

# **Detection of Starch Adulteration in Turmeric Powder Using FTIR Spectroscopy and Machine Learning Techniques**

**BACHELOR TERM PROJECT REPORT**

*by*

**Likhith Gaikwad D**

**(21ME10044)**

Under the supervision of

**Prof. Mamoni Banerjee**

**Rajendra Mishra School of Engineering Entrepreneurship**



**Rajendra Mishra School of Engineering Entrepreneurship**

**INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR**

**(Spring 2024-25)**

# Declaration

I certify that

- a. The work contained in this report has been done by me under the guidance of my supervisor.
- b. The work has not been submitted to any other Institute for any degree or diploma.
- c. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
- d. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

Name: Likhith Gaikwad D

Roll number: 21ME10044

Date: 02-05-2025

Place: Kharagpur



**Rajendra Mishra School of Engineering Entrepreneurship**

**INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR**

**CERTIFICATE**

This is to certify that the thesis entitled “**Detection of Starch Adulteration in Turmeric Powder Using FTIR Spectroscopy and Machine Learning Techniques**” submitted by Likhit Gaikwad D (21ME10044), Department of Mechanical Engineering, IIT Kharagpur, for the completion of the Bachelor’s Thesis Project 2 for the degree of Bachelors of Technology in Mechanical Engineering, is a record of bonafide work carried out by him under my supervision during the period, 01.01.2025 to 30.04.2025, as per the BTP guidelines for the Spring Semester 2024-2025. The report presented here has not been previously submitted, nor will it be submitted, partially or entirely, for the fulfillment of any other degree or diploma within this institution or any other educational establishment. This declaration complies with the regulations set forth by the University, and I am confident that it adheres to the requisite standards for submission.

Date: 02-05-2025

Place: Kharagpur

Prof. Mamoni Banerjee

Rajendra Mishra School of Engineering Entrepreneurship

Indian Institute of Technology, Kharagpur

## **Table of Contents**

	<b>Page No.</b>
<b>1. Abstract</b>	<b>1</b>
<b>2. Introduction</b>	<b>2</b>
<b>3. Literature Review</b>	<b>4</b>
<b>4. Methodology</b>	<b>10</b>
<b>5. Results and Discussion</b>	<b>17</b>
<b>6. Conclusion</b>	<b>28</b>
<b>7. Limitations</b>	<b>29</b>
<b>8. Recommendations</b>	<b>30</b>
<b>9. Future Directions</b>	<b>30</b>
<b>10. Potential Applications and Usage</b>	<b>31</b>
<b>11. Bibliography</b>	<b>31</b>
<b>12. Tables and Plots</b>	<b>34</b>

## Abstract

Turmeric is a widely consumed spice known for its therapeutic, nutritional, and culinary applications. However, its commercial value has made it susceptible to adulteration, with starch being one of the most common and economically motivated adulterants. This study presents a machine learning-assisted methodology for detecting and quantifying starch adulteration in turmeric powder using Fourier Transform Infrared (FTIR) spectroscopy. Spectral data were pre-processed using techniques such as Savitzky–Golay smoothing, Standard Normal Variate (SNV), and Extended Multiplicative Scatter Correction (EMSC) to enhance signal quality. Feature extraction and selection methods, including Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE), were employed to reduce dimensionality and highlight relevant spectral regions. Multiple regression models—Linear Regression (LR), Decision Tree Regression (DT), Support Vector Regression (SVR), and Artificial Neural Networks (ANN)—were trained and evaluated. Among them, the optimized ANN model demonstrated superior performance, achieving an  $R^2$  value exceeding 0.88 on the validation dataset. These findings underline the potential of combining FTIR spectroscopy with advanced machine learning algorithms to ensure the authenticity and safety of turmeric, offering a scalable solution for food quality control and adulteration monitoring.

**Keywords:** Turmeric adulteration, Starch, FTIR spectroscopy, Machine Learning, Artificial Neural Network, Food authentication, Spectral preprocessing.

## Introduction

**Turmeric**, a golden-yellow spice derived from the rhizome of *Curcuma longa*, is a staple ingredient in **culinary**, **medicinal**, and **cosmetic** applications due to its rich *curcuminoid*

*content* and *bioactive properties*. Revered in traditional systems of medicine like **Ayurveda**, it is valued for its **anti-inflammatory**, **antioxidant**, and **antiseptic** effects. With the growing global demand for turmeric in **food** and **nutraceutical industries**, concerns surrounding its **quality** and **authenticity** have intensified.

In recent years, **food adulteration** has emerged as a critical challenge affecting both **food safety** and **consumer trust**. Turmeric, due to its **powdered form** and **high market value**, is particularly susceptible to adulteration. Common adulterants include **metanil yellow**, **lead chromate**, and **starches** such as corn, rice, or tapioca powder. Among these, **starch adulteration** is prevalent because it is **difficult to detect visually** and is often used to **increase volume** at **lower costs**. This compromises not only the **nutritional integrity** of turmeric but also poses **health hazards**, especially for individuals with **dietary restrictions** or **allergies**.

Conventional methods of detecting adulteration—such as **microscopy**, **iodine tests**, and **chromatography**—although useful, often lack the **speed**, **sensitivity**, or **scalability** required for modern **food quality control**. In this context, **spectroscopic techniques** such as **Fourier Transform Infrared (FTIR) Spectroscopy** have gained traction. FTIR enables **rapid, non-destructive** analysis of chemical compounds and is capable of detecting **subtle molecular differences** introduced by adulterants.

**Fourier Transform Infrared (FTIR) spectroscopy** emerges as a vital tool in detecting **food fraud** due to its **rapid** and **reliable detection capabilities**. Operating in the **mid-infrared range** (2500 nm to 25,000 nm or  $4000\text{ cm}^{-1}$  to  $400\text{ cm}^{-1}$ ), FTIR provides **critical molecular insights** into food samples. This enables **accurate** and **reliable identification** of **trace amounts of adulterants** (Goyal *et al.*, 2024). Detecting adulteration in food samples requires correlating **spectral data** with **sample characteristics**, a process known as **calibration**.

However, **manual calibration** poses challenges due to **overlapping peaks**, **noise**, and **external factors** affecting spectral readings (*Valand et al., 2019*). Hence, more **robust**, **automated techniques** with strong **decision-making capabilities** are necessary.

However, interpreting FTIR spectra **manually** or using **classical statistical methods** can be challenging due to **non-linear patterns** in the spectral data. To overcome this, recent advancements advocate for the **integration of machine learning (ML) techniques** with spectroscopy. ML algorithms, particularly **Artificial Neural Networks (ANN)**, can model **complex non-linear relationships** between spectral features and adulterant concentration, thereby enhancing **prediction accuracy**.

Despite these technological advancements, there remains a **gap** in the application of **machine learning-assisted FTIR spectroscopy** for **quantitative adulteration analysis** in turmeric, particularly involving **starch**. Most existing studies are limited to **binary classification** (*adulterated vs. non-adulterated*), rather than providing a **quantifiable measure of adulteration percentage**.

The purpose of this study is to **bridge this gap** by developing a robust pipeline that integrates **FTIR spectroscopy** with advanced **machine learning models**—including **Linear Regression (LR)**, **Decision Tree (DT)**, **Support Vector Regression (SVR)**, **XG Boost** and **Optimized ANN**—for the **detection and quantification** of starch adulteration in **turmeric powder**. This work emphasizes **data preprocessing**, **feature selection**, and **model comparison**, thereby offering a **comprehensive approach** for future **food adulteration research**.

## Literature Review

Food adulteration, especially in powdered spices like turmeric, has garnered increasing attention due to its implications on consumer health and market fairness. Several studies have

identified the widespread use of starch-based adulterants in turmeric, often introduced to increase product volume and profit margins. These adulterants can be visually indistinguishable, making traditional detection techniques inadequate for precise quantification.

### **Spectroscopic Techniques in Adulteration Detection:**

Among the various analytical approaches, spectroscopic methods, particularly **Fourier Transform Infrared (FTIR) Spectroscopy**, have shown promise due to their non-destructive nature, rapid measurement, and sensitivity to chemical bonds. FTIR allows for the identification of functional groups that change in the presence of adulterants like starch, which typically introduces additional -OH and C-H absorption bands in the IR spectrum. Researchers such as Dhakal et al. (2016) and Kumar et al. (2019) have highlighted FTIR's potential in identifying adulterants in turmeric and other food matrices, especially when paired with chemometric or machine learning approaches.

