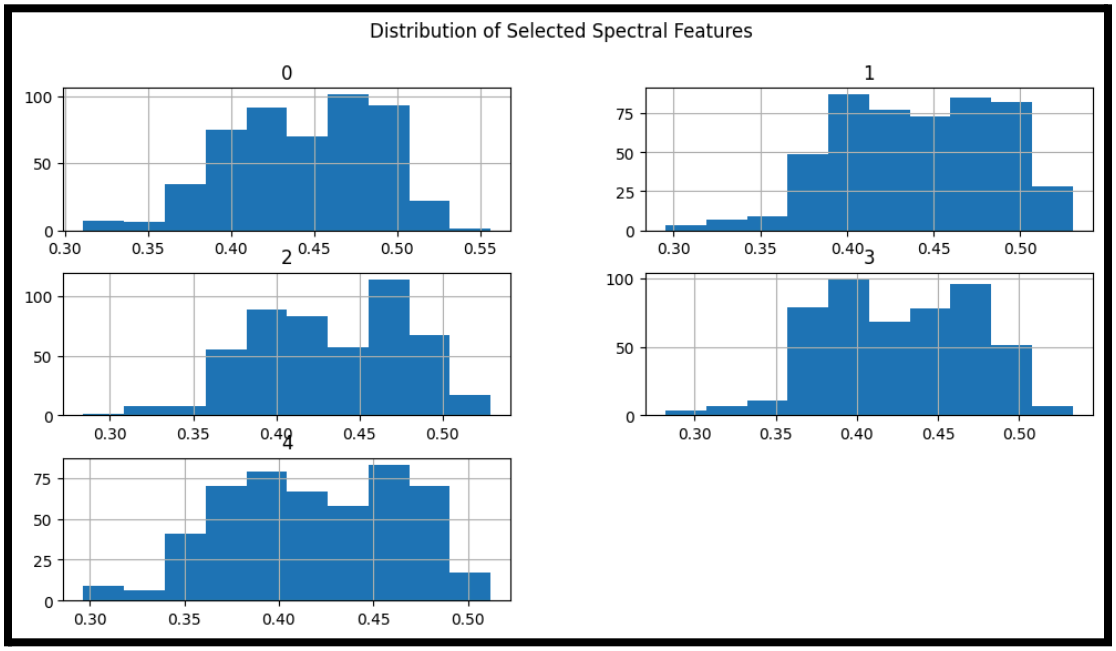
REPORT

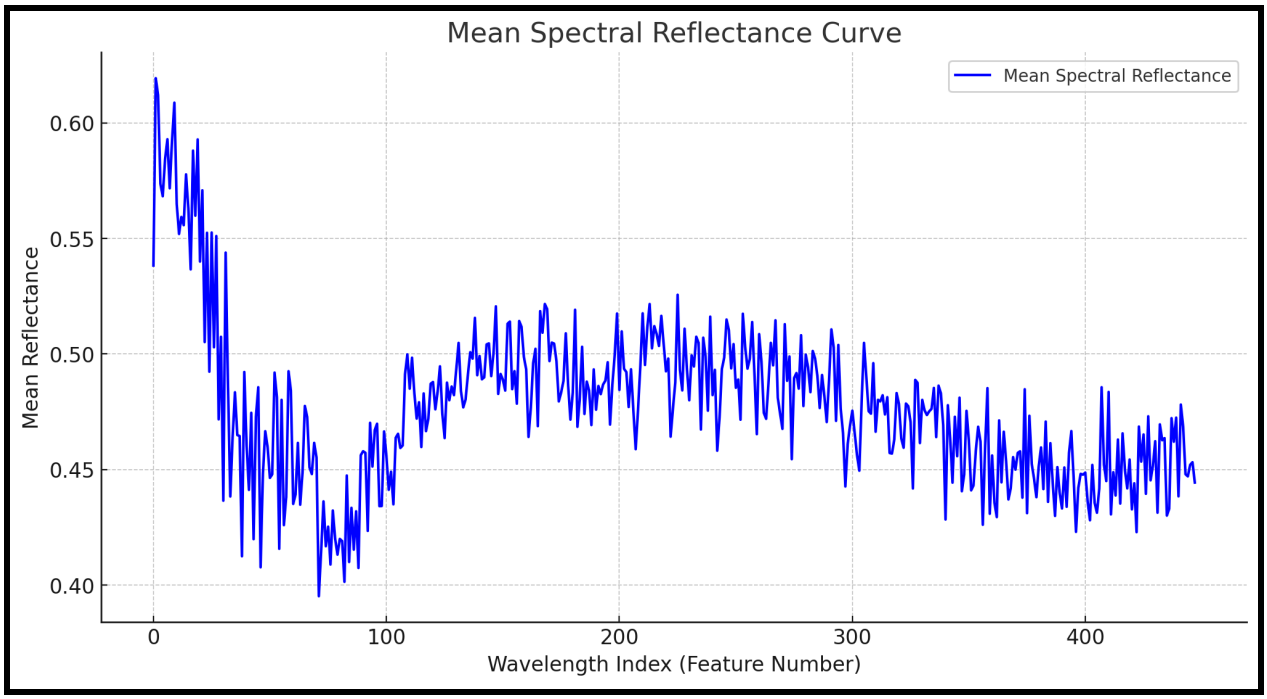
Understanding the dataset better:

Spectral reflectance features(0-477): Each column represents the intensity of reflected light at a specific wavelength.The values vary based on how the corn sample absorbs and reflects light at different wavelengths.

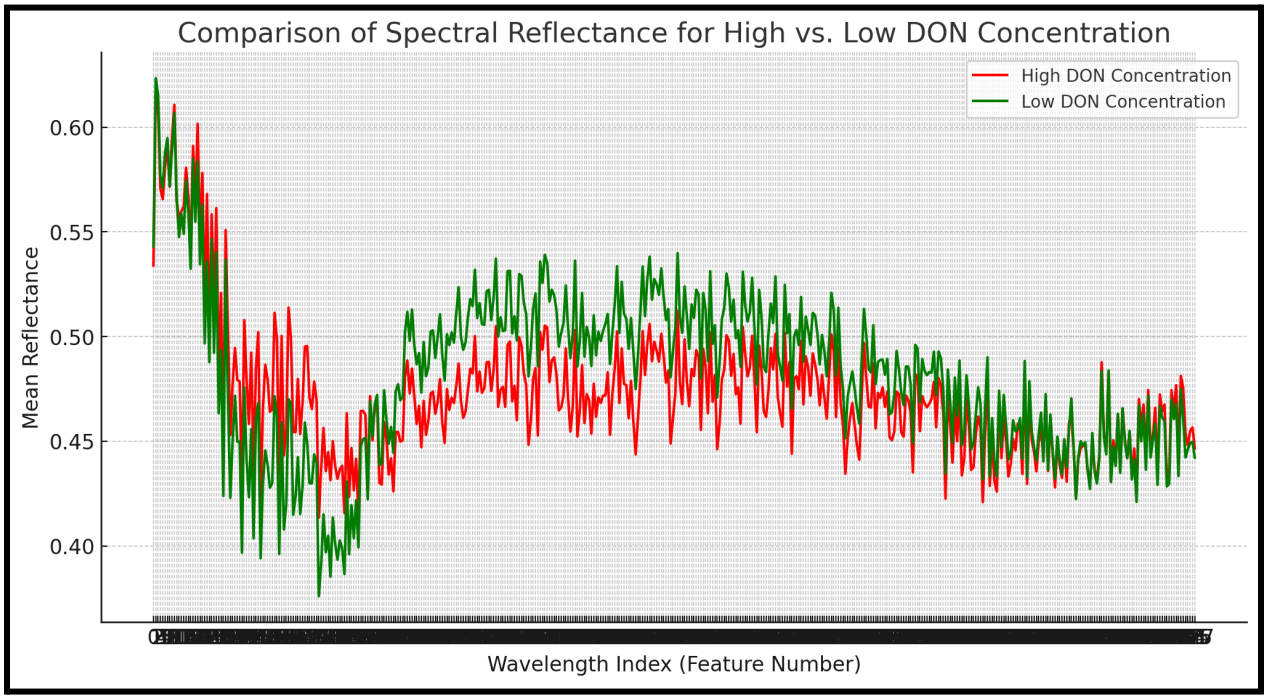
PREPROCESSING AND DATA EXPLORATION:

