Ain Shams Engineering Journal 16 (2025) 103565

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: www.sciencedirect.com

Full Length Article

Explainable AI-driven evaluation of plant protein rheology using tree-based and Gaussian process machine learning models

Mustafa Tahsin Yilmaz [a,b], Salman Badurayq [c], Kemal Polat [d,*], Ahmad H. Milyani [e,f], Abdulaziz S. Alkabaa [a], Osman Gul [g], Furkan Turker Saricaoglu [h]

[a] Department of Industrial Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah, Saudi Arabia
[b] Center of Research Excellence in Artificial Intelligence and Data Science (AIADS), King Abdulaziz University, Jeddah, Saudi Arabia
[c] Production Chief at Pladis Global, Phase 3, Industrial Area, Jeddah, Saudi Arabia
[d] Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah, Saudi Arabia
[e] Center of Excellence in Intelligent Engineering Systems (CEIES), King Abdulaziz University, Jeddah, Saudi Arabia
[f] Department of Food Engineering, Faculty of Engineering and Architecture, Kastamonu University, Kastamonu, Turkey
[g] Department of Food Engineering, Faculty of Engineering and Natural Science, Bursa Technical University, Bursa, Turkey
[h] Faculty of Engineering, Department of Electrical and Electronics Engineering, Bolu Abant Izzet Baysal University, Bolu, TR 14280, Turkey

ARTICLE INFO

Keywords:
Explainable artificial intelligence
Sesame protein isolates
Steady shear rheology
Tree-based machine learning models
Gaussian Process regressor

ABSTRACT

In this study, we conducted a comparative analysis of the explainability of Decision Tree Regressor (DTR) and Gaussian Process Regressor (GPR) models in predicting the shear stress and viscosity of sesame protein isolate (SPI) systems, employing explainable machine learning (EML) techniques to elucidate complex, nonlinear relationships among processing parameters. SPI samples were processed across pressure levels ranging from 0 to 100 MPa and ion concentration (IC) values from 0 to 200 mM. DTR model accurately predicted shear stress (R^2 = 0.999), while a GPR model achieved high performance for viscosity prediction (R^2 = 0.9925). Formally, the modeling task is framed as learning a predicting mapping function $f : \mathbb{R}^p \rightarrow \mathbb{R}$, where $x \in \mathbb{R}^p$ denotes the vector of predictors (pressure, IC, shear rate) and $y \in \mathbb{R}$ is the target variable (shear stress or viscosity), by minimizing a loss function such as mean squared error. Interpretation of model predictions using SHapley Additive exPlanations (SHAP), permutation importance, and partial dependence analysis revealed that pressure and IC are the most influential factors affecting shear stress and viscosity, with pressure inducing protein conformational changes that impact rheological properties. The shear rate exhibited a lesser direct impact within the systems examined. Partial Dependence Plots (PDPs) from the DTR model revealed strong, nearly linear positive relationships between pressure and shear stress, while the GPR model depicted more nuanced responses, highlighting the models' differing sensitivities. Variance-Based Sensitivity Indices (VBSIs) further quantified these influences, with pressure and IC showing higher sensitivity scores in the DTR model compared to the GPR model. Permutation importance and SHAP interaction analyses corroborated these results, emphasizing the dominant role of pressure and IC, both independently and interactively, in determining shear stress. In contrast, viscosity predictions were influenced by more distributed and subtle interactions among all features. Employing explainable machine learning techniques enables a comprehensive understanding of feature relevance in complex, nonlinear rheological systems, facilitating the elucidation of viscosity development in sesame protein systems through rheological indices. This approach ensures no bias toward formulation composition and applied pressure, offering valuable insights for optimizing formulation and processing conditions in food applications to enhance the functional properties of SPI-based products.

* Corresponding author.
E-mail addresses: myilmaz@kau.edu.sa (M.T. Yilmaz), Salman.badurayq@pladisglobal.com (S. Badurayq), kpolat@ibu.edu.tr (K. Polat), ahmilyani@kau.edu.sa (A.H. Milyani), allkabaa@kau.edu.sa (A.S. Alkabaa), osmangul@kastamonu.edu.tr (O. Gul), furkan.saricaoglu@btu.edu.tr (F.T. Saricaoglu).

https://doi.org/10.1016/j.asej.2025.103565
Received 9 February 2025; Received in revised form 17 May 2025; Accepted 3 June 2025
Available online 18 June 2025

1. Introduction

Plant-based proteins are gaining attention for their nutritional and functional benefits, with SPI standing out due to its high protein content (>90 %) and useful properties like emulsification, foaming, and gel formation [1,2]. Although sesame is mainly valued for its oil, its protein fraction is increasingly recognized for food applications such as dairy alternatives, meat substitutes, and baked goods, thanks to SPI's unique rheological behavior [3]. Rheological properties of SPI are influenced by factors like high pressure homogenization (HPH), IC, and shear rate. HPH, involving intense shear and cavitation, alters protein structures and affects viscosity and flow behavior [4,5]. This technique has been used to enhance plant protein isolates from tigernut, quinoa, lentil, hazelnut, and sesame [1,6–9]. Likewise, IC variations affect protein solubility and interactions [10]. Understanding these influences is key to optimizing SPI-based formulations.

Although SPI holds strong potential, predicting its rheological behavior under various conditions remains difficult due to complex interactions among proteins, lipids, carbohydrates, and its distinctive biochemical profile. SPI is rich in hydrophobic and sulfur-containing amino acids like methionine and phenylalanine, which promote aggregation and hydrophobic interactions, leading to flow characteristics distinct from soy, pea, or rice proteins [2]. Residual fats and fibers, depending on the extraction method, further influence rheology by increasing viscosity or altering molecular interactions. Its solubility profile is also unique–higher in acidic environments and lowest near the isoelectric point (~pH 4)–affecting suspension stability [2,11]. SPI exhibits both shear-thinning and shear-thickening behavior depending on concentration: it acts a pseudoplastic flow behavior at low concentrations and displays shear-thickening at higher ones, being rare among plant proteins [12]. This duality complicates modeling, as does its time-dependent flow behavior. Unlike soy, which transitions between thixotropy and rheopexy with shear rate [13], SPI shows unpredictable time-dependent responses likely due to distinct network formation mechanisms [12]. Thixotropy–where viscosity decreases over time under constant shear–is particularly relevant but underexplored in SPI systems [3,12]. Recent work suggests that these unpredictable responses may stem from the formation of various protein structures (e.g., amyloid-like fibrils or amorphous aggregates), which depend on conditions like pH and temperature [14]. These potentially irreversible aggregation paths challenge standard thixotropic models and require more sophisticated constitutive equations. Additionally, SPI's interaction with other native components, such as oleosomes, enhances viscosity and gel strength in composite systems [15], making single-component modeling insufficient for real-world food applications. Finally, the relatively large particle size of SPI compared to other plant proteins necessitates specialized rheological measurement tools, as standard protocols often fail to capture its full complexity [12]. These factors underscore the need for customized modeling strategies that account for both structural uniqueness and multi-phase interactions.

Characterizing the rheological behavior of sesame protein–especially under time-dependent and multi-phase conditions–poses significant challenges. In this context, ML is superior for complex, dynamic systems where shear rate variability and interdependencies dominate, provided sufficient data and validation are available, while traditional approaches retain utility in scenarios prioritizing interpretability or involving well-understood, linear relationships. Therefore, ML offers a compelling data-driven alternative to traditional modeling. As a cornerstone of the fourth paradigm of scientific discovery, ML uncovers complex patterns in large datasets and has revolutionized food research [16,17]. In food rheology, ML applications have surged in the past decade [16,18,19], proving effective in modeling diverse rheological properties [16,17]. However, its use in this domain still remains emergent [18]. Supervised learning dominates, with classification algorithms like random forest, decision tree, logistic regression, and SVMs [17,20], and regression methods including linear, multivariate, decision tree, and Lasso regression [17,21]. Artificial Neural Networks (ANNs) stand out for capturing nonlinear relationships in biological systems–for instance, predicting kiwi firmness from nutrient content or pomegranate viscoelasticity from stress-relaxation data [21,22]. Hybrid models, such as ANN combined with Adaptive Neuro-Fuzzy Inference Systems (ANFIS), have also proven effective in modeling complex systems like grape molasses and wild-flower honey [23,24]. ML has even supported product classification for elderly consumers using logistic regression on texture-derived stress values [20].

While machine learning has advanced food rheology through methods like neural networks, logistic regression, and hybrid models, their value depends on strengths such as the interpretability of tree-based models and the uncertainty estimation capabilities of Gaussian Processes (GPMs), alongside comparisons with SVMs and Bayesian optimization. Tree-based models, especially ensembles, are particularly effective on structured data due to their clarity and competitive performance [25,26]. In contrast, GPMs, as non-parametric models, excel in providing uncertainty estimates and flexibility without assuming fixed functional forms [27,28]. GPMs often outperform neural networks on medium-sized tabular data and offer simpler tuning processes [29,30]. Compared to SVR, GPMs provide full predictive distributions rather than point estimates, offering richer insights [31,32]. In food science, model performance varies by task: decision trees perform well in food quality prediction using hyperspectral imaging, while GPR excels in hyperspectral crop monitoring and NIR-based fruit trait prediction, offering lower RMSE and out-of-distribution detection [33,34]. For ingredient optimization, Bayesian optimization with GPs works well in continuous spaces, while Random Forests may be preferable for discrete or high-dimensional settings [35,36].

Nevertheless, strong predictive performance alone does not guarantee the practical utility of a model; particularly in food rheology, where interpretability is key to drawing scientifically valid and trustworthy conclusions. Therefore, in addition to selecting the right model, interpretability remains critical for translating ML predictions into actionable insights because model interpretability plays a crucial role in AI by revealing the hidden mechanisms of complex black-box models and enabling the development of more transparent and dependable systems [37]. Explainable AI (XAI) aims to provide insight into how artificial intelligence models make decisions, ensuring their processes are transparent and understandable [38]. Some approaches to achieving model explainability include techniques such as partial dependence plots (PDP) [39], SHapley Additive exPlanations (SHAP) [40] and permutation importance [41]. SHAP is an influential XAI method that leverages Shapley values from cooperative game theory to quantify each feature's contribution to a model's prediction [42]. This robust theoretical framework facilitates both local and global analyses of model outputs [43], providing contrastive explanations that enhance the interpretability of machine learning models [40]. The partial dependence plot (PDP) illustrates how a single feature–or a pair of features–affects the predicted output of a machine learning model, independent of other features [44]. It helps reveal whether the relationship between the target variable and a given feature is linear, monotonic, or exhibits more intricate patterns [45]. The permutation method is the most efficient model-agnostic approach for estimating feature importance. It works by systematically generating permutations of input features to simulate different feature combinations, allowing for an effective assessment of each feature's impact on model predictions [45]. While each of these techniques offers valuable insights on their own, their individual limitations become more apparent in complex domains like food rheology, highlighting the need for a combined interpretability approach. Each technique alone has critical limitations: conditional Shapley values can sometimes assign non-zero importance to features that the model does not actually use. This occurs because the conditional expectation method may attribute relevance to features that have no direct effect on the prediction—particularly when those features are correlated with others that do influence the output. As a result,

TreeSHAP can produce non-zero estimates even for irrelevant inputs due to underlying feature correlations **Sundararajan and Najmi** [46] and **Janzinget al.** [47]. Moreover, the path-dependent nature of *TreeSHAP* can lead to unintuitive feature attributions. While *TreeSHAP* avoids extrapolation to unlikely data points by altering the value function—shifting from marginal to conditional expectations—it also slightly changes the underlying game. This modification means that even features with no genuine impact on the model's prediction may receive non-zero attributions, highlighting a trade-off between statistical realism and theoretical fidelity [45]. Shapley values can be misinterpreted or deliberately used to create biased explanations that conceal the true behavior of the model [48]. Permutation importance, which assesses the increase in prediction error when a feature's values are randomly shuffled, can underestimate the importance of variables within correlated groups (e.g., moisture and fat content) because the model retains predictive power through associated features, leading to biased rankings. Partial Dependence Plots (PDPs), while useful for visualizing average marginal effects, do not inherently provide importance metrics and may obscure heterogeneous interactions that methods like SHAP can reveal [45].

To better capture the complex rheological behavior of SPI, our study leverages the combined strengths of SHAP, permutation importance, and partial dependence plots (PDPs), offering a more comprehensive and reliable interpretation of feature importance than any single method alone. Integrating these interpretability methods mitigates their individual limitations: PDPs validate SHAP-derived importance rankings by confirming that changes in a feature yield domain-aligned trends (e.g., shear-thinning), while SHAP dependence plots add depth by capturing instance-specific effects. Additionally, permutation importance, when cross-referenced with SHAP and PDP results, helps distinguish genuine feature contributions from artifacts caused by multicollinearity. Building on this integrated interpretability framework, we turn to the modeling approaches themselves, selecting algorithms well-suited to capturing the complex, nonlinear behavior of food systems. Recognizing the proven effectiveness of machine learning in modeling food systems, tree-based and GPR models are employed to predict shear stress and viscosity under varying pressure and IC conditions. Although the integration of these methods improves our insight into feature importance for predicting shear stress and viscosity, it remains important to consider the inherent shortcomings of some explainability approaches—particularly in the context of correlated inputs, as highlighted by the known limitations of TreeSHAP. This limitation underscores the need to explore alternative approaches, such as *KernelSHAP*, which offers unbiased feature importance estimates across various model types, including GPRs. Integrating *KernelSHAP* with GPRs could enhance the interpretability of complex models, especially in scenarios where input features exhibit significant interdependencies [49]. Therefore, investigating the application of *KernelSHAP* in conjunction with GPRs is crucial for achieving more accurate and meaningful explanations in such contexts.

Several studies have applied machine learning to food rheology, particularly using support vector machines (SVM), artificial neural networks (ANN), and linear regression. For example, **Jeong, Kim** [20] used logistic regression to classify products based on texture attributes, while **Torkashvand, Ahmadi** [21] applied ANN to model viscoelasticity of pomegranate juice. However, these approaches generally lacked interpretability or probabilistic insight. In contrast, our study leverages interpretable and probabilistically grounded models; DTR and GPR formalized as functions $f : \mathbb{R}^p \rightarrow \mathbb{R}$ optimized to minimize prediction error. To formalize the scope of the study, let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ represent the input feature matrix consisting of n observations and p predictor variables, including pressure—*P*, *IC*, and shear rate $(\dot{\gamma})$. Let the target variable $y \in \mathbb{R}^n$ denote either shear stress (τ) or viscosity (η). The objective is to learn a regression function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ that minimizes the expected loss over the data distribution:

$$\min_{f \in \mathcal{M}} \mathbb{E}_{(x,y)} \mathcal{D}[\mathcal{L}(y, f(\boldsymbol{x}))]$$

where L denotes a loss function such as MSE, MAE, or RMSE, and $\mathcal{M}$ is the set of candidate models, including DTR, XGBoost and GPR. While tree-based models like DTR make deterministic predictions by minimizing a loss function over splits [50], GPR adopts a probabilistic framework by defining a distribution over functions. A Gaussian Process Regressor assumes that function values follow a joint Gaussian distribution:

$$f(\boldsymbol{x}) \ \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$$

where $m(x)$ is the mean function (commonly set to zero) and $k(x,x')$ is the kernel function defining the covariance between inputs. The model makes predictions by conditioning this prior on the training data and estimating hyperparameters by maximizing the log marginal likelihood [51]. This formulation enables GPR to capture uncertainty and smooth trends in small or moderately sized datasets, which is particularly valuable in food science applications where data are limited or noisy. Model performance is assessed using metrics such as R^2, RMSE, MSE, MAE, and Mean Absolute Percentage Error (MAPE). This mathematical formulation defines the structure and objectives of the predictive modeling task and supports downstream explainability analysis. By integrating SHAP, PDP, and permutation importance, we go beyond black-box prediction to provide feature-level explanations [40], an approach not previously used in modeling rheological behavior of plant-based protein systems.

Building on this formal foundation, our study advances the modeling of food rheology by integrating interpretable and probabilistic learning techniques within a comprehensive explainability framework. In food rheology, where input variables often exhibit strong interdependence and mechanistic interpretability, this multi-method approach ensures a balanced understanding of global patterns, local behaviors, and robustness to correlation—ultimately enhancing trust in model-driven insights. Despite the growing application of machine learning in the food rheology domain, no published study to date has employed explainable machine learning (EML) techniques to interpret rheological properties. To the best of our knowledge, only one study [38] has utilized EML in the broader food science field, focusing on bacterial growth detection in pork meat rather than rheological behavior. Our study represents the first application of explainable machine learning (EML) to assess rheological properties in a food system, offering practical value for plant-based protein research by uncovering complex, data-driven interactions. It introduces an integrated framework combining SHAP, partial dependence plots (PDPs), and permutation importance—an interpretability approach that is still largely untapped in the field of food science and technology.

