Biomass Characterization through NIR Spectra

GALBOKKA HEWAGE ISHARA MADHAVI

I. INTRODUCTION

The determination of moisture content in biomass is a critical factor in optimizing its energy efficiency and ensuring quality control. This study employs machine-learning techniques to analyze Near-Infrared (NIR) spectral data for accurate moisture prediction. Traditional laboratory methods for moisture determination are time-consuming and expensive, whereas NIR spectroscopy, combined with advanced computational models, offers a rapid, non-destructive alternative.

This research applies chemometric preprocessing techniques and machine learning models, including Partial Least Squares (PLS), Support Vector Regression (SVR), and Artificial Neural Networks (ANN), to evaluate their predictive accuracy and robustness.

II. DATA ANALYSIS

As the initial step for developing Machine Learning models, an extensive data analysis step was undertaken.

1. Dataset Overview

The dataset consists of NIR spectral data and reference moisture content measurements collected from various biomass samples, including pine and spruce wood chips, bark, forest residues, and sawdust. The samples were processed through a Fourier-transform NIR (FT-NIR) spectrometer while moving at a velocity of 1 m/s.

The dataset includes:

- **Feature Vectors**: The spectral data consists of multiple columns representing absorption values at different wavelengths (1037). There exists 125 samples and each undergoes the measurement process around 5-7 times, creating a final sample set of 773.
- **Wavelength Range**: The features correspond to wavelengths ranging from 834 nm to 2500 nm (12000 – 4000 cm^{-1}).
- **Removed Columns**: The dataset initially contained an unnecessary 'Sample ID' column and an unnamed column, which were removed to focus on spectral data.
- **Target Variable**: Moisture content, which serves as the dependent variable for model training and evaluation.

2. Exploratory Data Analysis (EDA)

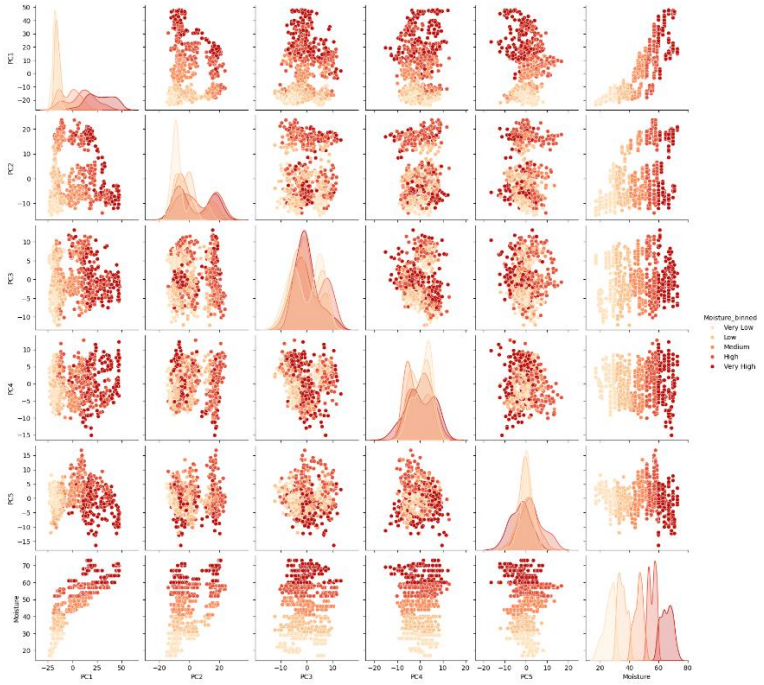

Figure 1. Pair-plot of Principal Components and Target Variable

To better understand the relationships between the spectral features and moisture content, an Exploratory Data Analysis (EDA) was conducted. One of the key visualization tools used was a pairplot, which allows for examining the distribution and relationships between principal components derived from the spectral data and the target moisture content, as illustrated in Figure 1.

Key Observations from the EDA:

- A clear separation in moisture content levels was observed across different principal components, indicating that moisture content influences spectral variations significantly.
- PC1 showcased the highest correlation with the target variable, Moisture due to the clear upward trend.
- The presence of overlapping clusters suggests that while spectral data is useful for moisture prediction, some preprocessing and dimensionality reduction are necessary to enhance separation.
- Some spectra exhibited extreme deviations from the general trend, which warranted **outlier detection and removal using PLS Q-residuals and Hotelling's T-squared statistics**.