1. Dataset does not contain any null values.
2. Using the boxplot we visualised the target variable(‘vimotoxin’) and it showed some outliers.
3. Normalised the dataset using Min-Max scaling to ensure consistent features range, though the data seems normalised but still to make sure.
4. The target variable is continuous thus making it a regression problem.
5. 

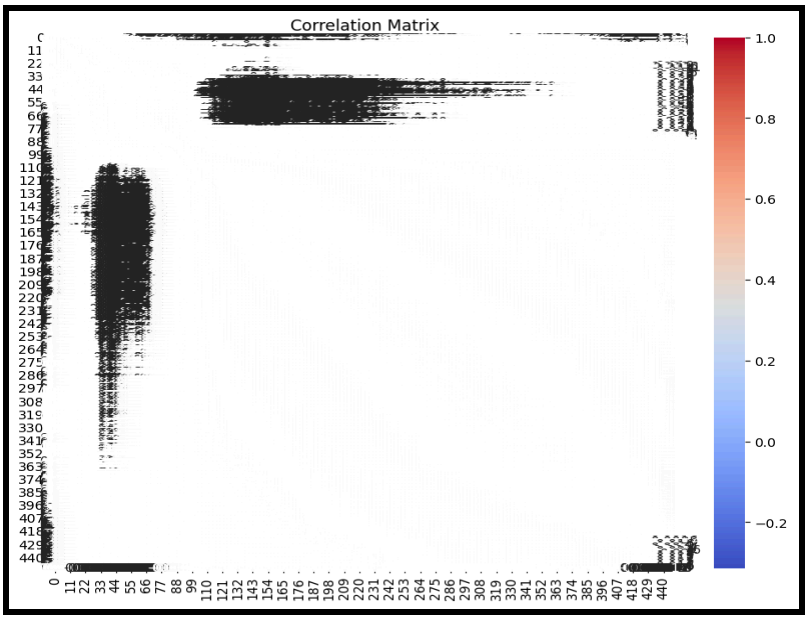
   The features show a bimodal distribution, this may be due to differences in spectral reflectance value for contaminated and non contaminated

6. 

The plot shows a varying mean reflectance, this surely ensures that with some particular wavelength contaminated and non contaminated act differently, to capture this particular wavelength range we plot a curve between low vs high DON samples

7. 

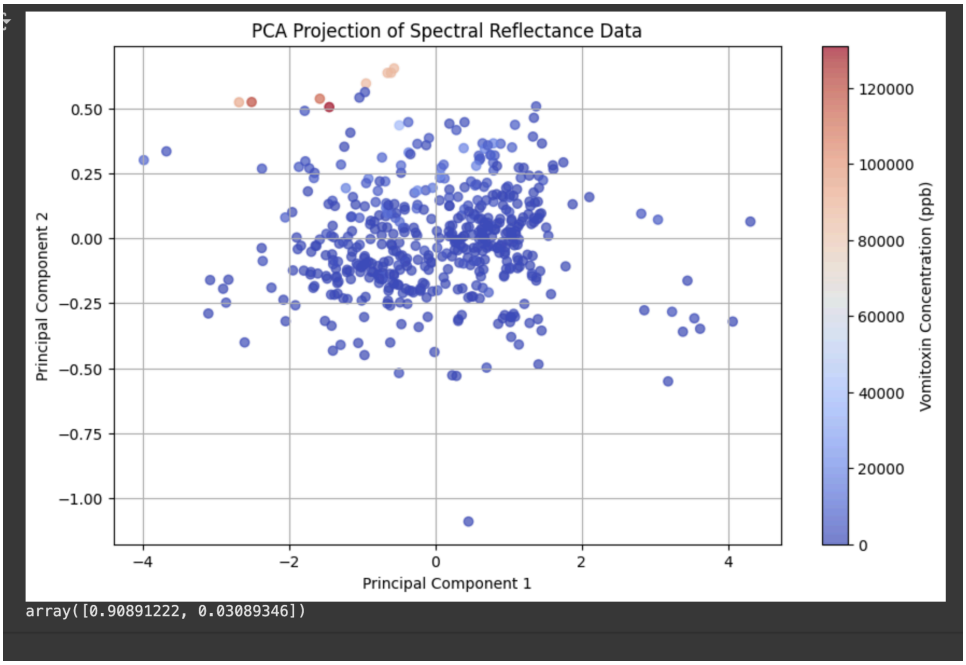
Some mid range wavelength shows greater separation between high and low DON samples, which seems to be important features for model training

