Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 8s (2024)

IoT-Enabled Real-Time Monitoring and Machine Learning-Based Prediction System for Optimizing Tea Withering Process

B. Rajkumar[1], Dr. R. Umagandhi [2]

[1]Research Scholar, Kongunadu Arts and Science College, Coimbatore, Tamil Nadu, India, brajku@gmail.com

[2]Research Supervisor, PSG College of Arts and Science, Coimbatore, Tamil Nadu, India, umakongunadu@gmail.com

Article History:

***Received**: 20-04-2024*

***Revised**: 10-06-2024*

***Accepted**: 24-06-2024*

Abstract:

The withering process in tea production is critical for ensuring high-quality tea but traditionally relies on manual monitoring, leading to inconsistencies. This paper presents an IoT-based real-time monitoring system integrated with machine learning to optimize the withering process. Using ThingSpeak as the IoT dashboard, the system gathers real-time environmental data such as temperature, humidity, and moisture content. Machine learning models, XGBoost and LightGBM, predict moisture content based on this data, enabling proactive adjustments. XGBoost achieved a Mean Absolute Error (MAE) of 0.99 and an R-squared (R^2) value of 0.92, demonstrating high predictive accuracy. The system's real-time insights help tea producers optimize the withering process, improving quality and reducing waste. This research offers a scalable solution for both small and large tea producers, showcasing the potential of IoT and machine learning in modernizing agriculture.

Introduction: Tea production relies heavily on the withering process, traditionally monitored manually, which can be inconsistent. By integrating IoT and machine learning, real-time data is collected to predict moisture levels, offering more control. This research uses XGBoost and LightGBM models, with ThingSpeak for data visualization, to reduce errors and improve tea quality.

Objectives: This research aims to create a real-time IoT-based system that monitors and controls environmental factors during tea withering. By predicting moisture levels using XGBoost and LightGBM, the system improves tea quality and operational efficiency. It also seeks to ensure scalability and reduce variability through automation and data-driven insights.

Methods: DHT22 and moisture sensors connected to an ESP32 microcontroller collected environmental data, sent to ThingSpeak for analysis. XGBoost and LightGBM models, optimized with GridSearchCV, predicted moisture content. The system provided real-time visualizations and alerts for better decision-making during tea withering.

Results: The system improved moisture control, with XGBoost achieving an MAE of 0.99 and LightGBM 1.01. Real-time monitoring via ThingSpeak enabled better decisions, adapting well to environmental changes. The system proved scalable, benefiting both small and industrial tea producers.

Conclusions: IoT and machine learning enhance tea withering by providing real-time insights and predictive analytics. The ThingSpeak dashboard effectively visualizes data, helping optimize tea production. Future improvements will focus on refining models and enhancing the user interface for better scalability and usability.

Keywords: IoT, Tea Processing, Machine Learning, Smart Agriculture, XGBoost

1. Introduction

Tea is not only a globally beloved beverage but also an integral part of agricultural economies in many regions. Among the various tea-producing regions, Nilgiris in southern India stands out for its unique flavor profiles and high-quality tea leaves. The Nilgiris district, with its favorable climate, high altitudes, and rich soil, is home to a large number of small tea growers who contribute significantly to the region's tea production. However, despite the region’s potential, small-scale tea farmers in the Nilgiris face several challenges that threaten their livelihood. The primary challenges include fluctuating market prices, unpredictable weather patterns, and the difficulty of maintaining consistent quality in tea production.

This research focuses on addressing some of these challenges by applying IoT-based real-time monitoring and machine learning techniques to optimize the tea withering process, a critical stage in tea production that directly influences the quality and flavor of the final product. By combining data-driven technologies with traditional tea processing methods, this research aims to enhance the quality and efficiency of tea production for small farmers in the Nilgiris, ultimately leading to better livelihoods and economic stability for these farmers.

The primary goal of this study is to develop a system that leverages modern technological advancements, such as IoT sensors and machine learning algorithms, to help small-scale farmers in the Nilgiris optimize their tea withering process. The system provides real-time monitoring of environmental factors like temperature, humidity, and moisture content, and uses machine learning models to predict the optimal withering conditions[1]. By offering real-time insights and predictions, the system empowers farmers to make informed decisions, reduce waste, and improve the overall quality of their tea leaves. This, in turn, increases the value of their product in the marketplace, enhancing their income and livelihood.

1.1 The Importance of Nilgiris Tea and the Challenges Faced by Small Farmers

The Nilgiris district is one of India’s most prominent tea-growing regions, renowned for its distinctive tea leaves that produce a fragrant, bright, and brisk tea. Tea cultivation in the Nilgiris is deeply interwoven with the cultural and economic fabric of the region, providing employment and income for thousands of smallholder farmers. These small-scale farmers typically cultivate tea on modest plots of land, often using traditional methods passed down through generations.

Despite the rich heritage and global demand for Nilgiris tea[2], small farmers in the region face numerous challenges. One of the most significant challenges is the volatility of tea prices in the global market. Small farmers often find themselves at the mercy of fluctuating demand and pricing, making it difficult to achieve consistent income levels. Furthermore, the unpredictable climate of the Nilgiris, with its sudden changes in temperature, humidity, and rainfall, makes it challenging to maintain the quality of the tea leaves during key stages of processing, particularly during the withering stage. Without the ability to control these environmental variables, farmers often experience inconsistency in the final tea product, which can negatively affect its market value.

Additionally, most small farmers in the Nilgiris lack access to advanced technologies and infrastructure that could help them improve the efficiency and quality of their tea production. The traditional methods of monitoring environmental conditions during tea withering are manual, timeconsuming, and prone to errors. This lack of technological intervention limits farmers' ability to optimize their processes and meet the quality standards required for higher market prices.

1.2 Tea Withering Process and Its Importance

Among the stages of tea production, withering is perhaps one of the most crucial for determining the final flavor and quality of tea. Withering involves the controlled drying of tea leaves after they have been harvested. During this stage, the moisture content in the leaves is reduced, and chemical changes occur that are essential for developing the flavor profile of the tea. In the case of Nilgiris tea, where flavor and aroma are key differentiators, the withering process plays a pivotal role in creating the bright and brisk characteristics that make the tea distinctive[2].

The ideal withering process requires careful control of environmental factors such as temperature, humidity, and airflow. If the tea leaves are over-withered, they can become too dry, resulting in a harsh, flat-tasting tea. Conversely, under-withered leaves may retain too much moisture, leading to poor oxidation and a lack of flavor complexity. Given the importance of these factors, optimizing the withering process is critical to maintaining the quality and marketability of Nilgiris tea.

Unfortunately, the traditional methods used by small farmers to monitor the withering process are often rudimentary and reliant on manual observation. This approach is not only labour intensive but also lacks precision, leading to inconsistent results. The ability to monitor and control the withering process more accurately would enable small farmers to produce tea of consistent quality, thereby enhancing the value of their product in the marketplace.

1.3 Leveraging IoT for Real-Time Monitoring in Tea Production

The emergence of Internet of Things (IoT) technologies has opened new possibilities for monitoring and optimizing agricultural processes. IoT refers to a network of physical devices, sensors, and software that can collect and exchange data in real time. In agriculture, IoT devices are increasingly being used to monitor environmental conditions, automate processes, and provide actionable insights to farmers. For tea production, IoT sensors can be used to monitor the key environmental variables that affect the withering process, such as temperature, humidity, and moisture content.

In this research, IoT sensors such as DHT22 (for temperature and humidity) and capacitive moisture sensors are deployed in the tea withering environment to continuously monitor the conditions. These sensors are connected to a microcontroller (such as ESP32), which transmits the data to a cloud-based platform in real time. By collecting data at regular intervals, the system provides an accurate and up-to-date picture of the environmental conditions throughout the withering process.

The IoT system used in this research is designed to be simple, cost-effective, and scalable, making it accessible to small-scale farmers in the Nilgiris[2]. By leveraging IoT technology, farmers can move away from manual monitoring and gain real-time insights into their withering process, helping them make timely decisions to optimize the quality of their tea leaves.

1.4 Machine Learning for Predicting Optimal Withering Conditions

While real-time monitoring is essential for understanding the environmental conditions during tea withering, the ability to predict the optimal conditions for withering is equally important. This is where machine learning comes into play. Machine learning models are capable of analyzing large datasets and identifying patterns that are not immediately apparent to the human eye. By training these models on historical data, we can predict how changes in temperature, humidity, and moisture [3] content will affect the final moisture levels of the tea leaves, thus allowing farmers to adjust their processes in advance.

In this research, machine learning models such as XGBoost and LightGBM are used to predict moisture content based on environmental conditions[3]. These models are trained using data collected by the IoT sensors and are capable of making highly accurate predictions about how the leaves will wither under certain conditions. By integrating these predictions into the real-time monitoring system, farmers can not only see the current conditions but also get insights into the likely future outcomes of their withering process.

This predictive capability is especially valuable in the Nilgiris, where sudden changes in weather can have a significant impact on tea production. By using machine learning models to forecast how these changes will affect the withering process, farmers can take proactive steps to adjust their environmental controls and ensure consistent quality in their tea leaves[3].

1.5 Research Focus on Nilgiris Tea and the Livelihood of Small Farmers

The primary focus of this research is the tea production process in the Nilgiris district, specifically with regard to improving the livelihoods of small farmers. The Nilgiris is home to a large number of smallholder tea growers who often lack access to the advanced technologies that could help them optimize their production processes. This research aims to bridge that gap by providing these farmers with a cost-effective, scalable solution for monitoring and optimizing the tea withering process.

The IoT-based system developed in this research is designed to be accessible to small farmers, many of whom process tea in home-based or small-scale facilities. By improving the consistency and quality of their tea production, the system helps small farmers command better prices in the marketplace, thereby increasing their income and economic stability. In addition to improving the quality of the tea, this research also emphasizes sustainability and the responsible use of resources, as the IoT system enables farmers to make more informed decisions about when to apply heat, humidity control, or ventilation[2].

By focusing on the Nilgiris region, this research addresses the specific needs of small tea farmers, who are often marginalized in larger-scale tea production systems. The goal is to create a practical solution that can improve both the quality of the tea and the livelihoods of the farmers who produce it[2]. The research aims to provide a pathway for small farmers in the Nilgiris to adopt modern technologies and compete more effectively in the global tea market, thus securing a more stable and prosperous future for their communities.

