OUTLIER DETECTION WITH MACHINE LEARNING IN WIRELESS SENSOR NETWORKS

M. A. Raza[1], A. Mustafa[2], I. Ahmad[3], Muhammad Gulraiz Zahid Awan[4], M. Gul[5], Shaista[6], Amina[7]

[1,2,3,6,7] The University of Lahore, Sargodha Campus
[5]University of Sargodha
[4]Electrical Engineering Department, National University of Computer and Emerging Sciences,Peshawar, Pakistan
Corresponding Author Email: iftikharcau@gmail.com

ABSTRACT: In recent decades, Wireless sensor networks have made significant strides, drawing interest from the scientific and industrial communities. Such networks' scattered sensor nodes function autonomously in challenging environments, leaving them susceptible to mistakes and attacks that could reduce the reliability and accuracy of sensor readings. Sensor readings are categorized as aberrant data, outliers/anomalies when they considerably vary from the predicted healthy behaviors. Such outliers can have a significant influence on the decision-making process and subsequent results in data analytics. As a result, the academic community has recognized the use of machine learning algorithms for outlier detection in WSNs as an innovative and promising methodology. On the basis of numerous viewpoints taken from the body of current research, we present a thorough definition of outliers in this work. We offer a novel and creative method to identify sensor irregularities by utilizing machine learning techniques. By utilizing pattern recognition and anomaly detection methods, machine learning enables us to analyze sensor data and find outliers. We give a comparative assessment of several approaches using machine learning paradigms for outlier detection in WSNs in order to provide a thorough understanding. For academics and practitioners looking to choose the best strategies for their unique application settings, this overview is an invaluable resource. In the end, we explore the main issues surrounding the identification of outliers in WSNs. The dynamic nature of WSNs, the finite resources of sensor nodes, the changing climatic conditions, and the requirement for real-time detection are only a few of the problems that these difficulties cover.

Key words:

INTRODUCTION

Humans constantly create novel technologies based on their requirements. The advancement in shrinking electronic parts, along with the incorporation of wireless capabilities, significantly impacts our daily lives. The widespread use of intelligent mobile devices, such as smartphones, laptops, and smart electronics, in the era after personal computers (PCs), has made information technology devices more accessible, portable, widely available, and prevalent in society. Presently, it is feasible to build a compact embedded system, comparable in capacity to a 1990s PC, in a wallet-sized form factor. (Yu, Krishnamachari, & Kumar, 2006) Many compact, affordable, and low-power smart sensor nodes the size of Nano computers make up ad hoc networks with a particular focus on wireless sensor networks. (Ha, 2006) Due of their real-time applications in a variety of domains, including essential military surveillance, battlefield operations, building security monitoring, forest fire detection, and healthcare, these networks have attracted a lot of study interest. These applications rely on the cooperation and dependability of every node in the network. However, in actual deployments, nodes are vulnerable to several attacks and incursions, which can seriously impair system performance and interrupt network functionality.

One of the best options for accurate environmentalist command and monitoring is a wireless sensor mesh. Recently, there has been a notable transformation in WSN as a result of the quick development of communication technology and sensor technology. This has led to the triumphant adoption of wireless sensor mesh technology in the field of watering and efficient methods to support, enhance, and fortify irrigation. A WSN is defined broadly as a network of nodes working cooperatively to collectively observe and perhaps modify the environment, enabling interaction between people or computing devices and the environment. (Buratti, Conti, Dardari, & Verdone, 2009)

a) Terrestrial WSN
b) Underground WSN
c) Underwater WSN
d) Multimedia WSN
e) Mobile WSN

a) **Terrestrial WSN:** We use the Terrestrial WSN for base station communication. An unstructured network built on nodes is generated in this kind of WSN network. a sensor-based ad hoc network. The main problem with WSN is battery power, although as a workaround, solar

cells are employed as a backup power source. Terrestrial WSNs are networks of hundreds to thousands of wireless sensor nodes that can communicate effectively with base stations and are deployed ad hoc or on purpose. The target region, which is liberated from the solid plane, is randomly covered with sensor nodes in the unstructured mode.

b) **Underground WSN:** Wireless sensor networks that operate underground keep an eye on a variety of subsurface resources, including water, oil, and soil. These sensor networks go by the name UGWSN. Underground wireless sensor networks are more expensive than terrestrial WSNs in terms of deployment, equipment costs, and careful planning. WSNs are made up of several sensor nodes that are buried underground to monitor conditions. To send data from the sensor nodes to the base station, more sink nodes are positioned above the ground. Wireless sensor networks that are buried underground are challenging to recharge.

c) **Underwater WSN:** Networks of underwater sensors. The environment and submerged objects are monitored by acoustic sensor nodes, sinks, and other gadgets. Water covers more than 70% of the world.

Data from these sensor nodes is collected by autonomous underwater vehicles. Underwater communication has issues with bandwidth, significant propagation delays, and sensor failures.

d) **Multimedia WSN:** Multimedia WSNs are the kind of WSNs that can take pictures, record movies, and record sounds. These WSNs are similarly pre-planned, and nodes are dispersed around the environment for coverage and other objectives. The majority of the country is submerged in water. These networks are made up of submerged vehicles and several sensor nodes. Sensor failure, throughput, and significant propagation delays are issues with underwater communication. The WSN has a small capacity battery that can't be changed or recharged underwater.

e) **Mobile WSN:** Mobile Wireless Sensor Networks (MWSN) Sensor nodes are crucial components of today's mobile real-world applications. Because they can employ sensor nodes in any circumstance and cope with quick topographical changes, MWSNs are more adaptable than regular WSNs. A radio transmitter and receiver, a battery, and several sensors (such as light, temperature, humidity, pressure, and motion) are used to power the microcontroller on the mobile sensor terminals. (Sadeghi, Soltanmohammadlou, & Nasirzadeh, 2022)

Architecture of WSN: Based to the data collection, WSNs have a few key sorts of structures, which we will outline for the reader in this paper. After reading it, the reader will comprehend the fundamental structure of WSNs as well as a variety of other WSN-related information. Heterogeneous sensor networks, homogeneous sensor networks, and hybrid sensor networks are all wireless.

