International Journal of Advanced Engineering Research and Science (IJAERS)
Peer-Reviewed Journal
ISSN: 2349-6495(P) | 2456-1908(O)
Vol-13, Issue-9; Sep, 2026
Journal Home Page Available: https://ijaers.com/
Article DOI: https://dx.doi.org/10.22161/ijaers.139.4

Feature Selection and Multivariate Regression for Paper Surface Roughness Prediction

Matheus Assis Domingues[1], Lílian Cristina Côcco[2], Fillemon Edilyn da Silva Bambirra Alves[2], Ricardo Henrique Moreton Godoi[3], Carlos Itsuo Yamamoto[1]

[1]Department of Chemical Engineering, Federal University of Paraná, Brazil
Email: matheus.assis@ufpr.br; ciyama@ufpr.br
[2]LACAUT- Department of Chemical Engineering, Federal University of Paraná, Brazil
Email : lilian.cocco@ufpr.br; fesbalves@gmail.com
[3]Department of Environmental Engineering, Federal University of Paraná, Brazil
Email : rhmgodoi@ufpr.br

Received: 16 Aug 2026,

Received in revised form: 16 Sep 2026,

Accepted: 20 Sep 2026,

Available online: 24 Sep 2026

***Keywords*— Paper Surface Roughness, Feature Selection, Lasso Regression, Principal Component Analysis (PCA), Industrial Process Prediction.**

***Abstract*— Paper surface roughness is an important quality parameter because it affects coating quality, printability, and the final appearance of paper. Accurate prediction of roughness remains challenging due to the large number of correlated variables monitored throughout the papermaking process. This study compares three feature selection methods for predicting average paper roughness from industrial data: Pearson correlation analysis, Principal Component Analysis (PCA), and a hybrid approach combining the Least Absolute Shrinkage and Selection Operator (Lasso) with Forward Selection. The dataset consisted of 7,145 hourly observations and 198 process variables collected from an industrial paper production process. After data preprocessing and min-max normalization, the selected variables were used to develop Multiple Linear Regression (MLR) models. The Pearson-based model achieved an R^2 of 0.696, whereas the PCA model reached an R^2 of 0.728 using seven principal components that explained 92.37% of the data variance. The best performance was obtained with the Lasso + Forward Selection approach, which achieved an R^2 of 0.789 using only 20 original process variables. Although PCA reduced the dimensionality of the dataset, the hybrid approach produced a more accurate model while preserving variables with direct physical meaning. These results indicate that combining Lasso with Forward Selection is an effective strategy for predicting paper roughness and can support process monitoring and quality improvement in industrial papermaking.*

I. INTRODUCTION

Industrial papermaking is a continuous, high-speed manufacturing process where the monitoring and control of surface properties are paramount to the final product's market value. Among these parameters, average paper roughness represents a critical quality metric, directly governing the structural uniformity of subsequent functional coatings, optical characteristics, gloss development, and microtopographical distribution [1]. The development of the paper's three-dimensional topography is a remarkably complex mechanism, inherently defined as a physical and chemical continuum that spans from the hydrodynamic fiber dispersion, transverse fiber wall collapse, and hydrogen bonding within the fibrous network in the wet-end section up to the structural consolidation and polymer application in film press units [2-5].

Consequently, deviations in sheet topography compromise both surface finish uniformity and printability, leading to significant economic losses.

Modern paper mills address this challenge by deploying extensive automated sensor networks that continuously record hundreds of operational variables, including stock consistency, refining freeness, headbox pressures, and multi-layer sizing chemical and mineral filler consumption rates. Particulate mineral fillers and chemical retention aids significantly alter the drainage, local flocculation, and ultimate consolidation of the sheet matrix, impacting native surface voids [6-7]. However, this high-dimensional industrial architecture often generates heavily collinear, redundant, and noisy datasets. In continuous integrated chemical processes, operational variables rarely act in isolation; coupled control loops and process recirculation phenomena create massive interdependencies that effectively mask the true predictors of surface quality [8]. Traditional univariate or bivariated statistical approaches, such as the Pearson filter correlation, frequently struggle in this scenario because they analyze process parameters in a fragmented manner, completely omitting the synergistic or compounding effect that multi-stage process variations exert on the finished sheet structure, such as the three-dimensional impact that irregular basestock formation exerts on the subsequent wetting expansion and polymer coating absorption [9].

To bridge this methodological gap, this study proposes a robust data-driven predictive framework that benchmarks distinct multivariate feature selection techniques to optimize paper roughness predictions. By evaluating and comparing a traditional univariated Pearson filter against advanced algorithms namely the orthogonal subspace projection of Principal Component Analysis (PCA) and a hybrid regularized scheme combining the Least Absolute Shrinkage and Selection Operator (Lasso) with sequential Forward Selection, in order to isolate the most informative process features out of high-dimensional industrial noise.

The use of Lasso is mathematically justified by the introduction of an L1 penalty term that forces redundant or collinear coefficients to zero, effectively mitigating overfitting in high-dimensional matrices [10-11]. Therefore, the core objective of this paper is to evaluate the ability of these feature selection methods to feed a Multiple Linear Regression (MLR) predictor. Through this benchmark, the aim is to deliver a model that achieves not only high mathematical predictive accuracy but also preserves direct physical interpretability and root-cause diagnostic capabilities, thereby offering a practical decision-making tool for real-time quality control and process optimization in modern papermaking systems.

II. METHODOLOGY

2.1 Data Collection and Preprocessing

The dataset used in this study was obtained from an industrial paper production process and consisted of hourly averaged sensor measurements collected from the process monitoring system. The database contained 7,145 observations and 198 process variables, including operational measurements throughout the production line, product characteristics, and process input variables. These variables were used for the development of predictive models aimed at estimating the variable of interest, namely paper roughness.

Data preprocessing was performed to improve data reliability and ensure the robustness of the developed models. This stage involved the removal of unreliable observations associated with sub-optimal operating conditions, flat sensor signals with little or no variation, and spurious measurements. In addition, the operating ranges of the input variables were evaluated to verify process consistency and identify anomalous values. Data corresponding to process transition periods and industrial trial intervals were also removed to avoid the inclusion of non-representative operating conditions in the modeling stage [12].

2.2 Pearson Correlation Analysis

Feature selection is an important step in predictive modeling because it identifies the variables that are most strongly associated with the response variable while reducing the dimensionality of the dataset [13]. In this study, Pearson correlation analysis was initially employed to evaluate the linear relationship between each predictor variable and the target variable. This method was selected because of its simplicity, computational efficiency, and ability to quantify the strength and direction of linear associations.

