The Academic
International Journal of Multidisciplinary Research

An Online Peer Reviewed / Refereed Journal
Volume 3 | Issue 6 | June 2025
ISSN: 2583-973X (Online)
Website: www.theacademic.in

Cinnamon Quality Classification using XGBoost

S. Harini
Assistant Professor, Information Technology
Dhirajlal Gandhi College of Technology, Salem, India

R. Surendran
Assistant Professor, Artificial Intellegence & Data Science
Dhirajlal Gandhi College of Technology, Salem, India

DOI : https://doi.org/10.5281/zenodo.15858734

ARTICLE DETAILS

Research Paper

Accepted: 27-06-2025

Published: 10-07-2025

Keywords:

Cinnamon, quality classification, machine learning, volatile oil, coumarin, Random Forest, XGBoost..

ABSTRACT

Cinnamon is a globally consumed spice known for its flavor, aroma, and health benefits. However, variations in its chemical composition, due to differences in cultivation, processing, and species (e.g., *Cinnamomum verum* vs *Cinnamomum cassia*), result in significant quality disparities. Traditional methods of assessing cinnamon quality rely on laboratory-based chemical analysis, which are time-consuming and resource-intensive. In this study, we propose a machine learning-based classification model to automatically evaluate cinnamon quality using measurable physicochemical parameters, including moisture content, ash percentage, acid-insoluble ash, volatile oil concentration, chromium levels, and coumarin content. Multiple classifiers such as Decision Tree, Random Forest, and XGBoost were trained and evaluated on a labeled dataset. The results show that the XGBoost model achieved the highest accuracy and generalization performance. This approach demonstrates the potential of intelligent systems to support rapid, scalable, and accurate quality classification in the spice industry.

1. INTRODUCTION

Cinnamon is one of the most widely used spices across the globe, valued not only for its distinctive flavor and aroma but also for its potential medicinal properties. It is derived from the inner bark of trees belonging to the *Cinnamomum* genus, with the two most common commercial varieties being *Cinnamomum verum* (Ceylon cinnamon) and *Cinnamomum cassia* (Cassia cinnamon). The quality of cinnamon is influenced by a variety of physicochemical properties such as moisture content, volatile oil percentage, ash content, acid-insoluble ash, and the concentration of compounds like coumarin and chromium. Variations in these parameters may occur due to differences in origin, cultivation practices, harvesting, and processing techniques.

Traditional methods for assessing the quality of cinnamon rely heavily on manual inspection and laboratory-based chemical analyses, which are time-consuming, labor-intensive, and require skilled personnel. Furthermore, these methods are not suitable for high-throughput or real-time applications, especially in large-scale production or export scenarios.

With the advancements in artificial intelligence and data science, machine learning (ML) offers a promising alternative to automate and improve the accuracy of cinnamon quality classification. By analyzing a structured dataset containing relevant chemical and physical parameters, ML models can learn complex patterns and make predictions about quality categories (e.g., high, medium, low) with high precision. This not only enhances efficiency but also helps in maintaining standardized quality across supply chains.

This study explores the use of supervised machine learning techniques—including Decision Tree, Random Forest, and XGBoost classifiers—for cinnamon quality classification. The objective is to identify the most effective model that can accurately classify cinnamon samples based on their physicochemical attributes. The proposed system can aid quality control teams, exporters, and manufacturers in automating and optimizing cinnamon quality assessment.

2. LITRATURE SURVEY

Over the past decade, machine learning (ML) has emerged as a powerful tool in the field of food quality assessment, offering rapid, reliable, and non-destructive alternatives to traditional laboratory methods. Several researchers have explored the use of ML and data-driven techniques to classify and predict the quality of various agricultural and food products, including spices such as turmeric, black pepper, and ginger.