All nodes in homogeneous sensor networks have the same amounts of power, storage, processing, and other features. Data aggregation in a flat network is accomplished using data-oriented routing, in which a base station typically floods sensor nodes with query messages, and sensor nodes that have data that matches the query. (Hamami & Nassereddine, 2020)

The deployment and topology management of heterogeneous WSNs is more difficult than it is for homogeneous WSNs. The deployment and topology management techniques for heterogeneous sensor nodes with various communication ranges and sensitivities are presented in this paper. To calculate the cost of building heterogeneous WSNs, we also offer a cost model. The suggested technique can offer a greater coverage rate and a cheaper construction cost for the same sensor node, according to the testing results. (Hamami & Nassereddine, 2020)

In a hybrid sensor network, a number of mobile base stations collaborate to deliver quick real-time data collection. In the scenario depicted, several mobile base stations will relay the data that has been acquired. A wireless network, such as a cellular network, and a wireless sensor network are combined to form a hybrid wireless sensor network. These networks are essential for getting around the highly limited transmission ranges and data rates of conventional sensor networks. This unique feature focuses on hybrid wireless sensor networks made up of base stations and wireless sensor nodes. (Hamami & Nassereddine, 2020)

Components of WSN: Actuator nodes (ANs) and sensor nodes (SNs) are the two different types of nodes. A wireless network without infrastructure is deployed ad hoc using a large number of wireless sensors to track system and physical or environmental parameters. Routers are used to bypass obstacles or increase communication range. In WSNs with integrated CPUs, sensor nodes are utilized to monitor and control the immediate environment. They are linked to a base station, which serves as the WSN network's central processing unit. To share data, a base station in a WSN setup is linked to the Internet.

Sensor Node: Capability to analyze data, gather data, and communicate with connected nodes v Sew. A sensor node's typical performance ranges from 4 to 8 MHz, with 4 KB of RAM, 128 KB of flash memory, and—best of all—a 916 MHz radio frequency. (Zheng & Jamalipour, 2009)

Relay node: This is a connecting node that facilitates communication with its neighbors. It is applied to improve network dependability. Reliability A node is a

special form of field device that lacks both a process sensor and a control device, as well as an interface with the process itself. (Zheng & Jamalipour, 2009)

Actor node: Based on the requirements of the application, this top node executes and constructs a decision. These nodes often have a lot of resources and have better processors, stronger gearboxes, and longer battery lives. Significantly, the gaming node has radio frequencies of 916 MHz, 16 KB of RAM, 128 KB of flash memory, and around 8 MHz processing performance. (Zheng & Jamalipour, 2009)

Cluster head: A high-bandwidth sensing node called a cluster head is employed in WSNs to carry out the data fusion and aggregation functions. Within a cluster, more than one cluster head may exist depending on the system needs and applications. (Zheng & Jamalipour, 2009)

Gate: An interconnection between external networks and sensor networks is called a gateway. The gateway node cluster head is more powerful than the sensor node in terms of program and data memory, utilized CPU, transceiver scope, and potential for extension by external Memory. (Zheng & Jamalipour, 2009)

Introduction to machine learning in WSN: In the late 1950s, the idea of machine learning (ML) was first put forth as a means of simulating artificial intelligence. (Ayodele, 2010) As time went on, the emphasis increasingly turned to the creation of algorithms that are both durable and computationally practical. Machine learning techniques have been widely used over the last ten years for a variety of tasks, such as categorization, prediction, and data analysis across a wide range of domains, including biological information processing, speech interpretation, identifying unwanted messages, visual perception, identifying fraudulent activities, and managing advertising networks. These methods combine algorithms and techniques from a wide range of disciplines, including computer science, mathematics, statistical analysis, and the study of the nervous system.

The essence of machine learning is best summed up by the following two definitions:

a) The improvement of computer programs that aid in learning, resulting in more efficient knowledge acquisition and better system performance. (Duffy, 1997)

b) By identifying and characterizing regularities and patterns in the training data, computational approaches are used to improve machine performance. (Langley & Simon, 1995)

Machine learning plays a crucial role in Wireless Sensor Network (WSN) applications due to the following primary factors:

a) Sensor networks typically observe dynamic surroundings that undergo swift changes over time. For instance, the position of a node may vary due to factors like soil erosion or turbulent sea conditions. The aim is to design sensor networks capable of adjusting and functioning optimally in such dynamic environments.

b) In exploratory applications, Wireless Sensor Networks (WSNs) are useful because they may collect important data from dangerous and unreachable locations. Due of the uncertainty of these contexts, system designers may be forced to create solutions that may not work as intended at first. In these situations, robust machine learning methods that can alter based on freshly learned information are preferred by system designers. (Paradis & Han, 2007)

c) Even while sensor network designers frequently have access to vast amounts of data, they could have trouble identifying significant correlations within. For instance, WSN applications usually specify basic requirements for data coverage, which must be achieved using finite resources of sensor equipment, in addition to the fundamental requirements of maintaining communication connectivity and ensuring energy sustainability. (Romer & Mattern, 2004)

d) Emerging applications and integrations of Wireless Sensor Networks (WSNs), such as in Cyber-Physical Systems (CPS), Machine-to-Machine (M2M) communications, and Internet of Things (IoT) technologies, have been introduced with the goal of encouraging better decision-making and enabling autonomous control. (Wan, Chen, Xia, Di, & Zhou, 2013)