The dependence between variables allows one variable to carry information about another. This relationship may be either strong or weak, as well as positive or negative. Pearson's correlation coefficient is a statistical measure used to evaluate the degree of linear dependence between two variables.

The Pearson correlation coefficient (r) ranges from −1 to +1. The sign of the coefficient indicates the direction of the linear relationship between the analyzed variables, while its magnitude indicates the strength of the correlation [14]. Therefore:

$r > 0$: indicates a directly proportional relationship between the variables, with $r = 1$ representing a perfect positive linear correlation;

$r < 0$: indicates an inversely proportional relationship between the variables, with $r = -1$ representing a perfect negative linear correlation;

$r = 0$: indicates the absence of linear correlation between the variables.

The Pearson correlation coefficient can be calculated as the ratio between the covariance of two variables and the product of their standard deviations, as shown equation 1 (1) below:

$$r_{XY} = \frac{\text{cov}(X,Y)}{S_X S_Y} \quad (1)$$

Substituting the covariance and standard deviation terms, the correlation coefficient may also be expressed as equation 2 (2):

$$r_{XY} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (2)$$

2.3 Lasso and Forward Selection

In addition to Pearson correlation analysis, a second feature selection approach based on the combination of the Least Absolute Shrinkage and Selection Operator (Lasso) and Forward Selection was evaluated. While Pearson correlation assesses the relationship between each predictor and the response variable individually, the Lasso + Forward Selection approach evaluates the contribution of variables within the context of the regression model. The combination of these methods allows an initial reduction of the predictor space, followed by the identification of the subset of variables that provides the best predictive performance [15-16].

Initially, Lasso regression with automatic hyperparameter tuning (LassoCV) was applied to the complete set of predictor variables. The optimal regularization parameter (λ) was determined through five-fold cross-validation. Lasso estimates the regression coefficients by minimizing the residual sum of squares while simultaneously applying an L1 regularization penalty [10]. The optimization problem is expressed as equation 3 (3):

$$\min_{\beta} \left(\frac{1}{2n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \quad (3)$$

Where n is the number of observations, p is the number of predictor variables, β_j is the regression coefficient associated with the j-th predictor variable, and λ is the regularization parameter controlling the amount of coefficient shrinkage. As the value of λ increases, the coefficients of less relevant variables are progressively reduced toward zero, allowing irrelevant predictors to be automatically removed from the model.

Although Lasso efficiently reduces the number of candidate variables, the selected subset is determined by the regularization process and may not correspond to the combination of predictors that provides the best regression performance. Therefore, the variables retained by Lasso were subsequently submitted to the Forward Selection algorithm [17].

Forward Selection is a sequential feature selection method that starts from an intercept-only regression model as expressed in equation 4 (4):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \quad (4)$$

Where y is the response variable, $x_1, x_2, \ldots, x_p$ are the predictor variables, β_0 is the intercept, $\beta_1, \ldots, \beta_p$ are the regression coefficients, and ε represents the random error term.

At each iteration, all remaining candidate variables are evaluated and the predictor producing the greatest improvement in the adjusted coefficient of determination (R^2 adjusted) is incorporated into the model. The procedure continues until the inclusion of additional variables no longer provides a statistically significant increase in predictive performance [18].

2.4 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was evaluated as an alternative dimensionality reduction technique for developing the multivariate regression model [18-19]. In datasets containing highly correlated predictors, multicollinearity can adversely affect the stability and interpretability of linear regression models. Unlike Pearson correlation, which evaluates the relationship between each predictor and the response variable independently and therefore does not account for correlations among predictors, PCA transforms the original variables into a new set of orthogonal variables, known as principal components. This transformation reduces the dimensionality of the dataset while preserving most of the information contained in the original variables and completely eliminates multicollinearity among the transformed predictors [19-20]. Lasso-based methods also mitigate multicollinearity by favoring the selection of a reduced subset of predictors, often retaining only one variable from groups of highly correlated features. In contrast, PCA addresses the problem by replacing the original correlated variables with uncorrelated principal components, which can improve the stability and robustness of regression models.

PCA is based on the eigendecomposition of the covariance matrix of the predictor variables as described in equation 5 (5) [19]:

$$\Sigma = Q\Lambda Q^{T} \qquad (5)$$

Where Σ is the covariance matrix, Q is the matrix whose columns correspond to the eigenvectors, and Λ is the diagonal matrix containing the associated eigenvalues. Each principal component is obtained as a linear combination of the original predictor variables as described in equation 6 [19]:

$$PC_k = Xq_k \qquad (6)$$

Where X is the predictor matrix, q_k is the eigenvector associated with the k-th principal component, and PC_k represents the transformed variable. The contribution of each principal component to the total dataset variability was quantified through its explained variance, described as equation 7 (7):

$$EV_k = \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_k} \qquad (7)$$

Where λ_k is the eigenvalue associated with the k-th principal component and p is the total number of original variables.

The principal components were ranked according to their explained variance, and the leading components accounting for most of the cumulative variance were retained [18]. These components were then used as predictor variables in the development of the multivariate linear regression model, allowing the predictive performance obtained with PCA to be compared with the models developed using Pearson correlation analysis and the Lasso + Forward Selection approach.

2.5 Multivariate Regression Model Development

The dataset was divided into training and testing subsets to evaluate the predictive performance and generalization capability of the developed model. The training subset was used for model fitting and parameter estimation, while the test subset was employed to assess model performance using previously unseen data [18].

Multivariate linear regression was applied to describe the relationship between the process variables and the response variable [18, 21]. The regression model can be expressed as equation 8 (8):

$$y = \Phi(x)\theta + \varepsilon \qquad (8)$$

Where y is the response variable, $\Phi(x)$ is the regression matrix containing the predictor variables, θ is the vector of regression coefficients, and ε is the residual error term.

The regression function is defined as equation 9 (9):

$$f(x) = \sum_{i=0}^{K-1} \theta_i x_i \qquad (9)$$

The model parameters were estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of squared residuals between observed and predicted values. The parameter vector was estimated according to equation 10 (10):

$$\theta = (\Phi^T\Phi)^{-1}\Phi^T y \qquad (10)$$

III. RESULTS AND DISCUSSION

3.1 Data preprocessing

The removal of unreliable measurements and unstable operating periods reduced the influence of noise and abnormal process behavior, resulting in a more consistent dataset for the subsequent statistical analyses and predictive modeling. This filtering step is critical in continuous industrial processes, where process disturbances and spurious sensor measurements may introduce artificial autocorrelation or obscure genuine physical relationships [22].