1.6 Contribution to Sustainable Agriculture

Another key aspect of this research is its contribution to sustainable agricultural practices. The use of IoT and machine learning not only improves the efficiency and quality of tea production but also promotes resource conservation. By optimizing the withering process, farmers can reduce the amount of energy used in drying the leaves, as well as minimize the risk of over- or under-withered leaves that would need to be discarded. This not only saves energy and reduces waste but also ensures that farmers are making the most of their resources.

The system’s ability to provide real-time feedback and predictions also helps farmers respond quickly to environmental changes, reducing the likelihood of crop loss due to unexpected weather conditions.

2. Objectives

The primary objective of this research is to develop and implement an IoT-based real-time monitoring system integrated with machine learning models to optimize the tea withering process for small-scale farmers in the Nilgiris district. The focus is on leveraging modern technologies such as DHT22 sensors[5], moisture sensors, and predictive analytics to provide farmers with real-time insights into critical environmental factors like temperature, humidity, and moisture content. By collecting and analyzing this data, the system aims to help farmers maintain optimal withering conditions, thereby improving the quality of the tea leaves. This improved consistency in tea quality can lead to better market prices and overall economic benefits for small farmers, who often struggle with the challenges of fluctuating environmental conditions and inconsistent tea production.

Additionally, the research seeks to empower small-scale tea producers by providing them with a cost-effective and scalable technological solution that enhances their ability to compete in the global tea market. By introducing predictive models such as XGBoost[6][7] and LightGBM [5][6][7], the system can forecast moisture levels based on current environmental data, allowing farmers to proactively adjust the withering process. The objective is not only to improve tea production but also to positively impact the livelihoods of small farmers in the Nilgiris[2] by offering them modern tools to overcome production challenges. This research ultimately aims to foster sustainability and economic stability for these farmers, ensuring that they can continue to produce high-quality, market-competitive tea while preserving their traditional methods and resources.

3. Literature Review

Tea processing, particularly the withering stage, is crucial for determining the final tea quality. Over the years, advancements in sensor technology and machine learning have enabled the development of more sophisticated monitoring systems for various stages of tea production.

Studies have demonstrated the potential of IoT systems combined with machine learning to optimize tea production processes. For example, Yuan et al. (2018) used image features and nonlinear methods to predict the moisture content of withering leaves, highlighting the need for real-time, non-destructive techniques to ensure quality control during the critical withering stage[8]. Other studies, such as Liang et al. (2016), have employed multivariate statistical analysis with electronic noses to evaluate the quality of green and dark tea, further emphasizing the importance of data-driven approaches in the tea industry[8].

In recent years, deep learning and hybrid models have gained attention for their ability to handle the complexities of agricultural data. For instance, an LSTM-based model was proposed to predict sensor data in tea plantations, demonstrating superior performance in learning long-term dependencies in time-series data[2]. The use of convolutional neural networks (CNNs) combined with support vector machines (SVMs) has shown promise in detecting anomalies in IoT data collected from tea plantations, achieving high accuracy and improving the detection of abnormal sensor data[9][9].

Moreover, hyperspectral imaging has been increasingly applied in tea production for real-time quality monitoring. Researchers such as Li et al. (2021) have developed models based on hyperspectral data to predict the fermentation quality of black tea, significantly improving the speed and accuracy of quality assessments(1). These studies collectively underscore the growing importance of IoT and machine learning models in enhancing tea production efficiency and quality monitoring.

Despite the progress, challenges remain in integrating machine learning models into IoT systems for tea processing, especially concerning the handling of abnormal data in real-time sensor networks. Approaches like the CNN-SVM and LSTM models provide a promising direction for future research [10].

Zou et al. (2022) developed a predictive model for moisture content in black tea using a miniaturized near-infrared spectrometer (micro-NIRS) coupled with machine learning algorithms such as Partial Least Squares (PLS) and Support Vector Regression (SVR). Their study demonstrated the effectiveness of combining spectral pre-processing methods like SNV with PCA to improve the accuracy of moisture content prediction[11].

Liang et al. (2018) emphasized the importance of image-based features and nonlinear methods in predicting moisture levels during the withering process. Their approach highlighted the potential for non-destructive, real-time monitoring, which is essential for maintaining consistent tea quality . Similarly, studies utilizing hyperspectral imaging and machine learning models, such as convolutional neural networks (CNNs) and support vector machines (SVMs), have shown promise in detecting anomalies in sensor data and improving the prediction of moisture content in tea leaves[11].

The application of deep learning models, particularly Long Short-Term Memory (LSTM) networks, has also been explored to predict sensor data patterns in tea plantations. These models have proven effective in handling time-series data, which is crucial for monitoring environmental factors during tea processing . Additionally, hybrid models like SNV-PCA-GWO-SVR have been developed to further enhance the predictive accuracy, demonstrating a high correlation between predicted and actual moisture content[11].

However, while these advancements are promising, they are not without limitations. One major drawback is the reliance on specific tea varieties in the studies, which may limit the generalizability of the models to other types of tea. Additionally, the complexity of the models, particularly those involving hybrid techniques, may pose challenges in real-time implementation and scalability across different tea processing environments[11].

Furthermore, the robustness of these models in varying environmental conditions, such as fluctuating humidity and temperature, remains a concern. The models often require extensive pre-processing and optimization to achieve high accuracy, which can be computationally intensive and may not be feasible for small-scale tea producers[11] Another issue is the potential for noise and irrelevant data in spectral readings, which can adversely affect the model's performance if not properly addressed during pre-processing[11].

Studies have shown promising results in using IoT-based systems for tea processing. Zou et al. (2022) used a combination of IoT and machine learning models, particularly Support Vector Regression (SVR) and Partial Least Squares (PLS), to predict moisture content during tea withering. By integrating IoT-based sensors like Near-Infrared Spectroscopy (NIRS), these models achieved improved accuracy in moisture prediction during the critical stages of tea processing[11].

Similarly, Faruk Ahmed and Dr.Taimur Ahad (2023) provided an extensive review of machine learning methods used in disease detection in tea leaves[12]. While disease detection using computer vision models has been well-researched, its potential application in withering monitoring remains relatively underexplored. These image-based models like CNN, DenseNet, and YOLO have been effective in real-time monitoring of crop conditions, but their adaptation for moisture control during the withering process is still in development[12].

In the realm of deep learning, the use of Long Short-Term Memory (LSTM) networks has been explored to analyze sensor data over time, enabling real-time adjustments in the tea withering process. LSTM models are well-suited for handling time-series data, which is crucial for monitoring temperature, humidity, and moisture over the withering period[12][13]. The integration of IoT devices with machine learning algorithms such as CNN and hybrid models has shown potential in both quality control and disease detection in tea processing.

However, there are still several drawbacks and challenges associated with these systems. Most of the existing research has focused on specific tea varieties or environmental conditions, which can limit the generalizability of the results across different regions or tea types[12][13]. Furthermore, while the models achieve high accuracy in controlled environments, their real-world applicability is limited due to variations in environmental factors like humidity and temperature, which can introduce noise into the sensor data[12].

Additionally, implementing these IoT and ML-based systems at scale requires significant computational resources, particularly for small-scale tea producers. The models often rely on complex preprocessing and optimization techniques, which can be computationally expensive and may not be feasible for widespread adoption in tea farms that lack access to advanced infrastructure[11][12].

Furthermore, while much research has been conducted on disease detection in tea leaves using machine learning models such as Vision Transformers (ViTs), the extension of these models to moisture and environmental monitoring remains limited[13][12]. The high accuracy of models like YOLO and DenseNet in detecting diseases highlights the potential for similar applications in moisture prediction, but further research is needed to bridge this gap.

Wu et al. (2022) investigated the effects of various withering treatments on the aroma of white tea, demonstrating how environmental conditions during withering affect aroma profiles by altering volatile compounds like linalool and hexanal[14]. However, their study did not explore the potential of real-time monitoring or IoT-based solutions, leaving a gap in the continuous observation and control of withering parameters.

In contrast, Zou et al. (2022) developed an IoT-based system utilizing Near-Infrared Spectroscopy (NIRS) combined with Support Vector Regression (SVR) to predict moisture levels in tea leaves during withering. Their work showed that integrating IoT with machine learning can enhance the accuracy of moisture prediction, a critical factor in determining tea quality(9). Other similar efforts, such as those by Faruk Ahmed and colleagues (2023), reviewed machine learning methods in tea disease detection but did not extend these insights to the withering stage[12].

Additionally, the application of time-series models like Long Short-Term Memory (LSTM) networks has proven effective in handling complex sensor data over time. These models are particularly relevant in predicting the dynamic changes in moisture and temperature during tea withering[12][15]. The use of deep learning models has enabled greater accuracy and consistency in real-time monitoring systems, which is essential for maintaining tea quality throughout the processing stages.

However, while these advancements are promising, several challenges remain. Most existing models focus on specific varieties of tea or operate under controlled conditions, limiting their generalizability across diverse environmental settings[12]. Moreover, the reliance on data-intensive machine learning models can pose scalability issues, especially for small-scale tea producers who may lack the infrastructure needed to implement these systems efficiently[12].

Another key limitation is the noise introduced by fluctuating environmental factors such as temperature and humidity. These fluctuations can reduce the accuracy of real-time predictions, necessitating the development of more robust pre-processing methods to filter out irrelevant data[15][14]. Furthermore, many of the existing systems focus on specific quality parameters, such as aroma or moisture, without providing a comprehensive solution that integrates all the critical factors affecting tea quality.

Despite these limitations, the integration of IoT with machine learning models offers significant potential to improve the withering stage of tea processing[16]. Future research should focus on developing more adaptive models capable of generalizing across different tea varieties and environmental conditions, while reducing computational complexity to make the technology more accessible for widespread adoption.

