# Prediction of Wood Powder Adulteration Level in Cilantro Powder

Rishabh Goyal, Zirun Ye

## 1 Abstract

Adulteration in food products poses significant health and economic risks globally. In particular, the adulteration of cilantro powder with wood powder is a common issue that compromises product quality. This project explores the application of artificial intelligence (AI) to detect and predict wood powder adulteration in cilantro powder using spectral data. We evaluated several machine learning (ML) and deep learning (DL) models, including Linear Regression (LR), Decision Tree (DT), Support Vector Regression (SVR), Artificial Neural Networks (ANN), one-dimensional Convolutional Neural Networks (1d-CNN), and Deep Neural Decision Trees (DNDT), to determine the most accurate and efficient approach.

Our methodology involved preprocessing spectral data, training models, optimizing performance, and evaluating each model using classification metrics. The models were compared based on their accuracy, ability to handle noise, and efficiency in processing. Findings indicate that DL-based approaches, particularly ANN and 1d-CNN, show superior performance in predicting adulteration levels. This work demonstrates the potential of AI in improving food quality assurance and provides a foundation for further research in automated adulteration detection.

## 2 Problem Definition

Food adulteration is a persistent and growing global issue affecting trust in the global food supply chain. Among the many adulterants, wood powder is commonly used to tamper with powdered spices due to its similar texture and appearance. Cilantro powder, which is frequently used in culinary applications, is often targeted for such adulteration. Detecting and quantifying this adulteration through traditional methods can be time-consuming, expensive, and not scalable.

### 2.1. Problem Statement

This work addresses the challenge of accurately predicting the level of wood powder adulteration in cilantro powder using AI-based models. The goal is to design and evaluate a model that can analyze spectral data and determine adulteration levels with high precision and reliability.

### 2.2. Importance

Solving this problem can improve food safety, reduce economic fraud, and assist regulatory agencies and manufacturers in ensuring product authenticity. The use of models in this context provides opportunities for scalable quality control solutions.

### 2.3. Constraints

The project uses FTIR spectroscopy data collected under controlled conditions, and it focuses only on cilantro powder adulterated with wood powder. Other adulterants are also added in the cilantro powder, but we are only focusing on wood powder.

# 3 Thought Process & Approach

The project began by researching real-world problems where AI could make a positive impact. Food adulteration appeared as a critical issue due to its widespread nature in global trade and the serious consequences it poses to health and the economy. After reviewing news articles, we learned that powdered spices, particularly cilantro powder, are frequently adulterated with wood powder, making them an ideal candidate for our project.

We brainstormed various ways to detect adulteration and found that spectral data analysis, commonly used in quality assurance and research for food products, provided a non-destructive and scalable method for detecting adulterants. From there, we set our objective to build a model capable of predicting wood powder adulteration levels in Cilantro powder using FTIR spectroscopy data input by collaborating with National Institute of Technology, India for conducting the experiment and provideing us the necessary data for applying the AI models.

Throughout the development process, multiple base models were used to determine which approach was most effective. Started with traditional ML models such as Linear Regression, Decision Trees, and Support Vector Regression. As we progressed, we realized that deeper, more complex models, like Artificial Neural Networks (ANN), one-dimensional Convolutional Neural Networks, and Deep Neural Decision Trees, provided greater potential for capturing complex patterns in spectral data.

SG smoothing was applied to spectral signals before implementing any model to minimize inconsequential noise and ensure consistent inputs across all models. Other design decisions, such as feature selection and extraction, were made after several iterations. Performance evaluation was used to refine approaches and narrow down the most promising techniques.

# 4 Solution Description

The project is designed to predict the level of wood powder adulteration in cilantro powder using 1-d spectral data as input. The solution consists of the following key components:

### 4.1. Data Acquisition & Preprocessing

The dataset used to train the models contains the FTIR spectroscopy data for different levels of wood powder in cilantro powder in different replication sizes. It takes light intensities as features and analyzes their relationship with the adulteration level. The noisy data was removed using the Z-score outlier method.

Preprocessing steps included baseline correction, Savitzky-Golay (SG) smoothing, moving average smoothing, multiple scattering correction (MSC), extended multiplicative signal correction (EMSC), stan-

dard normal variate (SNV), standardization, and min-max Normalization to ensure consistency and model readiness.

