Accurate Detection of IoT Sensor Behaviors in Legitimate, Faulty and Compromised Scenarios

Keshav Sood, Mohammad Reza Nosouhi, *Member, IEEE*, Neeraj Kumar, *Senior Member, IEEE*, Anuroop Gaddam, Bohao Feng, *Member, IEEE*, and Shui Yu, *Senior Member, IEEE*

Abstract—In smart farming sector, Internet of Things (IoT) based smart sensing systems are vulnerable to failure, malfunction, and malicious attacks. Also, sensors are deployed often in an alien and harsh environment. Here, the conditions are not well supportive which either causes the sensor to fail prematurely or gives unusual and erroneous readings, known as outliers. This effects the smart network's performance and decision-making ability in many ways. Therefore, it is important to accurately detect the IoT sensor behaviour in legitimate, faulty, and compromised or attack scenarios. To distinguish the sensor behaviour in different scenarios we have proposed a feasible approach using spatial correlation theory which is validated using Moran's *I* index tool. We have used Classification and Regression Trees (CART), Random Forest (RF), and Support Vector Machine (SVM) models to test our approach. For real-time anomaly detection we have used an edge computing technology. We have compared the proposed approach, using Forest Fire real dataset, with the three existing recent works. Our results are promising in terms of accurate detection of IoT sensor behaviours in real-time. This will assist the precision farming industry in making better decisions to securely manage IoT field network, increase productivity, and improves operational efficiency.

Index Terms—Internet of Things, IoT behaviour, outlier detection, security, precision farming

1 Introduction

Internet of Things (IoT) has provided potential opportunities to redesign powerful industrial systems and applications [1]. In precision agriculture sector (smart farming), the IoT technology has transformed the conventional farming ways at massive scale and has brought agricultural revolution with promising ways to significantly boost economy [2]. With the emergence of precision farming, growing trend in diverse IoT applications are being observed to optimize farming operations. Such as: sensing soil moisture and nutrients, determining the optimal time to plant and harvest, reporting weather conditions, climate monitoring and manufacturing purposes, real-time soil monitoring in order to collect moisture data to enhance crop conditions, increasing efficiency and crop yields. etc. [3], [4].

In the United States, the IoT applications are very much advanced. It is estimated that up to 40% of growers (producing 70% of the crop) are using IoT applications to monitor and manage their cropping program [5]. AgThentic report predicted that IoT device installations will jump to 75 million in 2022, from 30 million in 2015, with an average annual growth rate of 20% [6]. Fundamentally the sensors are the key elements embedded within IoT devices which are being used to collect the valuable raw data. This valuable raw data is used for various purposes such as: designing IoT access control model, controlling water usage for optimal plant growth and so on. We note that the correct acquisition of data plays a primary role in any network operation and network's overall performance [7].

Unfortunately, *firstly* the IoT sensors are deployed in harsh environments which affects the IoT sensor behaviour and eventually the complete network operation. *Secondly*, IoT devices are typically vulnerable to malicious events [8]. Further, the use of smart communication technologies has significantly increased the risk of cyber-attacks, threats and introduce vulnerabilities in smart farming environments [9], [10], [11]. These cyber-attacks may disrupt the economies of countries that are widely dependent on agriculture [11], [12], [13]. Therefore, guaranteeing the sensor's correct operation and predicting malfunctions, etc. is essential. To elaborate this, below we show three main scenarios in which the IoT sensors may generate faulty and compromised malicious sensor streams.

Firstly, the *Intrinsic Sensor Errors*, i.e., faulty readings or measurements coming from genuinely faulty sensors [14], [15]. We call this as legitimate behavior. *Secondly*, the *Sensor Events*, where unprecedented change is recorded, e.g., if a worm crawls on the sensor, the sensor will generate data that will show how moist and warm the worm is [16]. We call this as faulty behaviour which also impacts the intelligent decision-making ability of IoT sensing system. *Finally*, the *Intermittent Sensor Errors*, we call this as compromised scenarios. These are caused due to malicious activity or event such as: tampering the sensor device etc. [8], [17]. Here,

- *Keshav Sood, Mohammad Reza Nosouhi, and Anuroop Gaddam are with Deakin University, Geelong, VIC 3220, Australia. E-mail: {keshav.sood, m.nosouhi, anuroop.gaddam}@deakin.edu.au.*
- *Neeraj Kumar is with Thapar University, Patiala, Punjab 147004, India. E-mail: neeraj.kumar@thapar.edu.*
- *Bohao Feng is with Beijing Jiaotong University, Beijing 100044, China. E-mail: bhfeng@bjtu.edu.cn.*
- *Shui Yu is with the University of Technology, Sydney, Ultimo NSW 2007, Australia. E-mail: shui.yu@uts.edu.au.*

Manuscript received 23 Dec. 2020; revised 5 Nov. 2021; accepted 22 Nov. 2021. Date of publication 1 Dec. 2021; date of current version 16 Jan. 2023. This work was partially supported by ARC under Grant 200101374. (Corresponding author: Keshav Sood.) Digital Object Identifier no. 10.1109/TDSC.2021.3131991

almost in every scenario the IoT sensor generates conflicting information, which is to be processed by machine learning models to make intelligent decisions, e.g., autonomous precision farming, etc. [18], [19]. The data streams generated in these situations affects the ability of intelligence models to enforce decisions which may affects every aspect of farmers work – from livestock to crop farming [12], [20]. *Hence it is very important to understand the IoT sensors behaviour such as: legitimate behavior, faulty and compromised behavior.*

One of the ways to distinguish the sensor behaviour is "detecting outliers". To detect anomalies the existing research is more focused on Wireless Sensor Networks (WSN). Some key areas are fraud detection, network security breaches, target tracking, environmental and health monitoring etc. [8], [21]. We have observed that little efforts have been made to address this critical issue in IoT farming domain. This is due to the IoTs unique characteristics (the three scenarios we have discussed) in comparison to WSNs. Researchers in [22], [23], and [24] have also mentioned significant differences between WSN and IoT sensors. We encourage readers to follow [25], [26], and [27] also. They emphasized that the conventional approaches used for outlier detection in WSN domain are not suitable in IoT based scenarios. Apparently, we cannot use conventional outlier detection solutions already available in WSNs on IoTs.

Researchers have pointed out that *"IoT is a compendium of multiple-similar sensors embedded as a unit capable of producing huge amounts of spatial-temporal data at low latency, unlike the WSN"* [8]. Unfortunately, *"recent works have considered WSN applications as IoT applications without distinguishing the new features that characterize this denomination"* [24]. Therefore, using novel methods, in case of a) substantially high volume of data generated by the sensing unit within the IoT due to the failure or malfunctioning of one sensor out of a pair, b) identifying the faulty sensor in order to make that data redundant in real time, and c) at the same time allowing the data produced by other sensors is extremely essential for accurate functioning of the IoT sensing device. In [28] researchers have mentioned that detecting patterns in real time data streaming data is crucial and significant problem. Further massive growth of various types of sensor units and devices is being seen. Thus, it is vital to investigate and understand the regular and irregular patterns of IoT sensing networks, particularly in real-time scenarios. In IoT sensing domain, this helps to enable predictive analytics for automated notification and decision support.

To address this, we classify the existing research into three categories: Statistical Techniques, Nearest Neighbour Techniques, and Artificial Intelligence Techniques [29], [30]. Doubtlessly, there are hundreds of papers available to detect outliers in different domains. We emphasize that the specific domain and problem we have worked on is unique and to much extent is different than the problems already mentioned in IoT domain. Motivated from this, using spatial-correlation method we have proposed a promising approach to detect outliers in real-time. We are among the early ones addressing this issue in the specified problem domain. Further, majority of the methods are successfully applicable in WSN domain rather than in IoT domain. We have designed a prototype using Python 3.7.4, Eclipse Paho MQTT, and used Forest Fire real dataset. The proposed module resides at the edge of the field network, which we have shown, for real-time IoT network monitoring.

Our contributions in this paper are as below:

1) We have proposed an approach to accurately distinguish the legitimate, faulty, and malicious sensor outliers or streams/events. For experiments, we have used Forest Fire real data set to evaluate our approach in smart farming sector as a use case.
2) The results obtained from the comparison of our approach with three recent approaches show that our approach is highly effective, i.e., accurate in detecting IoT sensor events in real-time. This also testifies that the proposal is contributing a new knowledge in this domain.
3) We have also proposed that the module resides at the edge of the network to reduce the data processing and computation latency of sensor stream analyses. The performance evaluation of the approach has been shown to distinguish sensor behaviours in real-time.

Benefit. The results of this early work provides further insights to network administrators to distinguish faulty and malicious sensor units which eventually would save unnecessarily time and cost for IoT sensing network maintenance. Also, this can early alarm the risk of cyber-attacks and timely detection of anomaly connections in smart farming domain, eventually would ease the complex network management operation. The approach will be useful in multiple sectors and scenarios other than in the precision farming sector.

Novelty. In contrast to the existing works, we have noticed that the existing work in smart farming sector detect outliers using different techniques (in WSN domain). However, they do not integrally considered approaches to distinguish sensor failure and sensor attacks streams/outliers, which we have interestingly proposed here, in IoT sensing context. For experiments and proposals validation, the majority of the existing works use WSN platforms instead of IoT platforms. We emphasise that the prototype implementation, we have shown using Python 3.7.4 and Eclipse Paho MQTT, is a true reflection of IoT test bed. Therefore, results obtained from our prototype and our approach are approximate to real scenarios. The comparison of our approach with existing works validates this point.

