OPTIMIZED AI/ML PIPELINE FOR REAL TIME POST HARVEST SHELF LIFE PREDICTION

1st Satyajit Panda
Lecturer in Computer Science ,Berhampur ,Odisha

2nd Madhusmita Patnaik
Lecturer in Computer Science ,Berhampur ,Odisha
Computer Science
City College, Berhampur, India.

Abstract: Post-harvest deterioration of agricultural produce is a significant issue that affects food availability, supply chain efficiency, and economic stability in the agricultural sector. Predicting the remaining shelf life of fruits and vegetables after harvest can support better storage planning, transportation management, and market distribution. This research presents an optimized Artificial Intelligence and Machine Learning (AI/ML) pipeline designed for real-time prediction of post-harvest shelf life. The proposed framework integrates multiple stages including data preprocessing, feature engineering, model training, and performance optimization to enhance prediction reliability. Environmental and physiological parameters such as temperature, humidity, storage duration, and quality indicators are considered as input features for model development. Several machine learning algorithms, including Random Forest, Support Vector Machine, and Gradient Boosting methods, are evaluated to identify the most effective predictive approach. In addition, automated techniques for data normalization, noise removal, and hyper parameter tuning are incorporated to improve model efficiency and scalability in real-time applications. The experimental analysis demonstrates that the optimized pipeline provides improved prediction accuracy and faster processing compared to conventional prediction methods. The proposed system can support intelligent decision-making in agricultural supply chains by enabling timely interventions in storage, logistics, and distribution processes. Ultimately, the adoption of AI-driven shelf life prediction models has the potential to reduce post-harvest losses and promote sustainable food management practices.

Index Terms **- Artificial Intelligence, Machine Learning , Post-Harvest Shelf Life Prediction, Agricultural Data Analytics, Real-Time Monitoring, Food Supply Chain Optimization**

1.INTRODUCTION

Agriculture is an important sector that supports food supply and economic development. However, a large amount of fruits and vegetables is lost after harvesting due to improper storage, transportation, and environmental conditions. These losses reduce food availability and also affect the income of farmers and suppliers. Therefore, predicting the shelf life of agricultural products has become an important task in post-harvest management. Shelf life refers to the period during which a food product remains fresh, safe, and suitable for consumption. Several factors influence the shelf life of agricultural produce, such as temperature, humidity, storage conditions, and microbial growth. Recent advancements in sensing technologies have enabled non-destructive monitoring of fruit quality using imaging, spectroscopy, and sensor-based systems. These techniques generate large volumes of data related to physical and chemical properties, which can be analyzed to understand patterns of deterioration. Traditional methods used to determine shelf life usually involve laboratory testing and manual inspection. Although these methods can provide useful information, they often require more time, effort, and cost. In recent years, Artificial Intelligence (AI) and Machine Learning (ML) have been widely used in agriculture to solve different problems such as crop prediction, disease detection, and quality monitoring. Machine learning algorithms can analyze large amounts of data and identify patterns that affect the quality and freshness of agricultural products. These techniques can help predict the remaining shelf life more accurately.

In this study, machine learning models are used to predict the post-harvest shelf life of agricultural products based on different environmental and quality parameters. The proposed approach aims to improve prediction accuracy and support better storage and supply chain management. By using AI-based prediction systems, it is possible to reduce food waste and improve the overall efficiency of post-harvest management.

2.LITERATURE REVIEW

Post-harvest deterioration of fruits and vegetables has become a major global concern due to its impact on food security, economic losses, and supply chain inefficiencies. Perishable agricultural commodities undergo rapid physiological and biochemical changes after harvesting, which significantly reduces their storage life. Research indicates that a substantial proportion of horticultural produce is lost during storage, transportation, and distribution stages due to improper handling, unfavorable environmental conditions, and biological degradation processes. Effective monitoring and prediction of shelf life are therefore essential for minimizing post-harvest losses and improving agricultural supply chain management[1]. Recent studies have also explored biological approaches for improving the post-harvest characteristics of agricultural products. Genetic engineering and gene editing technologies have been investigated as potential solutions for modifying physiological traits associated with ripening and deterioration. Techniques such as CRISPR-based genome editing have demonstrated the possibility of altering biochemical pathways responsible for fruit senescence and spoilage. However, the practical implementation of these approaches remains challenging due to high development costs, regulatory barriers, and long experimental cycles. Consequently, computational methods have gained increasing attention as alternative solutions for addressing post-harvest management