### 4.2. Model Selection & Training

Each model was trained using all preprocessed spectral datasets to classify adulteration levels after implementation.

4.2.1. Linear Regression:

Linear regression establishes a linear relationship between target output and input features by finding the best-fitting straight line to minimize the loss function. This is achieved using negative mean squared error (MSE) as the loss function. Using negative MSE is consistent with standard optimization practices and yields better model fit.

4.2.2. Decision Tree:

Decision Tree split data to reduce variance and predict continuous variables. The model generates rules that are easy to interpret, which help in explaining the interaction between input features and the target variable. Negative MSE was used as the evaluation metric.

4.2.3. Support Vector Regression:

SVR maps input features to a high-dimensional space through a kernel function to find a non-linear relationship. It finds a hyperplane that best fits the data with a margin from the hyperplane to the nearest points. It is highly applicable on high-dimensional data like spectral features. Radial basis function kernel was used in this study.

4.2.4. Artificial Neural Network:

ANNs consist of connected layers of nodes called neurons. The layers include an input layer, hidden layers, and an output layer. The ANN in this study consisted of an input layer comprising 64 neurons, a hidden layer comprising 32 neurons, and an output layer comprising 1 neuron. ReLU activation function was applied to the input and hidden layers, and a linear activation function was applied to the output layer. The Adam optimizer with MSE loss was applied for optimization and early stopping was applied to prevent overfitting.

4.2.5. One Dimensional Convolutional Neural Network:

1d-CNN are designed to automatically learn patterns from one-dimensional sequential data, such as sensor readings or time series[2]. In this study, the 1d-CNN model included a convolutional layer with 16 to 128 filters and a kernel size ranging from 2 to 10, followed by a max pooling layer with a pool size between 2 and 5. After the convolutional and pooling layers, a dense layer with 32 to 256 units was added, followed by an output layer with 1 neuron. The rectified linear unit (ReLU) activation function was applied to the convolutional and dense layers, and a linear activation function was used for the output layer. The model was optimized using the Adam optimizer with a learning rate tuned between 0.001 and 0.01.

4.2.6. Deep Neural Decision Tree:

The Differentiable Neural Decision Tree (DNDT) is a neural network model that mimics the behavior of traditional decision trees[1]. In this study, the DNDT model consisted of a single DNDT layer, where the number of cut points per feature ranged from 1 to 4 and the temperature parameter varied between 1.0

and 20.0. Leaf scores and cut points were initialized and constrained within a specified range to stabilize learning. The model was optimized using the Adam optimizer with a learning rate tuned between 0.001 and 0.01, and the mean squared error (MSE) loss function was employed.

### 4.3. Optimization & comparison

Feature selection methods, like recursive feature elimination and random forest importance, were applied to the proposed datasets to identify the key features for models to predict the adulteration levels. Feature extraction methods, such as PCA and LDA, are utilized to see the impact of each feature and capture the main features that affect the accuracy of the result.

The hyperparameters (e.g., learning rate, depth of trees, kernel functions, number of hidden layers, dataset) were tuned and compared models to identify the best-performing approach.

## 5 Evaluation & Insights

Datasets were trained and tested on six different ML and DL models. The models were implemented using Python, with support from ML libraries such as Scikit-learn, TensorFlow, and PyTorch.

The results showed a clear progression in performance from simpler models to more complex ones. The simpler models, such as linear regression, decision trees, and support vector regression, while computationally simple, have lower accuracy and some showed a high mean squared error, indicating their inability to capture the nonlinear relationships in the data.

### 5.1. Dataset

We used a dataset consisting of one dimensional spectroscopy data from cilantro powder samples. Each sample was labeled based on the percentage of wood powder adulteration (e.g., 0%, 1%, 2%, ..., 20%, 30%, ..., 100%). The dataset was balanced to ensure fair representation across all adulteration levels.

