment, diagnostics, and/or prognostics [56, 9, 50, 67]. Diagnostics aim for specifying and quantifying an actual failure [58] while prognostics have the goal of anticipating failures. Prognostics consider the past events, in addition to the machine's current state, and operating conditions to estimate the RUL [34, 78, 67, 68, 48]. This estimation is done by studying the evolution of continuous measurements of parameters that need to be tracked in time to assess the machine's state. These parameters can be temperature, humidity, vibration, pressure, and so on. A monitored parameter has a fixed threshold. Once this threshold is reached, an alarm goes on indicating that a symptom of system deteriorating has been detected. And after that, a diagnosis of the state of the system is made, then the RUL is computed with an associated confidence limit. There are two causes for the uncertainties in RUL predictions: either the threshold value of monitored parameter, or the RUL prediction itself. Uncertainties concerning prediction of RUL are shown in Figure 3. Necessary prerequisites for reliable prognostics are proposed in [48].

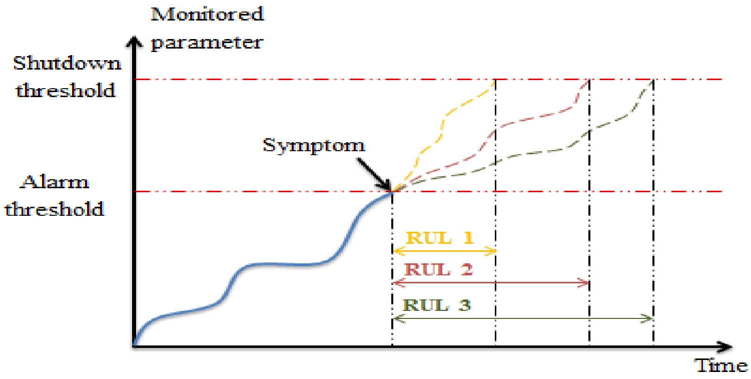


Figure 3: An illustration of RUL with uncertainties.

2.3. *Benefits of WSN in PHM*

Reliability is necessary in industry, or for any monitored area in general. This is the means to economic gain as well as client trust. For the past years, research in prognostics resulted in variety of tools and techniques that offer the possibility for plants to survey their systems, anticipate failures, and schedule maintenance activities. WSNs are mainly designed for surveillance purposes. They can be deployed in many fields such as military, automotive, agriculture, medicine, and so on [42]. Recently, a great deal of attention was given to WSN applications by industry. These sensor networks are used to monitor the machinery for maintenance scheduling. Furthermore, data will be provided by the sensors deployed to survey the system/component in order to assess the health, diagnose the system, and estimate the RUL. However, inaccuracy in the data will cause the prediction based on it to be irrelevant.

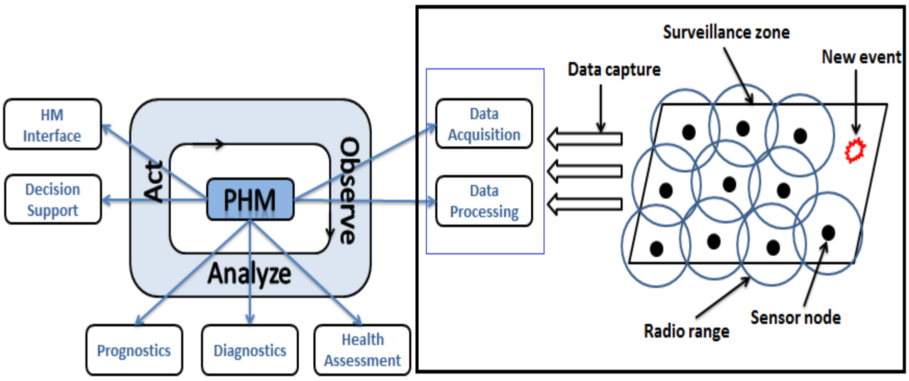


Figure 4: PHM steps with WSN monitoring.

WSN strategies evoked previously have important impact on the accuracy of data and therefore have an important impact on PHM [10, 20]. Then, before the network starts running, studying the strategies in details in WSNs need to be considered. As stated previously, our aim is to reveal the impact of such strategies on the accuracy of the captured data from the monitored area and therefore on PHM, using description and numerical simulations (see Figure 4). Indeed, since good predictions rely on real data, it is certain that the first step to be done in the research is ensuring a reliable source of information.

3. Numerical simulations

3.1. Machine learning algorithms

In this article, we focused on studying the diagnostic system (area monitoring) to examine its state: if failure is present or not at a certain time depending on the data captured by the sensors. As we mentioned before and as shown in Figure 2, diagnostic is a very important part in the PHM and the determining of RUL begins with identifying the system current state of health. For that, we use this step to evaluate the PHM, and observe how the process is affected by the strategies of WSNs.

The research in PHM is very broad and the authors working in this domain use several algorithms in order to perform the diagnostic of the system state. These methods are called machine learning algorithms in the literature. In machine learning, classification refers to identifying the class to which a new observation belongs, on the basis of a training set and quantifiable observations, known as properties. Machine learning displays a detailed study about the system and from it, an algorithm is built. These algorithms can be operated by building a model from example inputs, in order for the algorithm to be able to diagnose or take decision for new data.

We have chosen six machine learning algorithms to diagnose the system, which were used previously by several authors in the literature in order to evaluate their interest for PHM. We will then evaluate these six diagnostic algorithms in a WSN strategies situation. As stated previously, these algorithms are: Support Vector Machine (SVM) [27], Naive Bayes (NB) [76, 49], Gradient Tree Boosting (GTB) [14], Tree-Based Feature Selection (TBFS) [60], Nearest Neighbors (NN) [47], and Random Forests (RF) that are detailed in Section 3.2, as we are the firsts to use them for diagnostics based on data provided by a WSN [23]. These algorithms can be summarized as follows:

- **Support Vector Machine:** SVM is a learning technique that Vladimir Vapnik developed. In machine learning, SVMs are learning models that are supervised and are associated with learning algorithms that analyze data and recognize pattern, used to classify and regress pattern. A SVM training algorithm builds a model that attributes new examples in one category or the other relying on a certain set of data learning, each marked for belonging to one or two categories, making it a non-probabilistic binary linear classifier.

- **Naive Bayes:** they are direct acyclic graphs which are a synthesis of probability and graph theory that illustrates random variables and probabilistic inter-dependencies. They are constituted by a set of nodes that represent different states and directed edges that describe the transition probability between these states.

- **Gradient Tree Boosting:** GTB was introduced by Leo Breiman. It is a machine learning technique for problems related to regression and classification that produces a prediction model in the form of an entity of weak prediction models, typically decision trees. Based on the errors generated by the previous classifier, the distribution of the training set varies adaptively for each tree. In this study, we took GTB composed of 100 trees then the majority vote (by these 100 trees) is used to identify the class.

- **Tree-Based Feature Selection:** In machine learning and statistics, feature selection, also known as variable selection, attribute selection, or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection techniques are used for four reasons: (1) simplification of models to make them easier to interpret by researchers/users, (2) shorter training times, (3) to avoid the curse of dimensionality, (4) to enhance generalization by reducing overfitting (formally, reduction of variance). The central premise when using a feature selection technique is that the data contains many features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information. Redundant or irrelevant features are two distinct notions, since one relevant feature may be redundant in the presence of another relevant feature with which it is strongly correlated.

- **Nearest Neighbors:** NN is a non-parametric method used for classification. The input consists of the k closest training examples in the feature space. The output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor. NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The NN algorithm is among the simplest of all machine learning algorithms. NN can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of $1/d$, where d is the distance to the neighbor.

Finally we need a large and reliable data set in order to train these algorithms. And so, we can later diagnose the system (area monitoring) from the new data that will be captured by WSN.

3.2. The random forests algorithm

The RF algorithm is mainly the combination of Bagging [12] and random subspace [33] algorithms. It was defined by Leo Breiman as “a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest” [13]. This method resulted from a number of improvements in tree classifiers’ accuracy. This classifier maximizes the variance by injecting randomness in variable selection, and minimizes the bias by growing the tree to a maximum depth (no pruning). The steps of constructing the forest are detailed in Algorithm 1.

In a RF, the root of a tree i contains the instances from the training subset S'_i, sorted by their corresponding classes. A node is terminal if it contains instances of one single class, or if the number of instances representing each class is equal. In the alternative case, it needs to be further developed (no pruning). For this purpose, at each node, the feature that guarantees the best split is selected as follows.

1. The information acquired by choosing a feature can be computed through:
   (a) The entropy of Shannon, which measures the quantity of information

$$Entropy(p) = -\sum_{k=1}^{c} P(k/p) \times \log(P(k/p)) \tag{1}$$

where p is the number of examples associated to a position in the tree, c is the total number of classes, k/p denotes the fraction of examples associated to a position in the tree and labelled class k, $P(k/p)$ is the proportion of elements labelled class k at a position p.

Algorithm 1 Random forest algorithm

```
Input: Labeled training set $S$, Number of trees $T$, Number of features $F$.
Output: Learned random forest $RF$.
  initialize RF as empty
  for $i$ in $1..T$ do
    $S'_i \leftarrow$ bootstrap ($S$)
    initialize the root of tree $i$
    repeat
      if current node is terminal then
        affect a class
        go to the next unvisited node if any
      else
        select the best feature $f^*$ among $F$
        sub-tree $\leftarrow$ split($S'_i, f^*$)
        add (leftChild, rightChild) to tree $i$
      end if
    until all nodes are visited
    add tree i to the forest
  end for
```

(b) The Gini index, which measures the dispersion in a population

$$Gini(x) = 1 - \sum_{k=1}^{c} P(k/p)^2 \tag{2}$$

where x is a random sample, c is the number of classes, k/p denotes the fraction of examples associated to a position in the tree and labelled class k, while $P(k/p)$ is the proportion of elements labelled class k at a position p.

2. The best split is then chosen by computing the gain of information from growing the tree at given position, corresponding to each feature as follows:

$$Gain(p, t) = f(p) - \sum_{j=1}^{n} P_j \times f(p_j) \tag{3}$$

where p corresponds to the position in the tree, t denotes the test at branch n, P_j is the proportion of elements at position p and that go to position p_j, while $f(p)$ corresponds to either $Entropy(p)$ or $Gini(p)$.
The feature that provides the highest Gain is selected to split the node.

Finding the optimal training of a classification problem is often NP-hard. Tree ensembles have the advantage of running the algorithm from different starting points, and this can better approximate the near-optimal classifier.

3.3. A first case study oriented to intelligent manufacturing

We first intend to evaluate the 6 diagnostic algorithms previously detailed (SVM, NB, GTB, TBFS, NN, and RF) by comparing their efficiency on a real data set. To do so, we used data from the PHM08 Challenge Data Set available online [57]. This is a collection of data sets that have been donated by various universities, agencies, or companies. The data repository focuses exclusively on diagnostic data sets, *i.e.*, data sets that can be used for development and evaluation of diagnostic algorithms. There are mostly time series of data ranging from some nominal state to a failed one. Such data are divided into 2 parts: “data training” to learn the methods and “data testing” to test them. They consists of 26 columns of numbers, and a large number of rows. Each row is a snapshot of data taken during a single operational cycle. The columns correspond to:

1. Unit number
2. Time, in cycles
3. Operational setting 1
4. Operational setting 2
5. Operational setting 3
6. Sensor measurement 1
7. Sensor measurement 2
...
26. Sensor measurement 21

We have used these data to perform a first comparison of the efficiency of the 6 diagnostic algorithms presented previously. To do so, we used the python language and some of its modules available in [1] to compute the diagnostic algorithms. Then, the training and testing data mentioned earlier have been considered, and we evaluated the ability of these methods to predict the running time in cycles, given the remainder of each row (unit number, operational settings, and sensor measurements). The time needed to achieve a god training has been computed too, and obtained results are presented in Table 1.

Diagnostic algorithm	**Error rate (%)**	**Time to train (S)**
Support Vector Machine	20.3	1500.75
Naive Bayes	10.5	904.49
Gradient Tree Boosting	6.3	1850.52
Tree-Based Feature Selection	8.4	1233.85
Nearest Neighbors	17.6	720.14
Random forests	12.2	1011.63

Table 1: Error rates (%) in the prediction of the running time (in cycles) provided by various machine learning algorithms, and the time taken by these algorithms to perform a good learning stage (S).

We can notice first that the Support Vector Machine has the largest error rate (20.3 %). It is followed by Nearest Neighbors, Random forests, Naive Bayes, and Tree-Based Feature Selection, while the Gradient Tree Boosting is the most accurate method (having the lowest error rate, equal to 6.3 %). However, this Gradient Tree Boosting method takes more time than the others for the training stage on these real data (1850.52 ms). Considering again the time required to learn the data, GTB is followed by the Support Vector Machine, which definitively is malfunctioning on this set of data. Tree-Based Feature Selection and Random forests, for their part, achieve a good compromise between accuracy and speed, the former being a bit more accurate than the latter, but a bit slower too.