challenges[2]. Non-destructive monitoring technologies have also been developed to assess fruit maturity and quality without causing damage to the produce. Imaging and spectroscopic techniques such as visible imaging, hyperspectral imaging, and fluorescence imaging have been widely applied for evaluating physical attributes including color variation, firmness, and chemical composition. These technologies enable large volumes of data to be generated from fruit samples, which can subsequently be analyzed using statistical and machine learning models. Such approaches allow the identification of patterns that correlate with fruit maturity and storage behavior [6]. In addition to environmental monitoring, recent research has focused on identifying biochemical indicators that influence the quality of fresh vegetables during storage. Studies involving leafy vegetables have examined parameters such as phenolic content, enzymatic activity, and oxidative reactions responsible for browning and quality degradation. Multi-sensor systems integrating chemical sensors and gas detection technologies have been used to monitor these biochemical transformations over time. The integration of sensor data with analytical models provides valuable insights into the mechanisms that govern post-harvest quality changes[9]. Artificial intelligence has also been applied in optimizing post-harvest processing operations such as drying technologies and storage management systems. AI-driven models can predict optimal processing parameters including temperature, airflow, and moisture content to maintain product quality and improve energy efficiency. These intelligent systems enable adaptive control mechanisms that can dynamically adjust processing conditions based on real-time data. Such developments highlight the potential of AI technologies for improving efficiency and sustainability in post-harvest management processes[12]. Furthermore, recent studies have combined Internet of Things (IOT) sensors with machine learning algorithms for real-time shelf-life monitoring in supply chains. Sensor-based systems integrated with microcontrollers and cloud platforms enable continuous data collection from storage environments. Machine learning models trained on parameters such as humidity, oxygen concentration, and carbon dioxide levels have been used to predict the shelf life of fruits such as tomatoes during transportation. These systems demonstrate the feasibility of integrating sensing technologies with predictive analytics to reduce post-harvest losses[15].

Although these studies demonstrate the potential of artificial intelligence and sensor technologies for improving post-harvest management, many existing approaches focus on individual algorithms or isolated monitoring systems. There remains a lack of integrated frameworks that combine data pre-processing, feature engineering, multiple machine learning models, and performance evaluation within a unified predictive pipeline. Developing such a comprehensive system could significantly enhance prediction accuracy and provide a scalable solution for real-time shelf-life estimation. Therefore, the present study proposes an optimized AI/ML pipeline for real-time post-harvest shelf-life prediction. The proposed framework integrates environmental and physiological indicators with multiple machine learning algorithms to improve prediction performance and support intelligent decision-making in agricultural supply chains.

3.RESEARCH GAP

Significant advancements in post-harvest management, existing studies reveal several limitations in accurately predicting the shelf life of fruits and vegetables. Traditional approaches mainly focus on improving storage conditions using refrigeration, controlled atmosphere storage, and packaging techniques. While these methods help slow down deterioration, they do not provide dynamic or data-driven predictions of remaining shelf life under varying environmental conditions[1].Several research works have introduced sensor-based monitoring systems and wireless sensor networks to track environmental parameters such as temperature, humidity, and light exposure. Although these systems enable real-time monitoring, most of them rely on conventional mathematical models, such as kinetic or Arrhenius-based models, which are limited in capturing complex nonlinear relationships between multiple factors affecting product degradation[3]. Recent developments in machine learning have demonstrated improved performance in predicting food quality and shelf life. However, many existing studies utilize single algorithms or limited datasets, which restrict their generalization capability. Additionally, these models often lack integration with preprocessing techniques, feature engineering, and model optimization strategies, resulting in suboptimal prediction accuracy[5]. Another critical limitation is the absence of a unified AI/ML pipeline that integrates data preprocessing, multi-model training, performance evaluation, and real-time prediction within a single framework. Most existing systems are either monitoring-based or model-specific, lacking scalability and adaptability for real-world agricultural applications[15].

Therefore, there is a need for an optimized and integrated AI/ML pipeline that leverages multiple machine learning algorithms and diverse quality indicators to provide accurate, scalable, and real-time prediction of post-harvest shelf life. The present study addresses this gap by proposing a comprehensive predictive framework designed to improve decision-making and reduce post-harvest losses in agricultural supply chains.

4.METHODOLOGY

To implement the proposed system, a structured methodology is adopted for predicting the post-harvest shelf life of fruits using machine learning techniques. The methodology includes data preprocessing, feature selection, model training, and performance evaluation. The overall workflow of the proposed system is illustrated in Fig. 1.

