Biomass Characterization through NIR Spectroscopy

Lameya Islam[*]

[*] Faculty of Science & Engineering, Åbo Akademi University

1. INTRODUCTION

Chemometric analysis, particularly using near-infrared (NIR) spectroscopy, has gained prominence in material and quality assessment fields. This study aims to predict moisture content from spectral data using machine learning models. The primary challenge is handling high-dimensional spectral data with potential noise and correlations. This work evaluates the performance of three machine learning models—**Partial Least Squares Regression (PLSR), Support Vector Regression (SVR), and Artificial Neural Networks (ANN)**—under different preprocessing techniques.

2. METHODOLOGY

2.1 DATASET DESCRIPTION

The dataset consists of NIR spectral data collected from 125 biomass samples, including pine and spruce wood chips, bark, forest residues, and sawdust. The key dataset parameters are: Spectral range, Spectral resolution, Number of data points per spectrum, Measurements per sample.

The dataset used consists of spectral readings with **features representing wavelengths** and a **target variable—moisture content**.

But this type of dataset may lead to some challenges while performing different analysis on them. Key challenges include:

- **High Dimensionality**: Spectral data contain hundreds of wavelength values.
- **Collinearity**: Adjacent wavelengths are highly correlated.
- **Noise**: Raw spectral data include measurement artifacts.

2.2 DATA ANALYSIS

The dataset was loaded and preprocessed by removing unnecessary columns, handling missing values, and identifying outliers. A boxplot analysis of the moisture content was performed to detect anomalies. Outliers in spectral data were identified using the z-score method, where values exceeding a threshold of 3 standard deviations were flagged.

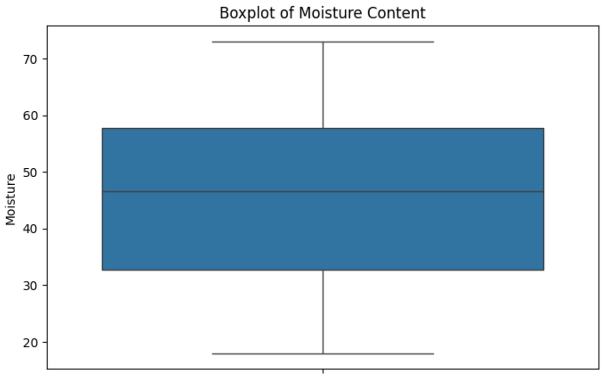

Figure 1: Outlier Analysis

The boxplot indicates that the median moisture content is approximately 45%, with values ranging from around 18% to 72%. The distribution appears fairly symmetrical, as the median is centered within the interquartile range, suggesting a relatively balanced spread of moisture content in the dataset.

2.3 DATA PREPROCESSING

The following preprocessing techniques were applied:

- **Savitzky-Golay Smoothing:** This technique applies a polynomial filter to the spectral data, reducing random noise while preserving important spectral features such as peaks and valleys. It ensures that the true underlying trends in the data are maintained without excessive distortion, leading to improved model accuracy.

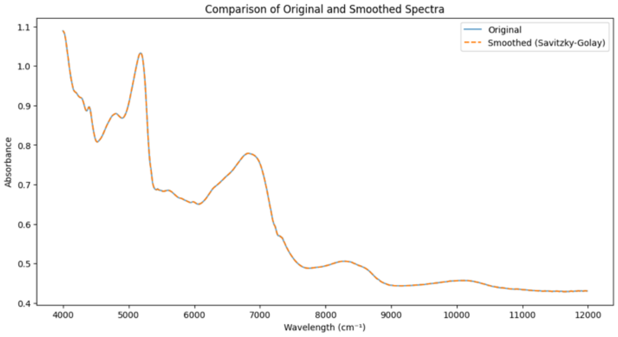

Figure 2: Comparison of Original and Smoothed Spectra

The plot compares the original and smoothed spectra, where the Savitzky-Golay filter is applied to reduce noise while preserving key spectral features. The original spectrum (blue) shows fluctuations due to noise, while the smoothed spectrum (orange dashed line) closely follows the original, highlighting a reduction in noise without distorting peak shapes. This smoothing technique enhances data quality, making it useful for spectroscopy and analytical applications.

- **Standard Normal Variate (SNV):** SNV is a normalization technique used to correct for scatter effects in spectral data. Variations in particle size, surface roughness, and sample density can cause inconsistencies in spectral readings. SNV transforms each spectrum by subtracting its mean and dividing by its standard deviation, thereby minimizing these inconsistencies and improving comparability across samples.

