Hyperspectral Data Analysis for Vomitoxin Prediction in Corn Samples

This technical report presents a comprehensive analysis of hyperspectral data for predicting vomitoxin (DON) concentration in corn samples. The implemented pipeline includes extensive data preprocessing, dimensionality reduction, and neural network modeling techniques to achieve accurate prediction performance from high-dimensional spectral features.

Preprocessing Steps and Rationale

The preprocessing pipeline was carefully designed to handle the complex nature of hyperspectral data:

Column-wise mean imputation was implemented for missing values, preserving the unique distribution characteristics of each spectral band rather than using global statistics. This approach maintains the integrity of wavelength-specific information critical for accurate analysis.

Feature engineering expanded the original 448 spectral bands into a richer feature set. First-order derivatives (447 features) were calculated between adjacent spectral bands to capture the rate of change in reflectance across the spectrum. These derivatives highlight absorption and reflection transitions particularly relevant for detecting biochemical compounds like vomitoxin.

Band ratios and normalized difference indices (NDIs) were computed to emphasize relationships between different parts of the spectrum and normalize for illumination differences across samples. The implementation carefully addressed potential division by zero issues by adding a small epsilon value (1e-6) to denominators.

Outlier detection employed the Interquartile Range (IQR) method with threshold factor of 3. Values beyond 3×IQR from quartiles were clipped rather than removed, preserving the overall sample size while reducing the impact of extreme values.

Target variable transformation using log transformation (np.log1p) addressed the right-skewed distribution of vomitoxin concentration values. This stabilized variance, made the distribution more symmetric, and improved model performance by making the target variable more amenable to linear modeling techniques while preserving zero values.

Robust scaling was applied to center features around the median and scale based on interquartile range instead of mean and standard deviation, providing resilience against remaining outliers in the spectral data.

Insights from Dimensionality Reduction

Principal Component Analysis (PCA) revealed that 10 principal components were sufficient to capture approximately 95% of the total variance in the dataset, significantly reducing the dimensionality from over 1100 features. The first principal component alone captured about 45% of the total variance, indicating a strong primary pattern in the data.

Feature importance analysis based on PCA loadings identified critical spectral regions for vomitoxin prediction. Features 1010, 1011, 900, and 901 demonstrated the highest total absolute loadings across principal components, suggesting these spectral regions contain the most relevant information for the prediction task. These regions likely correspond to molecular absorption features related to the mycotoxin or associated fungal presence.

Visualization of samples in PCA space revealed clustering patterns related to vomitoxin concentration levels, confirming the relationship between spectral characteristics and the target variable.

t-SNE analysis with 3 components provided complementary insights into the local neighborhood structures and revealed potential subgroups in the data that might be obscured in linear projections like PCA. Silhouette score analysis on the t-SNE results identified optimal clustering in the data that showed correlation with vomitoxin concentration levels.

Model Selection, Training, and Evaluation Details

Three neural network architectures were implemented and compared:

The LSTM model reshaped spectral bands as sequential data with 47 LSTM units in recurrent layers followed by 90 dense units. A dropout rate of approximately 0.40 was applied for regularization, with a learning rate of 0.00055. Bidirectional LSTM layers captured relationships in both directions of the spectral sequence.

The CNN model used a 1D convolutional architecture with 32 filters and kernel size 3 for feature extraction, followed by MaxPooling layers and 64 dense units. A dropout rate of 0.3 and learning rate of 0.0005 were optimized for stable convergence.

The standard neural network implemented 4 hidden layers with 211 neurons per layer, ReLU activation functions, and a dropout rate of 0.16 with a learning rate of 0.0002815.

Training methodology incorporated batch processing with size 32, custom early stopping with patience of 10 epochs, and gradient clipping with maximum norm of 1.0 to prevent exploding gradients, which is particularly important for LSTM networks.

Bayesian optimization with a Gaussian Process Regression and Expected Improvement acquisition function was used for hyperparameter tuning, with 5-fold cross-validation ensuring robust parameter selection.

Evaluation metrics included Mean Absolute Error (MAE) as the primary metric for assessing prediction error magnitude, Root Mean Squared Error (RMSE) for penalizing larger errors more heavily, and R^2 Score as a standardized measure for comparison across different target scales.

Key Findings and Suggestions for Improvement

The LSTM model demonstrated superior performance in capturing the sequential patterns in spectral data, as evidenced by the convergence of training and validation metrics over 100 epochs. The validation MAE stabilized around 0.5, indicating good predictive capability for log-transformed vomitoxin concentrations.

The actual vs. predicted plot shows a positive correlation between predicted and actual values, though with some scatter points deviating from the ideal prediction line, particularly for higher concentration values. This suggests the model performs better at lower concentration ranges and could be improved for higher values.

PCA loadings analysis revealed that certain spectral regions (especially features around 1000-1014) have disproportionate influence on the principal components, providing valuable targets for simplified sensor development.

For future improvement, implementing advanced spectral preprocessing techniques such as Savitzky-Golay filtering for noise reduction and continuum removal to normalize spectra would enhance feature quality. Exploring transfer learning approaches with pre-trained networks on similar hyperspectral datasets could improve performance with limited sample sizes.

Developing ensemble methods combining predictions from multiple architectural approaches (CNN, LSTM, Attention) would leverage their complementary strengths. Enhanced model interpretability through SHAP values and attention visualization techniques would provide better insights into decision-making processes.

Integration of additional data sources such as weather and field conditions that affect fungal growth would provide contextual information that could improve prediction accuracy. Implementing spectral data augmentation techniques would expand the training dataset and potentially improve model robustness.