Hyperspectral Imaging for DON Prediction – ML Intern Task

1. Introduction

The presence of mycotoxins such as Deoxynivalenol (DON) in corn is a critical concern for food safety and quality. Hyperspectral imaging (HSI) provides a powerful, non-destructive technique to analyze spectral reflectance data across hundreds of narrow wavelength bands, which can be used to estimate toxin levels. This project aimed to develop machine learning models that predict DON concentration using HSI data. We explored preprocessing techniques, dimensionality reduction, and regression modeling using both Feedforward Neural Networks (FNN) and Convolutional Neural Networks (CNN).

2. Data Preprocessing

The dataset included reflectance values across multiple spectral bands for individual corn samples, with DON concentration as the target variable. The first step in preprocessing involved removing non-informative identifiers and checking for missing values. No significant data quality issues were found. Next, we standardized all spectral features using StandardScaler to normalize the scale across bands. This ensures that the machine learning models treat each band equally during training, without bias from feature magnitude differences. We also visualized the average reflectance spectrum to understand general trends across wavelengths, and used a heatmap to observe sample-level variation. These exploratory steps confirmed that the spectral data showed measurable variance and contained a signal likely relevant to DON levels.

3. Dimensionality Reduction

Insights Due to the high number of input features (wavelength bands), dimensionality reduction was applied to simplify the learning task and enhance interpretability. Principal Component Analysis (PCA) was used to reduce the feature space while preserving 95% of the data variance. The first few principal components captured significant structure in the data, and scatter plots of the transformed samples revealed separation trends that hinted at the model's ability to discriminate DON levels based on spectral inputs. Page 1 t-SNE In parallel, t-SNE was used as a non-linear technique to visualize the clustering of samples in a 2D space. It reinforced the PCA findings by showing local grouping of samples with similar DON values, further validating that the spectral data holds predictive power.

4. Model Development and Evaluation

We trained two models: a basic Feedforward Neural Network (FNN) and a Convolutional Neural Network (CNN), each designed to regress DON values based on input reflectance. Feedforward Neural Network (FNN).

The FNN was built with several dense layers and trained using standard backpropagation. It was simple but effective, and performed well on this compact dataset. It served as a strong baseline model.

Convolutional Neural Network (CNN) Since hyperspectral data is naturally sequential, we built a 1D CNN to capture local spectral patterns across neighbouring wavelengths. This model used convolutional and pooling layers to extract higher-level features before feeding them into fully connected layers for regression.

Evaluation Metrics We assessed both models using:

- Mean Absolute Error (MAE) - measures average error magnitude.
- Root Mean Squared Error (RMSE) - penalizes larger errors more strongly.
- R^2 Score - measures the proportion of variance in DON explained by the model.

The CNN had a higher R^2 score, indicating a better ability to capture overall trends in DON concentration. However, the FNN had lower error metrics (MAE and RMSE), suggesting it made more accurate individual predictions.

5. Key Findings

- Hyperspectral reflectance data contains a strong signal for predicting DON concentrations.
- Dimensionality reduction via PCA and t-SNE confirmed that spectral patterns can differentiate samples by DON levels.
- While CNN generalized better, the FNN produced more precise predictions per sample on this dataset.

6. Limitations and Improvement Opportunities

Limitations:

- The dataset size was limited, especially for deep learning models like CNN, which benefit from larger samples.
- The spectral bands were used as-is, without band selection or domain-driven feature engineering.
- Overfitting risk remained due to high dimensionality and relatively small training data.

Suggestions for Improvement:

- Collect more samples or augment the dataset to improve model robustness.
- Apply spectral band selection using information gain or domain knowledge to reduce noise. Use cross-validation to better evaluate generalizability.
- Experiment with hybrid or ensemble models to combine the strengths of FNN and CNN.

7. Conclusion

This project successfully demonstrated the potential of machine learning-particularly neural networks-for predicting DON concentrations from hyperspectral imaging data. Both FNN and CNN models showed value, Page 3 each with unique strengths. With further improvements in data quality and modeling techniques, this approach can contribute significantly to real-time, non-destructive food safety monitoring