### **Use of Machine Learning in Spectral Analysis:**

Machine learning (ML) models, especially supervised learning algorithms, are increasingly being applied to spectral data for classification and regression tasks. Studies have shown that algorithms like **Support Vector Machines (SVM)**, **Decision Trees**, XG Boost and **Artificial Neural Networks (ANNs)** can effectively extract non-linear relationships from spectral features that are otherwise difficult to model using classical statistical approaches. In the context of food adulteration, these models have been employed for the detection of adulterants in milk, honey, tea, and spices. For example, Jain et al. (2020) demonstrated the use of ANN on NIR spectra for identifying starch adulteration in food powders, achieving promising accuracy levels.



### **Preprocessing and Feature Selection:**

Effective preprocessing of spectral data is crucial to improve signal-to-noise ratio and eliminate baseline shifts. Methods such as **Savitzky–Golay smoothing**, **Standard Normal Variate (SNV)**, and **Extended Multiplicative Scatter Correction (EMSC)** are widely recommended. Additionally, dimensionality reduction and feature selection techniques like **Principal Component Analysis (PCA)** and **Recursive Feature Elimination (RFE)** help reduce redundancy and enhance model performance. Research by Zhang et al. (2021) demonstrated that applying these techniques significantly improved adulteration detection accuracy in herbal powders.

### **Research Gap:**

While several studies have utilized FTIR and machine learning independently, the combination of optimized spectral preprocessing, robust feature selection, and advanced ML models like ANN for **quantitative prediction of starch adulteration in turmeric** is underexplored. Existing work either lacks comprehensive preprocessing pipelines or does not focus on regression-based adulteration quantification, limiting their application in real-world food safety monitoring.

## **Methodology**

### **Sample Preparation: Turmeric Powder Adulterated with Starch**

For this research, **250 grams Turmeric Powder** and **100 grams Corn Starch Powder** was purchased from the **Tech market (Indian Institute of Technology Kharagpur, West Bengal, India)**. To facilitate the development of a machine learning-assisted FTIR spectroscopy model for adulteration detection, turmeric powder samples were systematically adulterated with known proportions of starch. The adulterant selected was corn starch, a

commonly used, visually indistinguishable and inexpensive filler often found in fraudulent turmeric products.

Six sample groups were prepared to represent different levels of starch adulteration: **0%, 10%, 20%, 30%, 40%, and 50% (w/w)**. For each concentration, **1 gram of total sample weight** was maintained for consistency across FTIR analysis.

- **0% adulteration:** This served as the control sample, consisting of **1 gram of pure turmeric powder**, carefully weighed using a precision digital microbalance. The sample was transferred into a clean, labelled microcentrifuge tube and sealed to avoid contamination or moisture absorption.
- **10% adulteration:** For this batch, **0.9 grams of turmeric powder** was thoroughly mixed with **0.1 gram of corn starch**. The powders were homogenized using a glass stir rod on a sterile glass surface to ensure uniform mixing. The mixed sample was collected into a fresh microtube and labelled accordingly.
- **20% adulteration:** Here, **0.8 grams of turmeric** was combined with **0.2 grams of starch**. The mixture was homogenized manually for 2–3 minutes, ensuring even distribution of the adulterant. It was then transferred into a sterile storage tube.
- **30% adulteration:** **0.7 grams of turmeric** was blended with **0.3 grams of starch**. After mixing, the blend exhibited a slightly lighter yellow hue due to the dilution of turmeric pigments. The sample was stored in a clean tube and sealed.
- **40% adulteration:** For this set, **0.6 grams of turmeric** and **0.4 grams of starch** were combined. Extended homogenization was performed to avoid spectral inconsistencies during FTIR scanning.

- **50% adulteration:** An equal mix of **0.5 grams turmeric** and **0.5 grams starch** formed the most adulterated sample in the study. The colour, texture were markedly different from the pure sample, indicating a substantial alteration in composition.

All prepared samples were stored in small, airtight, and light-protective plastic microtubes, labelled with respective adulteration percentages. They were kept in a dry, ambient environment until subjected to FTIR spectral scanning. Each sample was scanned multiple times to minimize instrumental noise and confirm data reproducibility.

### **How FTIR Works – A Brief Overview**

FTIR (Fourier Transform Infrared) spectroscopy is an analytical technique used to identify and quantify chemical compounds based on how molecules absorb infrared (IR) light. When IR radiation is passed through a sample, specific wavelengths are absorbed depending on the molecular bonds present. These absorbed frequencies correspond to the **vibrational modes** of the molecules.

The core of the FTIR instrument is the **interferometer**, which modulates the IR light and produces an **interferogram**—a signal representing light intensity as a function of the position of a moving mirror. This signal is then mathematically converted using **Fourier Transform** into an IR spectrum, showing **absorbance vs. wavenumber ( $\text{cm}^{-1}$ )**.

Each compound has a **unique IR fingerprint**, allowing FTIR to detect functional groups and chemical changes, including adulterants, in complex samples like turmeric powder.

### **Pellet Press Overview and FTIR Scanning**

To prepare the turmeric-starch mixtures for FTIR analysis, each 1 g sample was finely ground and homogenized. The prepared powders were then transferred to a **pellet press** unit. In this setup, a precise amount of the sample was compressed under high pressure to form a thin,

compact disc (pellet). This ensures uniform thickness and optimal surface contact for infrared light transmission, minimizing scattering and enhancing spectral quality.



Pellet Press Machine



FTIR Machine



FTIR Machine

**Spectral data collection and analysis**

Once the pellets were formed, they were carefully placed in the **FTIR sample holder**. Each sample underwent **32 scans** in the mid-infrared range ( $4000\text{ cm}^{-1}$  to  $400\text{ cm}^{-1}$ ) using the transmission mode. This provided high-resolution spectral data with improved signal-to-noise ratio. The collected spectra were saved for subsequent preprocessing and machine learning analysis to detect and quantify the level of starch adulteration.

The collected spectral data were processed using **Jupyter Notebook**, a web-based application for machine learning, utilizing **Python 3.10.9**. The following updated libraries were employed for data handling and analysis: **NumPy 2.2.5**, **Pandas 2.2.3**, **Matplotlib 3.10.1**, **Seaborn 0.13.2**, **TensorFlow 2.19.0**, and **Scikit-Learn 1.6.1**. These tools provided the computational backbone for preprocessing, visualization, feature extraction, and model development. The analysis was conducted on a system running **Windows 11 Home Single Language** equipped with an **AMD Ryzen 7 5800H processor**, **16 GB RAM**, and **NVIDIA GeForce RTX 3050 GPU**, ensuring smooth execution of ML workflows. The raw FTIR data and corresponding adulteration percentages were compiled into an **MS Excel file**, referred to as the **original dataset "X\_raw"** throughout the manuscript.

### **Preprocessing of spectral data**

In this study, various preprocessing techniques were applied to the raw spectral data to enhance the quality of the data and improve the accuracy of the prediction models for food adulteration detection. The preprocessing methods implemented include Savitzky-Golay (SG) smoothing, Standard Normal Variate (SNV), and Extended Multiplicative Scatter Correction (EMSC), each addressing specific challenges in the spectral data.

The **Savitzky-Golay (SG) smoothing** technique was employed to reduce small-scale noise in the raw spectral data. This method is particularly effective for smoothing the absorption

bands in the spectra, which facilitates the identification of steep intensity changes while eliminating weaker noise. In this study, a window length of 11 and a polynomial order of 2 were used for SG smoothing. This technique helps in improving the spectral clarity and ensures that only the relevant features are highlighted for analysis.

To address unwanted variations in the spectra caused by scatter effects, **Standard Normal Variate (SNV)** preprocessing was applied. SNV standardizes the spectra by subtracting the mean and dividing by the standard deviation of each spectrum. This transformation removes spectral variations due to sample thickness and instrumental effects, making it easier to compare the spectra and enhancing the ability to interpret spectral features.

