A General Dataset Generator for Industrial Internet of Things Using Multi-sensor Information Fusion

Mian Wang[1], Jinlong Sun[1,*], Zhiyi Lu[2,*], Yu Wang[3], Yifan Zhang[1], Jie Zhang[1], Zhe Zhang[3], and Guan Gui[1,*]
[1]College of Telecommunications and Information Engineering, NJUPT, Nanjing, China
[2]Nanjing Great Information Technology Co. Ltd, Nanjing, China
[3]Nanjing Run-Time Electric Automation Co. Ltd, Nanjing, China
sunjinlong@njupt.edu.cn, luzhiyi6298@126.com, guiguan@njupt.edu.cn
*corresponding author

***Abstract*—The development of the fifth-generation (5G) communication technology is promoting the overall improvement of the Industrial Internet of Things (IIoT), and is accelerating the pace of the fourth industrial revolution. There lie problems in the IIoT, such as complex industrial field environment and diverse transmission information. It is difficult for us to collect and process massive real-time data for various devices in a timely, reliable, and efficient manner. In order to investigate and mitigate the above problems, we construct a general IIoT Dataset Generator based on a multi-sensor transmission and fusion architecture. The Dataset Generator has advantages such as good privacy, simple development and strong mobility. As an example, we use sound sensors and accelerometers to collect data for industrial scenes. The architecture is also equipped with a mature data preprocessing and visualization scheme. We demonstrate the effectiveness of the IIoT Dataset Generator and the processing algorithms under the considered scenario.**

***Keywords*—dataset Generator; IIoT; data acquisition; multi-sensor; information fusion amplifiers**

I. Introduction

Industrial Internet of Things (IIoT) is the application of Internet of things in industrial field [1]. It collects and processes data in the industrial environment, and realizes intelligent operations such as industrial monitoring, automation and control [2]. IIoT can effectively improve production efficiency and reduce production costs. It is considered to be the basis of future industrial systems. In the process of improving production efficiency, IIoT will produce massive, complex and heterogeneous data, which is called industrial data flow in the industry. Industrial scenes are often complex and diverse, facing the characteristics of difficult collector installation, bad environment, poor signal receiving strength, large amount of data and strong real-time. If the data collected and transmitted by IIoT devices are not effectively stored and managed, the data becomes incomplete, damaged or misleading [3], [4]. In order to solve the above problems, combined with the idea of multi-sensor information fusion, this paper proposes a multi-sensor data processing scheme for IIoT. In the later part of this paper, the proposed scheme is introduced in detail. The Dataset Generator designed in this paper can efficiently and cheaply collect data in the industrial field, and generate complete data-sets to make up for the lack of data-sets in the field of IIoT.

II. RELATED WORK

At present, there are two main schemes for IIoT Dataset Generator by domestic and foreign manufacturers. First, the embedded industrial gateway is used to access the industrial site, or data collection is carried out through serial ports, servers and other media. The other is to use the data transfer unit Data Terminal Unit (DTU) for transparent transmission to the cloud for data collection [5]. The main problems of these two main collection schemes can be summarized as follows. First, IIoT devices have different I/O mechanisms, development protocols, and data formats. The difference between hardware and software and communication is a difficult problem to handle. Second, developers need to redevelop new hardware and software. This will require a lot of cost and effort. Third, the control methods of Dataset Generator are data flow oriented. Data collection relies too much on the gateway, weakening the relative independence of each application. Fourth, privacy policies require purchase of hardware equipment from manufacturers, which is often expensive and strongly bound with the platform of the manufacturer. The voice of users is weakened, and the migration of data is limited [6]. Fifth, traditional method of collecting information from different sensors requires different equipment, which costs more and the operation process is more cumbersome.

