Internet of Things 32 (2025) 101611

Contents lists available at ScienceDirect

Internet of Things

journal homepage: www.elsevier.com/locate/iot

Research article

Virtual sensors for smart farming: An IoT- and AI-enabled approach

Athanasios Chourlias [a], John Violos [b,*], Aris Leivadeas [b]

[a] Department of Informatics & Telematics, Harokopio University of Athens, Omirou 9, Tavros, 177 78, Greece
[b] Department of Software and IT Engineering, École de technologie supérieure, 1100 Notre-Dame St W, Montreal, H3C 1K3, Canada

ARTICLE INFO

Keywords:
Physical sensors
Virtual sensors
IoT
Smart farming
Smart agriculture
Machine learning

ABSTRACT

Smart farming relies on precise environmental data to optimize agricultural practices, with key metrics such as air temperature, humidity, rain, ambient light, ultraviolet (UV) radiation and soil moisture to play a crucial role in agricultural decision-making. However, the vast spatial coverage of agricultural fields and the high cost of deploying numerous physical sensors pose significant challenges, particularly for small and medium-sized farms. To address these issues, virtual sensors – machine learning models that predict sensor values based on data from relevant physical sensors – offer a cost-effective and scalable alternative. In this research, a number of Arduino-based IoT devices are designed and deployed equipped with various physical sensors, a lithium-polymer battery which recharges continuously using a 6 W waveshare solar panel, and a Real-Time Clock (RTC) module that synchronizes data logging. The IoT devices operated across two agricultural fields over a span of almost three months. The data collected form the basis for evaluating multiple machine learning models as virtual sensors. Furthermore, the use of open weather data to develop a hardware-free solution is explored. Experimental results show that virtual sensors provide a cost-effective and accurate method for replacing physical sensors. The Light Gradient Boosting Machine emerged as the most accurate model for virtual sensors, achieving prediction errors of less than 1% in most of the cases. This makes it a valuable tool for enabling cost-effective and data-driven farming in resource-constrained IoT devices.

1. Introduction

Smart farming integrates advanced technologies such as Internet of Things (IoT), data analytics, and Artificial Intelligence (AI) to streamline and automate various aspects of agricultural production [1]. One of the key advantages of smart farming is its ability to analyze real-time data from diverse sources, such as sensor networks, drones, robots, and weather stations [2]. This empowers farmers with actionable insights, enabling them to make more informed and precise decisions about farm management.

Smart farming equips farmers with the tools to oversee and optimize every stage of the agricultural process, from irrigation and fertilizer application to plant disease prevention [3]. By incorporating technologies such as virtual sensors and autonomous systems, smart farming helps reduce resource waste, improving crop yields, enhance pest and disease management, and increase overall agricultural efficiency through real-time data analysis and automation [4]. IoT sensors positioned strategically across the field measure critical parameters like soil moisture, temperature, air humidity, and light intensity, providing a continuous stream of data [5]. This data is analyzed to deliver actionable insights, enabling precise and timely interventions tailored to the specific needs of each section of the field and enables farmers to respond more swiftly and effectively to changing environmental conditions, pest outbreaks, crop stress, and resource inefficiencies.

* Corresponding author.
E-mail address: ioannis.violos@etsmtl.ca (J. Violos).

https://doi.org/10.1016/j.iot.2025.101611
Received 17 January 2025; Received in revised form 17 March 2025; Accepted 8 April 2025
Available online 3 May 2025

Virtual sensors are becoming increasingly important in smart farming, using data from physical sensors to estimate conditions in areas lacking direct measurements [6]. This integration combines the precise, location-specific accuracy of physical sensors with the cost-effective, broad coverage offered by virtual sensors, albeit with slightly lower accuracy [7].

To address this challenge, in this paper, a Machine Learning (ML) model that establishes the relationship between physical sensor measurements and virtual sensor estimates is developed, by initially deploying physical stations equipped with sensors at locations designated for virtual measurements. After collecting time-series data from these sensors, an ML model is trained for the virtual sensors. Once the model is trained, the physical stations are replaced by virtual stations [8].

Additionally, a scenario is explored in which the model uses open data as input rather than physical sensors. In this configuration, all physical stations are removed entirely, and variables such as temperature, humidity, soil moisture, light, and UV levels are estimated solely through virtual sensors.

To investigate how monitoring data from physical sensors can be utilized to train virtual sensors, this study first designs and deploys IoT-enabled monitoring stations equipped with six physical sensors each across two agricultural fields in Anonymized city, Anonymized Country. The larger field initially hosted three stations, while the smaller field had two. However, during the 84-day experiment, natural disasters in the region resulted in the loss of one station [9]; three stations were recovered from the larger field, and only one from the smaller field.

Hence, this research explores the construction of physical stations using Arduino and integrated physical sensors, exchanging data via the LoRaWAN protocol into a centralized processing and storage node. It also analyzes the collected data using exploratory techniques such as time series plots, violin plots, and correlation matrices. Furthermore, multivariate, multi-output ML models are developed designed to function as virtual sensors, providing an efficient alternative to physical monitoring. Additionally, two representation and modeling approaches are examined, the time-series-based and the feature-vector-based wanted to see whether capturing temporal dependencies can improve the performance of virtual sensors.

Furthermore, this paper also tackles a common challenge in the field: the lack of publicly available, real-world datasets for training supervised and unsupervised ML models. Many organizations collecting such data do not share it openly, or access comes at a prohibitive cost, which underscores the importance of our contribution to advancing smart farming research. To overcome this limitation and promote innovation and collaboration in smart farming research, a comprehensive dataset has been made publicly available. This dataset, collected from two agricultural fields, includes measurements obtained from physical sensors, aiming to lower barriers for researchers and developers.

Thus, the major contributions of our work can be summarized as follows:

- This work demonstrates the process of designing and integrating physical sensors with IoT devices in a smart farming use case.
- Experiments are conducted in agricultural fields to collect data, resulting in a publicly available dataset.
- Various ML models are evaluated and compared to identify the most accurate for virtual sensor applications.
- Key insights and lessons learned from deploying smart agriculture IoT devices in real-world settings are shared.

The rest of the paper is organized as follows. Section 2 presents the background and related work regarding the topics of smart farming and virtual sensors. Section 3 presents our methodology to build virtual sensors. Section 4 presents the IoT implementation of the virtual sensors. Section 5 explores potential ML models that can serve as the foundation for a virtual sensor. Following the deployment of IoT devices in real-world fields, measurements are collected in order to compiled a dataset, detailed in Section 6. To develop the most accurate and efficient virtual sensors, experiments were conducted, which are detailed in Section 7. Section 8 presents the observations and lessons learned from our experiments, as well as insights gathered from discussions with farmers Finally, Section 9 concludes the paper giving some future directions for further exploration of virtual sensors for smart farming.

2. Related work

IoT-enabled monitoring in farms relies on a network of sensors strategically placed across the fields to measure vital parameters like temperature, humidity, light intensity, UV radiation, and soil moisture [10]. These sensors continuously collect real-time data, which is transmitted to a centralized system not only for informational and alert purposes but also for explanatory insights and predictive analysis [11]. With this infrastructure, farmers gain granular visibility into the environmental conditions of their farms without the need for manual inspections [12]. This real-time monitoring allows for timely interventions, such as adjusting irrigation schedules [13] automated machinery [14], disease [15] and pest [16] management, and precision farming [17]. Monitoring is essential in agriculture but poses challenges due to high costs, including system initialization (hardware like IoT devices and infrastructure) and operating expenses (like maintenance and sensor calibration), which must be balanced against the low profit margins and risks inherent in farming [18]. To address these economic constraints, replacing physical sensors with virtual sensors can be a cost-effective solution.

Virtual sensors, also known as soft sensors, are computational tools that estimate parameters or measurements by leveraging mathematical models, ML algorithms, and available sensor data, based on physical sensing devices that are deployed in a different location or measuring different parameters [19]. In the context of the IoT, virtual sensors play a pivotal role by integrating diverse data streams from IoT-enabled devices such as environmental sensors, robotic equipment, and agricultural machinery [20]. This synergy enhances real-time monitoring and predictive analytics in smart farming. The use of virtual sensors in agriculture offers numerous benefits beyond reducing the need for expensive physical sensors, including improved scalability for monitoring large areas and enhanced decision-making capabilities through pattern recognition techniques and historical data [21]. By providing

continuous, reliable estimates of critical parameters like soil moisture, crop health, and weather conditions, virtual sensors empower farmers to optimize resource use, improve yields, and promote sustainable farming practices [22].

Virtual sensors involve several essential steps in their development. First, a dataset is collected and prepared, addressing issues like missing data, outliers, and synchronization of multirate data [23]. Next, prepossessing techniques are applied to ensure data quality and alignment with application requirements [7]. Furthermore, feature selection is performed to identify relevant input variables through methods like correlation analysis or mutual information, ensuring that redundant features are eliminated [24]. During training, selected features are used to develop a predictive model, such as linear regressor, neural networks, or support vector machines, to capture relationships between easy-to-measure inputs and hard-to-measure targets [25]. Finally, validation ensures the virtual sensor's accuracy and generalization by employing techniques like cross-validation and performance metrics such as mean squared error and mean absolute error.

Closer to our work is the research by Wongchai et al. [8], which examines virtual sensors for predicting various crop parameters, such as yield and environmental conditions, using input features like soil moisture, temperature, and historical cultivation data collected via IoT devices. Their study employs a weight-optimized neural network with maximum likelihood for feature representation and utilizes an ensemble architecture combining a stacked autoencoder and kernel-based convolution network for making predictions. Another relevant study by Patrizi et al. [6] proposes virtual sensors for soil moisture prediction, utilizing input features such as ambient temperature, relative humidity, soil temperature, solar radiation, and rainfall data gathered through a wireless sensor network. Their study employs a long short-term memory (LSTM) network, a deep learning approach optimized for time-series data, to predict soil moisture levels accurately. Both approaches eliminate the need for expensive physical equipment, offering a cost-effective and efficient solution for smart farming applications.

In this research work several novel contributions to the development and application of virtual sensors in smart farming are presented. Unlike existing approaches that typically focus on single-input, single-output systems, virtual sensors are designed to integrate inputs from multiple heterogeneous physical sensors to predict multiple distinct outputs, enabling more comprehensive and versatile monitoring. Additionally, this study explores both time-series-based and feature-vector-based modeling approaches to assess their effectiveness in capturing temporal and contextual relationships in agricultural data. Our study also addresses a critical aspect of scalability: evaluating the performance of virtual sensors in predicting measurements not only within the same agricultural field but also across fields located many kilometers apart. Furthermore, this study explores the potential of leveraging open data as an alternative to physical sensor measurements, aiming to overcome challenges related to data accessibility and significantly reduce reliance on costly physical sensors required for virtual sensor input. These innovations are intended to broaden the scope, enhance the efficiency, and increase the applicability of virtual sensors in smart agriculture, addressing critical gaps in current research.

Monitoring systems are important components for optimizing agriculture resource management, improving crop yield, and adapting to weather changes [26]. The emergence of intelligent, programmable, and delay- and disruption-tolerant wireless networks has significantly advanced the development of scalable and real-time monitoring solutions that are not only cost-effective but also resilient, adaptable, and suitable for deployment in smart environments [27]. LoRa (Long Range) technology, known for its low power consumption and long-range communication capabilities, has proven especially effective for IoT-based agricultural applications, such as free-range cattle monitoring, where energy-efficient, secure, and mobile sensor networks are essential for reliable data collection in remote and unpredictable environments [28].

Recent advances in ML, remote sensing, and monitoring technologies have shown immense potential in addressing critical challenges in smart farming. Integrated approaches using remote sensing and wireless sensor networks enable early detection and effective management of insect pests, offering scalable solutions for sustainable crop protection [29]. The enhancement of irrigation efficiency and promotion of sustainable agriculture has been investigated through the integration of advanced soil moisture monitoring technologies, including remote sensing and IoT-based systems, which offer timely, cost-effective, and high-resolution data to support precision water management, particularly in resource-constrained agricultural settings [30]. Precision nitrogen management also benefits significantly from sensing and AI, allowing timely and non-destructive crop nitrogen assessments and improving fertilization strategies through dynamic modeling and in-field digital twins [31]. Similarly, AI-enhanced sensing and geographic information systems tools are enhancing soil erosion monitoring by enabling high-resolution mapping and predictive modeling, which support informed land management and conservation strategies [32]. Building on these developments, there is a growing trend toward harnessing the combined power of IoT, Big Data, and AI to advance Climate-Smart Agriculture and broader environmental goals through predictive analytics and real-time monitoring [33].

Despite these advancements, a common limitation across these studies is the reliance on open and accessible data. While some researchers overcome this constraint by leveraging satellite and drone-based remote sensing, many do not incorporate ground sensors, which offer higher accuracy in monitoring localized environmental conditions. This gap is largely due to regional data unavailability and the high cost associated with deploying ground-based sensor networks. Virtual sensors present a promising solution to this challenge, providing an opportunity not only for farmers to optimize their practices but also for researchers to conduct high-quality experiments while contributing valuable data to the broader agricultural community.

3. Proposed methodology for virtual sensors

In our proposed methodology physical and virtual stations are integrated. Physical stations are equipped with the necessary physical sensors, while virtual stations rely on data from these physical stations or open weather data APIs to simulate measurements at various points in the field using ML models. Table 1 provides an explanation for the terms physical/virtual sensors, stations and base stations.

Table 1
Definition of key terms.