Finally, **Extended Multiplicative Scatter Correction (EMSC)** was applied to address both additive and multiplicative variations in the spectral data. EMSC corrects for variations in sample thickness, instrumental drift, and scattering effects that could impact the spectral analysis. The method involves dividing the raw data by the mean spectrum, with safeguards to avoid division by zero using a small epsilon value ( $1e-10$ ). This ensures that the data is normalized while preserving the important features necessary for predicting the adulteration levels.

The datasets resulting from the application of SG smoothing, SNV, and EMSC preprocessing methods were labeled as “X\_sg”, “X\_snv”, and “X\_emsc”, respectively, in the manuscript, with each dataset representing a distinct preprocessing approach applied to the raw spectral dataset “X\_raw”.

### **Base models and their evaluation of performance**

Various ML algorithms are available for predicting food adulteration. For this study, common models of different classes including linear regression (LR), decision tree (DT), support vector regression (SVR), XG Boost and artificial neural network (ANN) were selected as

base models. LR and SVM were selected to understand the linear and non-linear relationship between the spectrum Journal Pre-proof Journal Pre-proof and the adulteration levels respectively. On the other hand, DT selected from tree type ML algorithms to understand the hierarchical structure, XGBoost was selected to capture complex **non-linear interactions** between spectral features and adulteration levels while maintaining computational efficiency. and ANN selected from neural network architecture for considering the complex non-linear relationship. The original and pre-processed datasets “X\_raw”, “X\_sg”, “X\_snv”, and “X\_emsc” were trained and tested with an 80:20 split from training set to testing set.

### 1. **Linear Regression (LR):**

- **Purpose:** LR is used to model the linear relationship between the input spectral data and the adulteration levels (target variable). It helps in understanding how well the features (wavelengths) relate to the adulteration percentage in a linear manner.
- **Method:** The model finds the best-fitting line that minimizes the residuals (errors) between predicted and actual values. In your case, it's used to predict the adulteration percentage of turmeric based on spectral data.

### 2. **Support Vector Regression (SVR):**

- **Purpose:** SVR is employed to capture non-linear relationships between the input spectral data and the adulteration levels. SVR is suitable for high-dimensional data like FTIR spectra and can capture complex patterns that linear models might miss.
- **Method:** SVR uses kernel functions (such as the radial basis function, RBF) to transform the input features into higher dimensions, allowing it to find non-linear relationships more effectively.

### 3. **Decision Tree (DT):**

- **Purpose:** DT is used to explore hierarchical structures in the data, identifying which features (spectral bands) have the most significant impact on the adulteration level. It works by splitting the data into distinct branches based on feature values, which makes it easier to interpret the decision-making process.
- **Method:** DT models create a tree structure that divides the data into subsets with the least impurity, offering transparency in understanding which spectral features contribute most to predicting the adulteration levels.

#### 4. **Extreme Gradient Boosting (XGBoost):**

- **Purpose:** XGBoost is employed to enhance prediction performance by combining the strengths of multiple **decision trees** in an **ensemble learning** framework. It is known for its speed and accuracy.
- **Method:** XGBoost builds a series of weak learners (trees) where each subsequent tree attempts to correct the errors made by the previous ones, using a technique called **gradient boosting**. It uses **regularization** to reduce overfitting and handles missing values efficiently

#### 5. **Artificial Neural Network (ANN):**

- **Purpose:** ANN is chosen for capturing complex, non-linear relationships between the input spectral data and adulteration levels. ANNs, with their ability to model intricate patterns and interactions, are particularly effective for high-dimensional data like FTIR spectra.
- **Method:** ANNs consist of layers of interconnected neurons, with input layers receiving the spectral data, hidden layers processing it, and output layers predicting



the adulteration level. The network is trained using an optimization process (like backpropagation) to minimize the prediction error.

In this study, an Artificial Neural Network (ANN) model was employed to predict turmeric adulteration percentage from FTIR spectral data. The ANN was implemented using the MLPRegressor with two hidden layers consisting of 64 and 32 neurons respectively. The rectified linear unit (ReLU) activation function was used in the hidden layers to effectively model non-linear relationships in the data. The output layer used a linear activation function suitable for regression tasks. The model was trained with a maximum of 1000 iterations and a fixed random state (42) to ensure reproducibility. This architecture was selected to balance the complexity and generalization performance while capturing important patterns from the pre-processed spectral features.

### **Performance evaluation**

Model performance was assessed using determination coefficient ( $R^2$ ), mean absolute error (MAE), and root mean square error (RMSE) on validation and testing sets.  $R^2$  measures variability explained by the model, RMSE evaluates the residual standard deviation, and MAE calculates average residual error. Models with high  $R^2$  and low RMSE were considered optimal. The two best-performing model-dataset combinations were selected for further feature engineering and optimization.

### **Feature selection and extraction for adulteration prediction**

To address the high dimensionality of FTIR spectral data, which comprised 7469 wavenumber features, this study employed dimensionality reduction strategies to improve computational efficiency and simplify machine learning models used for turmeric adulteration prediction. Two complementary techniques—feature selection and feature

extraction—were applied to the best-performing pre-processed datasets identified in the base modelling stage.

Feature selection was performed using Recursive Feature Elimination (RFE), a backward elimination method that identifies the most informative features by recursively removing those with the least impact on model performance. In this study, Support Vector Regression (SVR) with a linear kernel was used as the base estimator for RFE. The model was trained to iteratively rank the features based on their importance scores derived from regression coefficients. The top 15 features were retained as they showed the strongest contribution to predicting adulteration levels. These selected wavelengths are presumed to reflect critical spectral information tied to specific molecular bonds or functional groups relevant to pure and adulterated turmeric.

In this study, feature extraction was performed using Principal Component Analysis (PCA), a statistical technique that transforms high-dimensional data into a smaller set of uncorrelated variables, called principal components (PCs). Before applying PCA, the data underwent z-score normalization to standardize it, ensuring that all features contributed equally.

PCA was then applied, and six principal components were selected to represent the data. These components captured over 95% of the total spectral variance, significantly reducing the feature space while preserving essential chemical variability. The components were retained based on their explained variance ratio, which quantified each component's contribution to capturing variability in the dataset. By selecting components that together explained at least 95% of the variance, this approach minimized information loss and ensured that the most important spectral features were maintained.

To enhance the interpretability of the PCA-derived features, the top five most influential wavenumbers for each principal component were identified. This was done by examining the absolute weights of each original feature within the PCA loading matrix. For each of the six principal components, the five wavenumbers with the highest absolute contributions were extracted along with their corresponding weights. These dominant wavenumbers highlight the spectral regions that contribute most to the variation in the dataset and are likely associated with specific chemical structures or adulteration signatures in turmeric powder.

This interpretability step bridges the gap between the statistical transformation of the data and the domain-specific spectral analysis, offering insights into which spectral regions play a key role in detecting adulteration levels.

### **Model optimization by hyperparameter tuning for predicting the adulteration in Turmeric powder**

After identifying the best model among the base models, the selected model was optimized to fine-tune on 6 datasets. These included two best-selected datasets and their feature-selected and feature-extracted counterparts. The fine-tuning of the selected model was done by the selection of the best set of hyperparameters of the given model also termed as hyperparameter tuning. ANN, which provided the best results was chosen for hyperparameter tuning.

For hyperparameter tuning, **GridSearchCV** was used to identify the best-performing model across several preprocessing methods and feature engineering approaches. This process aimed at optimizing the performance of the **Artificial Neural Network (ANN)**, specifically

the **MLPRegressor** from **sklearn.neural\_network**, by varying key hyperparameters to improve prediction accuracy for food adulteration detection.

The primary hyperparameters optimized were:

1. **Number of hidden layers:** The grid search tested three configurations: a single layer with 32 neurons, a single layer with 64 neurons, and two layers with 64 and 32 neurons.
2. **Activation function:** The ReLU (rectified linear unit) activation function was selected for all models, as it is commonly used for non-linear regression tasks, providing better performance on large datasets.
3. **Alpha (regularization term):** Two values, 0.0001 and 0.001, were tested to control the amount of regularization applied during training, helping to prevent overfitting.
4. **Learning rate:** The learning rate was kept constant across models, ensuring stable convergence during training.