However, everything cannot be mastered in real data, like the real probabilistic law followed by each sensor. In particular, RULs are not available in this experimental set of data. As the other databases never provide all the information required for a complete performance analysis of the algorithms, we decided to perform further investigations on simulated data. Experimental protocols and obtained results are detailed below.

To explain more the data used and its importance, and confirm the results shown in Table 1, we studied the error rate of the six diagnostic methods in consideration, and by modifying the data training. Then, various percentages of the main data have been taken for training, and the impacts of this variation on the diagnostics have been studied, leading to the results shown in Figure 5. We can see that, as the percentage of data used for learning increases, the accuracy of the diagnostic of the six methods increases accordingly. Such results emphasize the well-known importance to have training data large enough for diagnostics.

Figure 6, for its part, shows in another way what we explained and clarified before. In this result, we studied the variation of the error rate with respect to the training time S. Indeed, in Table 1, we shown that each method takes a certain time t_t to perform a complete training. Our intention is now to study the variation of diagnosis accuracy for all methods with the time needed to do the training: as this time increases, error rates will obviously decrease until reaching the time t_t needed for a good training. Note that increasing the time over t_t is useless, which appears well in Figure 6: the error rate of a method is constant when the time given for training is greater than or equal t_t. For instance, the random forests algorithm needs 1011.63 second for its training, leading to an error rate equals to 12.2%.

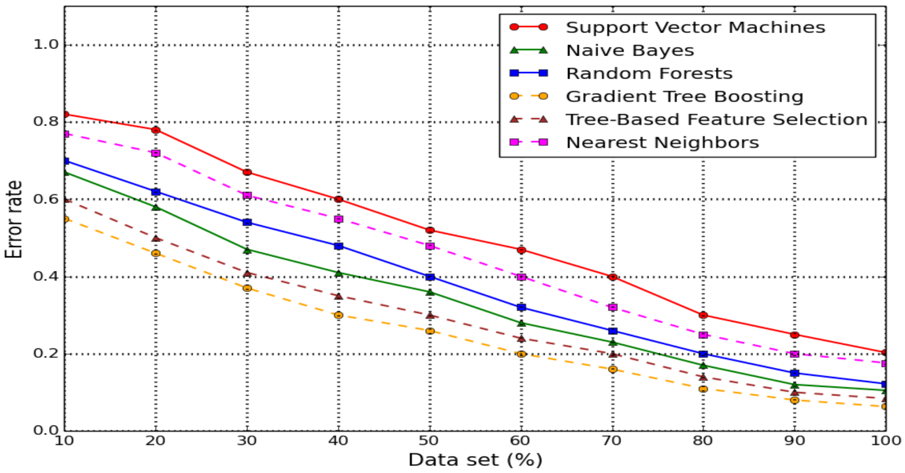


Figure 5: Error rate in diagnostics with respect to the percentage of data set to train.

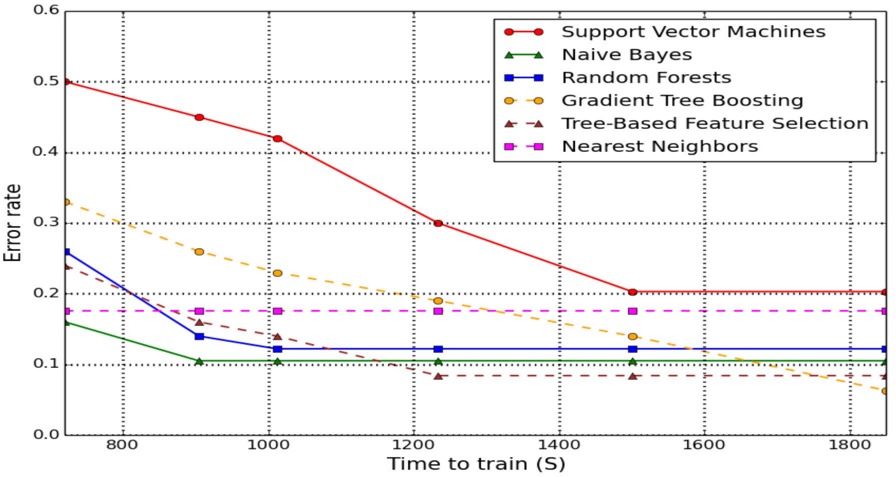


Figure 6: Error rate in diagnostics with respect to the time of training (S).

As can be seen in this situation, the error rate increases when the training time decreases, and conversely this error rate decreases as the time approached to t_t (it becomes constant when the training time is greater or equal to t_t). Therefore, these results illustrate on real data the importance of a good data training on the diagnostics accuracy.

3.4. Further experimental results on simulated data

In this section and by relying on our simulation, we will show the impact of various WSN strategies on diagnostics from several studies that we performed. These studies are shown and explained in the remainder of this paper.

3.4.1. WSN simulation

In order to simulate a WSN for monitoring the area under consideration and to show the impact of WSN strategies on the PHM, we used three types of sensing fields: temperature, pressure, and humidity. The number and parameters of each type of sensor depends on the type of study, as indicated in Section 3.4. Each sensor type captures specific data depending on the operating age t, and we consider that no level of correlation is introduced between the different features:

- Under normal conditions, temperature sensors follow a Gaussian law of parameter $(20\times(1+0.005t), 1)$. In case of a malfunction of the area in the range of this sensor, these parameters are mapped to $(350, 20)$. Finally, these sensors return the value 2 when they break down.

- The pressure sensors produce data following a Gaussian law of parameter $(5 \times (1 + 0.01t), 0.3)$ when they are sensing a well-functioning area. The parameters changed to $(20, 2.5)$ in case of area failure in the location where the sensor is placed, as long as the pressure sensors return 1 when they are broken down.

- The Gaussian parameters are $(52.5 \times (1 + 0.001t), 12.5)$ when both the area and the humidity sensors are in normal conditions. These parameters are set to $(80, 10)$ in case of area failure in the range of this sensor, whereas malfunctioning humidity sensors produce the value 3.

Each sensor follows a Poisson process (Pp) of parameter $(200 \times (1 - 0.01t) + 0.01)$, to determine if a breakdown occurs in the location where the sensor is placed. Subsequently all of these sensors execute the Algorithm 2.

Algorithm 2 Sensor algorithm

```
if $Pp < 1$ then
  the area and the sensors are in normal conditions
else
  if $1 \leq Pp < 100$ then
    the area is in failure (in the range of this sensor)
  else
    the sensor is broken down
  end if
end if
```

Each category of sensors has its own constant threshold, depending on the abnormality of the sensed data. If the data captured by the sensor in a specific category exceeded the threshold, this indicates that a symptom of system deteriorating has been detected. Then a diagnostic study aims at specifying and quantifying an actual failure (whether it failed or not). In this work, we used six algorithms for diagnosis which are mentioned in Section 3.1. Many applications of WSN exist like area monitoring, industrial monitoring, health care monitoring, environmental/earth sensing, etc. These applications have their own features and the threshold is related to those features. In this study, we chosen to simulate a WSN to monitor an area, to study later the strategies of WSN and their impact on PHM. Finally, we consider threshold values as follows: 26 degrees for temperature, 7 bars in pressure, and 80 percents of humidity.

The deployment strategy (manually or randomly) of sensors [52], the adjustment of the coverage radius of sensors [77], and the coverage in WSN (as we explained in Section 2.1.1) have an important impact on the accuracy of the data captured by WSN that will be used in PHM process. In order to study the impact of such strategies of WSN on PHM, in our simulation we consider the following hypotheses:

- Most of the times, the area to be monitored is hazardous and hard to access because of the difficulty in its geographical area like monitoring the forests, oceans, military zones, etc. Therefore we used random deployment for area monitoring.

- The region to be monitored is a rectangle of area $A = L \times W$, such that L and W are the length and width of the region respectively. The area of the coverage range of a sensor is mostly related to the area of the monitored region. Therefore we consider that the coverage area is set to be equal to 1% of the total area of the region. Subsequently, the coverage radius will be $R_c = 1/10 \times \sqrt{A/\pi}$. And we consider that the radio radius R_r equals double the coverage radius ($R_r = 2R_c$).

- We considered that at time $t = 0$ (when the WSN starts working after the deployment of sensors) the area is fully covered by the sensors used in networks to monitor this area.

3.4.2. Impact of scheduling mechanism in WSN on diagnostics

To study the effects of scheduling mechanism on diagnostics, we simulated a WSN composed of 200 sensors, sensing respectively the levels of temperature (70 sensors), pressure (70), and humidity (60 sensors). We considered that the WSN model is decentralized, for that we used 16 Cluster Heads to simulate the network on the area under consideration. For the learning stage, we take data consisting of N lines, each line is composed by T temperature data, P data of pressure, and H data of humidity to train these algorithms. All of these data are generated in the way mentioned in Section 3.4.1 (same type of data that will be captured by WSN during area monitoring).

The decision or diagnosis that is given by the algorithm is always related to the data learning of these algorithms. For example if the data learning is incomplete and it received a complete data to diagnose it, the error rate will be larger than if the data learning was complete and it received a complete data and vice versa. So, in brief, data learning as well as the data that will be used to diagnose the monitored area, determines the accuracy of the algorithm. For that, we used two types of data (complete and incomplete data) to train these algorithms, and to study for each data type, the variation of the error rate for each algorithm, with the variation of percentage of active sensors in WSN (scheduling).

The first study that was done in scheduling was to simulate a WSN in which all nodes are working at the same time: the percentage of active sensors is 100 %. Like each real WSN, the sensors on the long term will die (because of energy consumption) or break down (due to various causes as the operating age). This simulation was repeated 20 times on two different types of data learning for algorithms: complete data learning and incomplete data learning as shown in Figure 7 and Figure 8 respectively.

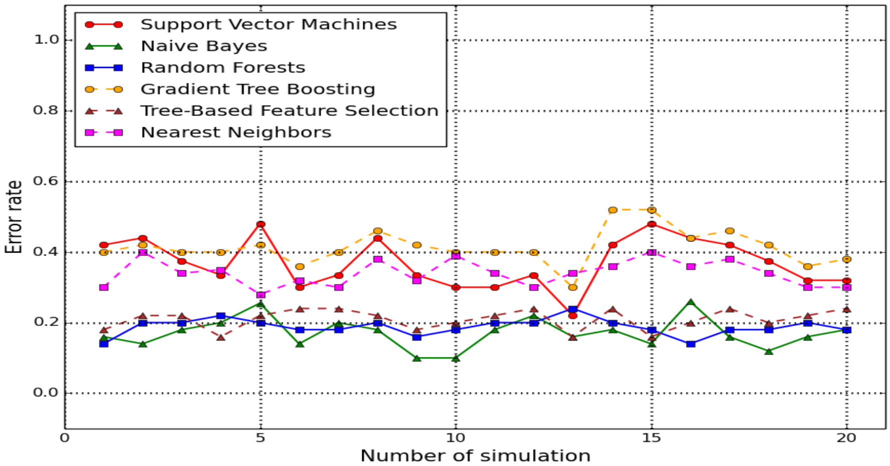


Figure 7: Error rate in diagnostics in the case of complete learning and the percentage of active sensors is 100 % with respect to the number of simulations.

Figure 7 indicates the variation of error rate for the six algorithms mentioned before, and during 20 simulations, in the particular case where the algorithms have complete data learning and all the sensors in WSN are active. As shown in the figure, during these 20 simulations, each algorithm presented a specific error interval (in %) as follows: $[22, 48]$ for SVM, $[10, 26]$ for NB, $[14, 24]$ for RF, $[30, 52]$ for GTB, $[16, 24]$ for TBFS, and $[28, 40]$ for NN.

The difference between Figure 7 and Figure 8 is that Figure 8 shows the results of algorithms training with incomplete data. During these 20 simulations, the error intervals (in %) of these algorithms changed to become as follows: $[9, 24]$, $[0, 14]$, $[6, 18]$, $[0, 8]$, $[2, 10]$, and $[8, 20]$ for SVM, NB, RF, GTB, TBFS, and NN respectively.