2 Related Work

In this section, we have classified the existing research into three categories; Statistical Techniques, Nearest Neighbour Techniques, and Artificial Intelligence Techniques. In the first technique, using stochastic distribution modelling outliers or faults are identified when the likelihood of the data instance is very low. Here, the model complexity also increases as the volume of the sensor data increase, this affects the model's performance to act in real-time situations. In the second approach, the sensor fault and outlier detection mainly rely on the notion of proximity. k-nearest neighbours model is the most common method to detect the outliers or to differentiate between abnormal and the correct data streams. Finally, the third approach is based on Artificial Intelligence (AI) based techniques. In this approach, Neural Networks and Fuzzy Logic based methodologies are currently very

popular [31], [32], [33]. These approaches are useful to improve decision making, enhance the clustering head selection, improve network security, etc.

In the three existing techniques there are some limitations [16], [34]. For example, the statistical techniques are not effective for data-intensive IoTs working in a real-time setting, as well they incur high computational cost of managing multivariate data produced. Nearest Neighbour Techniques have scalability concerns, especially in IoT context. They also give a high false-negative rate for sensor faults and outlier detection. Finally, the AI Techniques require extensive fine-tuning before being made operational in a real-life setting. Furthermore, as the work-flow of these approaches are based on rule-based strategies, here, if the number of sensor data variables increases this necessities to exponentially increase the number of rules. This poses an adverse effect on the performance of the IoT networks.

Further, in [9], an interesting smart farming ecosystem has been proposed to encode different farm–specific entities (e.g., sensors, machineries, workers, etc.) and their interactions. Authors have developed an attribute–based access control system that dynamically evaluates access control requests. However, the paper does not discuss the scenarios in which an attacker impersonates a network entity (e.g., a sensor). In fact, in such a scenario, the proposed access control mechanism will grant the attacker the same access rights that would be granted to the relevant impersonated entity. This is because it just determines what access rights (to different services) are granted to an entity that has already been authenticated. However, in our paper, we have provided a solution for the detection of false data injected by an attacker who has impersonated a legitimate sensor device in the network. Thus, we believe that integrating our solution into [9] would result in a secure smart farming architecture that is resilient to spoofing attacks.

In [34], authors have proposed an approach to detect anomaly in big IoT data streams using a one-class support Tucker machine (OCSTuM) and an OCSTuM based on tensor Tucker factorization and a genetic algorithm called GA-OCSTuM. These unsupervised anomaly detection in IoT domain approaches extend one-class support vector machines to tensor space. In [35], a learning–based attack detection mechanism has been proposed for IoT networks. In their proposed approach, network traffic features (e.g., source and destination addresses) are extracted and learned by a learning–based model to detect different types of attack. The authors have achieved detection accuracy of up to 99% for decision tree, RF, and Artificial Neural Network (ANN) classifiers. However, their work relies on the real–time extraction of features from network traffic and performing feature engineering tasks to make the features compatible with the input format of the deployed machine learning models (e.g., converting categorical features to their equivalent numeric vectors). In [32] an IoT architecture is proposed in order to detect the occurrence of both Error and Event in a forest environment with the help of statistical models. Authors have used Classification and Regression Trees (CART), Random Forest (RF), Gradient Boosting Machine (GBM) and Linear Discriminant Analysis (LDA).

Another interesting work in this field has been done by Pacheco *et al.* [36] in which an Intrusion Detection System (IDS) is developed for IoT networks. The proposed approach is based on Anomaly Behavior Analysis (ABA) and includes the use of sensor-DNA profiles (s-DNA). However, the authors have indicated that their approach is intended to protect those IoT networks with a limited number of sensors. In WSN domain, the work in [37] argued that the high false alarm rate alters the decision making of sensing systems. The false alarms occur due to the inaccurate and unreliable sensor streams. Therefore, to detect outliers in sensor data efficiently and accurately is an important problem. Authors have proposed an outlier detection algorithm (TSVDD) using model selection-based support vector data description (SVDD). Authors [38] have mentioned that the WSNs faults and outlier detection is very challenging for ensuring the quality of data analysis. To detect the outliers, authors have used support vector data description method (ID-SVDD). The Parzen-window algorithm and Mahalanobis distance (MD) approach is used to validate the outlier detection performance. The approach achieved high performance and successful in water quality monitoring.

Very recently, authors in [39] mentioned that the sensor data failure often occurs due to human intervention or changes in the environmental conditions. This eventually knock downs the IoT the sensing system. Researchers have proposed an online and credible data integrity monitoring (DIM) approach. In this approach the data failure is modeled by dividing it into format failure, timing failure and value failure. Also they have proposed heuristic policies to detect and isolate data failure, and show that the approach significantly improves the data quality.

Recent studies indicate that cloud–based services are getting popular but potentially vulnerable to different cyber threats that can cause severe disruptions in cloud–based IoT applications (e.g., DoS attacks, Malicious Insiders, Insecure API, etc.). In this regard, Garg *et al.* [40] have reviewed a number of cyber threats on cloud services and presented several real–world attacks on these systems. Further, researchers have proposed several anomaly detection techniques for cloud security. Sha *et al.* [41] proposed a multi–order anomaly detection technique based on the Markov chain model. They provided a new indicator of anomalies by monitoring the relative relations between results from different–order models. In [42], a practical framework is introduced that provides scalable and lightweight anomaly detection for sensor data. In the proposed framework, distributed fuzzy c–means (FCM) clustering is integrated with a lightweight cryptography technique to perform privacy–preserving operations on encrypted data.

The recent anomaly detection models for cloud services has been proposed in [43]. Stacked and Bidirectional LSTM models are utilized to perform anomaly detection in openstack cloud environment. A cross–dataset anomaly detection mechanism for cloud systems has been proposed in [44]. The anomalies in a new unlabelled target dataset are detected by training an anomaly detection model on an existing labelled dataset (source dataset). This is done by combining transfer learning and active learning models. In fact, transfer learning is used to transfer knowledge from the source dataset to the target dataset while active learning is utilized to determine the labels of the samples in the target dataset. Overall, the problem

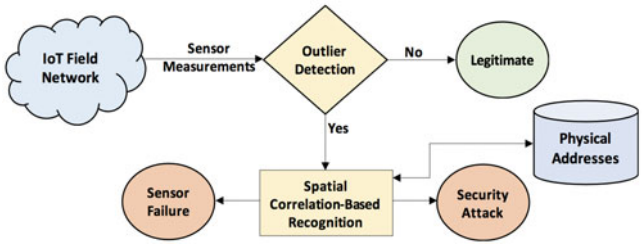


Fig. 1. The proposed high level system model.

of anomaly detection in cloud based IoT networks is a critical problem, and we note that with 5G infrastructure where the latency is very sensitive, the detection of sensor event accurately is a big problem. The work we presented is timely and offers important insights to the IoT domain. The use-case is well motivated by recent interest in precision agriculture techniques.

3 The Proposed Approach

In this section, we have presented our proposed approach for outlier detection in three cases, i.e., legitimate, faulty, and attack scenarios. We consider an IoT field network in which various IoT–enabled sensors have been deployed to measure environmental parameters (e.g., temperature). The sensors are installed at a fixed distance to each other. There are IoT gateways that collect measurements from the sensors throughout the field. We assume the sensor measurements are securely stored in either a fully trusted cloud-based database or on a secure blockchain–based database. Regarding the attacker capabilities, we assume the attacker device is equipped with the computation, communication, storage, and power resources larger than what the IoT devices have been equipped with. In addition, the attacker is assumed to be capable of either impersonating IoT sensors in the network through impersonation attack methods (e.g., IP and ARP spoofing), or forcing sensors to send incorrect data to the IoT gateway (through installing malwares on them). In all of these cases, the final target of the attacker is to inject false sensor data into the database.

In Fig. 1 we have shown a high-level view of our proposed approach which is based on spatial correlation between sensor measurements. The Algorithm 1 shows step wise execution of the proposed approach in which we have classified the approach in three phases, i.e., pre-processing phase, outlier detection phase, and spatial correlation-based classification phase. Firstly, the IoT sensors generated data is accumulated for data pre-processing. In the second phase, for outlier detection, this accumulated data or sensor readings collected from the IoT network are processed and analyzed. The behaviour of the IoT network is labelled as legitimate behaviour in case no outlier is detected. Otherwise, in the presence of any outlier in the IoT data analyses, the third phase of the Algorithm 1 is invoked. In this case, the IoT data is forwarded to the third phase of the proposed approach which uses Spatial Correlation-Based Recognition module which determines the spatial-correlation in the data eventually to accurately classifying the detected outliers into sensor failure or security attack clusters.