The tuning process was applied across various datasets resulting from different preprocessing methods. The dataset-specific models were evaluated using **tenfold cross-validation**, with an **80:20 split** between training and testing sets.

For each dataset, the **MLPRegressor** was trained within a pipeline that included **StandardScaler** for feature scaling, followed by the ANN model. The **GridSearchCV** conducted exhaustive search over all hyperparameter combinations, returning the best parameters that optimized **R<sup>2</sup>**, **RMSE**, and **MAE** on the validation data.

After hyperparameter optimization, the results for each dataset were captured, including the best hyperparameters, **R<sup>2</sup> score**, **RMSE**, and **MAE**, to assess the ANN's performance across different preprocessing techniques. This approach efficiently identified the most promising

feature transformations and model configurations, ensuring the best possible predictions for food adulteration detection.

## Results and discussion

### Spectral analysis of Turmeric powder and starch

#### Most Prominent Peaks

From the **raw FTIR spectra** and the **Savitzky-Golay smoothed spectra**, the most prominent absorbance peaks for your samples are observed at the following approximate wavenumbers:

- **$\sim 3400\text{ cm}^{-1}$** : Strong, broad peak
- **$\sim 2920\text{ cm}^{-1}$** : Medium peak
- **$\sim 2850\text{ cm}^{-1}$** : Medium peak
- **$\sim 1740\text{ cm}^{-1}$** : Sharp, strong peak
- **$\sim 1650\text{ cm}^{-1}$** : Medium peak
- **$\sim 1460\text{ cm}^{-1}$** : Medium peak
- **$\sim 1375\text{ cm}^{-1}$** : Small to medium peak
- **$\sim 1240\text{ cm}^{-1}$** : Medium peak
- **$\sim 1160\text{ cm}^{-1}$** : Medium peak
- **$\sim 1050\text{ cm}^{-1}$** : Medium peak
- **$\sim 720\text{ cm}^{-1}$** : Small but distinct peak

These peak positions are estimated visually from the provided spectra plots (especially the smoothed and raw spectra for clarity).

## Chemical Bond Assignments

The following table summarizes the likely chemical bond assignments for each major peak, based on standard FTIR interpretation:

Wavenumber (cm <sup>-1</sup> )	Typical Assignment	Possible Functional Group/Bond
~3400	O–H/N–H stretch (broad)	Alcohols, phenols, amines, water
~2920	C–H stretch (asymmetric, aliphatic)	CH <sub>2</sub> /CH <sub>3</sub> groups (alkanes, lipids)
~2850	C–H stretch (symmetric, aliphatic)	CH <sub>2</sub> /CH <sub>3</sub> groups (alkanes, lipids)
~1740	C=O stretch (strong, sharp)	Esters, aldehydes, ketones, carboxylic acids
~1650	C=C or C=O stretch (amide I, C=C)	Amides (proteins), alkenes, conjugated C=O
~1460	C–H bend (scissoring)	CH <sub>2</sub> /CH <sub>3</sub> groups (alkanes)
~1375	C–H bend (methyl, symmetric)	CH <sub>3</sub> groups
~1240	C–O stretch, C–N stretch	Esters, ethers, amines
~1160	C–O stretch	Esters, ethers
~1050	C–O stretch, C–N stretch	Alcohols, ethers, amines
~720	(CH <sub>2</sub> ) <sub>n</sub> rock (long chain methylene)	Alkanes (long chain), lipids

### Notes:

- The **broad peak at ~3400 cm<sup>-1</sup>** is characteristic of O–H stretching (hydroxyl groups, water) or N–H stretching (amines, amides).
- The **peaks at ~2920 and ~2850 cm<sup>-1</sup>** are typical for C–H stretching in aliphatic hydrocarbons (CH<sub>2</sub> and CH<sub>3</sub>).
- The **strong peak at ~1740 cm<sup>-1</sup>** suggests the presence of carbonyl groups, particularly esters or other C=O containing compounds.
- The **~1650 cm<sup>-1</sup>** region can indicate amide I (proteins/peptides) or C=C stretching in alkenes.

- Peaks in the **fingerprint region** (1500–500  $\text{cm}^{-1}$ ) are more specific to particular molecular structures but the above assignments are common.

### **Effect of preprocessing on adulteration prediction in Turmeric powder on base model performance**

In this study, three different types of preprocessing methods which include SG smoothing, SNV and EMSC were applied to correct the spectral scattering caused by the physical properties of the turmeric powder and starch powder mixture. The original and preprocessed datasets  $X_{\text{raw}}$ ,  $X_{\text{sg}}$ ,  $X_{\text{snv}}$ , and  $X_{\text{emsc}}$  were used to train the base models which include LR, DT, SVR, XGBoost and ANN. The base model performances for each dataset have been reported as shown in Table 1

Various studies reported that preprocessing techniques yield different results depending on the spectroscopic technique and adulteration challenge (Valand et al., 2019). For example, in adulteration detection of salted sea cucumber using hyperspectral imaging, SNV and variable sorting for normalization provided better detection result compared to other methods like SG smoothing and MSC (Zhang et al., 2021). Similarly, in olive oil adulteration detection with soybean oil using FTIR, first derivatives outperformed other techniques like MSC, SG smoothing, SNV, and wavelet denoising techniques (Meng et al., 2023).

### **Model Performance Analysis**

In our study, **SNV pre-processed dataset using the ANN model** achieved the highest performance with an  **$R^2$  score of 0.887201**, **RMSE of 1.679279**, and **MAE of 1.225200**, clearly outperforming other combinations. The **Raw dataset with ANN** and **SG-smoothed dataset with ANN** followed, both yielding identical  $R^2$  scores of **-0.504652**, indicating similar predictive performance.

Interestingly, **no preprocessing (Raw dataset)** and **SG smoothing** outperformed **SNV and EMSC pre-processed datasets** in most models, suggesting that excessive preprocessing may lead to information loss. For example, EMSC combined with any model yielded severely negative  $R^2$  values, particularly **EMSC with ANN (-3.312686)** and **EMSC with SVR (-34.796871)**, indicating poor model fit and predictive capability.

The **SVR model with SNV preprocessing** performed the worst among all with an  **$R^2$  of -35.800505**, **RMSE of 30.331729**, and **MAE of 29.919233**, indicating that SNV normalization introduced spectral distortions unsuitable for regression using SVR in this context.

The **XGBoost model**, yielded an RMSE of **14.14** and an  $R^2$  score of **0.31**. This reflects a limited capability of XGBoost to accurately model the relationship between spectral features and adulteration levels in turmeric powder. Compared to the ANN model on SNV-preprocessed data ( $R^2 = 0.887$ ), XGBoost underperformed significantly. This may be due to its reduced effectiveness in capturing the non-linear patterns present in FTIR spectra or insufficient tuning for this application.

## **Conclusion**

These findings clearly indicate the **best-performing datasets** are:

- **SNV pre-processed dataset with ANN**
- **Raw dataset with ANN**
- **SG-smoothed dataset with ANN**

This ranking aligns with the assumption that **Raw (X\_raw)** and **X\_snv** datasets retain more informative spectral features essential for accurate adulteration prediction in Turmeric



powder. On the other hand, **EMSC (X\_emsc)**, due to possible over-correction and signal distortion, demonstrated the **worst performance across all models**.

Hence, datasets O (X\_raw) and X\_snv were selected for **further feature selection and extraction** steps.

### **Selection of key features for adulteration prediction**

The key features (wavenumbers) were selected using the RFE method by finding the feature having lower MSE.