8. 

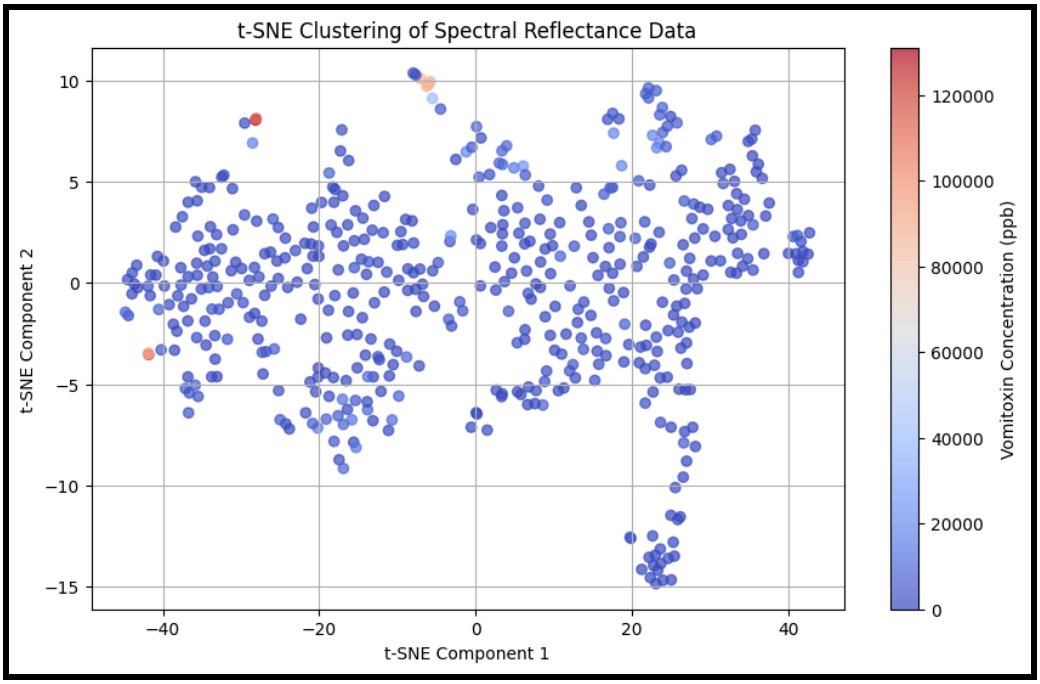
The correlation matrix with all features and vomitoxin doesnt reveal much but the dark regions definitely infer that some features are highly correlated.