2. Materials and methods

2.1. Materials

White sesame seeds were sourced from a tahini-producing company (Aslan Sesame and Tahini, Eskişehir, Turkey). Oil was extracted from the seeds using a cold pressing system (Kocmaksan KMS10, İzmir, Turkey), after which the resulting sesame pellets were ground into flour using a laboratory blender (Waring-80011 S, Stamford, CT, USA) and passed through a 212 μm sieve. The sesame cake flour obtained (with 4.32 % moisture, 38.78 % protein, 11.76 % lipid, 6.83 % ash, and 38.31 % carbohydrate content) was then utilized for protein extraction. All chemicals and solvents used in the extraction and analysis processes were of analytical grade.

2.2. Preparation of materials

2.2.1. Preparation of sesame protein isolate (SPI)

The protein extraction from sesame meal followed the method outlined in our previous study **(Gul et al., 2023)**. In this process, sesame meal was mixed with distilled water at a ratio of 1:9 (w/v) and homogenized using an ultra-turra (Velp OV5, Switzerland) at 9000 rpm for 3 min. To solubilize the proteins, the pH of the suspension was adjusted to 10.0 using 1 M NaOH, and the mixture was stirred at 600 rpm on a magnetic stirrer for 2 h. The soluble proteins were then separated by centrifugation (Hettich Universal 320R, Germany) at 9000 rpm for 15 min at 4 °C, producing a supernatant. After centrifuging the entire suspension, the supernatant was collected, and its pH was lowered to 4.5, the isoelectric point of the proteins, using 1 M HCl. The solution was centrifuged under the same conditions to collect the solubilized proteins as a precipitate. The resulting precipitate was stored at −18 °C overnight before undergoing freeze-drying (Teknosem Toros TRS4, İstanbul, Turkey) at −50 °C under vacuum (10^{-3} mbar). The dried proteins were then ground into a fine powder using a coffee grinder and stored in an amber jar at 4 °C until analysis. The protein content of the SPI (90.35 % protein) was measured using a Leco nitrogen analyzer (Leco FP828, MI, USA).

2.2.2. HPH treatment

The SPI was suspended in distilled water at a concentration of 4 % (w/v), and the pH was adjusted to 7.0 using either 1 M NaOH or 1 M HCl. The SPI suspension was then mixed with varying concentrations of NaCl (0, 50, 100, 150, and 200 mM) while being stirred on a magnetic stirrer for 2 h. The SPI suspensions were then processed through a high-pressure homogenizer (GEA Panda Plus 2000, Parma, Italy) at 0 (control), 50, and 100 MPa pressures. The SPI suspensions were collected in a beaker placed in an ice water bath to avoid any temperature rise. The temperature of the SPI suspension was recorded before and after homogenization, measuring 5 °C initially and 37 °C post-homogenization.

2.3. Rheological analysis

The rheological properties of the protein suspensions were evaluated

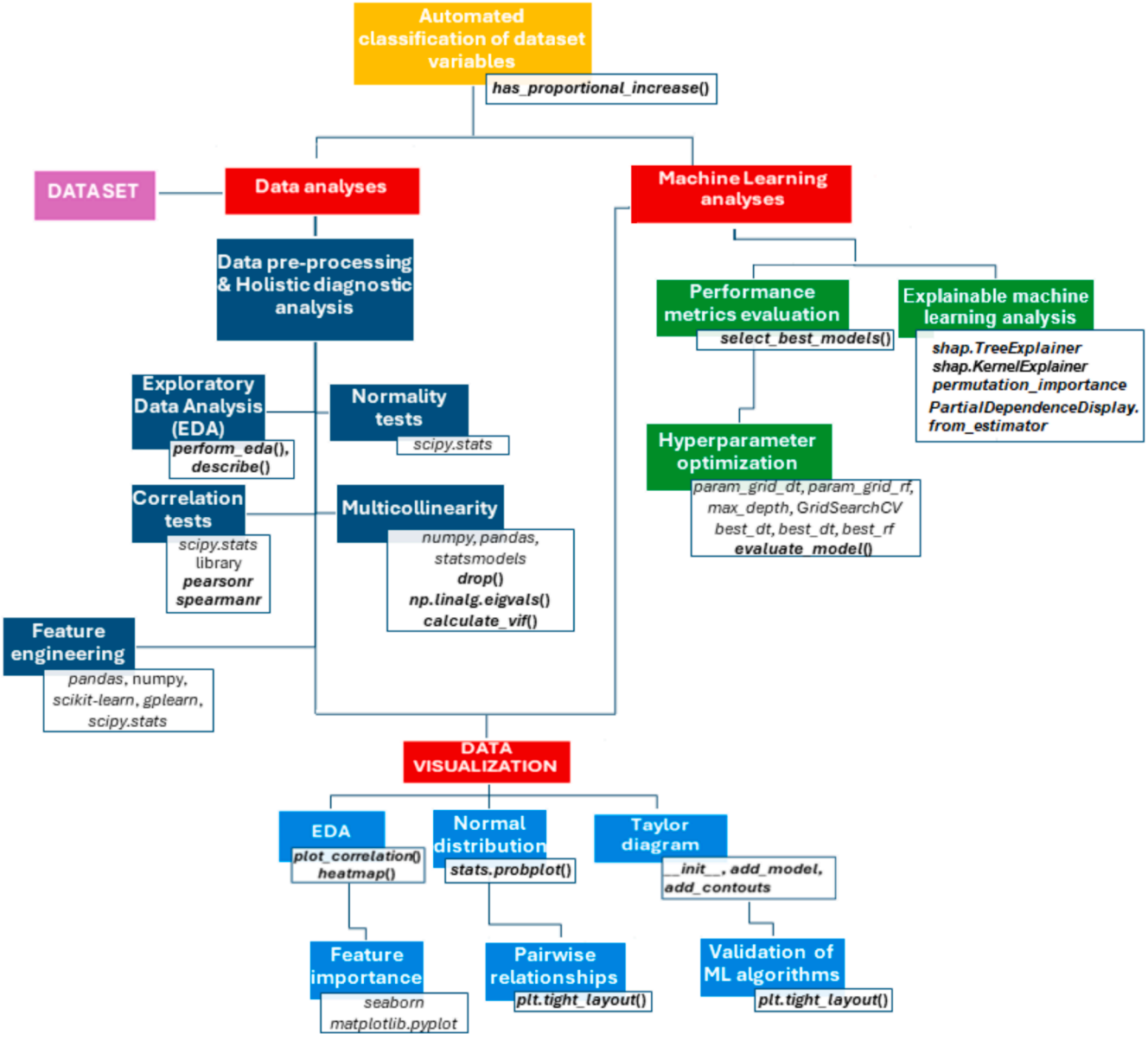


Fig. 1. Explainable machine learning pipeline showing key steps from data preprocessing and exploratory analysis to feature engineering, multicollinearity checks, model training, and data visualization.

using a stress–strain controlled rheometer (Anton Paar, MCR302, Austria) fitted with a Peltier heating system and a plate-plate geometry setup with a 25 mm diameter. The gap between the plates was set to 1 mm. Shear stress versus shear rate curves for the HPH–treated protein suspensions containing NaCl were measured within a shear rate range of 1 to 100 s^{-1} at 25 °C. To better understand the flow behavior of the suspensions, the curves were fitted to the Ostwald de-Waele model.

$$\tau = K(\dot{\gamma})^n \quad (1)$$

where τ, K, $\dot{\gamma}$ and n refer to shear stress (Pa), consistency index (Pa.s^n), shear rate (s^{-1}) and flow behavior index, respectively.

2.4. Methodology pipeline

A detailed methodology pipeline depicting the flow from automated classification of dataset variables to advanced data analysis and visualization is presented in Fig. 1. The pipeline showcases the application of key libraries for data preprocessing, machine learning applications, and data visualization.

2.5. Automated classification of dataset variables

The methodology automatically classified dataset columns as categorical or continuous based on data type and unique value patterns. Numeric columns with more than 10 unique values were labeled as continuous. For columns with 10 or fewer unique values, proportional value increases were used to distinguish continuous from categorical variables.

2.6. Data analysis

2.6.1. Data set

Table 1 presents the experimental design, showcasing the relationship between the input parameters–pressure (MPa), IC, and shear rate [1/s]–and the corresponding outputs–shear stress (Pa) and viscosity (Pa·s). The experiment includes 1162 observations, with changes in pressure—0, 50, and 100 MPa), IC—0, 50, 100, 150, and 200 mM, and shear rate introduced to examine their influence on the shear stress and viscosity of the tested material, resulting in a dataset shape (5 columns × 1162 rows).

2.6.2. Feature selection, data preprocessing, and holistic diagnostic analysis

The features; pressure (MPa), IC (mM), and shear rate [1/s] were selected based on their established mechanistic influence on the rheological and structural properties of plant protein isolates, as documented in colloidal and food science literature. Pressure was included to reflect its role in promoting particle-solvent interactions over particle–particle interactions by mechanically reducing protein particle size through macromolecular fragmentation, thereby improving the functional and rheological properties of plant-based proteins [1]. IC was chosen to evaluate electrostatic effects, where variations in IC screen charges on sesame protein surfaces, impacting solubility and aggregation kinetics via the Hofmeister series [52]. This parameter is critical for mimicking real-world processing conditions (e.g., in emulsified foods or protein gels) where ionic strength governs colloidal stability. Shear rate was incorporated as a fundamental rheological variable, as plant protein dispersions exhibit shear-dependent non-Newtonian behavior; this feature captures dynamic structural rearrangements (e.g., alignment or breakdown of protein aggregates) under flow, essential for modeling viscosity and shear stress These features collectively represent key physicochemical and processing variables that dominate the interfacial and bulk behavior of sesame protein isolate. Their selection aligns with prior studies on plant proteins, ensuring the machine learning models (tree-based and Gaussian process regressors) capture nonlinear interactions between pressure-induced structural changes, ionic environment, and flow dynamics. This approach enables the prediction of macroscopic rheological properties from microscale and process-driven variables, bridging theoretical colloidal science with data-driven modeling. Following the rationale for feature selection based on their physicochemical relevance, the dataset underwent a thorough preprocessing phase to ensure its readiness for machine learning analysis. During the data preprocessing phase, preliminary inspections verified that the dataset was complete and free from any missing or improperly formatted entries. All observations and variables were retained for modeling, ensuring full utilization of the experimental data. The dataset was used in its original form without outlier removal, in order to preserve the integrity of the experimental results and reflect real-world variability in rheological measurements. This ensured that the machine learning models learned directly from the unaltered physical behavior of the material under study. Prior to modeling, basic integrity checks were performed to verify the completeness and consistency of the dataset, confirming the absence of missing or corrupted values. Summary statistics including mean, median, range, and standard deviation were computed for all variables to understand their distribution and spread. Histograms and box plots were created for both input and output variables. These visualizations helped identify potential outliers and assess the overall data spread. However, all values were retained to reflect the true experimental outcomes.

Exploratory Data Analysis (EDA) was carried out to better understand the structure of the dataset, focusing on both categorical and numerical variables. A summary report was generated to highlight key statistics and insights. The normality of the target variable was assessed using a standard statistical test, which helped determine whether the data followed a typical distribution pattern. Relationships between variables were explored using common correlation measures to identify how strongly they were related. To check for potential issues with overlapping or highly related predictors (multicollinearity), the variance structure of the data was analyzed. Specific indicators, such as the condition number and eigenvalues, were used to flag any concerns. Additionally, a common technique called Variance Inflation Factor (VIF) was applied to measure how much each predictor was influenced by the others. The dataset was prepared accordingly, and a custom function was used to compute these values and classify each variable based on its level of multicollinearity.

2.6.3. Feature engineering

This stage involved applying a range of data transformation techniques to enhance the predictive performance of machine learning models. Transformations were applied individually to both input features and the target variable using standard methods, including Min-Max Scaling, Z-score Standardization, Log Transformation, Robust Scaling, Max Abs Scaling, Box-Cox transformation, Yeo-Johnson transformation, square-root transformation, rank transformation, and logit

Table 1
Experimental design with input parameters and corresponding output measurements.

Observations	Inputs			Outputs	
	Pressure (MPa)	IC (mM)	Shear rate [1/s]	Shear stress [Pa]	Viscosity [Pa·s]
1	0	0	3.89	0.0157	0.0040
2	0	0	5.16	0.0228	0.0044
3	0	0	6.42	0.0314	0.0049
4	0	0	7.69	0.0378	0.0049
5	0	0	8.95	0.0442	0.0049
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
1158	100	200	94.9	5.3097	0.0559
1159	100	200	96.2	5.2921	0.0550
1160	100	200	97.5	5.3275	0.0547
1161	100	200	98.7	5.3648	0.0543
1162	100	200	100.0	5.3540	0.0535

transformation. These transformations were applied to assess whether transformation of the input and output variables could improve the R^2 scores of the ML models. The goal was to enhance the distributional properties of the data and reduce the influence of outliers and skewness.

2.7. Machine learning analysis

The methodology involved evaluating and comparing various ML models to predict outcomes based on the input features, pressure, IC, and shear rate. The ML models used included both linear and non-linear algorithms. First, the dataset was split into training and testing sets with a 80–20 ratio. Each model was trained on the training data, and predictions were made on both the training and testing sets.

2.7.1. Performance metrics

The performance of the models was assessed using several metrics; R^2 (determination coefficient), root mean squared error (RMSE), mean squared error (MSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The methodology involved selecting and ranking the top-performing machine learning models based on various performance metrics. First, the performance of all the models was evaluated and normalized using a custom function.

2.7.2. Composite performance scoring

To evaluate and compare the predictive performance of various regression models, multiple error metrics—including *R*2, RMSE, MSE, MAE, and MAPE—were calculated separately for both the training and test datasets. Each metric was normalized across models to ensure comparability, with higher R2 and lower error values scaled appropriately. An overall performance score was then derived for each model by averaging the ten normalized metric values (five from training and five from testing), thereby summarizing the model's generalization capability into a single composite score. These scores were subsequently used to rank the models. For visual interpretation, Taylor diagrams were employed, mapping each model's overall score onto both the radial (standard deviation) and angular (correlation) axes to enable simultaneous assessment of accuracy, consistency, and similarity to the ideal reference.

2.7.3. Hyperparameter optimization for best-performing models

In this study, we applied hyperparameter optimization and performance evaluation for Decision Tree, XGBoost, and Gaussian Process regressors to model shear stress. The dataset was divided into training and testing sets. Grid search cross-validation, performed over five folds, was used to tune model hyperparameters for each model. For the Decision Tree Regressor, parameters included variations in *max_depth*, *min_samples_split*, and *min_samples_leaf*; for XGBoost, *n_estimators*, *max_depth*, *learning_rate*, and *subsample*. The depth of the decision trees in the code was determined through the grid search over a set of predefined max_depth values: [None, 5, 10, 15, 20]. For each value of *max_depth*, in combination with different *min_samples_split* values, the model was evaluated using 5-fold cross-validation on the training data. The selection criterion was based on the average cross-validation score across the folds–typically the R^2 score unless otherwise specified. The *max_depth* value that yields the highest average cross-validation performance was chosen as the optimal tree depth. For the Gaussian Process Regressor, combinations of kernels (*RBF*, *RationalQuadratic*, and *Matern*) with constant kernels, alongside values for *alpha* and *n_restarts_optimizer* were used to mitigate overfitting. The following kernel functions were selected because they each represent different assumptions about the structure and smoothness of the underlying function:

- RBF: for smooth and global behavior.
- RationalQuadratic: for capturing patterns across multiple length scales.
- Matern ($\nu = 1.5$): for rougher or more irregular functions.

All three in a grid search were used because this approach allows the model to adapt to the nature of the data, choosing the kernel that best fits the observed patterns during cross-validation. The most suitable kernel was selected based on yielding the best average cross-validation score (typically R^2) on the training data during grid search. These models were then re-evaluated using the performance metrics–R^2, MSE, RMSE, MAE, and MAPE, to assess predictive accuracy and error rates on both training and test sets. Results, including the best hyperparameters and cross-validated scores, were printed, offering a comprehensive comparison of the performance of each model.

2.8 SHAP analysis and explainable AI

In this study, a comprehensive set of explainable AI techniques—including SHAP analysis, Partial Dependence Plots, and Permutation Importance—was employed to yield complementary insights into both global feature effects and localized model behavior at the instance level, using a Decision Tree Regressor (DTR) and Gaussian Process Regressor (GPR) which were selected due to their superior overall performance score—as shown by Taylor diagrams—among the machine learning models tested.

