Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 46 No. 6
June 2025

Data Reduction Technique for Smart Agriculture in Wireless Sensor Networks using Machine Learning

Deepak Singh Rawat[1], Sunil Kumar[2], Dr Javalkar Dinesh[3]

1. Student Department of Electronics and communication Engineering, Lingaya's Vidyapeeth, Faridabad Haryana – 121002 INDIA
2. Student Department of Electronics and communication Engineering, Lingaya's Vidyapeeth, Faridabad Haryana – 121002 INDIA
3. Faculty Department of Electronics and communication Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana – 121002 INDIA

Abstract **-** "Smart farming is an advanced approach that combines technologies like the IoT, automation, drones, and artificial intellect to boost both the harvest and quality of farm products while tumbling manual work. By applying a wireless sensor network, which delivers self-governing energy, tracks valve and shift functionality, and achieves remote locations, high-quality crops can be produced consistently throughout the year. Wireless sensor networks, designed to collect data from all sensors with low energy consumption and extensive communication ranges, are a fundamental element of IoT. This paper proposes a system for monitoring soil moisture, temperature, and humidity in small-scale farms. Transmitting large quantities of data can lead to high energy and bandwidth consumption on the sensor nodes. To mitigate this, a machine learning algorithm is introduced, aimed at reducing data using the Data Reduction Algorithm (MLDR). MLDR's purpose is to efficiently gather commercial data. It works as a technique for dimensionality reduction, utilizing machine learning at the sensor network interface to minimize the amount of data sent to the central system while ensuring accuracy and relevancy."

Keywords – IOT, MLDR, WSN

I. INTRODUCTION

In many towns and villages across various nations, agriculture serves as the foundation of the local economy, providing a significant incentive for its development. Additionally, agriculture is the main source of income for several countries. Farms are divided into different zones based on various factors. The 2030 Agenda for Sustainable Development emphasizes the importance of gathering accurate data to combat drought, prevent famine, and support small-scale farmers. This initiative aims to enhance agricultural productivity in developing countries while simultaneously reducing food scarcity Severe weather events can result in significant losses for farmers and others who depend on crops for their livelihood. A useful solution has emerged in the form of a framework that can predict maize crop yields two to three months in advance, helping farmers avoid unexpected challenges.

This framework provides thorough, well-organized, and always available data, which can be accessed globally for both monitoring and financial management purposes. However, Wireless Sensor Networks (WSNs) face challenges due to the limited energy resources available for sensors, which reduces the overall longevity of the network. Despite the sampling frequency, each sensor continuously monitors its designated variable and transmits the data to the central system for processing.

In modern industrial agriculture, there is a demand for new, more efficient methods to address emerging challenges. Climate change and limited resources are adding to the complexity, creating a need for automated systems and smarter decision-making. WSNs are now being utilized in numerous sectors, including healthcare, industrial applications, environmental monitoring, and surveillance. These networks must contend with factors such as inconsistent network availability, high bandwidth consumption, and the constrained energy capacity of sensor nodes. As a result, redundant data transmission can become an issue,

especially when monitoring parameters remain stable. Managing large volumes of data generated periodically at the sensor level is a complicated task. For example, each node might capture and transmit temperature data every 30 minutes, generating 48 data points per hour, with multiple devices collecting data across various environmental parameters Aside from temperature, there are several other parameters that may fluctuate more frequently. To address this issue, researchers propose the implementation of a Machine Learning based Data Reduction Algorithm (MLDR) technique. This approach employs a training process to forecast future values of monitored characteristics, such as temperature, at either the sensor or the central system.

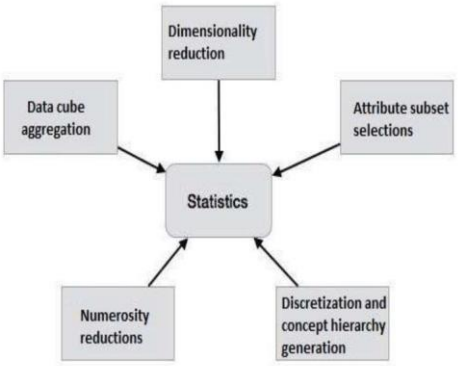
. By using a dual-prediction approach, MLDR enhances the system's efficiency and minimizes unnecessary data transmission from the sensor nodes. Consequently, this reduces the energy consumption during data communication. Rather than transmitting every piece of recorded data to the central system, the sensor nodes only send essential information, maintaining a balance between energy savings and data accuracy. The system's state is continuously monitored and updated. Following the learning phase, both the central system and sensor network commence the predictive process. The results indicate the successfulness of this approach, as it reduces data transmission by approximately 70% while maintaining high data reliability at the central system.