As depicted in the two figures, the error rate in Figure 7 (with complete data learning) is larger than the error rate shown in Figure 8 (with incomplete data learning). Depending on what we mentioned at the beginning of this section, the accuracy of the algorithm is related to data learning and the data captured by WSN to be treated. We can conclude that the data captured by the WSN in each simulation is incomplete, because simply the sensors on long term are dead or broken down, and this facts lead to not covered places in the area (coverage hole) and therefore to incomplete

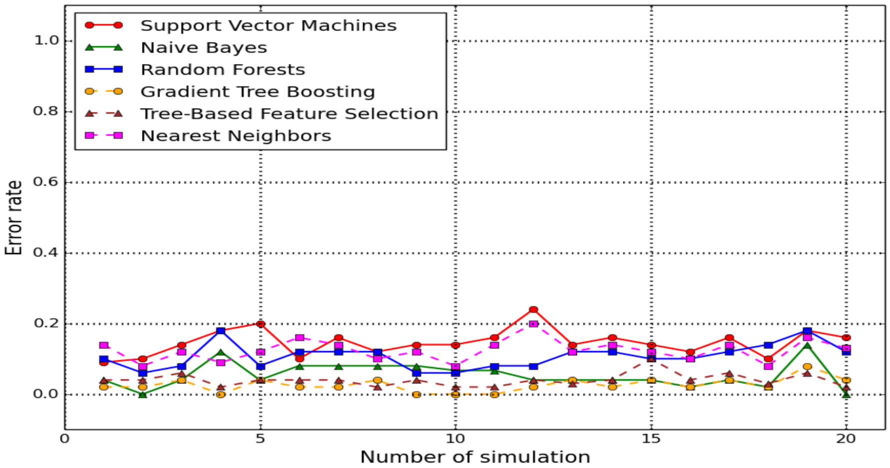


Figure 8: Error rate in diagnostics in the case of incomplete learning and the percentage of active sensors is 100 % with respect to the number of simulations.

data. We conclude from this study that the WSN on the long term will capture incomplete data and this will have an important impact on the diagnosis of the area state. Therefore the battery consumption of sensors in WSN is a very important strategy that has an impact on the work of networks over time, and taking it into consideration through our study is very important.

Scheduling mechanism strategy is one of the best solutions to preserve the energy of sensors for a longer time. Let us deepen the study of this strategy in WSN and its impact on diagnostics, in which the goal is to evaluate the variation of error rate with respect to the variation of the percentage of the active sensors, as shown in Figure 9 and Figure 10 with complete and incomplete data learning respectively. Each point in these figures is an average of error rates of a given algorithm on 20 simulations (for a certain percentage of active sensors in WSN). That is, for instance, the percentage of errors found in Figure 9 and Figure 10 in case where all sensors were working (100 %) are equal to the average of the percentage of errors found in Figure 7 and Figure 8 respectively.

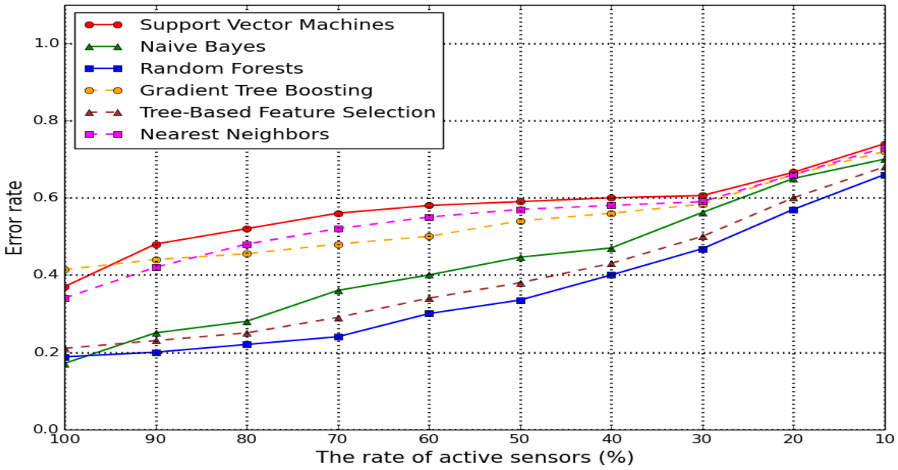


Figure 9: Error rate in diagnostics in case of complete learning with the variation of the percentage of active sensors.

Figure 9 contains the variation of error rate for six algorithms in case of complete learning with the variation of the percentage of the active sensors in WSN. As shown in the figure, as the percentage of active sensors in WSN decreases, the error rate of these algorithms increases, until the error rate of the algorithm (in case were only 10 % of sensors are working) reaches 74 % for SVM, 70 % for NB, 66 % for RF, 72 % for GTB, 68 % for TBFS, and 73 % for NN. Conversely, if the percentage of active sensors is 100 %, then the error rate is equal to 37 % for the SVM algorithm, 17 % for NB, 18.8 % for RF, 41.4 % for GTB, 21 % for TBFS, and 34 % for NN.

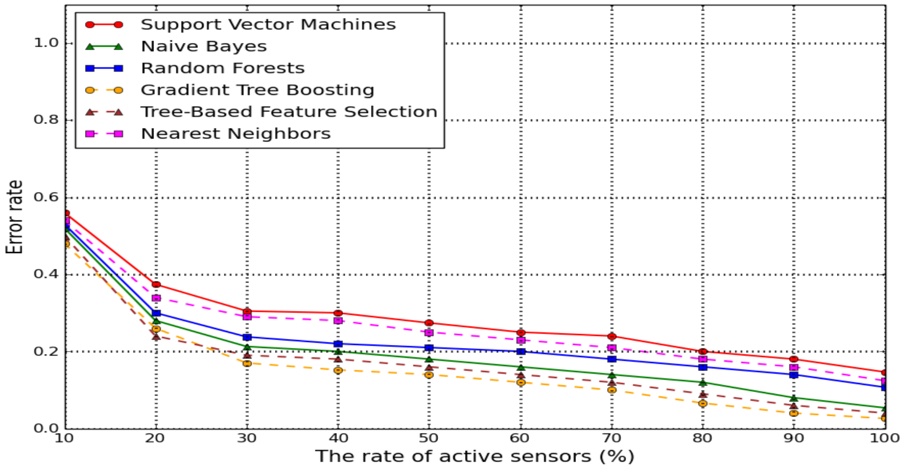


Figure 10: Error rate in diagnostics in case of incomplete learning with the variation of the percentage of active sensors.

The difference between Figure 9 and Figure 10 is that, in 10, we have studied the variation of error rate of the algorithm with respect to the variation of active sensors, in case where the training of these algorithms has been performed on incomplete data. As shown in Figure 10, which is similar to Figure 9, the error rates of the algorithms is decreasing as the percentage of active sensors in WSN is increasing. The percentage of errors (in case where only 10 % of the sensors are working) is equal to: 56 % (SVM algorithm), 52 % (NB), 53 % (RF), 48 % (GTB), 50 % (TBFS), and 54 % (NN). These percentages decrease to become 14.7 %, 5.4 %, 10.7 %, 2.6 %, 4 %, and 12.4 %, for SVM, NB, RF, GTB, TBFS, and NN algorithms respectively (when all sensors are active).

We deduce from these two figures that the error rate in Figure 9 is evolving in a way larger than the one in Figure 10, according to the modification of the number of active sensors. So, the error rate for each percentage in Figure 9 is larger than the one in Figure 10. Consequently the data captured by WSN is incomplete, despite the percentage of working sensors in WSN: incomplete data learning with incomplete captured data produces error rates lower than if the data learning was achieved with incomplete captured data, and *vice versa*. By relying on the change of curves in these two figures, we conclude that, as the percentage of active sensors in WSN increases, the coverage rate in the area monitoring increases accordingly, and so the data captured by WSN will finally be more accurate and precise for diagnostics. For that, the strategy of scheduling in WSN is very important and has a large impact on the efficiency of the network. Then this strategy must be taken into consideration in our study of networks, and we should rely on effective algorithms capable of using the sensors in an efficient manner, in order to increase the lifetime of the network and maximize the coverage over time.

3.4.3. Impact of coverage in WSN on diagnostics

In order to study the consequences of WSN strategies based on density and coverage, we now consider a network composed of 600 sensors, sensing respectively the levels of temperature (200 sensors), pressure (200 sensors), and humidity (200 sensors). And since WSN model is decentralized (like the study in Section 3.4.2), we used 48 Cluster Heads to simulate a WSN on the area under consideration.

In the first study, we simulated all the active sensors in the WSN at the same time with complete data learning of algorithms (for all the monitored area), in order to study the impact of the density of sensors on diagnostics. This study is like the one that was done in Section 3.4.2 (shown in Figure 7), but now we increase the number of sensors in WSN to be three times larger, therefore the density in WSN is improved. This simulation has been repeated 20 times just like the preceding studies, in order to evaluate the impact of the parameter we modified (number of sensors in WSN) on diagnostics, as shown in Figure 11.

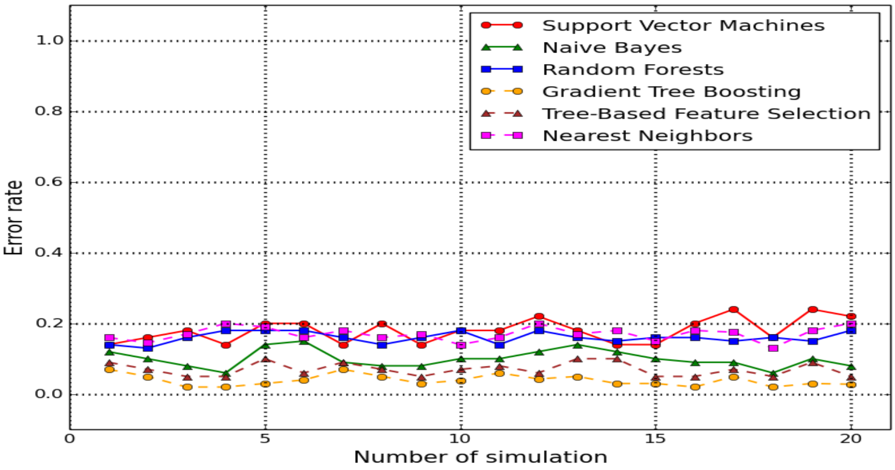
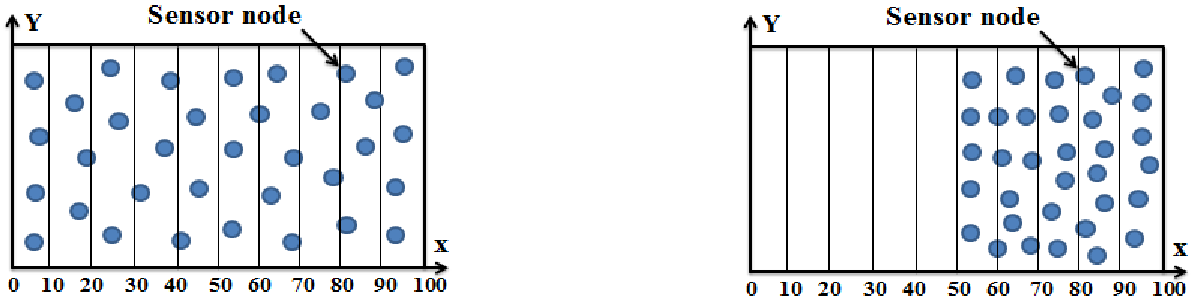


Figure 11: Error rate in diagnostics in the case of complete learning with respect to the number of simulations.

Figure 11 presents the error rate variations of our six algorithms during 20 simulations. As shown in this figure, during these 20 simulations, each algorithm shown a specific error interval as follows: [14, 24] for SVM, [6, 15] for NB, [13, 18] for RF, [2, 7] for GTB, [5, 10] for TBFS, and [13, 20] for NN. In comparison with Figure 7, the error rates of the six algorithms in Figure 11 have decreased compared to the rates obtained in Figure 7. From this comparison, we can verify the obvious fact that the density in WSN must be taken into consideration: density is a very important strategy, and has a great impact on coverage rate in an area, and therefore on the accuracy of the data with the time that will be captured by WSN, and used in diagnostics.



(a) Uniform distribution of sensors on an area (the asymmetrical density on abscissa is 0).

(b) Non-uniform distribution of sensors on an area (in case asymmetrical density on abscissa is 50).

Figure 12: asymmetrical density on abscissa.

We used the result obtained in Figure 11 as a starting point, in which we studied later the impact of diagnostics with the variation of the location of this density of sensors at the length of area monitoring (asymmetrical density on abscissas). We evaluated two different cases in this study. In the first one, we considered that the sensors are distributed uniformly and the area is fully covered, and we studied how error rates evolve with the modification of data learning algorithms at the length of monitored area (asymmetrical density on abscissas) as shown in Figure 13.

In the other (opposite) case, we considered complete data learning in which the data include all the information about the monitored area (we can call it uniform data learning), and we studied the variation of error rate of algorithms with the variation of the location of density of sensors in WSN at the length of area monitoring (asymmetrical density on abscissas) as shown in Figure 14. It is worth mentioning that Figure 12a is an example about uniform distribution (asymmetrical density on abscissa is 0) and Figure 12b is related to non-uniform distribution (in this case, the asymmetrical density on abscissa is 50). Each point in the Figures 13 and 14 is an average for the error rates of a certain algorithm on 20 simulations: the error rates of the algorithms in these two figures, in case where the asymmetrical density on the abscissa is 0, are equal to the average of the percentage of errors found in Figure 11.