A total of **15 key wavenumbers (in  $\text{cm}^{-1}$ )** were selected from the SNV pre-processed FTIR dataset (X\_snv) using **Recursive Feature Elimination (RFE)**, which identifies the most relevant features contributing to the prediction of starch adulteration in turmeric powder. The selected wavenumbers were:  **$1600.9 \text{ cm}^{-1}$ ,  $1600.10 \text{ cm}^{-1}$ ,  $1600.11 \text{ cm}^{-1}$ ,  $1600.12 \text{ cm}^{-1}$ ,  $1600.13 \text{ cm}^{-1}$ ,  $1600.14 \text{ cm}^{-1}$ ,  $1600.15 \text{ cm}^{-1}$ ,  $1600.16 \text{ cm}^{-1}$ ,  $1600.17 \text{ cm}^{-1}$ ,  $1600.18 \text{ cm}^{-1}$ ,  $1600.19 \text{ cm}^{-1}$ ,  $1600.20 \text{ cm}^{-1}$ ,  $1610 \text{ cm}^{-1}$ ,  $1610.1 \text{ cm}^{-1}$ , and  $1610.2 \text{ cm}^{-1}$** . These closely spaced wavenumbers fall within a narrow region of the mid-infrared spectrum, indicating a localized area where molecular vibrations—possibly associated with starch-specific or turmeric-characteristic functional groups—are most sensitive to changes in sample composition.

The datasets X\_rfe\_raw and X\_rfe\_snv were prepared from selected features on dataset X\_raw and X\_snv using RFE, respectively. These datasets were then used to train the optimized ANN model to reduce the complexity of models while still preserving acceptable results.

### **Feature extraction for adulteration prediction**

Feature extraction transforms the higher dimension of the dataset to a lower dimension while preserving its highest variation (information). **PCA** is a statistical technique that redraws the

correlated features into a new set of uncorrelated features or components, known as **principal components (PCs)**. These PCs are linear combinations of the original features and ordered by the variance explain in the data (Granato et al., 2018).

To identify the most influential spectral regions in differentiating starch adulteration levels in turmeric powder, **Principal Component Analysis (PCA)** was applied to the SNV-pre-processed dataset ( $X_{\text{snv}}$ ). For each of the top six principal components (PCs), the **top five contributing wavenumbers** were extracted based on the magnitude of their loading weights.

For **PC1**, the dominant spectral features were observed around **3840.11  $\text{cm}^{-1}$ , 3840.12  $\text{cm}^{-1}$ , 3840.9  $\text{cm}^{-1}$ , 3840.10  $\text{cm}^{-1}$ , and 3840.8  $\text{cm}^{-1}$** , possibly linked to overtone or combination bands of functional groups. **PC2** highlighted key contributions near **2080.2  $\text{cm}^{-1}$ , 2070.18  $\text{cm}^{-1}$ , 2070.17  $\text{cm}^{-1}$ , 2080.3  $\text{cm}^{-1}$ , and 2080  $\text{cm}^{-1}$** , which may correspond to characteristic vibrations associated with starch-based adulterants. For **PC3**, peaks at **966  $\text{cm}^{-1}$ , 966.1  $\text{cm}^{-1}$ , 3510.16  $\text{cm}^{-1}$ , 3510.17  $\text{cm}^{-1}$ , and 965.1  $\text{cm}^{-1}$**  were prominent, likely related to C–H or O–H bending vibrations.

In **PC4**, wavenumbers clustered around **1460.4  $\text{cm}^{-1}$  to 1460.2  $\text{cm}^{-1}$**  were dominant, potentially signifying aliphatic bending or deformation modes. Meanwhile, **PC5** revealed critical wavenumbers at the lower end of the spectrum: **400.19  $\text{cm}^{-1}$ , 400.18  $\text{cm}^{-1}$ , 400.17  $\text{cm}^{-1}$** , along with higher peaks at **925  $\text{cm}^{-1}$  and 924.1  $\text{cm}^{-1}$** , indicating a combination of lattice vibrations or fingerprint region features. Finally, **PC6** exhibited exceptionally high influence from **400.3  $\text{cm}^{-1}$** , along with strong opposing weights at **400.1  $\text{cm}^{-1}$ , 400.2  $\text{cm}^{-1}$ , 400.4  $\text{cm}^{-1}$ , and 400  $\text{cm}^{-1}$** , reinforcing the relevance of low-wavenumber features in distinguishing subtle compositional differences.

The datasets X\_pca\_raw and X\_pca\_snv have been prepared using the extracted features (PC1, PC2, PC3, PC4, PC5 and PC6) from dataset X\_raw and X\_snv. These datasets were utilized further to train the optimized ANN model.

### **Hyperparameter tuning for model optimization**

The Artificial Neural Network (ANN) model was fine-tuned using a grid search approach via GridSearchCV, optimizing key hyperparameters such as the number of hidden layers, the number of neurons per layer, activation function, and regularization term (alpha). This optimization was applied across six datasets—Raw, SNV, RFE-Raw, RFE-SNV, PCA-Raw, and PCA-SNV—each representing different stages of preprocessing and feature selection. For each dataset, a pipeline integrating feature scaling and the MLPRegressor model was used to ensure consistent data handling. The hyperparameter grid was explored using 3-fold cross-validation, and model performance was evaluated on the test set using  $R^2$ , RMSE, and MAE metrics. The results of optimized models have been mentioned in Table 2 using performance metrics.

It has been found that datasets X\_snv and X\_raw provide better results than other datasets X\_rfe\_raw, X\_rfe\_snv, X\_pca\_raw and X\_pca\_snv.

This demonstrates that adulteration prediction in Turmeric powder can be done without any feature engineering using the ANN model, which can also be seen from the figure

## **Conclusions**

This study underscores the potential of machine learning-assisted FTIR spectroscopy as a powerful tool for predicting starch adulteration in Turmeric powder. By employing various preprocessing techniques and base models, we have demonstrated the approach in accurately predicting adulteration levels. Our findings indicate that the RAW dataset and SNV dataset

consistently outperformed other preprocessing methods, particularly when utilized with ANN models without feature engineering.

Moreover, through feature selection and extraction methods such as RFE and PCA, we were able to identify key spectral features and reduce the dimensionality of the dataset. The identification of key spectral features through feature selection and extraction methods demonstrates the possibility of refining predictive models for more precise and targeted adulteration analysis.

## Limitations

1. **Sample Diversity:** The dataset may not represent the full variability in turmeric powder across different geographical origins, processing methods, or brands, limiting generalizability.
2. **Adulterant Type:** The study focused only on starch as the adulterant. Real-world turmeric adulteration can involve other substances such as metanil yellow, lead chromate, or chalk powder.
3. **Spectral Overlap:** FTIR spectra of turmeric and starch may have overlapping regions, making it difficult to isolate adulteration signals without information loss during preprocessing.
4. **Preprocessing Sensitivity:** The performance of the models was highly dependent on the preprocessing technique used, and some methods (e.g., EMSC) led to significant information loss, reducing prediction accuracy.
5. **Model Interpretability:** Advanced models like ANN and SVR, although accurate, function as black boxes and offer limited insight into which spectral features are most influential.

6. **Limited Hyperparameter Optimization:** Grid search, while effective, explores a limited hyperparameter space compared to Bayesian or evolutionary approaches.

## Recommendations

1. **Incorporate More Adulterants:** Future studies should expand to include multiple adulterants and mixtures to build a more robust and comprehensive detection model.
2. **Use of Advanced Optimization:** Employ Bayesian optimization or genetic algorithms for deeper and more efficient hyperparameter tuning of ANN and other models.
3. **Spectral Fusion Techniques:** Combine FTIR data with data from other spectroscopic methods (e.g., NIR, Raman) to improve classification and regression accuracy.
4. **Model Explainability:** Integrate techniques such as SHAP (SHapley Additive exPlanations) or LIME to improve the interpretability of complex models like ANN and SVR.
5. **Larger and More Diverse Dataset:** Expand the dataset by including samples from multiple sources, seasons, and processing techniques to improve model robustness.

6. **Deployable Application:** Develop a user-friendly interface or mobile application to apply the trained model in real-time food quality monitoring systems

## Future Directions

1. **Multi-Adulterant Detection:** Future research can extend the methodology to detect and quantify a range of common adulterants in turmeric, such as metanil yellow, chalk powder, and lead chromate, using multi-label or multi-output machine learning models.
2. **Integration of Deep Learning:** Employ convolutional neural networks (CNNs) or recurrent neural networks (RNNs) directly on FTIR spectral data to capture complex patterns and enhance accuracy.
3. **Portable Device Compatibility:** Integrate the model into portable FTIR instruments or smartphone-connected spectrometers for real-time, on-site adulteration detection.
4. **Cross-Sample Transferability:** Explore domain adaptation techniques to transfer models trained on one type of turmeric (e.g., regional or seasonal variants) to another with minimal re-training.
5. **Time-Series and Stability Studies:** Study how adulteration detection accuracy varies with storage time, temperature, and humidity to make the model robust for industrial deployment.