Related Work: Wireless sensor networks (WSNs) are susceptible to security flaws and intrusion assaults, which may compromise user privacy or reduce their overall performance and efficacy. As a result, there has been an increase in research projects aimed at creating effective intrusion detection systems (IDS) customized to the special features of sensor networks. To identify intrusions in WSNs, several researches have suggested machine learning-based IDS solutions. The majority of modern intrusion detection methods use offline learning techniques, including Support Vector Machines, Random Forest, Artificial Neural Networks, Decision Trees, and other tools of a like kind. It's interesting to note that very few research studies have examined the potential advantages of online learning as a substitute strategy for maximizing the benefits provided by these approaches.

The authors (Ifzarne, Tabbaa, Hafidi, & Lamghari, 2021) have presented the ID-GOPA intrusion detection model for wireless sensor networks (WSNs). This methodology was created with the explicit purpose of efficiently identifying intrusions within WSNs. ID-GOPA employs both the information gain ratio and the online passive aggressive algorithm to efficiently handle the continuous flow of data flowing across the network. The primary goal of this approach is to detect unusual activity by carefully analyzing all network events. An offline phase and an online phase are the two separate operating stages of ID-GOPA. Based on the cluster WSN

network architecture, the model not only detects the presence of an intrusion but also categorizes the exact sort of assault. By analyzing the work, we have determined that ID-GOPA has gained overall accuracy of 96% but it is still can be increase by combining an ensemble of algorithm to detect anomalies.

(Zidi, Moulahi, & Alaya, 2017) applies the Support Vector Machine (SVM) technique to classify received sensor data and detect faults using kernel functions.

Fault detection in WSNs presents significant challenges for several reasons. First, the limited resources of sensor nodes hinder the use of conventional techniques that require extensive computational resources. Additionally, the deployment of sensors can occur in hazardous and diverse environments. The detection process needs to be accurate and swift to minimize potential losses.

The use of SVM in WSNs for fault detection imposes no additional burden on the sensors. The entire procedure is carried out at the sink node, which is unrestricted in terms of resources. The cluster head receives the decision function when it has been established from the sink node. Consequently, our method and the cloud-based method reduce the use of sensor resources. In contrast, other techniques such as Bayes, HMM, and SODSEN require executing algorithms at both the cluster head and the sensors themselves to perform fault detection. This makes our technique highly efficient in terms of the constrained resource nodes of the sensors. The author holds the viewpoint that the anticipation of faults is a more efficacious approach in averting errors compared to uncovering them in the moment of occurrence.

In (Warriach & Tei, 2013) authors present a centralized methodology for fault detection in wireless sensor networks (WSNs). This technique relies on a statistical approach and leverages Hidden Markov Models (HMMs). As a supervised machine learning solution, the acquired data was divided into two categories: a training set and a test set. In practical scenarios characterized by offset faults, stuck-at faults, and gain faults, the proposed approach demonstrated commendable performance.

(Obst, 2014) introduces a distributed scheme for detecting faults in wireless sensor networks (WSNs) by employing a recurrent neural network. The author presents a unique methodology that involves training a Self-Organizing Deep Echo State Network (SODESN) to detect faults in WSNs. According to this method, sensor value predictions are based on data obtained from sensors on nearby nodes. The outcomes show how this strategy's distributed computation and local communication capabilities are robust in minimizing WSN link failures. In particular, SODESN performs exceptionally well at anomaly identification, especially in the presence of many faults and realistic link properties. Furthermore, the scalability of SODESN is noteworthy, as it efficiently accommodates an increasing number of WSN nodes in the network by relying solely on local communication with the nearest neighbors.

In (Titouna, Aliouat, & Gueroui, 2016) The author presents a fault detection strategy (FDS) for wireless sensor networks (WSNs) that makes use of both battery power and sensed data to find malfunctioning sensor nodes. Before deciding, each sensor node carefully assesses its state to establish its operational integrity. It then signals that decision to a higher level for secondary verification. A thorough comparison was made between the suggested scheme's performance and that of a meaningful FDWSN technique in terms of a number of parameters, including detection accuracy, false alarm rate, control overhead, and memory overhead. The FDS performs better than the FDWSN, according to simulation data. The FDS's simultaneous evaluation of sensed data and remaining node energy is one standout benefit. This holistic approach enhances the realism of the decision-making process, although the validation of the FDS was solely conducted through simulation.

The techniques discussed earlier in this section proved to be inadequate in meeting the specific constraints of wireless sensor networks (WSNs). Hence, it is advisable to adopt novel data analysis techniques that address the distinctive characteristics and requirements of WSNs, enabling more effective detection of failures.

METHODOLOGY

The research methodology employed in this study encompasses two distinct phases. In the first phase, the identification of outliers is carried out by carefully examining the perspectives of multiple authors derived from relevant literature. By thoroughly reviewing existing works, we aim to gain insights into the various viewpoints and approaches regarding outliers in the context of wireless sensor networks. Special emphasis is placed on discussing the primary sources or causes that lead to the occurrence of outliers. Understanding the outlier phenomena in the field of wireless sensor networks is based on this stage.