SHAP Analysis: DTR and GPR were trained to predict shear stress and viscosity, respectively, with SHAP used to interpret individual feature contributions. The Decision Tree, being a model-specific and transparent algorithm, was interpreted efficiently using the model-aware *TreeExplainer*. In contrast, the GPR–treated as a black-box, model-agnostic method due to its kernel-based structure–was analyzed using the more computationally intensive *KernelExplainer*. SHAP values were calculated on the test dataset using *shap.Explainer* for the models, and *shap.KernelExplainer* for the Gaussian Process Regressor (GPR). Four types of SHAP visualizations were produced:

(i) global SHAP summary bar plots were used to present the average absolute SHAP values across all features,

whereas **(ii)** 3D SHAP interaction scatter plots illustrated how IC, pressure, and shear rate jointly influenced the magnitude of SHAP values within a three-dimensional feature space.

$$\mathrm{SHAP}_i(f) = \sum_{S \subseteq F\{i\}} \frac{(|S|!(F) - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \tag{2}$$

where $\mathrm{SHAP}_i(f)$ represents the SHAP value for feature i in model. $S \subseteq F\{i\}$ means that S includes all features F excluding feature i. Consequently, the model estimates the mean influence of feature i.

In DTR, SHAP values can be computed efficiently using *TreeExplainer*, which leverages the internal structure of the tree to compute exact Shapley values without sampling or surrogate modeling [53]:

For a given input instance $x = (x_1, \cdots, x_M)^T \in \mathbb{R}^M$, SHAP explains the prediction f(x) as:

$$f(x) \approx \phi_0 + \sum_{j=1}^{M} \phi_j \tag{3}$$

where:

$\phi_0 = \mathbb{E}_x[f(x)]$ is the expected output of the model over the training data (the base value),

$\phi_0 \in \mathbb{R}$ is the SHAP value representing the contribution of feature j to the prediction for input x,

The sum $\phi_{0+}\sum_j \phi_j$ provides an additive decomposition of the prediction.

The global feature importance across the dataset can be obtained by averaging the absolute SHAP values:

$$I_j = \frac{1}{n} \sum_{i=1}^{n} \left| \phi_j^{(i)} \right| \tag{4}$$

where $\phi_j^{(i)}$ is the SHAP value of feature j for sample i, and n is the number

of instances.

To estimate SHAP values for black-box models, *KernelExplainer* fits a local linear surrogate model $g(z')$ around the prediction point x, where z' is a binary vector representing feature inclusion. The contribution of each sampled coalition z' is weighted using the kernel function [40]:

$$\pi_x(z') = \frac{(M-1)}{\binom{M}{|z'|} \bullet |z'| \bullet (M-|z'|)} \tag{5}$$

Despite the underlying model (GPR) being nonlinear, SHAP uses a locally linear surrogate model in the binary coalition space to estimate feature attributions. This model is defined as:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{6}$$

where $z'_j \in \{0,1\}$ indicates the inclusion of feature j. This formulation enables SHAP to compute additive explanations, consistent with Shapley value theory. Then, we train the linear model g by optimizing the following loss function L:

$$L(\hat{f}, g, \pi_x) = \sum_{z' \in Z} [\hat{f}(h_x(z')) - g(z')]^2 \pi_x(z') \tag{7}$$

where **Z** is the training data. z' is a binary vector indicating/presence/absence of features,

$h(z')$ maps binary masks to full feature inputs,

$\hat{f}(h_x(z'))$ is the model prediction with simulated missing values.

$\pi_x(z')$ is a kernel weighting function emphasizing proximity to x.

In our work, *KernelExplainer* estimates SHAP values by fitting a local linear surrogate model $g(z')$ around the instance x, using weighted sampling over feature coalitions. The SHAP values ϕ_j are obtained by minimizing a weighted least squares loss, ensuring that the surrogate model closely approximates the original model's predictions across different subsets of features.

$$\min_g \sum_{z'} \pi_x(z')[f(h(z')) - g(z')]^2 \tag{8}$$

where $\pi_x(z')$ is a Shapley kernel that emphasizes balanced subsets, and $h(z')$ is a function that fills in missing features using background data.

(iii) To move beyond standard feature attribution, we analyzed pairwise SHAP interaction effects (using TreeExplainer) for both target variables—shear stress and viscosity. This approach was necessary because SHAP interaction values could not be computed for the Gaussian Process Regressor (GPR) due to limitations of the *KernelExplainer*, which does not support interaction decomposition. In contrast, *TreeExplainer* is fully compatible with decision tree models and enables efficient and exact calculation of both main effects and pairwise SHAP interaction values. This method decomposes prediction contributions not only into individual (main) effects but also into pairwise interaction effects between features. It enables a richer understanding of how features jointly influence the model's output.

In pairwise SHAP decomposition, the SHAP interaction tensor $\phi_{i,j,k}$ was computed, where each entry quantifies the interaction between features j and k for observation i. These values were post-processed to produce two categories:

- Main effects ($j = k$): Capturing the isolated contribution of each feature.
- Interaction effects ($j \neq k$): Representing symmetric co-dependencies, computed as the sum of $\phi_{i,j,k} + \phi_{i,k,j}$. For $i \neq j$, the interaction value is given by [45]:

$$\phi_{i,j} = \sum_{S \subseteq F \setminus \{i,j\}} \frac{|S|!(M-|S|-2)!}{2(M-1)!} \delta_{ij}(S) \tag{9}$$

where,

$$\delta_{ij}(S) = \hat{f}_x(S \cup \{i,j\}) - \hat{f}_x(S \cup \{i\}) - \hat{f}_x(S \cup \{j\}) + \hat{f}_x(S) \tag{10}$$

(iv) To explore feature importance across varying operating conditions, we performed subgroup-specific SHAP analysis. The test dataset was partitioned based on unique combinations of pressure and IC values. For each subgroup, SHAP values were recomputed and visualized using beeswarm plots. This enabled the identification of local shifts in feature relevance that might be obscured in global analyses, providing deeper insight into model behavior under different regimes. Subgroup-specific SHAP plots, generated based on unique combinations of Pressure and IC levels, revealed how feature importance varied across different operating conditions.

Let D be the full test dataset; $G_{P=0,IC=0} = \{x \in D | x_{\text{Pressure}} = 0, x_{\text{IC}} = 0\}$, Then, the individual SHAP values plotted are $\phi_j^{(i)}$ for each $i \in G_{P=0,IC=0}$, so the plot does not show the aggregated importance $I\phi_j^{(G_k)}$ per se; rather it shows the distribution of SHAP values $\phi_j^{(i)}$ within that subgroup, as shown below:

$$I\phi_j^{(G_k)} = \frac{1}{|G_k|} \sum_{i \in G_k} \left| \phi_j^{(i)} \right| \tag{11}$$

Partial Dependence Plots (PDPs): Partial Dependence Plots (PDPs) illustrate the marginal effect of one or two features on a model's predicted outcome by averaging predictions over the dataset; in this study, PDPs were generated to visualize the individual contribution of each predictor to the model output. The PDP for feature x_j is defined as the expectation over the marginal distribution of the remaining features X_{-j}, as shown in Eq. (12). Here, $\hat{f}$ denotes the trained model, and X_{-j} represents all input features except X_j.

$$PDP_j(x_j) = \mathbb{E}_{X-j}[\hat{f}(x_j, \boldsymbol{X}_{-j})] \tag{12}$$

The partial dependence function for regression is further formalized in Eq. (13) [45], where $\hat{f}_S(\boldsymbol{x}_S)$

Denotes the PDP for a subset of features $\boldsymbol{x}_S$, and the expectation is taken over the complement feature subset $\boldsymbol{x}_C$. The term $\mathbb{P}(\boldsymbol{X}_C)$ represents the marginal distribution of the complement features:

$$\hat{f}_S(\boldsymbol{x}_S) = \mathbb{E}_{X_C}[\hat{f}(\boldsymbol{x}_S, \boldsymbol{X}_C)] = \int \hat{f}(\boldsymbol{x}_S, \boldsymbol{X}_C) d\mathbb{P}(\boldsymbol{X}_C) \tag{13}$$

In practical implementation, this expectation is approximated using the empirical average over the dataset, as shown in Eq. (14), where $\boldsymbol{x}_C^{(i)}$ is the i-th sample from the dataset:

$$\hat{f}_S(\boldsymbol{x}_S) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}\left(\boldsymbol{x}_S, \boldsymbol{x}_C^{(i)}\right) \tag{14}$$

Our approach is a deterministic, grid-based approximation of first-order (main effect) sensitivity using PDP variance. To quantify the global influence of each numerical feature on the model's output, a variance-based global sensitivity index was computed as the standard deviation of the model's partial dependence values across a range of input values, using the following formulation Eq. (15).

$$I(\boldsymbol{x}_S) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K} \hat{f}_S\left(\boldsymbol{x}_S^{(k)}\right) - \frac{1}{K}\sum_{k=1}^{K} \hat{f}_S\left(\boldsymbol{x}_S^{(k)}\right)^2} = \sqrt{Var(\hat{f}_S(\boldsymbol{x}_S))} \tag{15}$$

Here, the first term represents the mean squared PDP value, and the second term is the square of the mean PDP value. $\hat{f}_S\left(\boldsymbol{x}_S^{(k)}\right)$ represents the model evaluations at those points. This equation defines the PDP-based feature importance score $I(\boldsymbol{x}_S)I(\boldsymbol{x}_S)$, computed as the standard deviation of the PDP curve across K sampled values of $\boldsymbol{x}_S$.

Comparison of SHAP feature importance with PDP-based importance: To better understand and validate model interpretability, SHAP feature importance was also quantitatively compared with PDP importance. SHAP values were extracted from the explanation object or used directly, and global SHAP importance was computed as the mean of absolute values across all test samples. This comparison is essential, as SHAP captures local, instance-level contributions while PDP reflects global, average effects—highlighting potential discrepancies or complementarities between the two interpretability methods and ensuring a more robust understanding of feature influence.

Permutation feature importance (PFI): Permutation PFI was computed to assess each feature's impact on the model's predictive performance by measuring the increase in prediction error after permuting the feature's values—breaking its relationship with the true outcome. We follow the permutation importance framework introduced by [54], computing the drop in the model's R^2 score after randomly permuting each feature. To account for randomness, the permutation is repeated 30 times. The final importance score is defined as the mean drop in R^2, and the standard deviation across permutations is used as an uncertainty measure, as implemented in *sklearn.inspection.permutation_importance*. To formalize the computation of permutation feature importance (PFI), let $S(f, \mathscr{D})$ denote the performance of a trained model f on dataset $\mathscr{D}$, using a scoring function S, such as the coefficient of determination R^2. For a given feature j, let $\mathscr{D}^{\pi(j)}$ represent the dataset in which the values of feature j are randomly permuted according to a permutation π. The expected importance of feature j is then defined as the drop in performance due to permuting j:

$$FI_j = \mathbb{E}_\pi\left[S(f, \mathscr{D}) - S\left(f, \mathscr{D}^{\pi(j)}\right)\right] \tag{16}$$

In practice, this expectation is approximated by repeating the permutation process $\boldsymbol{K}$ times with independently drawn permutations $\pi_1, \cdots, \pi_K$, and computing the empirical mean:

$$\widehat{FI_j} = \frac{1}{K}\sum_{k=1}^{K}\left[S(f, \mathscr{D}) - S\left(f, \mathscr{D}^{\pi_k(j)}\right)\right] \tag{17}$$

To quantify uncertainty in the importance estimate, we compute the standard deviation across repetitions:

$$\hat{\sigma}_j = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\left[S(f, \mathscr{D}) - S\left(f, \mathscr{D}^{\pi_k(j)}\right)\right] - \widehat{FI_j}\right)^2} \tag{18}$$

This formulation reflects the procedure implemented in *sklearn.inspection.permutation_importance*, which we adopt in this study with $K = 30$ repeated permutations.

2.9. Interpreter and integrated development environments (IDEs)

Python served as the interpreter, allowing the code to be easily created, tested, debugged, refactored, and analyzed. Visual Studio Code, as the IDE, was utilized to develop the models.

3. Results and discussion

3.1. Classification of variables in data set preprocessing

Python is considered a valuable tool for data pre-processing [55]; therefore, it was used to analyze feature types, revealing that all variables in the dataset—pressure, IC, shear rate, shear stress, and viscosity—were continuous, with no categorical columns. Notably, pressure and IC had discrete but ordered and proportional levels (e.g., 0, 50, 100 MPa for pressure), confirming their treatment as continuous variables. This classification was important, as applying one-hot encoding would have disrupted the ordinal structure and weakened the model's ability to interpret treatment magnitudes. Since all variables were continuous, preprocessing was simplified, and models like linear regression and tree-based methods could directly exploit the quantitative relationships without the need for encoding, enhancing predictive accuracy in regression tasks.

3.2. Data pre-processing and diagnostic analyses

3.2.1. EDA (Exploratory data Analysis)

The dataset comprises 1,162 rows and 4 continuous variables: pressure, IC, and shear rate as inputs, with shear stress or viscosity as outputs. There are no missing values. Pressure and IC range from 0 to 100 MPa and 0 to 200 mM, respectively, while shear rate spans 0.1 to 100 1/s. Shear stress and viscosity range from 0.0007 to 5.66 Pa and 0.00334 to 0.6215 Pa·s, with notable outliers detected in both output variables. Handling these extreme values is essential for building reliable predictive models.

3.2.2. Normality tests

The analysis of shear stress and viscosity data using the Shapiro-Wilk and Kolmogorov-Smirnov tests indicated that both variables did not conform to a normal distribution. For shear stress, the Shapiro-Wilk test yielded a test statistic of 0.765 ($P < 0.01$), while the Kolmogorov-Smirnov test produced a test statistic of 0.199 ($P < 0.01$). Viscosity showed even more substantial evidence against normality, with a Shapiro-Wilk test statistic of 0.523 ($P < 0.01$) and a Kolmogorov-Smirnov test statistic of 0.301($P < 0.01$). The Q-Q plots for shear stress and viscosity further supported these findings (Fig. 2). In the Q-Q plot for shear stress, the data points deviated noticeably from the reference line, especially in the tails, indicating non-normality and skewness (Fig. 2**a**). Similarly, the Q-Q plot for viscosity showed a pronounced divergence from the theoretical quantiles, particularly in the upper tail (Fig. 2**b**), reinforcing the conclusion that viscosity did not follow a normal distribution. Given the precise non-normal distribution of shear stress and viscosity data, as evidenced by the Shapiro-Wilk and Kolmogorov-Smirnov tests and further supported by Q-Q plot deviations, conducting normality tests before machine learning algorithm selection is crucial. This step ensures that the algorithms chosen align with the data distribution characteristics, thereby enhancing model accuracy and reliability in capturing the actual rheological behavior of such variables [56].

3.2.3. Correlation tests

Pearson and Spearman correlation analyses (Fig. 2**c and d**) revealed no significant linear or monotonic relationships among the input variables—pressure, IC, and shear rate. All correlation coefficients were close to zero and statistically insignificant (e.g., pressure–IC: $r = -0.017$, $p = 0.561$), confirming their independence in the experimental setup. This independence is advantageous for modeling, allowing each variable to be adjusted independently without introducing confounding effects.

3.2.4. Multicollinearity tests

Multicollinearity analysis (based on eigenvalues, condition number, and VIF) indicated no serious collinearity issues among the independent variables—pressure, IC, and shear rate. Eigenvalues were close to 1, suggesting orthogonality among variables. The condition number was 1.0398, well below the threshold of 30, confirming a very low risk of multicollinearity [57]. VIF values for pressure (2.02) and IC (2.21) were within acceptable limits ($VIF < 5$), indicating only moderate and tolerable multicollinearity [57]. Overall, the results support the inclusion of all variables in the predictive model without redundancy concerns.