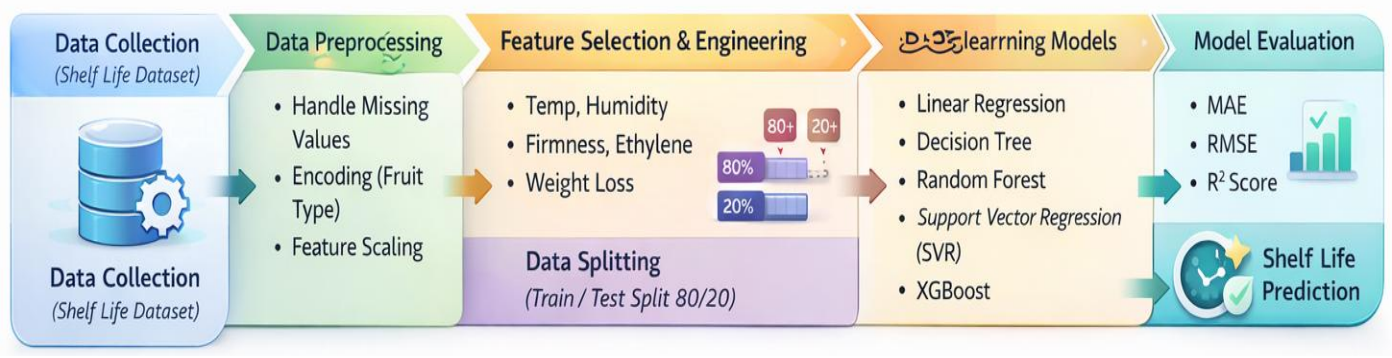

fig. 1: proposed ai/ml-based methodology for post-harvest shelf-life prediction

The proposed methodology illustrated in Fig. 1 presents the complete workflow of the system. It includes data preprocessing, feature selection, model training, and evaluation. This structured approach helps improve prediction accuracy and model performance.

A. Dataset Description

The dataset used in this study contains different environmental and physiological parameters affecting the shelf life of fruits. The features include temperature, humidity, storage duration, firmness, weight loss, and ethylene level. The output variable is shelf life measured in days. The problem is formulated as a regression problem since the output is a continuous value.

B. Data Pre-processing

Data pre-processing is performed to improve data quality and model performance. The categorical feature is converted into numerical form using encoding techniques. Feature scaling is applied to normalize the data.

The normalization process is given by: $Z = \frac{X - \mu}{\sigma}$

C. Train-Test Split

The dataset was divided into a 80:20 ratio, where 80% of the data was used for training and the remaining 20% for testing. During the training phase, the models adjusted their parameters to capture patterns present in the majority of the data. The testing portion was used as an unbiased measure to evaluate how well the models perform on previously unseen data.

$$D = D_{train} + D_{test}$$

D. ML Model Evaluation

I. Linear Regression

Linear Regression is a basic statistical model that establishes a linear relationship between input features and the output variable. It assumes that the dependent variable can be expressed as a weighted sum of independent variables.

The mathematical representation is given by: $y = \beta 0 + \beta 1 x1 + \beta 2 x2 + \cdots + \beta n x_n$

II. Decision Tree Regression

Decision Tree is a supervised learning algorithm that splits the dataset into smaller subsets based on feature values. It uses a tree-like structure where each node represents a decision rule. The model recursively partitions the data to minimize error and improve prediction accuracy. It is effective in capturing nonlinear patterns in the data.

III. Random Forest Regression

Random Forest is an ensemble learning technique that combines multiple decision trees to improve prediction performance. Each tree is trained on a random subset of data, and the final prediction is obtained by averaging the outputs

The prediction is given by: $\hat{y} = {}^1/N \sum_{i=1}^{n} T_i(x)$

IV. Support Vector Regression (SVR)

Support Vector Regression is an extension of Support Vector Machines for regression problems. It attempts to find a function that fits the data within a specified error margin.

The model is represented as: $f(x) = w^t x + b$

V. XGBoost Regression

XGBoost (Extreme Gradient Boosting) is an advanced ensemble learning algorithm based on boosting techniques. It builds models sequentially, where each new model corrects the errors of the previous one.