In the subsequent phase, an extensive analysis is conducted on multiple techniques that leverage machine learning paradigms for detecting outliers in wireless sensor networks. Various state-of-the-art methodologies and algorithms are studied in detail, considering their effectiveness, applicability, and performance in outlier detection. These techniques are carefully examined to understand their underlying principles, mechanisms, and strengths. Through this comprehensive analysis, we aim to identify and highlight the most promising and effective approaches for detecting outliers in wireless sensor networks. A comprehensive summary of the identified

techniques is presented within this section, providing a consolidated overview of the different machine learning-based methods employed for outlier detection in wireless sensor networks.

What are Outliers?: There are many different definitions of an outlier in the academic community, some of which are included here:

The initial definition for outlier comes from (Grubbs, 1969), "An outlier observation or outlier, is one that deviates markedly from other members of the sample in which it occurs".

(Breunig, Kriegel, Ng, & Sander, 2000) "Outliers are points that lie in the lower local density concerning the density of its local neighborhood".

(Jiang, Tseng, & Su, 2001) "Outliers are points that do not belong to clusters of data set or as clusters that are significantly smaller than other clusters".

(Hawkins, He, Williams, & Baxter, 2002) "Points that are not reproduced well at the output layer with high reconstruction error considered as outliers".

(Muthukrishnan, Shah, & Vitter, 2004) "If the removal of a point from the time sequence results in a sequence that can be represented more briefly than the original one, then the point is an outlier".

(Sadik & Gruenwald, 2011) "An outlier is a data point which is significantly different from other data points, or does not conform to the expected normal behavior, or conforms well to a defined abnormal behavior".

(Titouna, Aliouat, & Gueroui, 2015) "An observation that deviates a lot from other observations and can be generated by a different mechanism".

In this scenario, certain types of extraordinary occurrences, such as system malfunctions and natural calamities, demand distinctive attention. As we are unfamiliar with the appearance of outliers, we can construct a system to identify them based on deviations from the established and defined standard. Ultimately, an outlier within this framework is an exceptional entity that appears captivating and unnecessary simultaneously. This form of outlier detection deviates significantly from conventional methods. However, our objective is to uncover an unconventional relationship, comprehending what transpires and what warrants our scrutiny. Subsequently, we inform the anomaly detector to acknowledge these novel instances as commonplace, perpetuating the cycle for detecting natural events.

Outliers Types: Finding data examples that deviate from the predefined norm is the main goal underlying the development of outlier detection algorithms. (Gupta & Sinha, 2014) Outliers can be classified as either Global or Local outliers, depending on their relationship to and placement within the remaining dataset, keeping this purpose in mind. Global outliers are extraordinary occurrences that show a marked departure from the norm and include all available data points. Such outliers are easily detectable and may then be removed by using a variety of filtering procedures. (Hodge & Austin, 2004) Global outliers can be divided into two categories: Category 1 or second-order external outliers contains all of a sensor node's dataset as outliers in relation to other neighboring nodes. The third-order external outliers, or Category 2 outliers, on the other hand, identify a cluster or subtree of sensor nodes within the structure that may be classified as outliers. These kind of outliers are sometimes referred to as high-order external outliers. On the other hand, local outliers, also known as first-order outliers, classify data points as outliers based on their closeness to nearby local neighbors.

Ways of getting the outliers: In difficult locations where, conventional networks cannot be set up by human involvement, sensor nodes are often used. Sensor nodes are very prone to the emergence of outliers because of the many contextual variables and little resources. The use of outlier identification techniques in WSN is crucial for maintaining the reliability and integrity of data, which guarantees the data's quality. (Ayadi, Ghorbel, Obeid, & Abid, 2017).

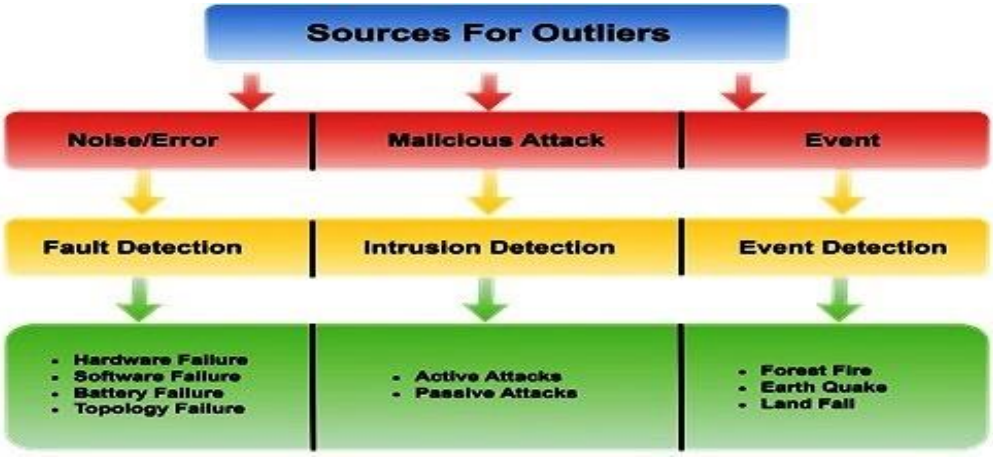

Figure 1: Sources of outliers

Figure 1: Sources of outliers Illustrates various ways of getting outliers.