DIMENSIONALITY REDUCTION:

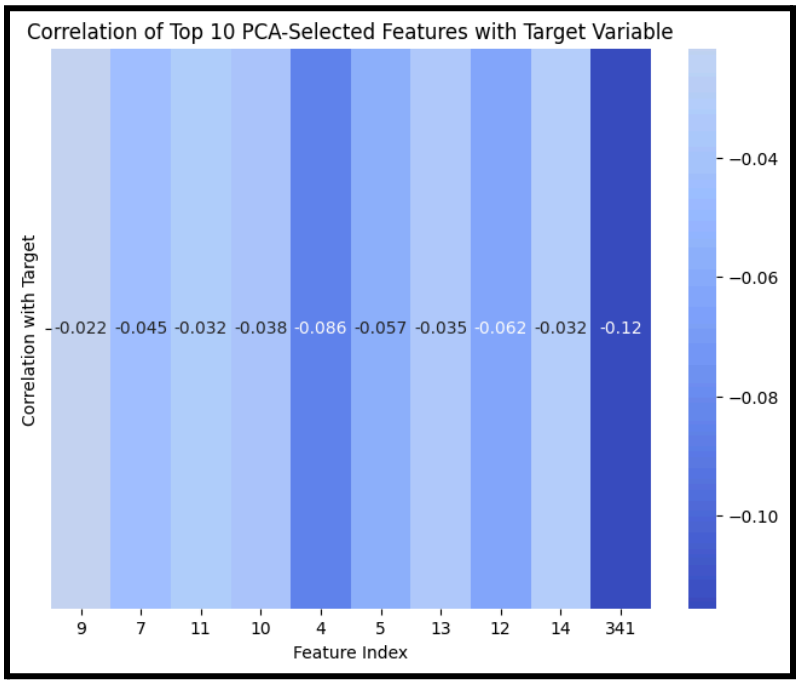
1. Implementing PCA:



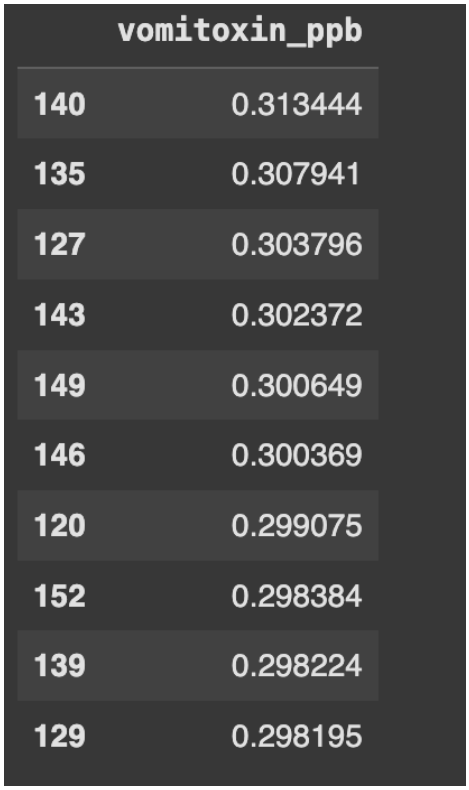
PC1=90.89% captures most of the variance in the data, so most of the spectral data is contained in PC1, so we can potentially just use few key spectral reflectance features instead of all 478

2. 

Doesnt reveal any new insights, most of the data points still overlap, there is some spectral separation between high and low concentration samples but not perfect, so

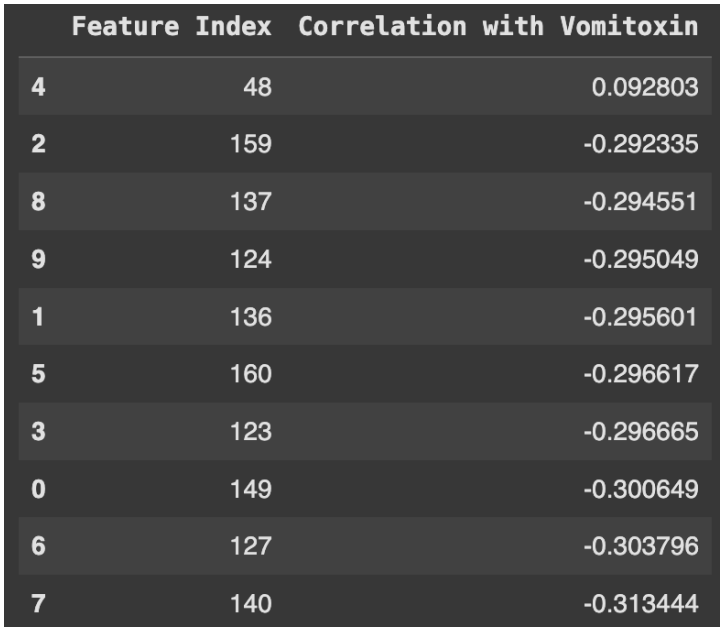
3. 

