# Mycotoxin Prediction Pipeline: Technical Report

Saketha Rama
email:**sakethram9999@gmail.com**, Ph:**+91 7989173987**

March 16, 2025

**Abstract**

This report presents a comprehensive analysis of the mycotoxin prediction pipeline developed to predict deoxynivalenol (DON) concentration in corn samples using hyperspectral imaging data. The pipeline encompasses data preprocessing, dimensionality reduction, model selection, training, evaluation, and interpretability analysis. The report details the methodologies employed, challenges encountered, and results achieved, along with recommendations for future improvements.

## 1 Introduction

Mycotoxins, particularly deoxynivalenol (DON, also known as vomitoxin), are secondary metabolites produced by fungi that can contaminate agricultural products and pose significant health risks to humans and animals. Traditional methods for detecting mycotoxin levels are time-consuming and expensive. This project aims to develop a machine learning pipeline to predict DON concentration in corn samples using hyperspectral imaging (HSI) data, providing a faster and more cost-effective alternative for mycotoxin screening.

## 2 Data Overview

The dataset consists of 500 corn samples with 448 spectral features derived from hyperspectral imaging. Each sample is labeled with its corresponding DON concentration (vomitoxin_ppb). The dataset exhibits the following characteristics:

- 500 samples with 448 spectral features and 1 target variable (vomitoxin_ppb)
- DON concentration ranges from 0 to 131,000 ppb
- Highly skewed distribution of DON values (mean: 3,410 ppb, median: 500 ppb)
- Presence of duplicate sample IDs (hsi_id) for some corn samples

## 3 Preprocessing Steps

### 3.1 Data Cleaning

Initial data exploration revealed several issues that required attention:

- **Duplicate Samples**: Three samples (imagoai_corn_395, imagoai_corn_385, and imagoai_corn_443) had duplicate entries. These duplicates were identified but retained for analysis as they might represent different measurements of the same sample.
- **Missing Values**: No missing values were detected in the dataset, eliminating the need for imputation strategies.
- **Outlier Analysis**: Using the Interquartile Range (IQR) method, 80 samples were identified as potential outliers in terms of DON concentration. However, following research by Aguinis et al. (2013) and studies on DON contamination in corn, these extreme values were considered scientifically significant rather than errors. Instead of removing outliers, a log transformation approach was adopted.

### 3.2 Feature Transformation

- **Log Transformation**: Due to the highly skewed distribution of DON concentration values, a log transformation (log1p) was applied to normalize the target variable. This transformation significantly improved the distribution's normality and made it more suitable for modeling.
- **Feature Scaling**: The spectral features were standardized using z-score normalization to ensure all features contributed equally to the model training process.

## 4 Dimensionality Reduction Insights

The hyperspectral data contained 448 spectral features, presenting a high-dimensional space that could lead to overfitting and computational inefficiency. SHAP (SHapley Additive exPlanations) analysis was used to identify the most important features contributing to the prediction of DON concentration.

Key insights from dimensionality reduction:

- The top 10 most important features (59, 70, 23, 215, 352, 170, 21, 342, 194, 285) accounted for a significant portion of the predictive power.
- Feature 59 had the highest importance score (0.098), followed by feature 70 (0.094) and feature 23 (0.071).
- Many features had minimal or zero importance, indicating redundancy in the spectral data.
- The importance scores followed a power-law distribution, with a small subset of features contributing most of the predictive information.

This analysis suggests that future models could potentially use a reduced feature set without significant loss in predictive performance.

## 5 Model Selection

Multiple modeling approaches were evaluated to determine the most effective method for predicting DON concentration:

### 5.1 Traditional Machine Learning Models

Several traditional machine learning models were initially explored, including:

- Linear Regression
- Random Forest
- Gradient Boosting
- Support Vector Regression

These models provided a baseline for performance comparison but were ultimately outperformed by neural network approaches.

### 5.2 Deep Learning Approach

A specialized neural network architecture was developed for this task:

- **DirectNonLinearNN**: A simple yet effective architecture consisting of a single linear layer followed by a sigmoid activation scaled to match the output range of log-transformed DON values.
- The model was designed to capture direct relationships between spectral features and DON concentration while maintaining interpretability.
- The architecture included weight initialization using Xavier uniform initialization to ensure stable training.

- The output layer used a sigmoid activation function scaled to a range of approximately 0-12 (matching the log-transformed DON values).

The neural network approach was selected as the final model due to its superior performance and ability to capture complex non-linear relationships in the data.

# 6 Training Process

The training process was carefully designed to ensure robust model performance:

- **Data Splitting**: The dataset was split into training (70%), validation (15%), and test (15%) sets using stratified sampling based on DON concentration ranges.
- **Loss Function**: Mean Squared Error (MSE) was used as the loss function to optimize the model.
- **Optimizer**: Adam optimizer with a learning rate of 0.001 and weight decay of 1e-4 was employed to minimize the loss function.
- **Learning Rate Scheduling**: ReduceLROnPlateau scheduler was implemented to reduce the learning rate when validation loss plateaued, with a factor of 0.5 and patience of 10 epochs.
- **Early Stopping**: To prevent overfitting, early stopping with a patience of 20 epochs was implemented, monitoring validation loss.
- **Batch Size**: A batch size of 32 was used for training, providing a good balance between computational efficiency and model convergence.

