Author/s:
Golzarijalal, M;Ong, L;Neoh, CR;Harvie, DJE;Gras, SL

Title:
Machine learning for the prediction of proteolysis in Mozzarella and Cheddar cheese

Date:
2024-03

Citation:
Golzarijalal, M., Ong, L., Neoh, C. R., Harvie, D. J. E. & Gras, S. L. (2024). Machine learning for the prediction of proteolysis in Mozzarella and Cheddar cheese. Food and Bioproducts Processing, 144, pp.132-144. https://doi.org/10.1016/j.fbp.2024.01.009.

Food and Bioproducts Processing 144 (2024) 132–144

Contents lists available at ScienceDirect

Food and Bioproducts Processing

journal homepage: www.elsevier.com/locate/fbp

Machine learning for the prediction of proteolysis in Mozzarella and Cheddar cheese

Mohammad Golzarijalal[a,b], Lydia Ong[a,b], Chen R. Neoh[a], Dalton J.E. Harvie[a], Sally L. Gras[a,b,*]

[a] Department of Chemical Engineering, The University of Melbourne, Parkville, VIC 3010, Australia
[b] The Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Parkville, VIC 3010, Australia

ARTICLE INFO

Keywords:
Gradient boosting
Random forest
Support vector regression
Cheese quality
SHAP analysis

ABSTRACT

Proteolysis is a complex biochemical event during cheese storage that affects both functionality and quality, yet there are few tools that can accurately predict proteolysis for Mozzarella and Cheddar cheese across a range of parameters and storage conditions. Machine learning models were developed with input features from the literature. A gradient boosting method outperformed random forest and support vector regression methods in predicting proteolysis for both Mozzarella (R^2 = 92%) and Cheddar (R^2 = 97%) cheese. Storage time was the most important input feature for both cheese types, followed by coagulating enzyme concentration and calcium content for Mozzarella cheese and fat or moisture content for Cheddar cheese. The ability to predict proteolysis could be useful for manufacturers, assisting in inventory management to ensure optimum Mozzarella functionality and Cheddar with a desired taste, flavor and texture; this approach may also be extended to other types of cheese.

1. Introduction

Proteolysis is a major biochemical event occurring in cheese during storage, where the protein network gradually breaks down, resulting in the release of proteins ranging from large peptides to free amino acids (Fox, 1989). The extent of proteolysis depends on the type of cheese, conditions of manufacture, type of ripening and the duration of storage. Mozzarella and Cheddar cheese share some similar manufacturing steps, with cooking and stretching steps involved in Mozzarella and not in Cheddar. They differ in composition and storage, however, resulting in different proteolytic profiles (Guinee, Harrington, et al., 2000). Low moisture Mozzarella cheese is usually consumed fresh, with a shelf life of up to 6 months (Mastromatteo et al., 2014), while Cheddar is commonly consumed after a period of storage, known as maturation or ripening, which lasts for up to 18 months or longer for extra mature Cheddar cheese (Muir, 1996). Additionally, Mozzarella cheese is typically stored at low temperatures (0–4 °C) or frozen (Golzarijalal et al., 2024) while Cheddar cheese is stored at relatively higher temperatures of about 7–13 °C (Walsh et al., 2020). Other factors that contribute to proteolysis include indigenous milk enzymes, coagulating enzymes, enzymes from starter bacteria, nonstarter bacteria, as well as secondary starters or adjunct cultures (for Cheddar cheese) (Fox and McSweeney, 1996).

The extent of proteolysis can be quantified by assessing the amount of soluble or extractable nitrogen (N) using various solvents, such as water, sodium acetate buffer or trichloroacetic acid (McSweeney and Fox, 1997). Two common indices of proteolysis are the level of water-soluble nitrogen (WSN), or the level of nitrogen extracted and quantified in buffers at pH 4.6 (pH 4.6-SN) (Kuchroo and Fox, 1982), which can account for the small and medium-sized peptides, free amino acids and degradation products resulting from proteolysis (McSweeney and Fox, 1997; Rank et al., 1985). Similar levels of soluble nitrogen have been reported for studies of internal bacterially ripened cheese types, such as Cheddar and Mozzarella, using either water or a buffer at pH 4.6 (Fox, 1989; Kuchroo and Fox, 1982; McSweeney and Fox, 1997), giving confidence in the similarity of these techniques. In some cases, however, higher levels of WSN have been reported during ripening (McSweeney and Fox, 1997).

Predicting the level of proteolysis is of importance to manufacturers, as it allows Mozzarella products to be distributed during a window of optimal functionality and Cheddar products to be marketed with favorable flavor and texture, as these traits are strongly influenced by proteolysis. In Mozzarella, functionality is commonly defined by stretchability and meltability (Fife et al., 1996), which decrease over time. Proteolysis is known to affect stretchability (Guinee et al., 2002;

* Corresponding author at: Department of Chemical Engineering, The University of Melbourne, Parkville, VIC 3010, Australia.
E-mail address: sgras@unimelb.edu.au (S.L. Gras).

https://doi.org/10.1016/j.fbp.2024.01.009
Received 29 September 2023; Received in revised form 15 January 2024; Accepted 24 January 2024
Available online 28 January 2024

Guinee, Harrington, et al., 2000; To et al., 2022), melt time and temperature (Guinee et al., 2001; To et al., 2022), flowability (Guinee et al., 2001, 2002; Henneberry et al., 2015), apparent viscosity (Guinee et al., 2001) and firmness (Guinee et al., 2001, 2002; Henneberry et al., 2015; To et al., 2022) in low moisture Mozzarella cheese. Proteolysis also impacts on the sensory characteristics of Cheddar (Fox and Wallace, 1997; Hannon et al., 2003a; Wilkinson et al., 1992), as well as the firmness and flowability (Fenelon and Guinee, 2000; Guinee, Auty, and Fenelon, 2000; Hou et al., 2014; McCarthy et al., 2016, 2017a).

One approach for predicting cheese proteolysis is to use multilinear regression. Linear regression successfully predicted proteolysis indices in Gamnonedo blue cheese as a function of storage over three months ($R^2 > 87\%$, for two batches, each containing 5 cheeses), although a second-order polynomial could more accurately predict water-soluble nitrogen ($R^2 = 0.98$) (de Llano González et al., 1993). The overall desirability of processed Gouda cheese could also be predicted as a function of storage time, storage temperature and several properties (e.g., texture, rheology, pH, sensory, etc.) using multilinear regression (with minimum $R^2 = 87\%$), however, in this case an artificial neural network was more accurate (with minimum $R^2 = 96\%$) (Stangierski et al., 2019). These examples highlighted above illustrate how multilinear regression can be less accurate. The existence of multicollinearity between variables (Coker et al., 2005) and nonlinear trends (Povak et al., 2014) could be the reasons for lower accuracy of linear regression technique, suggesting other methods should also be considered.

Non-linear models can offer more accurate predictions for proteolysis than linear regression, considering the effect of several interacting factors on proteolysis. These include cheese composition (McCarthy et al., 2016; Rudan et al., 1999), manufacturing conditions (Dave, McMahon, et al., 2003; Dave, Sharma, et al., 2003) and storage conditions (Feeney et al., 2001; Guinee et al., 1991), which can interact significantly (Dave, McMahon, et al., 2003; McCarthy et al., 2016) and also have collinearity and non-linear relationships with the level of proteolysis (Feeney et al., 2001; Ong and Shah, 2009). Previously, a model employing square root, logarithmic transformations and power polynomial procedures could predict the amount of total free amino acids (i.e., proteolysis products) in granular Cheddar cheese based on storage time, temperature and enzyme concentration ($R^2 > 0.96$) (J. C. C. Lin et al., 1987).