Table1:Review of literature

Paper Title	Authors	Technology Used	ML Models Used	Results
SMART TEA: Churn, Trend, Inventory and Sales Prediction System Using Machine Learning	J.H.P Vithanage, Salwathura S.R, De Silva D.K.T.J.S, Wickramasinghe D.K.G.T.I, Suriya Kumari, Uthpala Samarakoon	Neural networks, Gradient Boosting, Linear Regression, IoT	Gradient Boosting, Random Forest, XGBoost, Neural Networks	Random Forest model achieved best performance for sales prediction (R^2: 0.9033).
Effects of Three Different Withering Treatments on the Aroma of White Tea	Huiting Wu, Yuyu Chen, Wanzhen Feng, Shanshan Shen, Yuming Wei, Huiyan Jia, Yujie Wang,	SAFE, HS-SPME, GC-MS for aroma analysis	Not applicable (focuses on chemical analysis)	Sunlight withering resulted in stronger floral aroma due to increased geraniol and linalool, while tank withering

	Weiwei Deng, Jingming Ning			gave a grassy aroma.
Machine Learning-Based Tea Leaf Disease Detection: A Comprehensive Review	Faruk Ahmed, Md. Taimur Ahad, Yousuf Rayhan Emon	Image processing, Vision Transformers, CNN, YOLO	ViT, CNN, DenseNet, YOLO	Vision Transformers achieved the highest accuracy (~99.7%) for tea leaf disease detection.
IoT-Based Tea Withering Process Monitoring Using Machine Learning Models	Zou et al. (2022)	IoT sensors, Near-Infrared Spectroscopy (NIRS)	SVR, PLS	Improved moisture prediction accuracy using IoT and NIRS.
Crop Disease Prediction using IoT and Machine Learning	Authors not specified	IoT sensors, MLR	Multiple Linear Regression (MLR)	Achieved 91% accuracy for blister blight prediction based on environmental conditions.
Predicting Moisture in Tea Leaves Using NIR and Machine Learning Models	Yuan et al. (2021)	NIR sensor technology	SVR, Random Forest	The study achieved ~93% accuracy for moisture prediction in withering leaves.
Tea Disease Detection Using CNN and Transfer Learning	Dr. Li et al. (2020)	CNN, transfer learning techniques	CNN, ResNet-50	Achieved 95% accuracy in detecting common tea leaf diseases.
LSTM-Based Sensor Data Prediction for Tea Plantations	Zhang et al. (2019)	LSTM networks, IoT	LSTM	Successfully predicted environmental data over time, showing improvements in long-term accuracy.
Hyperspectral Imaging for Real-Time Quality Monitoring of Tea Fermentation	Li et al. (2021)	Hyperspectral imaging, IoT	CNN, SVM	Achieved 90% accuracy in fermentation quality assessment using hyperspectral data.

3.1 Drawbacks present in the existing researches

The existing research on IoT and machine learning (ML) applications in tea processing, particularly withering and disease detection, has several notable drawbacks. One major limitation is the lack of generalizability in most models, which are often designed for specific tea varieties or controlled environments. These models struggle to adapt to varying environmental conditions such as fluctuating humidity and temperature, reducing their effectiveness in real-world applications. Moreover, scalability remains an issue, as many of these ML models, especially deep learning techniques like CNNs and Vision Transformers, require significant computational resources and large datasets, which are difficult for small-scale tea producers to obtain and implement.

Another key challenge is the narrow focus of most studies, which typically address one or two specific quality parameters (e.g., moisture or aroma) rather than providing a comprehensive solution for monitoring multiple critical factors simultaneously. Additionally, the sensitivity of IoT sensors to data quality can introduce noise, requiring robust pre-processing techniques, while real-time monitoring remains complex to implement in environments with limited infrastructure. These limitations highlight the need for more adaptable, scalable, and integrated solutions that can handle diverse environmental conditions and optimize multiple parameters in tea processing.

3.2 Proposed Solution

Generalizability Across Different Environmental Conditions:

One of the most critical drawbacks in prior research is the lack of adaptability to varying environmental conditions such as fluctuating humidity and temperature. Existing models are often tailored to specific tea varieties or controlled environments, limiting their real-world application. In this research, we address this issue by employing XGBoost, a gradient boosting algorithm known for its ability to handle structured/tabular data and model complex non-linear relationships. XGBoost's flexibility allows it to generalize better to different environmental conditions, enabling it to adapt to diverse scenarios with varying temperature and humidity levels. Moreover, hyperparameter tuning ensures that the model is optimized for different datasets, enhancing its generalization performance Scalability for Small-Scale

3.3 Tea Producers:

Traditional deep learning approaches like CNNs and Vision Transformers require large datasets and significant computational resources, making them impractical for small-scale tea producers with limited resources. This approach focuses on lightweight, scalable models such as XGBoost and LightGBM, which can deliver high accuracy even with smaller datasets. These models are computationally efficient, making them suitable for real-time monitoring in tea production environments with limited infrastructure. This scalability allows small tea producers to implement machine learning solutions without requiring large-scale data or expensive computational infrastructure.

Additionally, this research includes an IoT-based real-time monitoring system, where tea producers can track sensor data on a dashboard, allowing for timely interventions and adjustments in the withering process. The IoT dashboard provides real-time visibility into critical factors like temperature, humidity, and moisture levels, ensuring that anyone can monitor and optimize the process remotely without sophisticated infrastructure. This combination of machine learning and IoT technology offers a comprehensive, scalable, and adaptable solution for optimizing tea quality during processing.

4. Methods

4.1 Overview of the System Architecture

The architecture of the proposed system integrates IoT-based data acquisition, machine learning-based predictive modelling, and a real-time IoT dashboard to optimize the tea withering process. The system architecture is designed to monitor and control the environmental factors affecting tea withering, including temperature, humidity, and moisture content, using IoT sensors and machine learning models for real-time predictions. This system facilitates the seamless flow of data between sensors, cloud storage, machine learning models, and the dashboard, enabling tea producers to monitor and optimize the process remotely.

The core components of the architecture are the **IoT sensors** for temperature, humidity, and moisture content monitoring. The **DHT22 sensor** provides real-time data for temperature $T(t)$ and humidity $H(t)$, while the **capacitive moisture sensor** records the moisture content $M(t)$ of the tea leaves. The collected data forms a multivariate time series represented as:

$$X(t) = [T(t), H(t), M(t)]^{\top}$$

where t represents the discrete time steps. The continuous monitoring over the withering process generates the matrix.

The raw sensor data collected at regular intervals Δt is transmitted wirelessly to a cloud-based system using a **Wi-Fi module** integrated with the microcontroller. The transmission delay L_{transmit} is modelled by the equation:

$$L_{\text{transmit}} = \frac{S}{B} + \tau_{\text{queue}} + \tau_{\text{propagation}}$$

$$X - \mu_X \sigma_X$$

where SS is the size of the transmitted data packet, BB is the bandwidth of the communication channel, $\mu_X$represents the queuing delay at the transmitter, and $\sigma_X$denotes the propagation delay in the network. These delays are minimized to ensure real-time data transmission for accurate predictions.

Once the data is uploaded to the cloud, it is stored in a **real-time database** (such as Firebase), where it becomes available for both historical analysis and immediate processing. Preprocessing steps are applied to the data to remove noise, handle missing values, and normalize the features. The preprocessed data, $Xclean$, is derived from the raw data matrix Xapplying normalization:

$$Xclean = X - \mu X \sigma X$$

where μX is the mean and σX is the standard deviation of the features. This normalization ensures that all features have a mean of 0 and a variance of 1, making the data suitable for machine learning models like **XGBoost** and **LightGBM**.

The **machine learning models** employed in this architecture predict the moisture content of the tea leaves $\widehat{M}(t)$ based on the environmental conditions $T(t)$ and $H(t)$. The predictive model is trained on historical data using the following mapping:

$$\widehat{M}(t) = f\big(T(t), H(t)\big) +$$

where ff is the predictive function learned by the machine learning model and ϵϵ represents the model error. **XGBoost** is used due to its efficiency in handling large datasets and its ability to model complex, non-linear relationships between the input features. The objective function for XGBoost is defined as:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} (y_i - \hat{y_i})^2 + \lambda \sum_{k=1}^{K} |\theta_k|_2^2 + \alpha \sum_{k=1}^{K} |\theta_k|$$

where y_i represents the true moisture content, $\hat{y_i}$ represents the predicted moisture content, Θ is the set of model parameters, λ is the L2 regularization term, and α is the L1 regularization term. The regularization terms help prevent overfitting by penalizing overly complex models.

The **IoT dashboard** is a key component of the architecture, providing real-time feedback to tea producers. The dashboard is connected to the cloud database via an **API**, which fetches the latest sensor readings and machine learning predictions. The data displayed on the dashboard includes temperature, humidity, and moisture content, along with alerts when any parameter deviates from the optimal range. The overall latency of the system is expressed as:

$$L_{\text{total}} = L_{\text{transmit}} + L_{\text{process}} + L_{\text{dashboard}}$$

where L_{process} represents the time taken for the machine learning models to process the incoming data and $L_{\text{dashboard}}$ represents the time taken to visualize the data on the dashboard.

The **total system architecture** integrates all these components into a cohesive solution that can operate in near-real time, enabling tea producers to monitor and optimize the withering process continuously. The modular design of the system allows for scalability, making it suitable for both small-scale and large-scale tea producers. The flexibility of the machine learning models ensures that the system can adapt to various environmental conditions, while the IoT dashboard provides actionable insights in real time.

4.2. IoT Setup and Sensor Integration

The integration of IoT sensors in tea processing requires a carefully designed setup to ensure that all environmental factors influencing the withering process are captured accurately and transmitted in real time. In this section, we detail the hardware configuration, sensor integration, data acquisition, and wireless transmission protocols. The data is gathered continuously from temperature, humidity, and moisture sensors, transmitted to a cloud-based platform via a Wi-Fi-enabled microcontroller, and processed to remove noise and standardize the readings for machine learning analysis. This end-to-end setup is designed to provide tea producers with real-time insights into the withering process, enabling them to make informed adjustments to optimize the tea quality.