Combined with the idea of multi-sensor information fusion, this paper proposes a multi-sensor data processing scheme for IIoT [7]. Compared with the existing data acquisition schemes, the advantages of the scheme proposed in this paper are as follows:

- **Low complexity of development:** The equipment of Dataset Generator is single, and its I/O mechanism, development protocol and data format are unified. It can easily realize the processing of software, hardware and communication.
- **High development reuse rate:** For different industrial sites, we only need to adjust and develop the shell of the Dataset Generator. There is no need to redevelop hardware, communication and software.
- **Good maintainability and mobility:** The size of the Dataset Generator is very small and the shell is highly plastic. Data acquisition no longer depends on the gateway, which enhances the relative independence of each link and is more conducive to software maintenance.
- **Good privacy:** There is no need to purchase the hardware equipment of the manufacturer, and the collected dataset is completely mastered by the user. The user's authority is strengthened, and the storage and transmission of data

2767-6684/22/$31.00 ©2022 IEEE
DOI 10.1109/DSA56465.2022.00034

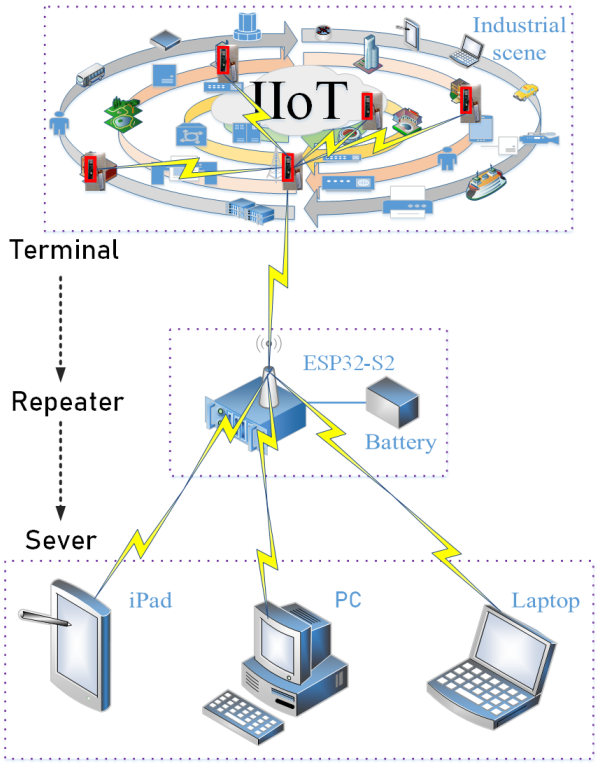

Fig. 1. System model of data acquisition in IIoT based on multi-sensor information fusion

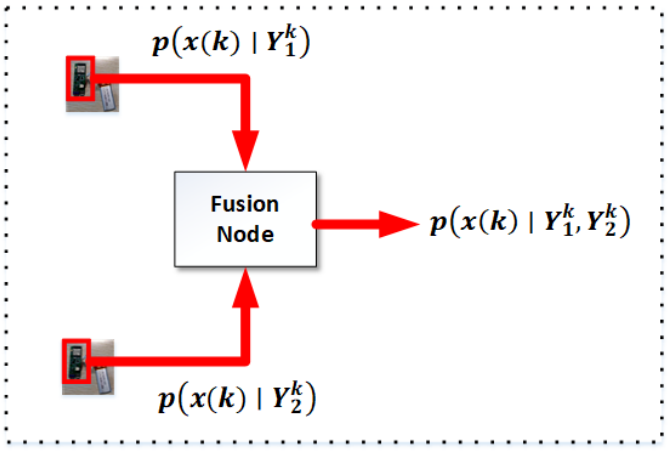

Fig. 2. Bayesian solution module of data fusion.

are not limited.

- **Diversity of datasets:** It can collect multi-sensor information at the same time, and the operation process is relatively simple. The mature synchronization algorithm makes the dataset more valuable.

III. SYSTEN MODEL AND ITS MODULES

The data acquisition model of IIoT is one of the most basic problems in the current research field. As shown in Fig. 1, we design a new system model of data acquisition in IIoT based on multi-sensor information fusion [8]. The model consists of data collector, repeater and server. In our model, the first part is the data acquisition module, which has two sub modules. One is the sensor sub module, and the other is the data fusion sub module.