The development of more generalizable models would be of academic and industrial interest as these could potentially address the limited design space and lack of generalizability common to many non-linear models that predict proteolysis to date. For example, only 20–30 observations for two types of enzymes, each at 2–3 concentrations, at two storage temperatures and five storage times (up to 112 days) were used to predict granular Cheddar proteolysis in a non-linear model (J. C. C. Lin et al., 1987). Similarly, other non-linear models have successfully predicted proteolysis but their applicability is often limited to specific cases, e.g., a polynomial used for probiotic goat's milk cheese made in 16 batches (R^2 not reported) (Gomes et al., 1998) and a quadratic response surface model ($R^2 = 0.77$) for white brine cheese made in a single batch (Alizadeh et al., 2006), making their predictions difficult to generalize. Proteolysis has successfully been predicted using partial least-square regression for Emmental cheeses of several origins for a relatively large design space involving 91 samples with reasonable accuracy ($R^2 = 0.77$), but this model relies on near infrared spectroscopy data (Karoui et al., 2006), which are not always readily available.

Machine learning techniques offer an alternative approach which could offer greater generalizability for prediction of cheese proteolysis. Machine learning techniques have been developed for other areas of dairy manufacturing to predict quality parameters, optimize processing conditions and to classify food products based on physical and chemical properties (Slob et al., 2021). Machine learning can also be used to extract patterns from datasets and develop flexible models that capture relationships between process variables, enabling collinear input variables or non-linear relationships to be considered (Chan et al., 2022; Y. Lin et al., 2022).

Successful examples of machine learning in the dairy industry include a support vector regression used to determine the somatic cell count of raw milk using dielectric spectroscopy (Zhu et al., 2022) or to quantify common adulterants in powdered milk, such as starch, whey and sucrose, using near infrared spectroscopy (Borin et al., 2006). Random forest methods have also been used to predict cheese sensory profiles (Granitto et al., 2007), assess cheese maturity using microbial composition and metabolome profiles (Afshari et al., 2020) and predict consumer acceptance based on sensory evaluations (Rocha et al., 2020). A gradient boosting machine method has also predicted milk adulteration, using differential scanning calorimetry (Farah et al., 2021) and milk composition, using near-infrared spectral (Sheng et al., 2022) and infrared spectroscopic data (Sheng et al., 2022), illustrating the broad potential of these approaches. Despite the advantages offered by machine learning, it has not yet been used for predicting cheese proteolysis.

The aim of this study was to investigate the ability of different machine learning techniques to predict the level of proteolysis in Mozzarella and Cheddar cheeses using data collected from the literature. Readily available input features including chemical composition, storage time and well-known manufacturing or storage factors, such as pH at milling stage, type of acidification, coagulating enzyme and storage temperature were selected, as this information is typically recorded during manufacture. A second aim was to explain the impact of each input feature on model predictions, to identify the features with the greatest impact. Predictions of proteolysis can potentially be useful for process optimization and inventory management.

2. Materials and methods

2.1. Data collection and pre-processing

A detailed literature review was conducted to collect data in peer-reviewed papers on soluble nitrogen content in Mozzarella and Cheddar cheeses and to determine how proteolysis can be affected by composition, manufacturing and storage factors. To increase the size of the dataset and potentially improve the model performance, data were selected from both cheese protein (or nitrogen) solubilized in water (Kuchroo and Fox, 1982) or buffers at pH 4.6 (Kuchroo and Fox, 1982), since both these methods can effectively measure the overall extent of proteolysis (McSweeney and Fox, 1997). Mean values reported for soluble nitrogen data were obtained either directly from tables or from plots via the open-source software PlotDigitizer v.3.0 (https://plotdigitizer.com). PlotDigitizer has been widely used in the literature as an accurate tool for data extraction from plots (Aydin and Yassikaya, 2022; Jelicic Kadic et al., 2016). The data were restricted to Mozzarella and Cheddar made using pasteurized milk and coagulated by rennet, excluding camel rennet, and Mozzarella samples stored (ripened) at 4 °C, as these processes are the most common (Barbano et al., 1994; Guinee et al., 2002; O'Keeffe et al., 1975). Generally, only data obtained from standard, or control experiments, were collected and innovations in cheese production, such as experiments using six times the coagulating enzyme concentration (Dave, Sharma, et al., 2003) or a mutant starter organism (Guinee et al., 1991) were not considered. Large variations in cheese composition, however, such as low fat or salt content or cheese made with the addition of secondary or probiotic cultures were considered, to provide a broad dataset on which to train the model. A total of 327 and 832 datapoints were extracted from 26 papers on Mozzarella and 52 papers on Cheddar cheese. The extracted datasets for Mozzarella and Cheddar cheese, along with the papers from which the data was extracted, are listed in Table A1 and A2, respectively.

Input features and response variables were determined for both cheese types. Potential input features selected for both Mozzarella and Cheddar cheese were four compositional properties, including protein (%w/w), fat (%w/w), moisture (%w/w) and salt content (%w/w), as well as three key manufacturing variables, including milling pH, coagulating enzyme concentration (measured in international milk-clotting units (IMCU)/kg of milk) and the storage time after production (measured in days). The response variable was soluble nitrogen, expressed as percentage of total cheese nitrogen (%TN). Two additional input features were selected for Mozzarella cheese, these were calcium content (%w/w) and the type of acidification process (i.e., enzymatic using a starter culture or chemical). These additional features were not considered for Cheddar cheese, as the calcium content was missing for the majority (75%) of data and most Cheddar samples (94%) were produced using a starter culture acidification, so the inclusion of other methods of acidification could bias predictions (Jeong et al., 2021). Instead, two other features of storage temperature (°C) and the addition or absence of adjunct culture were selected as relevant and important factors for Cheddar manufacture. Coagulating enzyme concentration was calculated by multiplying the enzyme activity with the dosage of enzyme used for cheesemaking (McCarthy et al., 2017b; Wang and Moraru, 2021). Enzyme activity for each experiment was either collected from the corresponding paper or found in the relevant manufacturers' catalog. The activity was calculated using the IDF Standard 157:2007 for calf rennet (IDF 157, 2007), details of which can be found elsewhere (Tabayehnejad et al., 2012). The final set of input features was refined to minimize high correlations between features and multicollinearity (Fig. 1).

Highly correlated features or features causing multicollinearity were identified, to ensure that features included in the model provide unique information. Multicollinearity occurs when there are two or more highly correlated input features (i.e., there is an approximately linear relationship between input features) and can lead to a misleading explanation of the effect of an input feature on the response variable (Chan et al., 2022). One way to assess the strength and direction of linear relationships between any pair of features, is to calculate the Pearson correlation as Eq. (1) (Rodgers and Nicewander, 1988):

$$r_{xz} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{m}(z_i - \bar{z})^2}} \tag{1}$$

where r_{xz} is the Pearson correlation between features x and z. $\bar{x}$ and $\bar{z}$ are mean of features in the entire dataset with size m (i.e., 327 and 832 for Mozzarella and Cheddar, respectively). Pearson correlation (r) varies between -1 and $+1$ and values in the range of $|r| \leq 0.35$, $0.35 < |r| \leq 0.67$ and $|r| > 0.67$ were considered low, moderate and high correlations, respectively using established criteria (Taylor, 1990).

Variance inflation factor (VIF) was calculated as an indicator of multicollinearity, calculated as follows for an input feature j (Chan et al., 2022):

$$VIF_j = \frac{1}{(1 - R_j^2)} \tag{2}$$

where R_j^2 is the coefficient of determination calculated from regression of the input feature j on other features. No formal value has been reported for VIF to determine multicollinearity but a value greater than 10 often indicates multicollinearity (Chan et al., 2022; Stine, 1995). To find the Pearson correlation and VIF for the two datasets, categorical input features were assigned numerical values of 0 and 1 for chemical and enzymatic (i.e., starter culture) acidification of Mozzarella or for inclusion or exclusion of an adjunct culture for Cheddar, respectively.