3.2.5. Pairwise relations

As shown in Fig. 3**a–d**, pressure was the dominant factor influencing both shear stress and viscosity, while IC had a comparatively minor effect. Fig. 3**a** illustrates a clear positive relationship between pressure and shear stress, whereas IC showed minimal impact. Fig. 3**b** confirms that

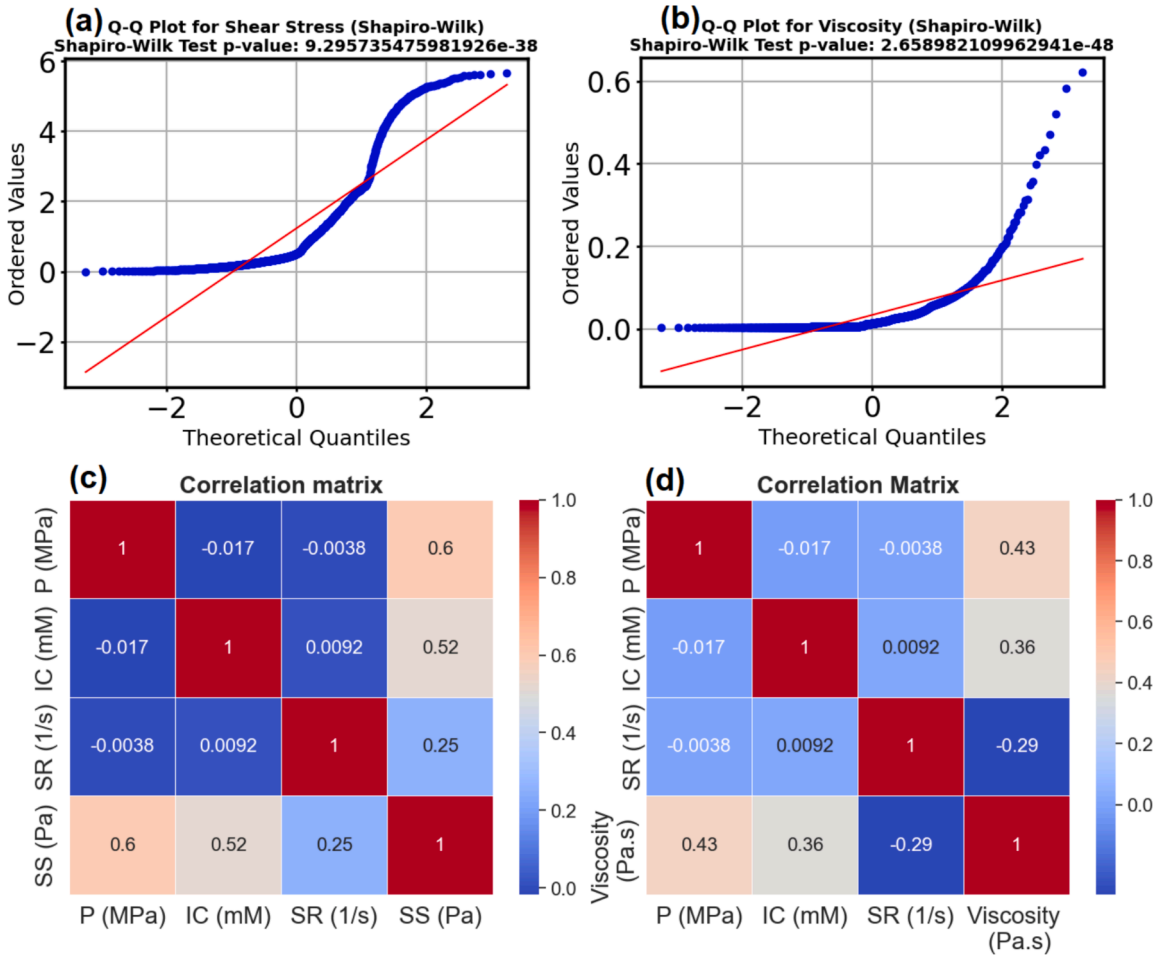


Fig. 2. Q-Q plots for **(a)** shear stress and **(b)** viscosity, illustrating deviations from normal distribution and correlation matrices showing relationships between input variables (pressure-P, ion concentration–IC, and shear rate-SR) for **(c)** shear stress-SS and **(d)** viscosity.

shear stress increases with shear rate, with pressure significantly amplifying this effect, while IC caused only slight variations. **In** Fig. 3**c**, viscosity also increased with pressure but was only marginally affected by IC. Fig. 3**d** reveals shear-thinning behavior across all conditions, with viscosity decreasing as shear rate increased—especially under high pressure—highlighting non-Newtonian fluid characteristics [58].

Fig. 4**a and 4b** illustrate the effects of pressure and IC on shear stress and viscosity across all shear rates. In Fig. 4a, shear stress increased nearly linearly with shear rate, with higher pressures producing steeper slopes, while IC had minimal impact within each pressure group. Fig. 4b shows clear shear-thinning behavior in viscosity for all combinations, with viscosity decreasing as shear rate increased. Higher pressure conditions (especially at 100 MPa) consistently led to higher viscosity values across shear rates, indicating increased resistance to flow and suggesting pressure as a key factor in controlling fluid stability.

The results from Figs. 3 and 4 confirm the fluid's pressure-dependent rheological behavior. Lower pressures led to reduced viscosities, indicating less flow resistance, while higher pressures increased viscosity, enhancing dissipation and potentially accelerating fluid heating during flow—consistent with findings of **Baranov** [59]. IC had a secondary effect, with only slight viscosity increases at higher concentrations. Overall, pressure emerged as the primary factor influencing both shear stress and viscosity, emphasizing its importance in applications requiring precise control over flow properties.

3.2.6. *Feature engineering*

Initial normality tests (Shapiro-Wilk and Kolmogorov-Smirnov) indicated non-normal distribution in the output data. Various transformation methods were applied to both input and output variables to enhance model performance, but none improved the R^2 scores. This suggests that the original data scale was more effective for capturing predictive patterns, and the analysis continued using the non-transformed variables.

3.3. *Machine learning analysis*

3.3.1. *Comparison of ML algorithms based on performance metrics*

Table 2 shows that the Decision Tree, XGBoost, and Random Forest Regressors achieved the highest performance in predicting shear stress, with R^2 scores close to 1, and very low RMSE, MSE, MAE, and MAPE values—for example, the Decision Tree Regressor yielded an RMSE of 0.0437 and a near-zero MAPE. These results reflect the models' strong ability to capture complex non-linear relationships. In contrast, Linear Regression showed poor performance across all metrics. For viscosity prediction, the Gaussian Process Regressor performed best, with R^2 scores of 0.9925 (train) and 0.9965 (test), and low error values including a test MAPE of 8.72. The Decision Tree and XGBoost Regressors also performed well, though with slightly higher error rates. These findings reinforce the suitability of tree-based and Gaussian Process models for accurately modeling both shear stress and viscosity, especially when accounting for multiple error metrics.

The Taylor diagrams in Fig. 5 visualize the overall performance of machine learning models for predicting shear stress and viscosity, respectively, by comparing correlation, standard deviation, and RMSE.

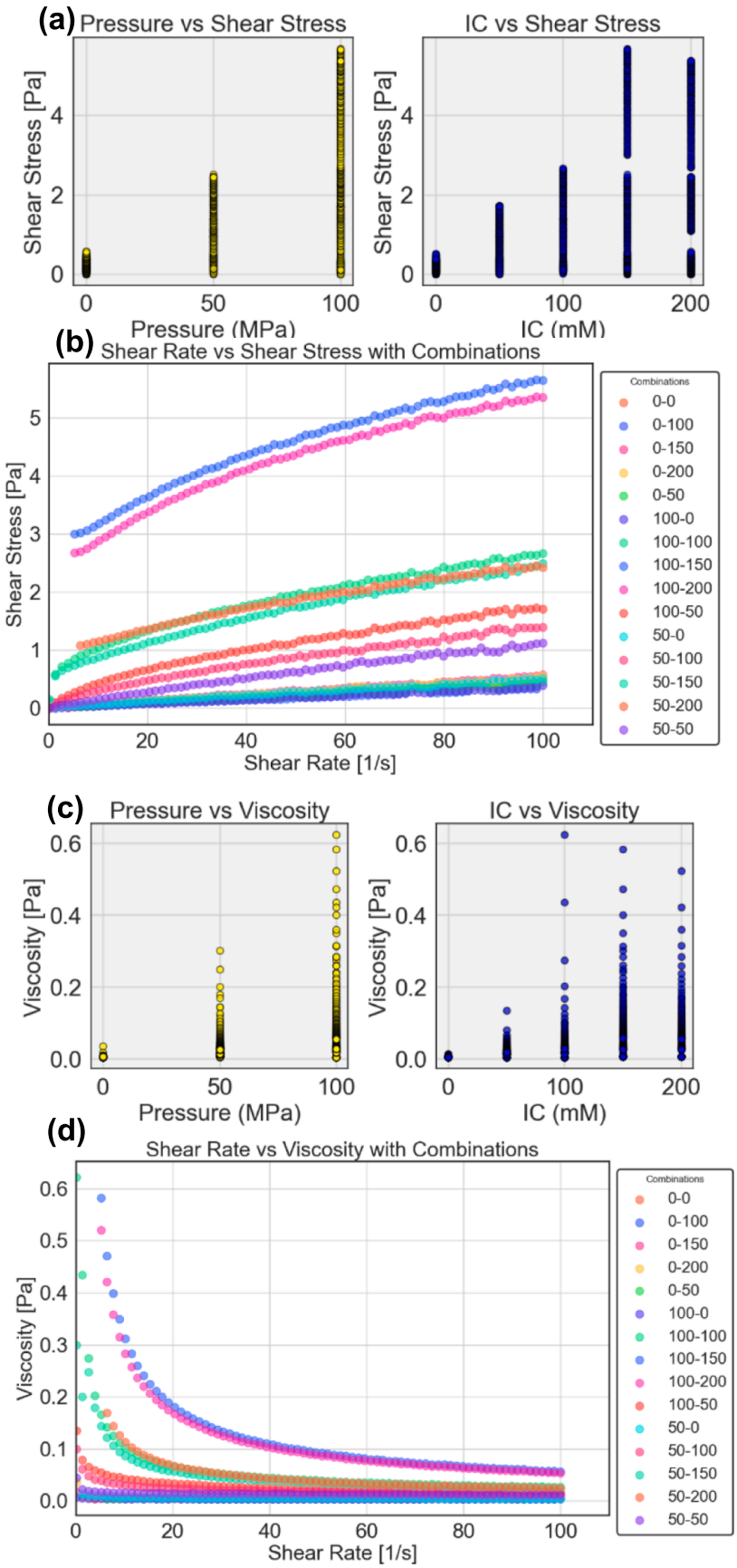


Fig. 3. The relationship between **(a)** pressure versus Shear Stress and IC versus Shear Stress, **(b)** shear rate versus shear stress at different combinations, **(c)** pressure versus viscosity and IC versus viscosity and **(d)** shear rate versus viscosity at different combinations.

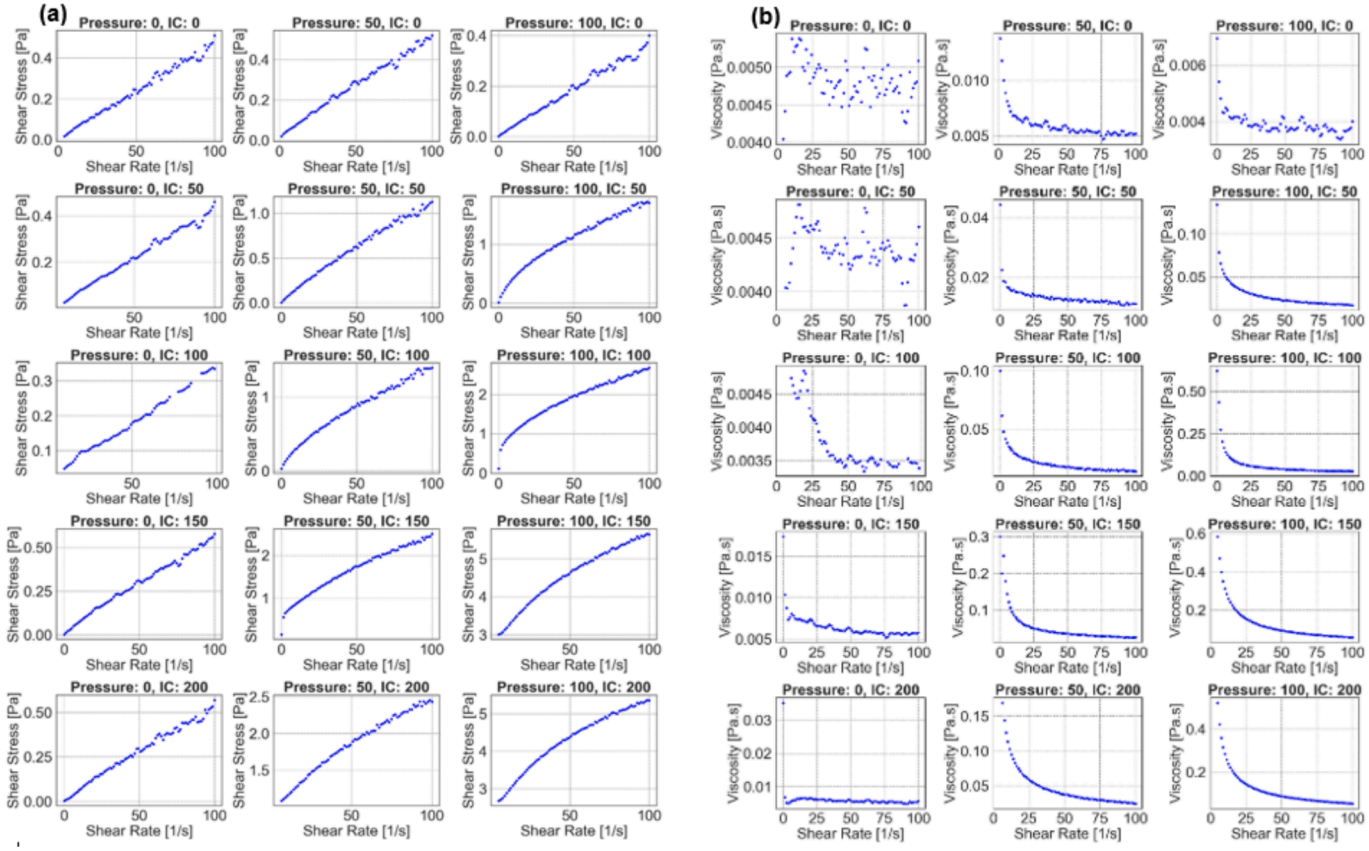


Fig. 4. Shear stress (a) and viscosity (b) versus shear rate across 15 unique combinations of pressure (0, 50, and 100 MPa) and IC levels (0, 50, 100, 150, and 200 mM). Each plot represents a distinct pairing of pressure and IC, highlighting the individual response of shear stress and viscosity to shear rate under the conditions.

Table 2
Performance metrics of various machine learning algorithms in predicting shear stress and viscosity, evaluated through training and testing datasets.

ML algorithms	Performance metrics† R2 Train	Test	RMSE Train	Test	MSE Train	Test	MAE Train	Test	MAPE Train	Test
Shear stress										
Decision Tree Regressor	1	0.999	0	0.0437	0	0.0019	0	0.0237	0	6.8842
XGBoost Regressor	1	0.9989	0.0079	0.0458	0.0001	0.0021	0.0054	0.0265	1.3423	6.684
Random Forest Regressor	0.9999	0.999	0.0122	0.0446	0.0001	0.002	0.0072	0.0205	3.2086	6.185
K-Nearest Neighbors Regressor	0.9997	0.9987	0.025	0.0496	0.0006	0.0025	0.0131	0.0196	10.8997	11.028
LightGBM Regressor	0.9995	0.9982	0.0332	0.0592	0.0011	0.0035	0.0189	0.032	30.8971	34.293
Gradient Boosting Regressor	0.9984	0.9956	0.0577	0.0925	0.0033	0.0086	0.0418	0.062	48.0288	43.167
Support Vector Regressor	0.9965	0.9949	0.0854	0.0995	0.0073	0.0099	0.0727	0.0752	50.7856	48.232
Multi-layer Perceptron	0.9628	0.9338	0.2799	0.3582	0.0784	0.1283	0.159	0.202	74.6607	71.878
Polynomial Regression	0.9368	0.9302	0.365	0.3677	0.1333	0.1352	0.2607	0.2621	206.658	186.854
Gaussian Process Regressor	1	0.5415	0	0.9425	0	0.8882	0	0.5039	0	41.014
Linear Regression	0.7022	0.6764	0.7924	0.7917	0.6278	0.6268	0.6196	0.6327	355.567	349.570
Viscosity										
Gaussian Process Regressor	1	0.9925	0	0.005	0	0	0	0.0018	0.0012	8.7176
Decision Tree Regressor	1	0.9443	0	0.0136	0	0.0002	0	0.0022	0	3.6982
XGBoost Regressor	0.9998	0.9443	0.0008	0.0136	0	0.0002	0.0005	0.0023	3.062	4.7552
Random Forest Regressor	0.9948	0.8851	0.0042	0.0196	0	0.0004	0.0008	0.0025	1.7793	3.6548
Gradient Boosting Regressor	0.9879	0.9333	0.0064	0.0149	0	0.0002	0.0037	0.0052	32.5987	41.7008
K-Nearest Neighbors Regressor	0.9612	0.8030	0.0114	0.0257	0.0001	0.0007	0.0017	0.0029	2.3681	3.5971
LightGBM Regressor	0.9297	0.7389	0.0154	0.0295	0.0002	0.0009	0.004	0.0064	30.0636	50.5231
Polynomial Regression	0.7123	0.5629	0.0312	0.0382	0.001	0.0015	0.0174	0.0172	159.894	159.760
Linear Regression	0.4353	0.3039	0.0436	0.0482	0.0019	0.0023	0.025	0.025	245.706	270.230
Support Vector Regressor	0.0834	−0.0873	0.0556	0.0603	0.0031	0.0036	0.0478	0.0491	578.504	594.309
Multi-layer Perceptron	−1.6219	−1.3922	0.094	0.0894	0.0088	0.008	0.0683	0.0656	566.418	521.112

† Metrics include R^2 (coefficient of determination), RMSE (root mean square error), MSE (mean squared error), MAE (mean absolute error), and MAPE (mean absolute percentage error).