Algorithm 1. The Process Flow Algorithm of the Proposed Approach

Pre–Processing Phase
Inputs: $dataset$ (history of sensor measurements),
r (train_test ratio)
Outputs: Train and test datasets: d_{train}, d_{test}
1: Data Cleaning: $dataset\_cleaned = dataset.clean()$
2: Data Scaling:
$dataset\_cleaned\_scaled = dataset\_cleaned.scale()$
3: Partitioning:
$d_{train}, d_{test} = train\_test\_split(dataset\_cleaned\_scaled, r)$
4: output d_{train} and d_{test}

Outlier Detection Phase:
Inputs: Learning model: $model \in \{CART, RF, SVM\}$
Train and test datasets: d_{train}, d_{test}
Evaluation metrics:
$M \in \{Accuracy, F-Score, Precision, Recall\}$
New measurement of sensor s_i: m_i
Outputs: $detection\_result \in \{0, 1\}$
1: Model Training: $model.fit(d_{train})$
2: Model Evaluation: $model.val(d_{test}, M)$
3: Outlier Detection: $detection\_result = model.predict(m_i)$
4: output $detection\_result$

Spatial Correlation–Based Classification Phase
Inputs: $detection\_result$, N (number of sensors), threshold T
$\forall s_i, i = \{1, 2, \ldots, N\}$:
Location data of s_i: $L_i = (x_i, y_i)$
m_i and $\bar{m}$ (mean value of the feature)
Output:
$outcome \in \{sensor\,failure, compromised\ measurement\}$
1: Moran's I index calculation:
$\forall i, j \in \{1, 2, \ldots, N\}\ \ w_{ij} = \frac{1}{d_{ij}}$,
where $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$
$\delta_i = \frac{\sum_{j=1, j\neq i}^{N} w_{ij}}{N-1} - \bar{m}$
$I_i = \frac{(m_i - \bar{m})}{\delta_i^2} \sum_{j=1, j\neq i}^{N} w_{ij}(m_j - \bar{m})$
2: **if** $I_i < T$:
$outcome = sensor\,failure$
else:
$outcome = compromised\ measurement$
4: output $outcome$

We emphasise that to accurately classify the outliers in different categories (*sensor failure*, *security attack*) spatial correlation theory is adopted to compute the spatial correlation between the detected outlier and the data readings of nearby sensors. We have considered two factors in our approach, i.e., IoT sensors for humidity and temperature measurement in field. We assume that the proposed approach resides at the edge of the IoT field network to minimize the latency of data processing and outlier detection. This is feasible using edge computing and other similar technologies. Our proposed approach (in phase 3) uses Classification and Regression Trees (CART), Random Forest (RF), and Support Vector Machine (SVM) machine learning models. Further detail is given below.

3.1 Stage 1: Outlier Detection

Legitimate Data Stream. Here, the basic idea we have used is derived from the physical phenomenon of data. We

have used two factors in our data set, i.e., humidity and temperature. It is fundamentally valid that over short distances these two factors would not change dramatically. Often, in the field networks the sensors are densely deployed, therefore, between the outliers we do not observe any spatial correlation. Naturally with the high probability we can observe that the spatial correlation of any sensor data with respect to its nearby sensors data would be very high. The data readings will be highly correlated. Eventually, the data stream is considered as *legitimate*. Overall, this behaviour of any IoT field network would be considered as *normal behaviour*.

3.2 Stage 2: Spatial Correlation–Based Classification

As we have mentioned that if no outlier is detected, the data stream is considered as legitimate. Otherwise, the data stream is forwarded to Spatial Correlation-Based Recognition module, as shown in Fig. 1.

Sensor Failure Data Stream. To accurately determine and to classify IoT sensor failure data streams from IoT security attack data streams we measure the correlation of any IoT field sensor from its neighbouring sensors. Again, it is natural that in scenarios where a sensor is legitimately faulty will generate faulty data. In such a case an outlier will be detected. However, the data streams generated from neighbours' sensors will not be the same because it is highly unlikely (the probability is very low) that at the same time the neighbours' IoT sensors will be faulty too. In this case, neighbours' sensors data points are not categorised as outlier as no considerable amount of correlation or similarity will be detected. With this way of anomaly detection, the IoT field network behaviour would be categorised as faulty behaviour. However, it can be argued that in scenarios where neighbours' sensors are faulty too there would be more outliers detected. In such scenarios, it is complex for the approach to accurately distinguish and categorise the faulty and attack behaviour of IoT field sensors. To alleviate this, below we have further extended an idea.

Attack Data Stream. We have taken a valid assumption that in case of a security attack (such as spoofing, man-in-the-middle, or jamming attacks) several neighbouring sensors are targeted by the attacker. As a result, faulty measurements are received from these sensors too. For example, in a jamming attack, an attacker affects the radio communication of all the sensors located in a specific geographical area by sending powerful fake radio signals. Moreover, in a man-in-the-middle attack, an attacker propagates jamming radio signals as well as spoofing signals to eavesdrop or alter the content of radio communication between IoT sensors and devices. In the mentioned examples, it is highly possible that several neighbourhood sensors placed near or next to the victim node (attacked node) will also be affected. Thus, in case of a security attack, spatial correlation between the outliers is detected at a significant level.

Now, the location data of the outliers detected in this case are fed into the classifiers. Here the classifiers would make the clustering decisions based on the possible spatial correlation between the outliers. It can be said that the outliers originated from the same area would be classified into the security attack class and into the sensor failure class. This IoT system behaviour would be considered as attack behaviour. Now, the next challenge is to determine the particularly attacked sensors as well as their location.

In our approach we consider that each IoT field sensor has its IP address through which the sensor returns the data streams with their x and y coordinates values. The complete geographical IoT field network under measurement is taken as a grid in which each sensor has its x and y coordinate values based on its physical location. The above consideration is valid for static IoT sensors. In real-world there are many scenarios where field sensors are mobile. We consider that in such scenarios the same process can be applied (to get data streams as well as sensor location in real-time) in a dynamic way since the IoT sensors location changes w.r.t time. Now, below a mathematical model of the approach is discussed to present the approach in a theoretical way and to better comprehend the Algorithm 1.

4 Mathematical Analysis

In this section, we have analyzed the proposed classification approach and discuss the efficiency and merit of the proposed spatial correlation-based approach. As we have discussed in Section 3, we have considered the spatial correlation between measurements made by neighbor sensors to distinguish a security attack from a sensor failure. To measure spatial dependency, we have used *Moran's I* index [45], [46] theory which is well known in statistics. Spatial dependency is the co-variation of properties within a geographic space, (IoT field network in our case). *Moran's I* index uses both feature locations and feature values to determine how close the IoT sensor measurements are in comparison with other IoT field network data streams. In the given set of features it evaluates whether the pattern expressed is clustered, dispersed, or random. Moran's I index has global and local versions which shows the level of spatial correlation. In global version we determine how different or similar the sensor measurements are from its all-neighbours' sensors in the field. One of the possible ways to compute this is to compare how much the IoT sensor values (of temperature or humidity in our case) are different from the mean of the variables (we are looking at) versus how much neighbours' sensors values are differed from the mean.

We clarify that the Global *Moran's I* compares the similarity of every object (data value/variable) with its neighbours. It then calculates the average of all these comparisons to show a global view about the spatial pattern of the variable. In some scenarios it is worth to investigate that which objects are similar or different to the objects in their neighbourhood. In such scenarios the Local *Moran's I* is very useful. In our work we consider global *Moran's I* version to determine the spatial correlation over an entire IoT field network which is computed as below

$$I = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \sum_{i=1}^{N} (x_i - \bar{x})^2}. \quad (1)$$

Here, N is the number of measuring sensors units and x_i is the variable of interest calculated by sensor s_i. $\bar{x}$ is the mean of feature x. In a geographical IoT field network it is important to define the neighbourhoods' IoT sensors. There are many standards being used by researchers which impact the analysis. We have used weighting scheme to define

neighbourhood IoT sensors as this is most suitable for our analysis and is most preferred by researchers. Using this w_{ij} weights approach we have defined neighbourhoods' IoT sensors. Using this approach if s_i and s_j are neighbor, then $w_{ij} = 1$, else, $w_{ij} = 0$. Besides, we have obtained the weights w_{ij} using d_{ij}. Here, d_{ij} is the physical distance between s_i and s_j. In other words, w_{ij} is the weight that determines the relationship between s_i and s_j and is calculated as below

$$w_{ij} = \frac{1}{d_{ij}^2}. \tag{2}$$

Now to better comprehend the above equation (1), we see from the numerator that it calculates the difference between x_i (reference sensor of interest) and the mean. This is multiplied by the difference between neighbours' sensors x_j and the mean. These two are multiplied by the weight w_{ij} given to the neighbours. The denominator is standardized our values. Generally, relatively high values of I and relatively low negative values of I indicates that the dataset is positively spatially correlated (or clustered), and negatively spatially autocorrelated, respectively. Mathematically, I near to 0 indicates no spatial correlation. And absolute value of I is higher than 0 shows high spatial correlation.

To get an ideal value of I which represents a random distribution, we compute the expected value of I notated as $E(I)$. This is known as Expected *Moran's I* which is computed as $E(I) = \frac{-1}{N-1}$ [45], [46]. For large data sets, example N approaches infinity, the $E(I)$ approaches zero. Usually values of $E(I)$ range from -1 to +1. Here, we emphasize that our aim is to know whether *Moran's I* is considerably different than random to observe the thresholds values of I on which the I is positively spatially correlated (or clustered). For this we compute $Z(I)$ which is a standard score to determine that if observed value or data point is above or below the mean value. The standardized $Z(I)$ is calculated using the following equation

$$Z(I) = \frac{I - E(I)}{\sqrt{E(I^2) - E^2(I)}}. \tag{3}$$

In the above equation, $\sqrt{E(I^2) - E^2(I)}$ is the variance ($V(I)$) of values in the dataset. Consider an example that in case $|Z| > 1.96$, we can say that the dataset is correlated with a degree of confidence is 95%. The global *Moran's I* index provides the global statistics of spatial correlation for entire dataset. The closer the values mean things are closely related whereas further the values show things are further away. Means Global *Moran's I* compares the behaviour (or similarity) of a sensor to its neighbours' sensors and then computes the average of this to illustrate an overall network behaviour to show the spatial pattern of the variable.