Term	Description
Physical sensors	These are tangible devices that directly measure environmental parameters, such as soil moisture, temperature, humidity, or UV radiation, in the field. They are part of the hardware in physical stations and provide real-time, location-specific data by interacting with the physical environment
Physical stations	These are hardware setups consisting of a microcomputer and four physical sensors. They are deployed in the field to directly measure environmental parameters such as soil moisture or temperature.
Base station	This physical station is designated as the central hub of the system, responsible for collecting data from other stations, processing it locally, or forwarding it to a cloud service. It serves as the primary anchor point within the monitoring network.
Virtual sensors	These are digital tools within virtual stations that replace individual physical sensors. They predict measurements (e.g., soil moisture or temperature) by analyzing data from existing physical sensors and applying ML algorithms.
Virtual stations	These are conceptual replacements for physical stations. Instead of using physical hardware to gather data, virtual stations rely on computational models to simulate the presence of a station in the field

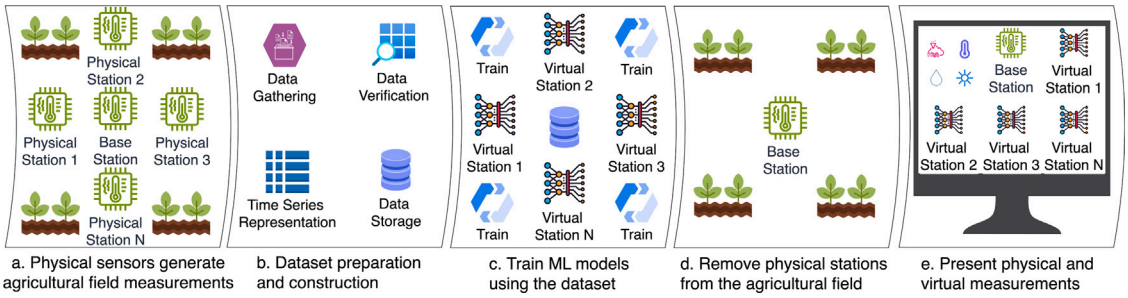

Fig. 1. The logical workflow of building virtual sensors.

3.1. System description

Fig. 1 illustrates the process followed to build the ML models for the virtual sensors that monitor the agricultural fields. The process starts with setting up and deploying physical stations at various locations across the agricultural field (Fig. 1a). Data collected from these physical stations is then gathered, verified, converted into time series, and stored to create a training dataset (Fig. 1b). Using this dataset, virtual sensors within virtual stations are trained. Each $virtualsensor_i$ learns the relationship between the base station and $physicalstation_i$ (Fig. 1c). This enables the removal of all physical stations, leaving only the base station in place (Fig. 1d). By utilizing the measurements from the base station and the virtual sensors, farmers can access highly accurate predictions for multiple locations within the agricultural field (Fig. 1e).

The proposed solution involves equipping physical stations with LoRa modules, specifically the SX1278 module, to transmit data to the base station which operates as central gateway. The Arduino MKR WAN 1310 (ABX00029) serves as a viable gateway option, capable of logging and forwarding data either to the cloud or to a centralized base station. The receiver utilizes the high-performance RFM95 W module, which has demonstrated exceptional range and reliability under real-world conditions. This setup ensures a centralized communication system, allowing all gathered data to be stored in a structured manner. Moreover, this system can function independently of cloud services, enabling local operation.

After the data collection phase, data validation, normalization, and transformation into time series format are performed. Based on these measurements a dataset is constructed that includes pairs of physical sensor time series between the measurements of a physical station with the other physical stations scattered in the agricultural fields. This dataset will serve to train the ML models of the virtual sensors. Following, the virtual sensors are integrated in virtual stations that operate on monitoring data of the physical stations, creating a robust pair of physical and virtual monitoring points. These steps will be explained in greater detail in the following subsections.

3.2. Centralized communication system

A centralized communication system connects the physical stations to a central processing unit (or base station). The preference for a centralized communication system stems from its ability to enhance the organization, efficiency, and accuracy of data management. By synchronizing field stations, a centralized system ensures systematic data collection and seamless transmission

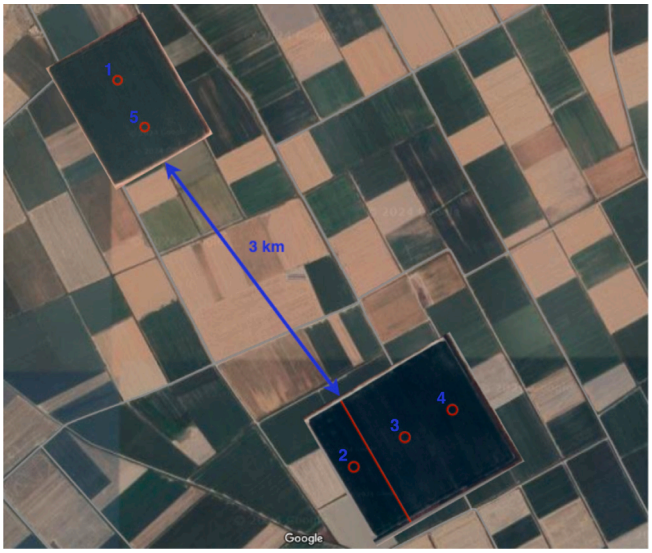

Fig. 2. Agriculture fields deployed with physical stations.

to the cloud. This enables real-time predictions to be generated directly in the cloud, removing the limitations imposed by resource-constrained models on edge devices. Moreover, the centralized setup supports the use of computationally intensive models, which can improve the overall performance of the system.

The centralized communication system can be realized using the LoRaWAN protocol, which is favored for IoT applications due to its unique benefits. It supports long-range communication, making it well-suited for large agricultural fields, and is optimized for low power consumption, ensuring extended battery life—an essential feature for remote field stations. The physical stations include a LoRaWAN transceiver that communicates the data to the base station which includes a LoRaWAN gateway.

3.3. Normalization and data transformation into time series format

Normalization methods are employed to ensure data comparability across different sensors. This includes standardizing measurements to common units and correcting discrepancies caused by varying measurement scales, which ensures that the dataset is ready for analysis. Raw data measurements from sensors can be organized into time series by associating each data point with a precise timestamp, ensuring temporal continuity. This process involves structuring the data into sequential intervals, allowing trends, patterns, and temporal dependencies to be analyzed effectively.

3.4. Training

The training process for our models starts with paired datasets collected from the base station and a physical station, both situated within the agricultural fields, as shown in Fig. 2. These datasets provide a foundation for training and evaluating the predictive capabilities of the virtual sensors.

Initially, the data from station 2 was designated as the reference dataset of the physical station which acts as base station, while the data from station 1, 3, and 4 were used to simulate virtual stations. The objective is to train models capable of predicting the readings of the virtual station based on the reference data. For each virtual sensor a multi-output and multi-variate ML model was trained using the input features of the physical station and the target values of the virtual stations. The features are detailed in Section 6. An alternative approach explored in this study involves training a model to predict a single output—namely a specific sensor value using input data from multiple other sensors. In addition, this study tested modeling multiple virtual stations simultaneously. However, as discussed in the experimental evaluation section, these two approaches yielded significantly lower performance.

3.5. Virtual stations

Virtual stations are ML models designed to predict sensor readings at specific locations leveraging data from physical stations. These models capture correlations arising from minor differences in soil properties, geographic positioning, or environmental factors such as the interplay of temperature and humidity. The most efficient and computationally lightweight approach utilizes a single model to simultaneously predict multiple values for one or more virtual stations, based on input from the physical station that acts as base station.

Fig. 3. The physical station setup before deployment (left) and actively measuring during operation (right).

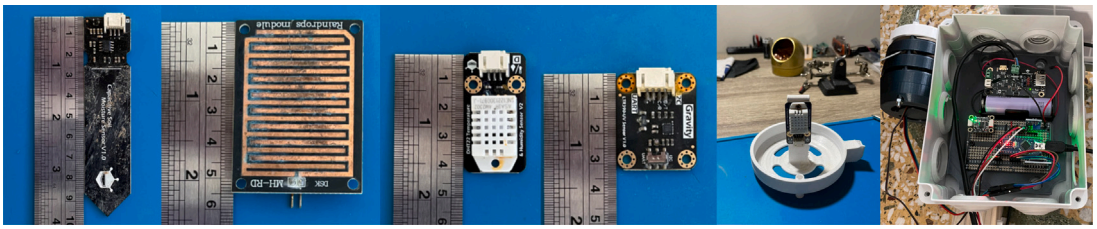

Fig. 4. From left to right: soil moisture sensor, rain detector sensor, temperature & humidity sensor, light & UV sensor. The last two images show the temperature & humidity sensor attached to the protective case and the electrical enclosure.

The model selection process involves careful evaluation of several factors, including the model's accuracy with the available dataset, its training and retraining efficiency, and its compatibility with different platforms, such as cloud services, local base stations, or resource-constrained MCUs like Arduino and STM32. This complexity is amplified for edge-level systems, where computational resources are limited. While larger models can improve performance on complex datasets, simpler tasks often benefit from lightweight models, which also require less computational power [34].

4. IoT infrastructure for smart farming sensors

This section outlines the technical details of our physical stations. Each station features an Arduino Nano microcontroller, which communicates with six sensors using Inter-Integrated Circuit (I2C) and Serial Peripheral Interface (SPI) protocols [35] and is synchronized with a real-time clock. The stations are powered by a cylindrical lithium-polymer (Li-Po) battery, continuously recharged by a Waveshare 6 W Solar Panel. Most components are securely housed within an electrical enclosure, while the soil moisture sensor is embedded in the soil. Additionally, a custom-designed protective case safeguards the temperature and humidity sensor, as well as the light and UV sensor. Fig. 3 illustrates the physical station both prior to deployment and during active operation. The components of the physical stations are depicted in Fig. 4 and further detailed in the following subsections.

4.1. Arduino Nano

The Arduino Nano is a compact microcontroller unit (MCU) based on the ATmega328 chip, widely used for embedded systems and prototyping due to its small size, ease of use, and versatility. It features digital and analog input/output pins that allow it to interface with various sensors, enabling the collection and processing of environmental data. The Arduino Nano connects to sensors using protocols like I2C, SPI, and analog inputs, allowing it to read data such as temperature, humidity, or light intensity. This makes it an ideal platform for sensor-based applications, where the MCU can receive sensor data, process it locally, and send it to other devices or cloud services for further analysis. Its low power consumption and compatibility with a wide range of sensors make it a popular choice for IoT and smart farming projects.

4.2. Soil moisture sensor

The physical stations were equipped with the V1 DFRobot Capacitive Soil Moisture Sensor which is a durable and efficient device designed for accurate measurement of soil moisture levels. Unlike traditional resistive sensors, it uses capacitive sensing technology to detect soil moisture without corrosion, ensuring long-term stability and reliability. The sensor operates with a voltage range

of 3.3–5.5 V, making it compatible with most MCUs, including Arduino. With a measurement depth of up to 22 cm, it provides precise readings that are unaffected by soil salinity. The sensor outputs analog signals, allowing seamless integration with MCU analog pins for real-time monitoring. Compact and easy to use, it is ideal for applications in smart farming, automated irrigation, and environmental monitoring.

For the soil moisture measurements, as well as the rain detector, which will be discussed in the next subsection, a unitless value is obtained, derived from the voltage generated by a variable resistance. The sensor generates an analog output, which is converted into a digital value by the processor using an analog-to-digital converter and calculated according to Eq. (1).

$$D = \left[\frac{V_{in}}{V_{ref}} \times (2^n - 1) \right] \tag{1}$$

where D is the ADC digital value, V_{in} is the analog voltage at the ADC pin, V_{ref} is the reference voltage of the ADC and n is the ADC resolution in bits.

4.3. Rain detection sensor

The rain sensor module is a simple and effective tool for detecting rain by measuring changes in resistance on its exposed copper traces. It operates with a voltage range of 3.3–5 V and provides both analog and digital outputs, making it compatible with various MCUs. The module is equipped with a sensitivity adjustment potentiometer, allowing fine-tuning for specific applications. Compact and easy to integrate, it is ideal for weather monitoring, automated irrigation, and smart home projects.

4.4. Protective case

For the temperature & humidity sensor and the light & UV sensor, which will be presented in the following subsections, a 3d printable protective case was designed. This protective case enables the sensors to obtain accurate readings from the external environment while safeguarding them from adverse external conditions. The light and UV sensor required an unobstructed view of the sky. The protective case was modified to include a large window in the top section. The window was made from large window in the top section. The window was made from laser-cut glass and affixed securely with thermal resistant silicone adhesive. The accuracy of the sensors within the protective case was tested using the MCU to validate and calibrate them prior to deployment for the experimental process.

4.5. Temperature & humidity sensor

For temperature and humidity measurements, the DFRobot DHT22 sensor was employed. It offers humidity measurement range of 0%–100% RH and a temperature range of −40 °C to 80 °C, providing accurate readings with minimal drift. The sensor features a factory-calibrated digital signal output, making it easy to integrate with MCUs via its simple single-bus interface. Compact and energy-efficient, the DHT22 offers a reliable solution for smart farming, weather stations, and indoor climate control systems. The units of measurement for temperature are degrees Celsius (C), while for humidity, they are expressed as a percentage (% relative humidity) which is a measure of the amount of water vapor in the air compared to the maximum amount the air can hold at a given temperature, expressed as a percentage.

4.6. Light & UV sensor

The Gravity LTR390UV-01 light sensor is a versatile and high-performance sensor designed for ambient light measurement within the wavelength range of 280–430 nm and UV radiation detection in the 450–700 nm range. Featuring both I2C and UART communication interfaces, it is easy to integrate with Arduino MCU. The sensor provides high precision and fast response, making it ideal for monitoring UV radiation and light intensity in applications like environmental monitoring, smart agriculture, and wearable devices. Based on the data sheet of the LTR390UV-01 Light Sensor[1] it measures light in lux, while measures UV levels using a unitless UV index (UVI) scale, derived from the sensor's digital output and a calibration factor called UV sensitivity. Specifically, the UVI is calculated as given by Eq. (2).

$$UVI_{Calc} = \frac{UVSensorCount}{UVSensitivity} \times W_{FAC} \tag{2}$$

where $WFAC$ (Window Factor) depends on the type of material covering the sensor (e.g., no window = 1) and the $UVsensitivity = 2300$.

[1] https://optoelectronics.liteon.com/upload/download/DS86-2015-0004/LTR-390UV_Final_%20DS_V1%201.pdf.

4.7. Real time clock

Enhancing the accuracy and reliability of the monitoring system was achieved by integrating the Gravity I2C DS1307 real-time-clock (RTC) module into the physical stations. The RTC ensures precise timekeeping, which is critical for off-grid deployments, enabling synchronization of data across all stations. This synchronization facilitates cross-referencing and validation of measurements, paving the way for real-time monitoring and dataset construction. Each physical station was configured with a three-minute counting interval, determined experimentally to balance memory usage and data granularity. This interval was sufficient for capturing meaningful changes in parameters like soil moisture, which typically evolve over several hours, thereby optimizing resource use while maintaining data integrity.