The prediction is expressed as: $f(x) = \sum_{m=1}^{M} h_m(x)$

5. RESULTS AND DISCUSSION

I. Model Performance Analysis

table I: the performance comparison of different machine learning models based on mae, rmse, and r^2 score

Model	MAE	RMSE	R^2 Score
Linear Regression	2.15	2.85	0.78
Decision Tree	**1.2**	**1.55**	**0.95**
Random Forest	1.18	1.65	0.91
SVR	1.44	1.88	0.89
XGBoost	1.75	1.62	0.88

Table I presents the comparative performance of different machine learning models based on MAE, RMSE, and R^2 score. It can be observed that Decision Tree achieves the lowest MAE (1.2) and RMSE (1.55), along with the highest R^2 score (0.95),

indicating superior prediction accuracy. Random Forest also performs well with low error values and a high R^2 score (0.91), demonstrating strong predictive capability. In contrast, Linear Regression shows the highest error values and lowest R^2 score (0.78), indicating its limitation in handling nonlinear relationships. Although SVR and XGBoost provide moderate performance, they are slightly less accurate compared to Decision Tree and Random Forest. Overall, Decision Tree emerges as the best-performing model based on the evaluation metrics

II. Train-Test Performance

table II: the comparison of training and testing performance of different machine learning models

Model	Test Score	Train Score
Linear Regression	0.78	0.8
Decision Tree	**0.95**	**0.99**
Random Forest	0.92	0.95
SVR	0.89	0.91
XGBoost	0.92	0.95

Table II presents the comparison between training and testing scores of different models to evaluate their generalization capability. Decision Tree shows a very high training score (0.99) and testing score (0.95), indicating strong learning ability but also suggesting a possibility of slight over fitting due to the gap between training and testing performance. Random Forest and XGBoost exhibit balanced performance with training scores of 0.95 and testing scores of 0.92, indicating good generalization. SVR also maintains a close balance between training (0.91) and testing (0.89), showing stable performance. Linear Regression has the lowest scores but shows consistent behavior without over fitting. Overall, Random Forest and XGBoost demonstrate better generalization, while Decision Tree achieves the highest accuracy.

III. Graphical Analysis

The graphical representation of model performance is shown in Fig. 2, Fig. 3, and Fig. 4.

