Moisture Prediction in Biomass Using Spectral Data

Mete Harun Akcay

I. Introduction

Near-Infrared (NIR) spectroscopy is a widely used technique for analyzing the chemical composition and properties of various materials, including biomass. It provides a non-destructive, rapid, and cost-effective means for determining the moisture content in biomass samples, which is crucial for optimizing processing, transportation, and storage in the biomass industry.

The purpose of this project is to evaluate and compare machine learning approaches for predicting biomass moisture content using NIR spectroscopy data. In particular, the study explores the effectiveness of different data pre-processing techniques and machine learning models to predict moisture content accurately.

The steps in this project include:

- Data analysis and exploration of the NIR spectral data.
- Application of various pre-processing techniques such as Savitzky-Golay filtering, Standard Normal Variate (SNV), and Multiplicative Scatter Correction (MSC) to improve the quality of the spectral data.
- Implementation of machine learning models, including Partial Least Squares (PLS), Support Vector Regression (SVR), and Artificial Neural Networks (ANN), to predict biomass moisture content.
- Evaluation and comparison of the models based on performance metrics such as R^2 and RMSE, with cross-validation applied to assess their accuracy and robustness.

II. Data Analysis

A. Dataset Description

The dataset used in this project consists of Near-Infrared (NIR) spectral data along with reference moisture content values for biomass samples. The NIR spectra were collected from solid biomass samples, including random blends of pine and spruce wood chips, bark, forest residues, and sawdust, sourced from biomass processing facilities. Spectral data were recorded using an FT-NIR spectrometer while the samples were moving on a turntable at a speed of 1 m·s^{-}1, with relative absorbance measured over a spectral range from 834 to 2500 nm.

The acquisition parameters for the NIR dataset are summarized in Table I.:

B. Data Exploration

The dataset consists of 773 rows and 1040 columns. Among these, two columns are designated for IDs, 1037 columns correspond to the spectral data points, and the final column contains the moisture content values for the biomass samples.

Parameters	**Value**
Recorded spectral range	834-2500 nm (12000–4000 cm^{-}1)
Number of measurements/sample	5-7
Spectral resolution	16 cm^{-}1
Number of data points in each spectra	1037
Number of tested samples	125

TABLE I: Parameters of the NIR dataset

There were no missing values in the dataset, ensuring its completeness for analysis.

There could be potential issues with the dataset. One of the challenges is high dimensionality; with 1037 spectral columns, the dataset may suffer from the ”curse of dimensionality,” which could lead to overfitting. Additionally, the data might be noisy due to various environmental or instrumental factors, making preprocessing necessary. The dataset may also contain redundant information, as many wavelengths may carry similar details, thus requiring dimensionality reduction techniques such as PLS or PCA. The relationship between spectral data and moisture content might also be non-linear, requiring more advanced regression techniques to accurately model this association.

To better understand the dataset, a graph showing the spectral data of 10 randomly selected samples is given in Figure 1, where x-axis represents the wavelength and the y-axis represents the absorbance values. As seen, spectral data for these samples follows a similar trend.

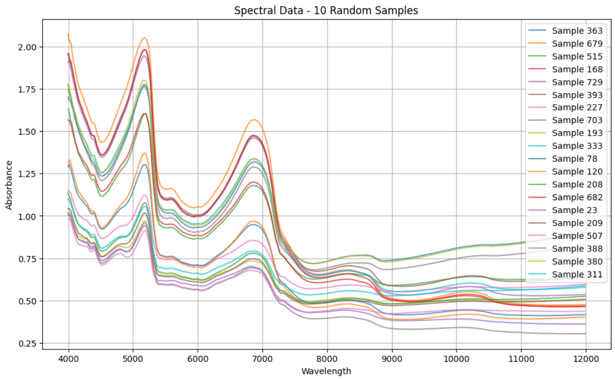

Fig. 1: Spectral data of 10 random samples.

In addition, the moisture content distribution in the dataset is visualized in Figure 2. The histogram of moisture values indicates a relatively bimodal distribution with some peaks around 30 and 60.