Furthermore this study implemented a data persistence function alongside periodic synchronization and communication among physical stations. In this configuration sensor data are saved to our SD card every hour to protect against data loss in the event of power failures or system malfunctions.

4.8. Battery & solar panel

Each physical station is powered by a single 18 650 cylindrical lithium-polymer (Li-Po) battery with a nominal voltage of 3.6 V and a capacity of 3400 mAh. To ensure continuous operation in off-grid environments, the battery is recharged via a Waveshare 6 W solar panel, enabling sustainable energy supply. Based on simulations conducted using Altium, the battery-panel combination can sustain the MCU for a minimum of six years before experiencing significant degradation [36].

The battery, charger, and panel combination was selected following an experimental evaluation in which the station's hardware was powered via a regulated supply, and the average current consumption was recorded. Notably, the data acquisition frequency was initially set to 1.5 min, which is twice the final resolution, thereby representing a higher-than-normal power demand. The measured average current consumption was 60 mA, and accounting for the buck-boost converter efficiency, which regulates power to the Arduino Nano, we determined that the station could remain operational for approximately 54 h without recharging.

Further analysis assessed the solar panel's power output under varying weather conditions. On a sunny day, the panel delivered a peak power output of 5 W, supplying ~ 850 mA to charge the battery while simultaneously powering the Arduino at ~ 60 mA. Under cloudy conditions, the panel produced 2–3 W, which, while lower, was still sufficient to prioritize battery charging when excess power was unavailable. Moreover, we calculated that the battery would need 3.4 h to fully recharge, which corresponds with our experiment where the battery charger in a span of 4.2 h, in sunny weather and in a span of 7 h in cloudy weather.

According to the battery manufacturer's specifications, the Li-Po battery maintains at least 500 charge–discharge cycles before its capacity significantly depletes. Given the observed charging and discharging currents, the battery capacity must fall below 900 mAh (37% health) before it fails to sustain overnight operation or prolonged cloudy periods. Extrapolating from current power demands and charge rates, the system is expected to remain operational for at least six years before requiring a battery replacement.

4.9. Calibration of sensors

Before initiating our experiments and deploying the stations, a series of calibrations were performed on the sensors to ensure accurate measurements. Firstly, the RTC was synchronized with the current date and time. In accordance to the manufacturers datasheet recommendations, a one-week experiment was conducted to measure the clock's drift. Based on the observed drift, a correction function was developed and programmed to run continuously, with weekly adjustments applied to compensate for the known drift.

Additionally, the LTR390 Light and UV sensor required calibration according to the datasheet specifications. Using the SPI protocol, two registers where configured using the SPI protocol to set the UV and ambient light readings under controlled conditions in a completely dark room

An optional calibration procedure was also undertaken for the soil moisture sensor. Measurements were recorded in two distinct environments: completely dry and fully submerged in water. These readings were used to map the raw ADC values to a percentage scale using standard Arduino library functions. This approach was implemented to improve interpretability and standardize the output.

4.10. Accuracy of sensors

The accuracy of physical sensors plays a vital role in the overall performance of a smart farming system, especially in the effectiveness of virtual sensors that depend on this data. Our approach focuses on using affordable, widely available sensor hardware to reduce costs and promote broader adoption of precision agriculture technologies among farmers. The sensors used in our system offer the following levels of accuracy: the DFRobot Capacitive Soil Moisture Sensor [37] has an accuracy of $\pm 4.2\%$, the rain detection sensor [38] $\pm 10\%$, the LTR390UV-01 light and UV sensor [39] $\pm 1\%$, and the DHT22 Temperature & Humidity sensor [40] maintains an accuracy of $\pm 0.5\%$ under harsh conditions and $\pm 0.1\%$ in normal environments.

Furthermore, although the sensors used in our system are low-cost, they provide sufficient reliability for a wide range of agricultural applications—particularly when combined with AI-based virtual sensors that can mitigate or correct minor inaccuracies over time. Machine learning models, especially those designed for time-series data in agriculture, have been shown to tolerate moderate levels of noise without significant performance degradation. For example, a 5% reduction in sensor data accuracy led

to only a 1.8% drop in crop yield prediction accuracy [41]. Similarly, increasing data accuracy from 95% to 99% resulted in just a 0.5% improvement in model performance for climate-based agricultural forecasts [42]. These findings support our decision to prioritize cost-effectiveness and scalability, especially in light of recent advancements in noise-resistant, low-cost sensors that offer a practical and reliable solution for large-scale deployment [43].

4.11. Long-term reliability measures in harsh environmental conditions

Ensuring the long-term reliability of virtual sensors in smart farming requires addressing two critical challenges: physical degradation of hardware components and value drift in sensor readings over time. To mitigate these risks, a combination of structural protections, protective coatings, and calibration strategies was implemented. These measures include sealing mission-critical components in IP65-rated enclosures,[2] applying conformal coatings to exposed electronics, using custom-designed protective casings, and performing initial and periodic sensor calibrations where necessary. Together, these protective and maintenance strategies enhance the durability and accuracy of the virtual sensor system, supporting consistent performance even in harsh and dynamic environmental conditions common in agricultural settings.

4.11.1. Degradation management

The physical degradation of sensor nodes, particularly in outdoor agricultural settings, is largely influenced by environmental exposure. Mission-critical components such as the MCU, charging board, and battery were securely housed within sealed IP65-rated electrical enclosures, as shown on the right of Fig. 4. This enclosure rating provides protection against dust ingress and resistance to water jets, effectively shielding internal electronics from moisture, debris, and other environmental stressors. Post-experiment inspections revealed no signs of corrosion or physical degradation, validating the effectiveness of this approach for medium-term deployments.

For components exposed to environmental conditions, including the soil moisture sensor, temperature and humidity sensor, and light and UV sensor, additional protective measures were necessary. A two-layer application of 419D acrylic conformal coating[3] was applied to exposed circuitry before deployment. This coating served to isolate sensitive surfaces from humidity and airborne contaminants, significantly reducing oxidation and mechanical wear. While certain elements such as connectors and switches could not be coated due to functional constraints, custom-designed 3D-printed protective casings were used to shield the temperature, humidity, and UV sensors without compromising measurement accuracy. For the soil moisture sensor, an enclosure was intentionally avoided based on preliminary testing, which showed that enclosed designs could trap water and accelerate connector degradation. Instead, a thicker layer of conformal coating was applied to ensure long-term durability during soil contact. For extended deployments, additional strategies such as selective conformal coating using masking techniques, the use of ruggedized connectors, and the inclusion of humidity-absorbing materials within enclosures can further enhance long-term protection and reliability.

4.11.2. Value drift management

In addition to hardware durability, maintaining sensor accuracy over time is essential. All sensors requiring calibration were calibrated prior to deployment, as described in Section 4.9. Components such as the RTC, which are known to exhibit drift over extended periods, are supported by software-based periodic recalibration functions. Other sensors, including the soil moisture and light/UV sensors, generally do not experience significant drift and only require an initial calibration to establish reference maximum values.

5. Regression models for virtual sensors

Regression analysis is used to predict one or more continuous dependent variables based on the values of one or more input variables. Regression models can be trained to map independent variables from the physical station, such as temperature and humidity, to the outputs of the virtual stations. The mapping between the independent variables (features) and the dependent variables (targets) is unknown, necessitating model learning to uncover underlying patterns. Factors such as geographical orientation, soil properties, vegetation cover, irrigation practices, microclimatic variations and exposure to sunlight or wind can influence the connection between the temperature and humidity of the reference and target spaces. These factors make it challenging to represent these relationships with a numerical equation. Consequently, nonlinear or data-driven approaches, including black-box models, are often preferred. The primary categories of ML and statistical regression models suitable for virtual sensors are detailed in the following subsections. These models will be applied and evaluated in Section 7 to identify the most effective prediction model.

[2] https://www.polycase.com/ip65-enclosures.

[3] https://www.e-praud.eu/conformal_coatings/acrylic_conformal_coatings/419d_mg_chemicals.

5.1. Linear regression

Linear regression (LR) models are among the most widely used techniques in ML and statistical modeling, particularly for regression tasks. These models operate on the principle that a linear equation can reasonably approximate the relationship between independent variables and dependent variables [44]. Their simplicity, interpretability, and efficiency make them foundational methods in many applications.

LR models assume that the response variable y can be expressed as a linear combination of the terms in the vector of independent variables, $X = [x_1, x_2, \ldots, x_n]$. The model parameters are typically estimated using methods like Ordinary Least Squares, which minimizes the sum of squared deviations between observed and predicted values. One of the key advantages of linear regression is its simplicity, offering a clear and interpretable relationship between variables. This makes it particularly valuable in scenarios where understanding and explaining the model's behavior is essential [45]. Additionally, linear models are computationally efficient, even when applied to large datasets.

5.2. Support vector regression

Support Vector Regression (SVR) is a ML technique designed for regression tasks, which involves predicting continuous values from input data [46]. A key strength of SVR is its ability to handle sparse data and efficiently address nonlinear problems. It works by identifying a regression function that generalizes well to the training data while minimizing the deviation between actual and predicted values. This is achieved by optimizing a symmetric loss function, which equally penalizes both large and small deviations.

For data that is not linearly separable, SVR leverages kernel functions to map the data into higher-dimensional spaces. This enables the model to handle nonlinear relationships effectively, providing flexibility for diverse types of data [47]. Additionally, it performs particularly well when the dataset is small compared to the number of input variables. Despite its advantages, SVR's performance is highly sensitive to the selection of hyperparameters, such as the kernel type and its parameters. This often necessitates careful tuning and extensive testing to achieve optimal results.

5.3. Bootstrap aggregating

Bootstrap Aggregating, or Bagging, is an ensemble learning technique that enhances the accuracy and stability of ML models by combining predictions from multiple models trained on different bootstrap samples—random subsets of the dataset created through sampling with replacement [48]. Bagging reduces variance and mitigates overfitting by leveraging diversity among models, making it effective for high-variance algorithms like decision trees. Random Forest (RF), a popular bagging method, builds multiple decision trees independently on different bootstrap samples and aggregates their predictions, improving robustness and reducing overfitting.

5.4. Boosting

Boosting is a ML technique aimed at improving the accuracy of predictive models. Similar to Bootstrap Aggregating, it operates on variations of the historical dataset. However, the key distinction lies in how these variations are generated. In Bagging, datasets are created by sampling with a uniform distribution, while in Boosting, data points with higher prediction errors are more likely to be sampled than those predicted correctly.

Boosting involves iterative phases of dataset creation and base model training [49]. Initially, all data points have an equal probability of being sampled, and the first base model is trained and evaluated on this dataset. For subsequent iterations, new datasets are generated where the sampling probability is weighted by the prediction errors of the previous model. This process of dataset generation, model training, and evaluation continues until the error rate falls below a predefined threshold or a specified number of base models are added to the ensemble.

One of the most well-known Boosting algorithms is AdaBoost. It works by training a base model on an initial dataset, then creating new datasets based on the prediction errors of the model, and iteratively refining accuracy through additional training cycles [50]. While highly effective, it is important to note that Boosting models, including AdaBoost, come with increased computational complexity and should be applied cautiously, particularly in resource-constrained environments.

Other popular Boosting algorithms include LightGBM (LGBM), XGBoost (XGB), and Histogram-Based Gradient Boosting (HGB). LGBM is optimized for speed and efficiency, using a histogram-based approach to split data and handling large datasets with lower memory consumption [51]. XGBoost, a highly scalable and flexible algorithm, employs advanced regularization techniques to prevent overfitting and has become a standard in many competitive ML tasks [52]. HGB, leverages histogram-based splitting to improve training efficiency and is particularly well-suited for large datasets [53]. These modern algorithms enhance Boosting by addressing computational challenges, making them more practical for real-world applications, though they still require careful tuning and consideration of resource limitations.

5.5. Feedforward neural networks

Regression Feedforward Neural Networks (FNN) are powerful tools in ML for predicting continuous dependent variables. Their core strength lies in the ability to represent and transform input data through a network of interconnected layers to model complex nonlinear relationships [54]. FNN utilize a hierarchical structure comprising three main types of layers: input, hidden, and output. Input data is processed through multiple hidden layers, each consisting of neurons that apply nonlinear activation functions such as the Rectified Linear Unit (ReLU) or softmax. These functions enable the network to capture intricate patterns in the data, facilitating accurate predictions.

Table 2
Overview of features and their respective descriptions.

ID	Feature	Description
t1	Timestamp	This feature represents, as the name implies, the timestamp of each and every measurement in a unix time format. This is recorded with a high precision RTC clock.
f1	Temperature	The temperature of the environment, measured from inside a waterproof weather station 3D printed box with appropriate venting. This is done in order to avoid false temperature readings when the sun is hitting it all day.
f2	Humidity	Measures the humidity of the air, this sensor is on the same board as the temperature sensor.
f3	Rain	This feature is the raw measurements from the rain sensor. By raw it is meant that the resistance value is read directly from the control module of the sensor and stored without further processing.
f4	Light	This feature represents the raw values recorded from the ambient light sensor, raw is the value recorded by the sensor in lux.
f5	UV	This sensor is in the same board as the light sensor and is located underneath a glass windows at the top of our station. These feature measurement unit is milliwatts per square centimetre.
f6	Soil moisture	This is a crucial feature for many applications this dataset can be used for and measures the humidity of the soil, at a depth of 15 centimetres. A capacitive sensor measures the dielectric constant of the soil. When the soil is irrigated, it becomes wetter and more conductive. This increase in conductivity can be used to scale the measurements from 0 to 100.

5.6. Time series neural networks

Time Series Neural Networks are specialized architectures designed to process and analyze sequential data, where observations are dependent on time. These networks aim to capture temporal patterns, trends, and dependencies in data, enabling tasks such as predicting sensor measurements for smart farming applications, including soil humidity, light intensity, and UV levels. Time Series Feedforward Neural Networks (TSFNN) rely on engineered lagged features, using fixed windows of past data as input but lack inherent sequential memory, making them suitable for simpler time series problems [55]. Time Series Recurrent Neural Networks (TSRNN) address this limitation by introducing loops within their architecture, allowing information to persist across time steps, thus effectively modeling sequential dependencies [56]. Time Series Long Short-Term Memory (TSLSTM) networks, an advanced form of TSRNNs, incorporates specialized gating mechanisms to handle long-term dependencies and mitigate the vanishing gradient problem, enabling them to excel in complex time series tasks where patterns span over long durations [57].