The process variables were grouped into four main categories: product composition, process variables, finished product properties, and chemical consumption variables. This categorization facilitated the interpretation of the influence of different production stages on paper roughness.

In addition, all variables were normalized using min-max normalization as shown in equation 11 (11) [23] to minimize distortions caused by differences in engineering units and variable magnitudes prior to the statistical analyses and regression modeling. Standardizing the scale of the variables is essential for predictive algorithms and multivariate statistical analyses, as it prevents variables with large numerical magnitudes, such as operating pressures or flow rates, from artificially dominating the mathematical model at the expense of smaller-scale variables that may have high physical relevance, such as chemical consumption rates [24].

$$x_n = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad (11)$$

After the preprocessing stage, the resulting dataset presented improved consistency, reduced noise influence, and better comparability among variables. These improvements were essential to ensure the reliability of the subsequent statistical analyses, particularly the Pearson correlation analysis, which was applied to investigate the linear relationships between the process variables and paper roughness. By minimizing distortions associated with unstable operating conditions and differences in variable scales, the preprocessing procedure contributed to a more accurate identification of the most influential process variables. Additionally, the paper engineering literature emphasizes that the prior removal of spurious measurements and unstable operating periods is an essential prerequisite for ensuring that statistical analyses and linear correlations reflect the actual physical relationships within the process, thereby preventing the masking of meaningful patterns by legitimate operational noise [25].

3.2 Pearson correlation analysis

Pearson correlation analysis was performed to evaluate the linear relationship between the process variables and the average paper roughness. The correlation coefficients were classified according to their magnitude as weak ($|\rho| \leq 0.3$), moderate ($0.3 < |\rho| < 0.6$), and strong ($|\rho| \geq 0.6$). Only variables presenting at least moderate correlation with average roughness were considered relevant for the subsequent analyses.

The results demonstrated consistent correlation patterns among the evaluated operating conditions, suggesting that common process mechanisms significantly influence the formation of paper surface roughness. Variables associated with fiber composition, process operating conditions, finished product properties, and chemical consumption showed the strongest correlations with the response variable.

3.2.1 Furnish composition variables

The variables related to fiber composition exhibited moderate correlations with average paper roughness, indicating that changes in the furnish composition influence the development of the paper surface. The variables associated with the proportions of raw materials 1 ($r = 0.32$) and raw materials 2 ($r = -0.50$) fibers in the middle-layer furnish, as well as their calculated percentages on a dry-mass basis ($r = 0.31$ and -0.48 respectively), showed the strongest correlations within this group.

In addition to the variables directly related to pulp composition, operational factors, such as machine chest consistency and coating line production rate ($r = -0.47$), also exhibited moderate negative associations with surface roughness. The consistency of the pulp suspension influences fiber packing and the initial flocculation dynamics, affecting the uniformity of jet deposition on the forming wire and, consequently, the final surface topography of the paper sheet [25].

The variables associated with raw materials 2 fibers exhibited moderate negative correlations with average roughness, whereas the raw materials 1-related variables presented moderate positive correlations. These results indicate that higher proportions of raw materials 2 fibers were associated with lower paper roughness, while higher raw materials 1 contents were associated with increased roughness. Raw materials 1 corresponds to long fibers, whereas raw materials 2 corresponds to short fibers, suggesting that the observed correlations are related to the distinct morphological characteristics of these fiber types.

The influence of fiber composition on paper roughness can be explained by the role that different fiber types play during sheet formation. Fiber morphology, including characteristics such as length, flexibility, fines content, and bonding ability, directly affects the arrangement and consolidation of the fiber network. Fibers with greater flexibility and conformability tend to pack more efficiently, filling voids and producing a denser and more homogeneous sheet structure [2]. This results in improved surface uniformity and lower roughness values.

In contrast, fibers with lower conformability or greater structural rigidity may promote a more heterogeneous fiber network, increasing surface irregularities and, consequently, paper roughness. [2,26].

3.2.2 Process variables

The process variables exhibited both positive and negative moderate correlations with average paper roughness, reflecting the influence of different operating conditions on sheet formation and surface development. The most relevant process variables were associated with coating suspension properties, stock preparation, refining conditions, and headbox operating parameters.

Among the evaluated variables, the coating headbox pH exhibited the strongest positive correlation with average roughness ($r = 0.59$). Variations in pH can influence the stability of the coating suspension and the interactions between chemical additives, potentially affecting coating application and surface uniformity. Consequently, changes in coating chemistry may contribute to variations in the final paper roughness. From a physicochemical perspective, pH plays a governing role in the surface charge (zeta potential) of mineral pigments and in the polymer conformation of binders and rheological additives. Deviations toward unstable pH ranges promote microflocculation of the coating color, significantly altering its viscosity and hydrodynamic behavior under the application blade, thereby inducing streak formation and irregular topographic deposition on the fibrous matrix. [5,7].

Machine chest consistency also presented a moderate negative correlation ($r = -0.46$). Variations in pulp consistency affect fiber concentration during stock preparation and can influence fiber distribution and sheet formation. More stable consistency conditions generally contribute to improved formation and greater surface uniformity, reducing paper roughness. Strict control of pulp consistency regulates the frequency of fiber–fiber interactions and prevents the formation of large flocs within the suspension, thereby ensuring a two-dimensionally uniform distribution of the stock jet over the forming wire [26].

The online freeness measured after the base refiner and the operating pressures at the headboxe also exhibited moderate correlations with paper roughness (r= -0.37). These variables are associated with refining intensity and stock delivery conditions, both of which influence fiber dispersion, sheet formation, and the development of a more uniform paper surface. The interdependence of these wet-end variables governs the formation quality of the base paper, particularly the uniformity of mass distribution across the sheet. Fluctuations in headbox pressure or inadequate refining lead to local variations in density and porosity. When the base sheet receives the aqueous coating color, these mass variations promote non-uniform fluid absorption and asymmetric fiber swelling (the wetting expansion effect), ultimately resulting in increased macroscopic surface roughness of the finished paper [2-9].

3.2.3 Finished product variables

The finished product variables exhibited moderate to strong negative correlations with average paper roughness, indicating that product characteristics are closely associated with the development of the paper surface. The applied basis weight at the SymSizer (an industrial film sizer used in papermaking to apply starch or pigment coatings onto a running paper or board web) and the calculated paper basis weight showed the strongest correlations within this group.

The SymSizer applied basis weight exhibited the strongest correlation observed in this group (r = -0.62), while the calculated paper basis weight also showed a moderate negative correlation (r = -0.53). These results indicate that higher basis weights were associated with lower paper roughness.