Next, data pre-processing steps were taken to increase model robustness. Firstly, 80% of datapoints were randomly assigned into a training set and the remaining 20% allocated into a test set. These same training and test sets were applied for all machine learning techniques. Pre-processing was then performed separately for the training and test sets, rather than on the entire dataset (training + test set), to prevent information leakage, which can lead to over-estimation of predictions (Kapoor and Narayanan, 2022). Missing values in the datasets (17% in milling pH and 20% in calcium content of Mozzarella and 3% in milling pH, 3% in salt content and 6% in enzyme concentration of Cheddar dataset) were treated by median imputation, to avoid bias or poor model performance (Emmanuel et al., 2021). Due to the skewed distribution of data, which contained several outliers, median imputation was used to replace missing data in both training and test sets. This is a common method for treating missing data in skewed distributions (Hasan et al., 2021), which involves replacing the missing values with the median of the available data in the training set.

Finally, data for all input features and response variables were normalized using the following equation (J. Li et al., 2020):

$$x_i^* = \frac{x_i - \mu}{s} \tag{3}$$

where x_i is the original value of the feature i, μ and s are the mean and the standard deviation of data of feature i in the training set, respectively and x_i^* is the normalized value of x_i.

2.2. Modeling methods

The set of potential input features were refined using the Pearson correlations (Eq. (1)) and VIF values (Eq. (2)) to identify highly correlated features and as an index of multicollinearity, respectively. A multilinear regression model and three different supervised machine learning algorithms, i.e., random forest, gradient boosting and support vector regression, were developed to find the best method for determining predictive relationships among the input features and response variable (i.e., level of proteolysis). MATLAB 2022R2 was used to build all four models. The general workflow of the study is shown in Fig. 1.

A 5-fold cross-validation method (Hosseinzadeh, Zhou, Altaee, et al., 2022) was performed using the training dataset to select the best hyperparameters for machine learning algorithms to improve robustness and predictability and to avoid overfitting (Sun et al., 2022). Training data were randomly split into 5 subsets of which 4 were used for training the model and 1 for this additional validation step. The training-validation process was repeated 5 times for different combinations (i.e., each time a different subset was chosen as the validation

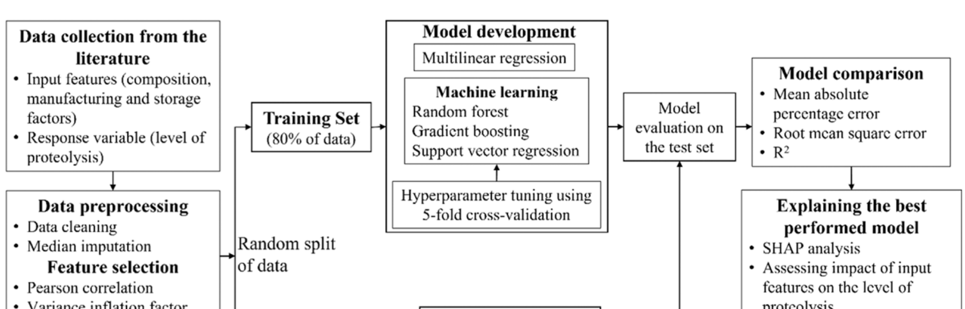

Fig. 1. General workflow of the study.

set) and the average root mean square error (RMSE) for the 5 assessments was considered for the model (Fig. A1). A Bayesian optimization method was next used to tune the hyperparameters for each machine learning algorithm, improving predictive performance. The hyperparameters are further described below in the relevant sections for each algorithm.

Performance of the multilinear regression and machine learning models was estimated on the external test set using measures of mean absolute percentage error (MAPE), root mean square error (RMSE) and the coefficient of determination (R^2), which are defined as follows (Hasan et al., 2021):

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n}} \tag{5}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{6}$$

where y_i and $\widehat{y}_i$ are the actual and predicted response variable of soluble nitrogen, respectively, $\overline{y}$ is the average value of actual soluble nitrogen and n is the number of datapoints in the test set.

Next, the potential for overfitting and underfitting was assessed, where a model performs very well on training data but fails to generalize on external, or unseen data such as the test set in this study, or when the dataset is too small, so the model performs poorly on both training and test sets, respectively (You et al., 2022). A learning curve was used as a diagnostic tool to assess machine learning model performance (i.e., mean squared error) on both the training and test sets, while holding the test set fixed and gradually increasing the size of the training data (Giola et al., 2021; Hosseinzadeh, Zhou, Zyaie, et al., 2022). This approach should not be confused with the same term used in artificial neural network literature (Viering and Loog, 2021). To generate learning curves for all machine learning algorithms, the original data was partitioned into training and test sets using an 80:20 ratio, as explained in Section 2.1, with the same test set used throughout the analysis.

2.2.1. Multilinear regression

A multilinear regression method was first used to fit the data; this statistical method can analyze the relationship between several input features, or independent variables, and a response, or dependent variable. In this method, the goal is to find a linear relationship between the input features and the response variable, using the following linear model (Uyanık and Güler, 2013):

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots \tag{7}$$

where Y is the response variable and $X_1, X_2,$. are input features. $b_0, b_1, \ldots$ are model coefficients, which can be estimated by applying a least square method, where the sum of the squared differences between the predicted and actual values of the response variable are minimized (Sergent et al., 1995).

2.2.2. Random forest

A random forest, a type of ensemble learning method that combines multiple decision trees, was applied next, where each tree acts in parallel as a single regressor and predictions are aggregated to derive the final model prediction. Each decision tree is trained individually using a bootstrap-aggregation method and feature-randomness technique (Prasad et al., 2006). The former reduces the prediction variance (J. Li et al., 2020) by selecting a random subset of training data with replacement (i.e., individual datapoints can be selected multiple times) and the latter ensures low correlation among decision trees by sending a random subset of input features into each decision tree (Prasad et al., 2006) (as shown schematically in Fig. A2).

Several hyperparameters were tuned using Bayesian optimization to enhance the random forest prediction, including the number of decision trees, the minimum leaf size (determining the minimum number of datapoints required to be at a leaf node), the maximum number of splits and the number of input features to select randomly for each split. Table A3 lists the hyperparameter ranges used to build the random forest model, selected using MATLAB, for Mozzarella and Cheddar cheese. A narrower range (i.e., [1,10]) was selected for the maximum number of splits, which determines how each decision tree can grow during the training, to avoid the formation of deep trees and risk of overfitting (Patange et al., 2022).

2.2.3. Gradient boosting

Gradient boosting, another type of ensemble learning method, that builds decision trees in series was also applied. In this method, each tree acts as a weak learner to create an arbitrarily strong learner (Friedman, 2001). The ensemble starts with a single weak learner, defined as the tree that can perform at least slightly better than a random guess on a given training set (Zhou et al., 2021) and new weak learners are added sequentially, with the aim of minimizing the overall error or loss of the model during training (as shown schematically in Fig. A3). Here, a least-squares regression algorithm (Friedman, 2001) was used to implement the mean squared error loss function (Elavarasan and Vincent, 2020) on the training dataset. The supplementary material also provides more details on the gradient boosting theory.

The performance of the gradient boosting model was enhanced by tuning a set of hyperparameters, using the method applied above for the random forest (see Table A3 for tuning ranges), where the hyperparameters were the number of decision trees, the learn rate, the minimum leaf size, the maximum number of splits and the number of input features to select randomly for each split.