3. Issues with the Dataset

- **Noise and Variability**: Raw NIR spectra contains unwanted noise due to environmental and instrumental factors.

- **Baseline Drift**: Variations in instrument response can lead to baseline shifts.
- **Overlapping Spectral Peaks**: The complex chemical composition of biomass leads to overlapping peaks, making it difficult to extract moisture-specific information.
- **Outliers**: Some spectra may deviate significantly due to measurement errors or sample inconsistencies.

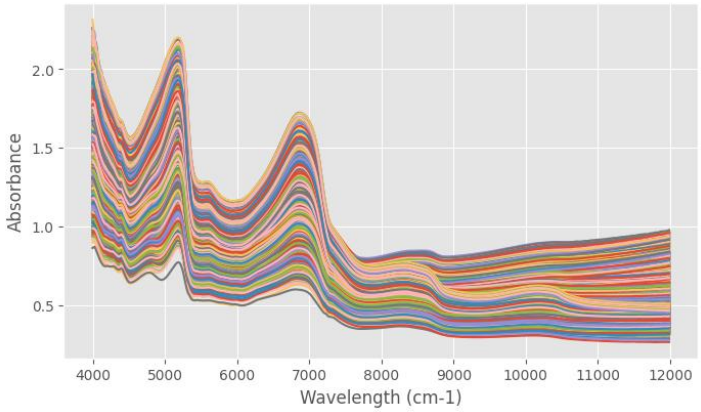

Figure 2. Original distribution of Absorbance values across the range of wavelengths

III. DATA PREPROCESSING

Preprocessing is crucial to remove noise, correct baseline drift, and enhance the predictive power of ML models. The following preprocessing techniques were tested, and Standard Scaler was used to scale feature values after every preprocessing technique.

1. Savitzky-Golay (SV-GOL) Smoothing

Characteristics: Savitzky-Golay smoothing is a polynomial filtering technique used to reduce noise while preserving the original shape and features of spectral data. Unlike simple moving average filters, it fits a low-degree polynomial to local segments of data (window), which helps retain peaks and fine spectral details [2].

Reason for Usage: This technique is essential in spectral preprocessing because it smooths out unwanted fluctuations that arise due to random noise while ensuring that valuable chemical information is preserved. This is particularly beneficial in NIR spectroscopy, where spectral noise can obscure small but significant variations in absorbance that relate to moisture content.

As displayed in Figure 2, the original absorbance value distribution is noisy, and by applying different SV-Gol filters with varying window sizes and polynomials, various degrees of smoothness in data were achieved.

As illustrated in Figure 3 and Figure 4, a section of the wavelength range is plotted, and the below smoothing patterns were observed.

- The blue line is the original spectrum.
- The red line shows a mild smoothing
- The green line represents more aggressive smoothing
- The magenta line is a more optimal choice of parameters.

Therefore, by analyzing these variations, the optimal window size and polynomial values were selected as 21 and 6 respectively to be applied in the SV-Gol smoothing filter as shown on Figure 5.

