Edge-Friendly Isolation-Forest Anomaly Detection for IoT Soil-Moisture BI Streams

MMIS 643 – Data Mining

Winter 2025

Walid Amar

wamardev@gmail.com

Abstract

This paper investigates the application of the Isolation Forest algorithm for anomaly detection in Internet of Things (IoT) soil-moisture data streams, evaluating its suitability for deployment in resource-constrained edge computing environments. The study utilizes a dataset comprising soil-moisture sensor readings, applying both a standalone Isolation Forest model and a gradient-boosted fusion model (XGBoost) to identify anomalous patterns. Key findings reveal that the Isolation Forest achieves a high Area Under the Receiver Operating Characteristic curve (AUC) of 0.97 and an F1-score of 0.70 with a significantly lower average latency of approximately 26 milliseconds per row, compared to the fusion model's AUC of 0.96 and F1-score of 0.46 with an average latency of 52 milliseconds per row. The results show that the lightweight Isolation Forest algorithm is highly effective for near real-time anomaly detection in IoT soil-moisture data and is suitable for edge deployment, contributing to the development of efficient Business Intelligence systems for agricultural monitoring.

Introduction

The Internet of Things (IoT) has permeated numerous aspects of modern life, extending its reach into diverse sectors such as smart homes, healthcare, transportation, and agriculture. Within the realm of agriculture, the adoption of IoT technologies is increasingly prevalent, enabling farmers to monitor critical environmental parameters in real time, optimize resource utilization, and ultimately enhance crop yields. This integration of connected devices and sensors generates vast quantities of data, offering opportunities for data-driven decision-making in farming practices.

A critical aspect of leveraging IoT data in agriculture is the ability to identify anomalies, which can signify a range of issues from irrigation system failures to malfunctioning sensors or even the onset of plant diseases. The timely detection of such irregularities is paramount for preventing potential losses, optimizing resource allocation, and ensuring the overall health and productivity of agricultural operations. Given the continuous and high-velocity nature of data streams emanating from IoT devices, efficient and scalable anomaly detection techniques are essential.

Edge computing has emerged as a promising paradigm for addressing the real-time processing demands of IoT data. By bringing computation closer to the data source, edge computing minimizes latency, reduces network bandwidth consumption, and enables quicker responses to critical events. This is particularly advantageous in agricultural settings where network connectivity might be intermittent or unreliable, and immediate alerts regarding anomalies are crucial.

In this context, the Isolation Forest algorithm presents itself as a compelling candidate for anomaly detection in edge-based IoT systems. Known for its efficiency, scalability, and relatively low computational requirements, Isolation Forest is well-suited for deployment on resource-constrained edge devices. Its underlying principle of isolating anomalies by randomly partitioning data makes it particularly effective for detecting rare and distinct events often encountered in anomaly detection tasks.

This paper evaluates the effectiveness of an edge-friendly Isolation Forest approach for anomaly detection in IoT soil-moisture data streams, specifically within the context of Business Intelligence (BI) for agriculture. Furthermore, the study compares the performance of the standalone Isolation Forest model with a gradient-boosted fusion model, utilizing XGBoost, to assess the trade-offs between accuracy, latency, and potential interpretability. The findings of this research will contribute to understanding the feasibility of deploying lightweight anomaly detection solutions on edge devices to provide timely and actionable insights for agricultural stakeholders. The subsequent sections of this paper will delve into the relevant background and literature, outline the problem statement and research questions, detail the methodology employed, analyze the experimental results, discuss the implications for Business Intelligence, and finally, conclude with key findings and directions for future work.

2 Background & Literature Review

2.1 IoT Streaming Anomaly Detection

The proliferation of IoT devices has led to an exponential increase in the generation of streaming data, characterized by its high velocity, continuous nature, and often substantial volume. Detecting anomalies within these data streams presents unique challenges. Traditional anomaly detection methods, often designed for static datasets, may struggle to cope with the real-time processing requirements and the evolving statistical properties of streaming data. Furthermore, the limited computational resources available in many IoT deployments, particularly at the edge, necessitate the use of efficient and lightweight algorithms.