a) Noise: Noise or errors, which refer to data entries coming from defective nodes, is one of the causes causing the creation of outliers. Typically, inaccurate data denotes arbitrary departures from the rest of the dataset. The main reasons for noise or mistakes are differences related to the environment, severity, and difficulties in the deployment zones. (Chen, Kher, & Somani, 2006)

b) Event: Another source of outliers is a circumstance that is regarded as an unexpected change that occurs within the installed parameters. Examples of such occurrences include chemical leaks, forest fires, floods, volcanic eruptions, earthquake activity, and significant changes in the climate. Events typically occur just once over a lengthy period of time, and when they do, they alter the whole historical pattern of sensed data. Getting rid of anomalous events might loss of great significance as a result secrets on the next event. (Ahmad et al., 2013)

c) Malicious attack: Malicious attacks typically modify the message's semantic content. By controlling a portion of the sensor nodes and supplying fabricated data to risk the integrity of the node or network, these assaults help the appearance of outliers. (Muna, Moustafa, & Sitnikova, 2018) These outliers can be divided into two groups: passive attacks and active attacks. Passive attacks, such as faked attacks, reply attacks, sinkhole attacks, and selective forward attacks, collect data without interfering with network traffic. Active attacks, such as man in the middle and denial of service attacks, collect data by interfering with the setup's normal operation. (Hadri, Chougdali, & Touahni, 2016) (Titouna, Nait-Abdesselam, & Khokhar, 2019)

Outlier detection techniques that adopt machine learning paradigms: Various techniques have been developed so far that are used to detect the outliers in wireless sensor networks. These techniques adopt the machine learning paradigms for detection. In this section we will discuss and compare some of those techniques.

a) Outlier detection using BBN:

A Bayesian Belief Network (BBN) is a directed graph accompanied by a corresponding collection of probability tables. The graph comprises nodes and arcs, where nodes represent variables that can be either discrete or continuous. The arcs within the BBN signify causal or influential connections among variables. The salient characteristic of BBNs lies in their ability to model and analyze uncertainty. In BBNs, we represent the dependence between uncertain variables by populating a node probability table (NPT), which encompasses the conditional probabilities of a node given the states of its parent nodes.

BBNs serve as a means to articulate intricate probabilistic reasoning. They find primary utility in situations necessitating statistical references alongside statements regarding event probabilities. In such cases, users possess certain evidence and aim to infer the probabilities of unobserved events. By leveraging probability theory and Bayes' theorem, one can update the values of all other probabilities in the BBN. BBNs independently allow us to model uncertain events and engage in debates concerning them. However, the true potency of BBNs emerges when we consistently apply the principles of Bayesian probability to propagate the impact of evidence on the probabilities of uncertain outcomes.

The entire BBN procedure can be categorized into three stages:

- Formulating Bayesian Belief Networks
- Acquiring knowledge of Bayesian Belief Networks
- Deriving conclusions from Bayesian Belief Networks

(Janakiram & Kumar, 2006) employed Bayesian belief networks to formulate an anomaly detection scheme. The inclusion of this phenomena to build conditional dependencies among the nodes measurements is justified given that the majority of nearby nodes show comparable readings. BBNs uncover potential anomalies in the collected data by inferring the conditional relationships between the observations. Additionally, this approach can be utilized to assess any missing values.

b) **Detecting outliers with K-nearest neighbor:** Using a simulated wireless sensor network, (Branch, Giannella, Szymanski, Wolff, & Kargupta, 2013) evaluated the behavior of the outlier identification system on real sensor data. These early results show our algorithm's potential by outperforming a strictly centralized strategy in important situations. The node obtaining this data and its nearest neighbors turn into a bottleneck for the whole system when the complete, unfiltered data from the entire sensor network is delivered to a single place.

This quick consequence can shorten the lifespan of the system since the battery-powered nodes located close to the collecting point will run out of energy while those farther away will still have a significant quantity of energy.

The emergence of traffic congestion, which causes message interferences and collisions, is the second effect of this imbalance.

When to modify the size of a sliding window or the number of neighbors in a distance-based outlier detection technique, the KNN (k-nearest neighbors) approach is very useful for determining the confidence of an outlier rating. Additionally, the mean value of the k-nearest neighboring nodes will be used to replace any

missing readings from nodes. However, in order to keep all the information obtained from the monitored environment, this non-parametric approach based on KNN requires a sizable amount of memory.

These particular use cases hold paramount importance in sensor networks with limited resources due to two primary factors. Firstly, communication represents a financially demanding operation. Secondly, the emergence of safety-critical applications relying on wireless sensor networks will necessitate the utmost precision in data, encompassing outlier information.

c) **Detection of selective forwarding attacks using SVM:** The main goal of the deployed sensor network, according to the (Kaplantzis, Shilton, Mani, & Sekercioglu, 2007) simulation of the application, is to immediately alert the central base station to the presence of a moving intruder. This is accomplished by every node initiating a packet with the base station as soon as its sensors identify a vehicle as being nearby. Through the analysis of these packets, the base station can examine the movement pattern and status of the vehicle.

The support vector machines (SVMs) nevertheless show a high level of accuracy in identifying such assaults in the event of an 80% selective forwarding attack. The detection precision of the SVMs, however, declines as the hacker's involvement in the network drops (with targeted source nodes lowered to 50% and 30%). This finding supports the idea that selective forwarding attacks are harder to detect precisely because they are more nuanced than black hole attacks.

In addition to having a flawless detection rate of 100% for black hole attacks, these approaches also have an approximate accuracy of 85% for selective forwarding attacks. Since just the base station is used for the intrusion detection process, there is no need for the sensor nodes to use any energy while yet providing this extra layer of protection.

According to the author, this work demonstrates an innovative use of support vector machines (SVMs) for WSN intrusion detection. Furthermore, it is the first analysis to solely analyze a distributed rather than a distributed intrusion detection system (IDS), eliminating any further effects on node power.

d) **Detecting outlier using SVM:** (Yang, Meratnia, & Havinga, 2008) described an online method that gives each sensor node in the network the ability to quickly categories incoming data readings as either normal or abnormal. Each node in a densely deployed wireless sensor network (WSN) has sufficient information to discover local outliers in real-time by taking use of the significant spatial correlations seen among the sensor readings of neighboring nodes.