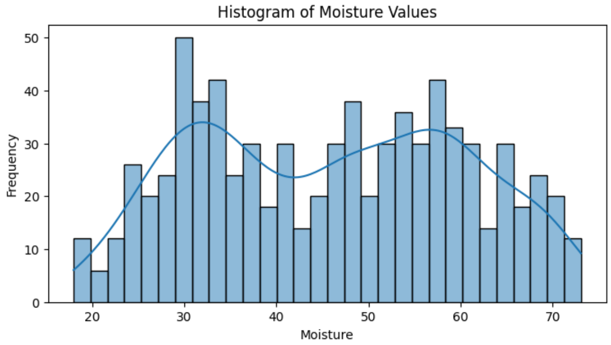

Fig. 2: Histogram of moisture values in the dataset.

Outlier detection was performed on the spectral data. Figure 3 shows boxplots of 10 randomly selected spectral columns. As can be observed, the middle lines of the boxes are generally closer to the lower edge, indicating a slight skew in the distribution, but no outliers were detected. After analyzing all the spectral columns, it was concluded that there were no columns exhibiting outliers.

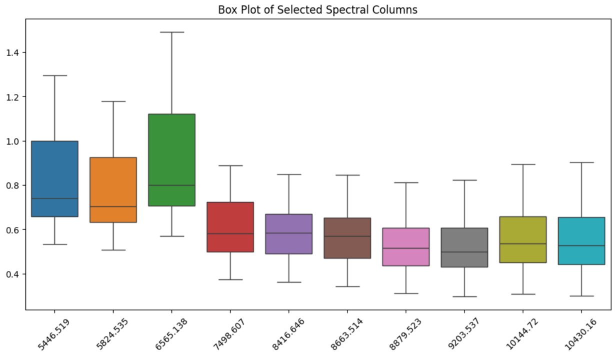

Fig. 3: Boxplots of 10 randomly selected spectral columns.

III. Pre-processing Techniques

The goal of preprocessing is to enhance the spectral data by reducing noise and correcting distortions, thus allowing the models to learn more effectively from the data. In this project, three different preprocessing techniques were applied to the spectral data to improve the accuracy and reliability of the machine learning models. These techniques—Savitzky-Golay filtering, Standard Normal Variate (SNV) transformation, and Multiplicative Scatter Correction (MSC)—were compared with each other and the raw data to evaluate their impact on model performance. The following subsections discuss each technique in detail, explaining their characteristics and why they were applied in this study.

A. Savitzky-Golay

The Savitzky-Golay filter is a well-known smoothing method used in chemometrics to reduce noise in spectral data while preserving important features. This method works by fitting a polynomial of a specified order to a moving window of data points, which smooths the data while maintaining sharp spectral features. It is often used in chemometrics due to its ability to smooth data without significantly altering the underlying patterns, making it ideal for spectral data analysis. Preprocessing with this technique helps improve the signal-to-noise ratio and enhances the ability of machine learning models to extract meaningful patterns from the data.

In this project, the Savitzky-Golay filter was applied with the following parameters: a window length of 21 and a polynomial order of 2. These parameters were selected to balance the reduction of noise and the preservation of key spectral features.

B. Standard Normal Variate (SNV)

The Standard Normal Variate (SNV) transformation is a normalization technique widely used in chemometrics to correct for multiplicative effects and baseline variations in spectral data. It works by subtracting the mean and dividing by the standard deviation for each individual spectrum, which results in a mean-centered, unit-variance data set.

SNV is used to remove variations caused by differences in sample size, shape, or instrumental effects, thus allowing for more consistent and comparable data. This normalization helps to focus the analysis on the true chemical variations present in the samples, which is crucial for improving the performance of machine learning models.

C. Multiplicative Scatter Correction (MSC)

The Multiplicative Scatter Correction (MSC) technique is used in chemometrics to remove scatter effects caused by variations in the physical properties of the samples, such as their size and shape. MSC works by first centering the spectra, then performing a linear transformation (scaling and shifting) to align the sample spectra with a reference spectrum, typically the mean spectrum of the entire dataset.