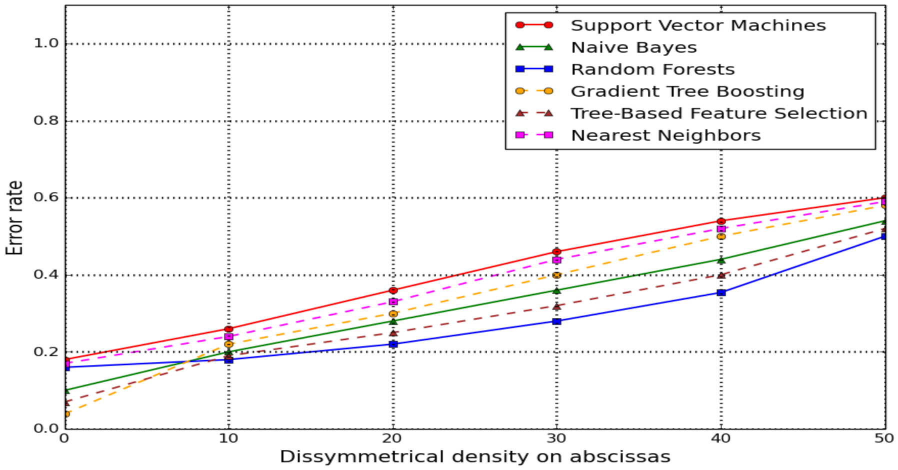


Figure 13: Error rate in diagnostics with the variation of data learning in a way of asymmetrical density on the abscissas (length of area monitoring).

Figure 13 shows the variation of error rate for the six considered algorithms, in case where the sensors are uniformly distributed in a fully covered area, with the variation of data learning of algorithms at the length of the monitored area (asymmetrical density on abscissas). It is worth mentioning that data learning is related to each region in the monitored area. For example if the data is complete, this means that the data learning includes information about all of the monitored area (asymmetrical density on abscissa is 0), as in Figure 12a. Conversely, Figure 12b illustrates the case where the learning process is realized when asymmetrical density on abscissa is 50, which means that the data learning includes information about only half of the area. As shown in Figure 13, as the asymmetrical density on abscissas increases (learning nodes are moved to the right of the area), the error rate of these algorithms increases until reaching (for an asymmetry of 50 %): 60 % (SVM), 54 % (NB), 50 % (RF), 58 % (GTB), 52 % (TBFS), and 59 % (NN). These values must be compared to the case where there is no asymmetrical density, in which the error rates are 18 % for SVM, 10 % for NB, 16 % for RF, 4 % for GTB, 7 % for TBFS, and 17 % for NN respectively.

Figure 14, for its part, is the result of the variation of error rate for six algorithms in case of complete data learning (the data includes information about all of the area monitoring), with the variation of the location of density of sensors in WSN at the length of area (asymmetrical density on abscissas). Figure 12a is when the uniform distribution of sensors in area monitoring (asymmetrical density on abscissa is 0): area is fully covered, while Figure 12b explains one case, if the distribution is non-uniform (in case asymmetrical density on abscissa is 50). As shown in Figure 14, as the asymmetrical density on abscissas decrease (sensors in WSN are more uniformly distributed), the error rate of these algorithms decrease until the error rate of the algorithm (only in the case if the asymmetrical density on abscissa is 0) equals 18 % for SVM, 10 % for NB, 16 % for RF, 4 % for GTB, 7 % for TBFS, and 17 % for NN. But if the asymmetrical density on the abscissa is 50, the error rate will be equal to 86 % for SVM algorithm, 82 % for NB, 74 % for RF, 84 % for GTB, 76 % for TBFS, and 85 % for NN.

Based on these simulations, we can conclude that the density level is an important strategy, and has a great impact on coverage, and therefore on diagnostics. Additionally, the location of nodes (uniformity in the density) is of importance in both the learning stage, and in the testing one (area monitoring after learning). These two elements

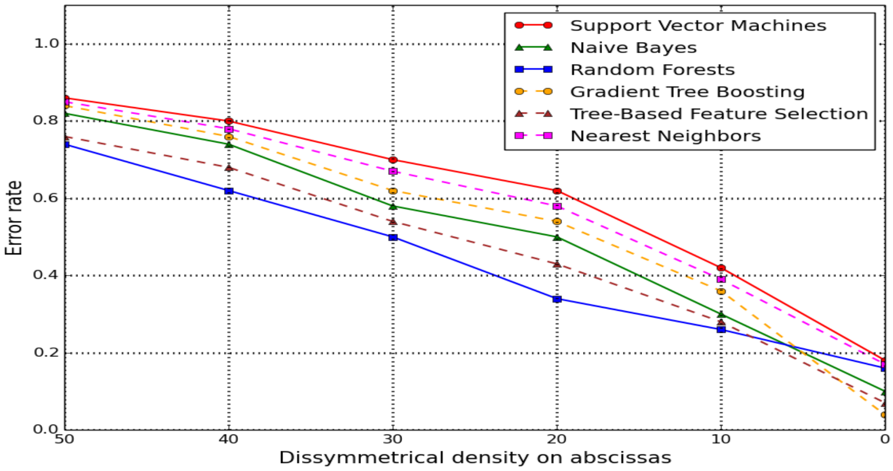


Figure 14: Error rate in diagnostics with the variation of area coverage by sensors in a way of asymmetrical density on the abscissas (length of area monitoring).

have a real impact on diagnostic and must be taken into consideration. Our experiments have illustrated that enlarging the uniformity of the node distribution improves the diagnosis quality of the monitored area, as a better coverage leads to a more accurate monitoring of the area. Since data learning is important for the accuracy of the algorithms in diagnosing, we studied the variation of data learning of the area that we need to monitor and we noticed that accuracy of diagnosis increases when the data learning contained all the information about the area. Finally, the best diagnostic happens when the data learning is complete and contains all the area monitoring. And also the distribution of sensors in an area must be uniform in a way that the area is covered to the maximum, as shown in Figure 13 and 14 in case where the asymmetrical density on the abscissa is 0, and this is what we have clarified before in Section 2.3.

3.4.4. Impact of topologies on diagnostics

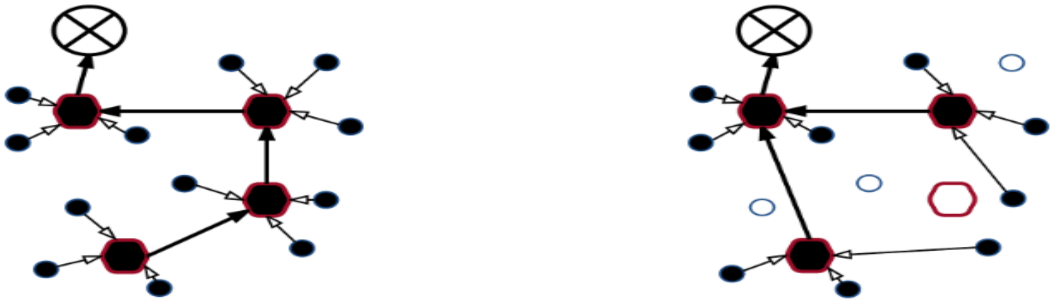
In order to illustrate the impact of topology strategies on the quality of data over time, and on the diagnostic of the state of the monitored area, we simulated four different topologies: decentralized topology, distributed topology, hierarchical topology, and centralized topology.

Decentralized topology:. In order to study the impact of decentralized topology on the diagnostics, we took WSN composed of 300 sensors, sensing respectively the levels of temperature (100 sensors), pressure (100), and humidity (100 sensors). We consider that the nodes are grouped into 30 clusters, each cluster being managed by a leader called cluster head (CH) or aggregator. The sensors capture the data from the area and send it to the CH, the latter aggregates the data and sends it to another CH or to the sink. In this study, we consider that the data aggregation at the CH happens as follows:

$$\sum_{i=0}^{S-1} D_i^c / S \tag{4}$$

where D is the data sent from the sensor to the aggregator, c is the type of sensor (temperature, pressure, or humidity), and S is the number of data that will be aggregated each time (for example, every 3 data from a certain type which are sent to CH from sensors, undergo aggregation).

The scenario of this topology is shown in Figure 15, the deployment of sensors is random, and the distribution and partition of CH on sensors follows K-means clustering method. Each sensor sends data to its CH. The latter, after aggregating these data, sends the aggregation to the closest CH on the condition that this CH is closer to the sink. If no CH meets this requirement, it will send it directly to the sink as shown in Figure 15a. After time $t = x$, the CH



(a) Sensors network at time $t = 0$.

(b) Sensors network at time $t = x$.

Figure 15: Scenario of decentralized topology.

and sensors may become inactive for several reasons, most importantly energy consumption or activity scheduling. If a CH became inactive, sensors in this cluster find other closest clusters to be in. In addition, CHs communicating with this inactive CH change their routes to the closest active CH to the sink. What is worth mentioning is that the black circles are the active sensors, the white circles are the inactive sensors, the black hexagons are the active CH, the white hexagons are the inactive CH, and finally the crossed circle is the sink.

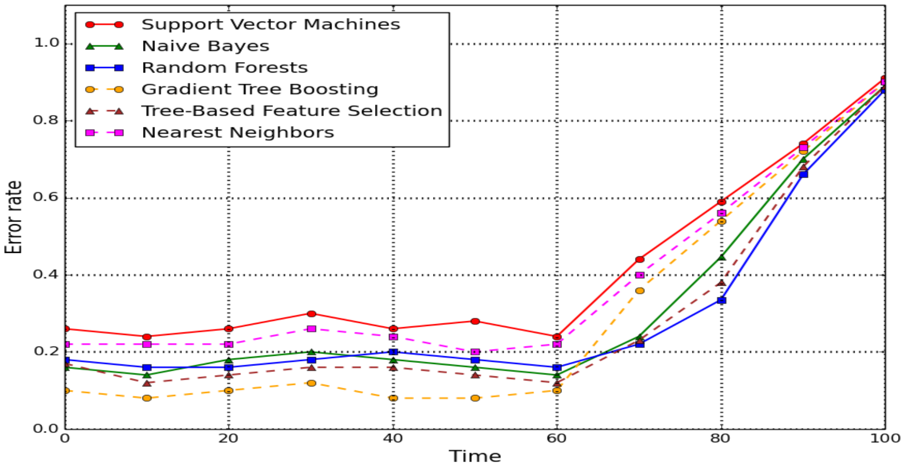


Figure 16: Error rate in diagnostics if the topology is decentralized with the variation of time.

As mentioned before, the topology may be dynamic, the sensors or CHs on the long term will die (because of energy consumption) or break down (due to various causes as the operating age). Figure 16 shows the variation of error rate for the six considered algorithms, in the case where the topology is decentralized, with the variation of time t (operating age). Each point in this figure is an average of error rates of a given algorithm on 20 simulations (for a certain t). As shown in the figure, during $t = 0..60$ (if $0 \leq t \leq 60$), each algorithm has a specific error interval (in %) as follows: $[24, 30]$ for SVM, $[14, 20]$ for NB, $[16, 20]$ for RF, $[8, 12]$ for GTB, $[12, 17]$ for TBFS, and $[20, 26]$ for NN. After that (if $t > 60$) the error rate for each algorithm increased significantly at these intervals to reach at $t = 70$, 44 % for SVM, 24 % for NB, 22 % for RF, 36 % for GTB, 23 % for TBFS, and 40 % for NN. This shows that at this time the sensors and CH in WSN are dying or breaking down, and this fact leads to the presence of uncovered places in the area (coverage hole) and therefore incomplete data for diagnostics. Then, when the WSN exceeds $t = 60$, the error rate of algorithms increases as time increases, to reach 91 % at $t = 100$ if the algorithm is SVM, 89 % for NB, 88 % for RF, 90 % for GTB, 89 % for TBFS, and 90 % for NN (approximately the whole network is inactive). Note that the error rate in this simulation (decentralized topology) is related to the method of data aggregation.

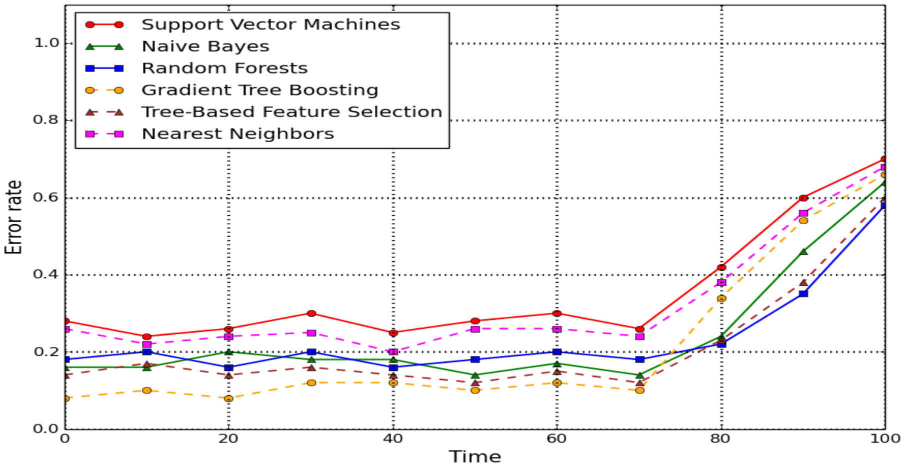


Figure 17: Decentralized topology with 2 × CH.