The sensor sub module is composed of multiple distributed heterogeneous sensors. Multisensor data provides more complete and reliable data, which can be analyzed to obtain more accurate results. In our research, we choose two heterogeneous sensors, sound sensor and acceleration sensor, to collect data. In order to reduce the consumption of channel resources and reduce the cost [9], the signals need to be preprocessed locally on two heterogeneous sensor nodes. The wireless signal $w(k)$ received by the hotspot can be expressed as

$$w(k) = s(k) + e(k), k = 0, 1, \ldots, N-1 \tag{1}$$

where $s(k)$ is a wireless signal, which is composed of amplitude modulation (AM) signal, frequency modulation (FM) signal, binary phase shift keying (BPSK) signal, quadrature phase shift keying (QPSK) signal, 8 phase shift keying (8PSK) signal, 2 phase shift keying (2ASK) signal, 4 phase shift keying (4ASK) signal, 2 phase shift keying (2FSK) signal and 4 frequency shift keying (4FSK) signals [10]. $e(k)$ represents the artificial noise in idwsn, which can be described as alpha stable noise.

The data fusion module in this paper will do a simple data preprocessing for the data of two different sensors. According to different fusion levels, multi-sensor information fusion technology can be divided into three categories: data level fusion, feature level fusion and decision level fusion. The data acquisition model designed in this paper reflects the data layer fusion of multi-sensor information fusion technology. This is a process of collecting information from different sources, detecting and analyzing the collected data signals and completing multi-sensor information fusion. Its main function is to clear invalid sensor data and synchronize different sensor data to obtain more accurate results. Data fusion uses a Bayesian solution, as shown in the figure, which requires that both sensors have available information.

The sum of the probability density function can be expressed as

$$p\left(x(k) \mid Y_1^k, Y_2^k\right) = p\left(y_1(k) \mid x(k)\right) p\left(y_2(k) \mid x(k)\right) \frac{p\left(x(k) \mid Y_1^{k-1}, Y_2^{k-1}\right)}{p\left(y_1(k), y_2(k) \mid Y_1^{k-1}, Y_2^{k-1}\right)} \tag{2}$$

where $p\left(y_1(k) \mid x(k)\right)$ and $p\left(y_2(k) \mid x(k)\right)$ respectively represent the available information of the two sensors. $p\left(x(k) \mid Y_1^k, Y_2^k\right)$ represents the result of data fusion of two sensors at time k . $p\left(y_1(k), y_2(k) \mid Y_1^{k-1}, Y_2^{k-1}\right)$ represents the joint likelihood function of two independent sensors [12].

The most important factor for fusion is to determine the following scores based on the information provided by the sensor

$$\frac{p\left(x(k) \mid Y_1^k\right)}{p\left(x(k) \mid Y_1^{k-1}\right)} \text{ and } \frac{p\left(x(k) \mid Y_2^k\right)}{p\left(x(k) \mid Y_2^{k-1}\right)} \tag{3}$$