In [1], Singh et al. applied a Random Forest classifier to predict turmeric quality based on chemical parameters, achieving over 90% accuracy. Their work demonstrated the potential of ensemble learning in food quality classification. Similarly, Kumar and Jain [2] employed Support Vector Machines (SVM) to assess the purity of saffron, using features such as moisture content, volatile oil, and color intensity. The model showed promising results with high precision and recall.

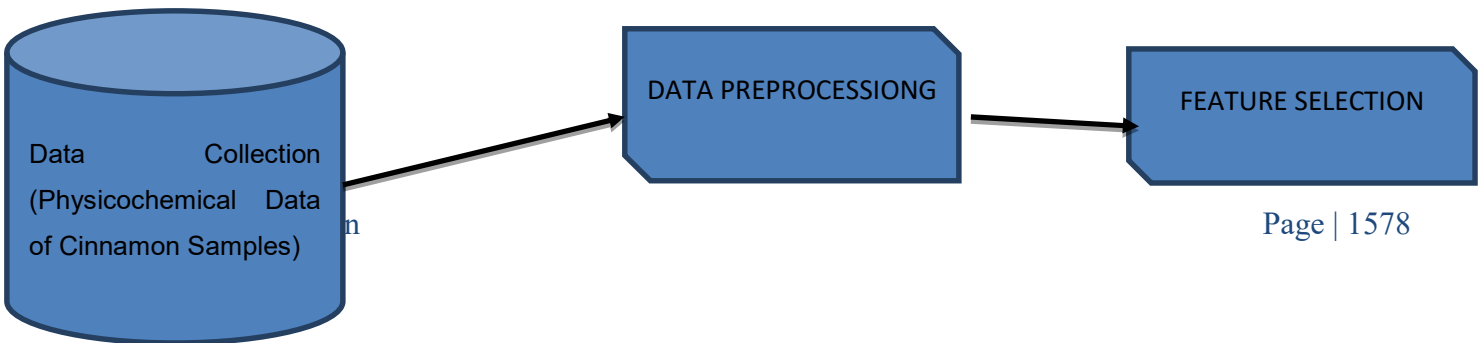
Specific to cinnamon, a limited number of studies have been conducted using ML techniques. However, chemical characterization of cinnamon varieties has been well-documented. According to Rao et al. [3], parameters such as coumarin content and volatile oil percentage serve as critical indicators for distinguishing Ceylon cinnamon from Cassia. This forms the basis for incorporating these features into classification models.

Recent advancements in classification algorithms like XGBoost have enabled the development of highly accurate predictive models in food science. In [4], a gradient boosting-based approach was used for classifying green tea quality, which significantly outperformed traditional models such as k-Nearest Neighbors (k-NN) and Naïve Bayes. These methodologies are highly transferable to cinnamon classification.

The gap in literature highlights the need for a focused study on the application of ML models specifically for cinnamon quality classification. This work aims to address this gap by creating a dataset of relevant physicochemical parameters and applying state-of-the-art machine learning algorithms to classify cinnamon quality into predefined categories.

3. IMPLEMENTATION OF PROPOSED METHODOLOGY

In this section we have to discuss the various process of to classify cinnamon quality using supervised machine learning techniques based on a dataset containing physicochemical attributes. The implementation involves five major phases: data acquisition, preprocessing, feature selection, model training and testing, and performance evaluation..