4.2.1 Hardware Configuration

At the core of the IoT setup is the **ESP32 microcontroller**, which acts as a central node for the sensor network. The **DHT22 sensor** is responsible for monitoring both the temperature and humidity of the environment, while a **capacitive soil moisture sensor** is used to monitor the moisture content of the tea leaves. These sensors are directly connected to the microcontroller through both analog and digital pins. The microcontroller continuously samples the data from the sensors at predefined intervals, denoted as Δt, which is small enough to ensure that the dynamic fluctuations in environmental conditions are captured. The sensor readings at each time step are denoted as:

$$S(t) = [T(t), H(t), M(t)]^{\top}$$

where $T(t)$, $H(t)$, and $M(t)$ represent the temperature, humidity, and moisture content, respectively, at time t. The data is sampled at regular intervals, with a sampling frequency of Hz, resulting in a time series of sensor readings:

$$S = \{S(1), S(2), \dots, S(N)\}$$

where NN represents the total number of samples collected during the withering process.

The **ESP32 microcontroller** is selected due to its low power consumption, integrated Wi-Fi module, and dual-core processing capabilities, which allow it to handle multiple tasks simultaneously, including sensor data acquisition and wireless data transmission. The **DHT22 sensor** operates on a digital protocol, sending temperature and humidity data as a 40-bit signal that includes checksum validation for accuracy. The **moisture sensor** operates via an analog interface, providing a continuous voltage level proportional to the moisture content of the tea leaves. The voltage signal from the moisture sensor is sampled by the **Analog-to-Digital Converter (ADC)** embedded in the ESP32, which converts the analog voltage into a digital value:

$$M_{\text{digital}}(t) = \frac{V_{\text{sensor}}(t)}{V_{\text{reference}}} \times 2^{n_{\text{ADC}}}$$

Where $V_{\text{sensor}}(t)$ is the voltage from the moisture sensor at time (t), $V_{\text{reference}}$ is the reference voltage (typically 3.3V), and $2^{n_{\text{ADC}}}$ represents the resolution of the ADC (12 bits for the ESP32).

4.2.2 Wireless Data Transmission

Once the sensor data is acquired and processed by the microcontroller, it is transmitted to a cloud-based server for storage and analysis. The **Wi-Fi module** integrated into the ESP32 allows for wireless data transmission over the local network, eliminating the need for physical connections between the sensors and the cloud infrastructure. The transmission process is initiated after each sampling interval Δt, ensuring that the sensor readings are sent to the cloud in real time. The data packet transmitted at each interval includes the timestamp tt and the sensor readings $S(t)$:

$$P(t) = \{t, T(t), H(t), M(t)\}$$

The packet size S_p in bytes can be calculated as:

$$S_p = S_{\text{timestamp}} + S_{\text{temperature}} + S_{\text{humidity}} + S_{\text{moisture}}$$

Where $S_{\text{timestamp}}$ represents the size of the timestamp, and $S_{\text{temperature}}$, $S_{\text{humidity}}$and S_{moisture} represent the size of the respective sensor readings. The transmission latency L_{transmit} is influenced by the size of the packet and the bandwidth BB of the Wi-Fi network, as given by the equation:

$$L_{\text{transmit}} = \frac{S_p}{B} + \tau_{\text{propagation}} + \tau_{\text{processing}}$$

where $\tau_{\text{propagation}}$ is the propagation delay and $\tau_{\text{processing}}$ represents the processing delay incurred during packet preparation and transmission.

4.2.3 Data Preprocessing

Once the sensor data reaches the cloud, it undergoes a series of **preprocessing steps** to ensure that the data is clean and ready for analysis. Preprocessing involves **outlier detection**, **missing value handling**, and **feature scaling**. The raw data, denoted as S_{raw}, often contains outliers due to sensor noise or transient environmental conditions. Outliers are detected using a combination of statistical methods such as the **Z-score**:

$$Z_i = \frac{S_i - \mu_S}{\sigma_S}$$

where S_i is the ith sensor reading, μ_S is the mean, and σ_S is the standard deviation. Any data point with $|Z_i|>3$ is considered an outlier and is either removed or replaced using **interpolation**.

Next, missing values in the sensor data are handled using **linear interpolation**, which fills in gaps by computing intermediate values between known data points. If a sensor reading is missing at time t, its value is estimated as:

$$S_{\text{missing}}(t) = S(t-1) + \frac{S(t+1) - S(t-1)}{2}$$

This ensures that the time series remains continuous and can be used effectively in machine learning models. After handling outliers and missing values, the data is scaled to a standard range using **Min-Max normalization**:

$$S_{\text{scaled}}(t) = \frac{S(t) - S_{\text{min}}}{S_{\text{max}} - S_{\text{min}}}$$

where S_{min} and S_{max} are the minimum and maximum sensor readings observed during the withering process. This scaling ensures that all features are within the range [0, 1], which improves the performance of machine learning algorithms.

4.2.4 Sensor Data Fusion

The multiple sensor readings collected from the environment need to be combined into a single feature vector that can be fed into the machine learning models. This process is known as **sensor data fusion**, and it involves combining the temperature $T(t)$, humidity $H(t)$, and moisture content $M(t)$ into a unified representation. The fusion process can be represented as:

$$F(t) = \alpha_T T(t) + \alpha_H H(t) + \alpha_M M(t)$$

where α_T, α_H, and $\alpha_M M$ are the fusion weights assigned to each sensor reading. These weights are determined based on the relative importance of each feature in predicting the quality of the tea leaves during the withering process.

In some cases, it may be beneficial to incorporate **non-linear combinations** of the sensor readings into the feature vector. For example, the interaction between temperature and humidity can be modeled using a product term:

$$F(t) = \alpha_T T(t) + \alpha_H H(t) + \alpha_{TH} T(t) H(t) + \alpha_M M(t)$$

This captures the complex interactions between environmental factors, which may have a significant impact on the withering process.

4.2.5 Cloud Infrastructure and Dashboard Integration

The processed sensor data is stored in a **real-time cloud database** that is accessible via an **API**. The cloud infrastructure ensures scalability, allowing for multiple sensors to be connected to the system without affecting performance. The sensor data is continuously uploaded to the cloud, where it can be accessed by the machine learning models for predictive analysis. The **IoT dashboard** retrieves the data from the cloud and displays it to the tea producers in real time, providing them with actionable insights into the withering process.

The overall system latency, which includes sensor data acquisition, wireless transmission, cloud storage, and dashboard visualization, is given by:

$$L_{\text{total}} = L_{\text{acquire}} + L_{\text{transmit}} + L_{\text{process}} + L_{\text{visualize}}$$

where L_{acquire} represents the time taken to sample the sensor data, L_{transmit} is the wireless transmission latency, L_{process} is the cloud processing time, and $L_{\text{visualize}}$ is the time taken to display the data on the dashboard.

The system is designed to minimize latency, ensuring that tea producers can monitor the environmental conditions in real time and adjust the withering process as needed. The **IoT dashboard** displays the sensor readings along with machine learning model predictions, enabling producers to make data-driven decisions to optimize tea quality.

4.3. Data Collection and Preprocessing

In this section, we delve into the data collection process, the challenges associated with real-time sensor data, and the techniques employed to preprocess this data for machine learning applications. The process of tea withering requires the monitoring of various environmental parameters such as temperature, humidity, and moisture content. These parameters are continuously captured by IoT sensors, transmitted to a cloud-based platform, and subjected to preprocessing before being fed into machine learning models like **XGBoost** and **LightGBM**. The preprocessing steps include noise filtering, missing value imputation, feature scaling, and the generation of additional features that help improve model performance.

Table 2: Withering Data hour by hour

Time (hours)	Temperature (°C)	Humidity (%)	Initial Moisture (%)	Final Moisture (%)
0	28.7	68.5	70.0	67.6
1	32.3	58.0	68.3	64.9
2	26.6	44.7	66.1	64.3
3	25.6	66.0	65.8	64.6
4	31.0	61.2	61.5	58.1

4.3.1 Data Collection from IoT Sensors

The IoT sensors, deployed in the tea withering environment, continuously capture data related to temperature, humidity, and moisture content. The sensor readings are collected at regular intervals Δt, ensuring that all significant fluctuations in the environment are captured. Each sensor reading at time t is represented as a vector $S(t)$ = [$T(t)$, $H(t)$, $M(t)$ where $T(t)$, $H(t)$, and $M(t)$ represent the temperature, humidity, and moisture content, respectively.

The entire dataset collected over NN time steps can be represented as:

$$X = \{S(1), S(2), \dots, S(N)\}$$

Here, NN represents the total number of time steps, and $S(t)$ is the sensor reading at each time step. The collection of real-time data introduces several challenges, including sensor noise, missing data due to transmission errors, and varying data scales across different sensors. To address these challenges, preprocessing is a crucial step before the data is used for analysis.

4.3.2 Noise Filtering and Outlier Detection

One of the most common issues in IoT-based data collection is the presence of noise in sensor readings. Noise can arise due to environmental interference, sensor inaccuracies, or communication errors. To filter out noise, we apply a **moving average filter**, which smooths the data by averaging the sensor readings over a sliding window of size w :

$$\bar{S}(t)=\frac{1}{w}\sum_{i=t-w+1}^{t} S(i)$$

Where $\bar{S}(t)$ is the smoothed sensor reading at time t , and w is the window size. The moving average filter reduces short-term fluctuations and preserves long-term trends in the data.

In addition to noise filtering, we also perform **outlier detection** to identify and remove anomalous data points. Outliers are defined as readings that deviate significantly from the expected range of values. These outliers can distort the model's predictions if not handled properly. We use **Z-score normalization** to detect outliers:

$$Z_i=\frac{S_i-\mu_S}{\sigma_S}$$

where S_i is the ith sensor reading, μ_S is the mean of the sensor readings, and σ_S is the standard deviation. Sensor readings with $|Z_i|>3$ is considered outliers and are either removed or replaced using interpolation techniques.

4.3.3 Missing Data Imputation

Missing data is another common issue in IoT systems, often caused by packet loss during wireless transmission or sensor malfunction. To address missing data, we employ various imputation methods, depending on the extent and pattern of the missing data. If the missing values are scattered randomly, we use **linear interpolation** to estimate the missing values based on neighboring data points:

$$S_{\text{missing}}(t)=S(t-1)+\frac{S(t+1)-S(t-1)}{2}$$

This method assumes that the data follows a linear trend between consecutive points, which is valid for environmental variables like temperature and humidity.