Now for deeper investigation to accurately determine which IoT sensors are similar or different to the other neighbourhood IoT sensors, the Local *Moran's I* is used eventually to determine the degree of correlation between neighbour sensors in the geographical IoT field network. This is similar to the global version, but the key difference is that each sensor location receives its own I, $Z(I)$, $E(I)$, and $V(I)$ value. The local *Moran's I* statistic for a particular sensor s_i is calculated using the formula given below

$$I_i = \frac{(x_i - \bar{x})}{\delta_i^2} \sum_{j=1, j\neq i}^{N} w_{ij}(x_j - \bar{x}), \quad where \tag{4}$$

$$\delta_i^2 = \frac{\sum_{j=1}^{N} w_{ij}}{N-1} - \bar{x}^2. \tag{5}$$

We note in the above formulas that the difference between the value of x_i and the mean is divided by the standard deviation (δ_i^2) of x_i. Fundamentally, the δ_i^2 determines that how much the spatial correlation values at neighbour IoT sensors vary from the mean.

Assume that the outlier (for sensor s_i) is detected by the phase 2 of Algorithm 1. Now to find a correlation (degree of spatial similarity) between s_i and its neighbors the local *Moran's I* is used. The value of *Moran's I* near to 0 will classify the data/IoT sensors as a legitimate sensor failure case. This indicates that the sensor s_i is faulty (and other IoT sensors operating at normal behaviour) as no correlation between the data values of s_i and its neighbors is determined. Now, if the analyses of s_i and its neighbors data streams observed spatially correlated, then it could be a case of security attacks on sensors which usually affect many neighbor IoT sensors.

So far, we have discussed our approach, motivational use cases for using machine learning models, and a mathematical model. Now in order to validate our approach we have conducted extensive experiments. A prototype implementation and results obtained are discussed in the following section which is divided into three sub-sections., i.e., test settings, our results, and comparison. To validate this approach, we have used three machine learning models as we have already mentioned before. We have used the accuracy and F-Score metrics as they are widely used for performance evaluation of classification models. The accuracy metric is used to evaluate the overall performance of the model while F-Score combines the precision and recall into a single metric that reflects the properties of both metrics.

5 Performance Evaluation

We have divided this section into three parts. Firstly we have provided our test settings following which we have discussed the experiments result. We have provided a third sub-section shows a comparison of our approach with existing works also.

5.1 Test Settings

In order to evaluate the performance of our proposal in real-world situations we have developed a prototype implementation of this approach. The implementation has been written in Python 3.7.4. Eclipse Paho MQTT Python library [47] has been used to set up an IoT field network. We have created a main client object using a simple Python script that enable applications to connect to an MQTT broker to publish messages, and to subscribe to topics and receive published messages as well. Forest Fire real dataset has been used in our work to evaluate the performance of our proposed solution [48]. It is publicly available on the Machine learning Repository of the University of California [48]. This dataset contains meteorological features such as Temperature and Relative Humidity as briefly has been given in Table 1. This

TABLE 1
The Forest Fire Dataset Details

Feature	Interpretation	Range	Mean value	Standard Deviation
X	x-Axis spatial coordinate of sensor locations	1–9	4.6	2.31
Y	y-Axis spatial coordinate of sensor locations	1–9	4.3	1.23
Month	Month	Jan–Dec	NA	NA
Day	Day of the week	Mon–Sun	NA	NA
Temp	Temperature (C)	2.2–33.3	18.89	5.81
RH	Relative Humidity (%)	15–100	44.29	16.32
Wind	Wind speed (km/h)	0.4–9.4	4.02	1.8
Rain	Rain (mm/m^2)	0–6.4	0.02	0.3

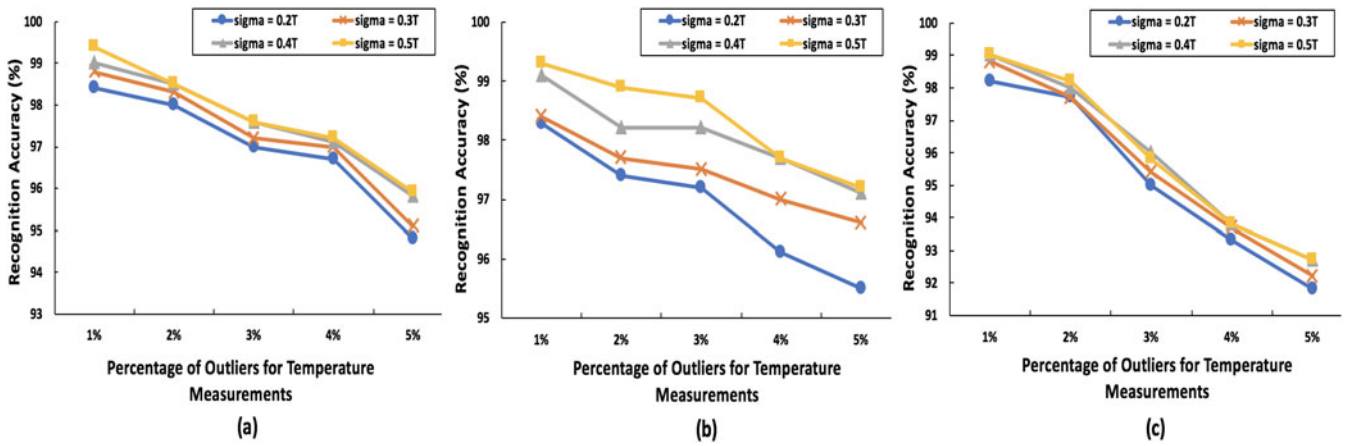


Fig. 2. Recognition accuracy of (a) Classification and regression trees, (b) Random forest, and (c) Support vector machine models for temperature outliers.

also includes spatial data x and y coordinates. The data was collected from January 2000 to December 2003 in the north-east region of Portugal.

We have leveraged Machine Learning algorithms (CART, RF, and SVM) for outlier detection and sensor failure/security attack recognition. In order to model outliers, we have intentionally added random Gaussian noise to the real sensor data values in the dataset. As we have discussed before in case of a sensor failure that this is very unlikely that the other sensors in its neighbourhood too would generate faulty measurements. However, in case of a security attack, several neighbouring sensors are affected by the attacker and as a result, faulty measurements could be received from several neighbouring sensors. In this case, the spatial correlation between the measurements is preserved. Thus, it is justified that to model the sensor failure scenario we have added artificial noise to the measurements of sensors those are located far from each other (have no spatial correlation along). To model the security attack scenario, we have randomly selected several IoT neighbours' sensors and then have added Gaussian noise to the data values of those sensors. For this reason, we changed the number of outliers (from the neighbour sensors) from 1% to 5% of total number of sensors, to observe its impact on the accuracy of the proposed approach. In other words, we changed the area that is targeted by the attacker from 1% to 5% of the whole field network area. Now, to further determine the intensity of the outliers, we have varied the Gaussian noise value. The level of the noise has been changed/varied by considering its standard deviation (*sigma*) as a percentage of the value of each measurement. Therefore, we have varied this value from $0.2T$ to $0.5T$ for both temperature and humidity outliers. Naturally as lower the level of noise would be the more difficult it would be to detect relevant outliers.

5.2 Results

We have shown the recognition accuracy of CART, RF, and SVM models in Figs. 2 and 3 for temperature and humidity outliers, respectively. It can be seen from these figures that having many outliers results in the lower recognition accuracy. Here, we note that the dataset is more skewed due to which outliers may create a separate legitimate cluster that convinces the classifier algorithms to consider them as inliers rather than outliers. However still the classifier considers them as outliers when there are only a few numbers of data point in such a cluster. Regarding the intensity of outliers, we say that higher level of the added noise results in higher accuracy, as we have had expected. Thus, assuming we have a fixed number of outliers in order to have more accurate results, when sigma increases. We have observed a similar behaviour for all the three algorithms. At lower number of outliers (1%) we have noticed that all three algorithms have given approximately similar performance with an accuracy which is in the range of 98.2% to 99.5% for different values of sigma. However, at larger number of outliers (i.e., 4% and 5%) the RF algorithm have performed better than CART and SVM with 96.1% and 95.1% accuracy, respectively. On the other hand, the RF classifier is more sensitive to the intensity of outliers than CART and SVM.