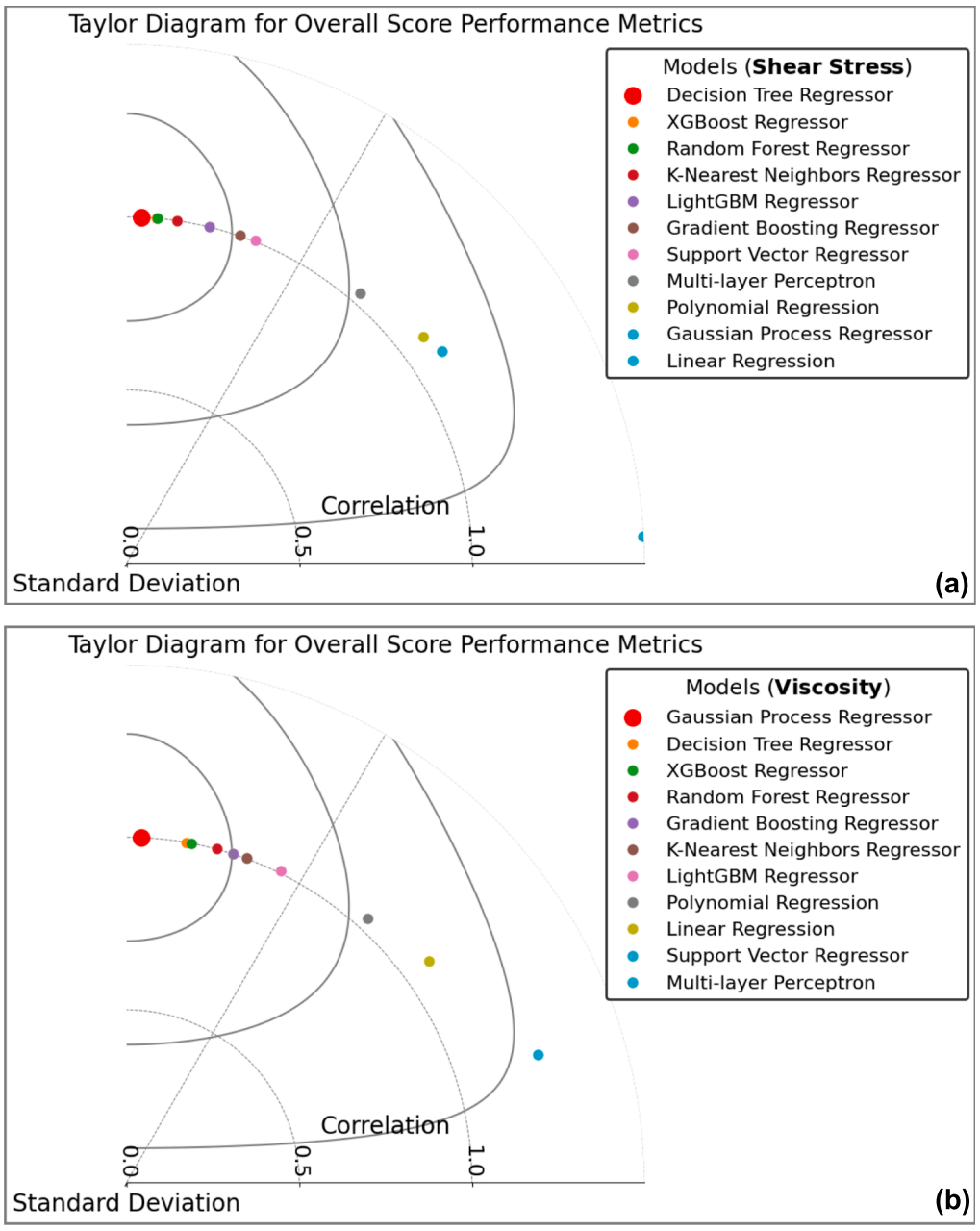


Fig. 5. Taylor Diagrams for overall score performance metrics of machine learning models. Performance of models predicting **(a)** Shear Stress and **(b)** Viscosity with various regression techniques, illustrating the relationship between predicted and actual values for different models.

In Fig. 5**a**, the Decision Tree Regressor achieved the highest overall score for shear stress prediction, showing the strongest correlation with actual values, lowest RMSE, and a standard deviation closest to the reference. XGBoost and Random Forest also performed well, though with slightly higher RMSE values. Simpler models such as Linear Regression and Gaussian Process Regressor showed lower correlations and greater deviations. In Fig. 5**b**, the Gaussian Process Regressor emerged as the best model for viscosity prediction, exhibiting the highest correlation, lowest RMSE, and excellent alignment with the reference standard deviation. Decision Tree and XGBoost Regressors also performed strongly but with marginally higher RMSE. Models like Support Vector Regressor and MLP underperformed across all metrics. These findings underscore the strength of tree-based models for shear stress and kernel-based models for viscosity, and highlight the importance of aligning model complexity with data characteristics. Hyperparameter tuning was further applied to enhance the performance of the top models.

When comparing the accuracy of our machine learning models to those of conventional empirical models, the results suggest that our models consistently deliver comparable performance. For example, the Herschel-Bulkley model yielded R^2 values of 0.99 for soy whey beverages [60], and high-protein soy desserts [61], while the Ostwald-de-Waele model showed $R^2 = 0.99$ in sesame protein isolate [1,62]. Similarly, Power-Law model showed $R^2 = 0.84$–0.99 for soy, pean and faba bean protein isolates [63], $R^2 = 0.94$–0.995 for soy for concentration of peanut protein isolate aggregation suspension [64], mixed soy protein isolate (SPI)–flaxseed gum (FG) dispersions [65] and the Carreau model achieved $R^2 = 0.90$–0.92 for pea protein-gluten blends. A direct comparison between empirical and machine learning models was conducted by **Herrmannet al.** [66] who investigated the rheological behavior of 34 commercial food dispersions. They employed both Herschel–Bulkley model and artificial neural networks (ANNs) to predict key parameters—yield stress (π_0), consistency coefficient (K), and flow behavior index

(*n*) using protein composition. Their results showed that ANNs with three hidden layers and two neurons per layer demonstrated high predictive accuracy across all Herschel–Bulkley parameters. On the other hand, empirical models, though effective, are constrained by fixed mathematical structures, limiting their ability to model intricate nonlinearities. In traditional empirical models, it is not possible to include the processing variables as the input. Therefore, in these traditional studies, shear rate is usually treated as an independent parameter and is often left out of the experimental design, where only processing variables are incorporated into the modeling framework. In contrast, machine learning approaches allow for the explicit inclusion of shear rate as an input variable in the experimental design, enabling a more accurate representation of SPI's non-Newtonian behavior, where viscosity depends on shear rate. Therefore, ML models can more effectively capture viscosity variations and reveal complex interactions among shear rate, pressure, and IC. This approach improves the predictive accuracy and realism of rheological models, overcoming the shortcomings of conventional methods that may overlook nonlinear or synergistic effects between processing parameters. This suggests ML approaches may better adapt to complex rheological behaviors in plant proteins, though differences in experimental conditions and protein types across studies warrant cautious interpretation.

3.3.2. Validation of machine learning algorithms

Fig. 6 shows a comparison of actual versus predicted values for shear stress and viscosity, employing the two top-performing machine learning models–Decision Tree and XGBoost for shear stress, and Decision Tree and Gaussian Process Regressor for viscosity (Table 2, Figs. 5 and 6). Each plot demonstrates the predictive accuracy of the models through the alignment of the predicted values (blue dots) with the actual values (represented by the red line). The R^2 scores on each plot quantified the goodness-of-fit, with values closer to 1 indicating higher predictive accuracy. The Decision Tree and XGBoost models demonstrated near-perfect accuracy in predicting shear stress, with R^2 scores of 0.9990 and 0.9992, respectively, indicating excellent model fit. For viscosity prediction, the Gaussian Process Regressor outperformed the Decision Tree, achieving an $R2$ of 0.9925 versus 0.9443, effectively capturing the data's non-linear patterns. These results highlight that while tree-based models are highly effective for shear stress prediction, Gaussian Process Regression is better suited for modeling the complexity of viscosity.

3.3.3. Hyperparameter optimization

Hyperparameter optimization aims to find the best settings to maximize the machine learning model's performance, improving accuracy, avoiding overfitting, and tailoring the model to specific data [67]. In this study, tuning was applied to enhance predictions for shear stress and viscosity. The optimized hyperparameters, shown in Table 3, provide settings that balance complexity, generalization, and computational efficiency. The Decision Tree model used an unrestricted depth and a minimum sample split of 2 for shear stress, allowing maximum complexity while managing overfitting. The XGBoost model, with a learning rate of 0.1, max depth of 10, and 100 estimators, balanced convergence speed with stability and accuracy, with subsampling at 80 % to reduce overfitting. For viscosity, the Decision Tree model was set to a max depth of 15, preventing overfitting while requiring at least five samples to split and one sample per leaf for better generalization on test data. The GPR used a *Matern kernel* and a low noise assumption (alpha = 1e-10), ideal for capturing non-linear patterns in viscosity data, with five optimizer restarts to find the best solution.

Hyperparameter tuning balanced model complexity and generalization, with target-specific adjustments—unlimited depth for shear

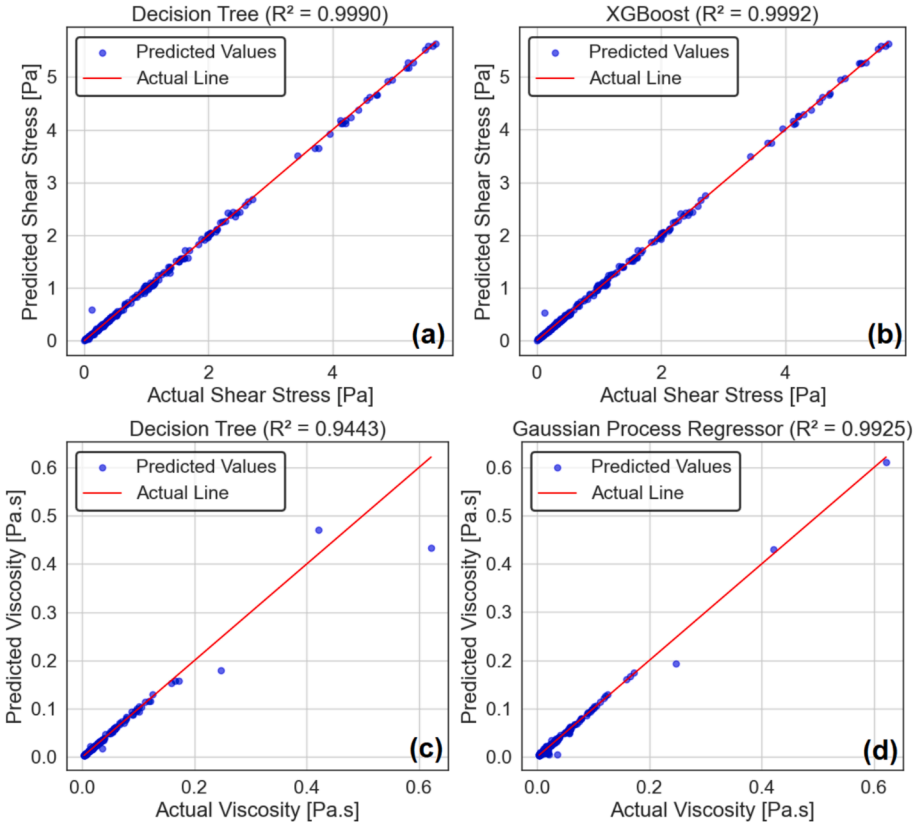


Fig. 6. Comparison of actual vs. predicted values for shear stress and viscosity using different machine learning models. (a) Decision Tree model for shear stress, (b) XGBoost model for shear stress, (c) Decision Tree model for viscosity, (d) Gaussian Process Regressor. The red line represents the ideal fit line where predicted values match actual values perfectly.

Table 3
Hyperparameters and performance metrics of top-performing machine learning models for predicting shear stress and viscosity.

Models	Best hyperparameters	Best result	Performance metrics									
			R2		MSE		RMSE		MAE		MAPE	
			Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Shear stress												
Decision Tree	{'max_depth': None, 'min_samples_split': 2}	0.9991	1	0.999	0	0.0019	0	0.0437	0	0.0237	0	6.8842
XGBoost	{'learning_rate': 0.1, 'max_depth': 10, 'n_estimators': 100, 'subsample': 0.8}	0.9993	1	0.9992	0	0.0015	0.0044	0.0384	0.0029	0.0216	1.4184	6.6242
Viscosity												
Decision Tree	{'max_depth': 15, 'min_samples_leaf': 1, 'min_samples_split': 5}	0.929	0.9897	0.8853	0	0.0004	0.0059	0.0196	0.0013	0.0031	1.7258	5.3225
Gaussian Process Regressor	{'alpha': 1e-10, 'kernel': 1**2 * Matern(length_scale = 1, nu = 1.5), 'n_restarts_optimizer': 5}	0.9842	1	0.9927	0	0	0	0.0049	0	0.0008	0	2.3529

stress and a max depth of 15 for viscosity in Decision Trees. XGBoost tuning significantly improved shear stress prediction, while the Gaussian Process Regressor, using a Matern kernel, proved highly effective for viscosity. Tuning notably enhanced performance metrics, particularly MAPE, for models like XGBoost and GPR. In contrast, Decision Tree models showed limited improvement for viscosity, indicating diminishing returns. Overall, optimization boosted predictive accuracy for advanced models, reinforcing their suitability for estimating complex material properties.

3.3.4. *Comparative analysis of explainability of ML models*

Fig. 7 presents a comprehensive comparative analysis of model interpretability for shear stress **(**Fig. 7**a,b,c,d,e,f)** and viscosity **(**Fig. 7**g, h,i,j,k,l)** predictions, utilizing DTR and GPR models through a variety of feature attribution and interaction methods. Applying SHAP, rooted in cooperative game theory, to the models provides a robust framework for interpreting ML predictions of shear stress and viscosity in protein-based systems. This method enables a detailed examination of how variables like pressure, IC, and shear rate influence model outputs, validates alignment with established rheological principles, and enhances transparency, thereby building trust in the model's predictive capabilities. In analyzing Fig. 7**(a) and (g)**, which illustrate the influence of input variables on shear stress and viscosity predictions in the SPI systems, pressure emerges as the most significant predictor for both targets, as indicated by the lengths of the bars in the feature attribution analyses. The SHAP analysis suggests that pressure significantly influenced shear stress and viscosity of SPI. Accordingly, high-pressure conditions can induce protein conformational changes, leading to denaturation, aggregation, or gelation, which in turn affect the rheological properties such as shear stress. A similar pattern was reported by [68], who examined the impact of dynamic high-pressure treatment (at 18, 100, and 150 MPa) on the shear stress of whey protein concentrate (WPC) and sodium caseinate (SC). Additionally, some studies have shown that pressure levels above 300 MPa can cause irreversible protein denaturation and gel formation, thereby altering rheological behavior [69,70]. SHAP analysis also revealed that IC played a significant role in the rheological behavior of SPI. The addition of ions can influence protein–protein interactions, leading to changes in viscosity and gelation properties. For example, the incorporation of Ca^{2+} ions was shown to modify the rheological behaviors of proteins, with increased protein levels and decreased pH enhancing the solutions' elastic behavior. Additionally, ion-specific effects can alter protein solubility and stability, impacting the overall rheology of the system [71]. Shear Rate presents the lowest SHAP value among the input variables, suggesting a lesser direct impact on shear stress and viscosity predictions within the examined protein-based systems.