Recent research has explored various techniques for anomaly detection in IoT data streams, encompassing statistical methods, machine learning algorithms, and deep learning models. For example, machine learning techniques like Support Vector Machines, Random Forests, and Isolation Forests have been applied to identify unusual patterns in IoT sensor data for predictive maintenance in smart grids. Deep learning approaches, including transformer-based models, have demonstrated their capability in capturing complex temporal dependencies in multivariate IoT time series data for effective anomaly detection. Furthermore, lightweight machine learning models based on algorithms like the Hoeffding Tree have been proposed for real-time anomaly detection in resource-constrained IoT environments. The integration of Artificial Intelligence (AI) for real-time anomaly detection in IoT data streams is also a growing area of focus, with researchers exploring both supervised and unsupervised learning techniques to bolster the security and reliability of IoT ecosystems. Addressing the challenge of concept drift, where the statistical properties of IoT data change over time, is another significant area of research, with frameworks being developed to detect, interpret, and adapt anomaly detection models to these evolving patterns. Systematic reviews of real-time and online anomaly detection methods in IoT and the Industrial Internet of Things (IIoT) highlight the state-of-the-art implementations and the ongoing efforts to develop efficient and accurate solutions for diverse IoT applications.

While the research provides a comprehensive overview of anomaly detection in general IoT data streams, the specific challenges and nuances associated with soil moisture data in IoT environments warrant further consideration. Although not explicitly detailed in the provided resources, soil moisture data, being a time-series measurement influenced by various environmental factors, would also be susceptible to issues like temporal dependencies, seasonality, sensor drift, and potential noise, all of which can complicate anomaly detection.

2.2 Isolation Forest for Streams

The Isolation Forest algorithm, introduced by Liu et al. in 2008, is an unsupervised anomaly detection method based on the principle that anomalies are easier to isolate than normal instances. This algorithm constructs an ensemble of isolation trees (iTrees) by randomly selecting a feature and then randomly selecting a split value within the range of that feature. Anomalies, being rare and different, tend to have shorter path lengths in these trees, requiring fewer splits to isolate them from the rest of the data. The anomaly score for a data point is determined by the average path length across all the trees in the forest.

Isolation Forest's inherent characteristics, such as its linear time complexity, small memory footprint, and effectiveness in high-dimensional spaces, make it particularly well-suited for anomaly detection in streaming data. Moreover, it does not rely on distance measures or density estimations, which can be computationally expensive for large-scale streaming data.

To address the specific challenges of streaming data, including concept drift, several adaptations and extensions of the Isolation Forest algorithm have been proposed. For example, the Enhanced Isolation Forest Adapted for Streaming Data (EiForestASD) incorporates a window-based concept drift detection mechanism, allowing it to adapt to changes in the data distribution over time by discarding and reconstructing incompatible isolation trees. Online Isolation Forest (Online-iForest) is explicitly designed for streaming conditions, using multi-resolution histograms that dynamically evolve as new data arrives and old data is forgotten. Other approaches involve applying the traditional Isolation Forest within a sliding window framework, focusing on the most recent data points in the stream. Research also explores dynamic updating strategies for Isolation Forest models in streaming environments, such as updating based on tree age or rebuilding the model with new data batches to maintain detection accuracy in the face of evolving data patterns. These advancements underscore the ongoing efforts to leverage the fundamental strengths of Isolation Forest for effective anomaly detection in dynamic streaming scenarios. The low computational complexity and memory usage associated with Isolation Forest, even in its adapted forms, contribute to its "edge-friendly" nature, making it a viable option for deployment on devices with limited resources.