This behavior can be explained by the influence of basis weight on sheet structure and surface development. Papers with higher basis weights generally contain a greater amount of fibrous material per unit area, promoting improved sheet consolidation and reducing the occurrence of surface irregularities. In addition, the material applied by the SymSizer contributes to filling surface voids and improving surface coverage, resulting in a smoother paper surface. Mechanistically, the SymSizer operates according to the metered film press principle, transferring a controlled polymer film of starch or other chemical formulations onto the moving paper web under pressure. A higher film application rate (applied basis weight or pickup) directly promotes the filling of the micro and macropores on the surface of the base paper, forming a continuous polymer layer that smooths surface irregularities and enhances the smoothness and printability of the finished product [5].

Overall, the observed correlations indicate that basis weight, both in the base sheet and in the material applied by the SymSizer, plays an important role in determining paper surface quality by contributing to a denser and more homogeneous sheet structure. Furthermore, a base paper with a higher calculated basis weight provides an inherently more stable and uniform three-dimensional fibrous network. This structural uniformity mitigates the localized fiber expansion and relaxation effects (moisture-induced roughening) caused by contact with the water contained in the polymer suspension during surface sizing, thereby preserving the topographical uniformity of the sheet. [2].

3.2.4 Chemical consumption variables

Among all evaluated variable groups, the chemical consumption variables exhibited the strongest correlations with average paper roughness, highlighting the importance of chemical control in the development of paper surface properties. The most significant correlations were associated with the consumption of rosin, AKD, PAC, calcined kaolin, and cooked starch.

The strong correlations observed for rosin and AKD indicate that internal sizing conditions are closely associated with paper roughness. Although these additives are primarily used to control water absorption, their dosage also affects the interactions between fibers and other chemical components, influencing sheet consolidation and the development of the paper surface. Notably, these sizing agents act through opposing topographical mechanisms. Rosin consumption exhibited a strong negative correlation with surface roughness (r = −0.72), indicating that conventional internal sizing based on resin emulsions promotes sheet consolidation and reduces surface voids. In contrast, specific AKD consumption showed a strong positive correlation (r = 0.67). This behavior suggests that high dosages of AKD, a long-chain alkaline synthetic sizing agent, tend to coat the fiber surfaces with hydrophobic chains prior to wet pressing, locally hindering the formation of interfiber hydrogen bonds. As a result, the three-dimensional fibrous network becomes less compact, leading to a rougher final surface microtopography [2,5]. Likewise, the correlation observed for PAC (r = -54) reinforces the importance of wet-end stability, since efficient retention of fines and fillers contributes to a more uniform fiber network and improved surface formation. PAC acts as a highly charged coagulant that neutralizes interfering anionic colloids (anionic trash) and enhances the retention of fines and additives on the forming wire. This stabilization prevents the asymmetric migration of fine particles induced by vacuum-assisted drainage (two-sidedness effect), ensuring a more homogeneous mass distribution throughout the thickness of the sheet and mitigating variations in surface roughness. [2].

The correlations obtained for calcined kaolin and cooked starch further demonstrate that additives involved in surface treatment and fiber bonding also influence paper roughness. However, the positive correlation observed for calcined kaolin (r = 0.39) deserves careful structural interpretation. Although fine mineral fillers contribute to opacity, calcined kaolin consists of rigid, porous aggregates with high structural integrity. When incorporated into the pulp suspension, these particles act as mechanical spacers that hinder the collapse and intimate contact between the cell walls of adjacent fibers during wet consolidation, resulting in a base paper with higher specific volume (bulk) and a macroscopically rougher surface texture. [6-7].

Calcined kaolin improves surface coverage by filling pores and surface irregularities, if applied exclusively in surface coating formulations [6] whereas cooked starch enhances fiber bonding and sheet consolidation. Together, these results indicate that maintaining appropriate chemical dosages is essential for obtaining stable process conditions and consistent paper surface quality.

Overall, the Pearson correlation analysis identified variables from different stages of the papermaking process that were associated with average paper roughness. Among all evaluated groups, the strongest correlations were observed for chemical consumption variables and finished product properties, suggesting that both wet-end chemistry and surface treatment play an important role in determining paper surface quality. These findings also demonstrate that paper roughness is governed by multiple interacting process variables rather than by a single operating parameter.

3.3 Multivariate Regression Model Using Pearson-Selected Variables

A Multiple Linear Regression (MLR) model was developed using the variables selected through Pearson correlation analysis to evaluate their ability to predict average paper roughness. Fig, 1 presents the relationship between the measured and predicted values obtained by the regression model.