The partial dependence plots (PDPs) derived from the DTR and GPR models provide complementary insights into how the key processing parameters affect the rheological properties of SPI, specifically shear stress and viscosity. For the DTR model **(**Fig. 7**b)**, where the target variable is shear stress, pressure emerges as the most influential factor, exhibiting a strong and nearly linear positive relationship with shear stress. This indicates that increasing pressure leads to a proportional rise in shear stress, likely due to enhanced molecular interactions or structural compaction under elevated pressure. The IC variable demonstrated a nonlinear pattern: shear stress increased with IC up to about 150 mM, after which it plateaus, suggesting a saturation effect. This could be attributed to ion-mediated protein aggregation reaching a maximum binding or charge-shielding capacity [72]. Meanwhile, shear rate shows a sublinear increase in shear stress, indicating diminishing returns at higher shear rates. These non-linear yet interpretable patterns are typical for tree-based models, which excel at detecting thresholds and complex interactions without requiring explicit functional forms. However, the GPR model **(**Fig. 7**h)**, trained to predict viscosity of SPI, reflects a more nuanced and smoothed response to the same set of input variables. Here, pressure also shows a positive relationship with

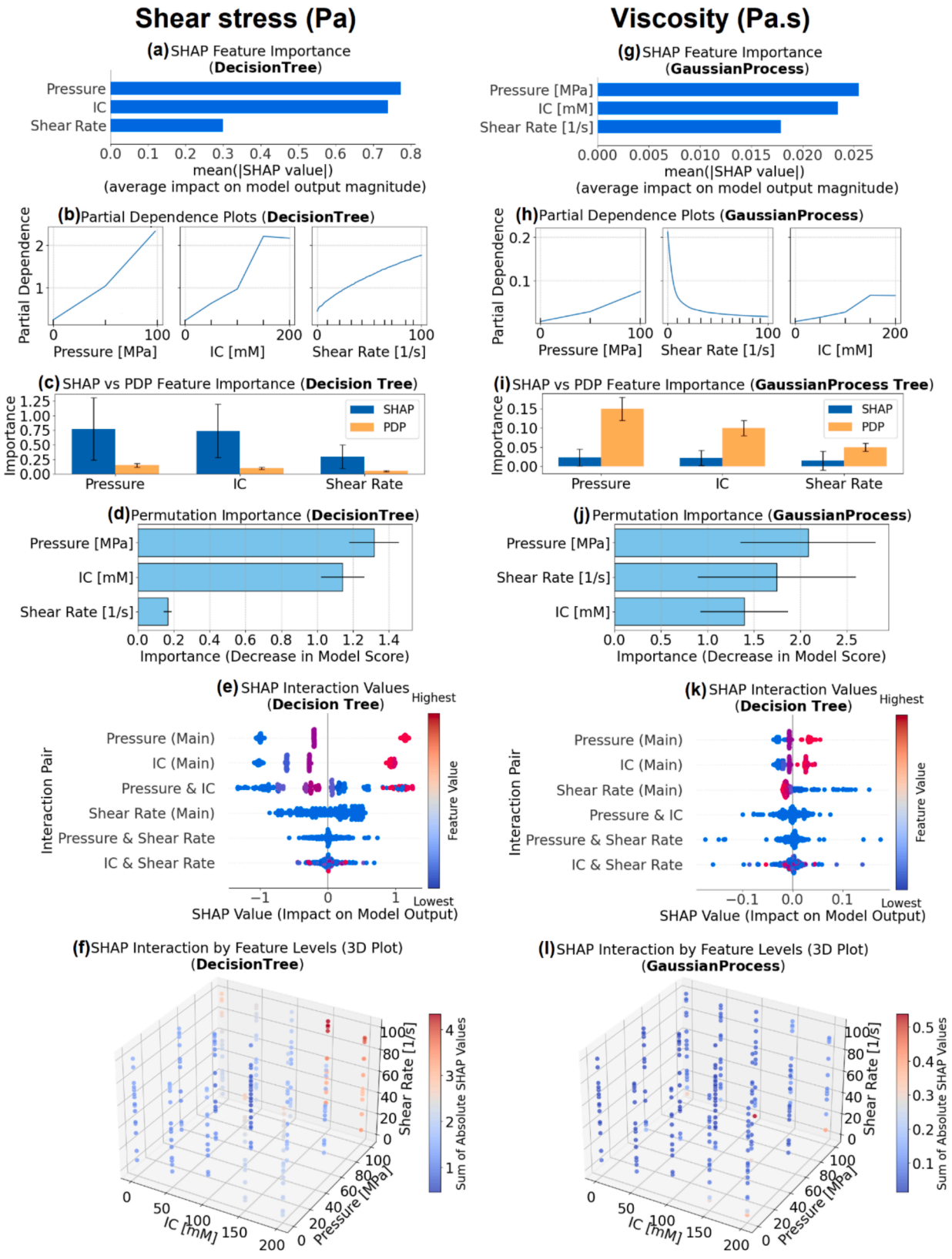


Fig. 7. Comparative model interpretation for shear stress and viscosity predictions using Decision Tree and Gaussian Process regressors. **(a, g)** SHAP feature importance plots displaying the average contribution of each feature to predictions. **(b, h)** Partial Dependence Plots (PDP) showing the marginal effects of each input feature on the model output. **(c, i)** Comparison of SHAP and PDP-based feature importances, highlighting differences in feature rankings. **(d, j)** Permutation importance analysis indicating the decrease in model score when each feature is shuffled. **(e, k)** SHAP beeswarm-interaction plots showing both main effects and pairwise interactions between input features. **(f,l)** 3D SHAP interaction plots illustrating how combinations of feature values influence the magnitude of SHAP values, reflecting the model's local sensitivity to input variation.

viscosity but with a far smaller magnitude, suggesting a more gradual influence. Low pressure and low IC (0 mM) were found to generally promote an increase in viscosity [73,74], possibly due to the ions that weaken the electrostatic affinity between protein molecules [73]. Similarly, **Baskıncı and Gul** [1] found that the viscosity of sesame protein isolate increased with pressure, likely due to HPH-induced reductions in particle size and improved protein solubility. The shear rate, notably, led to a sharp initial decrease in viscosity followed by a flattening trend, which is consistent with shear-thinning behavior commonly observed in protein solutions and dispersions. This is in contrast to the increasing trend seen in the DTR model's shear stress output. The IC variable again shows an increasing trend that levels off around 150 mM, reinforcing the idea of saturation kinetics, albeit with a subtler effect on viscosity. These smoother transitions are a hallmark of GPR models, which apply kernel-based regularization and often produce conservative estimates, especially when data are sparse or noisy [51]. The alignment of the IC response in both models—initial increase followed by a plateau—suggests a consistent physicochemical behavior likely rooted in protein-ion interactions [75]. While increased shear rate contributes positively to shear stress (as expected in shear-thickening or pseudo-plastic systems), it reduces viscosity due to alignment or breakdown of protein structures during flow. The divergence in feature impact and ranking between models, such as the interpretable DTR and the probabilistic GPR, highlights not only their contrasting modeling capabilities but also the differing mechanical properties—shear stress vs. viscosity—that respond uniquely to the same process conditions in complex, nonlinear systems.

To further strengthen these model-driven insights, integrating partial dependence plots with variance-based sensitivity indices (VBSIs) analysis allows for a more comprehensive understanding of how each processing parameter influences the rheological behavior of SPI. The combined analysis of PDPs and VBSIs provides a comprehensive view of how processing parameters influence the rheological behavior of SPI in terms of shear stress and viscosity. For the DTR model predicting shear stress, the sensitivity indices quantify the relative importance of input variables as follows: **pressure = 1.0822**, **IC = 0.9018**, and **shear rate = 0.3636**. These values strongly support the PDP trends. Pressure emerges as the most influential factor, aligning with the PDP which shows a steep and nearly linear increase in shear stress with rising pressure. Similarly, IC received the second-highest sensitivity score. In contrast, shear rate, though still positively correlated with shear stress in the PDP, had a significantly lower sensitivity index, indicating that its effect, while present, was relatively limited compared to pressure and IC. This affirms the notion that while increasing shear rate mechanically contributes to shear stress, its influence is moderated by the physicochemical effects governed by pressure and ionic strength. In the case of viscosity modeled by the GPR, the sensitivity indices are markedly smaller for all inputs: **pressure = 0.0318**, **IC = 0.0252**, and **shear rate = 0.0234**. These low values reflect the GPR's smoother and more conservative estimation behavior, which is also evident in the PDPs. The plots show only slight changes in viscosity across the range of input variables, with the most notable trend being a mild decrease in viscosity with increasing shear rate—indicative of shear-thinning behavior, a common property of protein dispersions. Pressure and IC both show gentle increases followed by plateaus, which is in line with their low sensitivity indices, suggesting minimal but consistent influence on viscosity. The striking divergence between the two models is not just in model architecture but also in the magnitude of sensitivity. The DTR model detects strong, nonlinear dependencies in the case of shear stress, reflecting the mechanical response of the system to processing conditions. In contrast, the GPR model's low sensitivity indices indicate that viscosity, as a more intrinsically averaged property, is less reactive to individual parameters within the explored range. This suggests that while shear stress is highly tunable by process conditions, viscosity may be governed by more subtle or synergistic mechanisms not easily captured by individual features. These findings reinforce the importance of model-specific interpretation and highlight how different rheological outputs respond to process conditions in distinct ways.

Fig. 7**(c) and (i)** shows the comparison of SHAP and PDP-derived feature importances, which further elucidates the relative influence of inputs on the rheological behavior of SPI. For the DTR model predicting shear stress, both SHAP and PDP analyses identify pressure and IC as the dominant features, with pressure slightly leading in both methods. SHAP values for both pressure and IC exceed 0.75, confirming their substantial contributions, while shear rate shows a noticeably lower SHAP value ~ 0.3 (Fig. 7**c**). This ranking and magnitude align closely with the VBSIs, where pressure (**1.08**) and IC (**0.90**) far surpass shear rate (**0.36**). However, a notable discrepancy arises in the PDP-derived importance values, which are significantly lower for all features—particularly for shear rate. This gap may result from PDPs' inability to fully account for interactions or non-additive effects, which SHAP values are designed to capture [40]. The large error bars in SHAP for pressure and IC also suggest heterogeneity in feature influence across the data space, potentially due to thresholds or non-linear effects typical in protein structuring under processing stress. In contrast, the GPR model predicting viscosity presents a more nuanced picture (Fig. 7**i**). PDP importance values are clearly higher than SHAP values across all features, most notably for pressure and IC, where PDP indicates importance levels of ~ 0.15 and ~ 0.10 respectively, while SHAP values remain below 0.03. This divergence underscores the over-smoothing tendency of GPR models, which may downplay feature impacts in local regions while still reflecting global trends in PDPs. Moreover, the VBSIs for viscosity were uniformly low (**pressure: 0.0318; IC: 0.0252; shear rate: 0.0234**), which aligns well with the low SHAP values, reinforcing the idea that no single input strongly governs viscosity within the given range. The discrepancy between PDP and SHAP in this case may be due to mild global curvature in the function captured by the PDP, but insufficient localized variance to register strong SHAP effects. Taken together, these results highlight the complementarity of **SHAP**, **PDP**, and **variance-based sensitivity analysis**. For highly nonlinear systems like shear stress response, SHAP and variance-based sensitivity provide consistent, interpretable insights into dominant features—clearly prioritizing Pressure and IC. Meanwhile, for smoother responses like viscosity, SHAP offers a more conservative picture, revealing limited feature influence, even when PDPs suggest modest directional trends. The disparities between PDP and SHAP in both models point to the importance of considering both global (PDP) and local (SHAP) explanations to avoid over- or underestimating feature relevance, particularly in models with complex interactions or non-stationary effects. Together, they help detect and mitigate bias, enabling both dataset-level and sample-level interpretation while revealing feature interactions.

The permutation importance results, shown in Fig. 7**(d) and (j)** offer additional insight into the relative contributions of input features by quantifying the decrease in model performance when each variable is randomly permuted. In the DTR model Fig. 7**(d)**, the permutation importance values show that pressure and IC have similarly high importance scores (~1.3 and ~ 1.15, respectively), with shear rate contributing far less (~0.15). This reinforces the earlier findings from PDPs and SHAP (Fig. 7**c**), where pressure and IC had the largest partial dependence and SHAP magnitudes, and from the VBSIs (1.08 for Pressure, 0.90 for IC). It is clear across all methods that pressure and IC were the dominant drivers of shear stress in SPI, while shear rate had only a marginal effect. The permutation results further support this conclusion by directly linking variable perturbation to a drop in model performance, confirming that the model's predictive power relies primarily on pressure and IC. In contrast, for the GPM model Fig. 7**(j)**, the permutation importance results differ markedly from the earlier SHAP and sensitivity index findings. Here, all three features show relatively high and similar importance values, ranging from approximately 1.3 to 2.1, with pressure slightly leading. This contradicts the low SHAP values (all below 0.03 in Fig. 7**i**) and VBSIs (all ~ 0.02–0.03), which had previously suggested only minimal influence of any individual feature on viscosity.

The PDPs for the GPR model (Fig. 7**h**), while exhibiting only minor changes in predicted output, do reveal weak but consistent directional trends—such as a mild increase with Pressure and IC, and a slight decrease with shear rate. Although the absolute effect sizes are small, these patterns indicate that the model captures some level of sensitivity to the inputs, particularly in the form of smooth, global trends. Therefore, the permutation importance results may be reflecting the model's reliance on distributed and possibly interaction-driven effects that are not readily apparent in SHAP or variance-based scores. However, given the small scale of change in the PDP curves, the influence of individual features on viscosity should still be interpreted as modest at best. On the other hand, this divergence can be understood in the context of how each method defines "importance". While SHAP and VBSI analyses decompose feature contributions additively or through variance partitioning, permutation importance captures global disruption to predictive structure, including feature interactions and nonlinear compensations. In models like GPR, where smooth, interaction-rich approximations are common, even variables with weak marginal effects can have large interdependent roles. That is, each variable may not be important in isolation, but the model still depends on their combined structure to produce accurate predictions. This is consistent with the broad confidence intervals seen in the GPR permutation bars (Fig. 7**j**), indicating uncertainty in variable-specific disruption due to interactions.

The SHAP interaction plots (Fig. 7**e**,**k**) provide further insight into the role of both main effects and feature interactions in the DTR model's prediction of shear stress and viscosity of SPI. For shear stress, the interaction plot (Fig. 7**e**) reinforces earlier findings that pressure and IC are the most influential features. Their main effects exhibit wide SHAP value ranges, with distinct bands of positive impact for high feature values (red). Importantly, the interaction between pressure & IC shows notable contribution, indicating that the joint effect of these two variables is not purely additive, but synergistic. This aligns with the previously observed high VBSIs, strong permutation importance, and high SHAP values for both features. Meanwhile, shear rate appears to have minimal interaction-based contribution, consistent with its lower importance across all previous interpretability methods. In contrast, for viscosity (Fig. 7**k**), SHAP interaction values are small in magnitude across all features and pairs, with most points clustered near zero. Although subtle main effects of pressure and IC are visible, the interactions—especially those involving pressure & shear rate and IC & shear rate—were weak. These observations are consistent with earlier SHAP and VBSIs results, which indicated very low individual contributions of features to viscosity. However, the presence of mild structured patterns supports earlier suggestions from PDPs and permutation importance that the model's predictive accuracy may still depend on weak, distributed interactions among variables. In summary, SHAP interaction analysis confirms that the shear stress model relies heavily on pressure and IC, both independently and interactively, while the viscosity model reflects weak, non-dominant interactions across all features, reinforcing its more diffuse dependency structure.

Fig. 7**(f) and (l)** illustrate the 3D SHAP interaction plots which provide a spatial visualization of how combinations of pressure, IC, and shear rate influence model predictions through cumulative feature interactions. For the DTR model (Fig. 7**f**), the sum of absolute SHAP values is highest when all three input variables are at elevated levels. This region—characterized by high pressure (~100 MPa), moderate to high IC (~100–200 mM), and high Shear rate (~100 1/s)—shows intense interaction activity (red-colored points), confirming that shear stress is most sensitive to joint effects under extreme processing conditions. This spatial pattern reinforces earlier findings from PDPs, SHAP values, and permutation importance, all of which highlighted strong main effects from pressure and IC, and interactions such as pressure × IC. It also aligns with the SHAP interaction plots **(**Fig. 7**e)**, which showed clear non-additive contributions between variables, especially under high feature values. On the other hand, the GPR model (Fig. 7**l**) shows much weaker SHAP interactions across the input space. The SHAP values remain low throughout, with only isolated, mild hotspots—typically at low shear rate and moderate IC levels—where weak interactions occur. This aligns with prior SHAP and variance-based results, which suggested that no individual variable exerts strong marginal influence. While permutation importance implied the presence of interaction effects, this 3D plot reveals that such interactions are spatially diffuse and not concentrated in any dominant region of the input space. Together, these 3D SHAP visualizations validate that the predictive power of the shear stress model is highly dependent on strong interactions among key features, particularly under intensive processing conditions. In contrast, the viscosity model exhibits more distributed and subtle interactions, supporting the earlier conclusion that it reflects a more diffuse dependency structure. All these findings highlight the importance of using multiple interpretability techniques to obtain a well-rounded understanding of feature relevance, especially in models that may rely heavily on nonlinear interactions or distributed information across variables.