Data Reduction:

Data reduction refers to the technique of reducing the amount of storage space required for data. This quantitative method enhances system capacity and helps lower operational costs. Manufacturers often focus on raw total capacity and alternate storage volume, which typically reflects data after it has been compressed. In terms of data storage, one common issue is running out of space due to excessive data accumulation. Data compression can significantly boost memory performance and efficiency by requiring less storage. To reduce the overall volume of data stored within a system, data reduction employs various strategies.

Data Reduction Strategies:

Fig. 1. Strategies in Data Reduction



a. Data Cube Aggregation: In the creation of digital images, aggregation operations are applied to the dataset.

b. Dimensionality Reduction: This process identifies and eliminates redundant features, thereby creating a smaller and more efficient dataset by which data will accurate.

c. Data Compression: It is utilized to optimize frameworks for data transmission, reducing influence dataset.
d. Numerosity Reduction: A quantitative method where data values are substituted or approximated by alternative representations, thus reducing the data's size.
e. Discretization and Concept Hierarchy Generation: This method involves replacing continuous data with higher-level classifications or grouped categories to simplify the dataset.

Data reduction plays a significant role in data mining, as it helps to simplify the analysis process. In modern agriculture, there is a growing need for innovative techniques due to several emerging challenges. Issues like global climate change and water scarcity have intensified the demand for better farming methods. Consequently, automation and intelligent decision-making technologies have become vital. Wireless Sensor Networks (WSNs) are increasingly used across various sectors as affordable and efficient measurement tools. In sustainable agriculture, these networks operate using a star topology to

collect environmental data from numerous sensor nodes, which is then transmitted to a central sink for processing.

Dimensionality reduction and clustering are considered unsupervised classification methods. Clustering is the process of grouping similar objects into clusters, ensuring that no two items within the same cluster are exactly identical. This method, being unsupervised, aims to uncover hidden patterns in the data. Each object is characterized by specific features, and the first step in clustering involves defining how to separate these objects. The success of the clustering algorithm largely depends on selecting an appropriate distance metric. Numerous clustering techniques are available, each presenting distinct advantages and limitations. Among these, the k (means) algorithm is one of the mostly used, as it iteratively selects k centroids to establish clusters. These centroids serve as representatives for the items within each cluster. The primary advantages and disadvantages of the k-means method include:

a. The determination of the number of clusters (k) must be made a priori, which poses a challenge when the exact number of clusters is indeterminate.

b. Owing to the iterative nature of the k-means algorithm, it may converge to a local optimum, potentially resulting in inaccurate outcomes.

c. The algorithm presupposes that clusters are spherical in shape, an assumption that does not universally hold true.

.The proposed system prioritizes energy efficiency through the incorporation of a data reduction mechanism that utilizes machine learning. Energy consumption plays a vital role in determining both the lifespan of batteries and the data storage capabilities within Wireless Sensor Networks (WSNs). Consequently, the system is engineered to decrease energy expenditure by minimizing the amount of data transmitted among sensor nodes. Over the years, data compression in WSNs has garnered significant attention. Common techniques for data reduction include clustering, rescheduling, compression detection, nonlinear inter-data compression, database associations, and pattern recognition, such as dual prediction models. Clustering techniques specifically help to lower energy consumption in WSNs by recognizing spatial-temporal patterns in the data. The authors suggest using a fuzzy inference system to re-cluster nodes for enhanced energy savings. Key factors such as the average distance between the sink and cluster head routers are considered to achieve optimal energy efficiency.

Machine learning techniques are applied to categorize similar data within each network element, thus minimizing the need for excessive data transmission. The paper also introduces a Bayesian Interpretation Methodology for detecting spatial correlation and data volume, designed to minimize redundant data transmission. By leveraging temperature and pressure data, this methodology enables data reduction, utilizing a large set of connected data points for more efficient transmission. The classification process is carried out using machine learning, focusing on similarities between nodes within a cluster, observed data rates, and the proximity of cluster members.