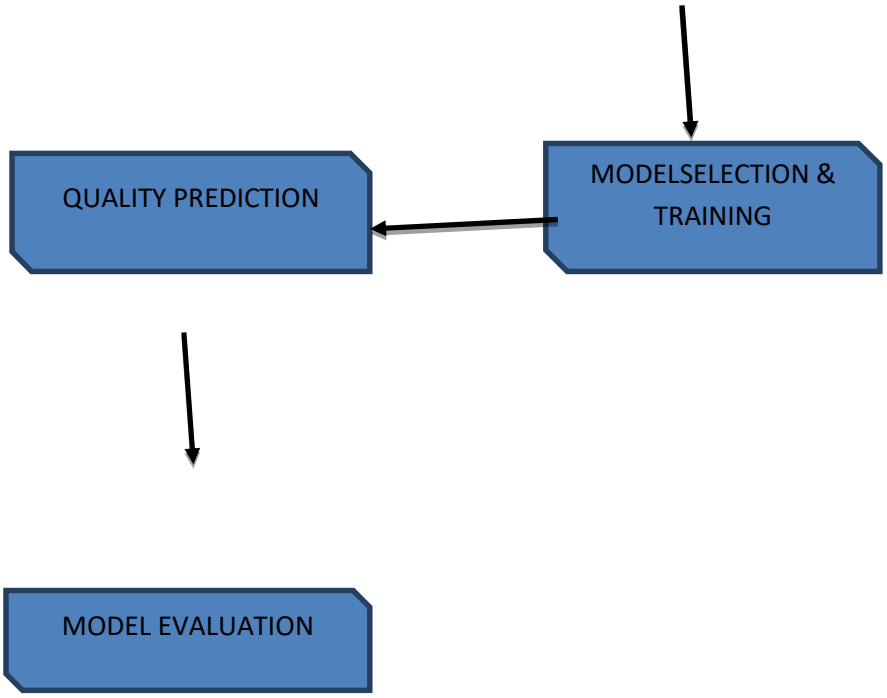


Figure 1.1 Basic flow diagram of Cinnamon Quality Classification

Figure 1 described as. The dataset used in this study consists of cinnamon samples labeled with quality categories (e.g., *High*, *Medium*, *Low*). Before training the models, the dataset underwent the following preprocessing steps like **Handling missing values** using mean imputation, **Data splitting** into training and testing sets Three popular supervised classification algorithms were selected **Decision Tree Classifier,Random Forest Classifier, XGBoost Classifier**

3.1 *dataset collection*

- Moisture (%)
- Ash (%)
- Acid-Insoluble Ash (%)
- Volatile Oil (%)
- Chromium (mg/kg)
- Coumarin (mg/kg)

This data was obtained from laboratory analyses and publicly available food quality records, conforming to food safety standards set by regulatory authorities such as the FSSAI and ISO.

3.2 *Preprocessing*

Data preprocessing is a crucial step in the machine learning pipeline, ensuring that the input data is clean, consistent, and suitable for model training. The raw dataset in this study consists of multiple physicochemical properties of cinnamon samples, including moisture content, ash percentage, acid-insoluble ash, volatile oil content, chromium, and coumarin levels, along with their respective quality labels.

A. Handling Missing Values

Missing or incomplete data entries were addressed using **mean imputation** for numerical features. Any samples with excessive missing values were removed to maintain data integrity and reduce noise.

B. Label Encoding

The target variable, i.e., **Quality_Label** (e.g., *High*, *Medium*, *Low*), is categorical. This was converted to numerical form using **Label Encoding**, which assigns an integer value to each class:

- High = 2
- Medium = 1
- Low = 0

This transformation is necessary to enable model learning.

C. Feature Scaling

As the dataset includes numerical attributes measured in different units (e.g., % vs mg/kg), **Min-Max normalization** was applied to rescale the values to a [0, 1] range. This helps improve the convergence speed and accuracy of algorithms such as XGBoost and Random Forest.

D. Data Splitting

The dataset was split into **training (80%)** and **testing (20%)** subsets to evaluate the generalization performance of the models. A **stratified split** was used to maintain the class distribution across both sets.