2.2.4. Support vector regression

The third machine learning method applied, support vector regression, is based on statistical learning theory (İskenderoğlu et al., 2020) and is suitable for linear and non-linear regression problems. In this method, a margin of error, ε, is defined by creating an ε-insensitive region, where the goal is to find a function that has at most ε deviation. A kernel function is normally useful to transform training datapoints from a low-dimensional to a higher-dimensional vector space, where a linear hyperplane can be found (Smola et al., 2004). Here, a Gaussian radial basis function was employed as the kernel function, since it has reportedly generated promising predictions in previous studies (Hosseinzadeh, Zhou, Altaee, et al., 2022; J. Li et al., 2020; Nguyen et al., 2021). Support vector regression attempts to formulate the kernel function as an optimization problem, aiming at finding the narrowest ε-insensitive region possible, i.e., reducing the prediction error. A soft margin loss function (Cortes and Vapnik, 1995) was also used here to introduce slack variables for predictions that are outside of the region (farther than ε from actual values). A regularization parameter (C) was also defined, which controls the penalty imposed on slack variables and helps to prevent overfitting (as shown schematically in Fig. A4). The supplementary material provides further details on the support vector regression theory.

Epsilon (ε) and the regularization parameter (C) were tuned to find the best support vector regression model. The kernel scale, or kernel coefficient parameter (Nguyen et al., 2021), was also tuned, as this is another important hyperparameter for a gaussian kernel that controls the width of the Gaussian and the flexibility of the model, which can affect the predictability of support vector regression (Ben-Hur et al., 2008). Ranges determined in MATLAB were used for hyperparameter tuning, as listed in Table A3.

2.3. Model explainability and the impact of input features

After developing the four models and evaluating their predictive performances, the model with the best performance (with respect to the metrics introduced in Eqs. (4)-(6)) was assessed to determine the importance of specific input features. The method of SHapley Additive exPlanations (SHAP), an extension of Shapley values from cooperative game theory (Shapley, 1953), was adopted as an accurate and consistent method (Lundberg et al., 2018) to help explain the importance of each input feature in the predictions of the selected model (Lundberg and Lee, 2017). Local SHAP values were calculated by determining the marginal contribution of each input feature to the prediction of every single observation in the training set (i.e., 262 and 666 datapoints for Mozzarella and Cheddar, respectively) considering all possible combinations of input features (Jeong et al., 2021; Lundberg and Lee, 2017). Local SHAP values can be positive or negative, indicating whether a feature has a positive or negative impact on the deviation of the prediction from the average (calculated from all possible combinations of input features), i.e., positive and negative values show a positive and negative contribution of the feature, respectively (Kruse et al., 2021; Lundberg and Lee, 2017). A score was assigned to each feature across all observations in the training set, reflecting its relative importance in each prediction and local interpretability. To make numerical representation of categorical features possible in SHAP summary plots, these features were encoded by assigning integers 0 and 1 to chemical and starter culture acidification for Mozzarella or for exclusion or inclusion of an adjunct culture for Cheddar, respectively. A global SHAP value was also calculated for each feature by adding absolute local SHAP values, which can be used to identify the most and least significant features contributing to the response variable across the set of input features.

3. Results and discussion

Data on Mozzarella and Cheddar cheese proteolysis from the literature were collected to identify potential input features into the multilinear regression and three machine learning models (random forest, gradient boosting and support vector regression), together with the corresponding level of soluble nitrogen. Input features included cheese composition, manufacturing variables and storage variables. Multicollinearity assessment was performed to exclude highly correlated features. The predictive performance of the four models was then compared and the best model was used to determine the local and global significance of each feature in predicting soluble nitrogen. Each of these steps and the results obtained from the four models are examined here in turn.

3.1. Exploration of input features for Mozzarella and Cheddar cheese models

To find linear correlations between the potential input features and assess multicollinearity, Pearson correlations (r) (Eq. (1)), as well as the variance inflation factor (VIF; Eq. (2)), were evaluated. Values greater than 0.67 for the Pearson correlation (Taylor, 1990) and 10 for VIF (Chan et al., 2022; Stine, 1995) were considered a measure of high correlation and a sign of multicollinearity, respectively.

A high negative Pearson correlation was observed between protein and fat for both Mozzarella and Cheddar (−0.75 and −0.88, respectively; Fig. A5 and A6). High negative correlations were also observed between fat and moisture for both cheeses (−0.69 and −0.81, respectively) and between protein and moisture for Cheddar (0.65). These high correlations were expected, as the proportions of different components are related, with a change in one component affecting another. A high negative correlation was also observed between the milling pH and the type of acidification employed to make Mozzarella ($r = -0.7$), since a higher milling pH of ~5.6 is often used for chemical acidification (Feeney et al., 2002; Henneberry et al., 2015; Sheehan and Guinee, 2004), compared to a typical pH of ~5.2 used for starter culture acidification. Multicollinearity was observed among the Mozzarella input features of protein, fat and moisture, with VIF of ~10, 21 and 10.9, respectively (Table A4) but was not observed in the dataset collected for Cheddar cheese, probably due to higher number of observations in the Cheddar dataset (i.e., 832) compared to that of Mozzarella (i.e., 327) (Guan and Fu, 2022).

Fat was selected as a feature to retain in the model and protein excluded, to reduce correlation between input features, as the fat content varies across a wider range for both Mozzarella (~4–28%w/w) and Cheddar cheese (~1–40%w/w), compared to the variation in protein (~23–36%w/w for Mozzarella and ~21–40%w/w for Cheddar). It should be noted that variations in protein could also be considered as substantial, however, this variation is less than the variations observed in the fat content. Fat variation is also reflected in different cheese categories, including low-fat (Fife et al., 1996), reduced fat (Rudan et al., 1998) and full fat cheese (Rudan et al., 1999). This selection reduced the high Pearson correlations and VIF, which could otherwise make the model unreliable (Chan et al., 2022) and led to a significant decrease in VIF with a maximum of ~6 and ~3 (Table A4) for the Mozzarella and Cheddar datasets, respectively (Fig. A5 and A6). Some linearly correlated features were still present in the dataset after feature selection, which can potentially reduce the accuracy of the models developed, specifically for multilinear regression, as this method is less effective than machine learning in handling correlated datasets (Chan et al., 2022).

A non-linear relationship was observed between the selected input features and soluble nitrogen for both cheese types, as shown by scatterplots of the data (Fig. A7 and Fig. A8). Only the storage time showed a slight linear trend with an R^2 of 0.44 and 0.46 (based on linear regression for a single parameter) for Mozzarella (Fig. A7g) and Cheddar (Fig. A8g), respectively. This observation aligns with the low Pearson correlations calculated for soluble nitrogen with most input features (< 0.27) and the relatively high correlation between soluble nitrogen and storage time (> 0.66) (Fig. A5 and A6).

Most input feature data collected fell in a similar range for both the Mozzarella and Cheddar cheeses, with a similar median occurring for both datasets as shown in Fig. 2. The median was similar for salt content (median of 1.6%w/w for Mozzarella and 1.7%w/w for Cheddar), enzyme concentration (median of 38.5 IMCU/kg of milk for Mozzarella and 44 IMCU/kg of milk for Cheddar) and milling pH (median of 5.2 for Mozzarella and 5.4 for Cheddar), with a few outliers as exceptions. Mozzarella typically contains a higher moisture content than Cheddar (median of 48.4%w/w and 39.3%w/w for Cheddar), which gives a higher meltability (Lefevere et al., 2000), although this also leads to a shorter shelf life (25 days median and 91 days maximum for Mozzarella compared to 60 days median and 270 days maximum for Cheddar; Fig. 2). Cheddar has a higher fat content than Mozzarella (31%w/w median and 21%w/w median) (Fig. 2) and most Mozzarella cheeses had calcium concentration ranging between 551 mg/100 g of cheese and 976 mg/100 g of cheese, with a few outliers slightly above or below this range (Fig. 2a). Within the categorical input features, use of starter culture acidification (or a starter culture) was more common than chemical acidification for manufacture of Mozzarella manufacture (Fig. 2a) and the absence of adjunct culture was more common for Cheddar manufacture (Fig. 2b).