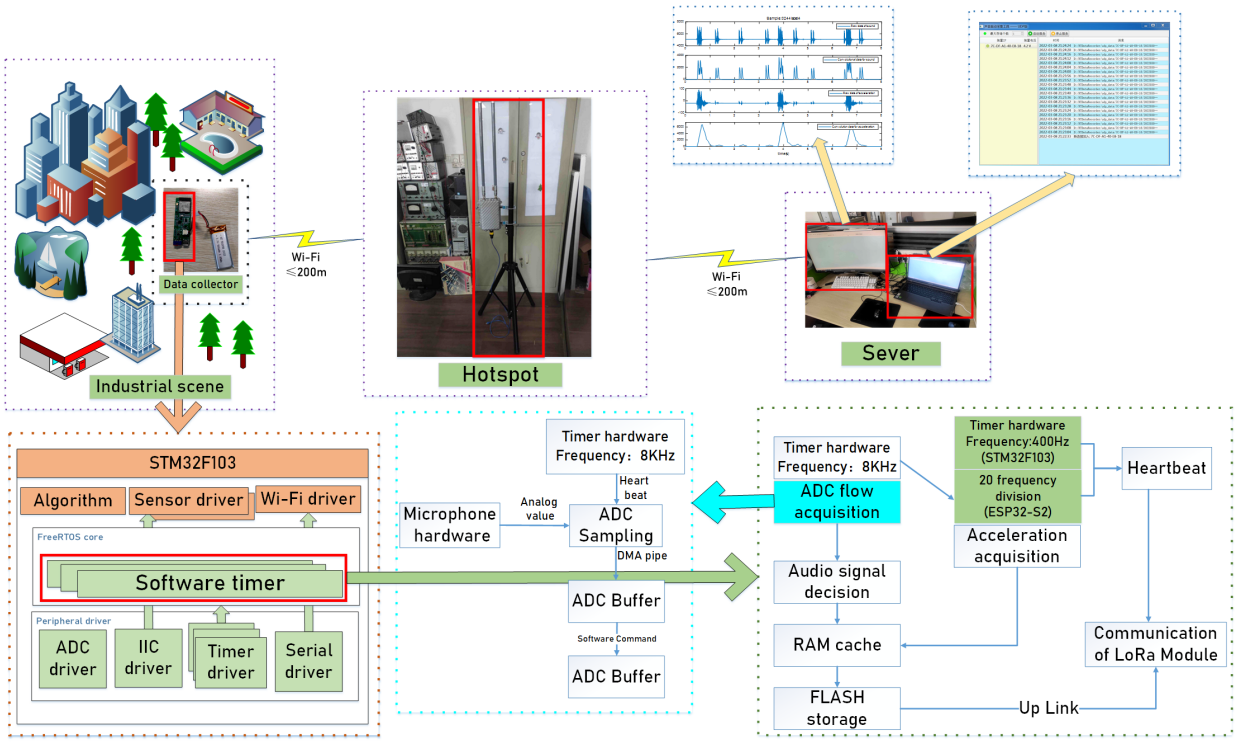

Fig. 3. Schematic diagram of data acquisition in an industrial scenario. The acquisition system uses the technical scheme of Inter-Integrated Circuit (IIC) and esp32-s2, where STM32F103 is the core chip of the data collector.

The advantage of data-layer fusion is large amount of information and high accuracy. We don't need to think about data loss [13]. It's easy to extract a lot of data details. It is important to fuse the outputs of these sensors in an efficient and intelligent way. The rest of this article will study data transmission and storage, as well as data analysis. At present, the data fusion module of this data generator is not mature enough, sometimes it needs manual assistance to confirm the accuracy of the data. Subsequent work will optimize the performance of this module and further develop feature layer fusion.

IV. System Architecture

A. Hardware Architecture

The Dataset Generator in this paper mainly implements the function of data collection in the industrial scene. Its data collection system uses the technical scheme of IIC and ESP32-S2, which has been verified to be technically feasible and can meet the needs of users [14].

In order to allow for data collection in a variety of environments, we also conducted field studies. This data set generator can customize the shell to suit the field situation. For example, data collection for underwater scenes will consider adjusting the shell's waterproof performance, density and other physical properties. Specifically, in order to collect data for large machines, we can upgrade the fixture of the modified housing so that the data set generator can be firmly fixed on the machine without affecting the operation of the machine.

First of all, in terms of hardware selection, we compare the performance of several different types of development board chips. The STM32F103 chip is used as the core chip of the data set generator [15]. The low power consumption, flash memory and built-in controller of STM32F103 perfectly meet the needs of this project. We use a dual timer to collect sound signals at 4 KHz and acceleration signals at 800 Hz. It uses an operational amplifier to convert analog signals to sound and a built-in analog-to-digital converter to collect them. Acceleration signals are collected using a 3D accelerometer and transmitted to the Microcontroller Unit (MCU) using the IIC communication protocol.