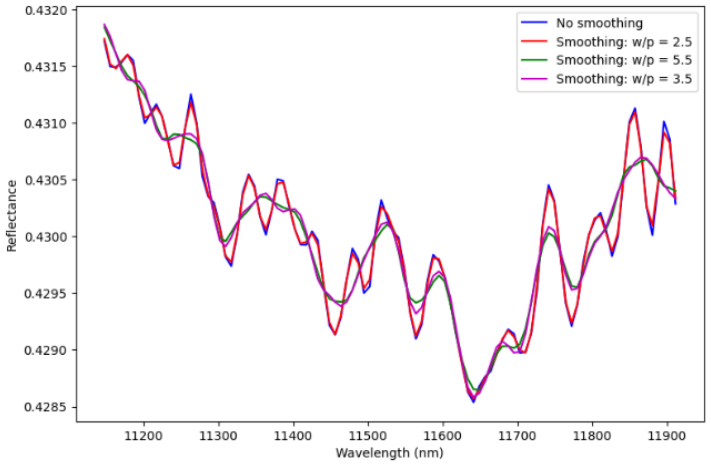

Figure 3. Comparison of smoothing abilities of different SV-GOL filters

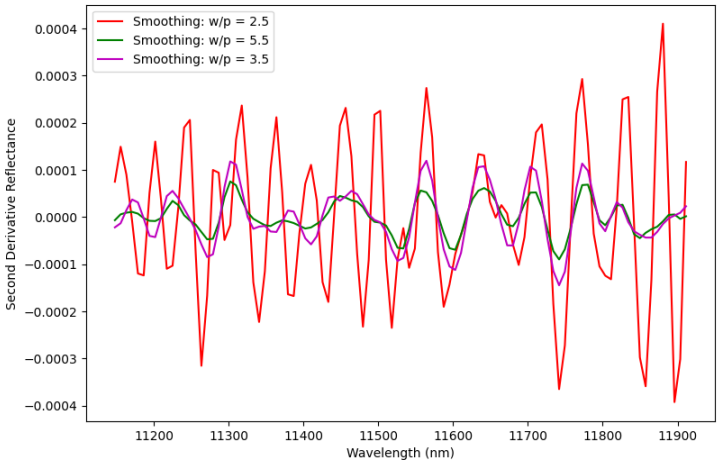

Figure 4. Comparison of smoothing abilities of different SV_GOL filters applied on the second derivatives of data

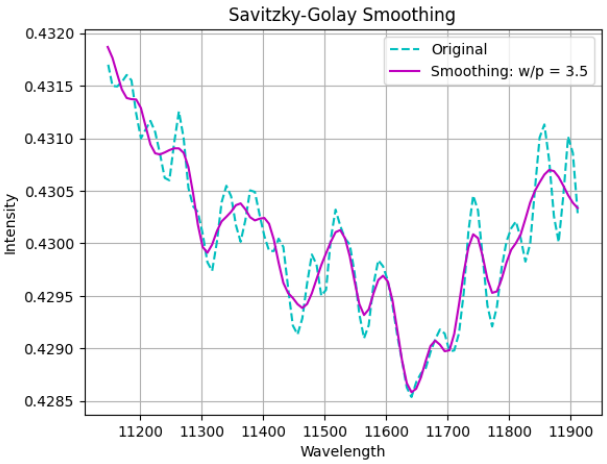

Figure 5. SV-GOL smoothing (window size 21 and polynomial 6)

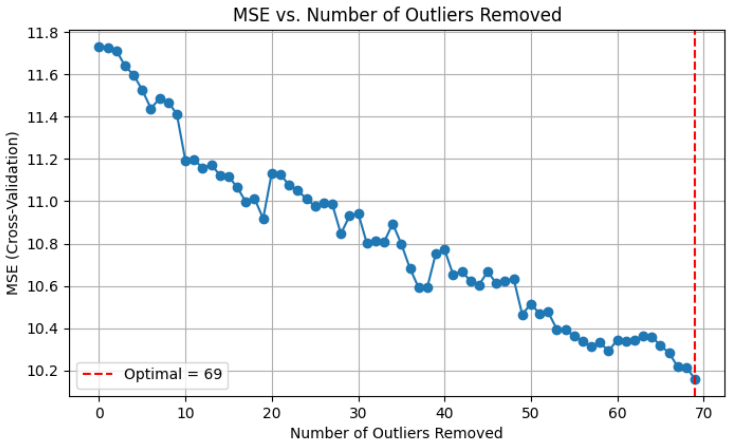

Figure 7. Technique to select the optimal count of outliers to be removed, based on MSE values from PLS Regression

2. Outlier Removal using PLS Regression

Reason for Usage: Outliers can distort the calibration and prediction accuracy of machine-learning models. In this study, Partial Least Squares (PLS) regression combined with two key statistical measures—Q-residuals and Hotelling's T-squared—were used to systematically detect and remove outliers. The outliers are detected using a 95% confidence interval and a scatter Q-residuals VS Hotelling's T-squared was created which marks the position of the confidence level in both axes as shown in Figure 6.

- **Q-Residuals Characteristics**: measure how well each observation is explained by the calibration model. Large Q-residuals indicate spectral points that contain noise or anomalies not captured by the model.
- **Hotelling's T-Squared Characteristics:** identifies deviations within the model space, pinpointing spectra that significantly differ from the majority of the dataset.

To optimize outlier removal, an iterative approach was implemented: one outlier was removed at a time, up to a maximum threshold, and the model's mean squared error (MSE) was monitored as illustrated in Figure 7. The process was stopped when MSE reached its minimum, ensuring that only genuine outliers were discarded without over-filtering useful data [3].