Key observations discussed in this paper include:

- **a**. To address the increasing risks posed by climate change, it is crucial to track weather predictions and soil conditions at the community level to facilitate prompt and informed decision-making.
- **b**. This paper presents the Machine Learning-based Data Reduction Algorithm (MLDR) and emphasizes its use in agricultural settings.
- **c**. MLDR is a method for data reduction that guarantees the retention of accurate data at the destination while minimizing the amount of data sent from sensor nodes by utilizing machine learning techniques..

Previous research, such as the work by Pradipkumar M et al. [1], proposed dynamic management strategies for groundwater resources in IoT-based environments. This research incorporates cloud computing with IoT systems, enabling global access to sensor data. Variables like pH, temperature,

turbidity, and dissolved oxygen can be monitored

using sensors, and the data is made accessible through cloud computing platforms for real-time analysis. Atif A et al. [2] explored the concept of

Sensor-Cloud Architectures, addressing the challenges of managing vast amounts of sensor data by utilizing scalable cloud infrastructure for data storage and processing. Nikhil Kedia et al. [3] proposed an affordable water quality monitoring system using sensor cloud technology, specifically designed for rural areas, which alerts authorities when water quality issues arise. R. Karthik Kumar et al. [4] presented a solar-powered advanced water quality management system using IoT technology, where Zigbee-based WSNs transmit sensor data such as pH and oxygen levels to a central system for analysis. Daiho Uhm et al. [5] examined a data reduction technique using singular value decomposition (PCA) and factor analysis (FA)

for more efficient data interpretation in data mining, helping to mitigate dimensionality issues and reduce processing time.

II. IMPLEMENTATION

The ML based Data Reduction Algorithm (MLDR) aims to improve data reduction processes. In both scenarios, the sensor node first sends the initial recorded value to the sink. Subsequently, each sensor network enters a learning phase, gathering n consecutive data points to identify trends or changes occurring among them.

$$tr1 = val1 - val0$$

$$tr2 = val2 - val1 \quad \ldots \quad (1)$$

$$tri = vali+1 - vali$$

In this context, **tr** represents the relationship between the data points, showing how specific parameters fluctuate over time.

$$d1 = |tr1 - tr0|$$

$$d2 = |tr2 - tr1| \quad \ldots \quad 2)$$

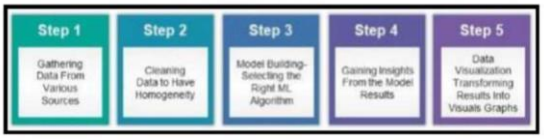
$$di = |tri+1 - tri|$$

d denotes the difference between the identified trends.

In the realm of WSN-based smart agriculture, machine learning is employed for data reduction. This paper outlines two primary approaches: Hold (MLDR-HOLD) and Buffer (MLDR-BUFFER), which vary based on the number of parameters being processed at the nodes. Furthermore, sensor networks have limited memory, meaning they cannot retain large amounts of data. Therefore, the choice between these two methods depends on the storage capacity of the node.

In the MLDR-Hold approach, when predictions become challenging, the system sends a "hold" command to the sink, prompting the training phase to continue and attempt to identify new trends based on the incoming data. Conversely, the MLDR-Buffer approach utilizes a buffer to temporarily store recent data, which is then processed to detect potential new patterns. If a pattern is found, it is immediately sent to the sink to enhance prediction accuracy. If no pattern is identified, the process returns to the training phase. This approach not only helps to conserve energy and reduce bandwidth consumption but also improves the efficiency of communication while maintaining data quality.

Fig. 2. Workflow of ML



Data Collection: The method employed for data collection is contingent upon the unique characteristics of the project. For instance, in a machine learning initiative that relies on real-time data, an Internet of Things (IoT) framework with several sensors is commonly established to gather the necessary information. This data may originate from diverse sources, including files, databases, or sensors. However, it often requires further processing before it can be analyzed effectively, as it may contain problems like missing values, significant outliers, improperly formatted text, or background noise. Therefore, data preparation becomes a crucial phase in the process.Data

Cleaning and Preprocessing: Data preprocessing involves transforming raw data, which is collected from different sources, into a clean and usable format. Data gathered in its raw state is often disorganized and unsuitable for direct analysis. The preprocessing phase encompasses various techniques aimed at cleaning and organizing the data to make it suitable for analysis.

Model Selection for the Data Type: The main aim here is to choose the most effective model based on the characteristics of the pre-processed data. The right model is selected after considering the type and structure of the data.

Training and Testing the Model: When training a model, three important datasets are used: "Training data," "Validation data," and "Testing data." Initially, the classifier learns from the "training data," and its parameters are fine-tuned using the "validation data." The classifier's performance is then assessed using "unseen test data." It is critical that the test data remains separate and is not used during the training phase. The test dataset should only be accessed during the evaluation phase after the training process has been completed.