3.2. Evaluation of model performance

The predictive performance of the models was compared, showing that the three machine learning algorithms outperformed multilinear regression in the prediction of soluble nitrogen, resulting in lower errors and higher R^2 values, as shown in Table 1. Figs. 3 and 4 also illustrate the high correlation between predicted and experimental data for Mozzarella and Cheddar using machine learning techniques, respectively.

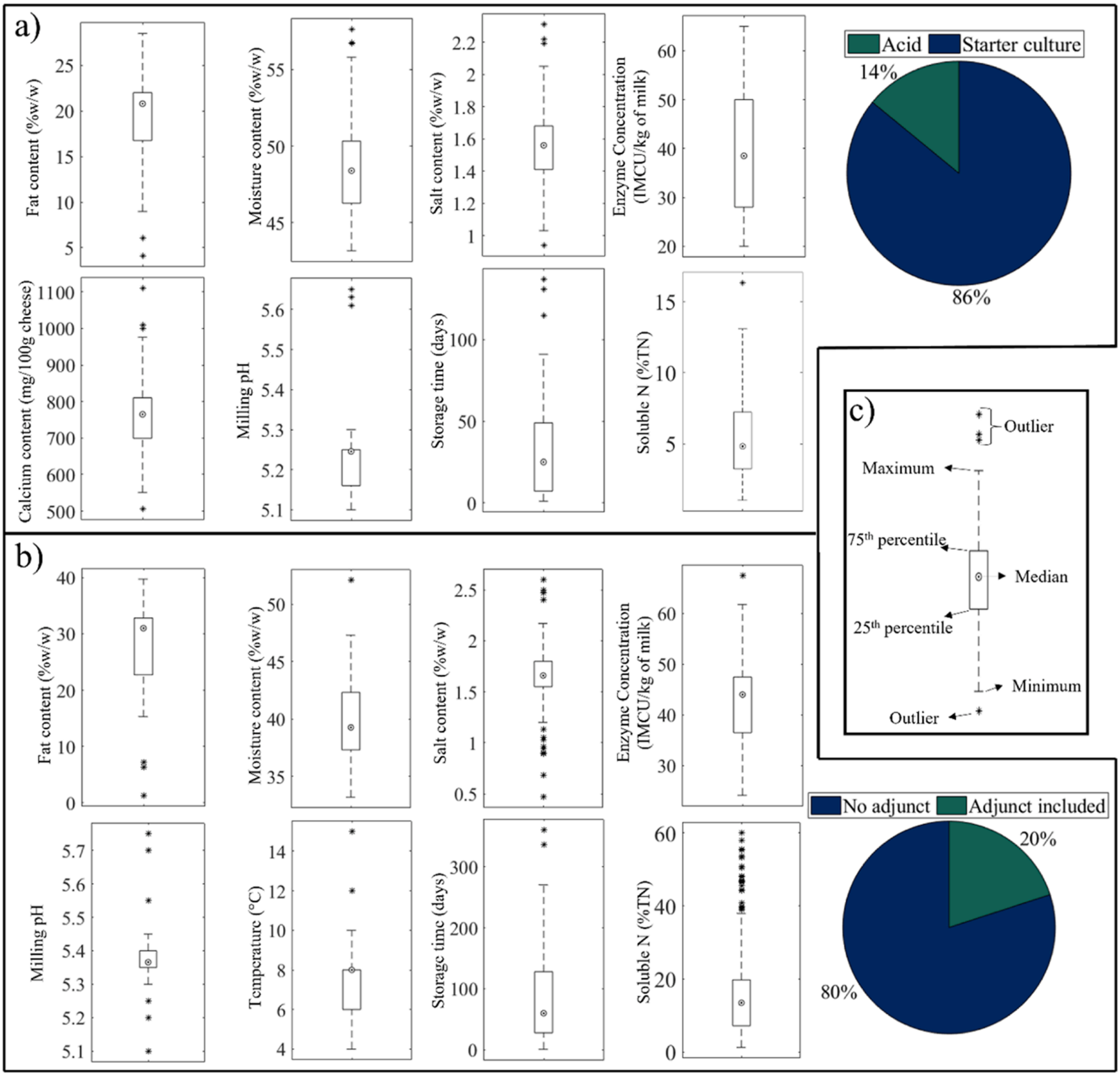

Fig. 2. The distribution of data for input features and the response variable (soluble N (nitrogen)) for (a) Mozzarella (n = 327) and (b) Cheddar cheese (n = 832) used to build the predictive models. Box plots and pi charts are used to represent the distribution of numerical and categorical data, respectively. (c) The box in each plot represents the interquartile range containing the middle 50% of the data, where the four lines from bottom to the top in each box plot are the minimum, the first quartile, the third quartile and the maximum of the statistical data, respectively. In addition, circles (⊙) and asterisks (*) on each boxplot show the position of the median and any outliers, respectively.

As it can be seen from Fig. 2, the response variable of soluble nitrogen was higher in Cheddar (13.48%TN median and 38%TN maximum) than in Mozzarella (4.83%TN and 13.1%TN, respectively). This is a direct result of the shorter shelf life of Mozzarella, where less proteolysis occurs, whereas the longer shelf life and greater proteolysis applied to Cheddar cheese assists in the development of desirable qualities, such as flavor (Hannon et al., 2003b). Processing and storage factors, such as the use of an adjunct culture (Lane and Fox, 1996) (20% of datapoints in Fig. 2b) or a higher storage temperature (Fenelon et al., 1999) (i.e., > than the 8 °C median in Fig. 2b), can also accelerate proteolysis in Cheddar. Additionally, comparing the soluble nitrogen data of Mozzarella (Fig. 3) and Cheddar (Fig. 4), the latter dataset has a different scale, with most data tightly clustered near the origin between 0–30% total nitrogen (TN) and some data spread at higher TN (Fig. 2).

Soluble nitrogen levels fell in a generally similar range when obtained by pH 4.6 and WSN analyses for Mozzarella and Cheddar cheese, as depicted in Fig. A9a and Fig. A9b, where, excluding the outliers, the data for pH 4.6 soluble nitrogen were from 1.03%TN to 12.06%TN and for WSN from 1.89%TN to 16.33%TN for Mozzarella and from 1.84%TN to 32.84%TN and from 1.31%TN to 39.91%TN for Cheddar respectively. This observation reflects trends reported elsewhere for internal bacterially ripened cheese types (Fox, 1989; Kuchroo and Fox, 1982; McSweeney and Fox, 1997). For Mozzarella, the range of the WSN was slightly larger than the pH 4.6 soluble nitrogen and for Cheddar the

Table 1
Predictive performance of the four different models using the metrics of root mean squared error (RMSE), mean absolute percentage error (MAPE) and the coefficient of determination (R^2) for Mozzarella and Cheddar cheese. The same set of features, test set and data split ratio of 80:20 (training:test data) were used for all four models.

Mozzarella cheese	Multilinear regression	Random forest	Gradient boosting	Support vector regression
RMSE[a] (%TN)	0.6	0.43	0.08	1
RMSE (%TN)	1.7	1.34	0.77	1.12
MAPE (%)	36	26	13	17
R^2 (%)	60	75	92	82
Cheddar cheese	**Multilinear regression**	**Random forest**	**Gradient boosting**	**Support vector regression**
RMSE[1] (%TN)	0.57	0.35	0.05	1.49
RMSE (%TN)	5	3.36	1.44	2.13
MAPE (%)	37	26	9	13
R^2 (%)	67	85	97	94

[a] The metric was measured on the training set.

range of WSN was also greater; higher WSN values have been reported previously (McSweeney and Fox, 1997) and in the case of Cheddar an increase in cheese pH during ripening could contribute to these changes. The two measures were assumed similar and effectively combined in the analysis and modeling performed here but could potentially be separated for aged cheese with extensive proteolysis in future modeling.