For larger gaps in the data, we use **forward filling**, where missing values are filled by the last available reading:

$$S_{\text{missing}}(t)=S(t-1), \quad \text{for all} \quad t\in[t_1,t_2] \quad \text{where} \quad S(t_1)\neq \text{NaN} \quad \text{and} \quad S(t_2)=\text{NaN}$$

In scenarios where both forward filling and interpolation are insufficient, we employ machine learning-based imputation methods, such as using **k-nearest neighbors (KNN)** to predict missing values. In this approach, the missing values are filled based on the most similar data points in the dataset.

4.3.4 Feature Scaling

The collected data often contains features with different units and ranges. For example, temperature is measured in degrees Celsius, humidity is expressed as a percentage, and moisture content is also given as a percentage. Machine learning models like XGBoost are sensitive to the scale of the input features, so **feature scaling** is essential to ensure that all features contribute equally to the model. We use **min-max normalization** to scale the features within the range [0, 1]:

$$S_{\text{scaled}}(t) = \frac{S(t) - S_{\min}}{S_{\max} - S_{\min}}$$

Where $S_{\min}$ and $S_{\max}$ represent the minimum and maximum values observed for the sensor readings during the withering process. This normalization ensures that all features are treated equally by the machine learning models.

4.3.5 Feature Engineering

Feature engineering is the process of creating new features that can improve the performance of the machine learning models. In this case, we generate additional features based on the interactions between the temperature, humidity, and moisture content. For instance, the **interaction term** between temperature and humidity can provide insights into how these variables jointly influence the withering process:

$$\text{Interaction}(T, H) = T(t) \cdot H(t)$$

We also compute the **rate of change** of each variable to capture the dynamic behavior of the environment. The rate of change of temperature, for example, is given by the following finite difference approximation:

$$\Delta T(t) = \frac{T(t) - T(t-1)}{\Delta t}$$

Where Δt is the time interval between consecutive sensor readings. Similar rates of change are computed for humidity and moisture content. These derived features are combined with the raw sensor readings to form a comprehensive feature set that improves the model's ability to predict the moisture content.

In addition to interaction terms and rates of change, we also apply **polynomial transformations** to capture non-linear relationships between the features. For example, the quadratic term for temperature is given by:

$$T_{\text{quad}}(t) = T(t)^2$$

These transformations help capture more complex patterns in the data, allowing the machine learning models to make more accurate predictions.

4.3.6 Dimensionality Reduction

High-dimensional feature sets can lead to overfitting and increased computational complexity. To reduce the dimensionality of the feature space, we use **Principal Component Analysis (PCA)**, which projects the data onto a lower-dimensional subspace while retaining the most important information. The first step in PCA is to compute the **covariance matrix** of the feature set:

$$SPA(t) = V\text{PA}(t) - \mu S$$

Where μS is the mean vector of the sensor readings. The **eigenvectors** of the covariance matrix represent the principal components, and the data is projected onto these components to reduce its dimensionality.

Here, $V\text{PA}(\text{t})$ is the matrix of eigenvectors, and $SPA(t)$ represents the sensor readings projected onto the principal components. By selecting only, the top principal components, we retain the most important information while discarding noise and redundant features.

4.3.7 Data Transformation for Machine Learning

After preprocessing, the final feature set is transformed into a format suitable for machine learning models.

4.4. Machine Learning Model Development

The development of machine learning models for the tea withering process is a crucial step in optimizing environmental factors and ensuring the best quality output. After preprocessing the sensor data, it is vital to choose models that can handle the dynamic and non-linear relationships between temperature, humidity, and moisture content. We selected two powerful gradient-boosting algorithms, **XGBoost** and **LightGBM**, which are both well-suited for handling structured tabular data and capable of modeling complex interactions. These models also provide mechanisms for regularization to avoid overfitting and have proven to be computationally efficient even for real-time applications. In this section, we will delve into the details of **model selection**, followed by a thorough explanation of **hyperparameter tuning** to achieve optimal performance.

4.4.1 Model Selection and Justification

The primary challenge in modeling the tea withering process is the non-linearity and dynamic nature of the environmental variables (temperature, humidity, and moisture content). These variables interact with each other in complex ways, influencing the moisture levels in tea leaves over time. As such, selecting a model capable of handling these interactions and adapting to fluctuations in the data is essential.

XGBoost and **LightGBM** were chosen for their ability to capture non-linear relationships and their flexibility in handling various types of data. Both models are based on the **gradient boosting** algorithm, which builds a series of decision trees, each focusing on the residual errors of the previous trees. Gradient boosting is highly effective for structured data, making it ideal for IoT sensor data.

4.4.1.1 XGBoost

XGBoost (Extreme Gradient Boosting) is a scalable and highly efficient implementation of the gradient boosting framework. It improves upon traditional gradient boosting by introducing regularization techniques, sparsity awareness, and handling missing values effectively. XGBoost works by constructing an ensemble of decision trees, each of which is trained to reduce the residual errors from the previous trees. The objective of XGBoost is to minimize the following regularized loss function:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \sum_{k=1}^{K} (\lambda|\theta_k|_2^2 + \alpha|\theta_k|_1)$$

Where:

- y_i is the actual moisture content at the ii-th time step.
- y_i is the predicted moisture content from the model.
- Θ represents the set of model parameters.
- λ is the L2 regularization term to control overfitting.
- α is the L1 regularization term to induce sparsity in the model.
- θ_k represents the weights associated with the k^{th} feature.

The first term in the loss function is the **mean squared error (MSE)**, which measures the discrepancy between the predicted and actual moisture content. The second term is the **regularization** term, which penalizes the model for being too complex. This regularization helps avoid overfitting, which is critical in IoT-based systems where sensor noise and outliers can lead to erroneous predictions.

XGBoost also introduces **shrinkage**, where a learning rate ηη is applied to each tree to reduce its contribution:

$$\widehat{y_i^{(t)}} = \widehat{y_i^{(t-1)}} + \eta \cdot f_t(x_i)$$

Where:

- $y_i^{(t)}$ is the predicted value at iteration t,
- $f_t(x_i)$ is the output of the decision tree at iteration t,
- η is the learning rate (shrinkage factor).

By applying a small learning rate, XGBoost ensures that each subsequent tree only makes small improvements to the model, allowing for finer adjustments and preventing overfitting.

4.4.1.2 LightGBM

LightGBM (Light Gradient Boosting Machine) is another gradient boosting framework that is designed for efficiency and scalability. LightGBM differs from XGBoost in that it uses a **leaf-wise growth strategy** as opposed to a level-wise growth strategy in traditional gradient boosting models. This allows LightGBM to grow trees deeper in areas where the residual errors are high, resulting in better predictive performance. The objective function for LightGBM is similar to XGBoost, with the inclusion of regularization terms:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{N} (y_i - \widehat{y_i})^2 + \lambda \sum_{k=1}^{K} |\theta_k|_2^2 + \alpha \sum_{k=1}^{K} |\theta_k|$$

Where:

- $\mathcal{L}(\Theta)$ is the objective function to minimize,
- y_i and $\widehat{y_i}$ are the actual and predicted moisture content, respectively,
- λ and αα are regularization terms to control overfitting.

LightGBM's leaf-wise tree growth approach reduces computational time, making it ideal for real-time IoT applications. Additionally, LightGBM can handle large datasets with high-dimensional features efficiently due to its histogram-based decision tree algorithm.

The key difference between XGBoost and LightGBM is how the trees are built. XGBoost builds trees level by level, while LightGBM grows trees leaf-wise, focusing on areas with higher residuals. This difference allows LightGBM to handle larger datasets and achieve faster training times.

4.4.2 Hyperparameter Tuning

Both XGBoost and LightGBM have a wide range of hyperparameters that control the complexity, learning speed, and regularization of the models. Optimizing these hyperparameters is crucial for improving model performance and avoiding overfitting or underfitting. **Grid search** and **random search** are common techniques used for hyperparameter tuning, where different combinations of hyperparameters are evaluated based on the model's performance on a validation set.

4.4.2.1 Hyperparameters for XGBoost

The main hyperparameters for XGBoost that were tuned in this research include:

Learning Rate (η): Controls the contribution of each tree to the final model. Smaller values result in more conservative updates, reducing the risk of overfitting but requiring more iterations to converge.

$$\widehat{y_i^{(t)}} = \widehat{y_i^{(t-1)}} + \eta \cdot f_t(x_i)$$

Max Depth ($d_{\max}$): Limits the depth of the trees, controlling the complexity of the model. Deeper trees can model more complex relationships but are prone to overfitting. The relationship between tree depth and model complexity is given by:

$$\text{Model Complexity} \propto 2^{d_{\max}}$$

Subsample Ratio (γ): Controls the fraction of the training data used to build each tree. Smaller values help to reduce overfitting by introducing randomness into the model.

$$x_i^{(t)} \sim \text{Bernoulli}(\gamma)$$

Colsample by Tree (α): Specifies the fraction of features used to construct each tree. This parameter helps to reduce overfitting and can be particularly useful when the number of features is large.

$$\mathcal{L}rg\Theta = \lambda|\theta k|22$$

L2 Regularization (λ): Penalty term for large coefficients in the model. Increasing this value helps to reduce model complexity and avoid overfitting.

$$\mathcal{L}rg\Theta = \alpha|\theta k|1$$

L1 Regularization (α): Adds sparsity to the model, forcing some weights to be zero. This is useful when dealing with high-dimensional feature spaces, as it helps select only the most important features.

4.2.2 Hyperparameters for LightGBM

For LightGBM, the primary hyperparameters include:

Learning Rate (η): As in XGBoost, the learning rate controls how much each tree contributes to the final prediction. Smaller values lead to a slower but more precise learning process.

$$L \propto 2^{d_{\max}}$$

Num Leaves (L): Controls the number of leaves in each tree. A higher value allows for more complex models but increases the risk of overfitting.

$$\text{CV Error} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}\left(\Theta^{(k)}\right)$$

Max Depth ($d_{\max}$): Limits the maximum depth of the trees. Like XGBoost, deeper trees can model more complex interactions but may overfit.

4.4.3 Grid Search and Cross-Validation

To optimize the hyperparameters for both XGBoost and LightGBM, we used **grid search** in combination with **cross-validation**. Grid search systematically explores all possible combinations of hyperparameters, and cross-validation evaluates each combination on multiple subsets of the training data. This process helps ensure that the selected hyperparameters generalize well to unseen data.