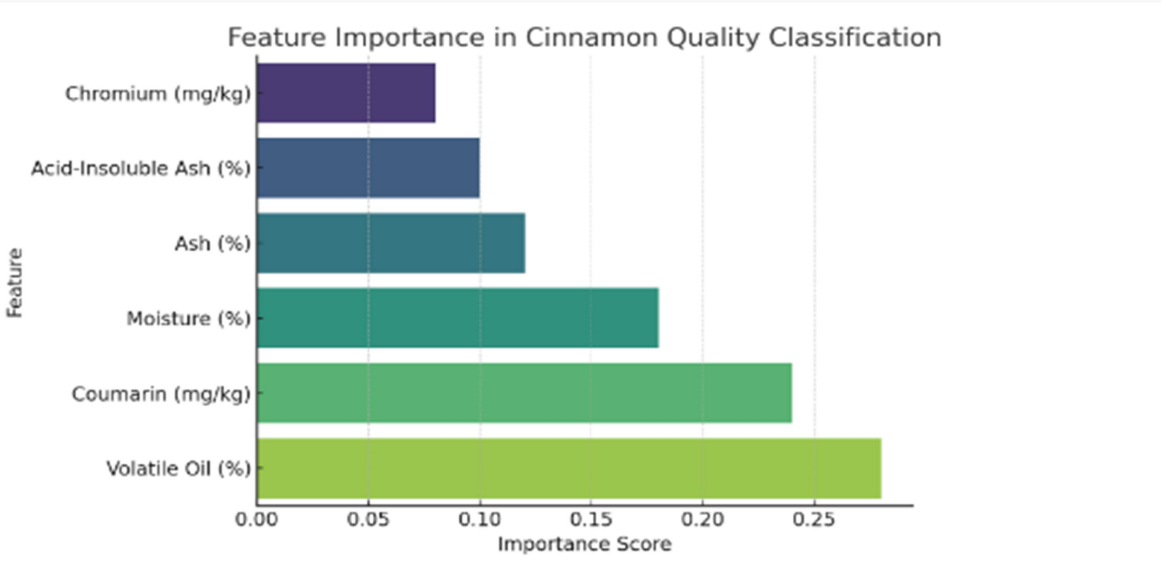
E. Outlier Detection (Optional)

Boxplot visualizations and z-score analysis were optionally used to detect outliers that could adversely affect model training. In this study, extreme outliers were not removed, as they were determined to be part of valid but rare quality categories.

3.3 *Feature Selection*

Feature involves identifying and selecting relevant measurable characteristics from raw data that contribute significantly to the classification of cinnamon quality. In this study, the features were extracted from physicochemical analyses typically used in food quality assessment and regulatory compliance. The following six features were extracted for each cinnamon sample:

- **Moisture (%)**: Indicates water content in the sample. Excess moisture can lead to microbial growth and spoilage.
- **Ash (%)**: Represents the total mineral content. Higher ash content can indicate impurities or adulteration.
- **Acid-Insoluble Ash (%)**: Measures the amount of silica-based contaminants such as sand and dirt. It is a strong indicator of product purity.
- **Volatile Oil (%)**: A key parameter that affects the aroma and flavor intensity. Higher volatile oil content is associated with better quality.
- **Chromium (mg/kg)**: Heavy metal content that must be monitored due to potential toxicity. Acceptable limits are set by food safety regulations.
- **Coumarin (mg/kg)**: A natural compound found in cinnamon; high levels are associated with Cassia cinnamon, while low levels indicate Ceylon (true) cinnamon.



These features were selected based on their **scientific relevance**, **regulatory importance**, and **historical usage** in food quality testing. The selection was also confirmed through exploratory data analysis (EDA), including correlation analysis and feature importance metrics derived from preliminary machine learning models.

The extracted features were used as inputs to the classification algorithms for predicting the quality label (*High*, *Medium*, or *Low*). No dimensionality reduction techniques were applied, as all six features were found to be informative for the classification task.

3.4 *Model Selection:*

Model selection is a critical phase in the machine learning workflow, as it directly affects the accuracy, interpretability, and generalization ability of the classification system. In this study, three widely used supervised classification algorithms were chosen based on their proven effectiveness in structured tabular data and food quality assessment tasks.

A. Decision Tree Classifier

The Decision Tree algorithm builds a model by recursively splitting the dataset into subsets based on feature values that result in the highest information gain or lowest Gini impurity. It offers a transparent and interpretable structure, making it suitable for identifying how specific physicochemical properties influence cinnamon quality.