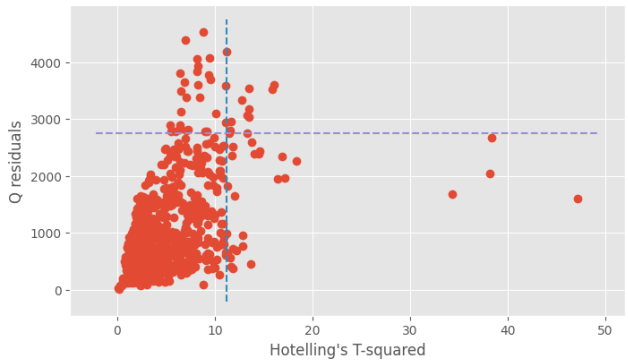

Figure 6. Outlier detection using a confidence interval of 95%

3. Scatter Correction Techniques

Scattering effects in NIR spectroscopy arise due to variations in sample particle size, measurement geometry, and differences in path length. These effects can introduce baseline shifts and distort spectral features, making it difficult to extract meaningful chemical information. Scatter correction techniques are designed to eliminate these effects, improving model robustness and accuracy [1].

3.1 SNV

Characteristics: Standard Normal Variate (SNV) is a normalization technique applied to individual spectra to remove additive and multiplicative scatter effects. It works by transforming each spectrum to have a mean value of zero and a standard deviation of one.

Reason for Usage: SNV correction is particularly useful in cases where scattering causes spectral variations unrelated to moisture content. By standardizing spectra, SNV ensures that any differences in absorbance are primarily due to chemical composition rather than physical variations in the sample.

Figure 8 shows the scatter corrected absorbance value distribution, after applying SV-GOL smoothing and outlier removal from the original data.

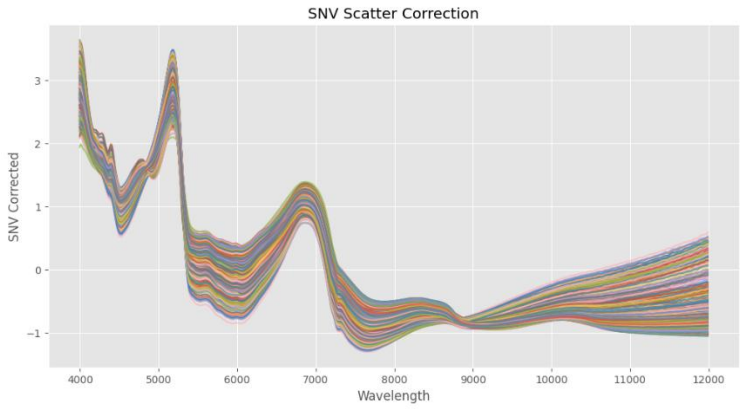

Figure 8. Scatter Correction with SNV, after SV-GOL smoothing and Outlier Removal

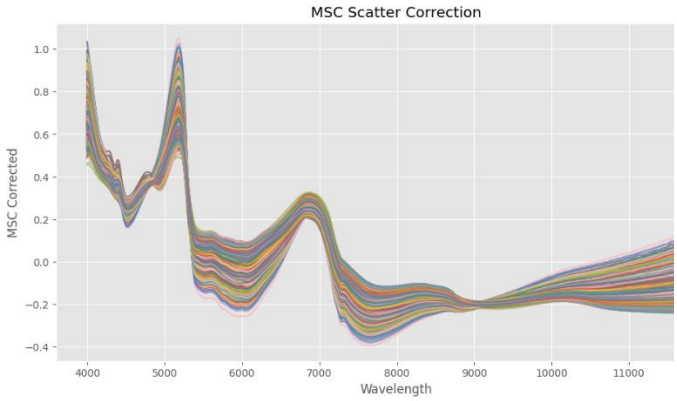

Figure 9. Scatter Correction with MSC, after SV-GOL smoothing and Outlier Removal

3.2 MSC

Characteristics: Multiplicative Scatter Correction (MSC) is another widely used method for correcting spectral distortions caused by scattering effects. In MSC, each spectrum is adjusted to match a reference spectrum, compensating for path length differences and particle-size variations.

Reason for Usage: MSC is effective in reducing the impact of sample heterogeneity on spectral data. By aligning spectra to a common reference, it improves model performance by ensuring that the differences observed are due to the analyte of interest rather than external scattering effects.