6. Exploratory data analysis

In this section, the structure and characteristics of the gathered sensor measurements are described. The dataset was collected from IoT-enabled monitoring stations deployed across the two agricultural fields depicted in Fig. 2: a larger field equipped with three physical stations and a smaller field with one operational physical station, as the second station became non-functional after the flood in Anonymized city. These data offer crucial insights into the environmental conditions of the fields, serving as the foundation for our analysis and modeling. After the dataset collection, missing values and outliers were inspected, while the timestamps and the values of the different sensors were cross checked, in order to determine the validity of the recorded data. Table 2 presents the features of the dataset along with their corresponding descriptions.

Each feature derived from a physical sensor at a distinct physical station has its own unique distribution, characterized by a set of different statistical properties. Time series plots and violin plots where used to observe the distribution of values. The time series plots visually represent how a variable changes over time, revealing trends, patterns, and fluctuations in the data. The shape of the violin plots illustrates the density of the data and revealing areas of concentration or spread. In the middle of each violin plot is a small box plot, where the rectangle marks the first and third quartiles, and the central dot or line represents the median.

The analysis begins by examining and analyzing the data collected from physical station 1 in agricultural field 1 (Section 6.1). Next, the data from physical stations 2, 3, and 4 in agricultural field 2 are briefly presented (Section 6.2). Following this, feature correlations within the same field (Section 6.3) and across different fields (Section 6.4) are analyzed. This analysis includes visualizing and interpreting the data to uncover significant patterns and relationships.

6.1. Physical station in agricultural field 1

Agricultural field 1 is the smallest of the experimental sites, spanning over 80 acres. It includes physical stations 1 and 5, though physical station 5 was destroyed by the natural disaster in Anonymized city. Fig. 5 displays the violin plots for temperature, humidity, light, UV, and soil moisture recorded at the first physical station. The plots highlight that each feature has a distinct distribution, varying in skewness and deviation from the median. The four time series plots in Fig. 6 include the segments of the dataset that

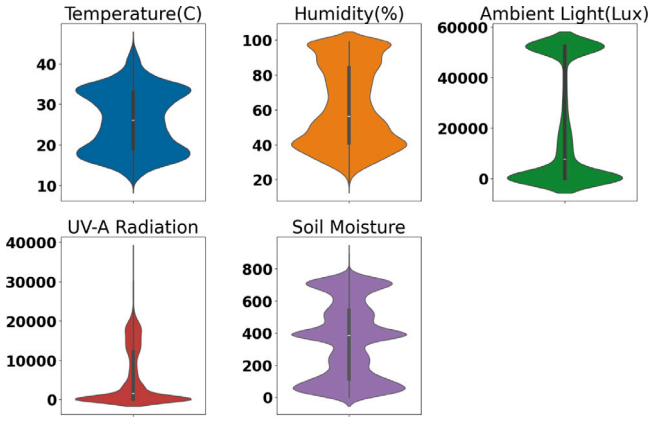

Fig. 5. Violin plots showing the distributions of features measured at physical station 1.

comes from four distinct irrigation periods. The x axis represents various timestamps (in a day-month-year and hour format), while the y axis shows the values of the soil moisture (blue), light intensity (red) and humidity (green).

A consistent pattern is observable across the time series plots. During the day, the air humidity tends to decrease, as expected, while at night it increases. This daily fluctuation is further amplified during irrigation periods, where atmospheric humidity levels are noticeably higher. Focusing on soil moisture, an interesting observation arises, particularly in the second and final time series plot: the soil moisture levels do not increase proportionally or immediately after irrigation starts. Instead, a delay of several hours is observed before any significant change occurs. This delay is attributed to the sensor's placement deep and near the plant's root zone, where it takes time for water from the irrigation system to reach.

It is important to note that the rain detector sensor operates by measuring changes in electrical resistance. When water comes into contact with the sensor's surface, it alters the module's resistance, which is then used to determine the presence and intensity of rain. A low resistance output typically indicates rain detection; however, this can result in false detections, as water may contact the sensor surface for reasons unrelated to rainfall. This issue can be mitigated by applying a Moving Average (MAVG) to the sensor's data points. As shown in Fig. 7, individual instances of low resistance often represent false rain signals, but using MAVG allows for more accurate detection of actual rainfall events.

Furthermore, natural rainfall can disentangle from irrigation events by using MAVG, as the consistent patterns associated with irrigation are more easily differentiated from the irregular and intermittent characteristics of rainfall. This type of distinctions is crucial for accurately analyzing water usage and can be used for optimizing irrigation strategies.

Fig. 8 presents the soil moisture, the temperature and the light over two irrigation periods and two heavy rainfalls, spanning approximately two weeks. It is observed that during the first irrigation period, there is high soil moisture. This matches expectations because farmers aim to keep the soil and plants with high moisture during this season. This trend persists throughout the dataset. At the start of the plot, a fluctuation in the light sensor values is observed indicating the presence of dense clouds obstructing daylight. These readings are corroborated by the rain sensor data, which confirms that the cloud cover was accompanied by rainfall.

During the first irrigation period, rainfall preceded the start of watering, as marked by the horizontal lines. The watering lasted approximately one day, and the combined effect of the rain and irrigation kept the soil at consistently high moisture levels. In the second irrigation period, the soil was notably dry before watering began. As irrigation commenced, a gradual increase in soil moisture was observed, reaching its peak at the second blue line, which indicates the end of the watering period.

Fig. 9 provides a more detailed view of the second watering period. The blue line represents the soil time series, showing a gradual increase in moisture levels after the start of irrigation, which aligns with expectations. While the air humidity decreases when the irrigation closes to the end and the light increases.

6.2. Physical stations in agricultural field 2

Agricultural field 2 is the largest of the experimental sites, covering over 160 acres and includes physical stations 2, 3, and 4. To evaluate whether all sensors record similar feature distributions, their violin plots are presented. As shown in Fig. 10, the violin plots reveal substantial differences compared to the violin plots of agricultural field 1 of Fig. 5. This indicates that the sensors measure significantly varied feature distributions. Furthermore, for the soil moisture, its noteworthy that the values varies significantly among the physical stations 2, 3 and 4 of the field 2.

Additionally, a rise in humidity with minor fluctuations was detected at station 3. Further investigation revealed that a small leak in the irrigation system persisted throughout the season, as confirmed by the responsible farmer. This finding underscores how sensor data and exploratory analysis can provide valuable insights into the impacts of both natural phenomena and human activities in smart farming.

For the sake of brevity, time plots and analysis for stations 2, 3 and 4 are not repeated in this article. The corresponding figures can be found in the GitHub repository of the article's first author.

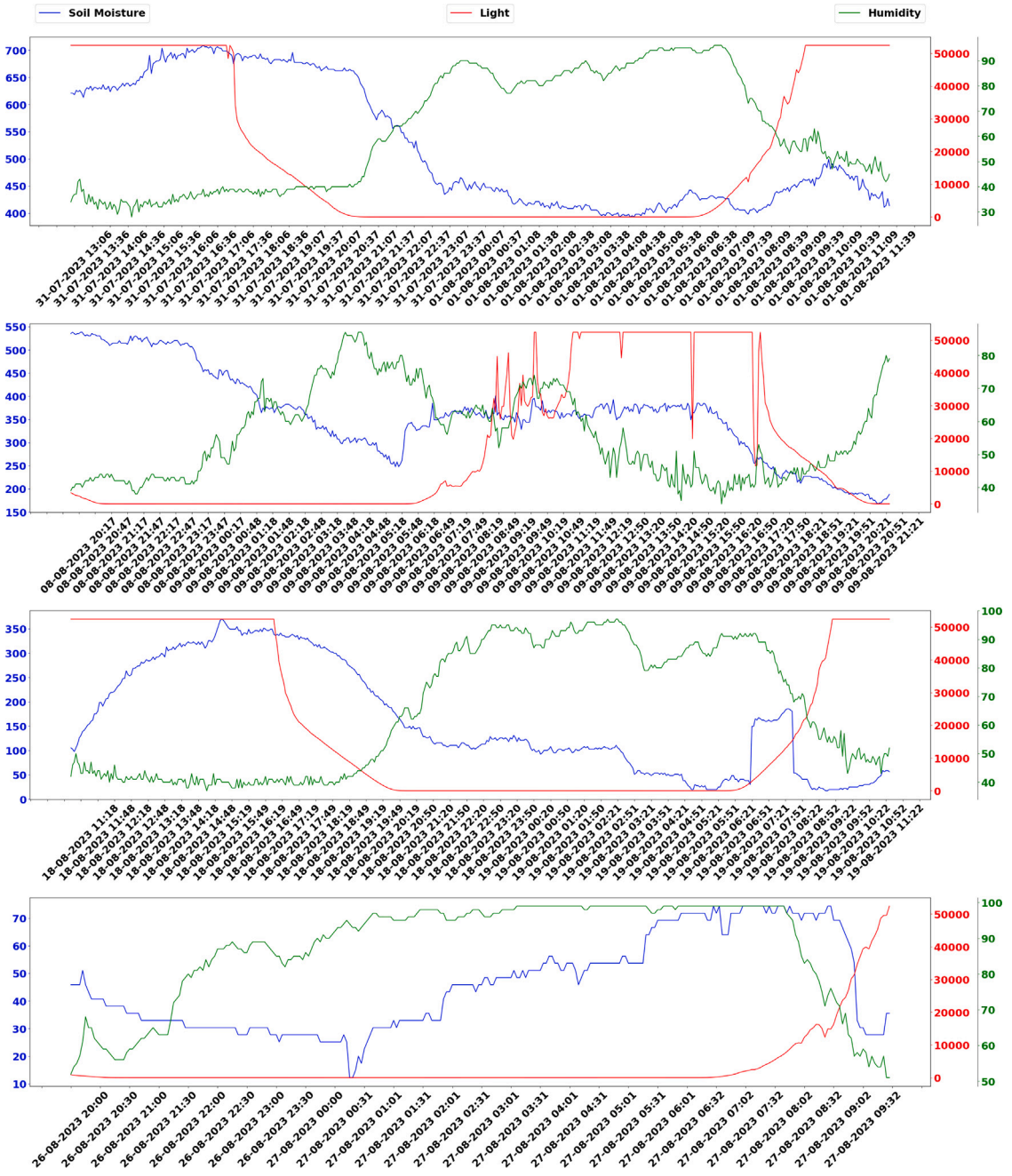

Fig. 6. Time series plots depicting soil moisture, air humidity, and light across four distinct irrigation periods.

6.3. Inner-field feature correlation

This subsection analyzes the correlation of features across the features of the physical stations 2, 3, and 4 within the agricultural field 2. The resulting correlation heat-map is displayed in Fig. 11. The value of correlation coefficient lies between −1 and 1. If there is no correlation between the features f_i and f_j then $\rho(f_i, f_j) = 0$. A perfect negative correlation is found if $\rho(f_i, f_j) = -1$ and a perfect positive correlation is found if $\rho(f_i, f_j) = 1$.

The results align with expectations, showing a high positive correlation among the features of the same type of sensors. Furthermore, there is a strong positive correlation between temperature and light alongside a strong negative correlation between temperature and humidity. Temperature and light are strongly positively correlated because higher light intensity directly heats

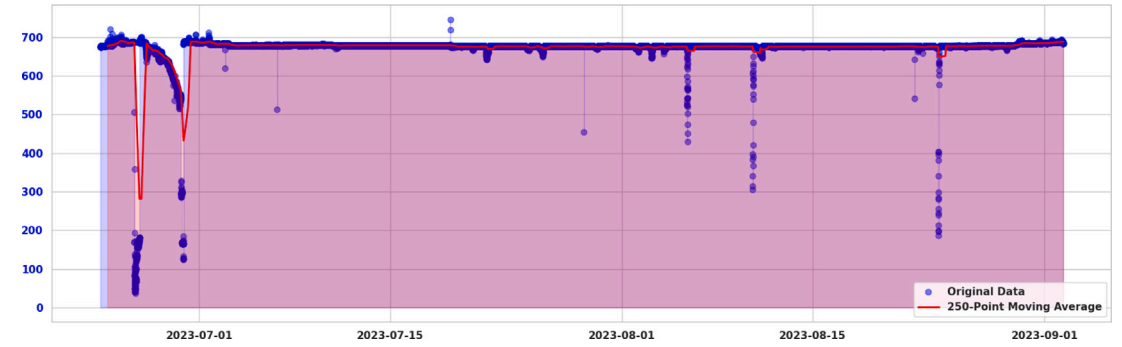

Fig. 7. Moving average of data points from rain detection sensors over time.

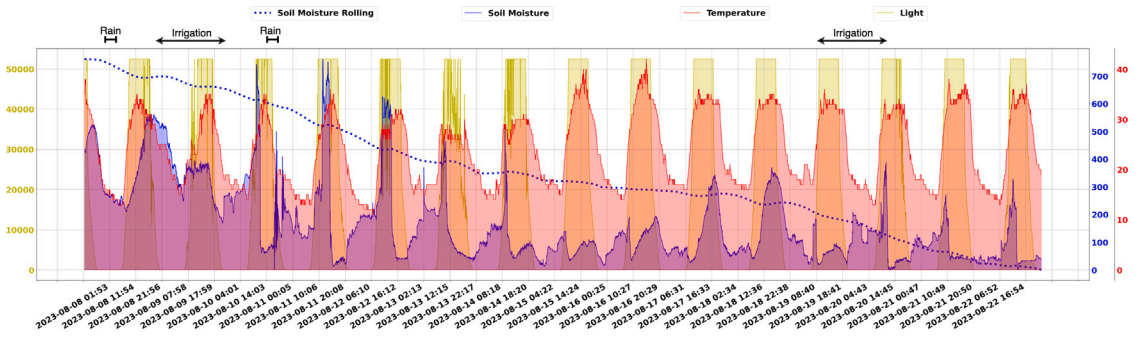

Fig. 8. Temperature, Soil humidity, Ambient light during heavy rainfall and Watering intervals.