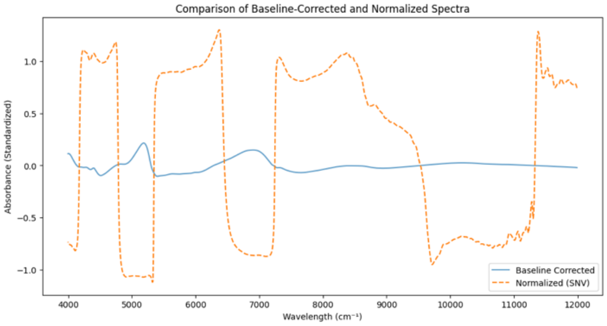

Figure 3: Comparison of Baseline-Corrected and Normalized Spectra

The plot compares baseline-corrected and normalized (Standard Normal Variate, SNV) spectra across a wavelength range of approximately 4000 to 12000 cm^{-1}. The baseline-corrected spectrum (solid blue line) appears smooth and retains overall spectral features with minimal distortion. In contrast, the SNV-normalized spectrum (dashed orange line) shows significant fluctuations, with sharp peaks and valleys indicating enhanced contrast in absorbance variations. This suggests that baseline correction preserves the general trend of the data, while SNV normalization amplifies variations, potentially improving the differentiation of spectral features.

- **Mean Centering and Scaling:** This step standardizes the spectral data by ensuring that all features have a mean of zero and are scaled to a common range. This is particularly useful for ML models, as it prevents certain wavelengths from dominating due to larger numerical values. By equalizing the scale of all spectral variables, the model learns patterns more effectively and generalizes better to new data.

2.4 Feature/Wavelength Selection

To optimize model performance, Partial Least Squares (PLS) Regression was used for feature selection. The absolute values of the PLS regression coefficients were analyzed to identify the most significant wavelengths contributing to moisture prediction.

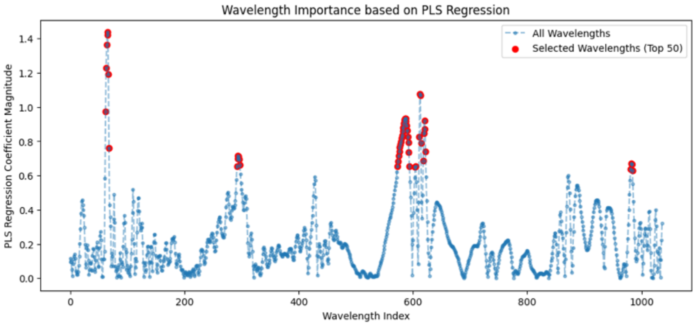

Figure 4: Important Wavelength

The top 50 wavelengths with the highest coefficient magnitudes were selected, reducing the dataset to these key features. This selection helps eliminate irrelevant data, improving model efficiency and accuracy.

3. Modeling

3.1 Selected Algorithms

Three models were selected based on their ability to capture different patterns in the data:

1. **Partial Least Squares (PLS):** PLS is a linear regression method that reduces data dimensionality while maximizing covariance between spectral data and moisture content. It is widely used in chemometrics due to its robustness in handling collinear data. The PLS model was trained using cross-validation to optimize performance. This method is particularly advantageous in high-dimensional correlated data, as it captures the most relevant information while avoiding overfitting.

2. **Support Vector Regression (SVR):** SVR is a non-linear regression approach that maps input features into a high-dimensional space using kernel functions. The SVR model was trained with radial basis function (RBF) kernels, and hyperparameters were optimized using cross-validation. This technique is well-suited for capturing complex nonlinear relationships in spectral data and provides robust predictions while performing effectively even on small datasets.

3. **Artificial Neural Network (ANN):** ANNs are powerful models capable of learning non-linear relationships. A feedforward neural network with one hidden layer and ReLU activation function was implemented. The model was trained using backpropagation with the Adam optimizer. ANNs are particularly effective for capturing intricate spectral patterns and nonlinear interactions; however, they require larger datasets to achieve optimal performance.

4. Results

The models were evaluated using **5-fold cross-validation**, with **R² (coefficient of determination) and RMSE (root mean squared error)** as performance metrics. Model's accuracy:

Model	**R^2**	**RMSECV**
PLS	0.95	3.1
SVR	0.96	3.0
ANN	0.95	3.2

Table 1: Result of the Models

4.1 Evaluation Analysis: RMSE Cross-Validation

Evaluating model performance through RMSE in cross-validation provides insights into prediction accuracy and consistency.