Multilinear regression gave an RMSE of 1.7%TN and 5%TN and failed to predict a fraction of the unseen data in the test set for Mozzarella (36%) and Cheddar (37%) (Table 1) (as measured by MAPE). The inferior performance of multilinear regression could be attributed to the presence of linearly correlated features (e.g., fat and moisture content; Fig. A5 and A6), as well as the existence of a non-linear relationship between all input features, except for the storage time, and soluble nitrogen, as identified by Pearson correlations smaller than 0.27 (Fig. A5 and A6), as well as in the corresponding scatter plots (Fig. A7 and A8). This finding is consistent with previous reports of the superior performance of machine learning methods applied to other foods, including in studies of the mechanical parameters of cranberries (Gorzelany et al., 2022), as well as predictions of total soluble solids and the pH of strawberries (Adedeji et al., 2022) or variation in the yield of maize (M. Li et al., 2021).

The gradient boosting and support vector regression algorithms were better able to predict soluble nitrogen compared to the random forest algorithm. While the training and test errors converged to a low value in learning curves plotted for the three algorithms (Fig. A10 and A11 for Mozzarella and Cheddar, respectively), signaling generalizable and well-fitted models that did not over- or underfit the data (You et al., 2022), a larger gap remained between the curves for the random forest algorithm (Fig. A10a and A11a for Mozzarella and Cheddar, respectively). This gap indicates that adding more than 100 datapoints into the training set did not further improve performance for the random forest algorithm. The increased training error observed in the learning curves of support vector regression for Mozzarella with the addition of more training samples could be due to a small epsilon-insensitive region (Fig A4). This could result in more training samples falling outside this region, leading to the observed greater error. The model for Mozzarella had a significantly smaller epsilon value (~0.001) than Cheddar (~0.37) (Table A3), consistent with this hypothesis.

Gradient boosting provided the most accurate prediction of soluble nitrogen, using the metrics defined as Eqs. (4–6), with an RMSE of 0.77%

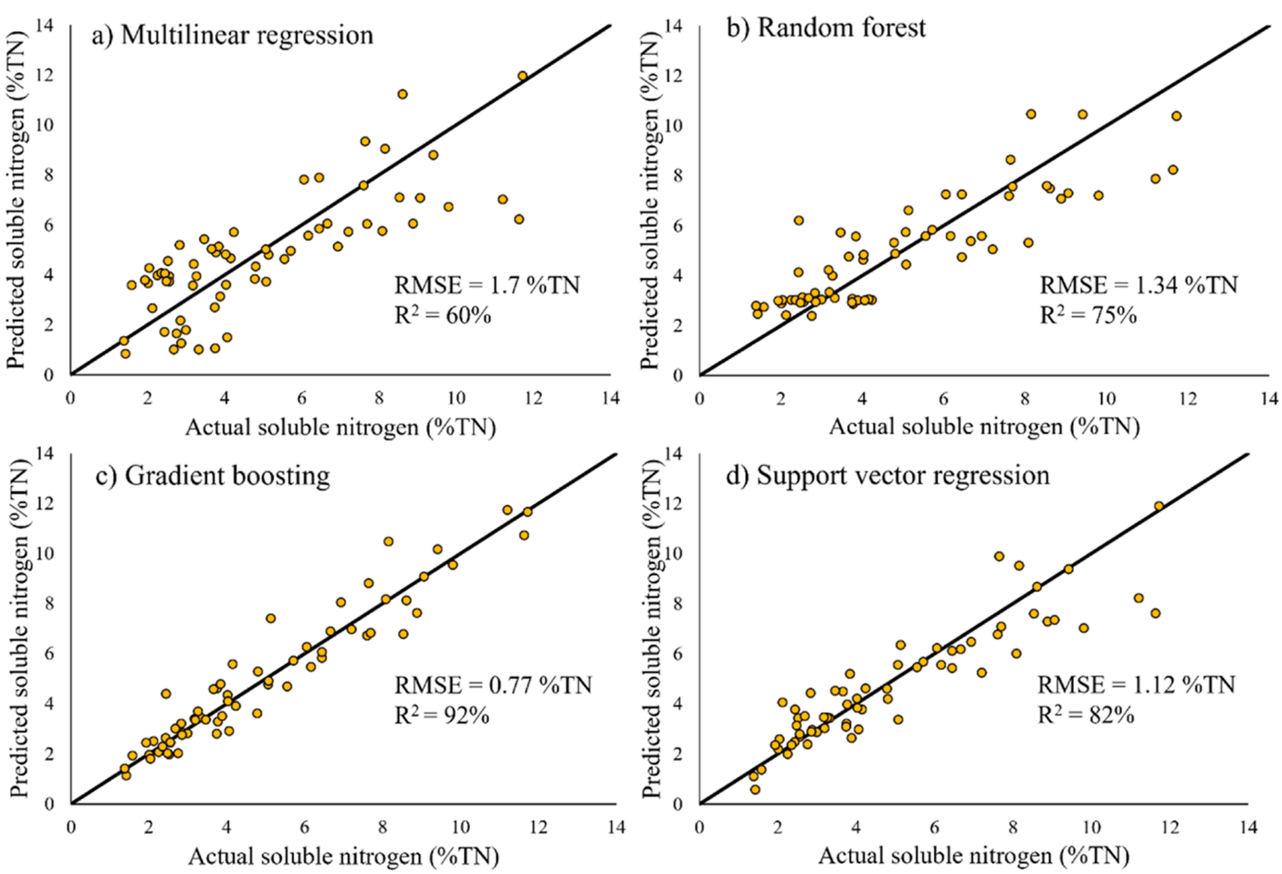

Fig. 3. Correlation between predicted and actual experimental data for soluble nitrogen determined using (a) multilinear regression, (b) random forest, (c) gradient boosting and (d) support vector regression models for Mozzarella cheese. The same set of features, test set were used for all four models and data were split with a ratio of 80:20 between training and test sets. RMSE is the root mean squared error (Eq. (5)) and R^2 the coefficient of determination calculated for the test set (Eq. (6)).