## Potential Usage and Applications

1. **Quality Control in Food Industry:** Food processing units and spice manufacturers can integrate the model into their QA/QC workflows to ensure product authenticity before distribution.
2. **Regulatory and Inspection Agencies:** Government food safety bodies can use this approach for rapid screening of turmeric samples for adulteration during market inspections.
3. **Consumer Empowerment Tools:** A simplified version of the model can be embedded into consumer apps to allow everyday users to test turmeric samples using handheld FTIR devices.
4. **Export and Import Authentication:** Helps in verifying the purity of turmeric in international trade, ensuring compliance with import-export food safety regulations.
5. **Research and Academic Use:** Can serve as a template methodology for future studies on food adulteration detection using spectroscopy and machine learning.

## References

1. Al-Awadhi, M. A., & Deshmukh, R. R. (2021). Detection of Adulteration in Coconut Milk using Infrared Spectroscopy and Machine Learning 2021 International Conference of Modern Trends in Information and Communication Technology Industry (MTICTI) <https://ieeexplore.ieee.org/document/9664764>
2. Calle, J. L. P., Barea-Sepúlveda, M., Ruiz-Rodríguez, A., Álvarez, J. Á., Ferreiro-González, M., & Palma, M. (2022). Rapid Detection and Quantification of Adulterants in Fruit Juices Using Machine Learning Tools and Spectroscopy Data. *Sensors*, 22(10). <https://doi.org/10.3390/s22103852>

3. Choudhary, A., Gupta, N., Hameed, F., & Choton, S. (2020). An overview of food adulteration: Concept, sources, impact, challenges and detection. *International Journal of Chemical Studies*, 8(1), 2564-2573.  
<https://doi.org/10.22271/chemi.2020.v8.i1am.8655>
4. Chung, T., Tam, I. Y. S., Lam, N. Y. Y., Yang, Y., Liu, B., He, B., Li, W., Xu, J., Yang, Z., Zhang, L., Cao, J. N., & Lau, L.-T. (2022). Non-targeted detection of food adulteration using an ensemble machine-learning model. *Scientific Reports*, 12(1). <https://doi.org/10.1038/s41598-022-25452-3>
5. Dutta, A. (2017). Fourier Transform Infrared Spectroscopy. In *Spectroscopic Methods for Nanomaterials Characterization* (pp. 73-93).  
<https://doi.org/10.1016/b978-0-323-46140-5.00004-2>
6. Gorde, P. M., Dash, D. R., Singh, S. K., & Singha, P. (2024). Advancements in sustainable food packaging: A comprehensive review on utilization of nanomaterials, machine learning and deep learning. *Sustainable Chemistry and Pharmacy*, 39. <https://doi.org/10.1016/j.scp.2024.101619>
7. Goyal, R., Singha, P., & Singh, S. K. (2024). Spectroscopic food adulteration detection using machine learning: Current challenges and future prospects. *Trends in Food Science & Technology*, 146. <https://doi.org/10.1016/j.tifs.2024.104377>
8. Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L., & Maggio, R. M. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends in Food Science & Technology*, 72, 83-90. <https://doi.org/10.1016/j.tifs.2017.12.006>
9. Lia, F., Morote Castellano, A., Zammit-Mangion, M., & Farrugia, C. (2018). Application of fluorescence spectroscopy and chemometric models for the



detection of vegetable oil adulterants in Maltese virgin olive oils. *Journal of Food Science and Technology*, 55(6), 2143-2151. <https://doi.org/10.1007/s13197-018-3131-0>

10. Lohumi, S., Lee, S., Lee, H., & Cho, B.-K. (2015). A review of vibrational spectroscopic techniques for the detection of food authenticity and adulteration. *Trends in Food Science & Technology*, 46(1), 85-98. <https://doi.org/10.1016/j.tifs.2015.08.003>
11. Mandal, S., & Mandal, M. (2015). Coriander (*Coriandrum sativum* L.) essential oil: Chemistry and biological activity. *Asian Pacific Journal of Tropical Biomedicine*, 5(6), 421-428. <https://doi.org/10.1016/j.apjtb.2015.04.001>
12. Martens, H., & Stark, E. (1991). Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, 9(8), 625-635. [https://doi.org/10.1016/0731-7085\(91\)80188-f](https://doi.org/10.1016/0731-7085(91)80188-f)
13. Meng, X., Yin, C., Yuan, L., Zhang, Y., Ju, Y., Xin, K., Chen, W., Lv, K., & Hu, L. (2023). Rapid detection of adulteration of olive oil with soybean oil combined with chemometrics by Fourier transform infrared, visible-near-infrared and excitation-emission matrix fluorescence spectroscopy: A comparative study. *Food Chemistry*, 405. <https://doi.org/10.1016/j.foodchem.2022.134828>
14. Nallan Chakravartula, S. S., Moschetti, R., Bedini, G., Nardella, M., & Massantini, R. (2022). Use of convolutional neural network (CNN) combined with FT-NIR spectroscopy to predict food adulteration: A case study on coffee. *Food Control*, 135. <https://doi.org/10.1016/j.foodcont.2022.108816>

15. Othman, S., Mavani, N. R., Hussain, M. A., Rahman, N. A., & Mohd Ali, J.

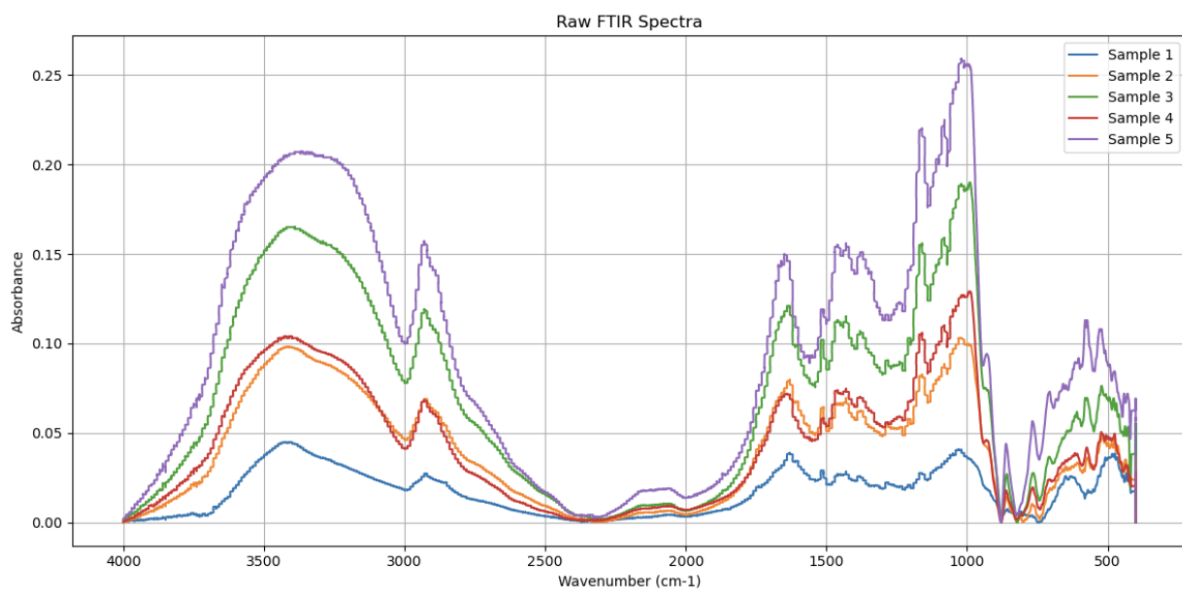
(2023). Artificial intelligence-based techniques for adulteration and defect

detections in food and agricultural industry: A review. Journal of Agriculture and