Another consideration after data collection is data storage. ESP32-S2 is selected as the repeater for the data storage system. The repeater can be transmitted up to 200 meters away and has industry-leading low power and radio frequency performance. It can transmit data to the platform at medium and long distance, which fully meets the requirements of this Dataset Generator. We use the AD conversion module to convert the collected sound and acceleration analog data into binary data. The converted binary data is then sent to the buffer via the DMA pipeline. Judged and cached by the CPU, then stored in FLASH. Finally, we use the WIFI module to transfer to the repeater ESP32-S2, and then to the server for storage. The overall hardware schematic diagram is shown in Fig. 3.

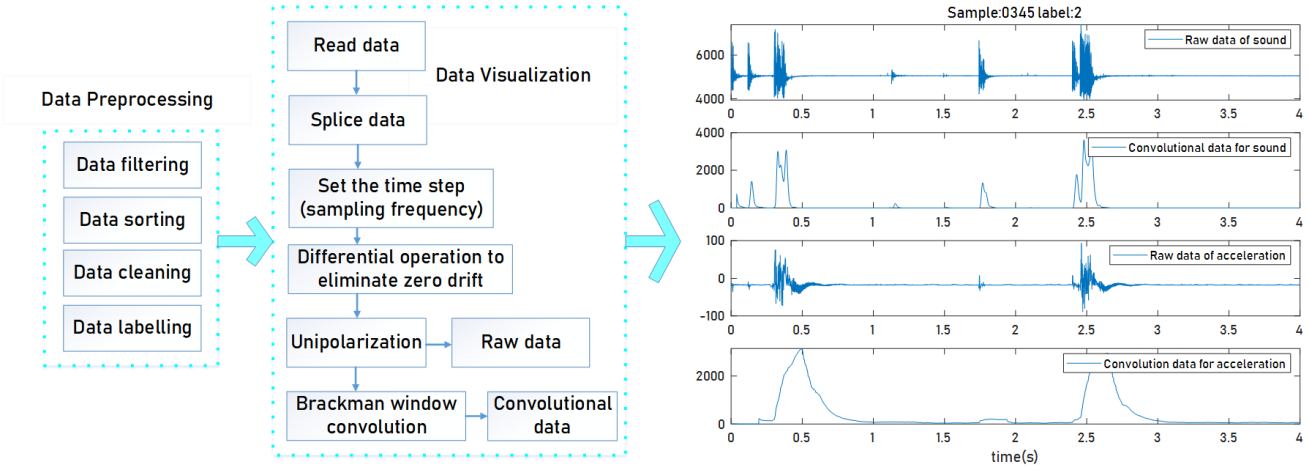

Fig. 4. Flow chart of data preprocessing and visualization, and an example of visualized results is displayed. The waveform of sound sensors and accelerometers can be displayed in real time.

B. Software System

The Dataset Generator model has a mature data preprocessing and visualization system [15]. The process of data preprocessing and visualization is shown in Fig. 4.

The data preprocessing system can perform data cleaning, data filtering, data sorting, data marking and other operations on the collected data sources. This data can then be generated into a common multi-sensor information dataset based on the IIoT scenario.

The pre-processing system will display the binary data collected in a variety of forms intuitively through the MATLAB platform. After the targeted monitoring algorithm is designed, high accuracy monitoring can be achieved for various industrial scenarios.

V. Applications of the IIoT Data

Using this data set generator, we collected data from the industrial site of a large arm. After data preprocessing, a high-quality dataset is generated. This Dataset Generator saves time and cost for dataset processing compared to previous schemes. Users have a higher voice and data privacy. Technicians are easier to operate. It also has a mature visualization system, which provides an important basis for enterprises to optimize production and decision-making.

The dataset generated by the Dataset Generator can be simply visualized, or the corresponding algorithm can be redesigned for more specific improvements based on different scenarios. Next, a subsequent algorithm improvement process for the dataset of a mechanical arm is shown. We used a Dataset Generator to collect 500 times the normal operation of a mechanical arm [16]. The acceleration and sound of the arm are very high when it is working. We analyze these two factors.

