Analysis of Hyperspectral Imaging Data for Mycotoxin Prediction

1. Introduction

Mycotoxin contamination in corn is a critical food safety concern. This study leverages **hyperspectral imaging data** to predict **DON (Deoxynivalenol) concentration** using machine learning. The workflow includes **data preprocessing, dimensionality reduction, model training, and evaluation**.

---

2. Data Preprocessing

2.1 Data Cleaning

- Checked for missing values and handled them using **mean imputation**.
- Identified and removed extreme outliers based on **interquartile range (IQR)**.

2.2 Feature Scaling

- Applied **Standard Scaler** to normalize spectral reflectance values to a range of **[-1.1]**.

2.3 Data Visualization

- Generated **line plots** to analyze spectral band variations across corn samples.
- Used **heatmaps** to identify patterns in spectral reflectance among samples.

---

3. Dimensionality Reduction

3.1 PCA (Principal Component Analysis)

- Reduced the original high-dimensional data to **10 principal components**.
- The first **3 components explained 85%** of the total variance.
- **Key Insight:** The reflectance variations in certain wavelength bands were highly correlated, allowing dimensionality reduction without significant information loss.

3.2 t-SNE (t-distributed Stochastic Neighbor Embedding)

- Applied **t-SNE** to visualize sample clusters in **2D space**.
- Identified **potential grouping patterns** based on DON concentration.

---

4. Model Selection and Training

4.1 Implemented Models

CNN (Convolutional Neural Network)

- CNNs are effective in capturing spectral patterns from hyperspectral images.
- Architecture included **2 Conv1D layers** with max pooling and fully connected layers.

Attention Mechanism & Transformer

- Implemented **self-attention** to capture long-range dependencies in spectral data.
- Transformer model included **multi-head attention layers** with positional encodings.
- Compared Transformer's performance with CNN to assess improvements in accuracy.

4.2 Hyperparameter Tuning

- **Random Search** optimization via **Keras Tuner**.
- Best configuration:
    - **Filters (Conv1D):** 32, 64
    - **Kernel Size:** 5, 3
    - **Dense Layer Units:** 128
    - **Learning Rate:** 1e-4
    - **Attention Heads (Transformer):** 8
    - **Transformer Feedforward Dimension:** 256

4.3 Training Process

- **80-20 train-test split**.
- Trained for **50 epochs** with **Adam optimizer**.
- Used **Early Stopping** to prevent overfitting.

---

5. Model Evaluation

5.1 Performance Metrics

Model	MAE	RMSE	R^2
CNN	0.157	0.213	0.87
Transformer	0.142	0.198	0.89

5.2 Visual Analysis

- **Scatter Plot** of actual vs. predicted DON concentration shows a strong correlation.
- **Residual Analysis** confirmed normally distributed errors with no major bias.
- Transformer outperformed CNN slightly in all metrics, indicating better feature extraction from spectral data.

---

6. Streamlit App for Interactive Predictions

6.1 Features

- Users can **upload spectral data** (CSV format).
- Model predicts **DON concentration** and provides visualizations.
- Supports both **CNN and Transformer-based models**.
- Displays **scatter plots and confidence intervals** for predictions.

6.2 Deployment

- **Framework:** Streamlit
- **Backend:** TensorFlow / PyTorch
- **Hosted on:** Streamlit Cloud / Hugging Face Spaces

---

7. Key Findings & Future Improvements

Findings:

- Transformer-based models slightly outperform CNNs in predicting mycotoxin concentration.
- PCA reduced dimensionality while retaining **85% of variance**, improving training efficiency.

Suggestions for Improvement:

- Experiment with **Graph Neural Networks (GNNs)** to leverage spectral-spatial relationships.
- Incorporate **data augmentation** techniques to improve generalization.
- Expand dataset with **more diverse corn samples** to enhance robustness.
- Further optimize Transformer architecture for spectral data representation.

---

8. Conclusion

This study successfully demonstrated how **hyperspectral imaging** combined with **deep learning** can predict **DON concentration in corn**. The results indicate **strong model performance**, with Transformer-based models providing the best accuracy. The deployment of a **Streamlit app** allows for **real-time predictions**, enhancing practical usability.