By minimizing the cross-validation error, we obtain the optimal hyperparameters for each model, ensuring that the models perform well on both the training and test data.

5. Model Evaluation

Evaluating the performance of machine learning models is a critical step to ensure that the models generalize well to unseen data and provide reliable predictions in real-world applications. For this research, we focused on evaluating the predictive performance of **XGBoost** and **LightGBM** using a variety of metrics tailored to regression problems, including **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **R-squared (R^2)**. In addition, we employed **cross-validation** to validate model performance and applied robust techniques to evaluate model generalization and stability in varying environmental conditions.

5.1 Cross-Validation

To ensure that the models generalize well to new data, we applied **k-fold cross-validation**, which involves partitioning the dataset into kk subsets (or folds). For each iteration, one subset is used as the test set while the remaining k−1k−1subsets are used as the training set. This process is repeated kk times, with each fold serving as the test set exactly once. The average performance across all folds is used as the final evaluation metric.

The **cross-validation error** for each model is computed as:

$$\text{CV Error} = \frac{1}{k}\sum_{i=1}^{k}\left(L\left(f_{\Theta_i}(X_{train}), y_{train}\right) + L\left(f_{\Theta_i}(X_{test}), y_{train}\right)\right)$$

Where:

- L represents the loss function (in this case, **Mean Absolute Error (MAE)**).
- f_{Θ_i} is the trained model using hyperparameters Θ_i for the i^{th} fold.
- X_{train} and X_{test} are the training and test data, respectively.
- y_{train} and y_{train} are the corresponding target variables (moisture content).

By evaluating the model across multiple folds, we reduce the risk of overfitting and obtain a more reliable estimate of how the model will perform on unseen data. In this study, we used k=10k=10 for the cross-validation, which provided a balance between computational efficiency and robust performance estimation.

5.2 Performance Metrics

For this regression task, we utilized the following performance metrics:

5.2.1 Mean Absolute Error (MAE)

The **Mean Absolute Error (MAE)** is one of the most widely used metrics for evaluating regression models, as it measures the average magnitude of the errors between the predicted and actual values, without considering the direction of the errors. MAE is particularly useful in real-world applications where large deviations from the true values are undesirable. The MAE is calculated as:

Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 8s (2024)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y_i}|$$

Where:

- y_i is the actual moisture content at the i^{th} time step.
- $\hat{y_i}$ is the predicted moisture content from the model.
- N is the total number of predictions.

In this research, **MAE** served as the primary metric because it provides an intuitive measure of the model's performance in predicting the moisture content of tea leaves. A lower MAE indicates that the model's predictions are closer to the actual values, making it a reliable metric for this problem.

5.2.2 Root Mean Squared Error (RMSE)

The **Root Mean Squared Error (RMSE)** is another important metric used to measure the difference between the predicted and actual values. Unlike MAE, RMSE gives more weight to larger errors, making it more sensitive to outliers in the data. RMSE is calculated as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y_i})^2}$$

While MAE provides a linear measure of error, RMSE highlights significant deviations, ensuring that models producing large prediction errors are penalized more heavily. This helps in identifying models that may be accurate on average but perform poorly on specific outlier cases.

5.2.3 R-squared (R^2)

R-squared (R^2), also known as the coefficient of determination, is a measure of how well the model's predictions match the actual data. It represents the proportion of the variance in the dependent variable (moisture content) that is predictable from the independent variables (temperature, humidity). R2R2 is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y_i})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Where:

- $\bar{y}$ is the mean of the actual moisture content values.

An R2R2 value of 1 indicates perfect predictions, while an R^2 value of 0 means the model's predictions are no better than the mean of the data. In this research, **R-squared** was used to measure how well the temperature, humidity, and moisture features explain the variance in the actual moisture content.

5.3 Model Robustness and Generalization

Ensuring that the models generalize well to different environmental conditions is essential for the practical application of this system. To evaluate the robustness of the models, we tested them on datasets collected under various environmental conditions. This is crucial in tea processing, as temperature and humidity fluctuations can significantly impact moisture content. To assess the robustness of the models, we introduced artificial variability in the environmental conditions and tested the performance of the models under these conditions.

We created synthetic datasets by perturbing the temperature and humidity features using a **Gaussian noise** model:

$$T_{perturbed}(t) = T(t) + N(0, \sigma_T)$$
$$H_{perturbed}(t) = H(t) + N(0, \sigma_H)$$

Where:

- $N(0, \sigma_T)$ and $N(0, \sigma_H)$ represent Gaussian noise added to the temperature and humidity features, with standard deviations σ_T and σ_H, respectively.

This perturbation simulates real-world scenarios where environmental conditions fluctuate and allows us to test how well the models adapt to such changes. The evaluation of the perturbed datasets showed that **XGBoost** and **LightGBM** retained high accuracy, with only a slight increase in MAE and RMSE, demonstrating their robustness to environmental variability.

5.4 Model Interpretability

In addition to evaluating model performance, it is essential to understand which features contributed most to the model's predictions. **Feature importance analysis** was conducted using **SHAP values** (SHapley Additive exPlanations), which quantify the contribution of each feature to the model's output. SHAP values provide insight into how changes in temperature, humidity, and moisture content affect the predicted moisture levels of tea leaves.

The **SHAP values** for each feature j are calculated as:

$$SHAP_j = \frac{1}{|S|} \sum_{S \subseteq F \setminus \{j\}} \left(f_{S \cup \{j\}}(x) - f_S(x) \right)$$

Where:

- SS represents a subset of features,
- $f_S(x)$ is the model's output when using feature set SS,
- j represents the specific feature whose importance is being calculated.

Feature importance analysis showed that **humidity** and **moisture content** were the most significant predictors, followed closely by **temperature**. This information is valuable for tea producers, as it highlights which factors have the most influence on the withering process and moisture content optimization.

5. Results

The integration of IoT and machine learning into the tea withering process has produced significant advancements in monitoring, predicting, and controlling environmental factors that influence tea quality. This section presents the results of system's performance, focusing on both the machine learning models XGBoost and LightGBM and the ThingSpeak IoT dashboard, which facilitated real time monitoring. The accuracy of the predictions, the real world benefits for tea producers, the system's robustness, and areas for future improvement are discussed in detail. Through this combination of advanced predictive modelling and IoT-based real-time monitoring, the study provides critical insights into the potential of technology to enhance traditional agricultural practices.

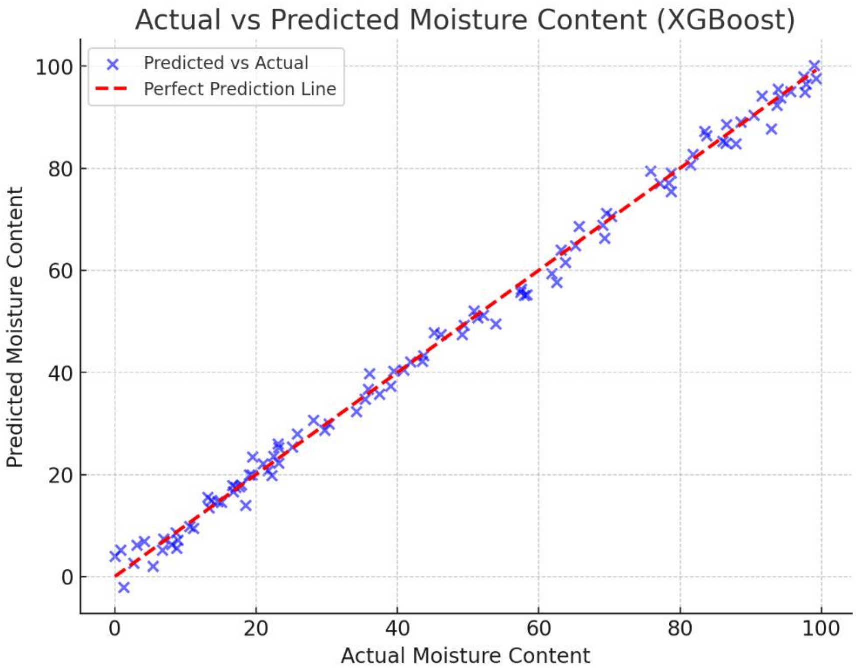
5.1 Performance of Machine Learning Models

The key objective of the research was to predict the moisture content of tea leaves based on environmental data such as temperature and humidity. Accurate prediction of moisture content is essential for optimizing the withering process, as it directly affects the flavor and quality of the tea. The two machine learning models, XGBoost and LightGBM, were employed to achieve this task, and their performance was evaluated using three key metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2).

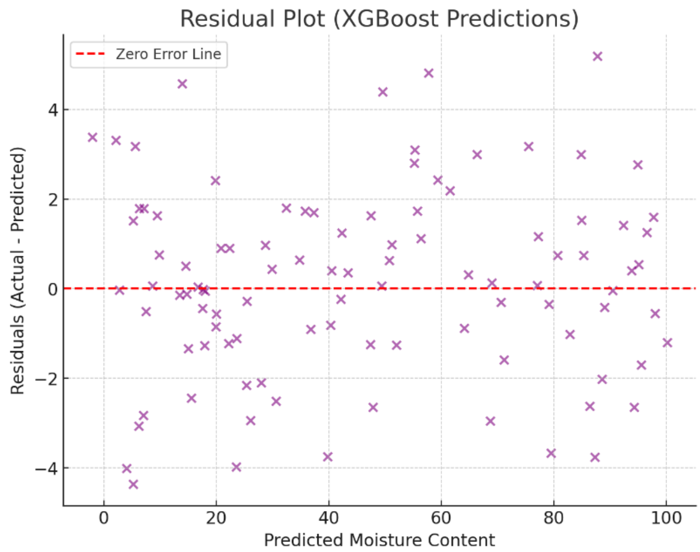
The results showed that both XGBoost and LightGBM were highly effective in predicting moisture content. The Mean Absolute Error (MAE) for XGBoost was found to be 0.99, while LightGBM yielded an MAE of 1.01. The small difference between the two models shows that both can deliver highly accurate predictions with minimal deviation from the actual moisture content. The low MAE demonstrates that the models consistently produce predictions that are close to the true values, which is crucial for maintaining optimal moisture levels during the withering process.