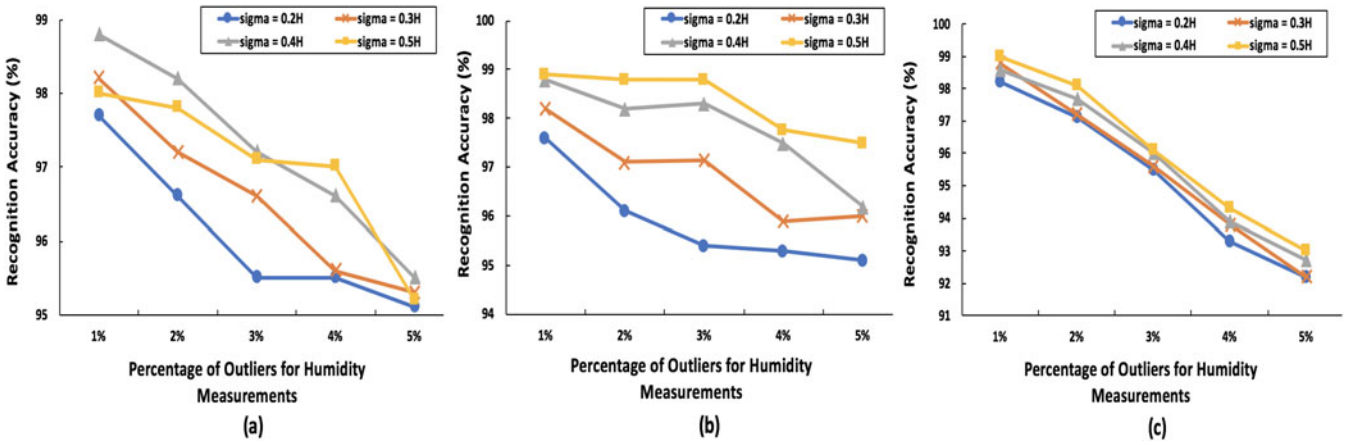


Fig. 3. Recognition accuracy of (a) Classification and regression trees, (b) Random forest, and (c) Support vector machine models for humidity outliers.

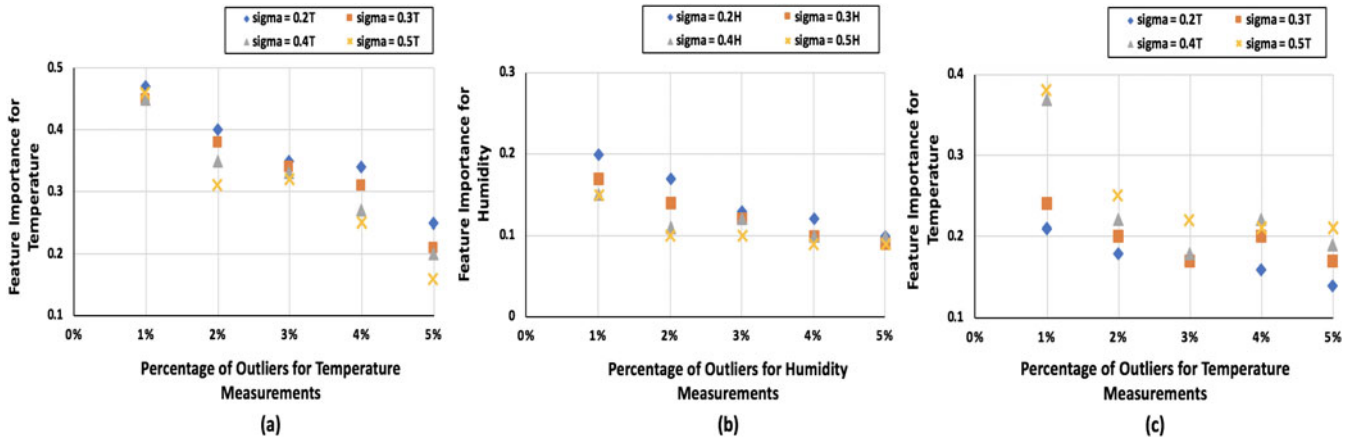


Fig. 4. Feature importance of the CART and RF classifiers. (a) Temperature in CART, (b) Humidity in CART, and (c) Temperature in RF.

In Fig. 4 we have shown the feature importance (for temperature and humidity features) in CART and RF classifiers. The metric of feature importance allocates a score to all the features of a dataset based on their usefulness in the process of target prediction, i.e., it shows how helpful a feature is in performing model predictions. In fact, feature importance metric highlights the top feature of the dataset in terms of the contribution to the predictions of a model. This is very helpful when we want to apply dimensionality reduction and feature selection processes to the dataset which results in the improvement of model efficiency in terms of different factors such as CPU and memory usage, accuracy, latency, etc. In our experiments, we obtain the importance of each feature to highlight the impact of number of outliers on the usefulness of dataset features in the classification process. In this regard, any reduction in the importance of a specific feature (e.g., temperature) indicates that the model has distracted from taking the most important feature into account (i.e. temperature).

We have not shown SVM feature importance, as it is not available for linear kernels. We have noticed in this figure that for larger number of outliers the importance of temperature and humidity feature reduces. We understand that this is expected because a large number of outliers may result in creating a specific cluster which is considered as inlier rather than outlier. This has reduced the importance of that feature in order to decide the type of outliers. But we have also observed that the temperature and humidity features remain the most important features, in case of temperature and humidity outliers, respectively.

To better comprehend this result, we have shown the decision tree of the CART algorithm for temperature outliers in Fig. 5. As seen in the figure that the first decision has been made by checking the *temp* feature of each record of the dataset (in the root node). Since the temperature measurements are all in a specific range (when CART algorithm is trained), then the measurements greater than 33.561 are decided as outlier. This clearly has cross-validated or confirms our results which show the temperature is the most important feature in this case. Otherwise, in the next level, the relative humidity (*RH*) feature is checked. As we can see that the x and y coordinates are also examined since spatial correlation can help the algorithm to perform more accurate classifications. This procedure continues until the leaf nodes of the tree are reached that cover all the possible situations. From extensive evaluation we have noticed a clear consistency in the results we have shown and discussed. To build more confidence in the proposal we have analysed it using three different algorithms also and real data sets. The results are consistent which justify that the proposed idea has a merit and needs further rigorous testing.

Now, as previously said we proposed that the module resides at the edge of the field network, we have designed a

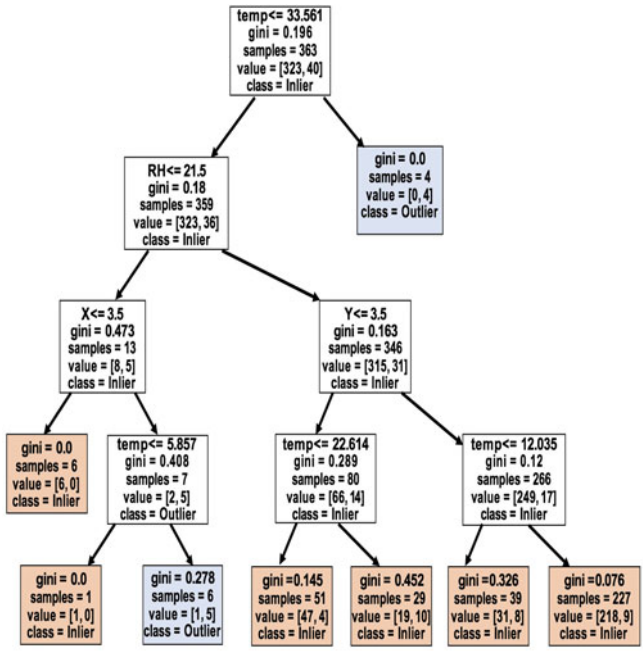


Fig. 5. Decision tree of the CART algorithm for temperature outliers.

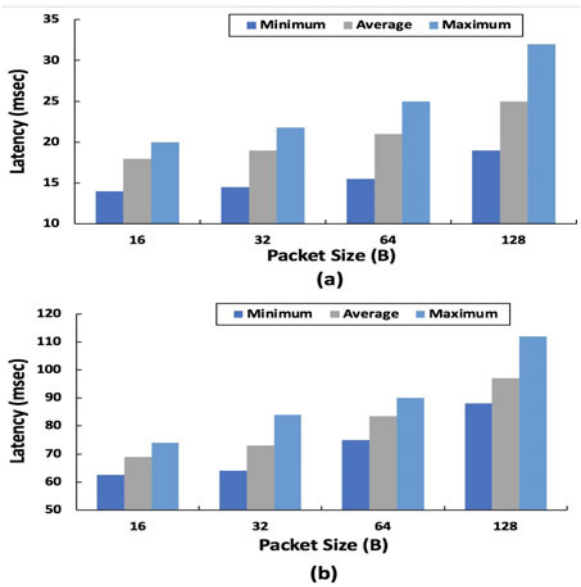


Fig. 7. Calculated latency between the IoT field sensor and (a) Edge server and (b) Cloud server.

prototype implementation to validate this proposal. A small IoT field network is deployed, and the sensor streams are analysed at the edge and at the cloud server to compare the performance of the proposal in terms of latency. The high-level view of the edge IoT network architecture implemented in this work is shown in Fig. 6. As seen in this Figure, the edge server resides near to the IoT network on which the outlier detection module runs. The same anomaly detection process runs at a cloud server located 280*Kms* away from the IoT field network. To securely provide network connectivity between the field network and cloud server we have used ZeroTier [49] client application. To calculate the latency between the servers and field network we have used PRTG Network Monitor software [50]. Other than the latency it can monitor various network and system conditions such as optimal use of bandwidth used by the network devices, routers, etc.

Using the above structural implementation of the prototype we have calculated the latency between the IoT field sensors/devices and servers. It can be seen from Figs. 7a and 7b that our anomaly detection scheme detects the outliers significantly faster if the proposed scheme resides at the edge of the network rather than to place it at the cloud server. This is evident from the results that the average latency is 23*msec* and 87*msec* obtained at edge server and at cloud server, respectively. This value is observed at the packet size of 64B. The IoT field network measurements are sent to cloud server for processing which introduce significant delay and hence is not effective to distinguish IoT network behaviour in real-time. Although this issue can be alleviated by deployed multiple cloud server, but this is not effective from resource and cost optimization perspective. Further, for large size IoT filed networks we may deploy more edge servers, but it does have this same issue also. Besides, the real-time anomaly detection time of the approach can be further enhanced by monitoring the network logs and events by a centralized sever(at the core segment) to obtain a global view of the network behaviour.