Figure 17 shows the result of WSN with decentralized topology under the same conditions and parameters that we explained in previous result (Figure 16), but here we used 60 CHs instead of 30 CHs. With these parameters, the error rate of the six algorithms varied in a significant way from what is shown in Figure 16 (with 30 CHs). More precisely, the error rate of the algorithms remained constant (along the intervals determined before) for $t = 0..70$, while in our previous simulation, it remained constant until $t = 60$. Then, the error rates start to increase over time, as numerous sensors in the network emptied their batteries, to reach the following rates: 70 % at $t = 100$ if the algorithm is SVM, 64 % for NB, 58 % for RF, 66 % for GTB, 60 % for TBFS, and finally 68 % for NN. We can remark that, at this time and with 30 CHs, the whole network became inactive in our previous simulation (see Fig. 16). We can deduce from these simulations that the difference between obtained results is due to two causes. The first reason, which is the most important one, is related to the increase of cluster heads (and thus of clusters) in the network. Indeed, the works that a CH must realize (receiving data from sensors, aggregation, and transmission) is now divided on more CHs, leading to a decrease of the energy consumption for each cluster head. The preservation of energy in each CH consequently leads to the increase of the network's lifetime, and to better diagnostic results in terms of error rates (see Fig. 16). The other cause is that the increase in number of CHs in the network decreases the distance between the sensors and their associated CH. Hence the consumption of energy of sensors decreased accordingly, which leads to a lifetime increase.

To determine which of these two causes has a greater impact on topologies and diagnostics with time, the number of CH in the network was taken as 30 (as the study shown in Figure 16), but now we increased the density of sensors in order to decrease the distance between the sensors and the CH and between the sensors themselves. In this study, we used 400 sensor nodes (100 more sensors) sensing respectively the levels of temperature (140 sensors), pressure (130 sensors), and humidity (130 sensors) and the result is shown in Figure 18. The error rate for the six algorithms in this figure varied significantly from what was shown in Figure 16 (with 300 sensors), but is opposite to what is shown in Figure 17. The error rate of diagnostics with 400 sensors remained constant (along the intervals determined before) during $t = 0..50$, while with 300 sensors, the error rate remained constant during $t = 0..60$. After that, the error rate of the algorithms started to increase with time and this shows that the sensors in WSN started dying, to reach 91 % at $t = 90$ if the algorithm is SVM, 90 % if NB, 89 % if RF, 91 % if GTB, 89 % if TBFS, and 90 % if NN (the whole network is inactive). Recall that with 300 sensors, the whole network become inactive at $t = 100$ as shown in Figure 16.

Even though the distance between the sensors and CH decreased after increasing the density of the sensors, the lifetime of WSN decreased. This is because of a very important reason which should be taken into consideration, that is, the number of sensors in each cluster in WSN (with decentralized topology) that are locally managed by CH. Normally, as the number of sensors in the clusters increases, the work of the CHs also increases, and therefore

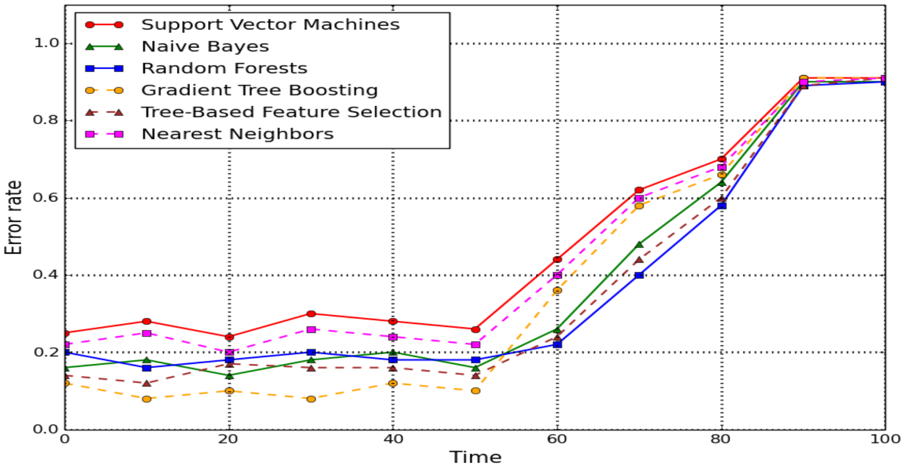
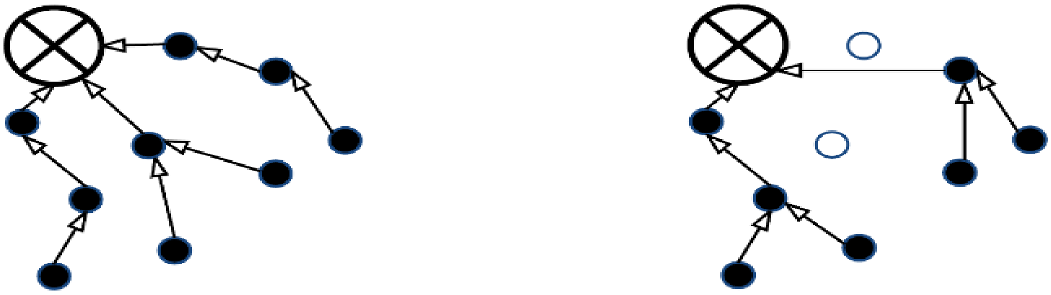


Figure 18: Decentralized topology with +100 sensors.

consumption of energy will also increase. Figure 18 explained this conclusion when we increased the number of sensors in the network, and we noticed the variation of the lifetime from previous studies. Finally, we can conclude the importance of dividing WSN to the largest possible number of clusters in order to divide the work on a larger number of CH, and therefore reduce energy consumption. And to get an accurate data from the area monitoring, the area should be fully covered for the longest achievable period, and here the density of sensors in WSN plays an important role. Consequently, we should offer a reasonable trade off between the number of sensors in WSN and between the largest possible number of clusters in this network, to get the longest lifetime and therefore an accurate data for diagnostics for a longer period.

Distributed topology:. In order to study this topology and its impact on diagnostics, we consider a WSN composed of 300 sensors, sensing respectively the levels of temperature (100 sensors), pressure (100 sensors), and humidity (100 sensors).

The scenario of distributed topology is shown in Figure 19, where all sensor nodes in the network have the same role and importance; *i.e.*, there is no aggregation role, no clusters, and no CHs. Data packets are forwarded in a hop-by-hop manner. Each sensor is able to discover its neighbors within a radio radius of $2R_c$ (R_c is the coverage radius). We assume that every node can access information about its neighbors, including their locations. The nodes choose neighbors to communicate with, and the latter should be closest to the sink within the sender's radio range. If the sensor is closest to the sink, the sensor will then send it directly to the sink.



(a) Sensors network at time $t = 0$.

(b) Sensors network at time $t = x$.

Figure 19: Scenario of distributed topology.

As explained in the scenario before, after certain time $t = x$, the sensors may become inactive and the routes always change in function of the closest neighbors to the sink as shown in Figure 19b. The black, white, and crossed circles represent the active sensors, inactive sensors, and the sink respectively.

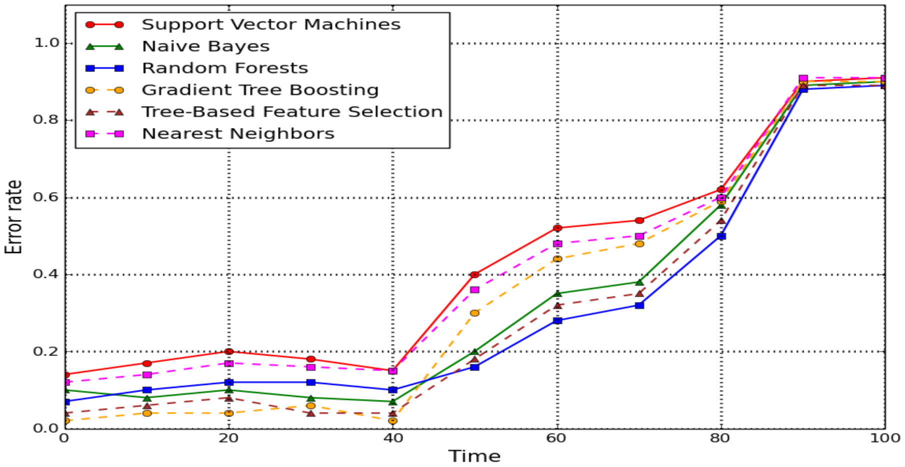


Figure 20: Error rate in diagnostics if the topology is distributed with the variation of the time.

Figure 20 presents the variation of error rate for the six considered algorithms in the case of distributed topology, with the variation of time t (operating age). Each point in this figure is an average of error rates of a given algorithm on 20 simulations (for a certain t). As shown in the figure, if $0 \leq t \leq 40$, each algorithm has a specific error interval (in %) as follows: $[14, 20]$ for SVM, $[7, 10]$ for NB, $[7, 12]$ for RF, $[2, 6]$ for GTB, $[4, 8]$ for TBFS, and $[12, 17]$ for NN. After that (if $t > 40$) the error rate for each algorithm increased significantly at these intervals to reach, at $t = 50$, 40 % for SVM, 20 % for NB, 16 % for RF, 30 % for GTB, 18 % for TBFS, and 36 % for NN. This shows that at this time the sensors in WSN are dying or breaking down, and this fact leads to the presence of uncovered places in the area (coverage hole) and therefore incomplete data for diagnostics. Then, when the WSN exceeds $t = 40$, the error rate of algorithms increases with time to reach 90 % at $t = 90$ if the algorithm is SVM, 89 % if NB, 88 % if RF, 90 % if GTB, 89 % if TBFS, and 91 % if NN (approximately the whole network is inactive).

At each data transfer, the energy of a sender is reduced regarding its distance from the recipient. So packet transfer distance is one of the most important issues in topologies, and play an important role in the variation of lifetime of network and therefore on diagnostics with time. In order to study the importance and impact of transfer distance strategy in network, we simulated a WSN with distributed topology but now in a scenario different from what was used in previous study (shown in Figure 20). In this simulation, each sensor in WSN chooses the closest neighbor to communicate with. At all times, if the sensor is closest to the sink, it will directly communicate with the sink. After certain time $t = x$, the sensors may become inactive and the routes always change in function of the closest neighbors, the result of this study is shown in Figure 21. The error rate for the six algorithms in this figure varied significantly from what is shown in Figure 20. When we reduced the transfer distance in the topology, the error rate of diagnostic remained constant (along the same intervals mentioned in previous study) during $t = 0..60$, while in a previous study, the error rate remained constant during $t = 0..40$. After that the error rate of algorithms started to increase over time, to reach 62 % at $t = 100$ if the algorithm is SVM, 58 % for NB, 50 % for RF, 59 % for GTB, 54 % for TBFS, and 60 % for NN. What is worth mentioning is that in previous study, the whole network became inactive at $t = 90$, as shown in Figure 20.

Thus the lifetime of WSN increased significantly when we decreased the transfer distance between sensors, and the network remains active for a longer time, and therefore the sink continues to receive information from the monitored area for a longer time. By relying on this study and on the comparison between these results (Figure 20 and 21), we can conclude to the importance of transfer distance strategy in WSN on lifetime, and thus on diagnostics with time.

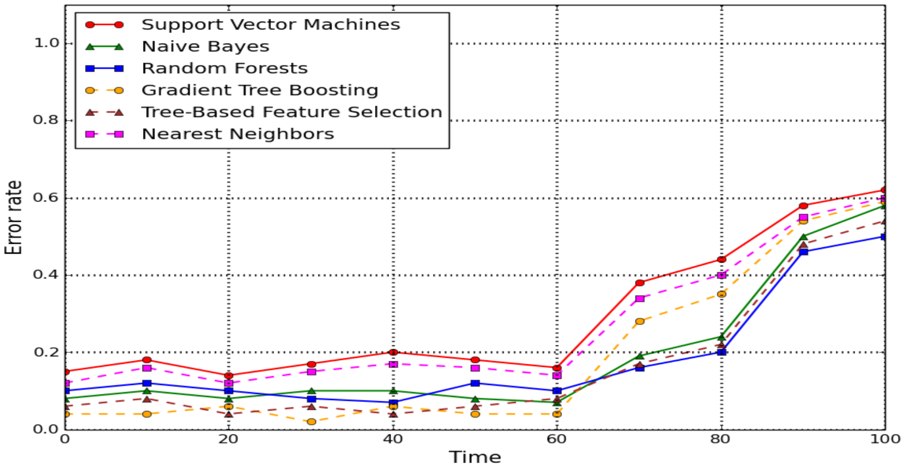
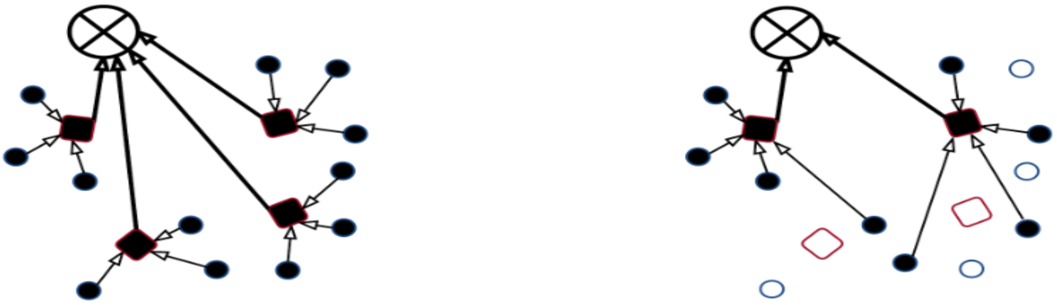


Figure 21: Distributed topology with closest transfer distance between sensors.

