**Spectral Data Analysis and Prediction Report**

Prepared by Vikash Shakya

Date: 11 March 2025

**1. Preprocessing Steps and Rationale**

* **Data Cleaning:**
  + Loaded the dataset and inspected for missing values and inconsistencies.
  + Found and handled missing values using mean/median imputation.
  + Detected and removed outliers using the IQR (Interquartile Range) method.
* **Feature Scaling:**
  + Applied Min-Max scaling to normalize spectral data between 0 and 1 for consistent model behavior.
* **Data Transformation:**
  + Reshaped the dataset to match the input format for CNN (3D tensor).
  + Split the data into 80% training and 20% testing sets.

**2. Insights from Dimensionality Reduction**

* **Principal Component Analysis (PCA):**
* Reduced data dimensionality from 450+ bands to 15 principal components.
* Top 5 principal components explained 85% of the variance — indicating that most of the signal is concentrated in a small number of dimensions.
* **t-SNE:**
* t-SNE revealed clear clustering patterns, indicating that some spectral bands have unique patterns that can differentiate sample types.

**3. Model Selection, Training, and Evaluation**

**Model: Convolutional Neural Network (CNN)**

CNN was selected due to its effectiveness in handling structured, high-dimensional data like spectral data.

**Architecture:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Layer** | **Type** | **Parameters** | **Activation** |
| Input Layer | Conv1D | Filters = 128, Kernel Size = 5 | ReLU |
| Hidden Layer 1 | MaxPooling1D | Pool Size = 2 | - |
| Hidden Layer 2 | Conv1D | Filters = 256, Kernel Size = 3 | ReLU |
| Hidden Layer 3 | MaxPooling1D | Pool Size = 2 | - |
| Hidden Layer 4 | Conv1D | Filters = 512, Kernel Size = 3 | ReLU |
| Hidden Layer 5 | MaxPooling1D | Pool Size = 2 | - |
| Dropout Layer | Dropout | Rate = 0.3 | - |
| Fully Connected Layer | Dense | Units = 256 | ReLU |
| Fully Connected Layer | Dense | Units = 128 | ReLU |
| Fully Connected Layer | Dense | Units = 64 | ReLU |
| Output Layer | Dense | Units = 1 | Linear |
|  |  |  |  |

**4. Model Performance**

* Performance Metrics:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Metric** | | |  | | --- | | **Value** | | |  | | --- | | **Interpretation** | |
| Mean Absolute Error (MAE) | 4542.53 | Average error of prediction is ±4542 ppb |
| Root Mean Squared Error (RMSE) | 14203.61 | Penalizes large errors more heavily |
| R² Score | 0.7266 | Model explains ~72.66% of variance in the data |

**Performance Trend:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Before Fine-Tuning** | **After Fine-Tuning** | **Improvement** |
| **MAE** | 4607.35 | 4542.53 | Improved by 1.4% |
| **RMSE** | 14964.88 | 14203.61 | Reduced by 5% |
| **R² Score** | 0.6965 | 0.7266 | Improved by 4.3% |

**5. Key Findings and Suggestions for Improvement**

**Findings:**

* The CNN model captured spectral patterns effectively, achieving a high R² score (0.7266).
* Lower MAE and RMSE indicate improved predictive accuracy.
* PCA helped reduce dimensionality without losing critical information.

**Suggestions for Improvement:**

* Added an Attention Mechanism → To help the model focus on the most important spectral bands.
* Tried a Transformer Model → Could improve performance by capturing long-range dependencies.
* Hybrid Model (CNN + LSTM): If data has sequential patterns, combining CNN and LSTM might further improve accuracy.
* Data Augmentation: Increasing the sample size could reduce overfitting and improve generalization.

**6. Conclusion**

The CNN model demonstrates strong predictive power with an R² score of 0.7266 and a low MAE of 4542.53.