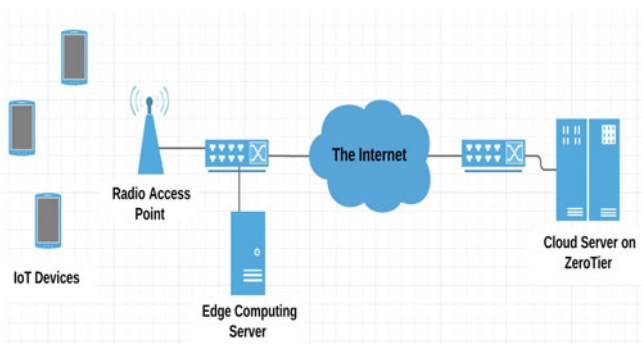


Fig. 6. The Edge IoT network architecture of our prototype implementation.

5.3 Comparison

To validate the effectiveness of the approach, we have compared it with the recent three works [32], [51], [52], the comparison result is given in Table 2. Accuracy and F-Score are the key metrics we have used for performance evaluation and comparison. The results have been taken at four different values of sigma when the percentage of outliers is 1%. Firstly, we have provided the work given in [32], [51], [52] to help reader to comprehend these schemes we have used as benchmarks. Authors of [32] have used CART, RF, Gradient Boosting Machine (GBM), and Linear Discriminant Analysis (LDA) machine learning models to detect the anomalies in a forest environment. Note that Table 2 only shows the best accuracy achieved in [32]. We note that the scheme has obtained 78% accuracy using the LDA model. Their models have failed to obtain the accuracies greater than 97.7% and 99.3% for temperature and humidity, respectively. Authors have not provided enough details related to the standard deviation (sigma) of the artificial noise they have added in dataset. We emphasize that the accuracy will higher at higher values of sigma which we

TABLE 2
Comparison of the Proposed Scheme With the Recent Similar Research Works

		Our Approach			[33]			[52]		[53]	
		CART	RF	SVM	CART	RF	GBM	EONF with ISSNIP	EONF with IBRL	RF	SVM
Temperature	Accuracy (%)	98.9	98.8	98.8	97.7	94.7	97.7	98.5	98.7	95	71
	F–Score (%)	98.2	98.2	98.4	95	89	95	Not Available	Not Available	94	70
Humidity	Accuracy (%)	98.2	98.4	98.7	99.3	99.3	99.3	98.5	98.7	95	71
	F–Score (%)	97.8	98	98.3	99	99	99	Not Available	Not Available	94	70

have validated and shown in Figs. 2 and 3. This is because in this case, the outliers are located in a farther distance to the normal data points than when a lower level of sigma is applied. Thus, it is more likely that the learning model identify them as outliers.

Further, to detect outliers', authors have used the Ellipsoidal Neighborhood Outlier Factor (ENOF) mechanism [53] in [51]. This is an anomaly scoring method which provides a score for each hyperellipsoidal model w.r.t the densities of its close neighborhood sensors. The anomaly score value is small for the data points clustered/belongs in a dense group of data points. The vice-versa is also true. We note that the model is required to obtain a threshold value using the standard deviation of the ENOF scores which is then used to accurately make decision for outlier detection. Authors have used ISSNIP [54] and IBRL [55] datasets which includes both temperature and humidity features. We have observed that their approach has high level of accuracy, however the F-Score metric is not calculated, thus it still outperformed by our approach. Further, the huge amount of information is exchanged in the scheme which makes vulnerable to many security and privacy threats.

A hybrid anomaly detection to detect anomalies in both application and network layers is investigated by authors in [52]. Using cloud computing technology, they train a neural network in a centralized manner to detect anomalies at local manner eventually to minimize the communication overhead and detection latency. Although there are certain advantages of the approach in terms of latency and communication overhead, but from our experiments we have shown that our approach still has higher accuracy. To our understanding the reason for this is the deviation between the predictions made by the server and the ones made locally at the IoT devices in the filed network. Moreover, frequent updates at local level are required from centralised cloud sever which may increase the latency and overhead at large scale IoT field networks.

Moreover, the learning–based attack detection mechanism proposed in [35] works based on the extraction of network traffic features (e.g., source and destination addresses). Although the authors have achieved detection accuracy of up to 99% for decision tree, RF, and Artificial Neural Network (ANN) classifiers, their work relies on the real–time extraction of features from network traffic and performing feature engineering tasks to make the features compatible with the input format of the deployed machine learning models (e.g., converting categorical features to their equivalent numeric vectors). In our approach however, there is no need to perform any feature extraction and engineering tasks since it works based on the received sensor measurements. In other words, the real–time sensor data are applied to the machine learning model as numeric features for the detection of anomalies. This brings several advantages including efficiency in terms of latency and processing power. Moreover, our solution can be implemented at the cloud server since it does not need to monitor the real–time traffic of the IoT network.

In addition, the other interesting work in this field done by Pacheco *et al.* [36] (which proposes an Intrusion Detection System (IDS) for IoT networks) is based on Anomaly Behavior Analysis (ABA) and includes the use of sensor-DNA profiles (s–DNA). In fact, a s–DNA data structure is first built (using the Discrete Wavelet Transform (DWT) approach) which accurately characterizes normal behavior of the sensors. Then, by continuous real–time monitoring of sensors' traffic, the runtime behavior of each sensor is obtained which is compared with the normal profile of the sensor for the detection of any abnormal behavior. The authors have shown that their s–DNA data structure can be deployed in IoT networks as an authentication mechanism for the sensors. They have achieved significant detection accuracy for both known and unknown attacks (98% and 97.4%, respectively) and with false positive rates as low as 0.5%. However, the authors have indicated that their approach is intended to protect those IoT networks with a limited number of sensors. For large scale scenarios, their approach should be tested in upper layers (e.g., service layer) and the outcome might be affected by other factors such as behavorial drift [36]. Moreover, their proposed approach needs to perform additional real–time processing tasks such as (1) continues monitoring of the IoT network traffic (at the end nodes layer), (2) real–time computation of DWT coefficients, and (3) discovery of the sensor type. This may result in significant practical issues since these tasks should be done at the end node layer at which there is usually a lack of enough computation, storage, and power resources. Contrary, our approach is simple and feasible.

The comparison of our approach with the existing recent three approaches indicates that the proposed approach in this paper is promising. We emphasize that the shown approach is simple, reasonable, feasible and practically possible. However, as said, to gain a deeper understanding and trust in this approach, still there are certain aspects need to be rigorously explored which we have discussed in the next section.

6 Further Discussion

In this study we have presented an early and timely investigation into the aspects surrounding outlier detection in farming context. The proposed solution can be deployed to secure every IoT service/platform in which similar-type sensors are located in a fixed position against each other. This is because in our analysis, we utilize the special correlation of sensor measurements to distinguish between a faulty sensor case and a security attack. For example, in environmental monitoring applications (e.g., meteorological systems, air pollution monitoring, bushfire detection, etc.), different measurements sent by sensors (located in a fixed distance from each other) are usually correlated. Moreover, there are industrial applications in which an array of similar-type sensors is deployed to monitor a specific parameter (e.g., in oil and gas pipelines, smart grid systems, etc.).

We are planing to further explore the issues, in different scenarios [56], [57], associated to this work for possible improvement of this study. We have outlined our concerns with this work below and we welcome research community to contribute knowledge on these. Firstly in our work we have used the static data set using which we have shown a merit of this work. However to get more insights the proposal needs to be tested on real-time IoT sensing field using real time scenarios. For real time quick detection it is essential to deploy the computing resources near to the farm in order to reduce latency which eventually effects the real-time performance of the proposal. Therefore, in our understanding the edge computing or federated learning technologies will help to address this issue. Although the edge computing is employed in this work, but to determine the computing and storage resources at the edge of the IoT sensor networks dynamically is challenge when the nodes are mobile in order to provide seamless connectivity between sensors and computing resources. Therefore we must consider the mobility of nodes into account as well as the dynamic deployment of edge servers instead of fixed or static deployed (as used traditionally).

Secondly, the current study lacks the evaluation of proposed approach using theoretical model which considers the heterogeneity of multi-domain networks and sensors data into account. We understand the importance of modelling also which help us to understand the approach in a much better way. Particularly this helps in the design and implementation of multi-vendor vertical heterogeneous IoT sensing architectures. In this context there are some theoretical models exists and we have provided a mathematical modelling in Section 4, but in this the main challenge we observe is the heterogeneity of data which will play critical role and to address this in modelling will be a great challenge. In our knowledge, the modeling in this context is not reported.

Finally, we have validated that the outliers originated from the same area would be classified into the security attack class and into the sensor failure class. This IoT system behavior would be considered as attack behavior. In contrary, we do accept that the assumptions or test setting in this work may not be ideal for other IoT domains. Example co-location may not be the criteria for all IoT domains as this is very specific to farming scenarios which are dependent on environmental conditions. In future, we plan to ease these assumptions so that the proposal can be made applicable out of this research domain and research settings.