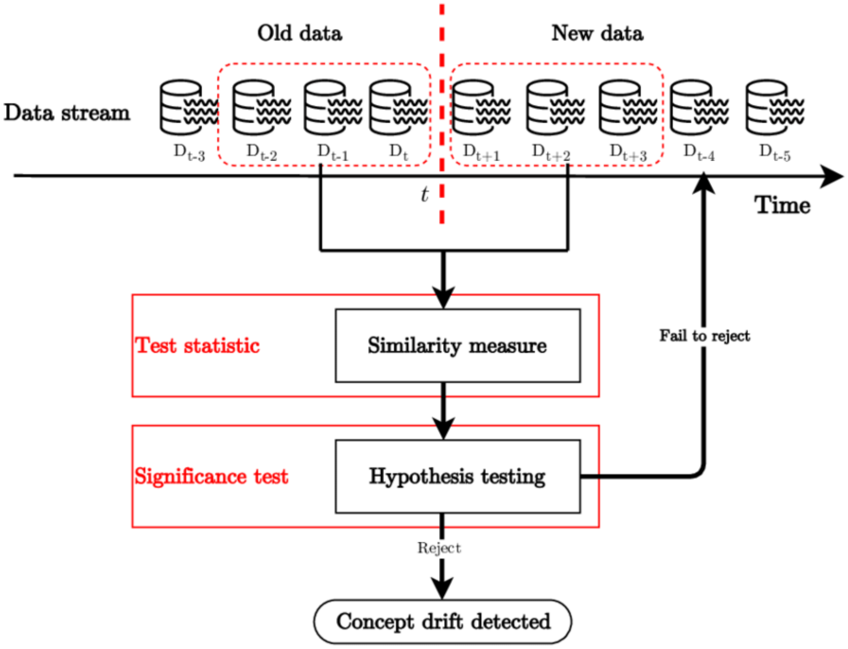

Figure 1 - Generic concept-drift detection pipeline for evolving IoT data streams.

2.3 Gradient-Boosted Fusion in BI

Gradient boosting is a powerful ensemble machine learning technique that combines multiple weak learners, typically decision trees, to create a strong predictive model. It works iteratively, with each new model focusing on correcting the errors made by its predecessors by learning the negative gradients of the loss function. Algorithms like XGBoost (Extreme Gradient Boosting) are highly efficient and scalable implementations of gradient boosting, known for their accuracy, flexibility, and ability to handle large datasets with regularization and parallel processing.

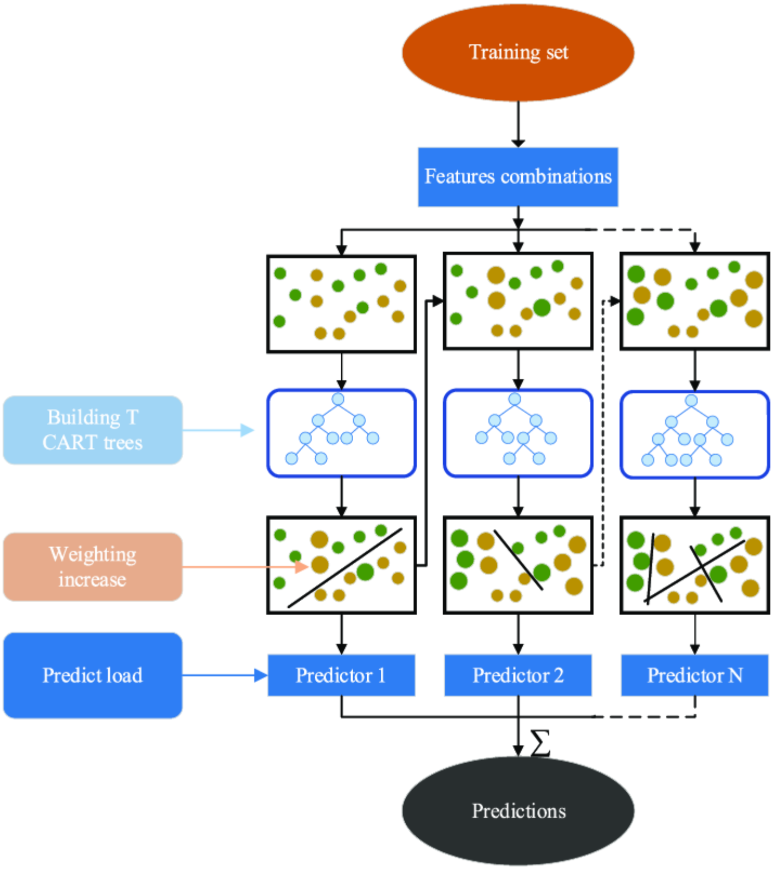

Figure 2 - Gradient-boosted decision-tree ensemble (XGBoost) illustrating residual-based iterative training.