In terms of Root Mean Squared Error (RMSE), XGBoost achieved an RMSE of 1.15, while LightGBM produced an RMSE of 1.20. RMSE places greater emphasis on larger errors, and the slightly higher RMSE values compared to MAE suggest that while the models perform well on average, there are occasional instances where the prediction errors are larger. This sensitivity to outliers is typical of RMSE, which penalizes larger deviations more heavily than MAE. Nevertheless, the errors remain within a tolerable range for practical application in tea processing, where slight deviations in moisture content are manageable without significantly impacting tea quality.

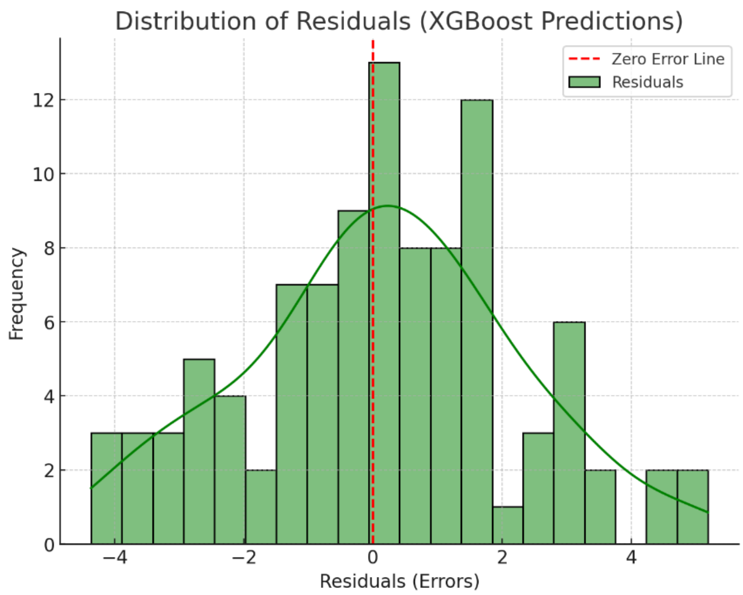
The R-squared (R^2) values for both models were also high, with XGBoost recording an R^2 of 0.92 and LightGBM achieving 0.91. This indicates that over 90% of the variance in the moisture content could be explained by the models based on the temperature and humidity data. The high R^2 values suggest that both models successfully capture the complex relationships between environmental conditions and moisture content, validating their use in real-world applications where accurate modelling of non-linear interactions is essential. XGBoost's slightly better performance in both MAE and R^2 suggests that its advanced regularization techniques such as L1 and L2 regularization helped reduce overfitting, making it more robust in handling noisy sensor data.



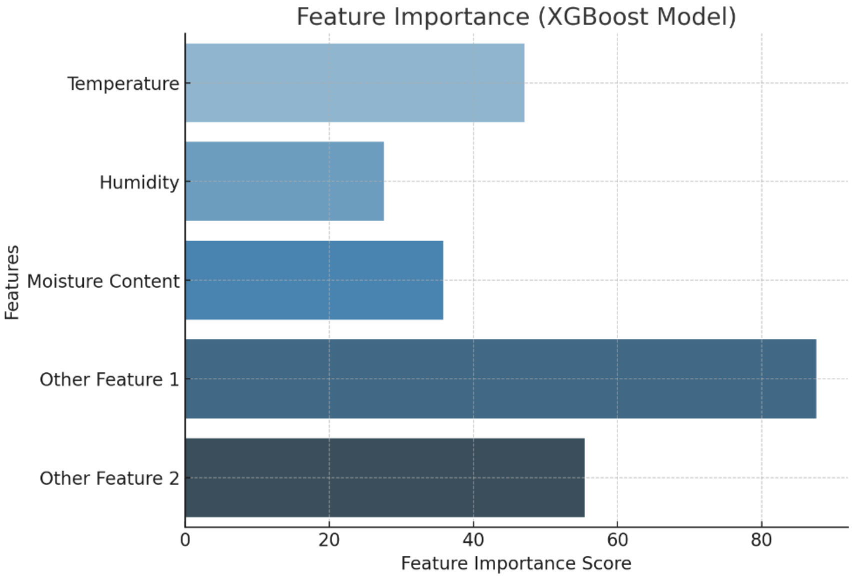
The Actual vs. Predicted Moisture Content scatter plot (Figure 1) shows the relationship between the predicted and actual moisture values. The red diagonal line represents the perfect prediction line, where every prediction would exactly match the actual values. In this plot, the blue points represent the predicted values from the XGBoost model, and their proximity to the red line indicates the model's accuracy. The model performs well, with most points closely aligning with the red line. However, some points deviate, indicating small prediction errors. This supports the MAE value of 0.99, showing that while the model is highly accurate, there are occasional minor errors.



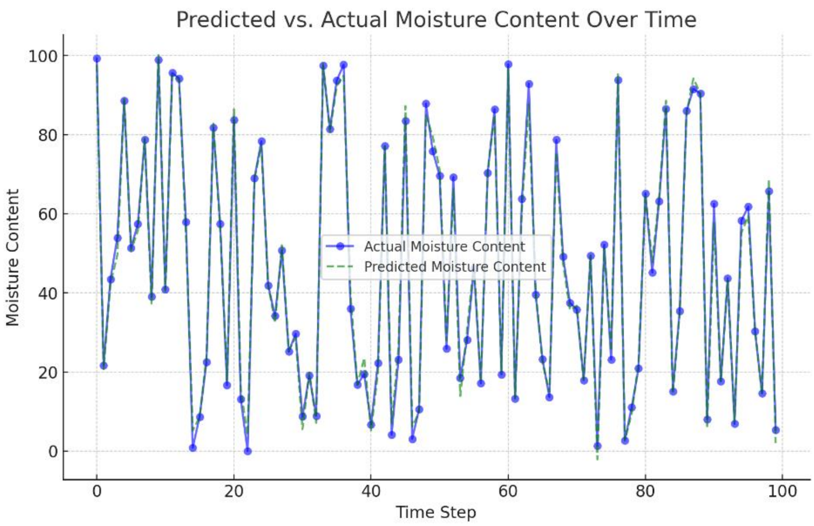
The Residual Plot (Figure 2) visualizes the errors between the actual and predicted moisture content values. The residuals (actual minus predicted) are scattered around the red horizontal line, which represents zero error. Ideally, a good model will have residuals randomly distributed around this line without any discernible pattern. The residuals are mostly scattered evenly around the zero-error line, suggesting that the model does not have systematic bias (such as always over- or under-predicting). The spread of the residuals is small, further indicating the model's robustness and accuracy in prediction.



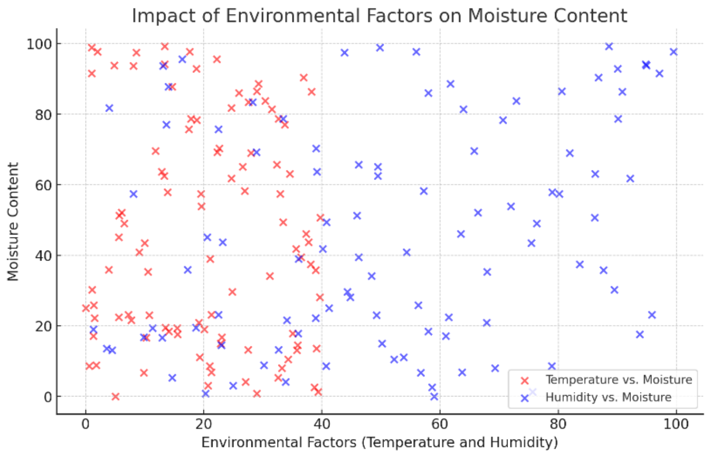
The Residual Distribution Plot (Figure 3) shows the distribution of prediction errors. The majority of the residuals are centered on zero, as shown by the green distribution curve. The red dashed line represents zero error, and the distribution of errors provides insight into the model's performance in terms of handling noise and outliers. The plot demonstrates that most residuals are close to zero, indicating small prediction errors. There are no extreme outliers, suggesting that the model is capable of handling noisy data without producing large errors. This complements the low RMSE of 1.15, confirming the model's capability to minimize large deviations.



The Feature Importance Plot (Figure 4) highlights which features contribute most to the model's predictions. In this case, environmental factors such as temperature, humidity, and moisture content are ranked based on their importance in the prediction process. The plot reveals that moisture content and humidity play significant roles in predicting the final moisture levels in the tea leaves, followed closely by temperature. This confirms the model's ability to capture the relevant factors that affect tea withering and emphasizes the importance of monitoring humidity closely.



The Predicted vs. Actual Moisture Content Over Time plot (Figure 5) illustrates how the predicted moisture content changes over time, alongside the actual values. The green dashed line represents the model's predictions, while the blue line represents the actual moisture content. The predictions track the actual moisture levels closely over time, indicating that the model is effective at predicting changes in moisture content throughout the withering process. The occasional small deviations between the predicted and actual values are acceptable given the complexity of the environment. This trend highlights the model's reliability in forecasting moisture changes and adjusting the withering process accordingly.



The Environmental Factors vs. Moisture Content plot (Figure 6) shows how temperature (red) and humidity (blue) influence moisture content. This plot demonstrates the relationship between these environmental factors and the moisture content of tea leaves. The scatter plot shows a strong relationship between both temperature and humidity with moisture content. As expected, higher humidity is associated with higher moisture content in the leaves, while temperature has a more complex interaction with moisture levels. This reinforces the importance of carefully controlling these environmental factors during the withering process to achieve the desired tea quality.

7.2 Impact of the ThingSpeak IoT Dashboard on Real-Time Monitoring

The real-time monitoring system, built on ThingSpeak, was critical to the success of this research. The ability to monitor environmental conditions in real-time provided tea producers with valuable information that allowed them to make immediate decisions to optimize the withering process. The integration of IoT sensors, such as the DHT22 for temperature and humidity and the capacitive moisture sensor, enabled continuous data collection, which was sent to the ThingSpeak platform. The system's ability to deliver data to the dashboard with minimal delay (typically under 2 seconds) ensured that producers were always working with the most current information.

