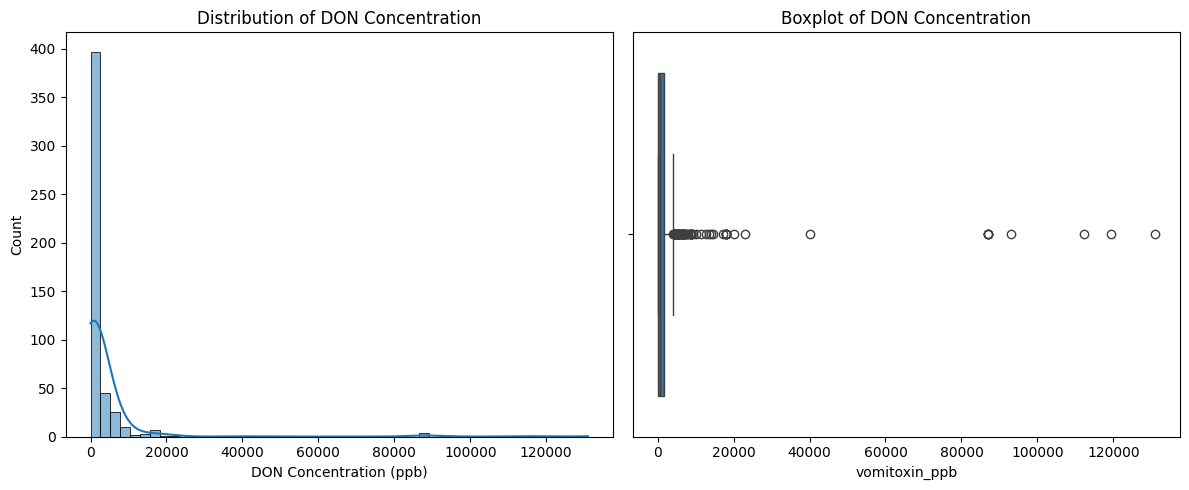
**Machine Learning Pipeline for DON Concentration Prediction**

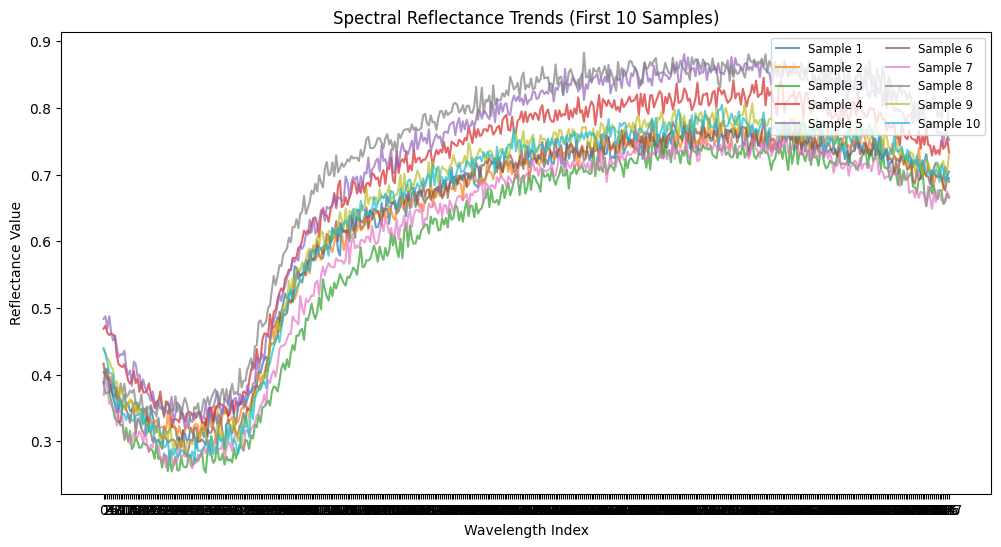
1. Data Pre-processing and Quality Checks

The dataset was analysed for missing values, and no missing entries were found.

The target variable, DON Concentration (vomitoxin\_ppb), was summarized as follows:

* Count: 500 samples
* Mean: 3410.006
* Standard Deviation: 13095.803
* Minimum Value: 0.000
* 25th Percentile (Q1): 137.500
* Median (Q2/50th Percentile): 500.000
* 75th Percentile (Q3): 1700.000
* Maximum Value: 131000.000

This box plot and Histogram visualised the target variable. As show, th data is extremely skewed, with many low values and few high values.   


This diagram shows the spectral reflectance of the first ten samples to visualize trends across wavelengths.  