The training process showed steady convergence, with both training and validation losses decreasing consistently over epochs.

# 7 Evaluation Metrics

The model was evaluated using multiple metrics to provide a comprehensive assessment of its performance:

## 7.1 Log Scale Performance

- $R^2$ Score: 0.1374
- Root Mean Squared Error (RMSE): 2.6463
- Mean Absolute Error (MAE): 2.0546

## 7.2 Original Scale Performance (after inverse transformation)

- $R^2$ Score: 0.7943
- Root Mean Squared Error (RMSE): 7583.01
- Mean Absolute Error (MAE): 2577.99

The significant difference between log scale and original scale metrics highlights the impact of the log transformation. While the $R^2$ score on the log scale appears modest, the model demonstrates strong predictive performance when evaluated on the original scale, explaining approximately 79% of the variance in DON concentration.

# 8 Interpretability Analysis

SHAP (SHapley Additive exPlanations) analysis was employed to interpret the model's predictions and understand feature importance:

### 8.1 Feature Importance

The SHAP analysis revealed that:

- Features 59, 70, and 23 were the most influential in predicting DON concentration, with importance scores of 0.098, 0.094, and 0.071 respectively.
- The top 10 features collectively accounted for over 50% of the model's predictive power.
- Many features (approximately 30%) had negligible importance (close to or equal to zero), suggesting potential for feature reduction.

### 8.2 SHAP Visualization

Two key visualizations were generated to aid interpretation:

- **SHAP Summary Plot**: Showed the distribution of SHAP values for each feature, highlighting both the magnitude and direction of feature effects.
- **SHAP Bar Plot**: Provided a clear ranking of features by their average absolute SHAP values, confirming the dominance of features 59, 70, and 23.

These visualizations offer valuable insights for domain experts to understand which spectral regions are most indicative of DON contamination in corn samples.

## 9 Key Findings

The analysis yielded several important findings:

- The highly skewed distribution of DON concentration necessitated log transformation for effective modeling.
- A simple neural network architecture with appropriate non-linearity outperformed traditional machine learning approaches.
- The model achieved an $R^2$ score of 0.79 on the original scale, demonstrating strong predictive capability.
- A small subset of spectral features (particularly features 59, 70, and 23) carried most of the predictive information.
- The model's performance suggests that hyperspectral imaging combined with deep learning can provide a viable alternative to traditional mycotoxin detection methods.

## 10 Limitations

Despite the promising results, several limitations should be acknowledged:

- **Sample Size**: The dataset of 500 samples is relatively small for deep learning applications, potentially limiting the model's generalizability.
- **Extreme Values**: The wide range of DON concentration values (0 to 131,000 ppb) presents challenges for accurate prediction across the entire spectrum.
- **Log Scale Performance**: While the original scale metrics are strong, the modest $R^2$ score on the log scale (0.14) indicates room for improvement in capturing the nuances of the transformed data.
- **Feature Interpretation**: While SHAP analysis identifies important features, connecting these to specific spectral bands requires domain expertise in hyperspectral imaging.

## 11 Suggestions for Improvement

Based on the analysis, several avenues for improvement are recommended:

- **Feature Engineering**: Develop domain-specific features based on spectral ratios or derivatives that might better capture DON-related signals.
- **Ensemble Approaches**: Implement ensemble methods combining multiple models to improve prediction robustness, especially for extreme values.
- **Advanced Architectures**: Explore more sophisticated neural network architectures, such as convolutional neural networks (CNNs) that can better capture spectral patterns.
- **Transfer Learning**: Leverage pre-trained models from related spectral analysis tasks to improve performance with limited data.
- **Data Augmentation**: Generate synthetic samples to increase the effective dataset size, particularly for underrepresented DON concentration ranges.
- **Stratified Modeling**: Develop separate models for different DON concentration ranges to improve accuracy across the entire spectrum of values.
- **Uncertainty Quantification**: Implement methods to quantify prediction uncertainty, providing confidence intervals for DON concentration estimates.
- **Dockerize and Deploy, add CI/CD** I was not able to dockerize and deploy properly due to the weights of model being deined by services like Heroku, Streamlit, Vercel. I would like to explore further by paying them and dockerizing to run smoothly.

## 12 Conclusion

The mycotoxin prediction pipeline demonstrates the feasibility of using hyperspectral imaging data combined with deep learning to predict DON concentration in corn samples. The model achieves good predictive performance, explaining approximately 79% of the variance in DON concentration on the original scale. The interpretability analysis provides valuable insights into which spectral features are most indicative of DON contamination.

While there are limitations and opportunities for improvement, the current pipeline represents a significant step toward developing faster, more cost-effective methods for mycotoxin screening in agricultural products. Future work should focus on addressing the identified limitations and implementing the suggested improvements to enhance the model's accuracy, robustness, and practical utility.