Other than the technical challenges there are some other hurdles; questions that might be raised here are: a) how can we protect deployed sensors in the landscape from fire or from interference by wildlife?, b) if deploying IoT sensors in a small field/zone, what factors are needed to be considered when determining the best place to deploy them, i.e., are we basing this on landscape features that create high risk of interference? From an ecologists point of view how we would protect sensors from interference from wildlife, the general public, natural disturbances is a key concerns? Also addition of other features in data sets are important angle to evaluate.

We fully acknowledge that the presented research work in this paper can only be considered as samples, not statistical guarantees. The proposed solution has its own benefits but also have few limitations. Therefore, further evaluation of this early work is needed. We reiterate that the conventional methods are not applicable to today's IoT network scenarios, researchers need to explore innovative, novel, simple and feasible solutions to address this issue we have discussed in our work. We plan to address these limitations in future.

7 Conclusion and Future Work

In smart farming domain we have discussed a feasible approach to distinguish IoT sensing behaviour. We have used Forest Fire real data set. Our approach is based on spatial correlation theory and we have validated the approach using CART, RF, and SVM algorithms. Our approach successfully detects condition/events such as detecting IoT sensors normal or legitimate behaviour, faulty, and malicious behaviour. This can a) early alarm the risk of cyber attacks, b) timely detection of anomaly connections, and c) helps to avoid any unplanned maintenance, eventually save costs. In future, we will investigate the issues using real IoT field network we have covered in further discussion section.

Acknowledgments

The authors would like to thank the Editor and anonymous reviewers for the insightful and constructive comments and suggestions which have greatly improved the quality of this work.

References

[1] L. D. Xu, W. He, and S. Li, "Internet of Things in industries: A survey," *IEEE Trans. Ind. Inform.*, vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

[2] Y. Liu, X. Ma, L. Shu, G. P. Hancke, and A. M. Abu-Mahfouz, "From industry 4.0 to agriculture 4.0: Current status, enabling technologies, and research challenges," *IEEE Trans. Ind. Inform.*, vol. 17, no. 6, pp. 4322–4334, Jun. 2021.

[3] C. Brewster, I. Roussaki, N. Kalatzis, K. Doolin, and K. Ellis, "IoT in agriculture: Designing a europe-wide large-scale pilot," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 26–33, Sep. 2017.

[4] A. Roy, P. Das, and R. Das, "Temperature and humidity monitoring system for storage rooms of industries," in *Proc. Int. Conf. Comput. Commun. Technol. Smart Nation*, 2017, pp. 99–103.

[5] Internet of Things. Accessed: Oct. 12, 2020. [Online]. Available: https://www.agrifutures.com.au/wp-content/uploads/publications/16–039.pdf
[6] Agrifutures australia annual report 2019–2020. Accessed: Oct. 12, 2020. [Online]. Available: https://www.agrifutures.com.au/product/agrifutures-australia-annual-report-2019–2020/
[7] K. Sood, S. Yu, and Y. Xiang, "Software-defined wireless networking opportunities and challenges for Internet-of-Things: A review," *IEEE Internet Things J.*, vol. 3, no. 4, pp. 453–463, Aug. 2016.
[8] A. Gaddam, T. Wilkin, and M. Angelova, "Anomaly detection models for detecting sensor faults and outliers in the IoT - a survey," in *Proc. 13th Int. Conf. Sens. Technol.*, 2019, pp. 1–6.
[9] S. S. L. Chukkapalli, A. Piplai, S. Mittal, M. Gupta, and A. Joshi, "A smart-farming ontology for attribute based access control," in *Proc. 6th IEEE Int. Conf. Big Data Secur. Cloud*, 2020, pp. 29–34.
[10] S. Sontowski *et al.*, "Cyber Attacks on Smart Farming Infrastructure," in *Proc. 6th Int. Conf. Collaboration Internet Comput.*, 2020, pp. 135–143.
[11] M. Gupta, M. Abdelsalam, S. Khorsandroo, and S. Mittal, "Security and privacy in smart farming: Challenges and opportunities," *IEEE Access*, vol. 8, pp. 34564–34584, 2020.
[12] J. Paull, "Organic agriculture in australia: Attaining the global majority (51%)," *Org. Agriculture Aust. Attaining Glob. Majority*, vol. 5, no. 2, pp. 70–74, 2019.
[13] M. Window, "Security in precision agriculture: Vulnerabilities and risks of agricultural systems," 2019. [Online]. Available: http://ltu.diva-portal.org/smash/get/diva2:1322203/FULLTEXT02.pdf
[14] A. B. Sharma, L. Golubchik, and R. Govindan, "Sensor faults: Detection methods and prevalence in real-world datasets," *ACM Trans. Sensor Netw.*, vol. 6, no. 3, pp. 1–39, 2010.
[15] J. Ye, G. Stevenson, and S. Dobson, "Detecting abnormal events on binary sensors in smart home environments," *Pervasive Mobile Comput.*, vol. 33, pp. 32–49, 2016.
[16] A. Gaddam, T. Wilkin, M. Angelova, and J. Gaddam, "Detecting sensor faults, anomalies and outliers in the Internet of Things: A survey on the challenges and solutions," *Electronics*, vol. 9, no. 3, p. 511, 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/3/511
[17] L. Yang, C. Ding, M. Wu, and K. Wang, "Robust detection of false data injection attacks for data aggregation in an Internet of Things-based environmental surveillance," *Comput. Netw.*, vol. 129, pp. 410–428, 2017.
[18] S. Kumar and V. K. Chaurasiya, "A strategy for elimination of data redundancy in Internet of Things (IoT) based wireless sensor network (WSN)," *IEEE Syst. J.*, vol. 13, no. 2, pp. 1650–1657, Jun. 2019.
[19] K. Sood, K. K. Karmakar, V. Varadharajan, U. Tupakula, and S. Yu, "Analysis of policy-based security management system in software-defined networks," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 612–615, Apr. 2019.
[20] A. H. Sarker, J. F. Bornman, and D. Marinova, "A framework for integrating agriculture in urban sustainability in australia," *Urban Sci.*, vol. 3, no. 2, p. 50, 2019. [Online]. Available: https://www.mdpi.com/2413-8851/3/2/50
[21] J. Chen and A. Yang, "Intelligent agriculture and its key technologies based on Internet of Things architecture," *IEEE Access*, vol. 7, pp. 77134–77141, 2019.
[22] N. Lu, N. Cheng, N. Zhang, X. Shen, and J. W. Mark, "Connected vehicles: Solutions and challenges," *IEEE Internet Things J.*, vol. 1, no. 4, pp. 289–299, Aug. 2014.
[23] P. Eugster, V. Sundaram, and X. Zhang, "Debugging the Internet of Things: The case of wireless sensor networks," *IEEE Softw.*, vol. 32, no. 1, pp. 38–49, Jan./Feb. 2015.
[24] J. A. Manrique, J. S. Rueda-Rueda, and J. M. T. Portocarrero, "Contrasting Internet of Things and wireless sensor network from a conceptual overview," in *Proc. IEEE Int. Conf. Internet Things IEEE Green Comput. Commun. IEEE Cyber, Phys. Social Comput. IEEE Smart Data*, 2016, pp. 252–257.
[25] F. Bu and X. Wang, "A smart agriculture IoT system based on deep reinforcement learning," *Future Gener. Comput. Syst.*, vol. 99, pp. 500–507, 2019.
[26] A. Somov *et al.*, "Pervasive agriculture: IoT-enabled greenhouse for plant growth control," *IEEE Pervasive Comput.*, vol. 17, no. 4, pp. 65–75, Oct.–Dec. 2018.
[27] I. Mohanraj, K. Ashokumar, and J. Naren, "Field monitoring and automation using IoT in agriculture domain," *Proc. Comput. Sci.*, vol. 93, pp. 931–939, 2016.
[28] S. Mahfuz, H. Isah, F. Zulkernine, and P. Nicholls, "Detecting irregular patterns in IoT streaming data for fall detection," in *Proc. IEEE 9th Annu. Inf. Technol., Electron. Mobile Commun. Conf.*, 2018, pp. 588–594.
[29] E. Siow, T. Tiropanis, and W. Hall, "Analytics for the Internet of Things: A survey," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–36, 2018.
[30] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, 2017.
[31] K. Zhang, K. Yang, S. Li, D. Jing, and H.-B. Chen, "Ann-based outlier detection for wireless sensor networks in smart buildings," *IEEE Access*, vol. 7, pp. 95 987–95 997, 2019.
[32] N. Nesa, T. Ghosh, and I. Banerjee, "Outlier detection in sensed data using statistical learning models for IoT," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2018, pp. 1–6.
[33] H. Fanaee-T and J. Gama, "Event detection from traffic tensors: A hybrid model," *Neurocomputing*, vol. 203, pp. 22–33, 2016.
[34] X. Deng, P. Jiang, X. Peng, and C. Mi, "An intelligent outlier detection method with one class support tucker machine and genetic algorithm toward big sensor data in Internet of Things," *IEEE Trans. Ind. Electron.*, vol. 66, no. 6, pp. 4672–4683, Jun. 2019.
[35] M. Hasan, M. M. Islam, M. I. I. Zarif, and M. Hashem, "Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches," *Internet Things*, vol. 7, 2019, Art. no. 100059.
[36] J. Pacheco and S. Hariri, "Anomaly behavior analysis for IoT sensors," *Trans. Emerg. Telecommun. Technol.*, vol. 29, no. 4, 2018, Art. no. e3188.
[37] Z. Huan, C. Wei, and G.-H. Li, "Outlier detection in wireless sensor networks using model selection-based support vector data descriptions," *Sensors*, vol. 18, no. 12, 2018, Art. no. 4328.
[38] P. Shi, G. Li, Y. Yuan, and L. Kuang, "Outlier detection using improved support vector data description in wireless sensor networks," *Sensors*, vol. 19, no. 21, 2019, Art. no. 4712.
[39] G. Liu, L. Shi, and D. Xin, "Data integrity monitoring method of digital sensors for Internet of Things applications," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4575–4584, May 2020.
[40] S. Garg, K. Kaur, N. Kumar, G. Kaddoum, A. Y. Zomaya, and R. Ranjan, "A hybrid deep learning-based model for anomaly detection in cloud datacenter networks," *IEEE Trans. Netw. Serv. Manage.*, vol. 16, no. 3, pp. 924–935, Sep. 2019.
[41] W. Sha, Y. Zhu, M. Chen, and T. Huang, "Statistical learning for anomaly detection in cloud server systems: A multi-order markov chain framework," *IEEE Trans. Cloud Comput.*, vol. 6, no. 2, pp. 401–413, Apr.–Jun. 2015.
[42] A. Alabdulatif, H. Kumarage, I. Khalil, and X. Yi, "Privacy-preserving anomaly detection in cloud with lightweight homomorphic encryption," *J. Comput. Syst. Sci.*, vol. 90, pp. 28–45, 2017.
[43] L. Girish and S. K. Rao, "Anomaly detection in cloud environment using artificial intelligence techniques," *Computing*, pp. 1–14, 2021. [Online]. Available: https://link.springer.com/article/10.1007%2Fs00607-021-00941-x
[44] X. Zhang *et al.*, "Cross-dataset time series anomaly detection for cloud systems," in *Proc. Annu. Tech. Conf.*, 2019, pp. 1063–1076.
[45] M. Yan *et al.*, "Outliers detection of cultivated land quality grade results based on spatial autocorrelation," in *Proc. 5th Int. Conf. Agro-Geoinformatics*, 2016, pp. 1–5.
[46] M. Das and S. K. Ghosh, "Measuring moran's i in a cost-efficient manner to describe a land-cover change pattern in large-scale remote sensing imagery," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2631–2639, Jun. 2017.
[47] Eclipse pahomqtt python library. Accessed: Oct. 12, 2020. [Online]. Available: http://www.eclipse.org/paho/downloads.php
[48] A. Asuncion and D. J. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLReposit-ory.html
[49] Zerotier: Securly connect any device, anywhere. Accessed: Aug. 18, 2020. [Online]. Available: https://www.zerotier.com/download/
[50] Prtg network monitor. Accessed: Aug. 18, 2020. [Online]. Available: https://www.paessler.com
[51] L. Lyu, J. Jin, S. Rajasegarar, X. He and M. Palaniswami, "Fog-empowered anomaly detection in IoT using hyperellipsoidal clustering," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1174–1184, Oct. 2017.