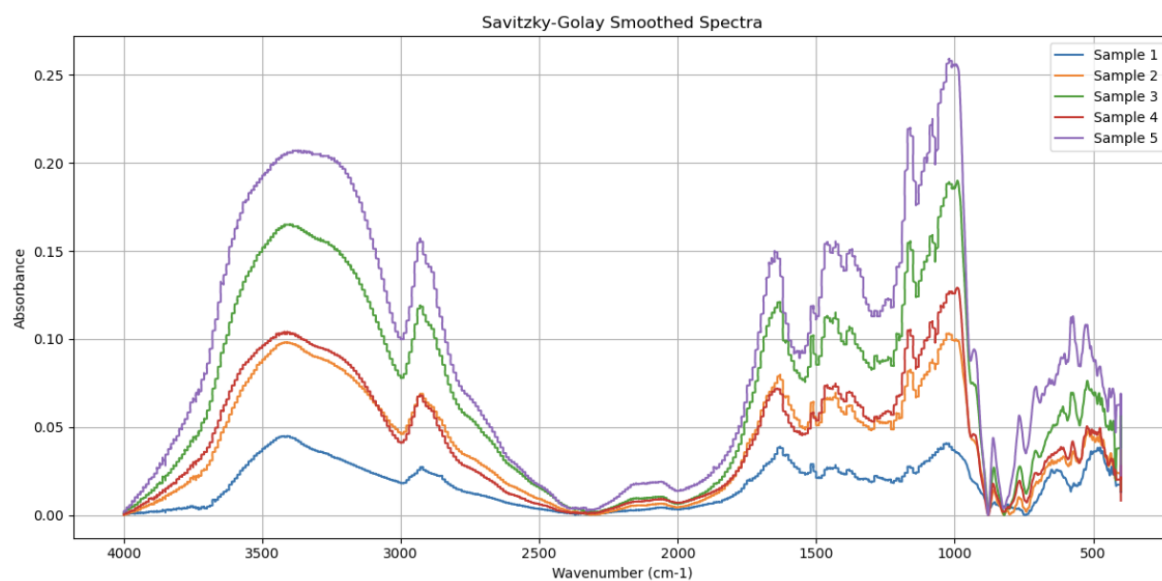
Food Research, 12. <https://doi.org/10.1016/j.jafr.2023.100590>

## Tables and Plots

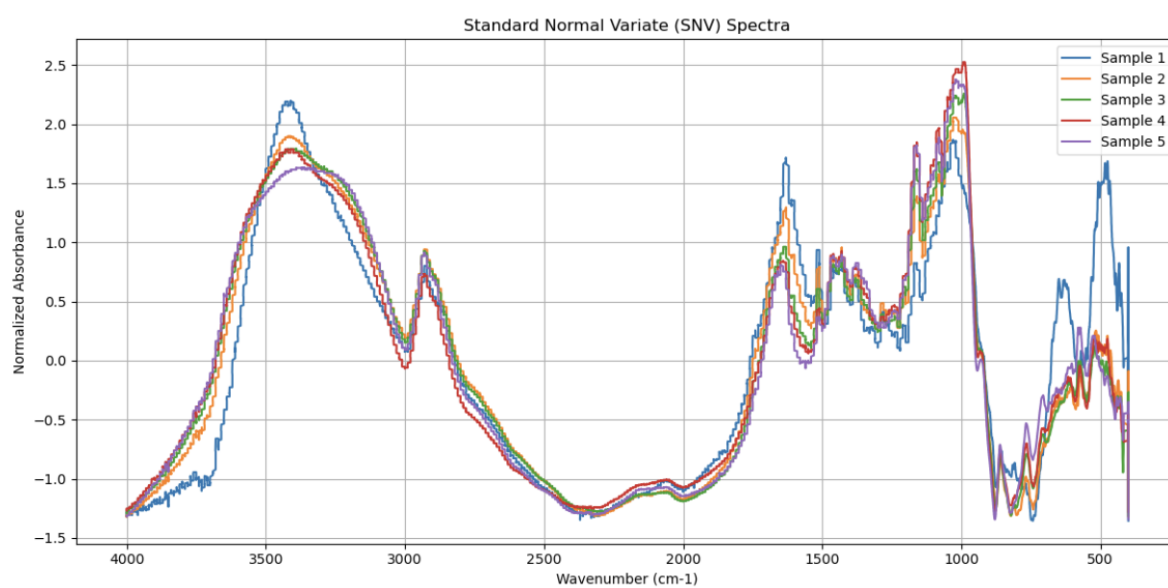
FTIR Spectra of Turmeric powder sample adulterated with different starch level (% w/w).



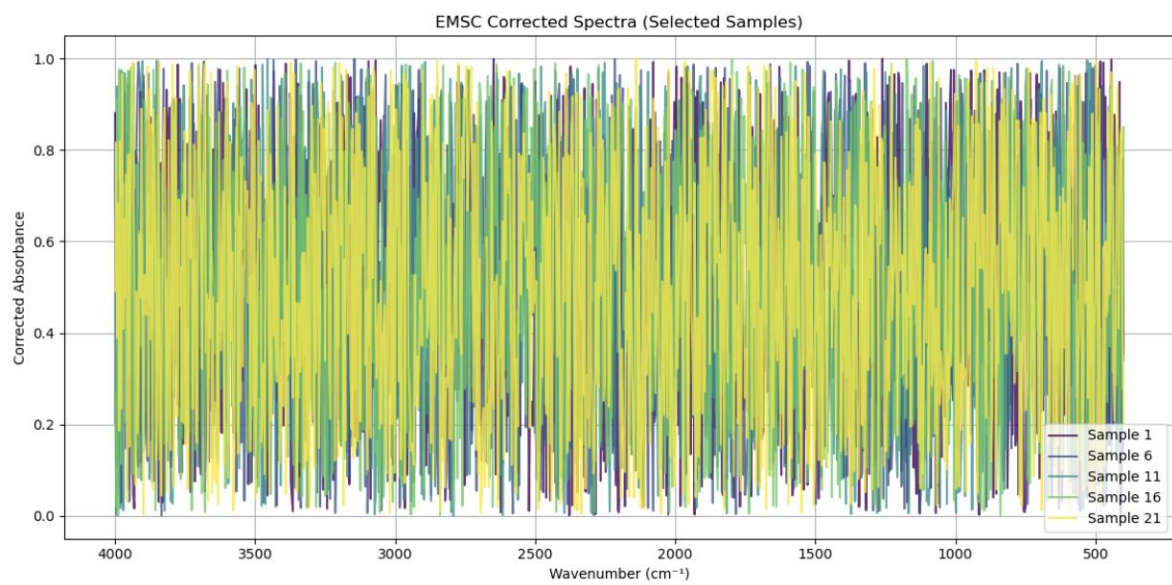
SG Smoothed FTIR Spectra of Turmeric powder sample adulterated with different starch level (% w/w).