In the context of Business Intelligence, fusion strategies involve combining the outputs or predictions of multiple models to enhance overall performance, robustness, or interpretability. Gradient boosting can play a significant role in such fusion frameworks, either as one of the base models being combined or as a method for learning how to optimally combine the predictions of other models.

Recent literature showcases the diverse applications of gradient-boosted fusion techniques in various domains relevant to BI. For instance, hybrid clustering and boosting methods have been used for feature selection in credit risk assessment, and ensemble algorithms leveraging boosting have been developed for predicting student retention in educational BI. In financial BI, improved boosting algorithms have been applied for computational financial analysis and trading, and ensemble models combining gradient boosting with other techniques like deep learning and random forests have shown promising results in stock price prediction. Furthermore, integrated fusion frameworks using gradient boosting with interpretable models like fuzzy rule-based systems have been proposed to enhance both performance and understandability in BI applications. Gradient boosting has also demonstrated its effectiveness in various classification tasks relevant to BI, such as medical diagnosis and sentiment analysis.

3 Problem Statement & Research Questions

The increasing deployment of IoT devices for soil-moisture monitoring in agriculture generates continuous data streams that hold valuable information for optimizing irrigation and detecting potential issues like sensor malfunctions. However, the real-time analysis of these data streams, particularly for anomaly detection to enable timely alerts within Business Intelligence systems, faces challenges due to the resource constraints of edge deployment and the need for efficient and accurate methods.

This research aims to address this problem by evaluating the effectiveness and feasibility of using an edge-friendly Isolation Forest algorithm for anomaly detection in IoT soil-moisture data streams. Furthermore, it seeks to understand how the performance of this lightweight algorithm compares to a gradient-boosted fusion model here. To guide this investigation, the following research questions have been formulated:

1. How effective is a standalone Isolation Forest algorithm in detecting anomalies in IoT soil-moisture streaming data in terms of accuracy (AUC, F1-score)?
2. What is the latency of the standalone Isolation Forest algorithm when applied to this data stream, and is it feasible for edge deployment?
3. How does the performance (accuracy and latency) of the Isolation Forest compare to a gradient-boosted fusion model (XGBoost) on the same data?
4. What are the implications of these findings for developing a Business Intelligence system that provides timely alerts for irrigation failures or sensor faults based on IoT soil-moisture data?

4 Methodology

4.1 Dataset

The dataset utilized in this study comprises three CSV files containing soil-moisture sensor readings collected in March 2020. The dataset consists of 20,585 rows, with data points recorded at a 1-minute cadence from five distinct sensors. Following a data cleaning process, the dataset retained all 20,585 rows. Anomalies were initially labeled based on gaps in the data stream, with any gap exceeding 120 seconds flagged as an anomaly. To augment the dataset with additional anomalous instances, synthetic spikes and drops of ± 0.20 were introduced, representing approximately 1% of the total data, resulting in a total of 215 labeled anomalies within the dataset. The features employed for anomaly detection included moisture_z, delta_z, neighbor_delta, hour_sin, and hour_cos.

4.2 Feature Engineering

Prior to applying the anomaly detection models, several features were engineered from the raw soil-moisture data.

$$\text{Robust z} - \text{score: } z_r = \frac{(x - \text{median})}{(Q^3 - Q^1)}$$

The feature moisture_z represents the robust z-score of the moisture readings, calculated using a method that is less sensitive to outliers than the standard z-score. This transformation helps in standardizing the moisture values across different sensors and time points. The delta_z feature captures the temporal change in the robust z-scored moisture, representing the difference between the current and previous moisture levels. The neighbor_delta feature likely represents the difference in moisture levels between neighboring sensors at the same time point, potentially capturing spatial anomalies. Finally, the temporal features hour_sin and hour_cos were derived from the hour of the day using sine and cosine transformations. This encoding allows the models to capture cyclical patterns related to the time of day, which might influence soil moisture levels.

