Analysis of Hyperspectral Data Models

March 13, 2025

1 Introduction

This report evaluates different machine learning models for predicting vomitoxin (DON) concentration using hyperspectral data. The objective is to analyze their performance, strengths, and weaknesses while identifying potential improvements.

2 Preprocessing Steps and Rationale

Hyperspectral data often contain redundant information and high-dimensional features. The preprocessing steps included:

- **Normalization**: Ensured all spectral bands had comparable scales.
- **Dimensionality Reduction**: PCA was applied to reduce feature space while retaining variance.
- **Feature Engineering**: Extracted meaningful features to enhance model interpretability.

3 Model Comparisons and Performance Analysis

The tested models include:

- **FeatureNet + XGBoost**: A feature extraction network combined with XGBoost.
- **CNN + Transformer + XGBoost**: Deep learning with convolutional and transformer layers.
- **Improved MLP + Optimized XGBoost**: A refined MLP with tuned XGBoost.

3.1 Performance Metrics

The models were evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score.

Model	**MAE**	**RMSE**	R^2 **Score**
FeatureNet + XGBoost	2265.12	8920.49	0.7153
CNN + Transformer + XGBoost	4234.03	17023.89	-0.0368
Improved MLP + Optimized XGBoost	**1638.36**	**3267.13**	**0.90+**

Table 1: Comparison of model performances.

4 Error Analysis and Model Weaknesses

4.1 FeatureNet + XGBoost

This model performed moderately well but had high RMSE, suggesting:

- Limited ability to capture spectral relationships.
- Need for enhanced feature extraction strategies.

4.2 CNN + Transformer + XGBoost

This model underperformed, exhibiting a negative R^2 score, implying poor generalization. Possible reasons include:

- Overfitting due to complex architecture.
- Inadequate training data for transformers.
- Suboptimal feature fusion strategies.

4.3 Improved MLP + Optimized XGBoost

This model achieved the best results with a high R^2 score (0.90+), due to:

- Effective feature selection and extraction.
- Optimization of XGBoost hyperparameters.
- Improved handling of high-dimensional data.

5 Key Findings and Recommendations

- Deep learning models require careful architecture selection to avoid overfitting.
- Feature extraction methods play a critical role in improving model performance.
- Optimized XGBoost combined with an improved MLP is the most effective approach.