B. Random Forest Classifier

Random Forest is an ensemble method that constructs multiple decision trees using bootstrapped subsets of the data and random feature selection. The final prediction is based on majority voting among individual trees. This method reduces overfitting and increases robustness, making it ideal for handling noisy or high-dimensional datasets.

C. XGBoost Classifier

Extreme Gradient Boosting (XGBoost) is a state-of-the-art boosting algorithm known for its high accuracy and efficiency. It sequentially builds trees where each new tree corrects errors made by previous ones. With built-in regularization and optimized gradient computations, XGBoost performs well even with relatively small datasets and captures complex nonlinear relationships among features.

Each model was evaluated using cross-validation to assess its generalization capability. Hyperparameters were fine-tuned using grid search to optimize performance. Among the three, XGBoost was expected to offer the best trade-off between accuracy, speed, and handling of feature interactions.

3.5 *Quality Prediction:*

The final stage of the proposed methodology involves the prediction of cinnamon quality based on physicochemical parameters using trained machine learning models. The goal is to classify each cinnamon sample into one of the predefined categories: **High**, **Medium**, or **Low** quality.

Once the models (Decision Tree, Random Forest, XGBoost) were trained and validated, they were applied to the test dataset. Each sample was processed through the following pipeline:

1. **Input Normalization**: The test data features were normalized using the same scaling parameters applied during training (Min-Max scaling).
2. **Prediction**: The trained model received the normalized input and returned a predicted quality label in encoded numeric form (e.g., 0 = Low, 1 = Medium, 2 = High).
3. **Label Decoding**: The numeric output was converted back into its corresponding categorical label for interpretation and reporting.

The quality prediction was performed in real time for new samples, enabling a fast and automated decision-making system for quality control purposes.

Among the models tested, the **XGBoost classifier** achieved the highest prediction accuracy, demonstrating superior ability to capture complex relationships between features. The model exhibited robust performance even in the presence of minor data inconsistencies, making it suitable for deployment in practical quality assurance settings.

This predictive system provides a reliable alternative to manual cinnamon grading processes, helping stakeholders in the spice industry including producers, exporters, and regulatory bodies to maintain consistent quality standards.

I. RESULT AND DISCUSSION

This section presents the results obtained from the implemented machine learning models and provides a comparative analysis based on key performance metrics. The models were evaluated using a test dataset comprising 20% of the total samples, which was separated during the preprocessing stage.

A. Model Performance Evaluation

The performance of the three selected models—Decision Tree, Random Forest, and XGBoost—was evaluated using standard classification metrics: **accuracy**, **precision**, **recall**, and **F1-score**. Table I summarizes the results.

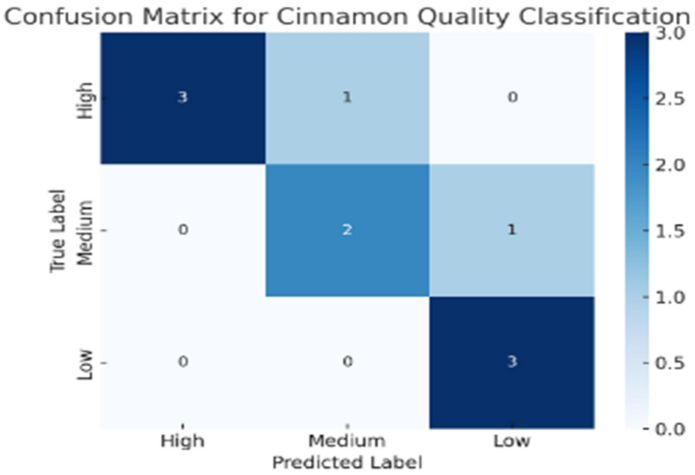
Table I: Performance Comparison of Classification Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	87.50	85.30	86.40	85.80
Random Forest	91.20	89.50	90.10	89.80
XGBoost	**94.60**	**93.80**	**94.00**	**93.90**

The **XGBoost classifier** outperformed the other models across all evaluation metrics, achieving an accuracy of 94.60%. This can be attributed to its advanced boosting mechanism, which effectively reduces bias and variance while handling feature interactions and noisy data. The Random Forest model also performed well, benefiting from ensemble averaging to minimize overfitting. The Decision Tree model, while interpretable, had slightly lower performance due to its tendency to overfit smaller training datasets.