As more deployed nodes are added to the network, this method exhibits good scalability. This is attributable to the local process that runs locally on each node and enables it to instantly categories incoming data measurements as normal or aberrant. This method relies only on the solution of a linear optimization problem, maintaining its low computing cost. After the optimization is complete, each node only retains the radius value and a select few of the initial data measurements in memory.

The author conducted a comparison examination of the suggested method against an offline system using both simulated and real data from the Intel Berkeley Research Laboratory. The experimental results demonstrate that the suggested technique outperforms the offline strategy in terms of parameter selection across various kernel functions, improving mining performance.

e) **Analyzing attacks with SOM:** Professor Teuvo Kohonen invented the SOM (self-organizing map) data visualization technique, which is represented as a grid-like network array of neurons. (Kohonen, Schroeder, Huang, & Maps, 2001) The SOM is represented as a multidimensional vector and is made up of neurons, which are a uniform lattice of map units placed on a regular low-dimensional grid.

The neurons used to build the map each have a unique k-dimensional weight vector or prototype vector $mi = [m_{i1}, ..., m_{id}]$, where d stands for the input vectors dimension. By virtue of their close proximity to other units, the neurons are linked to one another. The nodes carefully and gradually adapt to various classes or patterns of input signal. The software transforms nonlinear statistical correlations between data points in a high-dimensional environment into geometric connections between points on a two-dimensional map. By grouping related data elements together, this map successfully depicts how similar the data is.

(Avram, Oh, & Hariri, 2007) described a method for identifying abnormalities in a MANET network by looking at the routing protocol traffic. The underlying concept driving this method is to create baseline models of the regular network protocol behavior and then use these models to identify any anomalous behaviors caused by network attacks. The projected behaviors of the protocols are dramatically altered by these attacks, making them traceable using our method. The detection mechanism is constructed using the SOM model, which can be viewed as a learning algorithm capable of gathering statistical insights about network traffic and protocol behaviors and showing them in a two-dimensional geometric way. The simulation findings presented show the attainment of high detection rates, successfully identifying various network assaults aimed at wireless networks' DSR and AODV routing protocols. The precision and dependability of the detection system are guaranteed by these results, which are attained while maintaining low false alarm rates.

In the preceding section, we investigate different machine learning algorithms that address security concerns in WSNs. Table below provides a comprehensive summary of the examined methods.

Table 1:Summary of WSN outlier detection techniques that uses machine learning

Studies	Machine Learning Algorithm(s)	Predict Missing Data	Distributed/ Centralized	Complexity	Aim
Outlier detection using BBN	Bayesian	Yes	Distributed	Low	Detection of outliers
Detecting outliers with K-nearest neighbor	KNN	Yes	Distributed	Moderate	Detection of distributed outliers
Detection of selective forwarding attacks using support vector machine	SVM	No	Centralized	Moderate	Detect black hole and selective forwarding attacks
Detecting outlier using support vector machine	SVM	No	Centralized	Moderate	Online outlier detection
Analyzing attacks with self-organizing map	SOM	No	Distributed	Moderate	Detect anomalous behaviors

Issues in detecting outliers in WSN: Designing outlier identification systems is complicated and challenging due to the features of sensor data and the context of the sensor network. Although several methods for detecting outliers have been put forth in the literature, resource-constrained Wireless Sensor Networks (WSNs) make these strategies inappropriate. The majority of methods now in use try to find a balance between a high rate of detection and a low rate of false positives while consuming the least amount of energy. However, several difficulties must be taken into consideration in order to create effective outlier identification methods for WSNs. (Ayadi et al., 2017) (Chandola, Banerjee, & Kumar, 2009)

a) **Huge communication cost:** Numerous researches on Wireless Sensor Networks (WSNs) have shown that communication uses more energy than processing. (Gupta & Sinha, 2014) As a result, the computational cost of processing sensor node data is much outweighed by the cost of data transmission. The majority of established outlier identification techniques use a centralized strategy, in which sensor nodes gather data and send it all to a base station or cluster head for preprocessing. While some of these techniques have acceptable detection rates, their transmission costs are rising. However, distributed outlier detection methods only require little communication between sensor nodes, making them appropriate for nodes with limited resources. Significant obstacles, however, include propagation delay, signal absorption, lengthy paths, quickly shifting time-varying channels, noise, and diffusion severely hamper the ability to communicate within these limitations.(Sharma, Golubchik, & Govindan, 2010) As a result, transmission costs are a major obstacle for WSN outlier identification methods.

b) **Variable network topology:** Sensor networks are frequently subject to network outages since they are set up in unexpected locations for a predetermined amount of time. During the course of their assigned activities, certain sensor nodes may move, changing the processing and sensing capabilities. (Hodge & Austin, 2004) Network topology changes as a result of node mobility and communication issues. Additionally, the pre-deployed configuration of the network may change as a result of the addition or removal of additional nodes in accordance with the needs of certain applications. Node failures can, in some circumstances, also result in topological changes in the network. The standard reference model of outlier identification approaches is impacted by these dynamic changes. In addition, WSN uses a combination (thermal, infrared) of sensor nodes to execute various tasks. This kind of variability will further increase the difficulty of the algorithm used to find outliers. (Chirayil, Maharjan, & Wu, 2019)