Figure 9 shows the scatter corrected absorbance value distribution from MSC, after applying SV-GOL smoothing and outlier removal from the original data. The distribution shows a much smoother variation where the data scatter effect is reduced when compared to the original distribution in Figure 2. One of the advantages of Multiplicative Scatter Correction (MSC) is its ability to relate all spectra to a common reference. If the reference spectrum is close to a spectrum free from unwanted scattering, MSC serves as an effective correction method. However, if outliers are present, the mean spectrum may not accurately represent the reference, making MSC less reliable. In such cases, Standard Normal Variate (SNV) can be a better alternative, as it does not rely on a single reference spectrum and can effectively handle variations caused by outliers.

IV. DATA MODELLING

The study employs three machine-learning models: PLS [4], SVR [5], and ANN [6]. Each model has distinct characteristics suited for chemometric analysis.

1. Partial Least Squares Regression (PLS)

- **Characteristics**:
  - Projects spectral data onto a lower-dimensional space.
  - Captures covariance between spectral features and target moisture content.
- **Usage Scenarios**:
  - Effective for high-dimensional datasets with collinear variables.
  - Suitable for chemical and agricultural applications.
- **Benefits**:
  - Reduces dimensionality.
  - Robust against noisy spectral data.

2. Support Vector Regression (SVR)

- **Characteristics**:
  - Constructs a hyperplane that best fits the data in a transformed feature space.
  - Uses kernel functions (linear, RBF) for nonlinear regression.
- **Usage Scenarios**:
  - Ideal for datasets where relationships are nonlinear.
  - Effective when the dataset has high variance.
- **Benefits**:
  - Handles nonlinear relationships effectively.
  - Robust against overfitting, especially with small datasets.

3. Artificial Neural Networks (ANN)

- **Characteristics**:
  - ANN consists of multiple layers of neurons to model complex relationships.
  - Learns hierarchical representations from spectral data.
- **Usage Scenarios**:
  - Best for large datasets with nonlinear dependencies.
  - Suitable for feature extraction from high-dimensional data.
- **Benefits**:
  - High adaptability to complex spectral patterns.
  - Potential for superior accuracy with sufficient training data.

V. RESULTS

As indicated in Figure 10, Figure 11, and Figure 12, SV-Gol smoothing, Outlier removal technique, and the Advanced processing techniques (i.e. SNV and MSC Scatter Correction), incrementally improves the model performance by reducing the cross-validated RMSE value and Bias while increasing the R^2 score.

As illustrated in Figure 13, the performance of different preprocessing techniques and machine learning models was measured using **5-folds cross validation**, evaluating the key metrics: **R^2** (coefficient of determination), **RMSE** (Root Mean Square Error), and **Bias**. These metrics provide insights into model accuracy, generalization, and systematic deviations.