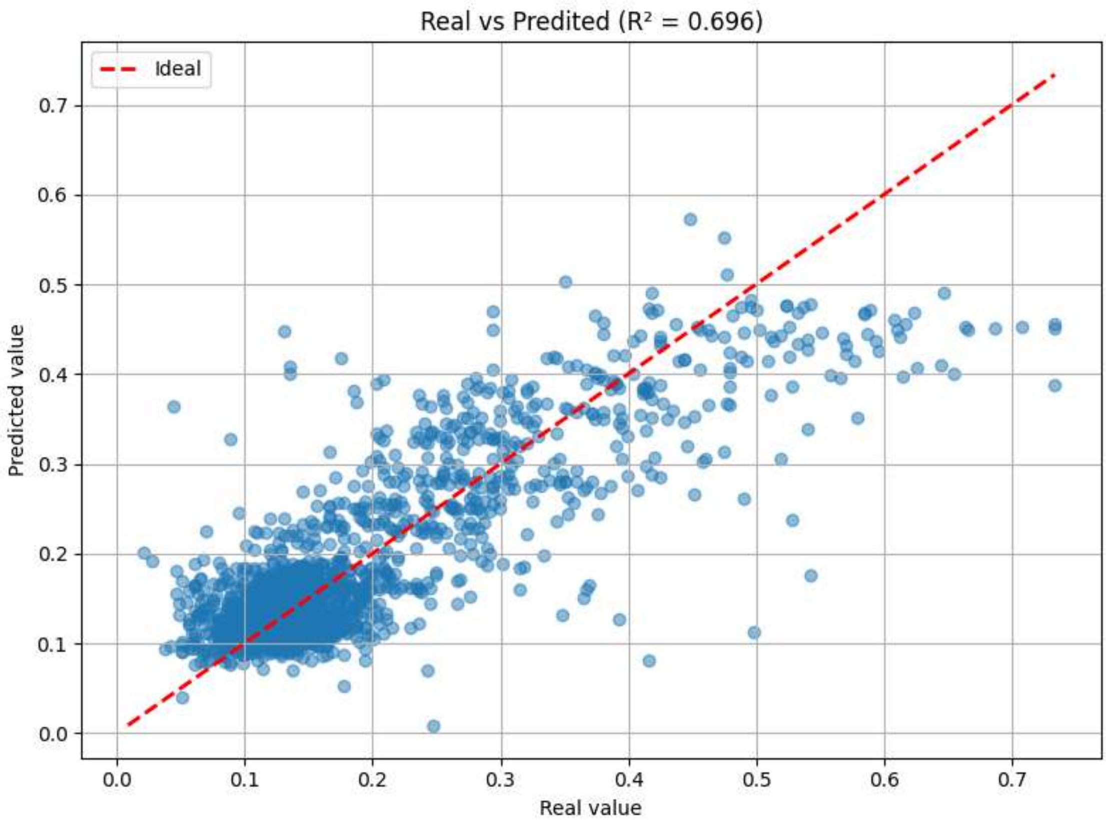

Fig. 1: MLR model using Pearson-selected variables.

The regression model achieved an R^2 of 0.696, indicating that nearly 70% of the variability in average paper roughness was explained by the selected variables. This result shows that the Pearson-based feature selection retained the most relevant information required to describe the linear relationship between the process variables and paper roughness.

A greater dispersion of the points can be observed for higher roughness values, where the model tends to underestimate some observations. This behavior suggests that the relationship between the process variables and paper roughness becomes more complex under these operating conditions and cannot be completely described by a linear model. This loss of linearity and the increased residual dispersion at the extreme roughness levels indicate that severe process disturbances or cumulative deviations within the industrial process trigger inherently nonlinear rheological and structural dynamics [27]. Under unstable operating conditions, synergistic interactions and saturation effects associated with chemical additives exceed the additive assumptions of global linear models, requiring robust feature selection approaches to identify the dominant predictors [11]. Overall, the results show that the variables selected through Pearson correlation retain the main information associated with paper roughness and allow the development of a regression model capable of representing the overall behavior of the process.

3.4 Lasso and Forward Selection

The Lasso and Forward Selection approach was applied to identify a reduced subset of variables for regression modeling. Initially, the Lasso algorithm retained 123 variables with non-zero coefficients, considerably reducing the dimensionality of the original dataset. The use of the Least Absolute Shrinkage and Selection Operator (Lasso) is widely established in process engineering because it applies an L1 regularization penalty that shrinks irrelevant or collinear coefficients to zero, thereby mitigating overfitting in highly redundant datasets [10]. From this subset, the 60 variables with the largest absolute coefficients were selected as candidate predictors for the Forward Selection procedure. After the Forward Selection procedure, the final regression model was composed of 20 variables associated with different stages of the papermaking process, including fiber composition, process operation, finished product properties, and chemical consumption.

The selected variables indicate that the prediction of paper roughness does not rely on a single process condition but on the combined contribution of variables from different process stages. In addition to variables previously identified by the Pearson correlation analysis, the hybrid approach also retained variables with lower individual correlations, suggesting that they provide complementary information when combined in the regression model. The reduction from the original dataset to only 20 predictor variables considerably simplified the regression model while preserving variables representative of the main process stages.

The 20 variables selected by the hybrid Lasso and Forward Selection approach represent different aspects of the papermaking process. The selected variables were Specific Rosin Consumption – Top Layer, Specific Rosin Consumption – Middle Layer, Specific AKD Consumption – Middle Layer, Specific PAC Consumption – Top Layer, Calculated Coating Weight by Flow Rate, Applied Coating Weight, Total Applied Coating Weight, SymSizer Applied Basis Weight, Coating Basis Weight Setpoint, Short Fiber Percentage, Long Fiber Percentage, Production Rate, Average Steam Pressure of CLS, Coating Headbox Level, Outlet Pressure, Refiner Load, Freeness at Base Refiner, Freeness at Top Refiner, Dry Solids – Bel-Bond Outlet, and Smoothing Press Bottom Roll. These variables encompass chemical consumption, coating application, fiber composition, production conditions, refining, drainage, drying, and sheet consolidation, covering different stages of the papermaking process and highlighting the multifactorial nature of paper roughness.

The large number of coating-related variables selected by the Lasso and Forward Selection approach highlights the dominant influence of the coating process on the prediction of paper roughness. Variables associated with coating weight, coating application stability, and coating delivery conditions (including the calculated and measured coating weights, total applied coating weight, SymSizer applied basis weight, coating basis weight setpoint, and coating headbox level) were simultaneously retained in the final model. This result suggests that not only the amount of coating applied but also the stability and uniformity of its application play a decisive role in determining the final surface topography. The coating layer fills surface pores and irregularities left by the fibrous substrate, reducing surface roughness and improving smoothness. Consequently, variations in coating application conditions may directly affect coating coverage, thickness uniformity, and defect formation, ultimately influencing the roughness of the finished paper [5,28-29].

From a physical standpoint, the inclusion of operational variables associated with drying and sheet consolidation, such as the Average Steam Pressure of the CLS, Dry Solids at the Bel-Bond outlet, and the Smoothing Press bottom roll, reinforces the importance of moisture removal and sheet densification in determining the final surface quality. The steam pressure supplied to the drying cylinders governs the evaporation rate and the relaxation of hygrothermal stresses within the paper sheet. Likewise, the dry solids content before drying reflects the efficiency of water removal, while the smoothing press contributes to the reduction of surface irregularities through mechanical densification of the sheet. Together, these variables indicate that the final surface topography is strongly influenced by the combined effects of moisture management, drying conditions, and surface finishing operations [2,5,9].

3.5 Multiple Linear Regression Using Lasso and Forward Selection

A Multiple Linear Regression (MLR) model was developed using the variables selected by the Lasso and Forward Selection approach. Fig. 2 compares the measured and predicted values of average paper roughness obtained with the regression model.