MSC is particularly useful for correcting data when there are multiplicative interferences or distortions that are not related to the chemical composition but are instead due to physical or instrumental factors. By applying MSC, signal-to-noise ratio is enhanced, making it easier to identify the chemical properties of the sample and improving the performance of predictive models.

D. Preprocessing Comparison

The graph in Figure 4 shows the raw data along with the preprocessed data using the three techniques for one of the rows (row 12) in the dataset. As observed, each of the preprocessing techniques modifies the spectral distribution differently.

The Savitzky-Golay filter smoothens the spectral data, reducing high-frequency noise while preserving the shape of the spectrum. The SNV technique, on the other hand, has a significant effect, notably standardizing the data by removing multiplicative scatter effects, which results in a more uniform baseline for the spectra. The MSC approach, similar to SNV, aims to remove scatter effects but with a different approach. Overall, each preprocessing technique significantly impacts the spectral data distribution, with SNV showing the most noticeable change.

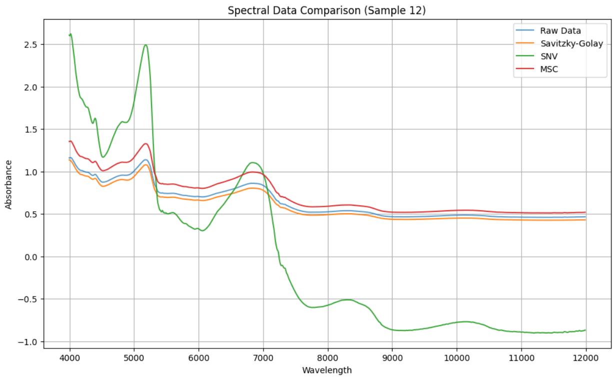

Fig. 4: Spectral data comparison for sample 12 with raw data and preprocessing techniques (Savitzky-Golay, SNV, MSC).

IV. Modeling

Three different machine learning approaches were used to model the biomass moisture content: Partial Least Squares (PLS), Support Vector Regression (SVR), and Artificial Neural Networks (ANN).

To ensure consistency across all datasets, each of the four datasets (raw data and three preprocessed datasets) was split into training and testing sets in an 80/20 ratio. This resulted in 618 samples in the training set and 155 samples in the test set for each dataset.

A. Partial Least Squares (PLS)

First modeling approach, Partial Least Squares (PLS), is a multivariate statistical method that combines principal component analysis (PCA) and multiple linear regression. It is especially useful when predictors are highly collinear or when the number of predictors exceeds the number of observations, which is common in spectral data analysis. PLS extracts latent variables (components) from the predictors that explain the most variance in both the predictors and the response variable, which are then used for regression.

PLS is widely used in chemometrics, spectroscopy, and other fields involving high-dimensional data with multicollinearity. It is effective in scenarios where traditional regression models struggle, and it is computationally efficient. One of its key benefits is its ability to handle multicollinearity and provide interpretable components for exploratory data analysis.

B. Support Vector Regression (SVR)

Support Vector Regression (SVR) is a machine learning algorithm derived from Support Vector Machines (SVM). It aims to find a hyperplane that best fits the data, balancing the complexity of the model with the ability to generalize to unseen data. SVR uses a kernel trick to map input data into a higher-dimensional space, allowing it to efficiently handle non-linear relationships between variables.

SVR is commonly used in regression problems, especially when the relationship between input and output variables is complex or non-linear. Its key advantage lies in its robustness to overfitting, especially in high-dimensional spaces, and its ability to provide a non-linear decision boundary, making it a strong candidate for tasks involving noisy or complex datasets.

C. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) are computational models inspired by the human brain, consisting of layers of interconnected nodes (neurons). ANN models learn to map inputs to outputs through training on a dataset, adjusting the weights of connections between nodes to minimize the error between predicted and actual values. The architecture and training process allow ANNs to capture complex, non-linear relationships in the data.