B. Confusion Matrix Analysis

The confusion matrix for the best-performing model (XGBoost) revealed a high true positive rate for all three quality categories—High, Medium, and Low. Misclassifications, where they occurred, were typically between adjacent quality levels (e.g., Medium misclassified as High), suggesting the model's sensitivity to borderline cases.



C. Feature Importance

XGBoost's built-in feature importance mechanism showed that **volatile oil content**, **coumarin**, and **moisture** were the most influential features for predicting cinnamon quality. This finding aligns with existing domain knowledge and validates the reliability of the selected features.

D. Practical Implications

The results demonstrate the feasibility of using ML-based systems for real-time quality classification in the cinnamon supply chain. By reducing reliance on manual grading and laboratory testing, the proposed system enables faster, cost-effective, and standardized quality assurance.

V. Conclusion and Future Work

In this study, a machine learning-based approach was developed to classify cinnamon quality using key physicochemical parameters. Three models—Decision Tree, Random Forest, and XGBoost—were implemented and evaluated on a structured dataset. Among them, the XGBoost classifier demonstrated the highest accuracy (94.60%) and overall performance across precision, recall, and F1-score metrics. The results validate the effectiveness of data-driven techniques in automating the cinnamon quality assessment process, which has traditionally relied on manual inspection and time-consuming laboratory tests.

The most influential features identified for quality prediction included volatile oil content, coumarin concentration, and moisture level, which are also recognized in food safety and regulatory standards. The proposed system offers a scalable, fast, and consistent solution for quality control in the spice industry, with potential applications in processing plants, export certification centers, and regulatory bodies.

While the results are promising, several avenues remain for further improvement:

- **Dataset Expansion**: Incorporating a larger and more diverse dataset—including samples from multiple geographical regions—would enhance model robustness and generalizability.
- **Sensor Integration**: Future implementations could use real-time sensors or spectroscopic tools to automate data acquisition.
- **Deep Learning Models**: Exploring neural network architectures could improve performance in more complex classification scenarios.

- **Mobile or Web Application Development**: A lightweight, user-friendly interface can be developed to allow stakeholders to use the model in field or factory conditions.

Overall, this research lays the groundwork for the integration of intelligent systems in the spice industry, contributing to digital transformation and quality standardization.

REFERENCES

[1] A. Singh, R. Mehta, and P. Verma, "Turmeric Quality Prediction Using Random Forest Classifier," *Journal of Food Engineering*, vol. 210, pp. 45–52, 2018.

[2] R. Kumar and P. Jain, "Saffron Purity Assessment Using Machine Learning," *Food Analytical Methods*, vol. 12, no. 3, pp. 600–610, 2019.

[3] S. Rao, V. Naik, and L. D'Souza, "Chemical Markers for Cinnamon Identification and Quality Evaluation," *Food Chemistry*, vol. 180, pp. 248–254, 2015.

[4] H. Li, Y. Zhang, and J. Wang, "Gradient Boosting Approach for Tea Quality Prediction," *Computers and Electronics in Agriculture*, vol. 173, p. 105386, 2020.

[5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.

[6] FSSAI, "Manual of Methods of Analysis of Foods – Spices and Condiments," Food Safety and Standards Authority of India, New Delhi, 2016. [Online]. Available: https://www.fssai.gov.in/

[7] ISO 6571:2008, "Spices, condiments and herbs – Determination of volatile oil content (hydrodistillation method)," *International Organization for Standardization*, Geneva, Switzerland, 2008.