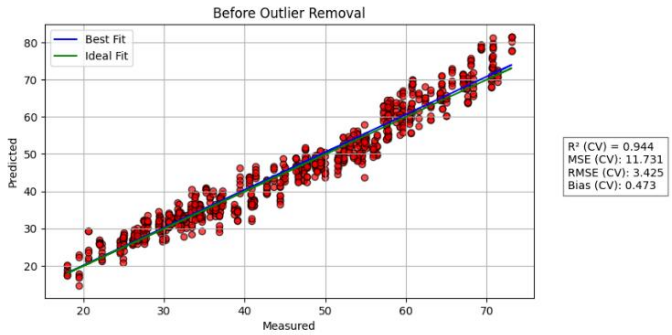

Figure 10. PLS Model performance before removing outliers

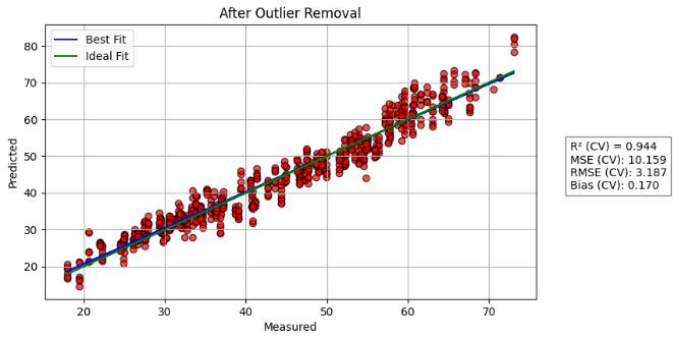

Figure 11. PLS Model performance after SV-Gol smoothing and removing outliers

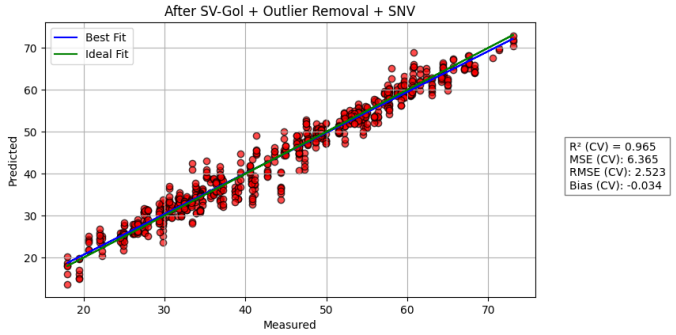

Figure 12. PLS Model performance after SV-Gol smoothing, removing outliers and SNV Scatter correction

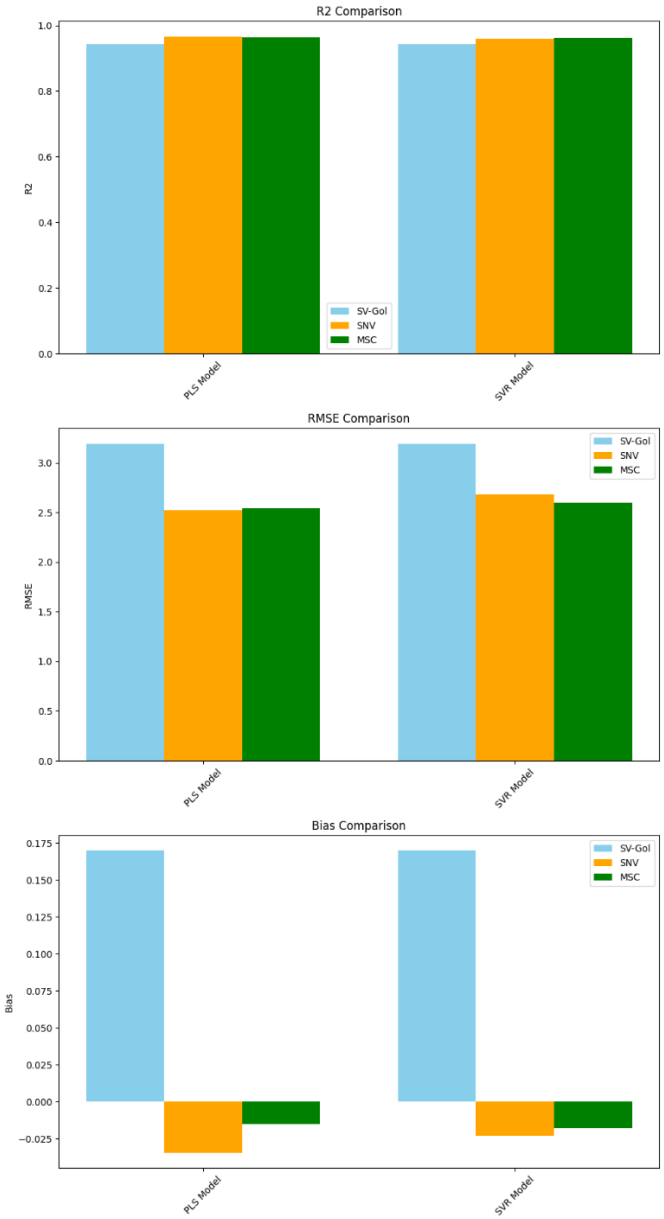

Figure 13. Performance comparison of PLS vs SVR models in terms of R2, RMSE and Bias

1. *R^2 Comparison*

The R^2 values across models and preprocessing techniques indicate how well the models explain the variance in moisture content. Higher R^2 values suggest better predictive capability.

- The PLS and SVR models both exhibited high R^2 values across all preprocessing methods.
- SV-Gol preprocessing (smoothing and outlier removal using PLS regression) resulted in slightly lower R^2 compared to SNV and MSC, suggesting that while noise reduction and outlier removal are helpful, additional scatter correction further improves prediction accuracy.
- **Applying SNV and MSC after SV-Gol increased R^2**, with SNV slightly outperforming MSC. This indicates that scatter correction techniques enhance spectral feature extraction for moisture prediction.