First, we analyze only the sound data and design an acoustic wave width (AWW) algorithm based on the working frequency of a mechanical arm. AWW algorithm calculates the total wave width of the working interval of the large manipulator, and counts the acoustic convolution data reaching the threshold [17]. In this dataset, there are 490 manipulator working events (MWE). There are 3 error detection and 10 missed detection in AWW algorithm, and the accuracy is 97.35%. The algorithm flow is as follows:

Algorithm 1: Acoustic wave width algorithm.

```
Input: Data sequence
Output: Number of MWEs
for Follow pointer ≠ Forward pointer do
    Calculate sound difference to eliminates zero drift
    Take absolute value
    Cache reallocation
    Convolute with the specified size
    Cache reallocation
    if No mark on the upper bound of the interval then
        Judge whether the upper bound threshold of
         the interval is reached
        else if The lower bound is not marked then
         Reset key point
        Reset status bit
            Find the maximun value in the in terval
        end
        Maintain follow pointer
    end
end
```

In order to further improve the accuracy of identification of the working state of a large manipulator, acceleration data is used to assist the identification on the basis of acoustic signal detection [18]. This eliminates the interference caused by the collision of the surrounding arm. To solve the asynchronous sound and acceleration characteristics generated by the Dataset Generator, the Finite State Machine Detection (FSMD) algorithm is designed. The status table of the FSMD algorithm is

TABLE I
STATE TABLE OF FINITE STATE MACHINE DETECTION ALGORITHM

Accelerate PS	Sound PS	Accelerate SS	Sound SS	Accelerate output	Sound output	MWE output
0	0	0	0	0	0	0
0	0	0	M	0	0	0
0	0	1	0	0	0	0
0	0	1	M	0	0	M
1	N	0	0	0	0	N
1	N	0	M	0	0	N
1	N	1	0	0	0	N
1	N	1	M	0	0	N

shown in Table. I. Among them, 0 and 1 are the results of acceleration recognition and M and N are the results of sound recognition. It matches the asynchronous sound characteristics with the acceleration characteristics by dividing the current state (PS) and the sub-state (SS). Ultimately, mismatched noise signals consisting of sound and acceleration can be filtered out. In this training, the large manipulator works 490 times normally, AWW algorithm combined with FSMD algorithm has one error detection and three miss detection, the accuracy rate is 99.18%, which has significantly improved.

VI. CONCLUSION

Combining the idea of multi-sensor information fusion, this paper presents a data set generator in the scenario of industrial Internet of Things. It has the advantages of convenient installation, strong signal, low delay and low energy consumption. The Dataset Generator generates a mature dataset that the user decides whether or not to expose. This not only improves the voice of users, maintains the privacy and security of data, but also makes up for the current lack of datasets in the industrial Internet of Things. At the same time, the Dataset Generator makes industrial data collection process no longer dependent on specific hardware devices such as gateways, reducing the cost of intelligent transformation of enterprises. The visualization of data enables enterprises to monitor real-time failures and timely repair projects in various industrial scenarios. It provides an important basis for enterprises to optimize production and make decisions.

VII. ACKNOWLEDGEMENTS

This work was supported by the National Key Research and Development Program of China under Grant No. 2021ZD0113003, National Natural Science Foundation of China under Grant No. 61901228, China Postdoctoral Science Foundation Project under Grant No. 2021M702466.

REFERENCES

[1] X. Hou, Z. Ren, K. Yang, C. Chen, H. Zhang, and Y. Xiao, "IIoT-MEC: A Novel Mobile Edge Computing Framework for 5G-Enabled IIoT," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, 2019, pp. 1–7.

[2] N. Koroniotis, N. Moustafa, F. Schiliro, P. Gauravaram, and H. Janicke, "The SAir-IIoT Cyber Testbed as A Service: A Novel Cybertwins Architecture in IIoT-Based Smart Airports," *IEEE Transactions on Intelligent Transportation Systems*,to be published, doi: 10.1109/TITS.2021.3106378.