ANNs are widely used in tasks such as pattern recognition, image processing, and time-series forecasting. Their ability to approximate complex functions makes them highly effective for large, unstructured datasets. However, they require considerable computational resources and data to avoid overfitting and to ensure generalization to unseen data.

V. Results

Three machine learning models—PLS, SVR, and ANN—were trained on the four datasets: Raw Data, Savitzky-Golay Preprocessed Data, SNV Preprocessed Data, and MSC Preprocessed Data. First, the results of each model will be discussed individually, followed by a general comparison based on RMSE and R² values.

A. PLS Results

A 5-fold cross-validation was applied for the PLS model. A graph showing the optimal number of components versus RMSE and R² is provided below in Figure 5. As observed in the plot, the optimal number of components varies based on the dataset, but it ranges between 12 and 17 for the best performance.

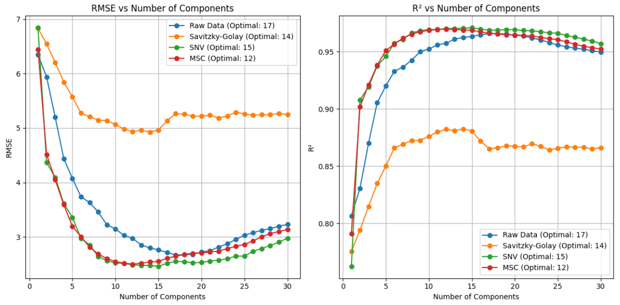

Fig. 5: Optimal Number of Components vs RMSE and R² for PLS

Next, Figure 6 displays the training and test RMSE values for each dataset. It can be observed that SNV preprocessing is slightly better with the RMSE value of 2.1282 than both MSC (2.2147) and raw data (2.5669), while the Savitzky-Golay data shows a significantly worse performance compared to the other datasets with RMSE = 4.5300.

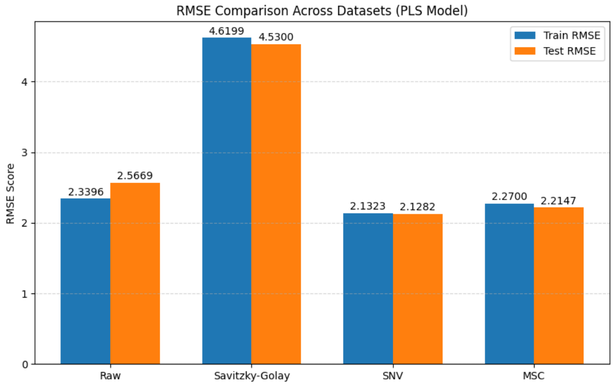

Fig. 6: Training and Test RMSE for Each Dataset in PLS

The scatter plots in Figure 7 illustrate the predicted versus actual values for the training and test sets of the best model, SNV with 15 components. The model demonstrates good performance, with most points closely following the ideal line, indicating accurate predictions.

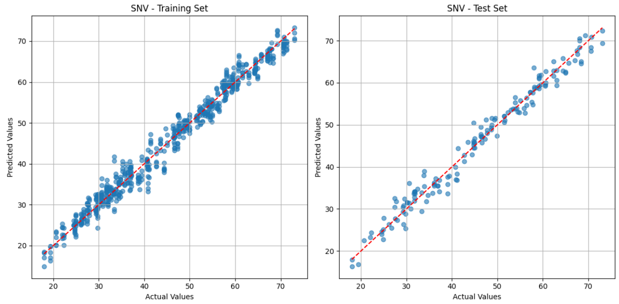

Fig. 7: Predicted vs Actual Values for PLS (Training and Test Sets)

Finally, Figure 8 presents the residual analysis for the SNV dataset on the training and test sets. The residuals show no discernible trends or patterns, suggesting that the model has successfully captured the underlying data relationships and that the errors are random and evenly distributed.