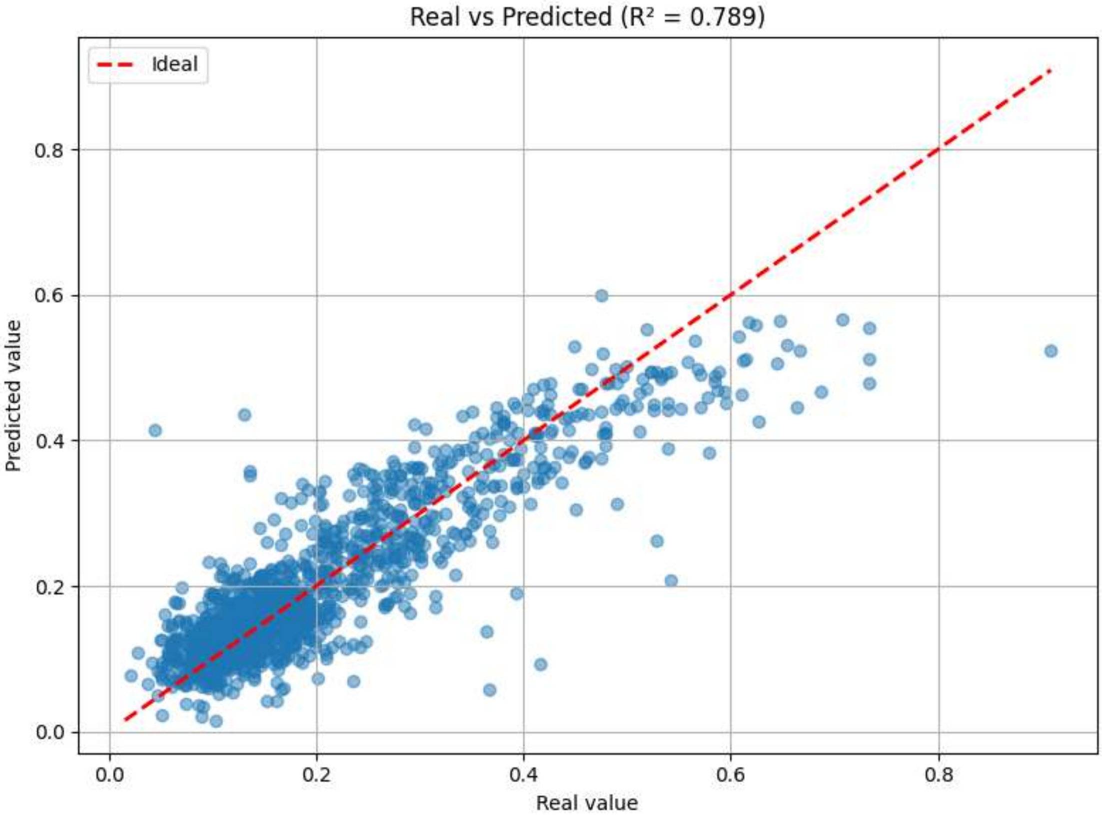

Fig. 2: MLR model using Lasso and Forward Selection.

The model achieved an R^2 of 0.789, showing an improvement compared with the model developed using the variables selected by the Pearson correlation analysis. Most of the predicted values followed the trend of the experimental data and were concentrated near the line of perfect agreement between the measured and predicted values. This substantial improvement in overall model fit reflects the superiority of statistical feature selection methods based on regularization and sequential search over bivariate correlation filters. While Pearson correlation analysis quantifies only the isolated linear relationship between each predictor and surface roughness, disregarding complex interdependencies, the hybrid approach evaluates the importance of each variable within a conditional and multivariate framework [27]. In complex industrial systems, highly collinear variables share overlapping variance. The Lasso algorithm addresses this redundancy through coefficient shrinkage, retaining only the subset of predictors with the greatest orthogonality and predictive power [11].

Although a larger dispersion is still observed for higher roughness values, the agreement between the measured and predicted values is greater than that obtained with the Pearson-based model. This result indicates that selecting variables according to their combined contribution to the regression model improves the predictive capability of the MLR model. The better performance obtained with the Lasso and Forward Selection approach suggests that some variables with relatively low individual correlation still provide relevant information when combined with other predictors. This behavior is characteristic of large-scale, continuous, multivariable industrial processes, such as papermaking, where process monitoring and quality control depend on the simultaneous interaction of multiple control loops [8]. As highlighted in the multivariate statistical process control literature, individual operating parameters may appear statistically insignificant when analyzed independently, yet they can serve as important hydrodynamic or thermal compensation factors through their interactions with variables from other sections of the machine [8]. As a result, the regression model was able to explain a larger portion of the variability in average paper roughness while using a reduced set of variables. Finally, the residual dispersion observed under high-roughness conditions highlights the physical limitations of global linear regression. Under extreme operating conditions or severe instabilities in chemical retention and starch dosing, the mechanisms governing fibrous web consolidation and stress relaxation undergo strongly nonlinear critical transitions that exceed the modeling capability of conventional additive structures [2,5].

3.6 Principal Component Analysis

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the original dataset before regression modeling. The number of retained components was automatically determined to preserve at least 90% of the cumulative variance of the original variables.

Table 1 summarizes the variance explained by each principal component. Seven principal components were sufficient to explain 92.37% of the total variance, reducing the dimensionality of the dataset while preserving most of the information contained in the original variables.

Table 1 – Eigenvalues and explained variance of the retained principal components.

Component	Explained Variance (%)	Cumulative Variance (%)
PC1	43.06	43.06
PC2	14.10	57.16
PC3	11.22	68.38
PC4	8.59	76.97
PC5	6.16	83.13
PC6	5.50	88.63
PC7	3.74	92.37

The first principal component explained 43.06% of the total variance, while the first three components together accounted for 68.38%. The remaining components contributed progressively smaller amounts of additional information, and seven components were sufficient to exceed the predefined threshold of 90% cumulative explained variance.

It is important to emphasize, however, that although PCA substantially simplifies the regression feature space by projecting the data onto only seven orthogonal components, it does not reduce the need for instrumentation or data acquisition in the industrial environment. In a real industrial setting, all original process variables must continue to be continuously monitored and acquired by physical sensors, since each principal component (PC) is a purely mathematical construct obtained from the linear combination of the entire set of raw process variables [8,19].

3.7 Multiple Linear Regression Using PCA

A Multiple Linear Regression (MLR) model was developed using the principal components obtained from the PCA. Fig. 3 presents the relationship between the measured and predicted values of average paper roughness.