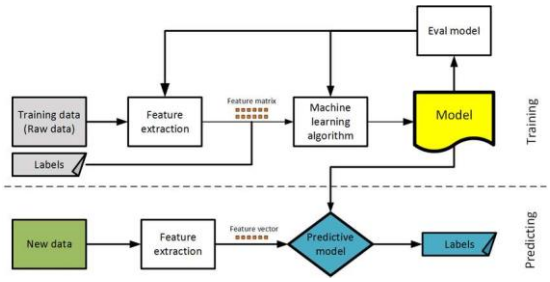


Fig. 3. Model

a. The fundamental concepts underlying the proposed machine learning-driven data reduction method (MLDR) are presented.

b. Within the MLDR framework, nodes embark on a learning journey, collecting data to comprehend the behavior of the parameters being monitored.

c. When a significant event is identified, the sensor node transmits the data along with the most recent reading from the learning phase to the source.

d. Subsequently, routine measurements are not sent to the sink unless a significant change, determined by established thresholds, occurs.

e. Both the sink and sensor nodes engage in a dual prediction process to estimate average values based on the latest data transmission and historical observations.

f. The sensing capabilities of each sensor node remain consistent over time. For every new reading received, the node assesses the estimated value against the actual measured value.

g. If a considerable discrepancy is detected (as per the learning phase), the monitoring server transmits the predicted value to the source as an adjustment signal, thereby preventing unnecessary predictions.

h. The node then reverts to the learning mode, seeking to identify a new pattern.

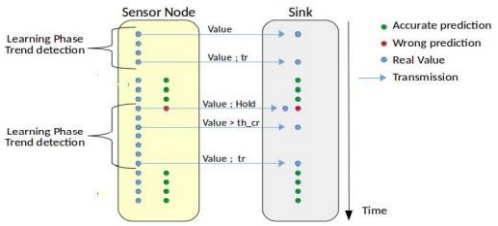


Fig. 4. The MLDR-H technique is used to model the behaviour of sensor nodes

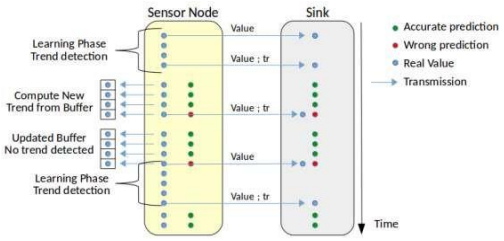


Fig. 5. Behaviour of sensor nodes employing the MLDR-B algorithm with a buffer size of n = 3

i. This research presents two unique variants: the Hold (MLDR-HOLD) and the Buffer (MLDR-BUFFER), each differing in computational effort required at the endpoints.

j. Due to the limited bandwidth of sensor networks, there is a need to conserve large volumes of historical data.

k. Given that sensor networks have limited bandwidth and cannot reliably store extensive past

data, two strategies can be employed based on the capacity of the node. In the Hold MLDR variant, when the sensor node identifies problems with its predictions, it transmits a hold signal to the sink. This action halts the predictions and initiates a new learning phase to identify emerging trends from the incoming data.

l. The Buffer MLDR version, on the other hand, temporarily stores the most recent data in a buffer, which is later used to calculate a new trend (if necessary).

m. If a new pattern is recognized, it is immediately transmitted to the sink to ensure the accuracy of predictions. If no new pattern is detected, the process returns to learning mode.