Hierarchical topology:. As what has been mentioned in Section 2.1.2, sensor nodes can be organized in several levels, making a hierarchical topology. The sensor nodes are organized in a tree hierarchy from the sink (being the root of a tree), until sensor nodes having no descendants (leaf nodes). In order to study the impact of hierarchical topology on diagnosis, and to compare this topology to other ones, we took WSN composed of 300 sensors (leaf nodes), sensing respectively the levels of temperature (100 sensors), pressure (100 sensors), and humidity (100 sensors). These sensors are considered as the access layer in this topology (third layer). We considered 30 nodes playing the role of the second layer in topology (the distribution layer), which implements policies and forward messages. These nodes are in charge to build the links between the leaf nodes towards the sink (core layer). Each of these 30 sensors has the same battery supply of leaf nodes; conversely, in a decentralized topology, CHs have received an extra supply, therefore the batteries last longer and we finally obtain more data for diagnostics.



(a) Sensors network at time t = 0.

(b) Sensors network at time t = x.

Figure 22: Scenario of hierarchical topology.

The scenario of this topology is shown in Figure 22. After a certain time $t = x$, the sensors in access layer or distribution layer may become inactive for several reasons, most importantly energy consumption. Unfortunately if a parent node (in distribution layer) become inactive, its children can no longer communicate with other nodes in the network. In this case, in order to keep connectivity, each sensor will then communicate with the closest active node in the distribution layer as shown in Figure 22b.

Figure 23 indicates the variation of error rate for the six considered algorithms under the same conditions of the previous studies, but here with a hierarchical topology. Again, each point in this figure is an average of error rates

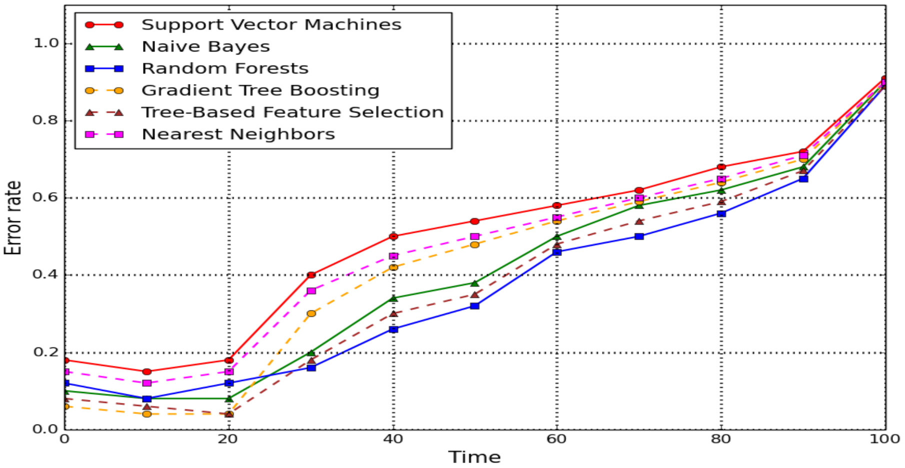


Figure 23: Error rate in diagnostics if the topology is hierarchical with the variation of time.

of a given algorithm on 20 simulations (for a certain t). As shown in the figure, if $0 \leq t \leq 20$, each algorithm has a specific error interval (in %) as follows: [15, 18] for SVM, [8, 10] for NB, [8, 12] for RF, [4, 6] for GTB, [4, 8] for TBFS, and [12, 15] for NN. These intervals are approximately the same where the topology is distributed (where the whole network is active), and this is because in these two topologies, there is no data aggregation. Then, for $t > 20$, the error rate for each algorithm increased significantly at these intervals to reach, at $t = 30$, 40 % for SVM, 20 % for NB, 16 % for RF, 30 % for GTB, 18 % for TBFS, and 36 % for NN. Such results show that at this time the sensors in WSN (in access or distribution layer) are dying or breaking down, and this fact leads to the presence of uncovered places in the area (coverage hole) and therefore incomplete data for diagnostics. Then, when the WSN exceeds $t = 20$, the error rate of algorithms increases with time, to reach 91 % at $t = 100$ if the algorithm is SVM, 90 % if NB, 89 % if RF, 90 % if GTB, 89 % if TBFS, and 90 % if NN. In other words, approximately the whole network is inactive.

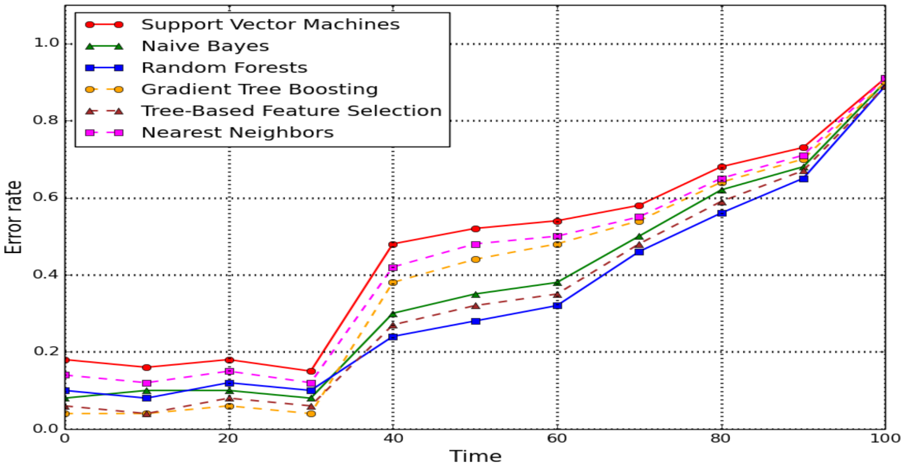


Figure 24: Hierarchical topology with 2× sensors in distribution layer.

Figure 24 shows the result of WSN with hierarchical topology under the same conditions and parameters that we explained in previous result (shown in Figure 23), but here we used 60 nodes in distribution layer instead of 30 nodes.

With these parameters, the error rate of the six algorithms varied significantly from what is shown in Figure 23 (with 30 nodes in distribution layer). In addition, the algorithms remained constant (along the intervals mentioned before) during $t = 0..30$, while in a previous study, the error rate remained constant during $t = 0..20$. After that, the error rate of the algorithms started to increase over time, and this shows that sensors in WSN started dying, to reach 91 % at $t = 100$ if the algorithm is SVM, 90 % for NB, 89 % for RF, 90 % for GTB, 89 % for TBFS, and 91 % for NN. Thus, when we used 30 more nodes in distribution layer, the lifetime of the network increased $10 \times t$, and this is what we obtained in Figure 23 and 24. What is worth to mention is that the batteries of the nodes in the distribution layer are like the ones in access layer, therefore the lifetime of WSN will increase more if the batteries in distribution layer were bigger. This difference is due to two basic reasons: first, the increase in number of nodes in distribution layer will divide the work, from receiving data from access layer and transmission of this data to the sink on more nodes, therefore the energy consumption of each node in this layer will decrease which will increase the lifetime. The other reason is the increase in number of nodes in distribution layer reduced the distance between these nodes and the nodes in access layer.

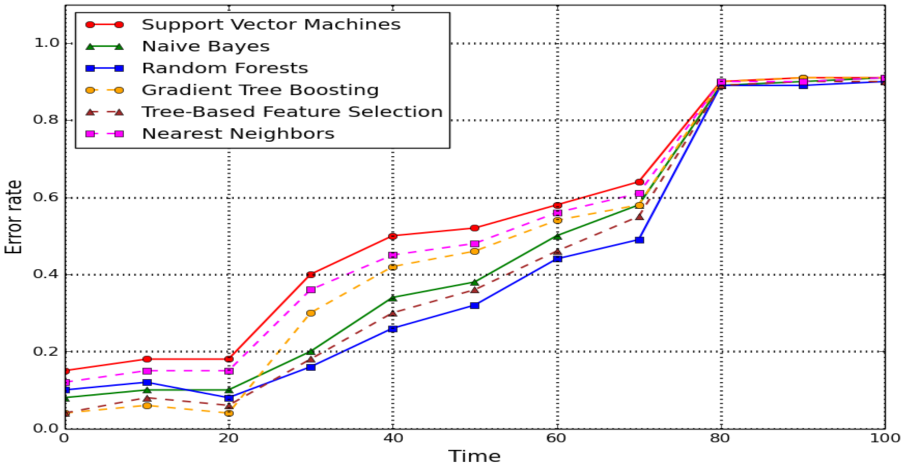
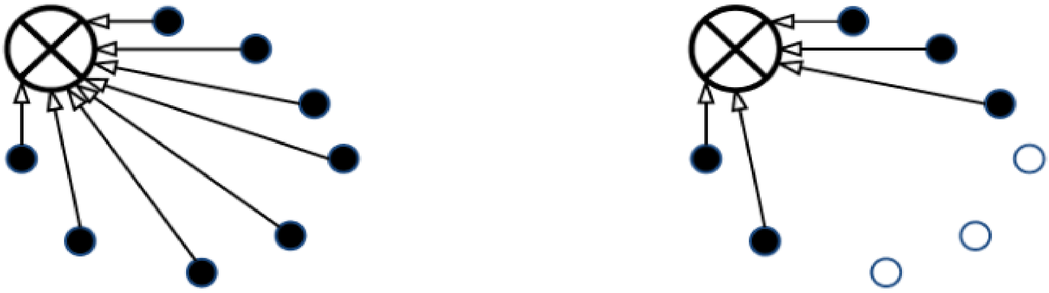


Figure 25: Hierarchical topology with +100 sensors in access layer.

To determine which of these two causes has the greatest impact on topologies and diagnostics over time, the number of sensors in distribution layer in the network was fixed to 30 (as the study shown in Figure 23). We increased the density of sensors in access layer (leaf nodes) in order to decrease the distance between the sensors in this layer and the sensors in distribution layer, and between the sensors themselves. In this study, we used 400 leaf nodes (100 more sensors) sensing respectively the levels of temperature (140 sensors), pressure (130 sensors) and humidity (130 sensors) and the result is shown in Figure 25. The error rate for the six algorithms in this figure varied differently from what was shown in Figure 23 (with 300 leaf nodes), and it is opposite to what is shown in Figure 24. These error rate of diagnostics remained constant (along the intervals determined before) during $t = 0..20$, after that the error rate of the algorithms started increasing with time, to reach 90 % at $t = 80$ if the algorithm is SVM, 89 % if NB, 89 % if RF, 90 % if GTB, 89 % if TBFS, and 90 % if NN (the whole network is inactive). With 300 leaf nodes, the whole network become inactive at $t = 100$ as shown in Figure 23.

The lifetime of WSN decreased even though the distance between the nodes in the second layer (distribution layer) and the nodes in the third layer (access layer) decreased after increasing the density. This is for the same reason that we studied in decentralized topology, which is the number of leaf nodes in access layer, by which each group of them is locally managed by parent node (in distribution layer). Normally, as the number of leaf nodes increases, the work of parent nodes also increases, and therefore consumption of energy will also increase. Figure 25 explained this conclusion when we increased the number of nodes in the network, and we noticed the variation of the lifetime from previous studies. Finally, we can conclude the importance of dividing WSN to the largest possible number of clusters, in order to divide the work of the network on largest possible number of parent nodes, so the energy consumption will be divided and reduced. Conversely, in order to get accurate data, the monitored area should be fully covered and for the longest possible time and here lies the importance of sensor density in WSN. So with hierarchical topology, we should offer a reasonable trade off between the number of sensors in access layer and between the largest possible number of parent node in this network, to get the longest lifetime and hence accurate data for diagnostics for a longer period.

Centralized topology:. In order to study the impact of this topology on diagnosis, we consider a WSN composed of 300 sensors, sensing respectively the levels of temperature (100 sensors), pressure (100 sensors), and humidity (100 sensors). In centralized topology, all the sensor nodes have the simple task of sensing new information and forwarding it to a central node, where all the data processing is done as shown in Figure 26a. In this topology, we can notice that, after $t = x$, the nodes that exhaust their energy first are the farthest from the sink. This is due to the long distance of packet transfer as shown in Figure 26b. The black and white circles are the active and inactive sensors respectively, and the crossed circle is the sink.



(a) Sensors network at time $t = 0$.

(b) Sensors network at time $t = x$.

Figure 26: Scenario of centralized topology.