c) **Resource limitations:** One example of a tiny microelectronic component with limited resources is the sensor node, which has limited power, transmitting, storage, and computing power. (Tran & Huong, 2017) However, in addition to high radio transmission bandwidth, many outlier identification methods created for Wireless Sensor Networks (WSNs) require significant memory for data processing, storage, and complex computational operations. Depending on the situation, sensor networks may only contain cheap sensors because of financial restrictions. As a result, it is extremely difficult to develop outlier detection methods for WSNs that can handle their limited memory for storing and computing duties while still using energy efficiently. (Hendrycks, Mazeika, & Dietterich, 2018)

d) **Challenge of distributed streaming data:** The difficulty of dynamic distributed streaming data in Wireless Sensor Networks (WSNs) is additional. To create a uniform reference for outlier identification in a distributed model, the sensed data must be streamed. It's possible that this information wasn't immediately accessible prior. Due to streaming's dynamic nature, which might change distribution patterns, only disseminated data is available for a set period of time and may not be useful for future study. Many currently used outlier identification methods were created using offline analysis of sensed data and successfully handle dispersed stream data. They might not, however, be appropriate for processing sensor data that is streaming online. (Kim, Choi, & Lee, 2015) Finding methods to assess scattered online stream data while developing outlier identification algorithms in WSNs is thus a major challenge for the academic community.

e) **Huge dimensional data:** In Wireless Sensor Networks (WSNs), a large number of sensed data points with various properties are included. Additionally, a larger network coverage area may result in higher dimensionality in the data. The computing costs of outlier identification methods that rely on these dimensions are greater, and they also place more stress on a sensor node's meagre resources. Additionally, from a performance standpoint, the expansion of data dimensions presents difficulties for the effectiveness of outlier detection. (Chirayil et al., 2019)

Conclusion: Wireless sensor networks differ from typical networks in a number of ways, necessitating the creation of specialized protocols and tools to solve the difficulties and constraints they provide. Utilizing machine learning techniques is one viable strategy since they provide a variety of approaches for improving wireless sensor network adaptation in dynamically changing settings and for detecting abnormal sensor behavior. Particularly, outlier detection serves a crucial and essential function across a variety of application areas since it makes it possible to identify anomalous observations that drastically differ from predicted patterns. In our research, we have examined many subcategories of outliers, considering their various manifestations and traits. Additionally, we have looked at how machine learning paradigms may be used to detect anomalous sensor behavior. Machine learning techniques may be used to effectively identify and indicate occurrences that differ from the norm by using the capability of pattern recognition and anomaly detection. We have developed a comparison summary table that lists numerous outlier detection methods in order to give a thorough insight. This overview helps researchers choose the best strategy for their unique application and requirements by facilitating a clear grasp of the advantages and disadvantages of each technique. Researchers must carefully examine the dataset they will use for testing since outliers might vary in kind, dimensionality, and amplitude. By choosing the right dataset, researchers may make sure that the outlier detection method they have chosen is in line with the particulars of the data at hand, producing findings that are more accurate and dependable. Despite the advancements achieved in outlier identification methods, there are still major obstacles to overcome and unanswered research problems. The dynamic nature of wireless sensor networks, changing environmental conditions, and the necessity for real-time identification are a few of the important concerns connected to outliers that we have addressed. To address these difficulties and improve the area of outlier identification in wireless sensor networks, further research must be done to provide novel strategies, formulas, and tools. By highlighting these challenges, we shed light on the areas that demand further attention and research in order to enhance the effectiveness and efficiency of outlier detection in WSNs. Through continued investigation and innovation, we can advance the field and unlock the full potential of WSNs in various domains. By tackling these unresolved difficulties, we may expand the potential of wireless sensor networks across a range of application domains.

REFERENCES

Ahmad, N., Hussain, M., Riaz, N., Subhani, F., Haider, S., Alamgir, K. S., & Shinwari, F. (2013). Flood prediction and disaster risk analysis using GIS based wireless sensor networks, a review. *Journal of Basic and Applied Scientific Research*, *3*(8), 632-643.

Avram, T., Oh, S., & Hariri, S. (2007). *Analyzing attacks in wireless ad hoc network with self-organizing maps.* Paper presented at the Fifth Annual Conference on Communication Networks and Services Research (CNSR'07).

Ayadi, A., Ghorbel, O., Obeid, A. M., & Abid, M. (2017). Outlier detection approaches for wireless sensor networks: A survey. *Computer Networks*, *129*, 319-333.

Ayodele, T. O. (2010). New advances in machine learning. *InTech: Rijeka, Croatia*, 19-49.

Branch, J. W., Giannella, C., Szymanski, B., Wolff, R., & Kargupta, H. (2013). In-network outlier detection in wireless sensor networks. *Knowledge and information systems*, *34*, 23-54.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). *LOF: identifying density-based local outliers.* Paper presented at the Proceedings of the 2000 ACM SIGMOD international conference on Management of data.

Buratti, C., Conti, A., Dardari, D., & Verdone, R. (2009). An overview on wireless sensor networks

technology and evolution. *Sensors, 9*(9), 6869-6896.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR), 41*(3), 1-58.

Chen, J., Kher, S., & Somani, A. (2006). *Distributed fault detection of wireless sensor networks.* Paper presented at the Proceedings of the 2006 workshop on Dependability issues in wireless ad hoc networks and sensor networks.

Chirayil, A., Maharjan, R., & Wu, C.-S. (2019). *Survey on anomaly detection in wireless sensor networks (WSNs).* Paper presented at the 2019 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD).

Duffy, A. H. (1997). The" what" and" how" of learning in design. *IEEE Expert, 12*(3), 71-76.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics, 11*(1), 1-21.

