Project Report: Predicting Vomitoxin Concentration in Corn

1. Introduction

Vomitoxin (DON) contamination in corn is a critical concern in agriculture and food safety. This project aims to develop a robust machine learning model to predict vomitoxin concentration based on spectral data features using CNNs, Random Forest, and XGBoost. The project involves data preprocessing, feature engineering, PCA for dimensionality reduction, and hyperparameter tuning for optimal model performance.

2. Preprocessing Steps and Rationale

- Removed outliers using IQR and replaced extreme values with the median.
- Applied log transformation and Box-Cox transformation to reduce skewness.
- Standardized numerical features using StandardScaler.
- Applied PCA to reduce dimensionality while retaining 95% variance.

3. Insights from Dimensionality Reduction

- PCA reduced features from 450 to the top 50 most significant ones.
- First principal component explained over 86% of the variance.
- Improved model efficiency without losing important information.

4. Model Selection, Training, and Evaluation

- Used CNN, Random Forest, and XGBoost for prediction.
- Performed Grid Search for hyperparameter tuning.
- CNN architecture: Conv1D layers with dropout and batch normalization.
- Random Forest performed best with RMSE ~12.48 and R^2 ~0.9998.

5. Key Findings and Suggestions for Improvement

- Removing outliers and using transformations improved model performance.
- CNN struggled with overfitting, whereas Random Forest performed best.
- Using more spectral data and fine-tuning PCA components could improve results further.