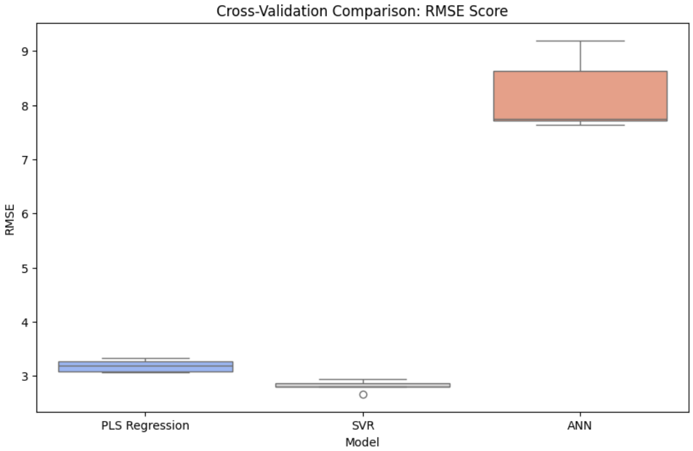

Figure 5: Cross-Validation Comparison

The boxplot compares the RMSE scores of three models—PLS Regression, SVR, and ANN—based on cross-validation results. The SVR model demonstrates the lowest RMSE with minimal variance, indicating superior performance and consistency. PLS Regression follows with slightly higher RMSE but still remains relatively stable. In contrast, the ANN model exhibits significantly higher RMSE with greater variance, suggesting that it struggles with prediction accuracy compared to the other two models. This result implies that simpler models like SVR and PLS Regression outperform ANN for this specific task, potentially due to overfitting, improper tuning, or data limitations affecting the ANN's learning process.

4.2 Model Performance Analysis

To further assess model accuracy, scatter plots and residual plots are analyzed to compare actual and predicted moisture content. These visualizations help evaluate how well each model captures data trends and identify potential biases or systematic errors in predictions.

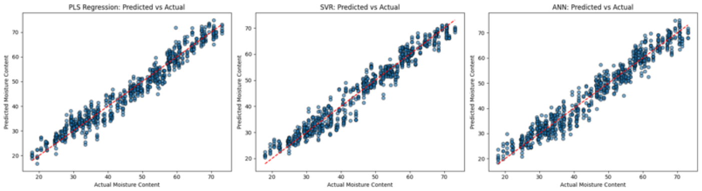

Figure 6: Scatter Plot of Actual vs Predicted Moisture

The scatter plots(Fig. 6) compare predicted vs. actual moisture content for PLS Regression, SVR, and ANN. All models show strong predictive performance, with SVR aligning closest to the ideal line, suggesting higher accuracy. ANN exhibits more variability, while PLS remains consistent.

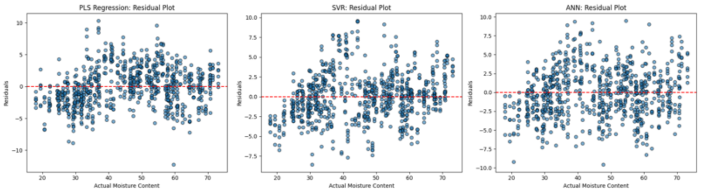

Figure 7: Residual Plot of Actual and Residuals Moisture

The residual plots(Fig. 7) show the distribution of errors for PLS Regression, SVR, and ANN. Ideally, residuals should be randomly scattered around zero, indicating no systematic bias. PLS Regression and SVR exhibit relatively balanced errors, while ANN shows more variability and potential heteroscedasticity. SVR has a tighter spread around zero, suggesting better overall predictive performance with fewer large errors.

5. Conclusion

The results of this study demonstrate that Support Vector Regression (SVR) achieved the highest accuracy for moisture prediction in NIR spectroscopy. Among the various preprocessing techniques tested, a combination of Savitzky-Golay filtering, normalization, and feature selection significantly enhanced the model's accuracy. Feature selection helped improve the performance of both SVR and Artificial Neural Networks (ANN) by reducing noise and dimensionality.

To improve further, future work could involve exploring deeper neural networks with larger datasets to boost prediction accuracy. Additionally, the development of ensemble models combining Partial Least Squares (PLS), SVR, and ANN could provide stronger predictive power. Experimenting with other preprocessing methods, such as derivative spectroscopy, could also improve results.

However, several challenges were encountered in this study. SVR, while providing the best performance, can be computationally expensive when dealing with large datasets, posing scalability challenges. Furthermore, ANN requires careful hyperparameter tuning to achieve optimal performance. Addressing these bottlenecks with more efficient tuning techniques or alternative methods for handling large-scale data could lead to even better results.