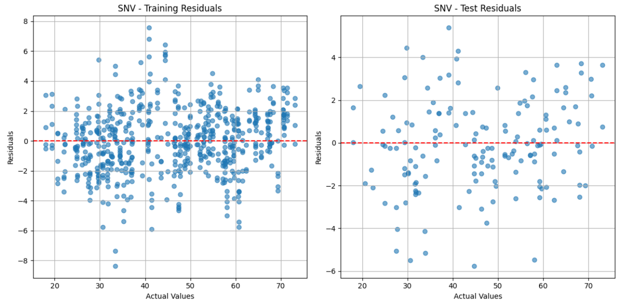

Fig. 8: Residual Analysis for PLS on SNV Dataset (Training and Test Sets)

B. SVR Results

The hyperparameter tuning for the Support Vector Regression (SVR) model was performed using a 5-fold cross-validation. The following parameter grid was used for tuning:

Hyperparameter	Values
C	0.1, 1, 10, 100
ϵ	0.01, 0.1, 0.5
Kernel	linear, rbf

TABLE II: Parameter grid used for hyperparameter tuning in SVR.

The best hyperparameters and corresponding cross-validation RMSE for each dataset are shown in the table below:

Dataset	Best Hyperparameters	Best CV RMSE
Raw Data	{C: 100, ϵ: 0.5, kernel: 'linear'}	3.3602
Savitzky-Golay Data	{C: 100, ϵ: 0.5, kernel: 'linear'}	5.1625
SNV Data	{C: 100, ϵ: 0.5, kernel: 'linear'}	2.3953
MSC Data	{C: 100, ϵ: 0.5, kernel: 'linear'}	2.6520

TABLE III: SVR Model: Best hyperparameters and cross-validation RMSE for each dataset.

Figure 9 below displays the R^2 performance for the SVR model across the datasets. Similar to PLS, this model also has SNV dataset as the best performing option with R^2 = 0.9750. It is followed by MSC and raw dataset, and again, Savitzky-Golay is behind them.

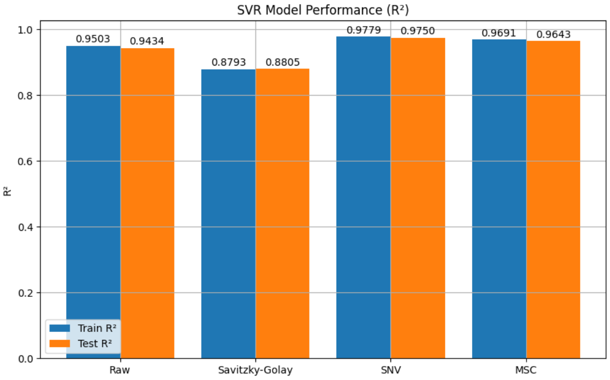

Fig. 9: SVR Model Performance (R²) across different datasets.

The predicted vs actual graph and the residuals plot for the SVR model are similar to those observed for the PLS model. As such, these plots are omitted for brevity, as they exhibit the same trends in terms of model performance. The residuals for both the training and test sets show no significant patterns, indicating that the model has effectively captured the underlying relationships between the features and the target variable.

C. ANN Results

The Artificial Neural Network (ANN) model was built using a feedforward architecture with two hidden layers. The model was compiled using the Adam optimizer with a learning rate of 0.001 and trained for 100 epochs with a batch size of 32. Early stopping with a patience of 10 was applied to prevent overfitting. The following architecture, shown in Figure 10 was used:

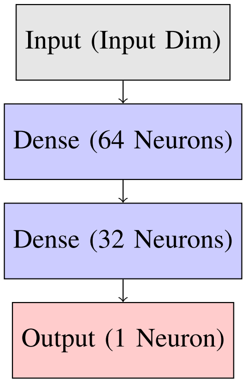

Fig. 10: Visualization of the ANN model architecture.