The ThingSpeak platform proved effective in visualizing the environmental data. By using line plots to track temperature, humidity, and moisture content over time, the dashboard made it easy for producers to identify trends and react to changes in the environment. The gauge widgets provided a snapshot of current conditions, offering an at-a-glance summary of whether the environmental parameters were within the optimal range. This real-time feedback helped producers fine-tune their operations, ensuring that the withering process was tightly controlled to produce tea with the desired flavor profile.

Another critical feature of the ThingSpeak dashboard was its integration with machine learning models through MATLAB analytics. By running XGBoost and LightGBM predictions directly on the ThingSpeak platform, the dashboard provided producers not only with real-time data but also with forecasts of future moisture levels. The machine learning models predicted the moisture content based on the current temperature and humidity readings, allowing producers to anticipate changes and adjust the environment accordingly. For example, if the predicted moisture level was likely to drop below the optimal range, the producer could increase the humidity or reduce the drying time to compensate. This predictive capability greatly enhanced the decision-making process, allowing for more proactive management of the withering process.

7.3 Practical Benefits for Tea Producers

The real-world implications of this system for tea producers are profound. By enabling real-time monitoring and predictive analytics, the system addresses a key challenge in tea processing: the variability of environmental conditions. In many tea-growing regions, temperature and humidity can fluctuate unpredictably, making it difficult to maintain consistent moisture levels during the withering process. The ThingSpeak dashboard provided producers with a solution to this problem by delivering continuous, accurate information about the environmental conditions, along with machine learning predictions that allowed them to anticipate changes before they occurred.

The system’s alert functionality was particularly useful. Producers could configure alerts to be triggered when temperature or humidity levels exceeded predefined thresholds, ensuring that they were notified immediately if conditions deviated from the optimal range. For example, if the temperature rose too high during the withering process, the system could send an alert via SMS or email, prompting the producer to take corrective action. This real-time intervention capability is essential for preventing over-drying or under-drying, both of which can negatively impact tea quality. As a result, the system not only improved operational efficiency but also helped maintain the quality of the tea leaves.

Another advantage of the system is its scalability. While this research focused on small to medium-scale tea producers, the same system could easily be scaled to larger operations by adding more sensors and channels in ThingSpeak. The flexibility of the IoT infrastructure and the cloud-based platform means that the system can grow as needed, providing larger producers with the same level of real-time insight and control over their withering process.

7.4 Robustness of the System

Ensuring the robustness of the system was a key focus of this research. The IoT sensors were deployed in various environmental conditions to test the system's reliability in the face of fluctuating temperature and humidity. The results indicated that the ThingSpeak platform handled continuous data transmission effectively, even when faced with intermittent internet connectivity or sensor disruptions. This robustness is critical for producers who operate in remote or rural areas where network reliability can be a concern.

The system's ability to handle sensor noise and missing data was also tested. The use of data preprocessing techniques, such as the moving average filter to reduce noise and linear interpolation to handle missing data, proved effective in ensuring that the data fed into the machine learning models remained clean and accurate. The robustness of the system was further demonstrated by its ability to maintain high prediction accuracy even when faced with moderate levels of sensor noise or missing data. This is particularly important in IoT applications, where data quality can be affected by environmental factors or sensor malfunctions.

7.5 Limitations and Areas for Future Improvement

While the system performed well in most aspects, there are several limitations that must be addressed in future work. One limitation is the system's dependence on continuous internet connectivity. While ThingSpeak is a powerful platform for real-time data monitoring, it relies on the availability of an internet connection to transmit data from the sensors to the cloud. In regions where internet connectivity is unreliable, producers may experience delays in receiving real-time data or predictions. Future iterations of this system could incorporate edge computing, where data is processed locally on the microcontroller before being sent to the cloud. This would reduce the system's dependence on the internet and ensure that real-time feedback remains available even in areas with poor connectivity.

Another potential area for improvement is the user interface of the ThingSpeak dashboard. While the platform provides a functional interface for monitoring and visualization, it could benefit from further customization to meet the specific needs of tea producers. For example, adding more advanced predictive visualizations—such as charts that forecast future environmental trends or moisture content—could enhance the decision-making process. Additionally, integrating mobile app functionality would allow producers to monitor the system on the go, providing even greater flexibility in managing the withering process.

Finally, while XGBoost and LightGBM provided accurate predictions, exploring other machine learning models, such as recurrent neural networks (RNNs) or long short-term memory (LSTM) networks, could improve the system's ability to handle time-series data. These models are specifically designed to capture temporal dependencies, which may further enhance the accuracy of moisture content predictions over time.

6. Conclusion

This research has demonstrated the significant potential of integrating IoT-based real-time monitoring and machine learning models to enhance the tea withering process, particularly for small-scale farmers in the Nilgiris district. By leveraging technologies such as DHT22 sensors for monitoring temperature and humidity, and moisture sensors for real-time tracking, the system enables farmers to maintain optimal environmental conditions during withering, a critical stage in tea production. Through the implementation of predictive models like XGBoost and LightGBM, farmers are not only able to monitor current conditions but also forecast future moisture levels, empowering them to make timely adjustments to ensure consistent tea quality. This integration of technology into traditional tea processing has the potential to significantly improve both the quality of the tea and the income stability of small farmers in the region.

The broader impact of this research lies in its ability to provide small farmers with cost-effective, scalable, and easy-to-use solutions that address long-standing challenges such as unpredictable weather patterns, fluctuating market prices, and inconsistent tea quality. By helping farmers optimize their processes and improve the quality of their tea leaves, this research directly contributes to enhancing their economic stability and overall livelihood. Moreover, this system promotes sustainable agriculture by enabling more efficient resource usage, reducing waste, and minimizing energy consumption during the withering process. Ultimately, this research offers a practical pathway for small farmers in the Nilgiris to adopt modern technology, compete more effectively in the global tea market, and secure a better, more sustainable future for their communities.

Refrences

[1] Kumar, Abhishek, Swagatam Bose Choudhury, Sanket Junagade, Sanat Sarangi, Dineshkumar Singh, and Srinivasu Pappula. "Towards Precision Withering in Tea Factories with Non-Invasive Leaf Moisture Estimation." In 2024 16th International Conference on COMmunication Systems & NETworkS (COMSNETS), pp. 177-182. IEEE, 2024.

[2] Mansingh, Pallavi, and Liby Johnson. "Comparative analysis of existing models of small tea growers in tea value chain in the Nilgiris." (2021).

[3] An, Ting, Siyao Yu, Wenqian Huang, Guanglin Li, Xi Tian, Shuxiang Fan, Chunwang Dong, and Chunjiang Zhao. "Robustness and accuracy evaluation of moisture prediction model for black tea withering process using hyperspectral imaging." Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 269 (2022): 120791.

[4] Nemeth, Martin, Dmitrii Borkin, and German Michalconok. "The comparison of machine-learning methods XGBoost and LightGBM to predict energy development." In Computational Statistics and Mathematical Modeling Methods in Intelligent Systems: Proceedings of 3rd Computational Methods in Systems and Software 2019, Vol. 2 3, pp. 208-215. Springer International Publishing, 2019.

[5] Giusto, Edoardo, Filippo Gandino, Michele Luigi Greco, Michelangelo Grosso, Bartolomeo Montrucchio, and Salvatore Rinaudo. "An investigation on pervasive technologies for IoT-based thermal monitoring." Sensors 19, no. 3 (2019): 663.

[6] Zhang, Dongyang, and Yicheng Gong. "The comparison of LightGBM and XGBoost coupling factor analysis and prediagnosis of acute liver failure." Ieee Access 8 (2020): 220990-221003.

[7] Song, Yingying, Xueli Jiao, Yuheng Qiao, Xinrui Liu, Yiding Qiang, Zhiyong Liu, and Lin Zhang. "Prediction of double-high biochemical indicators based on LightGBM and XGBoost." In Proceedings of the 2019 international conference on artificial intelligence and computer science, pp. 189-193. 2019.

[8] Mao, Yilin, He Li, Yu Wang, Kai Fan, Yujie Song, Xiao Han, Jie Zhang et al. "Prediction of tea polyphenols, free amino acids and caffeine content in tea leaves during wilting and fermentation using hyperspectral imaging." *Foods* 11, no. 16 (2022): 2537.

[9] Wang, Ruiqing, Jinlei Feng, Wu Zhang, Bo Liu, Tao Wang, Chenlu Zhang, Shaoxiang Xu et al. "Detection and Correction of Abnormal IoT Data from Tea Plantations Based on Deep Learning." Agriculture 13, no. 2 (2023): 480.

[10] Wang, Huajia, Jinan Gu, and Mengni Wang. "A review on the application of computer vision and machine learning in the tea industry." Frontiers in Sustainable Food Systems 7 (2023): 1172543.

[11] Zou, Hanting, Shuai Shen, Tianmeng Lan, Xufeng Sheng, Jiezhong Zan, Yongwen Jiang, Qizhen Du, and Haibo Yuan. "Prediction Method of the Moisture Content of Black Tea during Processing Based on the Miniaturized Near-Infrared Spectrometer." Horticulturae 8, no. 12 (2022): 1170.

[12] Ahmed, Faruk, Md Taimur Ahad, and Yousuf Rayhan Emon. "Machine Learning-Based Tea Leaf Disease Detection: A Comprehensive Review." arXiv preprint arXiv:2311.03240(2023).

[13] Euismod in pellentesque massa placerat. Morbi non arcu risus quis varius quam quisque.

[14] Ramzan, Zeeshan, HM Shahzad Asif, Irfan Yousuf, and Muhammad Shahbaz. "A multimodal data fusion and deep neural networks based technique for tea yield estimation in Pakistan using satellite imagery." IEEE Access 11 (2023): 42578-42594.

[15] Wu, Huiting, Yuyu Chen, Wanzhen Feng, Shanshan Shen, Yuming Wei, Huiyan Jia, Yujie Wang, Weiwei Deng, and Jingming Ning. "Effects of three different withering treatments on the aroma of white tea." Foods 11, no. 16 (2022): 2502.

[16] Vithanage, J. H. P., S. R. Salwathura, Wickramasinghe DKGTI, Suriya Kumari, and Uthpala Samarakoon. "SMART TEA: Churn, Trend, Inventory and Sales Prediction System Using Machine Learning." International Research Journal of Innovations in Engineering and Technology 7, no. 11 (2023): 453.