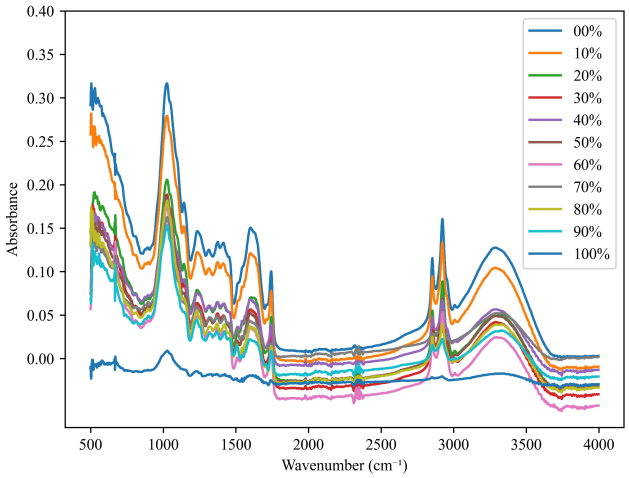


Figure 1: Wood powder level from 0 to 100% with 10% interval

Table 1: Performance comparison of different models across datasets for wood adulteration prediction in Cilantro powder

| Model | Dataset | R-squared | | RMSE | | MAE | |
|---|---|---|---|---|---|---|---|
| | | Validation | Testing | Validation | Testing | Validation | Testing |
| Logistic Regression | O | $-2.3162 \times 10^{18}$ | 0.3586 | $3.6845 \times 10^{8}$ | 0.1199 | $1.6107 \times 10^{8}$ | 0.1006 |
| | P1 | $-1.5235 \times 10^{18}$ | 0.4236 | $3.0172 \times 10^{8}$ | 0.1189 | $1.6589 \times 10^{8}$ | 0.0986 |
| | P2 | $-7.1075$ | 0.4223 | $6.4421 \times 10^{8}$ | 0.1260 | $3.5446 \times 10^{8}$ | $-0.1000$ |
| | P3 | $-236.0489$ | $-6.2386$ | 3.0791 | 0.3126 | 0.6067 | 0.2292 |
| Decision Tree | O | 0.8370 | 0.6137 | 0.1018 | 0.1371 | 0.0636 | 0.0811 |
| | P1 | 0.8215 | 0.6090 | 0.1004 | 0.1366 | 0.0629 | 0.0796 |
| | P2 | 0.8370 | 0.6137 | 0.1018 | 0.1371 | 0.0636 | 0.0811 |
| | P3 | 0.6351 | $-1.3657$ | 0.1686 | 0.1319 | 0.2111 | 0.0891 |
| SVR | O | 0.8289 | 0.6484 | 0.1113 | 0.1298 | 0.0794 | 0.1002 |
| | P1 | 0.8288 | 0.6489 | 0.1114 | 0.1298 | 0.0795 | 0.1002 |
| | P2 | 0.8222 | 0.4108 | 0.1129 | 0.1319 | 0.0835 | 0.1024 |
| | P3 | $-0.0671$ | $-1.1324$ | 0.3039 | 0.3288 | 0.2028 | 0.2331 |
| ANN | O | 0.9351 | 0.8852 | 0.0753 | 0.0853 | 0.0588 | 0.0673 |
| | P1 | 0.9350 | 0.8790 | 0.0756 | 0.0880 | 0.0590 | 0.0676 |
| | P2 | 0.6419 | 0.4678 | 0.1651 | 0.1728 | 0.1233 | 0.1339 |
| | P3 | $-2007.7344$ | $-6342.9477$ | 9.3222 | 14.0172 | 2.9833 | 5.1846 |
| 1-d CNN | O | 0.9650 | 0.9390 | 0.0559 | 0.0765 | 0.0381 | 0.0620 |
| | P | 0.9640 | 0.9585 | 0.0567 | 0.0631 | 0.0397 | 0.0460 |
| | OFS | 0.9234 | 0.9138 | 0.0827 | 0.0910 | 0.0562 | 0.0623 |
| | PFS | 0.8844 | 0.8989 | 0.1016 | 0.0986 | 0.0664 | 0.0695 |
| | OFF | $-0.0392$ | 0.7861 | 0.3049 | 0.1434 | 0.1965 | 0.0903 |
| | PFE | 0.8210 | 0.6828 | 0.1265 | 0.1747 | 0.0847 | 0.1141 |
| DNDT | O | 0.8554 | 0.8557 | 0.1137 | 0.1178 | 0.0769 | 0.0861 |
| | P | 0.8345 | 0.8521 | 0.1216 | 0.1193 | 0.0837 | 0.0876 |
| | OFS | 0.7847 | 0.8289 | 0.1387 | 0.1283 | 0.0978 | 0.0937 |
| | PFS | 0.8099 | 0.8340 | 0.1304 | 0.1264 | 0.0930 | 0.0916 |
| | OFF | 0.5671 | 0.5997 | 0.1967 | 0.1963 | 0.1395 | 0.1379 |
| | PFE | 0.5950 | 0.5701 | 0.1903 | 0.2034 | 0.1330 | 0.1529 |