The loss history for the four datasets is visualized in Figure 11. For the MSC dataset, the loss steadily decreased and converged towards the other datasets around epoch 60. The Raw and Savitzky-Golay datasets, on the other hand, experienced a significant reduction in loss within the first 5 epochs, after which there was little further decrease in loss. While SNV showed a slower decrease compared to Raw and MSC, it eventually achieved the best performance, mirroring the trends observed with the previous models.

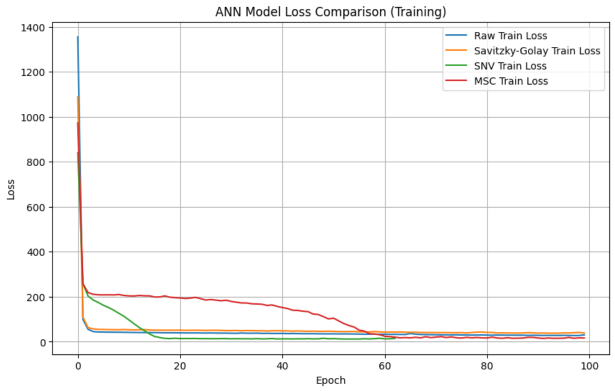

Fig. 11: Loss history of ANN model for the four datasets

The results are shown in a more detailed way in Table IV below. The table includes both the RMSE and R^2 for both the training and test sets.

Dataset	Train RMSE	Test RMSE	Train R^2	Test R^2
Raw Data	5.2374	5.3287	0.8691	0.8598
Savitzky-Golay	6.1360	5.9108	0.8204	0.8275
SNV	3.4502	3.6882	0.9432	0.9329
MSC	3.7922	3.9102	0.9314	0.9245

TABLE IV: ANN model performance for each dataset

From the table, it can be observed that the model performs the best on SNV data, achieving the lowest RMSE and the highest R^2 values for both the training and test sets. On the other hand, the Savitzky-Golay data exhibits the highest RMSE and lowest R^2 values, indicating that the model struggles with this preprocessed data. The Raw and MSC datasets show moderate performance, with MSC performing slightly better.

D. Final Comparison

The final comparison of model performance is shown in Figure 12, which presents the RMSE values for each of the datasets across the three models (PLS, SVR, and ANN). A total of 12 RMSE values are displayed, with each dataset having three corresponding values, one for each model.

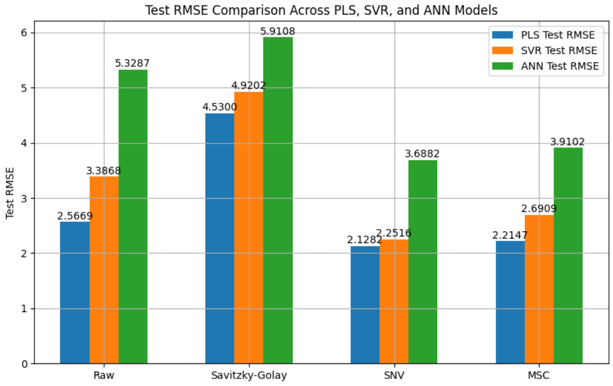

Fig. 12: Comparison of RMSE values across models and datasets.

Similarly, Figure 13 below shows the comparison of R^2 values of 12 dataset-model pairs.

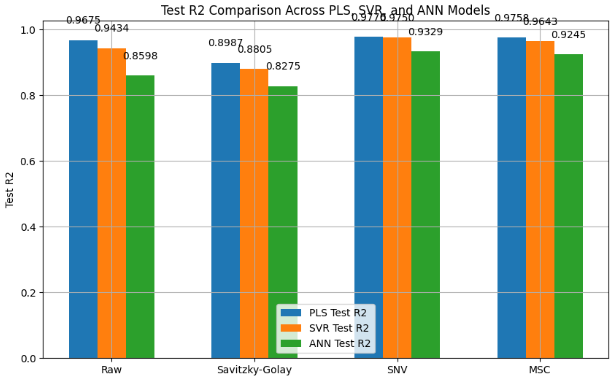

Fig. 13: Comparison of R^2 values across models and datasets

Results show PLS model performed the best overall, while SNV was found to be the most effective preprocessing technique. The best combination, PLS with SNV, achieved an RMSE of 2.1282 and an R^2 of 0.9776, making it the top-performing model and preprocessing pair. For a more detailed comparison, the following tables, Table V and Table VI, present the test RMSE and R^2 values for all twelve combinations of model and preprocessing technique.