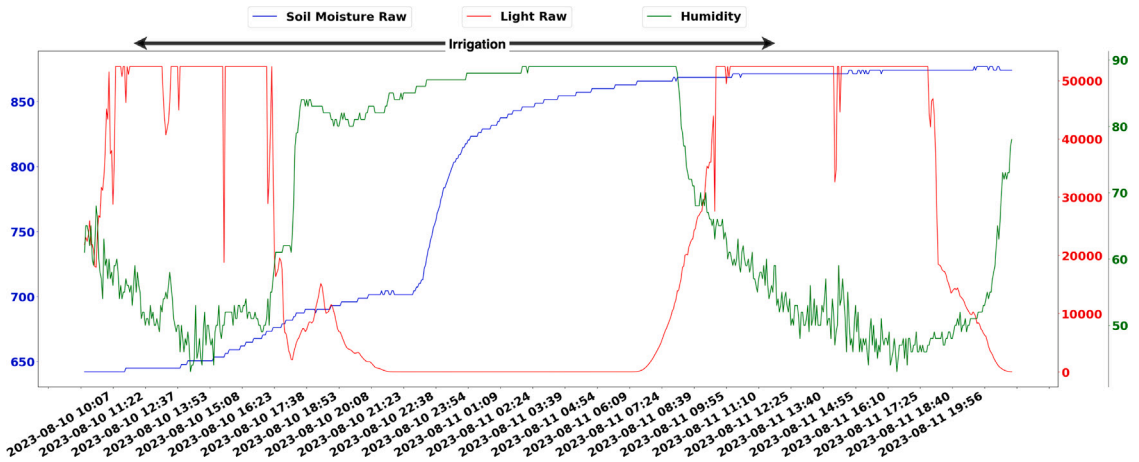

Fig. 9. Second day of irrigation.

surfaces and the air, while temperature and humidity are strongly negatively correlated because higher temperatures enhance evaporation, reducing relative humidity by increasing the air's capacity to hold moisture.

It is worth noting that the soil moisture readings from physical station 2 show almost no correlation with those from the other two stations. Upon investigation, it was discovered that this discrepancy is due to the irrigation schedule; the field monitored by the physical station 2 was irrigated only after the areas covered by the other physical stations had completed their irrigation cycles.

Additionally, weak yet noteworthy correlations where identified between soil moisture, temperature, and humidity. These relationships indicate that as soil moisture increases, both temperature and humidity tend to rise as well. Lastly, there is a positive correlation between soil moisture and atmospheric humidity for stations 2 and 3, while station 4 exhibits a negative correlation with the other stations.

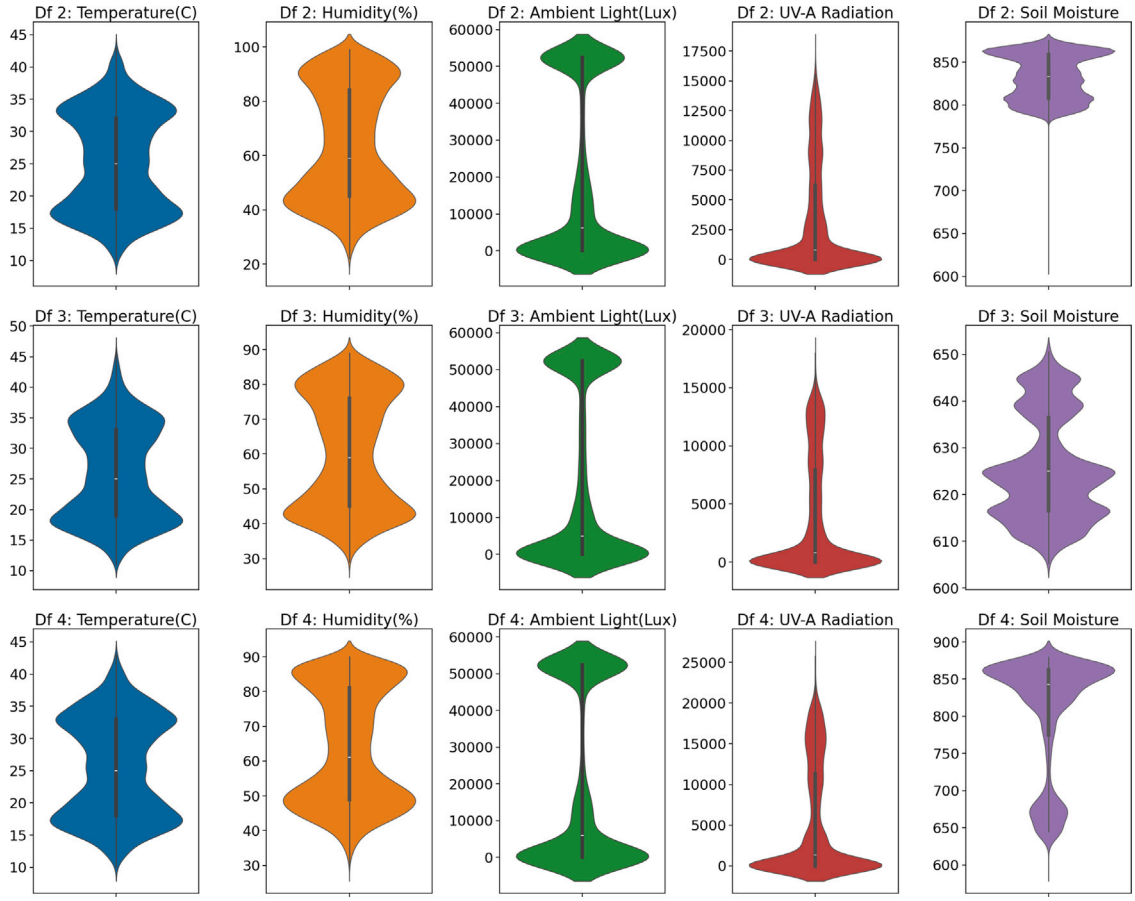

Fig. 10. Violin plots of the features from physical stations 2, 3, and 4 located in agricultural field 2.

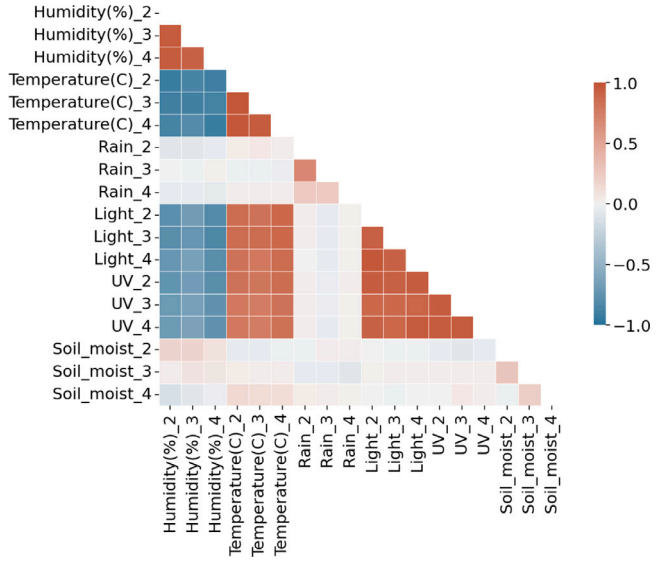

Fig. 11. Correlation Heat-map between physical sensors in the same field.

6.4. Cross-field feature correlation

The correlation between stations located in different geographical areas is examined and presented in Fig. 12. As shown in the correlation plot, the features exhibit similar relationships across fields separated by a distance of 3 km. This finding demonstrates

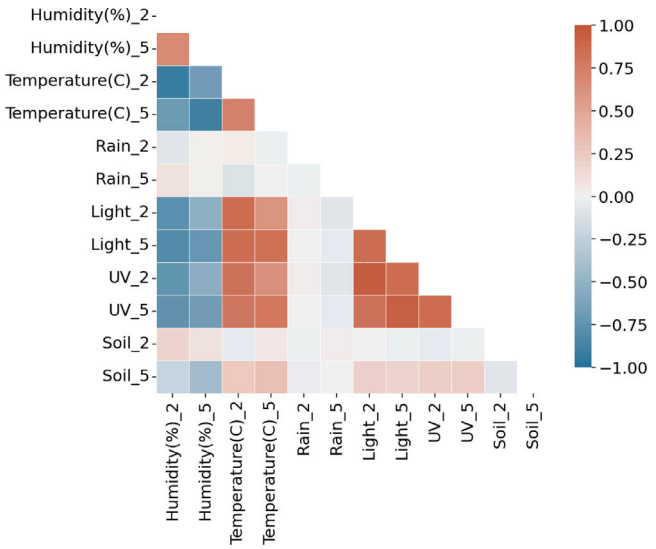

Fig. 12. Correlation heat-map of physical sensors across different fields.

that the underlying patterns can be captured effectively, allowing us to model not only nearby fields but also those further apart. This reinforces the applicability of the virtual sensor methodology across diverse geographical locations, without being limited by proximity.

7. Experimental evaluation

This section describes the experimental process that was followed to evaluate the performance of the virtual sensors and compare the various ML models they utilize. To accomplish the experimental evaluation the physical stations where deployed as explained in Section 4, and the models where trained as described in Section 5 using the dataset as outlined in Section 6. From the dataset 80% of the observations where used from each physical station to train the models and the remaining 20% for evaluation. Furthermore, experiments were carried out in which the virtual stations used open weather data as inputs instead of relying on the base station. The dataset collected from the physical stations, along with the associated code, is publicly available on the first author's GitHub repository [58].

7.1. Natural disaster during our experiment

A notable event during our experiment was the catastrophic flooding that occurred in the Anonymized region, which caused extensive damage across multiple sectors, including agriculture and livestock [9]. Many farmers not only lost their annual yields but also faced the destruction of costly equipment, including advanced smart farming systems. Our stations were partially affected by the disaster. Four out of the five deployed stations were working successfully. Unfortunately, the fifth station, located in a low-lying area, was submerged and could not be recovered. In the smaller field, only one station was operational, whereas all three stations in the larger field were functioning successfully.

7.2. Evaluation metrics

To evaluate the performance of the prediction models, three metrics are employed: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Hit Rate, with thresholds set at 0.5%, 1%, or 5%, depending on the specific feature. These metrics enabled us to evaluate the performance of each model and identify the one that delivered the best results. MSE measures the average of the squared differences between predicted and actual values. The goal is to minimize MSE, with 0 indicating a perfect model. MAE calculates the average of the absolute differences between predicted and actual values. The Hit Rate metric evaluates the percentage of predictions made by the ML model that falls within a specified error threshold. For instance, stating that the Hit Rate (1%) for Temperature using the LGBM method is 90.8% means that 90.8% of the predicted values have an error of less than 1%. The features vary in range and significance to farmers. Therefore, after consulting with them, specific thresholds where established, tailored to each feature.

Table 3
MAE and MSE metrics for virtual stations 1, 3, and 4.

Method	Station 1		Station 3		Station 4	
	MAE	MSE	MAE	MSE	MAE	MSE
LGBM	**0.089**	**0.021**	0.079	**0.022**	**0.064**	**0.017**
XGB	0.465	0.945	0.089	0.029	0.329	0.594
AdaBoost	0.535	0.935	0.087	0.037	0.352	0.621
HGB	0.734	0.945	0.088	0.029	0.331	0.464
LR	0.223	0.033	0.093	0.032	0.078	0.017
RF	0.123	0.032	0.103	0.029	0.086	0.017
SVR	0.097	0.028	0.094	0.035	0.074	0.018
FNN	0.421	0.315	0.457	0.363	0.368	0.377
TSFNN	0.262	0.366	0.285	0.323	0.298	0.139
TSRNN	0.018	0.092	0.088	0.021	0.073	0.011
TSLSTM	0.118	0.027	**0.071**	0.023	0.289	0.061

Table 4
Hit rate prediction of virtual station 1.

Method	Temp. (1%)	Hum. (5%)	Rain (0.5%)	Light (5%)	UV (5%)	Moist. (1%)
LGBM	**90.8**	**88.8**	99.8	**78.9**	**66.9**	**83.9**
XGB	64.3	44.6	93.4	15.5	5.5	12.4
AdaBoost	46.2	24.7	97.2	1.7	3.3	7.5
HGB	18.7	16.2	97.2	7.7	3.4	3.2
LR	46.1	39.1	98.9	8.7	2.9	2.3
RF	38.9	21.8	**99.9**	0.8	3.4	2.9
SVR	60.3	45.3	**99.9**	19.5	3.2	5.4
FNN	65.7	73.4	42.5	12.9	14.9	2.5
TSFNN	45.4	55.5	25.6	18.2	9.5	7.9
TSRNN	38.0	52.4	40.8	43.6	10.2	5.9
TSLSTM	49.8	57.9	45.7	37.8	28.9	47.9

Table 5
Hit rate prediction of virtual station 3.

Method	Temp. (1%)	Hum. (5%)	Rain (0.5%)	Light (5%)	UV (5%)	Moist. (1%)
LGBM	88.2	**93.2**	97.8	**84.3**	**85.4**	**96.4**
XGB	79.8	67.7	87.8	63.2	58.3	45.6
AdaBoost	76.5	65.4	95.6	43.2	36.5	88.7
HGB	81.2	64.5	87.4	54.9	41.5	89.4
LR	**88.9**	67.4	98.8	58.9	61.3	90.3
RF	56.7	33.4	89.9	34.3	23.4	80.9
SVR	84.5	56.4	**99.8**	55.6	45.6	88.9
FNN	86.9	62.9	79.1	45.4	43.6	76.9
TSFNN	63.8	71.2	89.9	24.5	60.8	70.9
TSRNN	47.5	67.4	86.4	35.4	57.6	65.3
TSLSTM	44.5	76.4	72.3	43.8	22.4	78.9

7.3. Outcomes & discussion

The performance of the virtual sensors is showcased following four different approaches. In Section 7.3.1, the use of ML models is evaluated to predict measurements for multiple sensors of a single virtual station. In Section 7.3.2, the performance of predicting virtual sensor measurements using a single-output methodology is examined. Section 7.3.3 details experiments involving a multi-output ML model to simultaneously predict sensor measurements for multiple physical stations. Lastly, Section 7.3.4 explores the feasibility of using satellite measurements as inputs for modeling a virtual station.