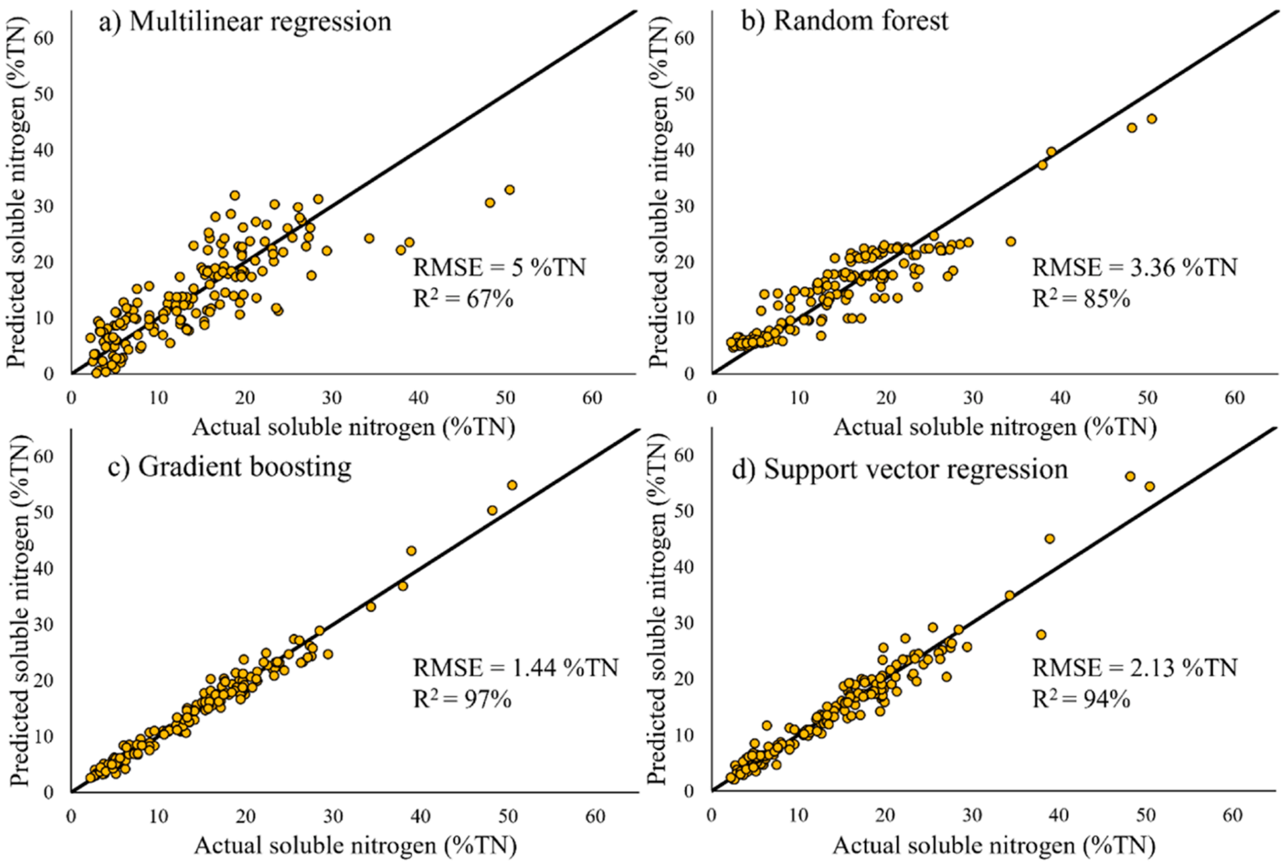

Fig. 4. Correlation between predicted and actual experimental data of soluble nitrogen determined using (a) multilinear regression, (b) random forest, (c) gradient boosting and (d) support vector regression models for Cheddar cheese. The same set of features and test set were used for all four models and data were split with a ratio of 80:20 between training and test sets. RMSE is the root mean squared error (Eq. (5)) and R^2 the coefficient of determination calculated for the test set (Eq. (6)).

TN, MAPE of 13%TN and R^2 of 92% for Mozzarella and an RMSE of 1.44%TN, MAPE of 9% TN and R^2 of 97% for Cheddar cheese. This superior performance was also evident in the denser scatter plots, where data clustered closer to the x = y line for Mozzarella (Fig. 3) and Cheddar (Fig. 4). The use of median imputation to replace missing data was also effective, as gradient boosting performed on the dataset with missing data resulted in a higher MAPE, i.e., 19% for Mozzarella and 12% for Cheddar. Gradient boosting has previously been reported to outperform support vector regression in the prediction of genomic breeding values (Ogutu et al., 2011) and to perform better than a random forest algorithm in determining milk adulteration using differential scanning calorimetry profiles (Farah et al., 2021).

The support vector regression algorithm performed better than the random forest algorithm, resulting in lower RMSE and MAPE, as well as higher R^2 for both Mozzarella and Cheddar cheese; the scatter was also lower in the plots of predicted and actual data (Figs. 3 and 4). Previously a support vector regression algorithm predicting fuel characteristics of hydrochar and pyrochar also performed better than a random forest alternative (J. Li et al., 2020). It should be noted, however, that the performance of machine learning algorithms can vary depending on the specific problem or data being analyzed (Awan et al., 2020). For instance, a random forest algorithm outperformed a support vector machine in predicting potential mining locations (Rodriguez-Galiano et al., 2015) and also outperformed a gradient boosting algorithm in the prediction of total soil nitrogen and carbon (Nawar and Mouazen, 2017), making it necessary to assess the best method for each application.

Moisture-in-nonfat substance (MNFS) could also potentially be another important feature, as it can significantly impact the level of proteolysis in Mozzarella (Rudan et al., 1999) and Cheddar (Fenelon et al., 2000) cheese. Including MNFS as an additional feature, however, did not improve the performance of the gradient boosting algorithm, giving an MAPE of 16% and 9% for Mozzarella and Cheddar, respectively. Similarly, replacing the moisture content with MNFS did not enhance the gradient boosting algorithm, with an MAPE of 13% and 11% for Mozzarella and Cheddar, respectively. Therefore, for the current dataset, the model is trained better when considering fat and moisture as features than when MNFS is included as a feature.

The gradient boosting algorithm, which performed best here, was further examined to ensure optimal performance by examining 40 different cases using 10 different seedings for random splitting of the data (numbered 0, 1, …, 9) and 4 different split ratios of the data (i.e., ratios of 75:25, 80:20, 85:15 and 90:10 for training: test set data). This approach, following a published method (Jeong et al., 2021), was applied to both Mozzarella and Cheddar and the same metrics used to assess the performance.

Good performance was observed for the gradient boosting algorithm for the expanded set of data cases, which resulted in only small changes in RMSE, MAPE and R^2 for both cheese types (Table A5 and A6). The RMSE ranged from 0.59%TN to 0.95%TN for Mozzarella and 1.17%TN to 1.89%TN for Cheddar, while the MAPE ranged from 7% to 16% and from 6% to 12% respectively, with the R^2 ranging from 85% to 97% and from 95% to 99% (Table A5 and A6). This data indicates that the tool developed is potentially generalizable and can predict the level of proteolysis in Mozzarella and Cheddar cheeses.

3.3. Model interpretation using SHAP analysis

The successful gradient boosting algorithm was selected for further analysis to assess the impact of input features on the level of soluble nitrogen predicted for Mozzarella and Cheddar cheese. To this end, local SHAP values were calculated for the observations in the training sets of Mozzarella (262 datapoints) and Cheddar (666 datapoints) to assess the impact of input features.

Storage time was found to be the most influential factor on both the Mozzarella and Cheddar cheese models using SHAP analysis, as shown in Fig. 5a and b, respectively. These SHAP summary plots present both the global patterns of each of the eight input features and individual variability in the level of soluble nitrogen predicted using each of the datapoints used in the training set for each model. Positive or negative values in Fig. 5a or b show a positive or negative contribution to the deviation from the average prediction (Kruse et al., 2021). The color scale also indicates a value within a normalized range for each input feature. For example, in Fig. 5a, a storage time feature with a blue −1 SHAP value indicates that the storage time was at its minimum value (i.e., 1 day) and that the datapoint represented contributed negatively to the prediction of soluble nitrogen level.

While storage time was of high importance for both models, the relative importance of other input features varied for the Mozzarella and Cheddar cheese models, as illustrated in Fig. 5c and d, where the impact of the input features is ranked in order of relative importance. After storage time, enzyme and calcium concentrations were next most important for Mozzarella, whilst fat and moisture content were of next highest importance for Cheddar, reflecting the different composition and manufacturing processes for the two cheeses.

The dominant positive effect of cheese storage time on the proteolytic predictions for both Mozzarella and Cheddar (Fig. 5) was expected, as proteolysis increases over time. It has been well documented in the literature that aging cheese results in greater bacterial and enzymatic activity, breaking down the protein network into smaller proteins and peptides, giving rise to an increase in the concentration of soluble nitrogen in both Mozzarella and Cheddar cheese (Costabel et al., 2007; Wilkinson et al., 1992), consistent with the importance assessed here. Without the inclusion of storage time as an input feature, the gradient boosting model performs poorly, with an MAPE of 64% for Mozzarella and 73% for Cheddar. Similarly, building a model solely based on storage time would result in inadequate MAPE values of 30% for Mozzarella and 34% for Cheddar, illustrating the importance of a complete set of input features that includes storage time and other features.