While global interpretability methods such as PDPs, SHAP summary plots, VBSI analysis, and permutation importance provide valuable insights into average model behavior, they often obscure critical local variations and context-specific interactions. To address this, we conducted SHAP beeswarm subplot analysis to offer a condition-specific and localized perspective on feature impact.

Fig. 8 presents SHAP beeswarm-by-condition plots that illustrate how input features affect model predictions for shear stress and viscosity under various fixed operating conditions. Each subplot corresponds to a specific pressure–IC pair, where the levels of pressure and IC were held constant (gray dots), and only shear rate varies (represented by a red–blue color gradient from high to low). Importantly, the vertical order of features in each subplot reflects their relative impact (ranked top to bottom by decreasing SHAP importance) on the model output under that specific condition, while the horizontal spread and color of dots capture both magnitude and direction of contribution of shear rate. This design isolates the effect of shear rate while assessing the contribution of the fixed values of pressure and IC to the model output. In the shear stress model (Fig. 8**a**), pressure and IC generally appear in the top-ranked positions, especially under higher values (e.g., P = 100 MPa, IC ≥ 100 mM), confirming earlier findings from PDPs, SHAP values, and sensitivity indices. However, the role of shear rate is condition-dependent: In intermediate regimes (e.g., P = 50 MPa, IC = 50–200 mM), it often ranks among the top two features. Notably, in these cases, higher shear rate values (red) tend to align with positive SHAP values, while lower values (blue) align with negative or near-zero SHAP values—indicating that increasing shear rate contributes to increased shear stress, as demonstrated by **O′ Flynnet al.** [76] in their study on soy protein isolate dispersions. They observed that at pH 6.9, the shear stress of unheated soy protein isolate dispersions rose from 0.8 to 18.5 Pa over a shear rate range of 9.6 to 300 s^{-1}. This effect diminishes generally under extreme conditions where pressure and IC dominate, suggesting the DTR model learns an interaction-sensitive structure where shear rate matters most under moderate regimes.

In the viscosity model (Fig. 8**a**), shear rate becomes the dominant feature, particularly under high Pressure and IC conditions (e.g., P = 100 MPa, IC = 150–200 mM). In these cases, shear rate is consistently ranked at the top, and higher SR values (red) correspond to negative SHAP values, confirming the model's learning of shear-thinning behavior—i.e., viscosity decreases as shear rate increases. A similar observation was made by **Chonghaoet al.** [77], who reported that the viscosity of soy protein isolate decreased with increasing shear rate. This trend holds across multiple subplots and is in agreement with earlier PDPs (Fig. 7**h**) and interaction plots (Fig. 7**k**). Together, the SHAP beeswarm-by-condition plots provide valuable, context-aware insights into model behavior by combining feature ranking, directionality, and value-driven attribution. They reveal how the influence of input features—pressure, IC, and shear rate—varies across processing conditions, showing that increases in Shear Rate can lead to higher shear stress in

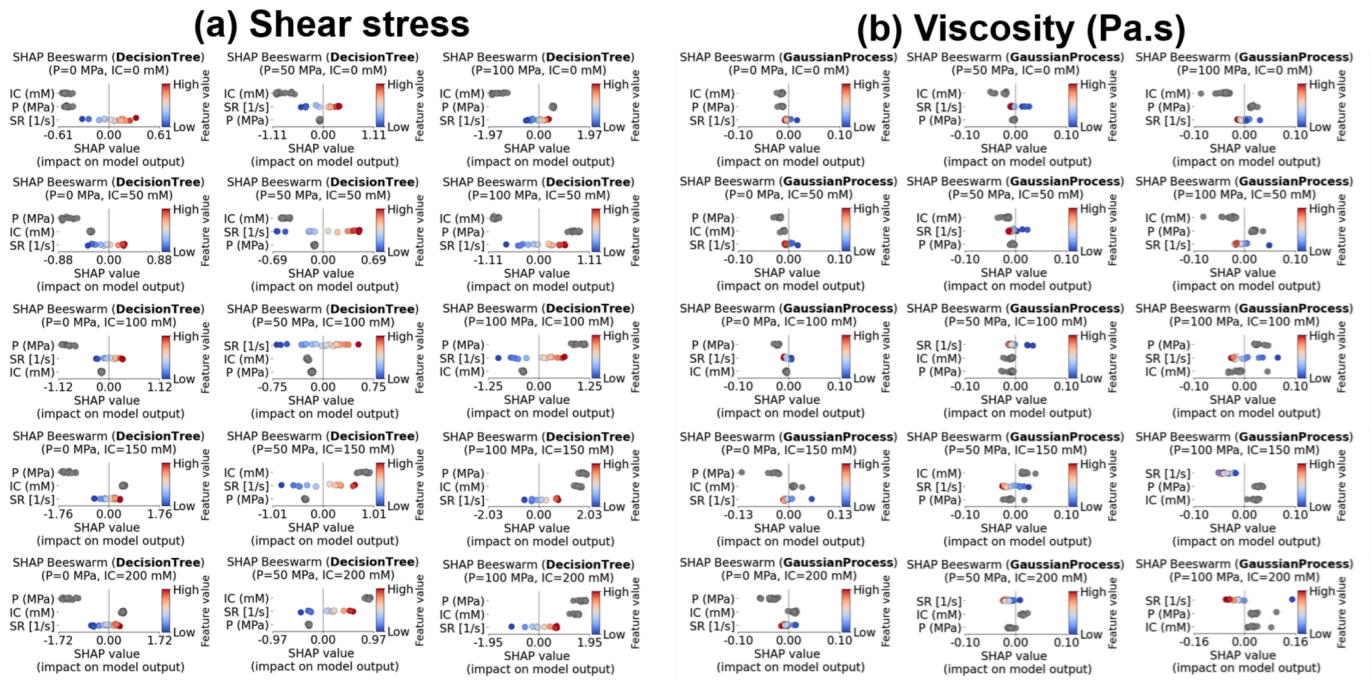


Fig. 8. SHAP beeswarm-by-condition plots illustrating impact of input features on model predictions for **(a)** shear stress (Pa) and **(b)** viscosity (Pa·s). Each subplot corresponds to a specific condition defined by fixed pressure (P: 0–100 MPa) and ion concentration (IC: 0–200 mM). Within each condition, the plots highlight how variations in the levels of shear rate—ranging from low (blue) to high (red)—affect the model output.

certain regimes and reduced viscosity in others. These nuanced, condition-specific attributions deepen the understanding gained from global interpretation methods, capturing interaction-driven effects in complex rheological modeling and highlighting the essential role of localized SHAP analysis in interpreting nonlinear, feature-interactive systems.

Beyond condition-specific insights, it is also important to evaluate whether the learned patterns and feature attributions uncovered through explainable AI techniques are system-dependent or have potential for generalization across other plant proteins and experimental conditions. The SHAP and permutation importance analyses used in this study provide unbiased, model-agnostic interpretations of the factors driving rheological behavior in sesame protein isolate (SPI). These techniques reveal consistent, domain-aligned effects of pressure, IC, and shear rate effects that are grounded in physical mechanisms such as protein unfolding, ionic shielding, and shear-induced restructuring. Because these mechanisms are conserved across many plant proteins (e.g., soy, pea), and because SHAP and permutation importance analyses are robust to model type and feature correlations, the learned patterns in feature relevance are likely transferable. Although retraining will be required to capture system-specific responses, we hypothesize that the model's interpretability structure provides a stable foundation for generalization. Future studies will explore this by applying the same pipeline to soy and pea proteins, and by incorporating temperature-dependent data to expand applicability under variable thermal processing conditions.

3.3.5. Real-world impact of explainable rheology model results

The interpretability provided by SHAP beeswarm plots significantly enhances the practical application of rheology models across various domains by offering detailed, context-specific insights into how input variables influence model predictions. This level of transparency is particularly valuable in scientific and engineering fields where precise control over operating conditions is crucial.

In food product development, achieving desired texture and stability depends heavily on controlling rheological properties like viscosity and shear stress. The interpretability analysis (e.g., SHAP, PDPs) reveals that pressure and ionic strength critically influenced shear stress, while shear rate dominated viscosity. These findings help formulators pinpoint which variables to adjust for target textures [78]. For example, if SHAP values show that low IC increases shear stress, developers can manipulate ion levels to control gel strength or mouthfeel. Similarly, understanding the shear-thinning behavior through PDPs guides the use of thickeners that flow under agitation but hold shape at rest, such as xanthan gum. These insights reduce trial-and-error by focusing formulation adjustments on the most influential parameters.

While formulation choices are central to defining rheological properties, their behavior can vary significantly during processing, highlighting the need to optimize operational parameters like mixing speed and applied pressure [79]. Shear rate and pressure emerged as dominant features influencing viscosity and shear stress, respectively. This directly supports optimization of mixing and pumping operations. For instance, PDPs showing strong shear-thinning behavior suggest ideal mixer speeds that achieve uniformity without overmixing. Likewise, SHAP values indicating viscosity increases under high pressure help engineers would allow avoiding excessive pump stress or energy loss in pipelines. When IC is important, operators can maintain consistent flow by controlling additive dosing or solvent purity. These model insights enable proactive adjustments to operating conditions, leading to improved process stability, energy efficiency, and reduced equipment wear–consistent with real-world applications in fluid handling systems and drilling fluids.

Beyond optimizing existing processes, these insights are equally valuable in the early-stage design of novel materials, where rheological performance must be tailored to meet specific functional or structural requirements. When designing materials such as hydrogels, inks, or biopharmaceuticals, understanding how formulation parameters affect rheology is essential [43]. Model interpretability identifies key variable–like pressure or IC–that modulate flow resistance or mechanical strength. For example, if IC is found to reduce viscosity, researchers can tune IC levels to design injectable formulations or print-friendly materials. PDPs showing nonlinear viscosity trends guide the selection of optimal polymer or protein concentrations, avoiding excess use of costly

ingredients. SHAP-based feature importance also supports rational material design [80]. This allows developers to fine-tune properties with fewer experiments and better mechanistic understanding.

Once materials and processes are in place, maintaining consistency and reliability becomes essential, making model interpretability a powerful tool for quality assurance and predictive maintenance in industrial settings [37,81]. In production environments, rheology serves as a sensitive indicator of product quality and equipment condition. The models' interpretability allows real-time identification of factors driving rheological deviations. For instance, a spike in viscosity due to increased IC might indicate contamination, prompting immediate intervention. SHAP values also help isolate root causes, such as pump inefficiencies linked to changes in shear rate. By identifying which variables most affect rheological outcomes, QC teams can establish tighter process controls and maintenance thresholds. Additionally, models can flag harmful interactions–like simultaneous high pressure and shear rate––that may predict future equipment failures.

4. Conclusions

The comprehensive application of interpretability techniques—including SHAP analysis, PDPs, VBSIs, and permutation importance—has provided a multifaceted understanding of the factors influencing shear stress and viscosity in sesame protein isolate (SPI) systems. Consistently, pressure and IC emerged as primary determinants of shear stress, with SHAP values exceeding 0.75 and VBSIs of 1.08 and 0.90, respectively. This underscores the significant impact of these variables on the rheological behavior of SPI. While shear rate exhibited a lesser effect, its influence was more pronounced under specific conditions, as revealed by SHAP beeswarm analyses. These findings align with established rheological principles, where pressure-induced conformational changes and ion-mediated protein interactions critically affect shear stress. In contrast, the viscosity model demonstrated a more subdued response to individual input variables. Both SHAP values and VBSIs were notably low across all features, indicating a diffuse dependency structure. However, permutation importance analyses suggested that viscosity predictions rely on subtle, distributed interactions among pressure, IC, and shear rate. This highlights the necessity of employing multiple interpretability methods to capture the nuanced interplay of factors affecting viscosity. The integration of these techniques not only validates the models' predictive capabilities but also enhances transparency, fostering trust in their application to protein-based systems.

Leveraging Decision Tree Regressor and GPR models, coupled with interpretability analyses, enables accurate prediction of SPI's rheological behavior under specific processing conditions. This predictive capability allows food scientists to optimize formulations to achieve desired texture, stability, and processability, thereby reducing reliance on extensive trial-and-error experimentation. Such an approach enhances product quality and increases consumer acceptance by ensuring SPI-based foods meet targeted textural attributes.

CRediT authorship contribution statement

Mustafa Tahsin Yilmaz: Conceptualization, Methodology, Software, Writing-original draft, Editing. **Salman Badurayq:** Investigation, Formal analysis. **Kemal Polat:** Investigation, Review, Editing. **Ahmad H. Milyani:** Visualization, Review. **Abdulaziz S. Alkabaa:** Investigation, Review. **Osman Gul:** Data Acquisition. **Furkan Turker Saricaoglu:** Data Acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The project was funded by KAU Endowment (WAQF) at king Abdulaziz University, Jeddah, Saudi Arabia. The authors, therefore, acknowledge with thanks WAQF and the Deanship of Scientific Research (DSR) for technical and financial support.

References

[1] Baskıncı T, Gul O. Modifications to structural, techno-functional and rheological properties of sesame protein isolate by high pressure homogenization. Int. J. Biol. Macromol. 2023;250:126005.
[2] Yüzer M, Gençcelep H. Sesame seed protein: Amino acid, functional, and physicochemical profiles. Foods and Raw Materials 2023;11(1):72–83.
[3] Gómez-Arellano A, et al. Rheological behaviour of sesame (Sesamum indicum L.) protein dispersions. Food Bioprod. Process. 2017;106:201–8.
[4] Luo L, et al. Impact of high-pressure homogenization on physico-chemical, structural, and rheological properties of quinoa protein isolates. Food Struct. 2022; 32:100265.
[5] Yu JWang L, Zhang Z. Plant-based meat proteins: processing, nutrition composition, and future prospects. Foods 2023; 12(22): 4180.
[6] Melchior S, et al. High pressure homogenization shapes the techno-functionalities and digestibility of pea proteins. Food Bioprod. Process. 2022;131:77–85.
[7] Saricaoglu FT. Application of high-pressure homogenization (HPH) to modify functional, structural and rheological properties of lentil (Lens culinaris) proteins. Int. J. Biol. Macromol. 2020;144:760–9.
[8] Saricaoglu FT, et al. Effect of high pressure homogenization (HPH) on functional and rheological properties of hazelnut meal proteins obtained from hazelnut oil industry by-products. J. Food Eng. 2018;233:98–108.
[9] Zhang A, et al. Effects of high pressure homogenization on the structural and emulsifying properties of a vegetable protein: Cyperus esculentus L. Lwt 2022;153: 112542.
[10] Idowu AO, et al. Functional properties of sesame (Sesamum indicum Linn) seed protein fractions. Food Prod. Process. Nutr. 2021;3:1–16.
[11] López G, et al. Development of a liquid nutritional supplement using a Sesamum indicum L. protein isolate. LWT-Food. Sci. Technol. 2003;36(1):67–74.
[12] Rafe A, et al. Dynamic rheological properties of sesame protein dispersions. Legume Sci. 2023;5(2):e177.
[13] Tiziani S, Vodovotz Y. Rheological effects of soy protein addition to tomato juice. Food Hydrocoll. 2005;19(1):45–52.
[14] Eze FN, et al. Upcycling of Defatted Sesame seed Meal via Protein Amyloid-based Nanostructures: Preparation, Characterization, and Functional and Antioxidant Attributes. Foods 2024;13(14):2281.
[15] Sheikh F, et al. Enhancing rheological and textural properties of gelatin-based composite gels through incorporation of sesame seed oleosome-protein fillers. Gels 2023;9(10):774.
[16] Guiné R. The use of artificial neural networks (ANN) in food process engineering. Int. J. Food Eng. 2019;5(1):15–21.
[17] Schmidt J, et al. Recent Advances and Applications of Machine Learning in Solid-State Materials Science Npj Computational Materials 2019;5(1):83.
[18] Nnyigide OS, Hyun K. A comprehensive review of food rheology: Analysis of experimental, computational, and machine learning techniques. Korea-Australia Rheology Journal 2023;35(4):279–306.
[19] Ore Areche F, et al. Formulation, characterization, and determination of the rheological profile of loquat compote Mespilus Germánica L. through sustenance artificial intelligence. J. Food Qual. 2023;2023(1):3344539.
[20] Jeong SKim H, Lee S. Rheology-based classification of foods for the elderly by machine learning analysis. Applied Sciences 2021; 11(5): 2262.
[21] Torkashvand AMAhmadi A, Nikravesh NL. Prediction of kiwifruit firmness using fruit mineral nutrient concentration by artificial neural network (ANN) and multiple linear regressions (MLR). Journal of integrative agriculture 2017; 16(7): 1634-1644.
[22] Saeidirad MHRohani A, Zarifneshat S. Predictions of viscoelastic behavior of pomegranate using artificial neural network and Maxwell model. Computers and Electronics in Agriculture 2013; 98: 1-7.
[23] Al-Mahasneh MRababah T, Ma'Abreh A. Evaluating the Combined Effect of Temperature, Shear Rate and Water Content on Wild-Flower Honey Viscosity Using Adaptive Neural Fuzzy Inference System and Artificial Neural Networks. Journal of Food Process Engineering 2013; 36(4): 510-520.
[24] Toker OS, Dogan M. Effect of temperature and starch concentration on the creep/recovery behaviour of the grape molasses: modelling with ANN, ANFIS and response surface methodology. Eur. Food Res. Technol. 2013;236:1049–61.
[25] Grinsztajn LOyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? Advances in neural information processing systems 2022; 35: 507-520.
[26] Gross K. *Tree-Based Models: How They Work. Accessed on 28.03.2025, https://blog.dataiku.com/tree-based-models-how-they-work-in-plain-english*. 2020.
[27] GeeksforGeeks. *Gaussian Process Regression (GPR). Accessed on March, 29,2025. https://chatgpt.com/c/67e6be13-27b8-8012-8f22-9f4b27d26018*. 2025.
[28] Howell EC, Hanson J. Development of a non-parametric Gaussian process model in the three-dimensional equilibrium reconstruction code V3FIT. J. Plasma Phys. 2020;86(1):905860102.