7.3.1. Multi-output modeling for a single virtual station using a physical base station

In this approach, the physical station 2 was set as the base station and use its features to predict the values at stations 1, 3, and 4 based on separate multi-output models. Each multi-output model represents a separate virtual station, and each output corresponds to the metrics of a virtual sensor. Table 3 provides a comparison of performance metrics with MAE and MSE aggregated for all features, across all models. The first double column represents models trained and evaluated to predict station 1, while the latter columns correspond to models trained and evaluated to predict station 3 and 4. The evaluation reveals that LGBM has significantly better evaluation outcomes. This occurs due to its leaf-wise tree growth strategy, which effectively minimizes loss and captures complex patterns in the data with high computational efficiency. Extensive experimental research was conducted in order to explore

Table 6
Hit rate prediction of virtual station 4.

Method	Temp. (1%)	Hum. (5%)	Rain (0.5%)	Light (5%)	UV (5%)	Moist. (1%)
LGBM	**99.9**	**97.7**	**99.2**	**85.1**	**89.5**	91.4
XGB	76.6	66.7	89.7	22.3	55.3	76.5
AdaBoost	71.6	53.2	97.5	4.3	53.2	88.2
HGB	75.2	52.7	97.1	21.9	53.4	79.8
LR	79.5	60.1	37.3	18.2	9.8	58.4
RF	54.3	58.8	**99.2**	18.5	7.7	43.5
SVR	79.3	63.5	88.6	23.7	8.4	53.4
FNN	78.5	70.9	98.6	16.2	61.9	82.2
TSFNN	50.3	70.4	56.5	20.4	63.6	**98.7**
TSRNN	38.1	60.3	70.4	55.6	28.6	75.2
TSLSTM	46.7	73.4	72.5	39.8	20.5	80.0

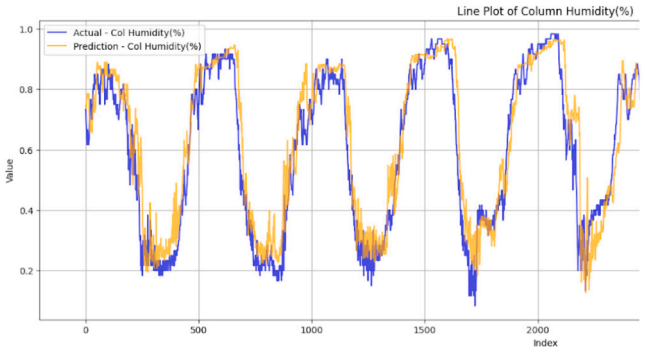

Fig. 13. Actual (blue line) and predicted (green line) humidity of base station 1 with multi-output model.

different topologies for time-series models, with a focus on LSTM, utilizing Bayesian optimization to fine-tune hyperparameters. From these experiments we concluded that LGBM using feature vectors outperforms time series methods like LSTM. LGBM efficiently captures complex feature interactions and handling high-dimensional data, making it less reliant on sequential dependencies and more robust to short-term noise.

The values in Table 3 represent aggregated errors, which do not provide stakeholders with a clear understanding of the prediction accuracy. Therefore, this study presents the Hit Rate in the subsequent tables for better interpretability. Table 4 summarizes the outcomes for the features of the virtual sensor 1, Table 5 summarizes the outcomes for the features of the virtual sensor 3, and Table 6 summarizes the outcomes for the features of the virtual sensor 4.

Bold entries indicate the best performance for each feature, and it is evident that LGBM consistently outperforms other models in most cases. Even in scenarios where LGBM ranks second, such as rainfall at Station 1 or temperature at Station 2, its predictions are very close to those of the top-performing model. Furthermore, considering that Station 4 is located approximately 50 m from Station 2, while Station 1 is about 3 km away, the experimental results validate the effectiveness of virtual sensors in both neighboring and distant field locations.

Additionally, to provide a visual representation of the predictions, Fig. 13 illustrates the predicted humidity values at station 1 in orange alongside the actual values in blue. The model exhibited high accuracy, with predictions staying within a 2% margin of the actual values in most cases.

7.3.2. Single-output modeling for a single virtual station using a physical base station

In the previous experiments, a multi-output model was employed that uses data from all sensors at the base station to predict all the measurements of a virtual station simultaneously. In this section, a single-output approach is investigated, where a separate model is trained to predict a single variable, such as soil moisture, using inputs from multiple sensors. However, this approach proved less effective than the multi-output method. Fig. 14 illustrates the performance of the single-output model, showcasing its limitations in accurately predicting individual variables.

These outcomes made us to conclude that a multi-output model is more accurate than a single-output model. This occurs because the multi-output model leverages shared learning across outputs, capturing common patterns and relationships in the data that benefit all predictions. The shared representation allows the model to generalize better, as features learned for one target can inform others, especially when outputs are correlated or interdependent. Additionally, multi-output learning acts as a form of regularization, reducing overfitting by encouraging the model to balance multiple objectives rather than focusing narrowly on a single target. By utilizing all available information more efficiently, a multi-output model can achieve improved accuracy and robustness.

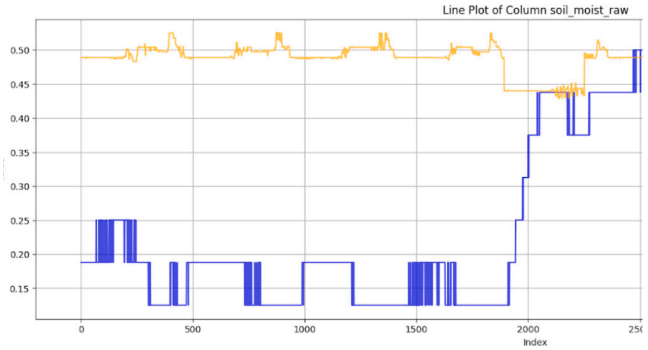

Fig. 14. Actual (blue line) and predicted humidity (green line) with single-output model.

Table 7
Hit rate predictions of an ML model concurrently outputting virtual stations 3 and 4.

Method	Station 3						Station 4					
	Temp. (1%)	Hum. (5%)	Rain (0.5%)	Light (5%)	UV (5%)	Moist. (1%)	Temp. (1%)	Hum. (5%)	Rain (0.5%)	Light (5%)	UV (5%)	Moist. (1%)
LGBM	89.5	**83.4**	90.3	27.6	17.5	35.1	52.1	43.2	74.1	21.9	9.1	7.3
XGB	85.8	78.9	89.9	**56.8**	43.8	73.4	76.6	67.5	63.7	22.5	**34.3**	94.3
AdaBoost	87.8	73.2	70.4	35.9	20.7	59.9	78.9	**70.9**	67.7	26.6	8.0	98
HGB	89.9	79.8	81.2	28.9	14.9	64.3	75.5	69.9	77.9	23.1	9.3	**98.1**
LR	**91.6**	73.5	19.8	37.8	**26.5**	53.2	**85.6**	68.8	8.5	20.8	9.5	96.3
RF	52.3	27.2	98.2	26.4	11.2	65.6	55.4	58.9	**98.9**	11.5	7.6	88.9
SVR	90.6	57.5	**98.5**	30.6	15.1	98	80.2	67.1	**98.9**	25.2	7.2	86.9
FNN	64.4	59.9	49.8	25.1	19.9	55.6	65.4	63.2	48.5	19.8	19.3	76.3
TSFNN	45.9	56.4	11.5	24.5	3.8	**97.2**	49.3	69.8	23.4	**34.3**	6.9	87.3
TSRNN	40.5	55.6	21.3	39.9	34	64.3	45.4	68.9	22.9	33.5	23.4	86.3
TSLSTM	56.7	39.6	21.6	27.4	15.2	79.8	60.5	53.3	68.9	26.4	25.5	97.8

7.3.3. Multi-output modeling for multiple virtual stations using a physical base station

The effectiveness of multi-output modeling inspired us to explore whether a multi-output model could predict the measurements for all virtual stations simultaneously. In these experiments, features from the base station were used as input, and features from all other stations were used as target outputs. Numerous experiments were conducted, testing various configurations and hyperparameters, but the results were not promising. Table 7 presents some evaluation outcomes, showing that while the model performs well on certain features, it struggles with others. These findings led us to conclude that employing separate multi-output prediction models for each virtual sensor is a more effective approach.

7.3.4. Multi-output modeling for a single virtual station using open data

In the final set of experiments, the use of weather observations from an open-data platform was explored as an alternative to measurements from the physical base station. This approach involved training the virtual sensors to model the relationship between open-data measurements and sensor data gathered from agricultural field stations. To acquire the open-data measurements, the Weather API provided by Visual Crossing [59] was used. This API provides access to both historical and forecast weather data on a global scale, sourced from over one hundred thousand worldwide stations, including satellite and maritime sources.

Visual Crossing offers access to a wide range of features. For our experiments, the following variables where selected as inputs: 1. Temperature, 2. Humidity, 3. Precipitation, 4. Dew Point, 5. Solar Radiation, 6. Solar Energy, 7. Wind Speed at 50 m, 8. Soil Moisture at 0.35 m, and 9. Soil Moisture at 0.4–1 m while the target features where kept the same. Table 8 presents the evaluation results. Temperature, humidity, precipitation, UV index, and soil moisture features can be predicted with high accuracy. However, light intensity shows lower prediction accuracy, primarily due to its variability caused by the unpredictable movement of clouds, resulting in patterns that do not align with those captured by the open-data measurements.

Although the accuracy of virtual sensors using open data is lower than that achieved with a physical base station, this approach offers a hardware-free solution. It enables the estimation of virtual sensor measurements without the need to deploy any physical devices in agricultural fields.

8. Observations & lessons learned

Through the development and deployment of physical and virtual sensors in real agricultural fields, integrated with IoT devices and AI models, and through direct engagement with farmers, valuable insights where gained into the functioning of these cyber–physical systems within the context of smart farming. These important findings are detailed in this section.

Table 8
Hit rate predictions using open-data.

Method	Temp. (5%)	Hum. (5%)	Rain (5%)	UV (5%)	Light (5%)	Moist. (5%)
LGBM	**84.3**	76.7	**89.3**	**54.8**	30.4	67.6
XGB	63.4	**65.4**	41.6	19.2	39.5	54.5
AdaBoost	55.4	73.6	74.3	16.3	**43.4**	75.4
HGB	56.4	71.2	76.3	29.2	34.5	78.2
LR	60.3	43.4	64.3	22.2	23.2	60.0
RF	67.4	71.2	78.3	22.3	24.3	**87.4**
SVR	48.9	60.3	90.2	20.2	28.3	76.3
FNN	64.3	32.5	33.4	28.3	28.3	43.3
TSFNN	43.3	29.3	38.4	22.3	11.2	43.4
TSRNN	45.6	33.4	12.3	18.2	18.6	45.5
TSLSTM	46.6	37.8	14.1	8.2	18.2	56.4

First, a significant challenge in the agricultural sector is the limited availability and high cost of sensors. Electrical conductivity (EC) sensors, which are valuable for assessing soil quality, face issues of accessibility and integration. Many commercially available EC sensors are expensive and come with inadequate documentation for interfacing with MCUs or microprocessing units. This lack of transparency complicates their use in cost-effective and adaptable farming systems, limiting opportunities for innovative, farmer-friendly solutions. For instance, most soil moisture sensors are restricted to a measurement depth of only 15 cm. This shallow range hinders effective monitoring of soil moisture at greater depths, which is essential for optimizing water usage and supporting plant growth. The demand for sensors capable of measuring moisture at depths of at least 30 cm remains unmet due to their high cost.

There is a need for developing and supporting open-source technologies in agriculture. While advanced irrigation control and automation systems from companies like Netafim[4] offer promising capabilities, they are often prohibitively expensive and require specialized expertise to implement and maintain. Sensors, controllers, and software necessary for such systems often cost tens of thousands of euros per field. There is a growing demand for simpler, more affordable solutions that farmers can develop, adapt, and manage independently. Unfortunately, the scarcity of open-source alternatives and detailed documentation hampers farmers' ability to harness technology tailored to their specific needs.

A further issue that should be taken into consideration is the inconsistency in sensor performance; for example, soil moisture sensors exhibited significant discrepancies between controlled laboratory conditions and real-world applications. Temporary fixes, such as applying conformal coating to protect sensitive sensor components.

Bridging the gap between complex AI models and practical agricultural use is essential for ensuring that smart farming solutions are both trustworthy and actionable. In our work, explainability plays a key role in making machine learning models more transparent and interpretable for end users, particularly farmers, who need to understand the reasoning behind the system's predictions. Instead of relying solely on data-driven outputs, explainability techniques help contextualize predictions within real agronomic knowledge. This is especially important in agriculture, where understanding cause-and-effect relationships is critical for informed decision-making. Without explainability, AI models may be perceived as black-box systems, which can reduce user trust, obscure potential biases, and make it difficult to troubleshoot unexpected results [60].

In our implementation, we observe that vector-based models such as LGBM, XGB, LR, SVM, and RF deliver highly accurate results. This strong performance enables the effective use of well-established explainability techniques, including Shapley Additive Explanations (SHAP) and Permutation Importance. SHAP values offer both global and local interpretability by quantifying the contribution of each input feature, such as sensor readings, to individual model predictions [61]. This makes SHAP particularly valuable for applications such as soil moisture estimation, humidity prediction, and anomaly detection. Permutation Importance, while conceptually simpler, provides an efficient method for evaluating feature relevance by measuring the impact on model performance when individual input variables are randomly shuffled [62]. By applying these techniques, farmers can gain clearer insights into how virtual sensor outputs are generated from physical sensor data or satellite-derived metrics, thereby improving transparency, building trust, and supporting more informed decision-making in the field.

Another important issue to highlight is that, although the rain sensor performed well during our experiments, its measurements are likely to become inaccurate over time due to significant degradation. This degradation can be attributed to several factors, including the highly saline water in the area, exposure to a mixture of various chemicals sprayed onto it, and its unprotected placement outdoors. An alternative to capacitive sensors for rain detection could be considered in future work. For instance, the same technology employed in automotive rain detection systems [63] could be utilized. This type of sensor and board provides categorical outputs, such as light, moderate, or heavy rainfall, in contrast to our capacitive sensor, which delivers raw values that must be mapped to a scale, such as 0 to 100 or 0 to 1.