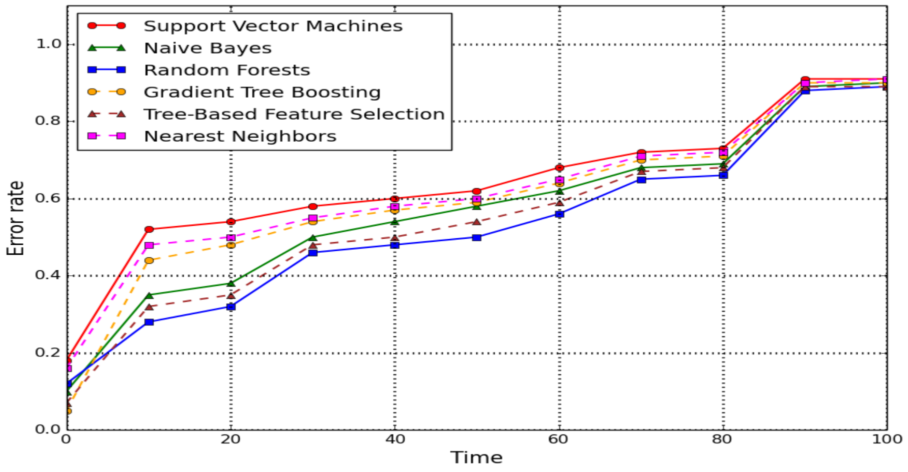


Figure 27: Error rate in diagnostics if the topology is centralized with the variation of time.

Figure 27 shows the variation of error rate for the six considered algorithms in the case of a centralized topology. Each point in this figure is an average of error rates of a given algorithm on 20 simulations. As we can see, at $t = 0$ (when the WSN starts working), each algorithm has a specific error rate (in %) as follows: 18 % for SVM, 10 % for NB, 12 % for RF, 5 % for GTB, 7 % for TBFS, and 16 % for NN. During the work of the network, with time, the sensors that are located farthest from the sink start dying first because they consume more energy than the others, and this is due to the long distance of packet transfer. For that, at $t = 10$ the error rate increased in a noticeable way to become 52 % with SVM, 35 % with NB, 28 % with RF, 44 % with GTB, 32 % with TBFS, and finally 48 % with NN. This is a proof that the data became incomplete for diagnostic, as the regions far from the sink are no longer covered by sensors. Then, when the WSN exceeds $t = 10$, the error rate of algorithms increases with time to reach 91 % at $t = 90$ if the algorithm is SVM, 89 % for NB, 88 % for RF, 90 % for GTB, 89 % for TBFS, and 90 % for NN.

Discussion:. Now we want to explain and compare the results obtained in our study for these four topologies, but with the same number of sensors used in WSN (300 sensors): the results of these studies are shown in Figures 16, 20, 23, and 27, for decentralized, distributed, hierarchical, and centralized topology respectively. The aim of this comparison is to focus on several elements or issues related to topologies that have a great impact on diagnostics.

We remark, in Figure 20 (with distributed topology), the noticeable variation of error rate of the algorithms with time from what is shown in Figure 16 (with decentralized topology). We can note that in Figure 20 the sensors, after $t = 40$, started dying or breaking down, and the whole network became inactive at $t = 90$, while in Figure 16, the sensors or CH started dying or breaking down after $t = 60$, and the whole network became inactive at $t = 100$. Moreover we can notice that in Figure 16, during $t = 0..60$ (the whole network is active) the error rate is evolving in a way larger than the one in Figure 20 during $t = 0..40$ (idem). From this study and based on this comparison we can conclude that the lifetime of the networks with decentralized topology is greater than if it were a distributed one, which is due to the fact that the data aggregation reduces the number of packet transfer, and therefore it further reduces the overall energy consumption in the network. But the error rate of diagnosis is greatly related to the data aggregation method (if the topology is decentralized) because data aggregation always reduces the data accuracy that will be used for diagnosis, and this is shown and clarified in these two figures where the whole network is active.

We had a different scene in Figure 23 (with hierarchical topology) because the variation of error rate of the algorithms varied with time in a significant way from what is shown in Figure 16 and 20 (with decentralized and distributed topology). We notice that, in Figure 23, the sensors after $t = 20$ started dying or breaking down, while in previous studies, the sensors became inactive after this time. Based on this study, we can conclude that the lifetime of the network with hierarchical topology is smaller than if it were with a decentralized or distributed topology (the network lifetime defined as time until the first node dies). Furthermore, we note that the whole network with hierarchical topology became inactive at $t = 100$ when considering the decentralized topology, while with distributed one the whole network became inactive at $t = 90$. Based on these results, we can conclude to the importance, when deploying a wireless sensor network for diagnostics purposes, of dividing the WSN in area monitoring into regions which are locally managed by a central node (or parent node).

If we suppose that the network lifetime can alternatively be defined as the time until the first node dies, then by relying on the curve modifications, we can conclude that the lifetime of the network with centralized topology is smaller than if it were a decentralized, distributed, or hierarchical one. Moreover, the whole network with hierarchical and decentralized topology became inactive at $t = 100$, while with distributed and centralized topology the whole network became inactive at $t = 90$. Based on these four results, we confirm what we mentioned before about the importance of dividing the WSN in area monitored into regions, which are locally managed by a parent node (the network remains active for a longer time, therefore the sink continues to receive information from the monitoring area for a longer time). Based on this work, we were able to notice the importance and impact of each type of topology in WSN on diagnostics with the increase of operating age of WSN, and focus on several issues related to these types of topologies.

3.4.5. Algorithm complexities

The last aspect to investigate in the choice of machine learning algorithms in a PHM based on data provided by a WSN is the algorithm complexities. This complexity varies according to the methods used for data classification, to learn the machine, and to perform diagnosis or to take decision from new data captured by the network. Indeed, efficiency of the machine learning algorithms depends both on its accuracy in diagnostic and on its complexity. By relying on our simulation, we studied the complexities for all the algorithms mentioned before, and the results are shown in Table 2.

Table 2 shows the calculation time of each diagnostic algorithm, in seconds and for three kinds of iteration numbers ($t = 1, 10$, and 100 iterations). This is the time taken by a given simulated method to perform the complete diagnostic

Diagnostic algorithm	**Real-time performances (s)**		
	t = 1	***t = 10***	***t = 100***
Support Vector Machine	0.27	2.5	23
Naive Bayes	0.21	1.9	18
Gradient Tree Boosting	0.3	2.8	26
Tree-Based Feature Selection	0.25	2.3	21
Nearest Neighbors	0.18	1.7	15.5
Random forests	0.22	2	19

Table 2: The calculation time of diagnostic algorithms (second) with respect to the operating age (t).

process. For instance, a Support Vector Machine used for PHM on WSN takes $0.27s$. to perform one iteration over the network, while $23s$. are needed for $t = 100$ iterations. By relying on the results shown in Table 2, we can notice that the Nearest Neighbors has a lower complexity than the other algorithms. It is followed by Naive Bayes, Random forests, Tree-Based Feature Selection, Support Vector Machine, while the most complex tool is the Gradient Tree Boosting.

What is worth mentioning is that diagnosis accuracy and complexity are two objectives that cannot be reach together. For instance, Nearest Neighbors outperforms the other methods in terms of complexity, but it has in general a worse error rate than the other algorithms in the various situations investigated previously. Support Vector Machine, for its part, appears as a bad choice when dealing with these two objectives. Conversely, Random Forests have a reasonable complexity, while exhibiting a low error rate during our simulation. This method appears as a good compromise for prognostic and health management based on data provided by a wireless sensor network.

4. Conclusion and Future Work

The WSNs provide PHM with a new way of distributed data collection and wireless transmission for diagnosing the state of an area and be informed if it is in failure or not. WSNs strategies are important factors for achieving QoS in WSNs application. In this paper, we explained the relation between WSN strategies and their impact on area diagnostic, and therefore on PHM. We mentioned and studied several important strategies in WSN that have an important influence on QoS of WSN, and we proved from this study that they also have an important impact on diagnostics.

In this work, we studied the variation of the accuracy of diagnosing for six algorithms, and by relying on this variation we determined and focused on several issues related to these strategies. We can conclude that the data captured by the WSN on long term is incomplete because simply the sensors become inactive for several reasons, most importantly energy consumption. Therefore the battery consumption of sensors in WSN is a very important strategy that has an impact on the work of networks with time, and taking it into consideration through our study is very important. Scheduling mechanism is one of the best solutions to preserve the energy of sensors for a longer time. For that, we studied this strategy and we proved that as the percentage of active sensors in WSN is decreasing (because of scheduling of sensors), the accuracy of area diagnosis will decrease due to the reduction of the coverage rate of WSN, so the accuracy of the captured data that will be used to diagnose will decrease. In a second step we studied the impact of the density of sensors and we shown that if the density of sensors increases, the accuracy of the area state will increase, and this shows that the coverage rate will increase. Then, the data accuracy will increase with the increase of number of sensors in monitored area. Additionally, the location of nodes (uniformity in the density) is of importance in both the learning stage, and in the testing one (area monitoring after learning). These two elements have a real impact on diagnostic and must be taken into consideration.

Moreover, we studied the topology effects in WSN through four different topologies, each one of them belonging to a certain type as follows: distributed, hierarchical, centralized, and decentralized topology. From this work and by relying on these topologies, we were able to prove that topologies have a great impact on the accuracy of the data and therefore on PHM, and this impact varies according to the type and parameters of topologies in WSN. As a future work, we will study effective and reliable algorithm that relies on scheduling mechanism to increase the network lifetime, and also to increase the coverage rate to the maximum with the use of certain amount of active sensors in network (density). This algorithm must be compatible with the different types of topologies in WSN. By relying on this algorithm, the data diagnostics will be accurate for the longest possible time. Finally, we have to perform this study on a real WSN to monitor a real area and to really observe the impact of topologies on PHM from one side, and to compare with the results that we obtained by our numerical simulations side.

This paper is partially funded by the Labex ACTION program (ANR-11-LABX-01-01 contract) and by the Interreg RESponSE project.

References

[1] http://scikit-learn.org.

[2] Localization and coverage for high density sensor networks. *Computer Communications*, 31(4):770 – 781, 2008.

[3] Sachin Adlakha and Mani Srivastava. Critical density thresholds for coverage in wireless sensor networks. In *Wireless Communications and Networking, 2003. WCNC 2003. 2003 IEEE*, volume 3, pages 1615–1620. IEEE, 2003.

[4] Muhammad Hammad Ahmed, Syed Wasi Alam, Nauman Qureshi, and Irum Baig. Security for wsn based on elliptic curve cryptography. In *Computer Networks and Information Technology (ICCNIT), 2011 International Conference on*, pages 75–79. IEEE, 2011.

[5] Özgür B Akan and Ian F Akyildiz. Event-to-sink reliable transport in wireless sensor networks. *IEEE/ACM Transactions on Networking (TON)*, 13(5):1003–1016, 2005.

[6] Ian F Akyildiz, Tommaso Melodia, and Kaushik R Chowdhury. A survey on wireless multimedia sensor networks. *Computer networks*, 51(4):921–960, 2007.

[7] Ian F Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci. A survey on sensor networks. *IEEE Communications magazine*, 40(8):102–114, 2002.

[8] Habib M Ammari. A unified framework for k-coverage and data collection in heterogeneous wireless sensor networks. *Journal of Parallel and Distributed Computing*, 89:37–49, 2016.

[9] Sanghoon Bae, Hanju Cha, and Youngsuk Suh. Study on condition based maintenance using on-line monitoring and prognostics suitable to a research reactor. In *European conference of the prognostics and health management society*, 2014.

[10] Jacques Bahi, Wiem Elghazel, Christophe Guyeux, Mohammed Haddad, Mourad Hakem, Kamal Medjaher, and Noureddine Zerhouni. Resiliency in distributed sensor networks for prognostics and health management of the monitoring targets. *The Computer Journal*, 59(2):275–284, 2016.

[11] Jacques M Bahi, Christophe Guyeux, Abdallah Makhoul, and Congduc Pham. Low-cost monitoring and intruders detection using wireless video sensor networks. *International Journal of Distributed Sensor Networks*, 2012, 2012.

[12] Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[13] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[14] Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, pages 477–505, 2007.

[15] Mihaela Cardei and Jie Wu. Coverage in wireless sensor networks. *Handbook of Sensor Networks*, 21, 2004.

[16] David W Carman, Peter S Kruus, and Brian J Matt. Constraints and approaches for distributed sensor network security (final). *DARPA Project report,(Cryptographic Technologies Group, Trusted Information System, NAI Labs)*, 1(1), 2000.

[17] Ruay-Shiung Chang and Shuo-Hung Wang. Deployment strategies for wireless sensor networks. In *Handbook of Research on Developments and Trends in Wireless Sensor Networks: From Principle to Practice*, pages 20–37. IGI Global, 2010.

[18] Jungeun Choi, Joosun Hahn, and Rhan Ha. Short paper_. *Journal of Information Science and Engineering*, 25:273–287, 2009.

[19] Wiem Elghazel, Jacques Bahi, Ahmad Farhat, Christophe Guyeux, Mourad Hakem, Kamal Medjaher, and Noureddine Zerhouni. Random forests for industrial device functioning diagnostics using wireless sensor networks. In *IEEE AEROSPACE CONFERENCE, 2015.*, pages 1–9, Big Sky, Montana, USA, mar 2015.