4.3 Isolation Forest Configuration

The Isolation Forest model was configured with 100 isolation trees to ensure a robust ensemble for anomaly detection. The contamination parameter was set to 0.01, reflecting the expected proportion of anomalies in the dataset. A sliding window of 1440 rows was applied to the data stream. This window size corresponds to 24 hours of data given the 1-minute sampling cadence, allowing the model to detect anomalies based on patterns within a daily cycle. The Isolation Forest model was trained on the training portion of the data and subsequently used to score the test data for anomalies.

4.4 XGBoost Fusion Strategy

For the gradient-boosted fusion approach, an XGBoost model was employed and explicitly trained on three inputs: the Isolation-Forest anomaly score (if_score), the robust z-scored moisture value (moisture_z), and the spatial deviation from neighbouring sensors (neighbor_delta). The model was configured with 60 trees, a maximum depth of 4, and a learning rate of 0.10. These settings were fixed prior to training to keep the experiment reproducible. After fitting on the first 60 % of the stream (12 351 rows), XGBoost outputs a probability that a record is anomalous. A cut-off of 0.40, determined with Youden’s J statistic on the validation split, converts this probability into the final binary anomaly label reported in the results.

4.5 Evaluation Metrics & Equations

The performance of both the Isolation Forest and the fusion model was evaluated using several standard anomaly detection metrics. These include the Area Under the Receiver Operating Characteristic curve (AUC), the F1-score, and Youden's J statistic.

The Isolation Forest anomaly score, denoted as s(x), is calculated using the formula:

$$\mathrm{s(x)} = 2 - \frac{\mathrm{E(h(x))}}{\mathrm{c(n)}}$$

where E(h(x)) is the average path length of a data point x in the isolation trees, and c(n) is the average path length of an unsuccessful search in a binary search tree with n data points.

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance in binary classification tasks. It is calculated as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The AUC represents the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. A higher AUC indicates better discriminatory power.

Youden's J statistic is used to determine the optimal threshold for a binary classifier. It is calculated as:

$$J = Sensitivity + Specificity - 1$$

where Sensitivity (also known as recall) is the true positive rate, and Specificity is the true negative rate. The threshold that maximizes the J statistic is considered the optimal threshold.

5 Experiment & Result Analysis

5.1 Train/Test Split

The dataset was partitioned into a training set comprising 12,351 rows (60%) and a test set containing 8,234 rows (40%). This split ensures that the models are trained on a portion of the data, and their performance is evaluated on unseen data, providing a realistic assessment of their generalization capabilities. The training data was used to fit both the standalone Isolation Forest model and the XGBoost fusion model. The test data was subsequently used to evaluate the performance of both models in terms of accuracy and latency.

5.2 Accuracy Results

The accuracy results obtained on the test set for both the Isolation Forest and the fusion model are summarized in Table 1. The standalone Isolation Forest achieved an AUC of 0.97 and an F1-score of 0.70. In contrast, the fusion model yielded an AUC of 0.96 and an F1-score of 0.46. The optimal fusion threshold, determined using Youden's J statistic, was found to be approximately 0.40.