Furthermore, the station's placement is important. Positioning it near the plant root provides highly accurate measurements of moisture at the plant. However, because the sensor was placed close to a single plant rather than between plants, it fails to capture a comprehensive view of the overall field moisture. When the sensors are close to the plants the measurements are affected by the plants. During the day, higher temperatures cause plants to uptake and recycle more water, and irrigation amplifies this effect.

[4] https://www.netafim.com/en/.

Conversely, at night, the plants' water requirements diminish, and evaporation combined with deeper water movement results in a decrease in observed soil moisture. These dynamics highlight the complex relationship between temperature, plant behavior, and irrigation in influencing soil moisture.

The virtual sensors developed in this system are data-driven machine learning models that operate independently of specific regional, climatic, soil, or crop conditions. Rather than relying on the physical characteristics of the environment, these models learn from the patterns and relationships within the sensor data itself captured over time. This allows the virtual sensors to generalize well and perform accurately across diverse agricultural environments. Their agnostic nature means they can adapt to different micro-climates, soil types, and crop varieties, as long as relevant sensor data is available for training. This flexibility was validated through successful deployments in fields with varying conditions and through additional experiments using open weather data, showing that virtual sensors can maintain strong predictive performance without being constrained to a specific location or setup.

The virtual sensor methodology has been evaluated in a variety of settings to assess its generalizability. The virtual sensor methodology was deployed and validated across two agricultural fields that were geographically distant from each other. These fields exhibited different environmental conditions, soil characteristics, and irrigation schedules, allowing us to assess the adaptability of the system. Prior to field deployment, all base stations were thoroughly tested in a controlled laboratory environment to ensure accurate sensor functionality. Additionally, the virtual sensor approach was successfully applied in a smart home environment, where it also demonstrated high accuracy [7]. Across all these scenarios, the system consistently delivered reliable results, confirming its robustness and general applicability across varied domains.

In this work, the Arduino platform was selected primarily due to its extensive community support and the availability of plug-and-play modules. This makes it particularly suitable for the prototyping phase, a critical step in the development of most products. To transition from prototype to product development, the recommended approach would involve designing a custom printed circuit board (PCB). This can be accomplished using open-source tools like KiCad[5] or professional design software such as Altium Designer.[6] Prototype PCBs could then be fabricated by manufacturers such as JLCPCB,[7] with components sourced from distributors like Mouser[8].

For this stage, the STM32 MCU platform would be a more appropriate choice. Designing a PCB around the STM32 platform allows for extensive customization, enabling the integration of features such as an on-board RTC or more precise sensors, such as the BME280 Bosch humidity and pressure sensor, directly onto the PCB. Furthermore, this approach allows for the selection of the MCU based on the specific requirements of the desired sensors. This flexibility eliminates the constraints imposed by preconfigured platforms like the Arduino Nano, which require choosing sensors compatible with the available pins. Instead, sensor selection can be prioritized and then tailor the STM32 MCU and PCB design to meet these needs.

Our research directly contributes to multiple United Nations Sustainable Development Goals (SDGs),[9] particularly SDG 2 – Zero Hunger, SDG 1 – No Poverty, and SDG 12 – Responsible Consumption and Production. By enabling cost-effective smart farming through the development and deployment of virtual sensors, we significantly enhance the precision and scalability of agricultural monitoring. Farmers can make more informed decisions using real-time and predictive data on soil moisture, humidity, temperature, and other key environmental variables. These insights lead to better crop yields and more efficient farming practices, ultimately increasing food production [64]. This has a direct impact on food security (SDG 2) and supports the livelihoods of farmers, especially in low-income regions, thus contributing to poverty alleviation (SDG 1).

In addition, our system promotes responsible consumption and production (SDG 12) by optimizing water and energy use in irrigation systems. Using sensor-based irrigation models, farmers can reduce water consumption by 36% to 47% without affecting crop yield, as demonstrated in this study [65]. Given that annual irrigation on farms requires between 3.2 and 9.1 megaliters of water per hectare,[10] the potential savings in water use are both environmentally and economically significant. Our approach also reduces dependency on costly hardware, democratizing access to precision agriculture by leveraging open-source tools, standardized communication protocols, and explainable AI. This not only facilitates sustainable farming practices but also fosters innovation and infrastructure development, aligning with SDG 9 – Industry, Innovation, and Infrastructure by bridging the gap between emerging technologies and practical, affordable agricultural solutions.

9. Conclusions & future work

This research has successfully demonstrated the integration of IoT-enabled physical sensors with AI-driven virtual sensors to optimize agricultural monitoring. Virtual sensors proved to be cost-effective and scalable, accurately predicting parameters such as soil moisture, temperature, and light intensity. The adaptability of these systems across both nearby and distant fields highlights their potential for widespread adoption in smart farming. Key contributions include insights into sensor placement, calibration, and data preprocessing, which address practical challenges while laying the foundation for more sustainable, data-driven agricultural practices.

Future work could explore integrating the proposed approach into a smart irrigation system that dynamically manages water supply using both physical and virtual sensors. This integration would optimize water usage by providing precise, data-driven

[5] https://www.kicad.org/.
[6] https://www.altium.com/.
[7] https://jlcpcb.com/.
[8] https://gr.mouser.com/.
[9] https://sdgs.un.org/goals.
[10] https://agriculture.vic.gov.au/farm-management/water/irrigation/variation-in-irrigation-requirements-of-forages-northvic.

irrigation strategies tailored to specific field conditions. Additionally, future efforts should focus on developing more affordable, reliable, and flexible sensors, alongside robust open-source platforms with comprehensive documentation. Empowering farmers to build and maintain their systems through these tools could significantly enhance agricultural efficiency and sustainability. Such advancements would not only provide cost-effective solutions for resource management and productivity improvements but also foster innovation and resilience in the agricultural sector.

CRediT authorship contribution statement

Athanasios Chourlias: Visualization, Software, Data curation. **John Violos:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Aris Leivadeas:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), grant No. RGPIN-2019-05250.

Data availability

Data will be made available on request.

References

[1] T. Ayoub Shaikh, T. Rasool, F. Rasheed Lone, Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming, Comput. Electron. Agric. 198 (2022) 107119, http://dx.doi.org/10.1016/j.compag.2022.107119, URL https://www.sciencedirect.com/science/article/pii/S0168169922004367.

[2] C. Marwa, S.B. Othman, H. Sakli, IoT based low-cost weather station and monitoring system for smart agriculture, in: 2020 20th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), (ISSN: 2573-539X) 2020, pp. 349–354, http://dx.doi.org/10.1109/STA50679.2020.9329292, URL https://ieeexplore.ieee.org/abstract/document/9329292.

[3] M. Balasubramaniyan, C. Navaneethan, Applications of internet of things for smart farming – a survey, NCRABE, Mater. Today: Proc. 47 (2021) 18–24, http://dx.doi.org/10.1016/j.matpr.2021.03.480, URL https://www.sciencedirect.com/science/article/pii/S2214785321025359.

[4] M. Javaid, A. Haleem, R.P. Singh, R. Suman, Enhancing smart farming through the applications of agriculture 4.0 technologies, Int. J. Intell. Networks 3 (2022) 150–164, http://dx.doi.org/10.1016/j.ijin.2022.09.004, URL https://www.sciencedirect.com/science/article/pii/S2666603022000173.

[5] P. Rajak, A. Ganguly, S. Adhikary, S. Bhattacharya, Internet of things and smart sensors in agriculture: Scopes and challenges, J. Agric. Food Res. 14 (2023) 100776, http://dx.doi.org/10.1016/j.jafr.2023.100776, URL https://www.sciencedirect.com/science/article/pii/S2666154323002831.

[6] G. Patrizi, A. Bartolini, L. Ciani, V. Gallo, P. Sommella, M. Carratù, A virtual soil moisture sensor for smart farming using deep learning, IEEE Trans. Instrum. Meas. 71 (2022) 1–11, http://dx.doi.org/10.1109/TIM.2022.3196446, URL https://ieeexplore.ieee.org/abstract/document/9849699 Conference Name: IEEE Transactions on Instrumentation and Measurement.

[7] G. Stavropoulos, J. Violos, S. Tsanakas, A. Leivadeas, Enabling artificial intelligent virtual sensors in an IoT environment, Sensors 23 (3) (2023) 1328, http://dx.doi.org/10.3390/s23031328, URL https://www.mdpi.com/1424-8220/23/3/1328 Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[8] A. Wongchai, S.K. Shukla, M.A. Ahmed, U. Sakthi, M. Jagdish, R. kumar, Artificial intelligence - enabled soft sensor and internet of things for sustainable agriculture using ensemble deep learning architecture, Comput. Electr. Eng. 102 (2022) 108128, http://dx.doi.org/10.1016/j.compeleceng.2022.108128, URL https://www.sciencedirect.com/science/article/pii/S0045790622003780.

[9] Anonimized: Catastrophic floods | UNICEF URLanonimized.

[10] B.B. Sinha, R. Dhanalakshmi, Recent advancements and challenges of internet of things in smart agriculture: A survey, Future Gener. Comput. Syst. 126 (2022) 169–184, http://dx.doi.org/10.1016/j.future.2021.08.006, URL https://www.sciencedirect.com/science/article/pii/S0167739X21003113.

[11] S.I. Hassan, M.M. Alam, U. Illahi, M.A. Al Ghamdi, S.H. Almotiri, M.M. Su'ud, A systematic review on monitoring and advanced control strategies in smart agriculture, IEEE Access 9 (2021) 32517–32548, http://dx.doi.org/10.1109/ACCESS.2021.3057865, URL https://ieeexplore.ieee.org/abstract/document/9350242 Conference Name: IEEE Access.

[12] S.D. Alwis, Z. Hou, Y. Zhang, M.H. Na, B. Ofoghi, A. Sajjanhar, A survey on smart farming data, applications and techniques, Comput. Ind. 138 (2022) 103624, http://dx.doi.org/10.1016/j.compind.2022.103624, URL https://www.sciencedirect.com/science/article/pii/S0166361522000197.

[13] E. Bwambale, F.K. Abagale, G.K. Anornu, Smart irrigation monitoring and control strategies for improving water use efficiency in precision agriculture: A review, Agricult. Water. Manag. 260 (2022) 107324, http://dx.doi.org/10.1016/j.agwat.2021.107324, URL https://www.sciencedirect.com/science/article/pii/S0378377421006016.

[14] C. Prakash, L.P. Singh, A. Gupta, S.K. Lohan, Advancements in smart farming: A comprehensive review of IoT, wireless communication, sensors, and hardware for agricultural automation, Sensors Actuators A: Phys. 362 (2023) 114605, http://dx.doi.org/10.1016/j.sna.2023.114605, URL https://www.sciencedirect.com/science/article/pii/S0924424723004545.

[15] N.G. Rezk, E.E.-D. Hemdan, A.-F. Attia, A. El-Sayed, M.A. El-Rashidy, An efficient IoT based framework for detecting rice disease in smart farming system, Multimedia Tools Appl. 82 (29) (2023) 45259–45292, http://dx.doi.org/10.1007/s11042-023-15470-2.

[16] C.-J. Chen, Y.-Y. Huang, Y.-S. Li, C.-Y. Chang, Y.-M. Huang, An aIoT based smart agricultural system for pests detection, IEEE Access 8 (2020) 180750–180761, http://dx.doi.org/10.1109/ACCESS.2020.3024891, URL https://ieeexplore.ieee.org/abstract/document/9200475 Conference Name: IEEE Access.

[17] S.V. Gaikwad, A.D. Vibhute, K.V. Kale, S.C. Mehrotra, An innovative IoT based system for precision farming, Comput. Electron. Agric. 187 (2021) 106291, http://dx.doi.org/10.1016/j.compag.2021.106291, URL https://www.sciencedirect.com/science/article/pii/S0168169921003082.

[18] A. Tzounis, N. Katsoulas, T. Bartzanas, C. Kittas, Internet of things in agriculture, recent advances and future challenges, Biosyst. Eng. 164 (2017) 31–48, http://dx.doi.org/10.1016/j.biosystemseng.2017.09.007, URL https://www.sciencedirect.com/science/article/pii/S1537511017302544.

[19] F.A.A. Souza, R. Araújo, J. Mendes, Review of soft sensor methods for regression applications, Chemometr. Intell. Lab. Syst. 152 (2016) 69–79, http://dx.doi.org/10.1016/j.chemolab.2015.12.011, URL https://www.sciencedirect.com/science/article/pii/S0169743915003263.

[20] F.K. Shaikh, S. Karim, S. Zeadally, J. Nebhen, Recent trends in internet-of-things-enabled sensor technologies for smart agriculture, IEEE Internet Things J. 9 (23) (2022) 23583–23598, http://dx.doi.org/10.1109/JIOT.2022.3210154, URL https://ieeexplore.ieee.org/abstract/document/9903855 Conference Name: IEEE Internet of Things Journal.

[21] K. Paul, S.S. Chatterjee, P. Pai, A. Varshney, S. Juikar, V. Prasad, B. Bhadra, S. Dasgupta, Viable smart sensors and their application in data driven agriculture, Comput. Electron. Agric. 198 (2022) 107096, http://dx.doi.org/10.1016/j.compag.2022.107096, URL https://www.sciencedirect.com/science/article/pii/S0168169922004136.

[22] V. Kumar, K.V. Sharma, N. Kedam, A. Patel, T.R. Kate, U. Rathnayake, A comprehensive review on smart and sustainable agriculture using IoT technologies, Smart Agric. Technol. 8 (2024) 100487, http://dx.doi.org/10.1016/j.atech.2024.100487, URL https://www.sciencedirect.com/science/article/pii/S2772375524000923.

[23] G. Anagnostopoulos, G. Stavropoulos, J. Violos, A. Leivadeas, I. Varlamis, Enhancing virtual sensors to deal with missing values and low sampling rates, in: 2023 11th IEEE International Conference on Mobile Cloud Computing, Services, and Engineering (MobileCloud), (ISSN: 2573-7562) 2023, pp. 39–44, http://dx.doi.org/10.1109/MobileCloud58788.2023.00012, URL https://ieeexplore.ieee.org/abstract/document/10229478.

