# Short Report: Predicting Vomitoxin Levels in Corn

## 1. Preprocessing Steps & Rationale

- **Data Cleaning**: Removed missing/null values and normalized spectral features.
- **Feature Engineering**: Used **Principal Component Analysis (PCA)** to reduce dimensionality while retaining variance.
- **Train-Test Split**: **80% training, 20% testing** for model evaluation consistency.

## 2. Insights from Dimensionality Reduction

- **PCA Visualization**: Showed a concentration of variance in the first few components, confirming the dataset's high dimensionality.
- **Feature Selection**: Reduced input size while maintaining predictive power, improving model efficiency.

## 3. Model Selection, Training & Evaluation

- **Models Used**:
  - **Traditional ML**: Random Forest, XGBoost
  - **Deep Learning**: MLP, CNN, LSTM
  - **Tuned Models**: Grid Search applied to **MLP, Random Forest, and XGBoost**
- **Performance Summary**:
  - **Best Model (MLP Tuned) → MAE: 578.23 ppb**, **RMSE: 925.02**, **$R^2$: 0.56**
  - **Random Forest & XGBoost performed comparably** but had slightly higher errors.
  - **Deep models (LSTM, GRU) struggled**, likely due to insufficient sequential dependencies in spectral data.

## 4. Key Findings & Suggestions for Improvement

**Best validation MAE** achieved with MLP (Hyperparameter-Tuned).
**Dimensionality reduction via PCA helped** reduce complexity while preserving performance.
**Deep models struggled**, suggesting that simpler architectures may be more suitable.

📌 **Future Improvements**:

- **More hyperparameter tuning** (batch size, activation functions, learning rate decay).
- **Experiment with hybrid models** (e.g., CNN + MLP).
- **Increase data augmentation or apply synthetic data generation** to improve generalization.