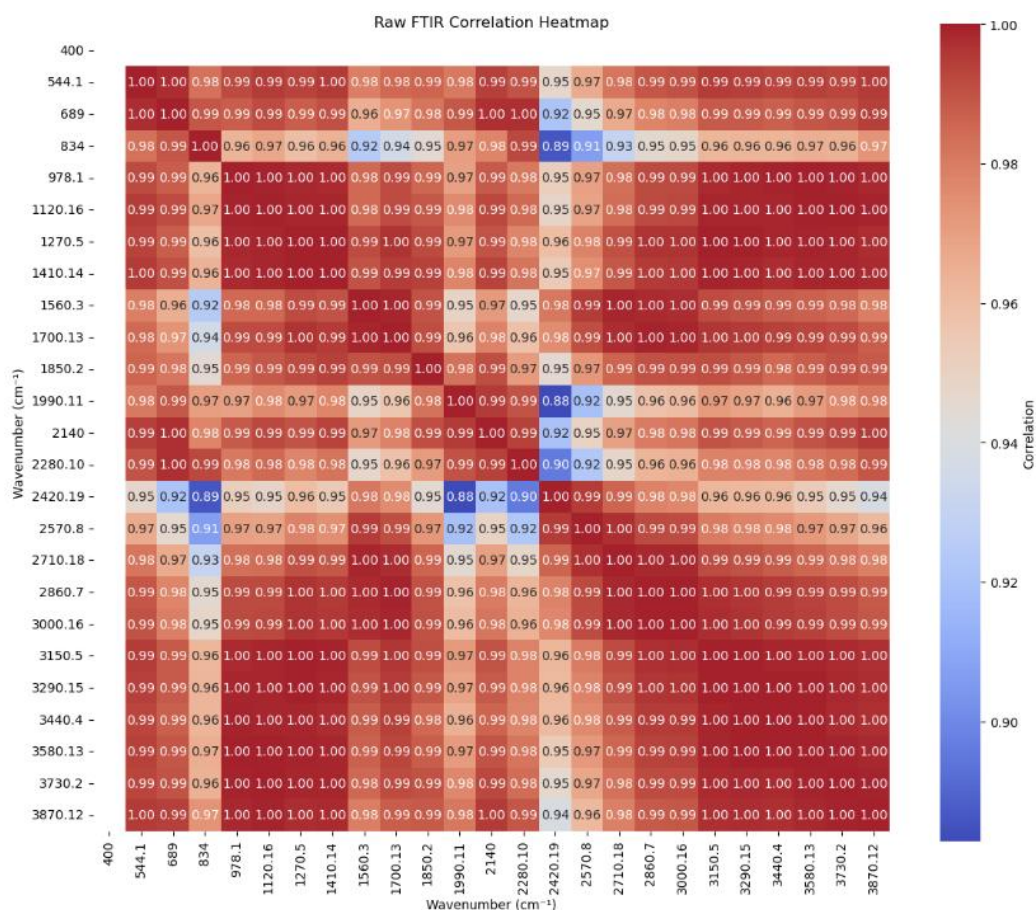
SNV FTIR Spectra of Turmeric powder sample adulterated with different starch level (% w/w).



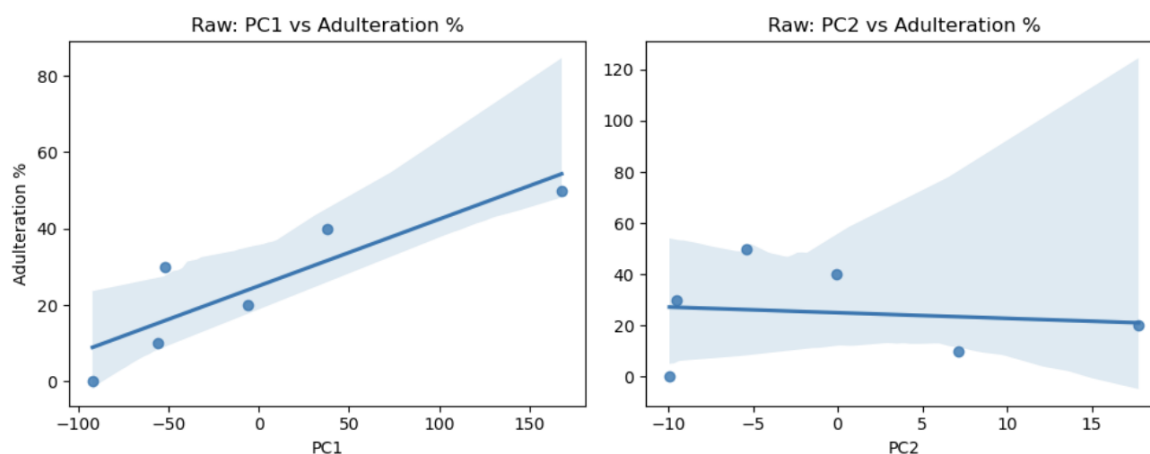
EMSC FTIR Spectra of Turmeric powder sample adulterated with different starch level (% w/w).



Correlation heat map for selected features for Original dataset X\_raw



Regression plot between each selected principal components (PC1 and PC2) and target adulteration level for Original Dataset X\_raw



Regression plot between each selected principal components (PC1 and PC2) and target adulteration level for SNV Dataset

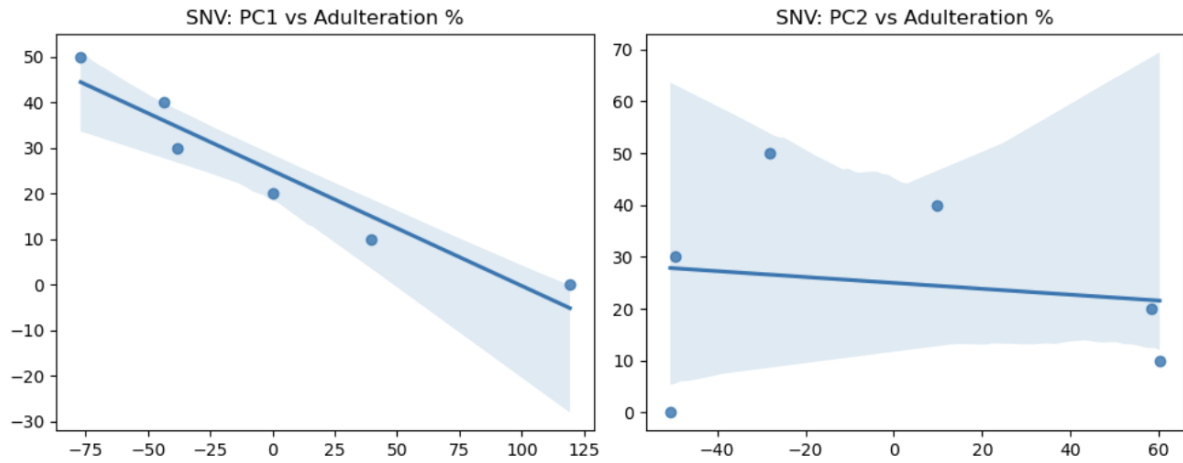


Table 1. Base models performance of different datasets for adulteration prediction.

All Model Performances:

	Dataset	Model	R2 Score	RMSE	MAE
11	SNV	ANN	0.887201	1.679279	1.225200
3	Raw	ANN	-0.504652	6.133212	6.105248
7	SG	ANN	-0.504682	6.133274	6.105315
8	SNV	LR	-0.931808	6.949475	6.527058
15	EMSC	ANN	-3.312686	10.383504	10.340999
9	SNV	DT	-9.000000	15.811388	15.000000
12	EMSC	LR	-11.724220	17.835512	16.645530
0	Raw	LR	-15.849238	20.523912	19.027355
4	SG	LR	-15.907352	20.559275	19.056946
1	Raw	DT	-25.000000	25.495098	25.000000
5	SG	DT	-25.000000	25.495098	25.000000
13	EMSC	DT	-25.000000	25.495098	25.000000
14	EMSC	SVR	-34.796871	29.915243	29.463931
6	SG	SVR	-34.978608	29.991085	29.564042
2	Raw	SVR	-34.978617	29.991089	29.564047
10	SNV	SVR	-35.800550	30.331729	29.919233

Top 2 Datasets:

Dataset

SNV 0.887201

Raw -0.504652

Name: R2 Score, dtype: float64

Best Performing Model: ANN on SNV Dataset

Actual vs Predicted:

Actual: 0.00, Predicted: 2.37

Actual: 10.00, Predicted: 9.92

Table 2. Optimized ANN models performance of different datasets for adulteration prediction

Final Fine-Tuning Results Across 6 Datasets:			
	Dataset	Best Params	R2 Score \
3	RFE-SNV	{'mlp__activation': 'relu', 'mlp__alpha': 0.00...	-3.474121
5	PCA-SNV	{'mlp__activation': 'relu', 'mlp__alpha': 0.00...	-16.398610
2	RFE-Raw	{'mlp__activation': 'relu', 'mlp__alpha': 0.00...	-38.344462
0	Raw	{'mlp__activation': 'relu', 'mlp__alpha': 0.00...	-54.027830
4	PCA-Raw	{'mlp__activation': 'relu', 'mlp__alpha': 0.00...	-61.251961
1	SNV	{'mlp__activation': 'relu', 'mlp__alpha': 0.00...	-145.487215
	RMSE	MAE	
3	10.576059	9.643301	
5	20.855821	17.462103	
2	31.362582	30.920700	
0	37.090373	35.701962	
4	39.449956	39.449654	
1	60.515952	54.254640	