This shows the correlation of top 10 features found using PCA which captures 90 percent of variation and their relation with target variable, the above plot clearly shows that the most important PCA features are weakly related to the target variable, so thats of no use

4. 

	vomitoxin_ppb
140	0.313444
135	0.307941
127	0.303796
143	0.302372
149	0.300649
146	0.300369
120	0.299075
152	0.298384
139	0.298224
129	0.298195

The top 10 correlated features out of all 447 only shows a maximum correlation of 0.313, this indicates that the target variable does not depend strongly on few variables but has a complex relationship with most of the variables, this insight may help us choose the model.

5. 

	Feature Index	Correlation with Vomitoxin
4	48	0.092803
2	159	-0.292335
8	137	-0.294551
9	124	-0.295049
1	136	-0.295601
5	160	-0.296617
3	123	-0.296665
0	149	-0.300649
6	127	-0.303796
7	140	-0.313444

This show correlation of top 10 features obtained using random forest regressor with target variable, here also we don't see any strong correlation.

Model Selection, Training, and Evaluation

Models Compared:

1. Convolutional Neural Network (CNN)
2. Fully Connected Neural Network (Optimized MLP)

Model Training Details:

- Hyperparameter tuning was performed using grid search for both models.
- CNN Architecture: 2 Conv1D layers, MaxPooling, Flatten, and Dense layers.
- Neural Network (MLP) Architecture: Fully connected layers with ReLU activations, dropout for regularization.
- Optimizer: Adam
- Loss Function: Mean Squared Error (MSE)

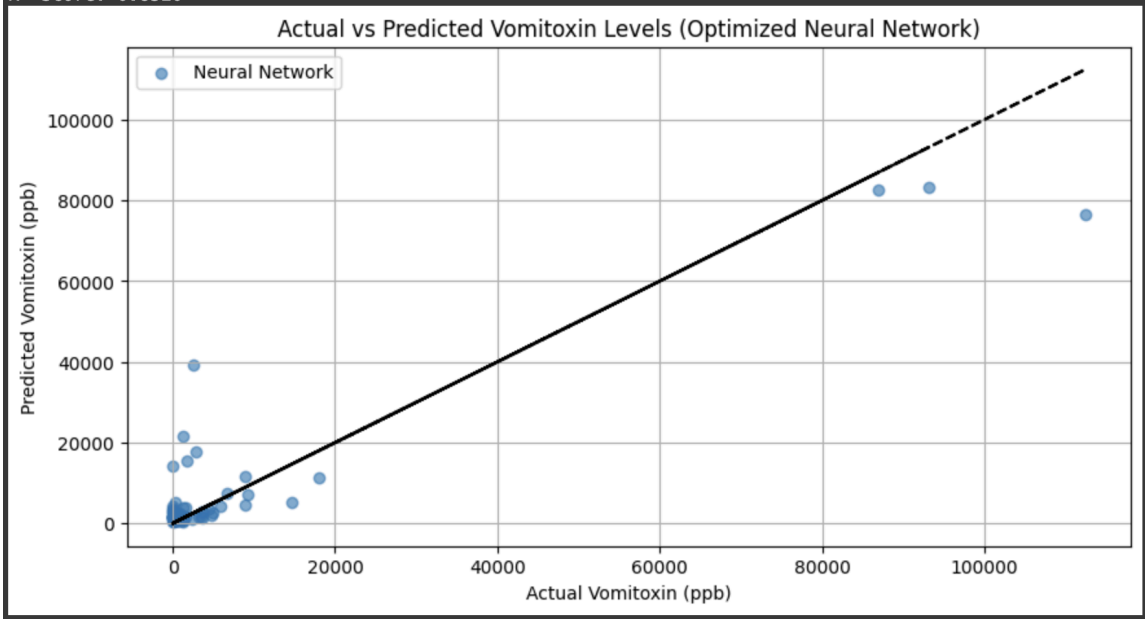
Evaluation Metrics:

- Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R^2 Score

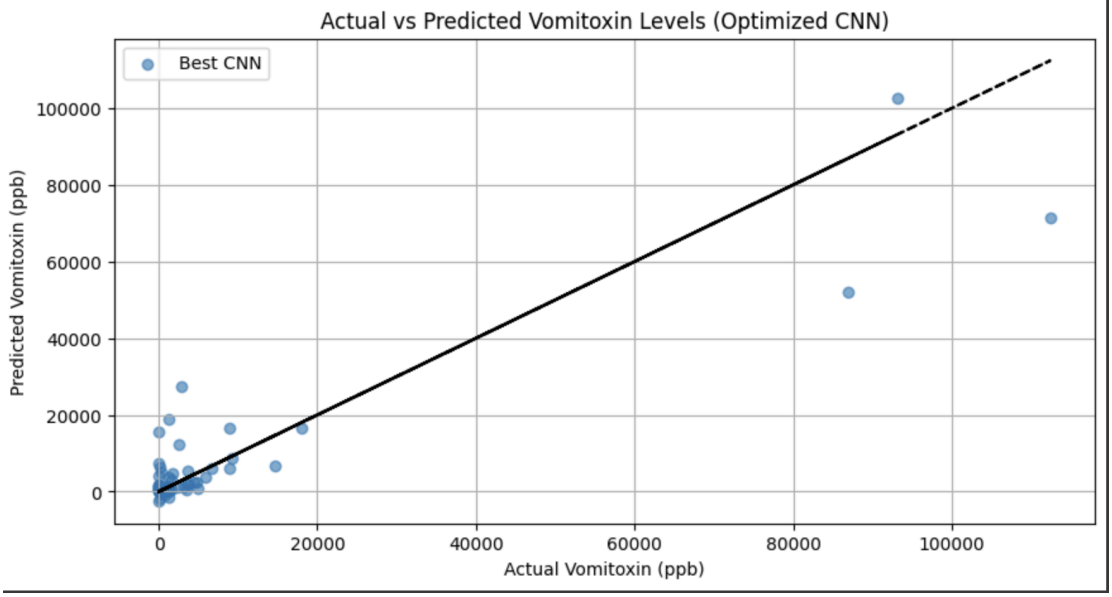
Performance comparison:

MODEL	MAE	RMSE	R^2 score
Neural net(optimized)	2818.8861	6431.3703	0.8520
CNN(optimized)	2673.623	6815.993	0.838

```
Model: Optimized Neural Network
MAE: 2818.8861
RMSE: 6431.3703
R² Score: 0.8520
```




```
Final Best CNN Model -> MAE: 2673.6233, RMSE: 6815.9953, R²: 0.8338
```



KEY FINDINGS AND SUGGESTIONS FOR IMPROVEMENT:

FINDINGS:

1. Neural network outperformed CNN, achieving lower MAE and RMSE and a better R^2 score.
2. CNN didn't generalise well with this dataset, possibly due to the lack of spatial relationships in spectral reflectance data.

3. Using all 447 features provided the best results, reducing the feature set did not improve performance, to prevent overfitting used drop regularization, batch normalisation and early stopping.

POTENTIAL IMPROVEMENTS:

1. Feature engineering instead of using raw data.
2. Further fine-tuning of dropout rates, learning rates, and layer sizes.
3. Could use a transformer model as there are many redundant features and transformers are robust to noise