[1]O: Original dataset; P: Preprocessed; OFS: Original+Feature Selection; PFS: Preprocessed+Feature Selection; OFF: Original+Feature Extraction; PFE: Preprocessed+Feature Extraction; P1: SG smoothing; P2: SNV; P3: EMSC; RMSE: Root mean square error; MAE: Mean absolute error; SVR: Support Vector Regression; ANN: Artificial Neural Network; CNN: Convolusional Neural Network; DNDT: Deep Neural Decision Tree.

### 5.2. Key Insights

- Results showed that ANN and 1-d CNN outperformed other models, indicating the strength of DL in capturing intricate spectral features.
- Overfitting was a challenge in the ANN, but mitigated via validation split, early stopping, and Bayesian optimization.
- Feature selection and extraction yielded inferior results compared to using the original and preprocessed datasets.
- Despite the higher dimensionality, the best-performing model required fewer hidden layers in this work.
- The DNDT produced inferior results compared to both the ANN and the 1-d CNN in this study.
- Imbalanced sets will lead to biased results.

# 6 Discussion

Our solution demonstrated several notable strengths. The most significant was the high classification accuracy achieved by the DL models, particularly the Artificial Neural Network (ANN) and the one-dimensional Convolutional Neural Network (1d-CNN), both of which exceeded 90% accuracy. These models also showed strong robustness to noise in the spectral input data, which is a common issue in real-world signal collection. In addition, the overall framework we developed is scalable; once trained, the models can quickly classify new samples, making this approach suitable for practical applications in quality control and food safety. Another strength of the project lies in the diversity of models tested, which gave us insight into the relative advantages and trade-offs of traditional ML versus DL techniques.

Surprisingly, the Deep Neural Decision Tree (DNDT) yielded inferior results compared to both the Artificial Neural Network (ANN) and the 1D Convolutional Neural Network (1D CNN). This performance gap may be attributed to overfitting, a common issue in decision tree-based models. Future work could focus on addressing this limitation through techniques such as regularization and dropout to enhance the model's performance.

In exploring alternative approaches, we considered techniques such as Principal Component Analysis (PCA) for dimensionality reduction. However, we ultimately decided to retain the full spectral input to preserve as much useful information as possible for better predictions.

# 7 Conclusion

In this project, we developed and evaluated a system for predicting wood powder adulteration in cilantro powder using spectral data and AI models. Our primary goal was to determine which model could most accurately predict adulteration levels based on FTIR spectroscopy data input. Through a thorough comparative analysis of six different models – Linear Regression, Decision Tree, Support Vector Regression, Artificial Neural Network, 1-d Convolutional Neural Network, and Deep Neural Decision Tree – we found that DL techniques, particularly ANN and 1-d CNN, provided the highest levels of accuracy and robustness.

Our investigation highlighted the importance of preprocessing, model selection, and hyperparameter tuning in working with spectral data. We also gained valuable insights into the challenges of building models that balance performance with generalizability. Despite the complexity of the models, the results were promising and suggest that AI can be a powerful tool in tackling food adulteration.

Looking ahead, future improvements could involve enhancing the implementation of the DNDT model, expanding the dataset to include more samples from diverse environments, and evaluating the model's generalization across various food products.

# References


[1] Aston Zhang et al. "Dive into Deep Learning". In: *arXiv preprint arXiv:1806.06988* (2018). URL: `https://arxiv.org/pdf/1806.06988.pdf` (visited on 03/20/2024).

[2] Aston Zhang et al. *Dive into Deep Learning: Convolutional Neural Networks*. 2023. URL: `https://d2l.ai/chapter_convolutional-neural-networks/index.html` (visited on 03/20/2024).