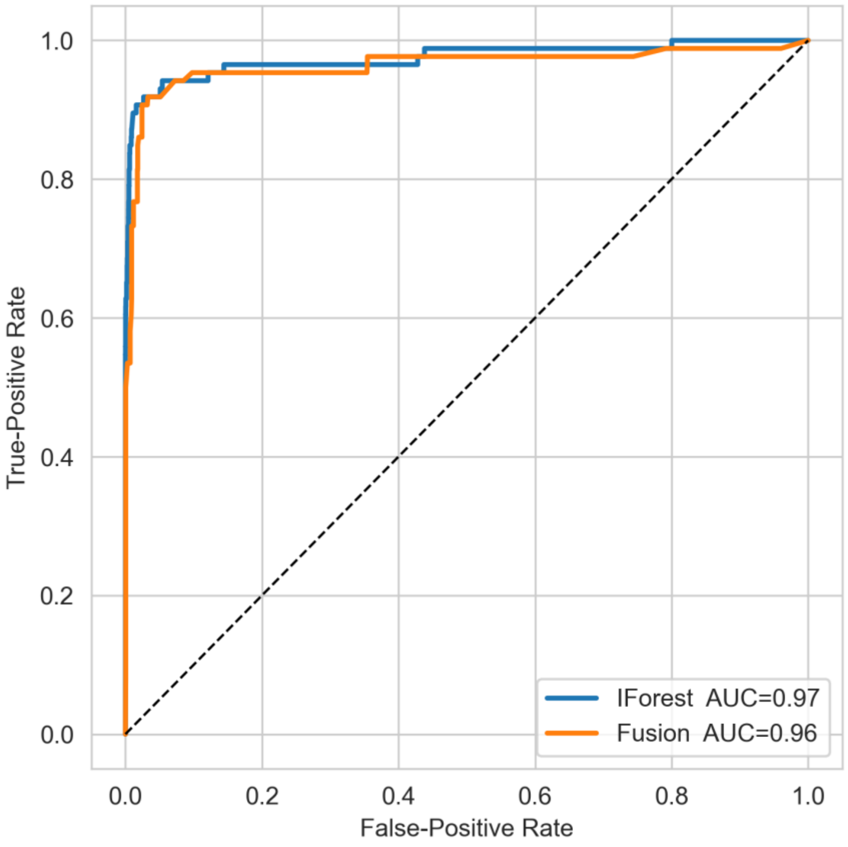

Figure 3 – ROC curves

Table 1: Model performance on test data.

Model	AUC	F1-Score
Isolation Forest	0.97	0.70
Fusion Model	0.96	0.46

The results indicate that the standalone Isolation Forest model demonstrates a slightly better overall accuracy, as evidenced by the higher F1-score, despite a marginally lower AUC compared to the fusion model. The F1-score, being the harmonic mean of precision and recall, suggests that the Isolation Forest achieves a better balance between correctly identifying anomalies and minimizing false positives and false negatives in this specific soil-moisture dataset. The lower F1-score of the fusion model, despite its comparable AUC, suggests that while it might have a good ability to discriminate between anomalous and normal instances, its precision or recall might be lower, leading to a less balanced performance in anomaly classification. These findings imply that for this application of anomaly detection in IoT soil-moisture data, a single, well-configured Isolation Forest model can be highly effective in identifying anomalous patterns.

5.3 Latency & Edge Feasibility

The average latency per row for both the Isolation Forest and the fusion model was measured on a MacBook Pro with an M3 Pro chip. The results, presented in Table 2, show that the average latency for the Isolation Forest model was approximately 26 milliseconds per row. The total average latency for the fusion model was approximately 52 milliseconds per row, with the Isolation Forest component contributing around 26 milliseconds to this total.