[29] McCaffrey JD. *Showdown: Gaussian Process Regression vs Neural Network Regression. Accessed on March 29, 2025. https://jamesmccaffrey.wordpress.com/2023/06/27/showdown-gaussian-process-regression-vs-neural-network-regression/*. 2025.
[30] Tsymbalov E, et al. Deeper connections between neural networks and Gaussian processes speed-up active learning. arXiv preprint arXiv:1902.10350 2019;.
[31] Bishop CM, Tipping M. Variational relevance vector machines. arXiv preprint arXiv:1301.3838 2013;.
[32] Mamat RCRamli A, Bawamohiddin AB. A Comparative Analysis of Gaussian Process Regression and Support Vector Machines in Predicting Carbon Emissions from Building Construction Activities. 2025; 131(1): 186-196.
[33] Chen X, et al. Gaussian process regression for prediction and confidence analysis of fruit traits by near-infrared spectroscopy. Food Qual. Saf. 2023;7:fyac068.
[34] Żelazny WRKusnierek K, Geipel J. Gaussian Process Modeling of In-Season Physiological Parameters of Spring Wheat Based on Airborne Imagery from Two Hyperspectral Cameras and Apparent Soil Electrical Conductivity. Remote Sensing 2022; 14(23): 5977.
[35] van Hoof J, Vanschoren J. Hyperboost: Hyperparameter optimization by gradient boosting surrogate models. arXiv preprint arXiv:2101.02289 2021;.
[36] White GMSiegel AP, Tovar A. Optimizing Thermoplastic Starch Film with Heteroscedastic Gaussian Processes in Bayesian Experimental Design Framework. Materials 2024; 17(21): 5345.
[37] Ucar AKarakose M, Kırımça N. Artificial intelligence for predictive maintenance applications: key components, trustworthiness, and future trends. Applied Sciences 2024; 14(2): 898.
[38] Ince V, et al. Machine learning-based prediction of Clostridium growth in pork meat using explainable artificial intelligence. J. Food Sci. Technol. 2025:1–14.
[39] Molnar C, *Interpretable machine learning*. 2020: Lulu. com.
[40] Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions "a Unified Approach to Interpreting Model Predictions" 2017;30:4768–77.
[41] Molnar C, et al. Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. Data Min. Knowl. Disc. 2024;38(5): 2903–41.
[42] Brusa E, et al. Explainable AI for machine fault diagnosis: understanding features' contribution in machine learning models for industrial condition monitoring. Appl. Sci. 2023;13(4):2038.
[43] Nadernezhad A, Groll J. Machine learning reveals a general understanding of printability in formulations based on rheology additives. Adv. Sci. 2022;9(29): 2202638.
[44] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann. Stat. 2001:1189–232.
[45] Molnar C. *Interpretable Machine Learning: a Guide for making Black Box Models Explainable*. (3rd ed.). 2025.
[46] Sundararajan M, Najmi A. *The many Shapley values for model explanation*. in *International conference on machine learning*. 2020. PMLR.
[47] Janzing DMinorics L, Blöbaum P. *Feature relevance quantification in explainable AI: A causal problem*. in *International Conference on artificial intelligence and statistics*. 2020. PMLR.
[48] Slack D, et al. *Fooling Lime and Shap: Adversarial Attacks on Post Hoc Explanation Methods*. and Society; 2020.
[49] Wang J. An intuitive tutorial to Gaussian process regression. Comput. Sci. Eng. 2023;25(4):4–11.
[50] Breiman L, et al., *Classification and regression trees*. 2017: Routledge.
[51] Rasmussen CE, Williams C. Gaussian processes for machine learning the mit press. Cambridge, MA 2006;32:68.
[52] Zhang Y, Cremer PS. Interactions between macromolecules and ions: the Hofmeister series. Curr. Opin. Chem. Biol. 2006;10(6):658–63.
[53] Lundberg SMErion GG, Lee S-I. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888 2018;.
[54] Fisher ARudin C, Dominici F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research 2019; 20(177): 1-81.
[55] Aqajari SAH, et al. pyeda: an open-source python toolkit for pre-processing and feature extraction of electrodermal activity. Procedia Comput. Sci. 2021;184: 99–106.
[56] Singh D, Singh B. Investigating the impact of data normalization on classification performance. Appl. Soft Comput. 2020;97:105524.
[57] Kim JH. Multicollinearity and misleading statistical results. Korean J. Anesthesiol. 2019;72(6):558–69.
[58] Ansari S, et al. Measurement of the flow behavior index of Newtonian and shear-thinning fluids via analysis of the flow velocity characteristics in a mini-channel. SN Appl. Sci. 2020;2(11):1787.
[59] Baranov A. Influence of temperature and pressure on viscoelastic fluid flow in a plane channel. J. Eng. Phys. Thermophys. 2020;93(5):1296–302.
[60] Punoo HARather JA, Muzaffar A. Development of soy whey fortified orange juice beverages: Their physicochemical, rheological, antioxidant, and sensory properties. Exploration of Foods and Foodomics 2023; 1(4): 206-220.
[61] Arancibia CBayarri S, Costell E. Effect of hydrocolloid on rheology and microstructure of high-protein soy desserts. Journal of food science and technology 2015; 52: 6435-6444.
[62] Liu P, et al. Rheological properties of soy protein isolate solution for fibers and films. Food Hydrocoll. 2017;64:149–56.
[63] Badjona A, et al. Gelation and rheological properties of ultrasound-extracted faba bean protein: a comparative study with commercial plant proteins. Food Hydrocoll. 2025;162:110997.
[64] Zhu Y-dLi D, Wang L-j. Dynamic rheological properties of peanut protein isolate and aggregation suspension and acid-induced gel. Powder Technol. 2019;358: 95–102.
[65] Bi C-h, et al. Characterization of non-linear rheological behavior of SPI–FG dispersions using LAOS tests and FT rheology. Carbohydr. Polym. 2013;92(2): 1151–8.
[66] Herrmann J, et al. Development of a rheological prediction model for food suspensions and emulsions. J. Food Eng. 2013;115(4):481–5.
[67] Feurer M, Hutter F. Hyperparameter optimization. Automated machine learning: Methods, systems, challenges 2019: 3-33.
[68] Venir E, et al. Dynamic high pressure–induced gelation in milk protein model systems. J. Dairy Sci. 2010;93(2):483–94.
[69] De Maria SFerrari G, Maresca P. Rheological characterization bovine serum albumin gels induced by high hydrostatic pressure. Food and Nutrition Sciences 2015; 6(9): 770-779.
[70] Apichartsrangkoon A. Effects of high pressure on rheological properties of soy protein gels. Food Chem. 2003;80(1):55–60.
[71] Quintero J, et al. Effect of the concentration, pH, and Ca2+ ions on the rheological properties of concentrate proteins from quinoa, lentil, and black bean. Foods 2022; 11(19):3116.
[72] Thorarinsdottir KA, et al. Effects of different pre-salting methods on protein aggregation during heavy salting of cod fillets. Food Chem. 2011;124(1):7–14.
[73] Hong TIwashita K, Shiraki K. Viscosity control of protein solution by small solutes: a review. Current Protein and Peptide Science 2018; 19(8): 746-758.
[74] Roberts D, et al. The role of electrostatics in protein–protein interactions of a monoclonal antibody. Mol. Pharm. 2014;11(7):2475–89.
[75] Batoulis H, et al. Concentration dependent ion-protein interaction patterns underlying protein oligomerization behaviours. Sci. Rep. 2016;6(1):24131.
[76] O' Flynn TD, et al. Rheological and solubility properties of soy protein isolate. Molecules 2021; 26(10): 3015.
[77] Chonghao B, et al. Effects of salt ions on rheological properties of SPI-GG hybrid system. Int. J. Agric. Biol. Eng. 2017;10(5):234–41.
[78] Bose P. *Why Rheology is Important in Food Technology. Accessed on March 30, 2025. https://www.azom.com/article.aspx?ArticleID=20575*. 2025.
[79] Raj AS, et al. Predicting rheological properties of wheat dough from flour properties using NIR coupled with artificial neural network. Journal of the ASABE 2024;67(4):1023–35.
[80] Wang J, et al. MIC-SHAP: an ensemble feature selection method for materials machine learning. Mater. Today Commun. 2023;37:106910.
[81] Aminzadeh A, et al. A Machine Learning Implementation to Predictive Maintenance and monitoring of Industrial Compressors. Sensors 2025;25(4):1006.

Dr. Mustafa Tahsin Yilmaz earned his Ph.D. in Food Engineering from Selcuk University, Turkey, in 2009. Between 2012 and 2013, he completed post-doctoral studies at the University of Illinois at Urbana-Champaign (UIUC), United States. Since 2018, he has been serving as a Professor in the Department of Industrial Engineering at King Abdulaziz University. He has published more than 300 articles in internationally recognized journals on topics such as experimental design, artificial intelligence (AI), and quality control in chemistry, food science and engineering. He has received numerous awards for his contributions to research and product development from various institutions. Dr. Yilmaz has participated in over 70 interdisciplinary research projects. His current research interests include the application of machine learning and deep learning algorithms, and design of experiment (DOE) techniques for improving quality control and advancing the fields of chemistry, medicine and food science and engineering. He is particularly interested in the use of AI and machine learning in medicine, food biotechnology, safety and preservation, process optimization, new product development, and the integration of AI into quality control systems to enhance efficiency and accuracy in food production.

With over 7 years of experience leading production lines at Pladis Global Co., **Salman Badurayq** holds a Bachelor's degree and he is currently pursuing a Master's in Industrial Engineering. Additionally, He is a certified Project Management Professional (PMP). His expertise lies in driving operational excellence within FMCG, focusing on KPI analysis, labor management, and process optimization. Through his education and hands-on experience, he has developed strategies that enhance production performance, from meticulous scheduling to fostering a high-performance culture. In collaboration with labor coordinators, he prioritizes talent development and streamline operations to consistently exceed production goals. He champions continuous improvement, creating operating procedures that blend theoretical insights with real-world applications, ensuring both operational reliability and alignment with business objectives.

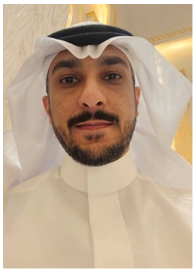
Dr. Kemal Polat graduated from the Electrical-Electronics Engineering Department at Selcuk University with a B.Sc. degree in 2002 and from the Electrical-Electronics Engineering Department at Selcuk University with an M.Sc. degree in 2004. He completed his Ph.D. in Electrical and Electronic Engineering at Selcuk University in 2008. He completed his post-doctoral degree in the Department of Electrical and Computer Engineering at the University of Houston between 2015 and 2016. In his post-doctoral work, he designed mathematical modeling of memory performance by designing various experiments on visual memory. He is now working as a Professor in the Electrical and Electronic Engineering Department, Engineering of Faculty, Bolu Abant Izzet Baysal University, since September 2011. He has 200 articles published in SCI journals and 100 international conference papers. His research interests include biomedical signal classification, control systems, electronics, statistical signal processing, visual memory, neuroscience, brain-computer interface, PPG signal, medical electronics, digital signal processing, pattern recognition, and classification. His Google h-index is 59. He is an IEEE senior member. He is the associate editor of Computers and Electrical Engineering (SCI), Elsevier.

Dr. Ahmad H. Milyani is an Associate Professor in the Department of Electrical and Computer Engineering at King Abdulaziz University, Saudi Arabia. He earned his Ph.D. in Electrical Engineering from the University of Washington in 2019, and both his M.S. and B.S. degrees in Electrical and Computer Engineering from Purdue University in 2013 and 2011, respectively.

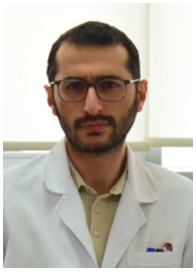
Dr. Milyani's research interests include power systems, energy markets, carbon emissions policy, and renewable energy integration. His work has been presented at leading IEEE conferences, including PowerTech and the Power & Energy Society General Meeting. He is a Certified LabVIEW Associate Developer and a member of IEEE and the IEEE Power and Energy Society since 2019.

At King Abdulaziz University, he plays an active role in academic and institutional development, serving as Vice Chairman of the General Courses Unit, Program Champion for ABET accreditation, and a member of multiple strategic committees. He is also affiliated with the Center of Research Excellence in Renewable Energy and Power Systems and Automation Laboratories.

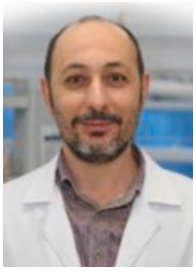
Dr. Milyani has received several honors, including the Outstanding Scholar Award from the Saudi Arabian Cultural Mission and the Barrett F. Robinson Outstanding Team Award from Purdue University.



Dr. Abdulaziz S. Alkabaa is an associate professor and department head of the Industrial Engineering Department at King Abdulaziz University, Jeddah. He received his PhD in ISE and a second master's degree in statistics from the University of Tennessee, and achieved several academic recognitions during his graduate studies. Dr. Alkabaa's research interests include stochastic processes, industrial statistics, design of industrial experiments, health care engineering, prediction modeling, and high-dimensional data mining.



Dr. Osman Gül graduated from Selçuk University, Faculty of Agricultural, Department of Food Engineering. He received the M.Sc. and Ph.D. degrees in Food Engineering from the Ondokuz Mayıs University, in 2009 and 2015, respectively. He is currently a Professor at the Department of Food Engineering, Faculty of Engineering and Architecture, Kastamonu University, Kastamonu, Türkiye. He has published more than 60 international papers. His research interests are food technology including plant based proteins, drying, microencapsulation, and powder technology.



Assoc. Prof. Dr. Furkan Türker Sarıcaoğlu completed his B. Sc., M.Sc., and Ph.D. in Food Engineering at Ondokuz Mayıs University, Turkey. He conducted part of his doctoral research at the USDA-WRRC with TÜBİTAK 2214-A support. Since 2018, he has been a faculty member at Bursa Technical University and leads the Saricaoglu Research Group.

His research focuses on biodegradable food packaging, plant protein modification, and nanotechnological food applications. He is the founder of Furya Nanolif R&D Inc., supported by TÜBİTAK-1512, and has led multiple national (TÜBİTAK 1001, 3501) and international (PRIMA, SEA-EU JFS) projects. He received the 2021 BTU Project Award and the 2024 TÜBA-GEBIP Award.

Dr. Sarıcaoğlu has 74 scientific publications, including 58 in international journals, and holds three patents. He has over 1400 citations and an h-index of 22 (Scopus). He serves as an editor or board member for several journals, including *Food Nutrition* (Elsevier) and *European Food Science and Engineering*.