2. *RMSE Comparison*

The RMSE values indicate the average error between predicted and actual moisture content.

- SV-Gol alone resulted in the highest RMSE, showing that while it removes noise and outliers, some spectral distortions remain.

- **Applying SNV or MSC after SV-Gol reduced RMSE significantly**, with **MSC achieving the lowest RMSE across models**. This highlights the importance of scatter correction in improving predictive accuracy.

- PLS and SVR showed comparable RMSE trends, reinforcing the effectiveness of SNV and MSC in enhancing model robustness.

3. Bias Comparison

Bias represents systematic errors in model predictions.

- SV-Gol preprocessing alone exhibited the highest bias, suggesting that while smoothing and outlier removal improve predictions, residual scattering effects introduce small but systematic deviations.

- **Applying SNV and MSC after SV-Gol effectively reduced bias**, with MSC showing slightly lower bias than SNV.

- PLS and SVR models performed similarly in terms of bias reduction, indicating that scatter correction plays a more significant role than the choice of regression model.

VI. CONCLUSION

1. Results achieved

This study successfully demonstrated that **machine learning models can accurately predict biomass moisture content using NIR spectroscopy by using various preprocessing techniques**.

By evaluating different preprocessing techniques and regression models, key processing strategies that influence model accuracy and generalization were identified.

2. Effectiveness of processing techniques and models

- Basic preprocessing with **Savitzky-Golay (SV-Gol) smoothing** and **outlier removal** using PLS regression was essential in removing noise and enhancing feature clarity.

- **Advanced preprocessing techniques** such as **MSC and SNV significantly improved model accuracy** by addressing scattering effects.

- **PLS regression provided the best accuracy** across all models, as indicated by its highest **R^2 of 0.965**, **lowest RMSE of 2.523**. While SVR also showed competitive performance, PLS remained the most robust and reliable choice for this dataset.

- Among the Advanced preprocessing techniques, **MSC Scatter Correction showed slightly better results,** with the lower bias, RMSE and higher R^2 values. The potential reasons for this observation could be as follows.

    - MSC aligns spectra to a chosen reference, which enhances the consistency of the data and improves correlation with true chemical concentrations.

    - By modeling and removing scattering effects through linear regression, MSC retains the original spectral characteristics better than SNV. In contrast, SNV normalizes each spectrum individually, which can distort spectral relationships and increase variability.

3. Bottlenecks and future improvements

The project was challenging in terms of the following aspects.

- Spectral noise and overlapping peaks in biomass samples made feature extraction complex. To overcome this challenge, a set of basic and advanced preprocessing techniques were applied, and it required careful curation of the most suitable processing technique.

- Outlier detection was necessary but required careful tuning to avoid removing relevant data. A strategic approach of fine-tuning the optimal number of outliers, based on the reported error values was utilized as a result.

- Computational demands of advanced ML models like ANN required careful hyperparameter optimization and the process was time-consuming.

The points below remain to be addressed as potential future improvements to this study, to make better predictions.

- Hyperparameter tuning can be performed to retrieve better results.

- Hybrid preprocessing approaches combining multiple correction techniques could be explored.

- Deep learning methods such as convolutional neural networks (CNNs) could be tested for automated feature extraction.

- Expanding the dataset with more biomass types and environmental conditions could further validate model robustness.

REFERENCES

[1] Two scatter correction techniques for NIR spectroscopy in Python. Available at: https://nirpyresearch.com/two-scatter-correction-techniques-nir-spectroscopy-python/.

[2] Savitzky–Golay Smoothing Method • Nirpy Research. Available at: https://nirpyresearch.com/savitzky-golay-smoothing-method/

[3] Outliers detection • nirpy research. Available at: https://nirpyresearch.com/data-operations-plotting/outliers-detection/.

[4] Partial least squares regression (2025) Wikipedia. Available at: https://en.wikipedia.org/wiki/Partial_least_squares_regression.

[5] Application of near infrared spectroscopy combined with SVR algorithm in rapid detection of camp content in red jujube, Optik. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0030402619309404.

[6] Ali, H. et al. (2023) Machine learning–enabled NIR spectroscopy. part 2: Workflow for selecting a subset of samples from publicly accessible data - AAPS pharmscitech, SpringerLink. Available at: https://link.springer.com/article/10.1208/s12249-022-02493-5.