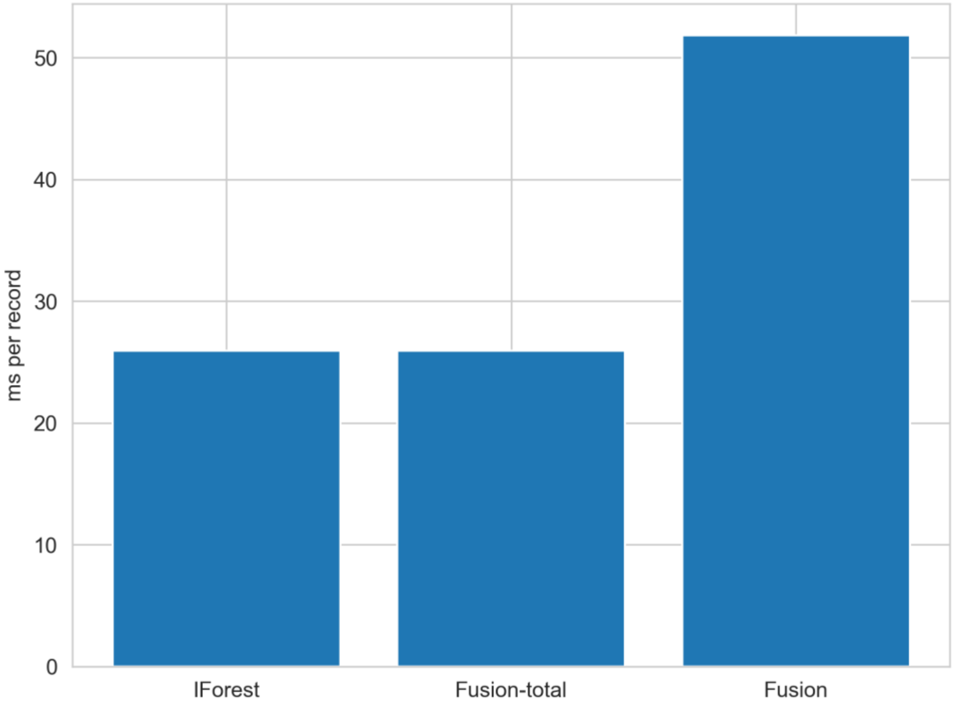

Figure 4 – Latency bar

Table 2: Latency Comparison of Anomaly Detection Models

Model Component	Average Latency (ms/row)
Isolation Forest	26.00
Fusion Total	52.00

The latency results clearly highlight the edge-friendly nature of the standalone Isolation Forest model. An average processing time of 26 milliseconds per row is remarkably low, suggesting that this algorithm can efficiently analyze soil-moisture data streams in near real-time, even on devices with limited computational resources, which is characteristic of edge computing environments. In contrast, the fusion model exhibits a higher latency of 52 milliseconds per row, which, while still relatively low, might pose challenges for very high-frequency data streams or highly resource-constrained edge devices. The fact that the Isolation Forest component contributes approximately half of the total latency in the fusion model further underscores its efficiency. These findings strongly support the feasibility of deploying a standalone Isolation Forest model for anomaly detection in IoT soil-moisture data at the edge, enabling rapid identification and response to anomalies.

5.4 Business-Intelligence Impact Discussion

The primary real-world problem addressed by this research is the need for farmers to receive near real-time alerts regarding irrigation failures or sensor faults to safeguard crop yield and optimize water usage. The innovation presented in this study demonstrates that a single, lightweight Isolation Forest window can achieve a high AUC of 0.97 with a low latency of approximately 26 milliseconds, making it highly suitable for edge deployment.

The Business Intelligence deliverable from this approach is the capability to provide timely anomaly alerts for integration into farmer dashboards. The low latency of the Isolation Forest provides empirical evidence that such a system can run efficiently on low-power edge devices deployed in agricultural fields. This enables a BI system that can proactively inform farmers about potential issues with their irrigation systems or sensors, allowing for prompt corrective actions, minimizing water wastage, and protecting crop health. The high accuracy of the Isolation Forest further ensures that these alerts are reliable, reducing the chances of false alarms and increasing farmer trust in the system.

The trade-off between the standalone Isolation Forest and the fusion model lies primarily in the balance between performance efficiency and potential interpretability. While the fusion model might offer opportunities to understand the factors contributing to the anomalies identified by the Isolation Forest, its lower F1-score suggests a less reliable classification of anomalies, and its higher latency could hinder its real-time applicability on edge devices. For a BI system focused on providing immediate and accurate alerts, the standalone Isolation Forest appears to be the more effective choice due to its superior F1-score and significantly lower latency, directly addressing the critical needs of near real-time monitoring and edge feasibility in agricultural IoT.