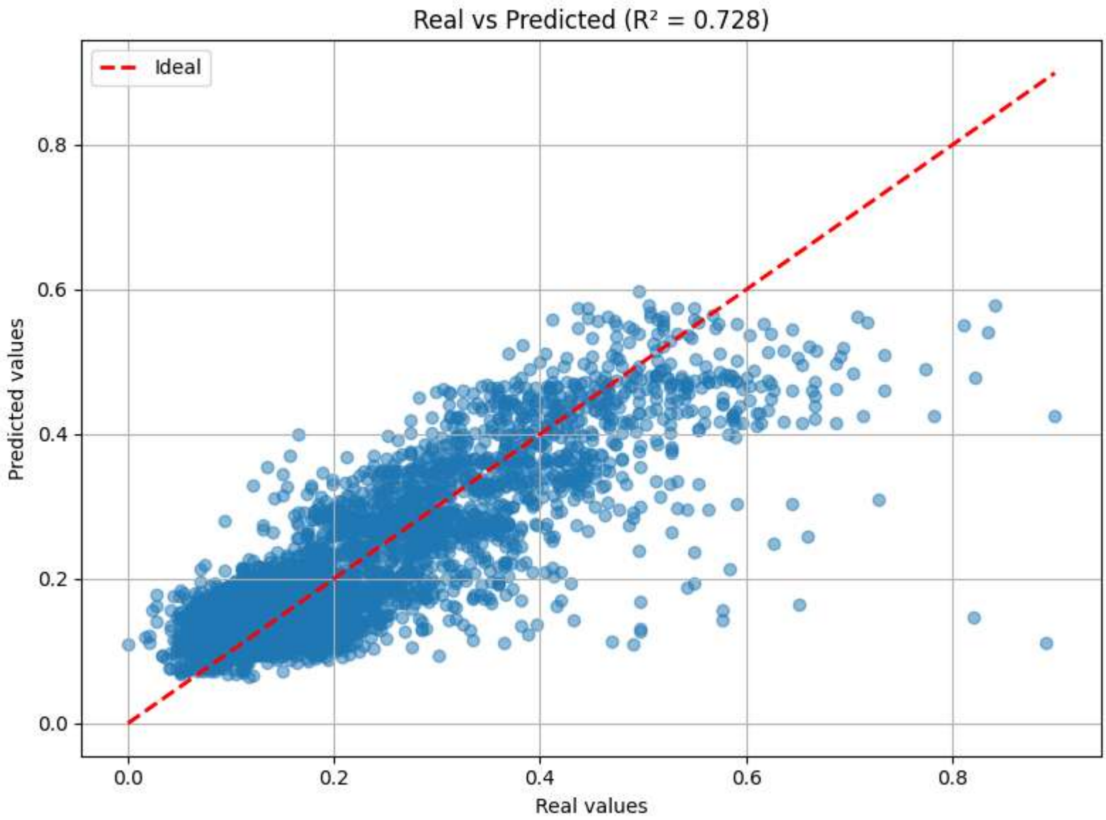

Fig. 3: MLR model using PCA.

The PCA-based model achieved an R^2 of 0.728, demonstrating predictive performance comparable to the models developed using Pearson correlation analysis and Lasso combined with Forward Selection. Most of the predicted values followed the general trend of the measured data. However, consistent with the behavior observed in the previous models, increased residual dispersion and a systematic tendency to underestimate high roughness values remain evident, as illustrated in Figure 3.

This recurrent behavior across distinct mathematical formulations suggests that part of the structural variability associated with extreme surface irregularities cannot be fully captured by linear models. In the papermaking industry, PCA is particularly effective at reducing redundancy and mitigating the effects of the extensive collinearity among process variables measured by industrial sensors. [8]. However, because PCA performs only a linear rotation of the feature space, the persistence of systematic errors under extreme operating conditions demonstrates that the physical transitions governing the deterioration of surface smoothness, such as local failures in filler retention or thermal instabilities during drying, give rise to bottleneck dynamics and synergistic effects that are inherently nonlinear and therefore beyond the modeling capability of global MLR models [8,27]. This reduction in dimensionality allowed the development of a regression model capable of representing the overall behaviour of average paper roughness while considerably reducing the number of input variables in the final predictive equation.

3.8 Comparison of the Feature Selection Methods

Table 2 summarizes the performance of the regression models developed using the three feature selection strategies evaluated in this study.

Although the three models achieved satisfactory predictive performance, differences were observed in both the coefficient of determination and the prediction errors. The Pearson-based model presented the lowest predictive performance (R^2 = 0.6958), whereas the Lasso and Forward Selection model achieved the highest R^2 (0.7893) while simultaneously producing the lowest MSE, RMSE, and MAE. The PCA-based model showed intermediate performance, with an R^2 of 0.7280, demonstrating that transforming the feature space through orthogonal projections can mitigate the detrimental effects of multicollinearity and improve predictive accuracy compared with traditional bivariate filtering methods, although the computation of these mathematical components still depends on the continuous acquisition of data from the entire set of original process sensors [8,27].

Tabela 2 – Comparison of regression model performance.

Method	Number of Inputs	R²	MSE	RMSE	MAE
Pearson Correlation	20	0.6958	0.004517	0.0672	0.0474
Lasso + Forward Selection	20	0.7893	0.003192	0.0565	0.0401
Principal Component Analysis	7	0.7280	0.003864	0.0622	0.0430

Note: for PCA, this represents the number of principal components, which are derived from all original variables.

The relatively small differences between the three models indicate that the variables selected by each approach were able to capture the main process characteristics affecting paper roughness. However, the consistent reduction in all error metrics obtained with the Lasso and Forward Selection approach suggests that this strategy identified a more informative subset of predictors for regression modeling. These results reinforce that selecting variables according to their combined contribution to the model is more effective than considering only their individual relationship with the response variable. Additionally, the statistical superiority of the hybrid approach (Lasso + Forward Selection) lies in its ability to simultaneously perform feature selection and coefficient regularization, identifying operational synergies that remain undetected by univariate analyses while avoiding overfitting [11].

In addition to achieving the highest predictive performance, the Lasso and Forward Selection model required only 20 original process variables, whereas the PCA model represented the dataset using seven principal components. Although PCA provided a more compact representation of the data, the resulting components are linear combinations of the original variables, making their physical interpretation less straightforward. From an industrial perspective, keeping the number of original process variables to a minimum is advantageous because it facilitates the identification of the operating conditions that influence paper roughness and supports process monitoring and optimization. In papermaking practice, models based on directly measurable process variables substantially reduce the computational burden associated with database maintenance and facilitate immediate root-cause diagnosis in the control room. When deviations in sheet surface topography occur, the direct interpretation of variables such as chemical dosing rates or dewatering parameters enables operators to implement corrective actions in real time. In contrast, deviations expressed as abstract PCA scores require the mathematical decomposition of complex loading vectors to identify the underlying process disturbance [8].

Overall, the results indicate that all three approaches are suitable for modeling average paper roughness. Although the metrics were computed on a single test split, the consistent reduction in all error metrics (MSE, RMSE, and MAE) strongly suggests that the Lasso + Forward Selection strategy identified a more robust and informative subset and provided the best balance between predictive performance, model simplicity, and interpretability, making it the most appropriate strategy for developing predictive models based on industrial process data.