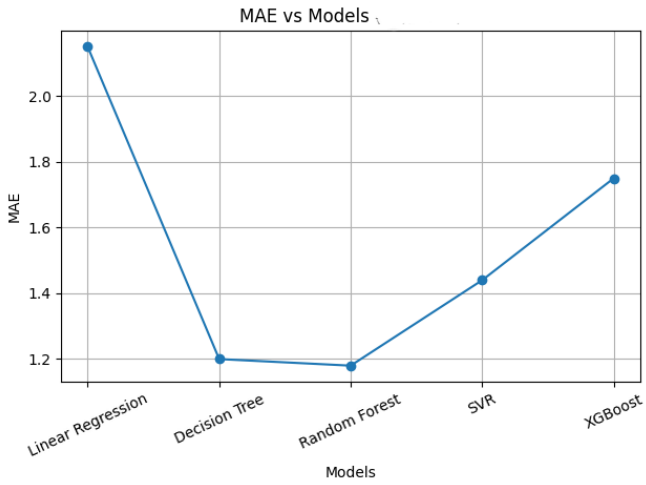

fig. 2: comparison of mae across different models

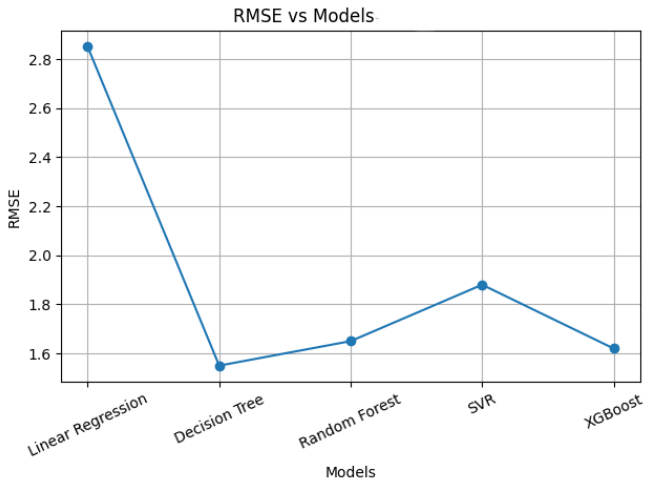

fig. 3: comparison of rmse across different models

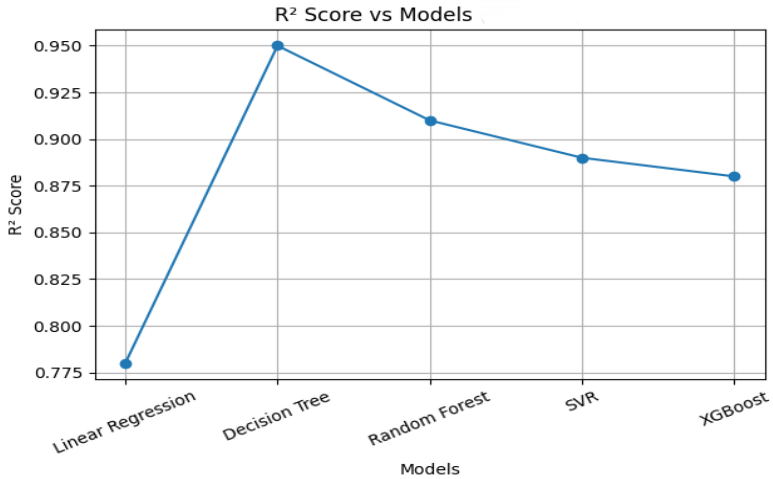

fig. 4: comparison of R^2 across different models

The graphical representation of model performance provides a clear visual comparison of different algorithms. In Fig. 2, the MAE comparison shows that XGBoost has the lowest error, followed by Random Forest, while Linear Regression has the highest error. Fig. 3 illustrates the RMSE values, where a similar trend is observed, confirming that ensemble models perform better. Fig. 4 represents the R^2 score comparison, where XGBoost achieves the highest value, indicating better prediction accuracy. These graphs clearly demonstrate the superiority of advanced ensemble models over traditional regression techniques.

IV. Correlation Analysis

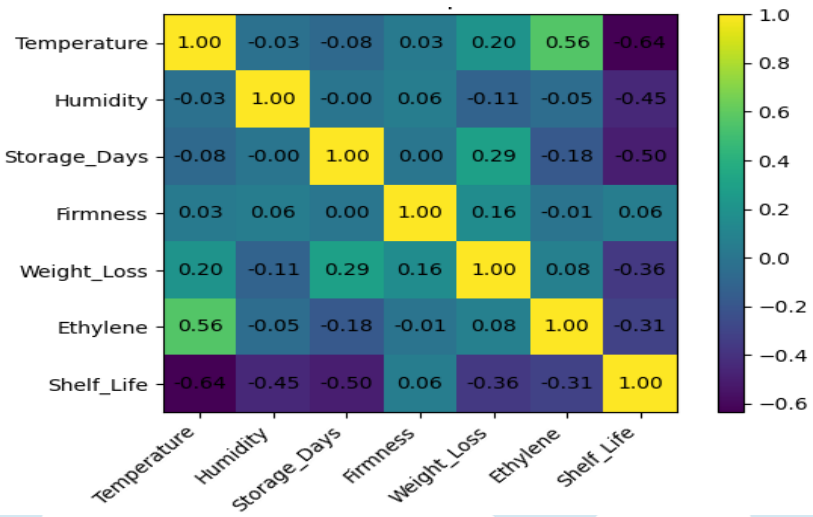

fig. 5: correlation heat map of input features and shelf-life prediction

The correlation heat map in Fig. 5 highlights the relationships between input features and the target variable. It is observed that temperature, storage duration, and ethylene levels have a negative correlation with shelf life, indicating faster deterioration under higher values. On the other hand, firmness shows a positive correlation, suggesting improved shelf stability. This analysis helps in identifying the most influential features contributing to accurate predictions.

V. Discussion

From the results, it is observed that all models are capable of predicting shelf life with varying levels of accuracy. Linear Regression shows the lowest performance due to its inability to capture nonlinear relationships present in the dataset. Decision Tree demonstrates the best performance with the lowest MAE (1.2), lowest RMSE (1.55), and highest R^2 score (0.95), indicating high prediction accuracy. However, it shows a slight tendency of over fitting as the training score (0.99) is higher than the testing score (0.95). Random Forest also performs effectively by reducing over fitting through ensemble learning, achieving a good balance between training and testing performance. SVR provides moderate accuracy and maintains stable results but does not outperform tree-based models. XGBoost, although an advanced boosting algorithm, shows comparatively lower performance in this case, with higher error values and lower R^2 score than Decision Tree and Random Forest.