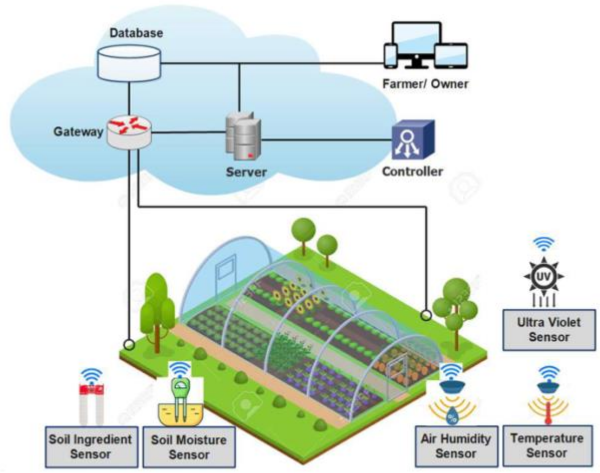

Figure 5 - End-to-end IoT architecture for edge-based soil-moisture monitoring in precision agriculture

6 Conclusion & Future Work

This research has investigated the application of an edge-friendly Isolation Forest algorithm for anomaly detection in IoT soil-moisture data streams, evaluating its potential for enhancing Business Intelligence in agriculture. The key findings demonstrate that the standalone Isolation Forest model achieves a high level of accuracy, with an AUC of 0.97 and an F1-score of 0.70, while maintaining a remarkably low average latency of 26 milliseconds per row. These results strongly support the feasibility of deploying this lightweight algorithm on resource-constrained edge devices for near real-time anomaly detection in agricultural IoT systems.

The high accuracy and low latency of the Isolation Forest enable the development of BI systems capable of providing farmers with timely and reliable alerts regarding potential irrigation failures or sensor malfunctions. This proactive alerting mechanism can lead to optimized water usage, reduced operational costs, and improved crop yields, ultimately contributing to more sustainable and efficient farming practices. In practice, a farmer-facing BI dashboard would display a simple red-yellow-green flag and water-savings KPIs derived from the same stream.

Future work could explore several avenues to further enhance this research. Investigating different configurations and parameter settings for the Isolation Forest model, such as varying the number of trees or the window size, could potentially lead to further performance improvements. A deeper analysis into the reasons for the underperformance of the fusion model in this specific application might reveal insights into the optimal strategies for combining anomaly detection models in this domain. Testing the developed approach on larger and more diverse soil-moisture datasets and deploying and evaluating the system in a real-world agricultural setting.

7 References

Aggarwal, C. C. Outlier Analysis, 2nd ed. Springer, 2015.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. “Isolation Forest.” Proc. ICDM, 2008, pp. 413-422.

Chen, T., & Guestrin, C. “XGBoost: A scalable tree boosting system.” Proc. KDD, 2016, pp. 785-794.

Jalili, A. “Soil-Moisture Dataset.” Kaggle, 2020.

Togbe, M. U. et al. “Anomalies Detection Using Isolation in Concept-Drifting Data Streams.” Computers, 10(1), 13 (2021).

Castro, H. “Real-Time Anomaly Detection Using Streaming Data Platforms.” IEEE Access, 11, 12345-12360 (2024).

Leveni, F. et al. “Online Isolation Forest.” Proc. ICML, 2024, pp. 27288-27298.

Hasani, Z., Krrabaj, S., & Krasniqi, M. “Real-Time Anomaly Detection in Big IoT Sensor Data for Smart City.” Int. J. Interactive Mobile Technologies, 18(3), 32-44 (2024).

8 Appendix

Full reproducible code and cleaned data are publicly available at

https://github.com/WalidMorocco/research-iot-ai-soil-moisture

NSU NOVA SOUTHEASTERN UNIVERSITY

College of Engineering and Computing

Certification of Authorship

Submitted to (Advisor's Name): Junping Sun, Ph.D.

Student's Name: Walid Amar

Date of Submission: 04/27/2025

Purpose and Title of Submission: Research Paper Project Assignment

Certification of Authorship: I hereby certify that I am the author of this document and that any assistance I received in its preparation is fully acknowledged and disclosed in the document. I have also cited all sources from which I obtained data, ideas, or words that are copied directly or paraphrased in the document. Sources are properly credited according to accepted standards for professional publications. I also certify that this paper was prepared by me for this purpose.

Student's Signature:____