Dataset	PLS RMSE	SVR RMSE	ANN RMSE
Raw	2.5669	3.3868	5.2469
Savitzky-Golay	4.5300	4.9202	5.9361
SNV	2.1282	2.2516	3.4060
MSC	2.2147	2.6909	3.9931

TABLE V: Test RMSE values for PLS, SVR, and ANN across four preprocessing techniques.

Dataset	PLS R²	SVR R²	ANN R²
Raw	0.9675	0.9434	0.8641
Savitzky-Golay	0.8987	0.8805	0.8261
SNV	0.9776	0.9750	0.9427
MSC	0.9758	0.9643	0.9213

TABLE VI: Test R² values for PLS, SVR, and ANN across four preprocessing techniques.

VI. Conclusion

In this project, machine learning approaches were employed to predict the moisture content of biomass using Near-Infrared (NIR) spectroscopy. The dataset comprised 773 samples with 1037 spectral data points and corresponding moisture values. The primary goal was to evaluate and compare the performance of three machine learning models—Partial Least Squares (PLS), Support Vector Regression (SVR), and Artificial Neural Networks (ANN)—on raw and preprocessed data. Various preprocessing techniques, including Savitzky-Golay smoothing, Standard Normal Variate (SNV), and Multiplicative Scatter Correction (MSC), were applied to the data to improve model performance.

All twelve model-dataset pairs are ranked based on their RMSE and R^2 scores in Table VII below.

Model	Dataset	Test RMSE	Test R^2
PLS	SNV	2.1282	0.9776
PLS	MSC	2.2147	0.9758
SVR	SNV	2.2516	0.9750
PLS	Raw	2.5669	0.9675
SVR	MSC	2.6909	0.9643
SVR	Raw	3.3868	0.9434
ANN	SNV	3.4060	0.9427
ANN	MSC	3.9931	0.9213
PLS	Savitzky-Golay	4.5300	0.8987
SVR	Savitzky-Golay	4.9202	0.8805
ANN	Raw	5.2469	0.8641
ANN	Savitzky-Golay	5.9361	0.8261

TABLE VII: Ranking of model-data pairs by

The results indicated that the PLS model combined with SNV preprocessing performed the best, achieving the lowest RMSE and highest R² score. Among the preprocessing techniques, SNV provided the best preprocessing technique, significantly improving model performance over the raw data, with MSC being the second-best option. On the other hand, Savitzky-Golay smoothing resulted in worse performance compared to the raw data, likely due to its over-smoothing effect, which reduced the variability in the spectral data.

Among the models, PLS proved to be the best-performing algorithms. This could be attributed to PLS's ability to handle multicollinearity in spectral data effectively. SVR also performed well, though not as effectively as PLS, likely due to its sensitivity to hyperparameters and the complexity of the dataset. ANN, despite its flexibility and ability to capture non-linear relationships, performed poorly in comparison, likely due to the lack of sufficient data for proper generalization and potential overfitting.

Several bottlenecks were encountered during this project. One key limitation was the relatively small dataset, which may have impacted the performance of more complex models like ANN. Increasing the dataset size could potentially improve the performance of such models. The Savitzky-Golay filter underperformed compared to other preprocessing techniques, and further exploration of its parameters or alternative techniques could yield better results. Hyperparameter tuning for models, especially SVR and ANN, was limited, and more exhaustive techniques like grid search could improve model performance. Lastly, feature selection and dimensionality reduction were not explored but could enhance model efficiency and accuracy. Future work could focus on these aspects for improved performance.

VII. Accessing the Project

The work for this project can be found in notebook. To reproduce the results, the dataset should be uploaded to a personal Google Drive and the variable file_path should be updated accordingly.