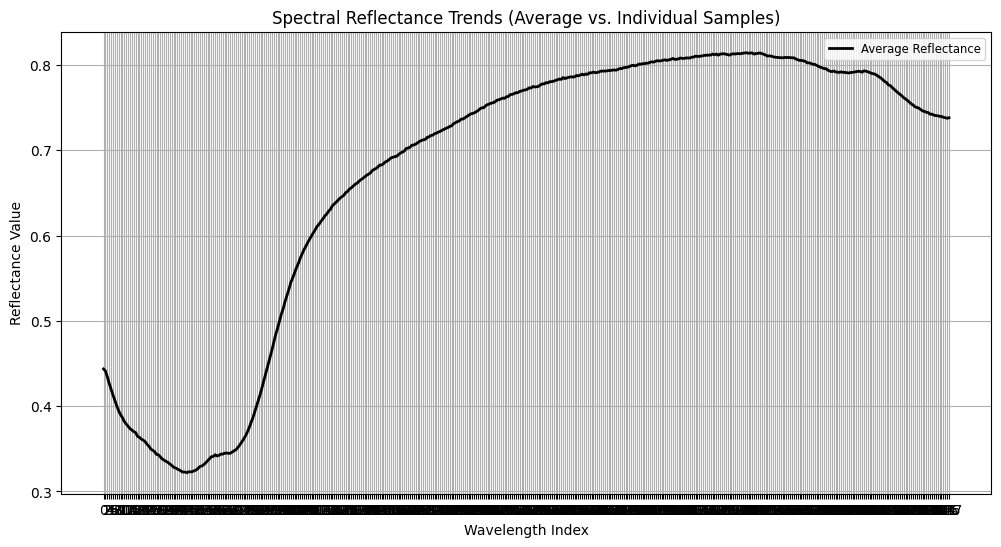
Outlier Detection and Removal:

Z-score Method identified 8 outliers, which were removed for further analysis.

IQR Method flagged 80 samples as potential outliers.

Isolation Forest detected 25 anomalies.

Based on the most reasonable approach, 8 outliers were removed using the Z-score method.

This diagram shows the average reflectance with the outliers removed.   
  


Data Consistency Checks:

16 inconsistent samples were identified and removed from the dataset.

This step ensured that sensor drift and inconsistencies in the spectral data did not impact model performance.  
A graph with blue dots

Description automatically generated

2. Model Training and Optimization

A simple neural network was chosen as the baseline model.

Performance of the baseline model:

* Mean Absolute Error (MAE): 1598.6073
* Root Mean Squared Error (RMSE): 2444.3378
* R² Score: -0.0420

Hyperparameters were optimized using Bayesian Optimization.

The best hyperparameters found:

Units per layer: 87

Number of layers: 1

Dropout rate: 0.2999

Optimized Neural Network Performance:

* MAE: 1621.5109
* RMSE: 2488.0714
* R² Score: -0.0797 (showing underfitting)

XGBoost method was also used:

* MAE: 57.3225
* RMSE: 215.7421

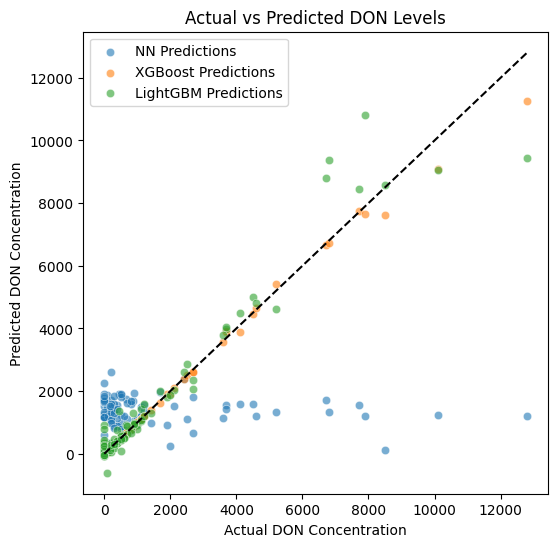
R² Score: 0.9919 (Best performance)

And LightGBM:

* MAE: 299.7755
* RMSE: 640.8835
* R² Score: 0.9284

XGBoost performed the best among all models, with the highest R² score and the lowest error rates.

Visualizations confirmed the effectiveness of XGBoost in making accurate predictions.



SHAP Analysis

SHAP (SHapley Additive exPlanations) was used to explain feature importance.

The most influential features:

Feature 448 had the strongest impact on predicting DON concentration.

Feature 392 also significantly influenced the model’s predictions.

Several features had little to no impact, suggesting they could be removed for model simplification.