[20] Wiem Elghazel, Jacques Bahi, Christophe Guyeux, Mourad Hakem, Kamal Medjaher, and Noureddine Zerhouni. Dependability of wireless sensor networks for industrial prognostics and health management. *Computers in Industry*, 68:1–15, 2015.

[21] Wiem Elghazel, Jacques Bahi, Christophe Guyeux, Mourad Hakem, Kamal Medjaher, and Noureddine Zerhouni. Prognostics and health management based on dependable wireless sensor networks. In *SENSORNETS 2015, 4th Int. Conf. on Sensor Networks*, pages ***–***, Angers, France, feb 2015.

[22] Wiem Elghazel, Kamal Medjaher, Christophe Guyeux, Mourad Hakem, Noureddine Zerhouni, and Jacques Bahi. Dependable wireless sensor networks for prognostics and health management : a survey. In *Annual Conference of the Prognostics and Health Management Society, PHM'14,*, volume 68, pages 1–15, Fort Worth - Texas - USA, sep 2014.

[23] Wiem Elghazel, Kamal Medjaher, Noureddine Zerhouni, Jacques Bahi, Ahmad Farhat, Christophe Guyeux, and Mourad Hakem. Random forests for industrial device functioning diagnostics using wireless sensor networks. In *Aerospace Conference, 2015 IEEE*, pages 1–9. IEEE, 2015.

[24] Deborah Estrin, Ramesh Govindan, John Heidemann, and Satish Kumar. Next century challenges: Scalable coordination in sensor networks. In *Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*, pages 263–270. ACM, 1999.

[25] GaoJun Fan and ShiYao Jin. Coverage problem in wireless sensor network: A survey. *JNW*, 5(9):1033–1040, 2010.

[26] Jessica Feng, Farinaz Koushanfar, and Miodrag Potkonjak. System-architectures for sensor networks issues, alternatives, and directions. In *Computer Design: VLSI in Computers and Processors, 2002. Proceedings. 2002 IEEE International Conference on*, pages 226–231. IEEE, 2002.

[27] D Galar, U Kumar, J Lee, and W Zhao. Remaining useful life estimation using time trajectory tracking and support vector machines. In *Journal of Physics: Conference Series*, volume 364, page 012063. IOP Publishing, 2012.

[28] Hassan Hareb, Abdallah Makhoul, and Raphaël Couturier. An enhanced K-means and ANOVA-based clustering approach for similarity aggregation in underwater wireless sensor networks. *IEEE Sensors Journal*, 15(10):5483–5493, 2015.

[29] Hassan Hareb, Abdallah Makhoul, Ramy Tawil, and Ali Jaber. Energy-efficient data aggregation and transfer in periodic sensor networks. *IET Wireless Sensor Systems*, 4(4):149–158, 2014.

[30] TC He, WM Cao, and WX Xie. Coverage analyses of plane target in sensor networks based on clifford algebra. *Acta Electron. Sin*, 37(8):1681–1685, 2009.

[31] Mohamed Hefeeda and Hossein Ahmadi. Energy-efficient protocol for deterministic and probabilistic coverage in sensor networks. *IEEE Transactions on Parallel and Distributed Systems*, 21(5):579–593, 2010.

[32] Aiwina Heng, Sheng Zhang, Andy CC Tan, and Joseph Mathew. Rotating machinery prognostics: State of the art, challenges and opportunities. *Mechanical Systems and Signal Processing*, 23(3):724–739, 2009.

[33] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[34] Andrew KS Jardine, Daming Lin, and Dragan Banjevic. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical systems and signal processing*, 20(7):1483–1510, 2006.

[35] Yiannis Kantaros and Michael M Zavlanos. Distributed communication-aware coverage control by mobile sensor networks. *Automatica*, 63:209–220, 2016.

[36] Gaurav S Kasbekar, Yigal Bejerano, and Saswati Sarkar. Lifetime and coverage guarantees through distributed coordinate-free sensor activation. *IEEE/ACM transactions on networking*, 19(2):470–483, 2011.

[37] L Krishnamachari, Deborah Estrin, and Stephen Wicker. The impact of data aggregation in wireless sensor networks. In *Distributed Computing Systems Workshops, 2002. Proceedings. 22nd International Conference on*, pages 575–578. IEEE, 2002.

[38] Soonmok Kwon, Jae Hoon Ko, Jeongkyu Kim, and Cheeha Kim. Dinamic timeout for data aggregation in wireless sensor netwoks. *Computer Networks*, 55:650–664, 2011.

[39] Lei Li, Baoxian Zhang, and Jun Zheng. A study on one-dimensional k-coverage problem in wireless sensor networks. *Wireless Communications and Mobile Computing*, 13(1):1–11, 2013.

[40] Mo Li, Zhenjiang Li, and Athanasios V Vasilakos. A survey on topology control in wireless sensor networks: Taxonomy, comparative study, and open issues. *Proceedings of the IEEE*, 101(12):2538–2557, 2013.

[41] Mo Li and Baijian Yang. A survey on topology issues in wireless sensor network. In *ICWN*, page 503, 2006.

[42] Zhijun Li and Guang Gong. Survey on security in wireless sensor. *Journal of the Korea Institute of Information Security and Cryptology*, 18(6B):233–248, 2008.

[43] Junbin Liang, Ming Liu, and Xiaoyan Kui. A survey of coverage problems in wireless sensor networks. *Sensors & Transducers (1726-5479)*, 163(1), 2014.

[44] KM Uma Maheswari, S Kanchana Devi, and S Govindarajan. Data aggregation in wireless sensor networks. *Wireless Communication*, 3(6):476–480, 2011.

[45] Abdallah Makhoul, Hassan Harb, and David Laiymani. Residual energy-based adaptive data collection approach for periodic sensor networks. *Ad Hoc Networks*, 35:149–160, 2015.

[46] Abdallah Makhoul, David Laiymani, Hassan Hareb, and Jacques Bahi. An adaptive scheme for data collection and aggregation in periodic sensor networks. *International journal of sensor networks*, 18(1/2):62–74, 2015.

[47] Ronald E McRoberts. Estimating forest attribute parameters for small areas using nearest neighbors techniques. *Forest Ecology and Management*, 272:3–12, 2012.

[48] ISO Condition Monitoring. Diagnostics of machines-prognostics part 1: General guidelines. *ISO13381-1: 2004 (e). vol. ISO/IEC Directives Part 2, IO f. S*, page 14.

[49] Selina SY Ng, Yinjiao Xing, and Kwok L Tsui. A naive bayes model for robust remaining useful life prediction of lithium-ion battery. *Applied Energy*, 118:114–123, 2014.

[50] Gang Niu and Bo-Suk Yang. Intelligent condition monitoring and prognostics system based on data-fusion strategy. *Expert Systems with Applications*, 37(12):8831–8840, 2010.

[51] Al-Sakib Khan Pathan, Hyung-Woo Lee, and Choong Seon Hong. Security in wireless sensor networks: issues and challenges. In *Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference*, volume 2, pages 6–pp. IEEE, 2006.

[52] AK PATIL and AJ PATIL. Issues of connectivity and coverage in wireless sensor networks. *International Journal of Electrical and Electronics Engineering Research (IJEEER)*, 1(3):249–258.

[53] Ying Peng, Ming Dong, and Ming Jian Zuo. Current status of machine prognostics in condition-based maintenance: a review. *The International Journal of Advanced Manufacturing Technology*, 50(1-4):297–313, 2010.

[54] Congduc Pham, Abdallah Makhoul, and Rachid Saadi. Risk-based adaptive scheduling in randomly deployed video sensor networks for critical surveillance applications. *Journal of Network and Computer Applications*, 34(2):783–795, 2011.

[55] Kishore Raja, Ioannis Daskalopoulos, Hamadoun Diall, Stephen Hailes, Tom Torfs, Chris Van Hoof, and George Roussos. Sensor cubes: A modular, ultra-compact, power-aware platform for sensor networks. In *International Conference on Information Processing in Sensor Networks (IPSN SPOTS)*. Citeseer, 2006.

[56] S RUSSELL and Peter Norvig. Artificial intelligence: A modern approach. [sl]: Pearson education. *Inc*, 22:25–26, 2003.

[57] A Saxena and K Goebel. Phm08 challenge data set. *NASA Ames Prognostics Data Repository (http://ti.arc.nasa.gov/project/prognostic-data-repository), NASA Ames Research Center, Moffett Field, CA*, 2008.

[58] JZ Sikorska, Melinda Hodkiewicz, and Lin Ma. Prognostic modelling options for remaining useful life estimation by industry. *Mechanical Systems and Signal Processing*, 25(5):1803–1836, 2011.

[59] Ivanovitch Silva, Luiz Affonso Guedes, Paulo Portugal, and Francisco Vasques. Reliability and availability evaluation of wireless sensor networks for industrial applications. *Sensors*, 12(1):806–838, 2012.

[60] V Sugumaran, V Muralidharan, and KI Ramachandran. Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing. *Mechanical systems and signal processing*, 21(2):930–942, 2007.

[61] Bo Sun, Rui Kang, and Jin-song XIE. Research and application of the prognostic and health management system [j]. *Systems Engineering and Electronics*, 10:041, 2007.

[62] Amirhosein Taherkordi, Majid Alkaee Taleghan, and Mohsen Sharifi. Dependability considerations in wireless sensor networks applications. *Journal of Networks*, 1(6):28–35, 2006.

[63] Di Tian and Nicolas D Georganas. A node scheduling scheme for energy conservation in large wireless sensor networks. *Wireless Communications and Mobile Computing*, 3(2):271–290, 2003.

[64] Di Tian and Nicolas D Georganas. Location and calculation-free node-scheduling schemes in large wireless sensor networks. *Ad Hoc Networks*, 2(1):65–85, 2004.

[65] Di Tian and Nicolas D Georganas. Connectivity maintenance and coverage preservation in wireless sensor networks. ad hoc networks journal. In *Ad Hoc Networks Journal*. Citeseer, 2005.

[66] Jie Tian, Xiaoyuan Liang, and Guiling Wang. Deployment and reallocation in mobile survivability-heterogeneous wireless sensor networks for barrier coverage. *Ad Hoc Networks*, 36:321–331, 2016.

[67] DA Tobon-Mejia, Kamal Medjaher, and Noureddine Zerhouni. Cnc machine tool's wear diagnostic and prognostic by using dynamic bayesian networks. *Mechanical Systems and Signal Processing*, 28:167–182, 2012.

[68] Diego Alejandro Tobon-Mejia, Kamal Medjaher, Noureddine Zerhouni, and Gerard Tripot. A data-driven failure prognostics method based on mixture of gaussians hidden markov models. *IEEE Transactions on Reliability*, 61(2):491–503, 2012.

[69] Javad Akbari Torkestani. An adaptive energy-efficient area coverage algorithm for wireless sensor networks. *Ad hoc networks*, 11(6):1655–1666, 2013.

[70] John Paul Walters, Zhengqiang Liang, Weisong Shi, and Vipin Chaudhary. Wireless sensor network security: A survey. *Security in distributed, grid, mobile, and pervasive computing*, 1:367, 2007.

[71] Lan Wang and Yang Xiao. A survey of energy-efficient scheduling mechanisms in sensor networks. *Mobile Networks and Applications*, 11(5):723–740, 2006.

[72] Xiaorui Wang, Guoliang Xing, Yuanfang Zhang, Chenyang Lu, Robert Pless, and Christopher Gill. Integrated coverage and connectivity configuration in wireless sensor networks. In *Proceedings of the 1st international conference on Embedded networked sensor systems*, pages 28–39. ACM, 2003.

[73] Wei Ye, John Heidemann, and Deborah Estrin. An energy-efficient mac protocol for wireless sensor networks. In *INFOCOM 2002. Twenty-first annual joint conference of the IEEE computer and communications societies. Proceedings. IEEE*, volume 3, pages 1567–1576. IEEE, 2002.

[74] Jennifer Yick, Biswanath Mukherjee, and Dipak Ghosal. Wireless sensor network survey. *Computer networks*, 52(12):2292–2330, 2008.

[75] Fei Yuan, Yiju Zhan, and Yonghua Wang. Data density correlation degree clustering method for data aggregation in wsn. *IEEE Sensors Journal*, 14(4):1089–1098, 2014.

[76] Harry Zhang. The optimality of naive bayes. *AA*, 1(2):3, 2004.

[77] Chuan Zhu, Chunlin Zheng, Lei Shu, and Guangjie Han. A survey on coverage and connectivity issues in wireless sensor networks. *Journal of Network and Computer Applications*, 35(2):619–632, 2012.

[78] Enrico Zio and Francesco Di Maio. A data-driven fuzzy approach for predicting the remaining useful life in dynamic failure scenarios of a nuclear system. *Reliability Engineering & System Safety*, 95(1):49–57, 2010.

[79] Dimitrios Zorbas, Dimitris Glynos, Panayiotis Kotzanikolaou, and Christos Douligeris. B {GOP}: An adaptive algorithm for coverage problems in wireless sensor networks. In *13th European Wireless Conference, EW*, 2007.