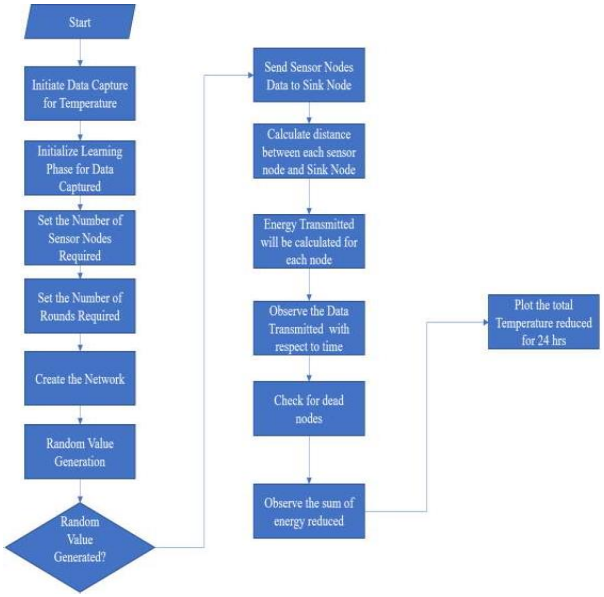


Fig. 6. Flow chart of Implementation

The process flow for the suggested system is outlined in the diagram above. Upon system initialization, temperature data will be processed, and machine learning will be used to monitor the data. The next step involves defining the number of nodes needed for transmission, which will facilitate the setup of a network composed of Sensor and Sink nodes. Subsequently, random temperature readings will be generated and assigned to each sensor node, which will transmit the data to the Sink node. The distance between nodes will be calculated, which is essential for energy consumption calculations. These calculated values will be evaluated in each cycle, with differences between consecutive values being determined. These differences will then be compared to a set threshold (5V). The number of transmitted data packets will be tracked, and the number of dead nodes (or data failing to meet the threshold) will be measured. Finally, the temperature graph will illustrate the data after applying the reductions based on the earlier criteria.

III. RESULTS

MATLAB is used to generate random data, where 24 data points are created. Each temperature value is then sequentially fed into the network, following a pipelined approach as shown in Fig. 7.

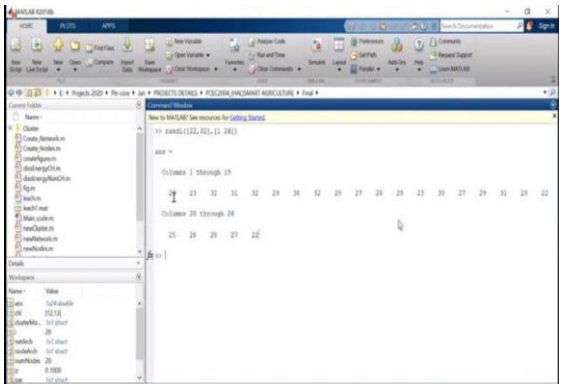


Fig. 7. Data

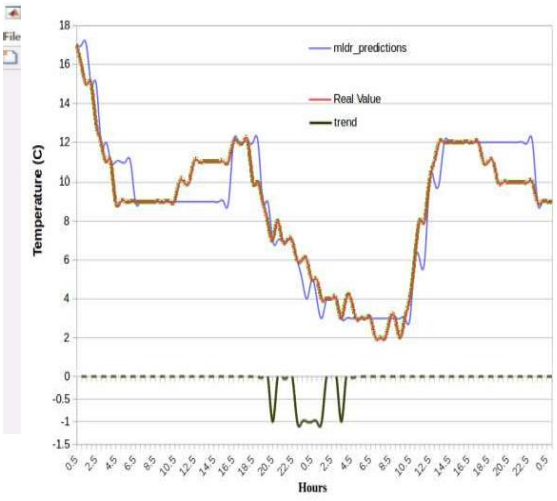
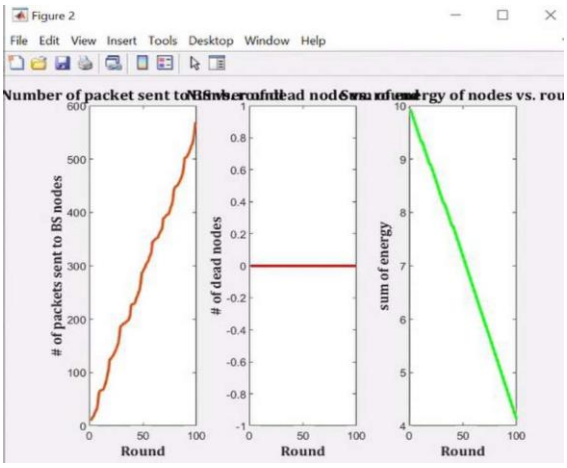


Fig. 8. Network

Figure 8 illustrates a WSN network arranged in a 20x20 grid, where the green node represents the sink node and the blue nodes act as sensor nodes, each transmitting data to the sink. The sink node

possesses the most energy, as it is responsible for sending the data to the processing server.

Fig. 9. Operations



The first graph illustrates the count of transmitted nodes, while the second graph shows the number of nodes that failed during the data transmission process. The third graph presents the total energy consumption of the nodes across various rounds. Some potential limitations include node failure due to battery depletion and errors that may arise in the collected data.