[52] A. Ayad, A. Zamani, A. Schmeink and G. Dartmann, "Design and Implementation of a Hybrid Anomaly Detection System for IoT," in *Proc. 6th Int. Conf. Internet Things Syst., Manage. Secur.*, 2019, pp. 1–6.
[53] S. Rajasegarar *et al.*, "Ellipsoidal neighbourhood outlier factor for distributed anomaly detection in resource constrained networks," *Pattern Recognit.*, vol. 47, no. 9, pp. 2867–2879, 2014.
[54] V. K. Sachan, S. A. Imam, and M. T. Beg, "Energy-efficient communication methods in wireless sensor networks: A critical review," *Int. J. Comput. Appl.*, vol. 39, no. 17, pp. 35–48, 2012.
[55] IBRLDataset, *City of Melbourne Open Data Homepage*, Sep. 4, 2020. [Online]. Available: https://data.melbourne.vic.gov.au/Environment/Sensor-readings-with-temperature-light-humidity-ev/ez6b-syvw
[56] Case studies. Accessed: Feb. 19, 2020. [Online]. Available: https://www.austrade.gov.au/agriculture40/case-studies
[57] A. Kumari, S. Tanwar, S. Tyagi, N. Kumar, R. M. Parizi, and K.-K. R. Choo, "Fog data analytics: A taxonomy and process model," *J. Netw. Comput. Appl.*, vol. 128, pp. 90–104, 2019.

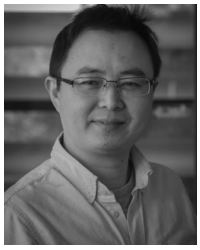
Keshav Sood received the PhD degree from Deakin University, Australia, in 2018. He is currently a lecturer with the Centre for Cyber Security and Innovation, School of IT, Deakin University, Melbourne and the deputy director of bachelor of cyber security course, Deakin University. He was a research fellow with Advanced Cyber Security Engineering Research Centre, The University of Newcastle, NSW, Australia. He was on the project funded by Defence Science and Technology Group, Australia. He was also a mentor in online future learn course on cyber security issues in small and medium scale enterprises and for a short time, with the Terminal Ballistic Research Laboratory (TBRL, DRDO, and Ministry of Defence), Chandigarh, India, in 2006. He is currently a professional engineer with Engineers Australia accreditation. His research interests include quality of service and security in heterogeneous next generation networks.

Mohammad Reza Nosouhi (Member, IEEE) received the master's degree in telecommunications engineering from the Isfahan University of Technology, Isfahan, Iran, and the PhD degree from the University of Technology Sydney, Ultimo, NSW, Australia, in 2020. He is currently a research fellow with the Centre for Cyber Security Research and Innovation Deakin University, Australia. He worked for more than ten years in the ICT industry, Iran.

Neeraj Kumar (Senior Member, IEEE) is currently a full professor with the Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology (Deemed to be University), Patiala, Punjab, India and an adjunct professor with Asia University, Taiwan, King Abdul Aziz University, Jeddah, Saudi Arabia, and Charles Darwin University, Australia. He has authored or coauthored more than 400 technical research papers. His research interests include green computing and network management, IoT, Big data analytics, deep learning, and cyber-security. He has also edited or authored ten books with international or national publishers, imncluding IET, Springer, Elsevier, and CRC. He is currently the editor of *ACM Computing Survey*, *IEEE Transactions on Sustainable Computing*, *IEEE Transactions on Network and Service Management*, *IEEE Network Magazine*, *IEEE Communication Magazine*, *Elsevier Journal of Networks and Computer Applications*, *Elsevier Computer Communication*, *and Wiley International Journal of Communication Systems*. He organized various special issues of journals of repute from IEEE, Elsevier, and Springer. He was the workshop chair at IEEE Globecom 2018, IEEE Infocom 2020 , and IEEE ICC 2020 He was the recipient of Best Paper Award from *IEEE Systems Journal* in 2018, 2020. Elsevier JNCA, IWCMC 2021, and IEEE ICC 2018, Kansas-city in 2018 and the Best Researcher Award from parent organization every year from last eight consecutive years. (2019, 2020, 2021 highly-cited researcher from WoS).

Anuroop Gaddam received the PhD degree in electronics, information communication engineering from Massey University, New Zealand, in 2012. He is currently a lecturer with the School of Information Technology, Deakin University, Australia. His research interests include experience in the analysis of Internet of Things spatial-temporal data, sensor fault or outlier detection, IoT based technologies for health monitoring, health informatics, designing developing smart sensors, embedded systems and wireless sensor networks, smart IoT monitoring system for precision agriculture, environmental monitoring with predictive analysis. He was the recipient of multiple international awards for his research on developing a Wireless Sensor Network based Smart Home for Elder-Care.

Bohao Feng (Member, IEEE) received the BS and PhD degrees from Beijing Jiaotong University in 2011 and 2017, respectively. He is currently an associate professor with the School of Electronic and Information Engineering, Beijing Jiaotong University. His research interests include service function chains, network security, and satellite communications. He was a TPC Member of a number of international conferences, including IEEE ICC and IEEE GLOBECOM. He has participated in several national research programs of China including the 973 Program and 863 Program.



Shui Yu (Senior Member, IEEE) recevied the PhD degree from Deakin University, Australia, in 2004. He is currently a professor with the School of Computer Science, University of Technology Sydney, Australia. He has authored or coauthored three monographs and edited two books, more than 400 technical papers, including top journals and top conferences, including *IEEE Transactions on Parallel and Distributed Systems*, *IEEE Transactions on Computers*, *IEEE Transactions on Information Forensics and Security*, *IEEE Transactions on Mobile Computing*, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Emerging Topics in Computing*, *IEEE/ACM Transactions on Networking*, and *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops*. His research interests include Big Data, security and privacy, networking, and mathematical modeling. His h-index is 58. He initiated the research field of networking for Big Data in 2013, and his research outputs have been widely adopted by industrial systems, such as Amazon cloud security. He is currently on a number of prestigious editorial boards, including the area editor of *IEEE Communications Surveys and Tutorials*, I*EEE Communications Magazine*, and *IEEE Internet of Things Journal*. He is a member of AAAS and ACM, a distinguished lecturer of IEEE Communications Society, and an elected member of Board of Governor of IEEE Vehicular Technology Society.