[24] F. Curreri, G. Fiumara, M.G. Xibilia, Input selection methods for soft sensor design: A survey, Futur. Internet 12 (6) (2020) 97, http://dx.doi.org/10.3390/fi12060097, URL https://www.mdpi.com/1999-5903/12/6/97Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[25] Q. Sun, Z. Ge, A survey on deep learning for data-driven soft sensors, IEEE Trans. Ind. Informatics 17 (9) (2021) 5853–5866, http://dx.doi.org/10.1109/TII.2021.3053128, URL https://ieeexplore.ieee.org/abstract/document/9329169Conference Name: IEEE Transactions on Industrial Informatics.

[26] D. Weraikat, K. Šorič, M. Žagar, M. Sokač, Data analytics in agriculture: Enhancing decision-making for crop yield optimization and sustainable practices, Sustainability 16 (17) (2024) 7331, http://dx.doi.org/10.3390/su16177331, URL https://www.mdpi.com/2071-1050/16/17/7331Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.

[27] L.A. Grieco, G. Boggia, G. Piro, Y. Jararweh, C. Campolo (Eds.), Ad-hoc, mobile, and wireless networks: 19th international conference on ad-hoc networks and wireless, ADHOC-NOW 2020, bari, Italy, october 19–21, 2020, proceedings, Lecture Notes in Computer Science, vol. 12338, Springer International Publishing, Cham, 2020, http://dx.doi.org/10.1007/978-3-030-61746-2, URL https://link.springer.com/10.1007/978-3-030-61746-2.

[28] D. Heeger, M. Garigan, E.E. Tsiropoulou, J. Plusquellic, Secure energy constrained lora mesh network, in: L.A. Grieco, G. Boggia, G. Piro, Y. Jararweh, C. Campolo (Eds.), Ad-Hoc, Mobile, and Wireless Networks, Springer International Publishing, Cham, 2020, pp. 228–240, http://dx.doi.org/10.1007/978-3-030-61746-2_17.

[29] S. Saran, S.S. Hiremath, A. Kumar, P. Ashoka, H. Singh, S. Chakraborty, V. Kashyap, A.K. Tiwari, S.K. Pandey, Remote sensing and automated monitoring systems for insect pest detection and surveillance, UTTAR PRADESH JOURNAL ZOOLOGY 46 (2) (2025) 155–171, http://dx.doi.org/10.56557/upjoz/2025/v46i24771, URL https://mbimph.com/index.php/UPJOZ/article/view/4771.

[30] X. Zhang, G. Feng, X. Sun, Advanced technologies of soil moisture monitoring in precision agriculture: A review, J. Agric. Food Res. 18 (2024) 101473, http://dx.doi.org/10.1016/j.jafr.2024.101473, URL https://www.sciencedirect.com/science/article/pii/S2666154324005106.

[31] Y. Zhou, Y. Ma, S.T. Ata-Ul-Karim, S. Wang, I. Ciampitti, V. Antoniuk, C. Wu, M.N. Andersen, D. Cammarano, Integrating multi-angle and multi-scale remote sensing for precision nitrogen management in agriculture: A review, Comput. Electron. Agric. 230 (2025) 109829, http://dx.doi.org/10.1016/j.compag.2024.109829, URL https://www.sciencedirect.com/science/article/pii/S0168169924012201.

[32] S.A.H. Selmy, D.E. Kucher, A.R.A. Moursy, S.A.H. Selmy, D.E. Kucher, A.R.A. Moursy, Integrating remote sensing, GIS, and AI technologies in soil erosion studies, IntechOpen, 2025, http://dx.doi.org/10.5772/intechopen.1008677, URL https://www.intechopen.com/online-first/1198031.

[33] N. Ahmed, N. Shakoor, Advancing agriculture through IoT, big data, and AI: A review of smart technologies enabling sustainability, Smart Agric. Technol. 10 (2025) 100848, http://dx.doi.org/10.1016/j.atech.2025.100848, URL https://www.sciencedirect.com/science/article/pii/S2772375525000814.

[34] B. Yu, J. Yao, Q. Fu, Z. Zhong, H. Xie, Y. Wu, Y. Ma, P. He, Deep learning or classical machine learning? An empirical study on log-based anomaly detection, in: 2024 IEEE/ACM 46th International Conference on Software Engineering, ICSE, (ISSN: 1558-1225) 2024, pp. 403–415, http://dx.doi.org/10.1145/3597503.3623308, URL https://ieeexplore.ieee.org/document/10549148?utm_source=chatgpt.com.

[35] F. Leens, An introduction to I2C and SPI protocols, IEEE Instrum. Meas. Mag. 12 (1) (2009) 8–13, http://dx.doi.org/10.1109/MIM.2009.4762946, URL https://ieeexplore.ieee.org/abstract/document/4762946Conference Name: IEEE Instrumentation & Measurement Magazine.

[36] A Guide to SPICE Simulation URL https://www.altium.com/documentation/altium-designer/spice-simulation-guide.

[37] J. Hrisko, Capacitive soil moisture sensor theory, calibration, and testing, 2020, http://dx.doi.org/10.13140/RG.2.2.36214.83522.

[38] C. Clemens, A.E. Jobst, M. Radschun, J. Himmel, O. Kanoun, Signal processing and calibration of a low-cost inductive rain sensor for raindrop detection and precipitation calculation, Measurement 227 (2024) 114286, http://dx.doi.org/10.1016/j.measurement.2024.114286, URL https://www.sciencedirect.com/science/article/pii/S0263224124001702.

[39] LTR-390UV-01 Lite-On | Mouser URL https://gr.mouser.com/ProductDetail/Lite-On/LTR-390UV-01?qs=g5ciJ0jwZaECcUd5i6p7%252Bg%3D%3D.

[40] Gravity: DHT22 Temperature and Humidity Sensor - DFRobot URL https://www.dfrobot.com/product-1102.html?srsltid=AfmBOooBEo5QKfN1JBH5j6048QnHvyUzNOZ5ryT4bjb-ksx0dlwWtosz.

[41] N. Parashar, P. Johri, A. Khan, N. Gaur, S. Kadry, An integrated analysis of yield prediction models: A comprehensive review of advancements and challenges, Comput. Mater. Contin. 80 (1) (2024) 389–425, http://dx.doi.org/10.32604/cmc.2024.050240, URL https://www.techscience.com/cmc/v80n1/57370Publisher: Tech Science Press.

[42] S.A. Haider, S.R. Naqvi, T. Akram, G.A. Umar, A. Shahzad, M.R. Sial, S. Khaliq, M. Kamran, LSTM neural network based forecasting model for wheat production in Pakistan, Agronomy 9 (2) (2019) 72, http://dx.doi.org/10.3390/agronomy9020072, URL https://www.mdpi.com/2073-4395/9/2/72Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

[43] R. Nandi, D. Shrestha, Assessment of low-cost and higher-end soil moisture sensors across various moisture ranges and soil textures, Sensors 24 (18) (2024) 5886, http://dx.doi.org/10.3390/s24185886, URL https://www.mdpi.com/1424-8220/24/18/5886Number: 18 Publisher: Multidisciplinary Digital Publishing Institute.

[44] C.G. Mattera, J. Quevedo, T. Escobet, H.R. Shaker, M. Jradi, Fault detection and diagnostics in ventilation units using linear regression virtual sensors, in: 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), 2018, pp. 1–6, http://dx.doi.org/10.1109/ISAECT.2018.8618755, URL https://ieeexplore.ieee.org/abstract/document/8618755.

[45] T. Hastie, J. Friedman, R. Tibshirani, Linear methods for regression, in: T. Hastie, J. Friedman, R. Tibshirani (Eds.), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, NY, 2001, pp. 41–78, http://dx.doi.org/10.1007/978-0-387-21606-5_3.

[46] H. Kaneko, K. Funatsu, Adaptive soft sensor based on online support vector regression and Bayesian ensemble learning for various states in chemical plants, Chemometr. Intell. Lab. Syst. 137 (2014) 57–66, http://dx.doi.org/10.1016/j.chemolab.2014.06.008, URL https://www.sciencedirect.com/science/article/pii/S0169743914001294.

[47] K. Smets, B. Verdonk, E.M. Jordaan, Evaluation of performance measures for SVR hyperparameter selection, in: 2007 International Joint Conference on Neural Networks, (ISSN: 2161-4407) 2007, pp. 637–642, http://dx.doi.org/10.1109/IJCNN.2007.4371031, URL https://ieeexplore.ieee.org/document/4371031.

[48] A.G. Putrada, N. Alamsyah, S.F. Pane, M.N. Fauzan, D. Perdana, Virtual sensors method and architecture for a smart home environment with random forest, in: 2023 10th International Conference on ICT for Smart Society, ICISS, 2023, pp. 1–6, http://dx.doi.org/10.1109/ICISS59129.2023.10292065, URL https://ieeexplore.ieee.org/abstract/document/10292065.

[49] R.E. Schapire, The boosting approach to machine learning: An overview, in: D.D. Denison, M.H. Hansen, C.C. Holmes, B. Mallick, B. Yu (Eds.), Nonlinear Estimation and Classification, Springer, New York, NY, 2003, pp. 149–171, http://dx.doi.org/10.1007/978-0-387-21579-2_9.

[50] H. Tian, A. Wang, Z. Mao, A new soft sensor modeling method based on modified AdaBoost with incremental learning, in: Proceedings of the 48h IEEE Conference on Decision and Control (CDC) Held Jointly with 2009 28th Chinese Control Conference, (ISSN: 0191-2216) 2009, pp. 8375–8380, http://dx.doi.org/10.1109/CDC.2009.5400292, URL https://ieeexplore.ieee.org/abstract/document/5400292.

[51] Y. Li, C. Yang, H. Zhang, C. Jia, A model combining Seq2Seq network and LightGBM algorithm for industrial soft sensor, 21st IFAC World Congress, IFAC- Pap. 53 (2) (2020) 12068–12073, http://dx.doi.org/10.1016/j.ifacol.2020.12.753, URL https://www.sciencedirect.com/science/article/pii/S2405896320310752.

[52] P.M.L. Ching, X. Zou, D. Wu, R.H.Y. So, G.H. Chen, Development of a wide-range soft sensor for predicting wastewater BOD5 using an extreme gradient boosting (xgboost) machine, Environ. Res. 210 (2022) 112953, http://dx.doi.org/10.1016/j.envres.2022.112953, URL https://www.sciencedirect.com/science/article/pii/S0013935122002808.

[53] G. Marvin, L. Grbčić, S. Družeta, L. Kranjčević, Water distribution network leak localization with histogram-based gradient boosting, J. Hydroinformatics 25 (3) (2023) 663–684, http://dx.doi.org/10.2166/hydro.2023.102.

[54] J. Corrigan, J. Zhang, Developing accurate data-driven soft-sensors through integrating dynamic kernel slow feature analysis with neural networks, J. Process Control 106 (2021) 208–220, http://dx.doi.org/10.1016/j.jprocont.2021.09.006, URL https://www.sciencedirect.com/science/article/pii/S0959152421001554.

[55] G.K. Vishwakarma, C. Paul, A.M. Elsawah, A hybrid feedforward neural network algorithm for detecting outliers in non-stationary multivariate time series, Expert Syst. Appl. 184 (2021) 115545, http://dx.doi.org/10.1016/j.eswa.2021.115545, URL https://www.sciencedirect.com/science/article/pii/S0957417421009520.

[56] H. Hewamalage, C. Bergmeir, K. Bandara, Recurrent neural networks for time series forecasting: Current status and future directions, Int. J. Forecast. 37 (1) (2021) 388–427, http://dx.doi.org/10.1016/j.ijforecast.2020.06.008, URL https://www.sciencedirect.com/science/article/pii/S0169207020300996.

[57] B. Xu, C.K. Pooi, K.M. Tan, S. Huang, X. Shi, H.Y. Ng, A novel long short-term memory artificial neural network (LSTM)-based soft-sensor to monitor and forecast wastewater treatment performance, J. Water Process. Eng. 54 (2023) 104041, http://dx.doi.org/10.1016/j.jwpe.2023.104041, URL https://www.sciencedirect.com/science/article/pii/S2214714423005603.

[58] T. chourlias, Athanasioschourlias/IoT-AI-virtual-sensors-for-smart-farming, 2025, URL https://github.com/Athanasioschourlias/IoT-AI-Virtual-Sensors-For-Smart-Farming original-date: 2025-01-15 T09:39:01Z.

[59] Weather Data & Weather API | Visual Crossing URL https://www.visualcrossing.com/.

[60] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 1135–1144, http://dx.doi.org/10.1145/2939672.2939778.

[61] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017, URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[62] G. Elkhawaga, O. Elzeki, M. Abuelkheir, M. Reichert, Evaluating explainable artificial intelligence methods based on feature elimination: A functionality-grounded approach, Electron. 12 (7) (2023) 1670, http://dx.doi.org/10.3390/electronics12071670, URL https://www.mdpi.com/2079-9292/12/7/1670Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

[63] A. Cord, D. Aubert, Towards rain detection through use of in-vehicle multipurpose cameras, in: 2011 IEEE Intelligent Vehicles Symposium (IV), (ISSN: 1931-0587) 2011, pp. 833–838, http://dx.doi.org/10.1109/IVS.2011.5940484, URL https://ieeexplore.ieee.org/abstract/document/5940484.

[64] N.K. Madzhi, M.A. Nor Akhsan, Control of plant growth by monitoring soil moisture, temperature and humidity in dry climate, IOP Conf. Ser.: Mater. Sci. Eng. 1192 (1) (2021) 012027, http://dx.doi.org/10.1088/1757-899X/1192/1/012027, Publisher: IOP Publishing.

[65] M. Palumbo, M. D'Imperio, V. Tucci, M. Cefola, B. Pace, P. Santamaria, A. Parente, F.F. Montesano, Sensor-based irrigation reduces water consumption without compromising yield and postharvest quality of soilless green bean, Agron. 11 (12) (2021) 2485, http://dx.doi.org/10.3390/agronomy11122485, URL https://www.mdpi.com/2073-4395/11/12/2485Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.