IV. CONCLUSION

This study successfully demonstrated that paper surface roughness in a large-scale industrial manufacturing continuous continuum cannot be attributed to an isolated process parameter, but is rather governed by a complex hydrothermal, chemical, and mechanical interplay stretching from stock preparation to final surface sizing.

Among the data-driven feature selection strategies evaluated, the hybrid Lasso combined with Forward Selection approach provided the most robust and accurate framework for predicting average paper roughness, achieving the highest coefficient of determination (R2 = 0.7893) and the lowest overall prediction errors (MSE, RMSE, MAE). While Principal Component Analysis (PCA) achieved a highly compact orthogonal space (7 components, R2 = 0.7280), it inherently required the uninterrupted monitoring of all original variables and lacked straightforward physical interpretability.

From a practical and industrial perspective, the Lasso + Forward Selection model offers a decisive operational advantage: by condensing a high-dimensional industrial dataset into a highly informative subset of only 20 original process variables (such as, steam pressure and specific rosin/AKD chemical consumption rates), it preserves the physical meaning of the predictors. This allows control room operators to perform immediate real-time root-cause diagnosis and execute physical corrective adjustments when surface quality deviations occur, a task that is mathematically bottlenecked when dealing with abstract PCA scores.

Finally, the recurrent underestimation and larger residual dispersion observed in all models under high roughness regimes delineate the intrinsic boundary of global linear models. These extreme windows represent highly volatile operational states where non-linear chemical microflocculation, wetting expansion, and thermal stress relaxation interactions dominate. Consequently, future research should focus on implementing non-linear machine learning architectures, such as random forests or deep neural networks, specifically conditioned on the high-dimensional physical features isolated by the hybrid approaches developed in this work.

It is important to emphasize that this study was conducted using data obtained from a specific industrial paper manufacturing system and its associated operating conditions. Although the proposed methodology can be adapted and applied to other paper machines and production systems, the process variables influencing surface roughness may differ according to equipment configuration, raw material characteristics, operational conditions, and process control strategies. Therefore, the direct transfer of the variables selected in this study to other manufacturing systems should not be assumed. Instead, the proposed feature selection and modeling framework can be reapplied to each specific process, allowing the relevant process variables to be identified and the predictive model to be appropriately adapted and recalibrated to the characteristics of the target equipment.

ACKNOWLEDGEMENTS

This study was financed in part by the Fundação da Universidade Federal do Paraná (FUNPAR) and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. The authors also gratefully acknowledge the support provided by the Laboratório de Análises de Combustíveis Automotivos (LACAUT) at the Federal University of Paraná (UFPR).

REFERENCES

[1] Chinga, G. (2004). Detailed characterization of paper surface structure for gloss assessment. Journal of Pulp and Paper Science, 30(8), 222–227.

[2] Alava, M., & Niskanen, K. (2006). The physics of paper. Reports on Progress in Physics, 69(3), 669–723.

[3] Sirviö, J. (2011). Fibres and bonds. In K. Niskanen (Ed.), Paper physics (Papermaking Science and Technology, Book 16, pp. 59–90). Paper Engineers' Association.

[4] Lorbach, C., Hirn, U., Kritzinger, J., & Bauer, W. (2012). Measurement results for fiber wall thickness, fiber coarseness and fiber collapse. Nordic Pulp & Paper Research Journal, 27(2), 264–271.

[5] Hubbe, M. A. (2024). Size press practices and formulations affecting paper properties and process efficiency: A review. BioResources, 19(1), 1925–1994.

[6] Bown, R. (2018). A review of the influence of pigments on papermaking and coating. In C. F. Baker (Ed.), The fundamentals of papermaking materials (pp. 83–137). FRC.

[7] Hubbe, M. A., & Gill, R. A. (2016). Fillers for papermaking: A review of their properties, usage practices, and their mechanistic role. BioResources, 11(1), 2886–2963.

[8] Ge, Z., & Song, Z. (2007). Process monitoring based on independent component analysis-principal component analysis (ICA-PCA) and similarity factors. Industrial & Engineering Chemistry Research, 46(7), 2054–2063.

[9] Hua, X., Tanguy, P. A., Li, R., & Van Wagner, J. S. (1996). Effect of basestock formation on paper coating. TAPPI Journal, 79(5), 112–115.

[10] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.

[11] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

[12] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.

[13] Kuzudisli, C., Bakir-Gungor, B., Bulut, N., Qaqish, B., & Yousef, M. (2023). Review of feature selection approaches based on grouping of features. *PeerJ, 11*, e15666.

[14] Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia, 126*(5), 1763–1768.

[15] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys, 50*(6), Article 94.

[16] Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection – A review and recommendations for the practicing statistician. *Biometrical Journal, 60*(3), 431–449.

[17] Villa-Blanco, C., Bielza, C., & Larrañaga, P. (2023). Feature subset selection for data and feature streams: A review. *Artificial Intelligence Review, 56*(Suppl 1), 1011–1062.

[18] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.

[19] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), Article 20150202.

[20] Lever, J., Krzywinski, M., & Altman, N. (2017). Principal component analysis. *Nature Methods, 14*, 641–642.

[21] Kane, M. T., & Mroch, A. A. (2020). Orthogonal regression, the Cleary criterion, and Lord's paradox: Asking the right questions. *ETS Research Report Series, 2020*(1), 1–24.

[22] Smith, C. A., & Corripio, A. B. (2005). *Principles and practice of automatic process control* (3rd ed.). Wiley.

[23] Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

[24] Hair, J. F., Jr., Babin, B. J., Anderson, R. E., & Black, W. C. (2019). *Multivariate data analysis* (8th ed.). Cengage.

[25] Alonso, A., Negro, C., Blanco, A., & San Pío, I. (2009). Application of advanced data treatment to predict paper properties. *Mathematical and Computer Modelling of Dynamical Systems, 15*(5), 453–462.

[26] Hubbe, M. A., & Koukoulas, A. A. (2016). Wet-laid nonwovens manufacture: Chemical approaches using synthetic and cellulosic fibers. *BioResources, 11*(2), 5500–5552.

[27] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis* (6th ed.). Wiley.

[28] Dahlström, C., & Uesaka, T. (2009). New insights into coating uniformity and base sheet structures. *Industrial & Engineering Chemistry Research, 48*(23), 10472–10478.

[29] Du, Y., Liu, J., Wang, B., Li, H., & Su, Y. (2018). The influence of starch-based bio-latex on microstructure and surface properties of paper coating. *Progress in Organic Coatings, 116*, 51–56.