Regarding another experiment referenced in [1], temperature data from October 27th to 30th, 2019, in Lille, France, was considered. The data was sampled every 30 minutes, with temperature variations ranging from 2 to 17 degrees Celsius. A total of 192 samples were gathered at the nodes. However, in order to preserve data accuracy, the amount of data sent to the sink node was reduced to 25 using MLDR-H and 27 using MLDR-B. Although MLDR-B is less energy-efficient, it maintains data integrity better than MLDR-H.

The difference between the values predicted by the MLDR model and the actual measurements varied from 0 to 2 degrees Celsius. Figure 10 illustrates the relationship between the MLDR predictions and the observed values, along with the variations in trends for different data points.

Fig. 10. Observation [1]

The values captured by the nodes each day, along with the number of values transmitted to the sink node, are presented in TABLE I. It also displays the percentage of data reduction achieved by MLDR-H and MLDR-B separately.

TABLE I. Experimental Results

Day	All data	MLDR-H	MLDR-B
27^{th}	48	8	9
28^{th}	48	10	11
29^{th}	48	5	5
30^{th}	48	2	2
Total	192	25	27
Data Reduction	0%	87%	85%

IV. CONCLUSION & FUTURE SCOPE

This research investigates a method for data reduction in agriculture, utilizing a fundamental machine learning approach within wireless sensor networks (WSN) to identify anomalies in climate variations that may adversely impact farming practices. The simulations reveal a data reduction exceeding 75% of the total dataset, achieving up to 50% less data compared to alternative techniques. Future studies could incorporate additional elements, such as spatial and temporal data correlation or variable sampling rates, into this data reduction framework to tackle the challenges associated with the substantial data influx from wireless sensors, ultimately resulting in a more robust solution. A scheduling mechanism could also be implemented to optimize the sensing process during periods of trend changes, or the network's data rate could be adapted to achieve the same objectives.

REFERENCES

[1] Christian Salim, Nathalie Mitton. Machine Learning Based Data Reduction in WSN for Smart Agri-culture. AINA 2020 - 34th International Conference on Advanced Information Networking and Ap-plications,

Apr 2020, Caserta, Italy. hal-02463167

[2] Balducci, F., Impedovo, D., Pirlo, G.: Machine

learning applications on agricultural datasets for smart farm enhancement. Machines 6(3) (2018)

[3] Monteiro,L.C.,Delicato,F.C.,Pirmez,L.,Pires,P. F.,Miceli,C.:Ducas: Data prediction with cubic adaptive sampling for wireless sensor networks. In: International Conference on Green, Pervasive, and Cloud Computing. pp. 353–368. Springer (2017)

[4] Radhika, S., Rangarajan, P.: On improving the lifespan of wireless sensor networks with fuzzy based clustering and machine learning based data reduction. Applied Soft Computing 83 (2019).

[5] R. Sharma, S. S. Kamble, A. Gunasekaran, V. Kumar, and A.kumar review on machine learning applications for sustainable agriculture supply chain performance compute oper. Res., vol 119 Jul. 2020, Art. no. 104926.

[6] K.G. Liakos, P. Busato, D. Moshou, S.pearson, and D. Bochits, "Machine learning in agriculture: A revie, "Sensors, bol. 18, no.8. p. 2674, 2018.

[7] C.F. Gaitan, "Machine learning applications for agricultural impacts under extreme events," in Climate Extremes and their implications for Netherlands: Elsevier, 2020, pp. 119-138.

[8] E.Acar,M.S.Ozerdem and B.B. Ustundag,"Machine learning based regression model for prediction of soil surface humidity over moderately vegetated fields," in Proc. 8th Int. Conf. Agro-Geoinformat, Istanbul, Turkey, Jul. 2019, pp. 1-4.

[9] X. Wang, W. Hu, K. Li, L. Song, and L. Song, "Modeling of soft sensor based on DBN-ELM and its application in measurement of nutrient solution composition for soilless culture." In Proc. IEEE Int. Conf. Saf. Produce Information (IICSPI), Chongqing, China, Dec. 2018, pp. 93-97.

[10] R. Andarde, S. H. G. Silva, D. C. Weindrof, S. Chakraborty, W.M. Faria, L. F. Mesquita, L. R. G. Guilherme, and N. Curi, "Assessing models for prediction of some soill chemical properties from portable X-ray Fluorescence spectrometry data in Barzallian coastal plants." Geoderma, vol. 357, Jan. 2020, Art. No. 11395