Gupta, S. K., & Sinha, P. (2014). Overview of wireless sensor network: a survey. *Telos, 3*(15µW), 38mW.

Ha, R. W. K. (2006). A Sleep-Scheduling-Based Cross-Layer Design Approach for Application-Specific Wireless Sensor Networks.

Hadri, A., Chougdali, K., & Touahni, R. (2016). *Intrusion detection system using PCA and Fuzzy PCA techniques.* Paper presented at the 2016 International Conference on Advanced Communication Systems and Information Security (ACOSIS).

Hamami, L., & Nassereddine, B. (2020). Application of wireless sensor networks in the field of irrigation: A review. *Computers and Electronics in Agriculture, 179*, 105782.

Hawkins, S., He, H., Williams, G., & Baxter, R. (2002). *Outlier detection using replicator neural networks.* Paper presented at the Data Warehousing and Knowledge Discovery: 4th International Conference, DaWaK 2002 Aix-en-Provence, France, September 4–6, 2002 Proceedings 4.

Hendrycks, D., Mazeika, M., & Dietterich, T. (2018). Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review, 22*, 85-126.

Ifzarne, S., Tabbaa, H., Hafidi, I., & Lamghari, N. (2021). *Anomaly detection using machine learning techniques in wireless sensor networks.* Paper presented at the Journal of Physics: Conference Series.

Janakiram, D., & Kumar, A. (2006). *Outlier detection in wireless sensor networks using Bayesian belief networks.* Paper presented at the 2006 1st International conference on communication systems software & middleware.

Jiang, M.-F., Tseng, S.-S., & Su, C.-M. (2001). Two-phase clustering process for outliers detection. *Pattern recognition letters, 22*(6-7), 691-700.

Kaplantzis, S., Shilton, A., Mani, N., & Sekercioglu, Y. A. (2007). *Detecting selective forwarding attacks in wireless sensor networks using support vector machines.* Paper presented at the 2007 3rd International Conference on Intelligent Sensors, Sensor Networks and Information.

Kim, S., Choi, Y., & Lee, M. (2015). Deep learning with support vector data description. *Neurocomputing, 165*, 111-117.

Kohonen, T., Schroeder, M., Huang, T., & Maps, S.-O. (2001). Springer-verlag new york. *Inc., Secaucus, NJ, 43*(2).

Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM, 38*(11), 54-64.

Muna, A.-H., Moustafa, N., & Sitnikova, E. (2018). Identification of malicious activities in industrial internet of things based on deep learning models. *Journal of information security and applications, 41*, 1-11.

Muthukrishnan, S., Shah, R., & Vitter, J. S. (2004). *Mining deviants in time series data streams.* Paper presented at the Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004.

Obst, O. (2014). Distributed fault detection in sensor networks using a recurrent neural network. *Neural processing letters, 40*, 261-273.

Paradis, L., & Han, Q. (2007). A survey of fault management in wireless sensor networks. *Journal of Network and systems management, 15*, 171-190.

Romer, K., & Mattern, F. (2004). The design space of wireless sensor networks. *IEEE wireless communications, 11*(6), 54-61.

Sadeghi, S., Soltanmohammadlou, N., & Nasirzadeh, F. (2022). Applications of wireless sensor networks to improve occupational safety and health in underground mines. *Journal of safety research*.

Sadik, S., & Gruenwald, L. (2011). *Online outlier detection for data streams.* Paper presented at the Proceedings of the 15th Symposium on International Database Engineering & Applications.

Sharma, A. B., Golubchik, L., & Govindan, R. (2010). Sensor faults: Detection methods and prevalence in real-world datasets. *ACM Transactions on Sensor Networks (TOSN), 6*(3), 1-39.

Titouna, C., Aliouat, M., & Gueroui, M. (2015). Outlier detection approach using bayes classifiers in wireless sensor networks. *Wireless Personal Communications, 85*, 1009-1023.

Titouna, C., Aliouat, M., & Gueroui, M. (2016). FDS: fault detection scheme for wireless sensor networks. *Wireless Personal Communications, 86*, 549-562.

Titouna, C., Nait-Abdesselam, F., & Khokhar, A. (2019). *A multivariate outlier detection algorithm for wireless sensor networks.* Paper presented at the ICC 2019-2019 IEEE International Conference on Communications (ICC).

Tran, K. P., & Huong, T. T. (2017). *Data driven hyperparameter optimization of one-class support vector machines for anomaly detection in wireless sensor networks.* Paper presented at the 2017 International Conference on Advanced Technologies for Communications (ATC).

Wan, J., Chen, M., Xia, F., Di, L., & Zhou, K. (2013). From machine-to-machine communications towards cyber-physical systems. *Computer Science and Information Systems, 10*(3), 1105-1128.

Warriach, E. U., & Tei, K. (2013). *Fault detection in wireless sensor networks: A machine learning approach.* Paper presented at the 2013 IEEE 16th International Conference on Computational Science and Engineering.

Yang, Z., Meratnia, N., & Havinga, P. (2008). *An online outlier detection technique for wireless sensor networks using unsupervised quarter-sphere support vector machine.* Paper presented at the 2008 International Conference on Intelligent Sensors, Sensor Networks and Information Processing.

Yu, Y., Krishnamachari, B., & Kumar, V. P. (2006). *Information processing and routing in wireless sensor networks*: World Scientific.

Zheng, J., & Jamalipour, A. (2009). *Wireless sensor networks: a networking perspective*: John Wiley & Sons.

Zidi, S., Moulahi, T., & Alaya, B. (2017). Fault detection in wireless sensor networks through SVM classifier. *IEEE Sensors Journal, 18*(1), 340-347.