[3] A. C. Panchal, V. M. Khadse, and P. N. Mahalle, "Security Issues in IIoT: A Comprehensive Survey of Attacks on IIoT and Its Countermeasures,"in *2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN)*, 2018, pp. 124–130.

[4] V. Sklyar and V. Kharchenko, "ENISA Documents in Cybersecurity Assurance for Industry 4.0: IIoT Threats and Attacks Scenarios," in *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 2019, pp. 1046–1049.

[5] A. Gabriel, W. P. Nwadiugwu, J. M. Lee, and D. S. Kim, "Energy-Aware Routing Scheme for Large-Scale Industrial Internet of Things (IIoT),"in *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, 2019, pp. 608–611.

[6] Y. H. Lai, Y. H. Huang, C. F. Lai, S. Y. Chen, and Y. C. Chang, "Dynamic Adjustment Mechanism based on OPC-UA Architecture for IIoT Applications," in *2020 Indo–Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN)*, 2020, pp. 335–338.

[7] H. Sun, Y. Jin, M. Fu, J. He, H. Liu and W. A. Zhang, "A Multisensor-Based Tightly Coupled Integrated Navigation System," in *2022 5th International Symposium on Autonomous Systems (ISAS)*, 2022, pp. 1–6.

[8] J. Xie, S. Huang, D. Wei and Z. Zhang, "Scheduling of Multisensor for UAV Cluster Based on Harris Hawks Optimization With an Adaptive Golden Sine Search Mechanism," *IEEE Sensors Journal.(Indo-Taiwan ICAN)*, 2020, pp. 335–338.

[9] F. P. Martins, J. A. R. Paixão, C. M. de Farias, and F. C. Delicato, "Hercules: A Context-Aware Multiple Application and Multisensor Data Fusion Algorithm," in *2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 2021, pp. 197–200.

[10] M. Liu, K. Yang, N. Zhao, Y. Chen, H. Song, and F. Gong, "Intelligent Signal Classification in Industrial Distributed Wireless Sensor Networks Based Industrial Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 4946–4956, July 2021.

[11] Z. Liu, G. Xiao, H. Liu and H. Wei, "Multi-Sensor Measurement and Data Fusion," *IEEE Instrumentation Measurement Magazine*, vol. 25, no. 1, pp. 28–36, February 2022.

[12] K. R. Shahi, P. Ghamisi, B. Rasti, P. Scheunders and R. Gloaguen, "Unsupervised Data Fusion With Deeper Perspective: A Novel Multisensor Deep Clustering Algorithm," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 284–296, 2022.

[13] E. Sisinni, A. Saifullah, S. Han, U. Jennehag and M. Gidlund, "Industrial Internet of Things: Challenges, Opportunities, and Directions," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 4724–4734, Nov. 2018 .

[14] M. Aazam, S. Zeadally and K. A. Harras, "Deploying Fog Computing in Industrial Internet of Things and Industry 4.0," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4674–4682, Oct. 2018.

[15] Z. Li, J. Kang, R. Yu, D. Ye, Q. Deng and Y. Zhang, "Consortium Blockchain for Secure Energy Trading in Industrial Internet of Things,"*IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3690–3700, Aug. 2018.

[16] J. Huang, L. Kong, G. Chen, M. Wu, X. Liu, and P. Zeng, "Towards Secure Industrial IoT: Blockchain System With Credit-Based Consensus Mechanism," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3680–3689, June 2019.

[17] X. Li, J. Niu, M. Z. A. Bhuiyan, F. Wu, M. Karuppiah and S. Kumari, "A Robust ECC-Based Provable Secure Authentication Protocol With Privacy Preserving for Industrial Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3599–3609, Aug. 2018.

[18] F. Alturjman and S. Alturjman, "Context-Sensitive Access in Industrial Internet of Things (IIoT) Healthcare Applications," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 6, pp. 2736–2744, June 2018.