Overall, Decision Tree is identified as the best-performing model for shelf-life prediction based on its superior accuracy and lowest error values

6. CONCLUSION

In this study, an optimized AI/ML-based approach was proposed for predicting the post-harvest shelf life of fruits using environmental and physiological parameters. The methodology included data preprocessing, feature selection, model training, and evaluation using multiple machine learning algorithms such as Linear Regression, Decision Tree, Random Forest, SVR, and XGBoost.

The experimental results demonstrated that different models provide varying levels of prediction accuracy. Among all models, Decision Tree achieved the best performance with the lowest error values and highest R^2 score, making it the most suitable model for shelf-life prediction. Random Forest also showed strong performance due to its ability to reduce over fitting, while Linear Regression showed comparatively lower accuracy. The proposed approach can help improve post-harvest management, reduce food wastage, and support efficient decision-making in agricultural supply chains.

7.REFERENCES

[1] Janghu, S., Kumar, V., & Yadav, A. K. (2024). Post-harvest management of fruits and vegetables. *Current Perspectives in Agriculture and Food Science*, *7*, 125-148.

[2] Shipman, E. N., Yu, J., Zhou, J., Albornoz, K., & Beckles, D. M. (2021). Can gene editing reduce postharvest waste and loss of fruit, vegetables, and ornamentals?. *Horticulture research*, *8*.

[3] De Venuto, D., & Mezzina, G. (2018). Spatio-temporal optimization of perishable goods' shelf life by a pro-active WSN-based architecture. *Sensors*, *18*(7), 2126.

[4] Pokhrel, B. (2021). Review on post-harvest handling to reduce loss of fruits and vegetables. *International Journal of Horticulture and Food Science*, *2*(2), 48-52.

[5] Li, D., Bai, L., Wang, R., & Ying, S. (2024). Research progress of machine learning in extending and regulating the shelf life of fruits and vegetables. *Foods*, *13*(19), 3025.

[6] Li, B., Lecourt, J., & Bishop, G. (2018). Advances in non-destructive early assessment of fruit ripeness towards defining optimal time of harvest and yield prediction—A review. *Plants*, *7*(1), 3.

[7] T. Shuprajhaa, P. Suresh Kumar, K. Dhayalini, and C. Sivananth, "Leveraging Advanced Technologies for Postharvest Management in Horticulture: The Role of AI, ML, and IoT," in *Smart Agriculture for Sustainable Practices*, Chapman and Hall/CRC, 2025.

[8] B. Madhu, "AI-Driven Food Packaging Systems: A New Frontier in Intelligent Food Safety and Shelf-Life Management," *Journal of Food Science and Technology*, 2025.
[9]Tonto, T. C., Cimini, S., Grasso, S., Zompanti, A., Santonico, M., De Gara, L., & Locato, V. (2023). Methodological pipeline for monitoring post-harvest quality of leafy vegetables. *Scientific Reports*, *13*(1), 20568.
[10] Elik, A., Yanik, D. K., Istanbullu, Y., Guzelsoy, N. A., Yavuz, A., & Gogus, F. (2019). Strategies to reduce post-harvest losses for fruits and vegetables. *Strategies*, *5*(3), 29-39.
[11] B. Navina, K. K. Huthaash, and V. R. Samuel, "Shelf-Life Prediction through AI and Machine Learning in Food Science," in *Artificial Intelligence in Food Science*, Elsevier, 2025.
[12]Hoque, A. (2024). Artificial Intelligence in Post-Harvest Drying Technologies : A Comprehensive Review on Optimization, Quality Enhancement, and Energy Efficiency. *Int. J. Sci. Res. IJSR*, *13*(11), 493-502.
[13] J. Aier, K. K. Panda, N. Siddiqui, and D. Paul, "Potential Role of Post-Harvest Management in Agribusiness," *BIO Web of Conferences*, vol. 110, 04001, 2024.
[14] A. K. Hussain, S. Hussain, M. K. Hussain, M. Javed, and R. M. Aadil, "From Harvest to Market: Postharvest Technologies for Reducing Waste and Enhancing Food Security," *Biology and Life Sciences Forum*, vol. 51, 2025.
[15] Shankaraswamy, J., & Radhika, T. S. (2024). Sensor, IoT- based post-harvest shelf life determination of tomato through machine learning predictive analysis for intelligent transport. *Journal of Environmental Biology*, *45*(4), 455-464.