Importantly, the effect of the other next important features on the models can also be explained by observations in the literature. Increasing the concentration of coagulating enzyme had a positive impact on the level of soluble nitrogen (Fig. 5a), which can be due to an increase in the concentration of residual rennet (Dave, McMahon, et al., 2003). This is consistent with previous reports that the level of coagulating enzyme in Mozzarella made with 0.25X, 1X and 4X rennet (where X = 50 IMCU/kg of milk) reached a stage of 50% intact β-casein 56, 21 and 6 days after production and storage at 4 °C, respectively (Dave, McMahon, et al., 2003). Reducing Calcium, impacted positively on proteolysis in Mozzarella (Fig. 5a), likely as a lower calcium concentration can enhance the susceptibility of casein micelles to proteolysis, leading to a greater degree of primary and secondary proteolysis (Feeney et al., 2002; Guinee et al., 2002). Additionally, the manufacturing process of reduced calcium Mozzarella involves lower pH levels at the setting and whey drainage stages, which leads to increased moisture content and rennet retention, leading to enhanced proteolysis (Feeney et al., 2002; Guinee et al., 2002).

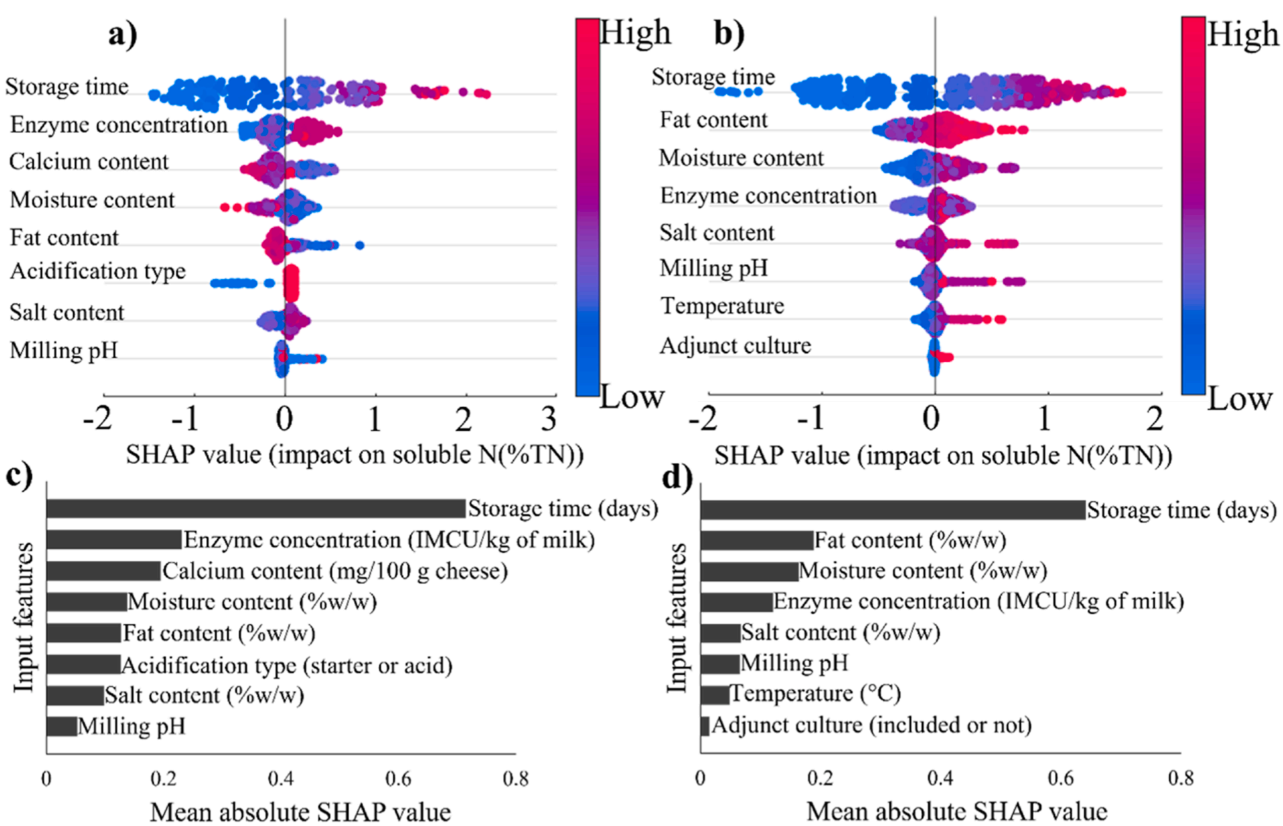

Fig. 5. Assessment of input feature importance on the predictions of soluble nitrogen (%TN) made by the gradient boosting model, as assessed by SHAP analysis. Each of the plots in a and b represent the impact of each datapoint for each input feature on the soluble nitrogen predicted for (a) Mozzarella and (b) Cheddar cheese, where each dot represents an observation in the corresponding training set. Data were all normalized to a similar scale. Positive and negative values correspond to high and low variable values in the plots in a and b, respectively. The mean absolute SHAP values for each feature are also presented in ranked order of importance for (c) Mozzarella and (d) Cheddar. SHAP values were calculated using the gradient boosting model, where the data were split with a ratio of 80:20 between training and test sets.

For the Cheddar model, an increase in fat content impacted positively on the level of soluble nitrogen (Fig. 5b), due to an increase in the level of moisture-in-nonfat substance (Fenelon et al., 2000; Guinee, Auty, Mullins, et al., 2000), a lower degree of casein aggregation or a higher enzyme activity-to-protein ratio (McCarthy et al., 2016). Similarly, an increase in moisture content had a positive impact on the level of soluble nitrogen (Fig. 5b), as previously reported (Phelan et al., 1973), which can arise due to the environment being more favorable for enzymatic and bacterial proteolytic activity, due to the reduced concentration of dissolved salt (i.e., salt-in-moisture concentration) (Upreti et al., 2006).

While intuitive, the local SHAP values should be treated carefully, as they can generate misleading results when features interact (Lundberg et al., 2020) or are correlated (Hiabu et al., 2023), which is true for some features of both Mozzarella and Cheddar datasets examined here. Moisture and fat are highly correlated for both cheeses, while acidification type is also highly correlated with moisture content and milling pH for Mozzarella (Fig. A5 and A6). Several input features also have significant two- or sometimes three-level interaction effects on soluble nitrogen level for each cheese type, e.g., a three-level interaction of storage time, coagulating enzyme concentration and fat content (Dave, McMahon, et al., 2003) or a two-level interaction of storage time and calcium or fat (Henneberry et al., 2015) for Mozzarella and a two-level interaction of storage time and salt-in-moisture (Upreti et al., 2006) or fat content (McCarthy et al., 2016) for Cheddar. Therefore, local SHAP values for less significant features can be distorted by the impact of other feature values (Hiabu et al., 2023), which can explain differences or unexpected patterns observed for the impact of moisture or fat content for Mozzarella (Fig. 5a), as well as enzyme concentration or salt content for Cheddar (Fig. 5b). Hence, confidence is higher in the top three input features identified by SHAP in Fig. 5c and d, as their trends align with previous experimental observations.

3.4. Industrial implications

This study described two useful outputs for industry: 1) a set of machine learning models demonstrating a proof-of-concept approach and 2) an associated SHAP analysis. The machine learning approach demonstrated can be used to assist cheese manufacturers by accurately predicting the level of proteolysis in cheese using readily available input features, potentially aiding process optimization and quality assurance. Insights from SHAP analysis can be used to help optimize formulation, processing techniques and storage conditions by highlighting the relative importance of each of these factors.

The models could be applied in several ways. The time course of proteolysis can be predicted for a new test set of desired process conditions at a given storage timepoint. All the input features considered in this study (explained in Section 2.1) will be required by the proposed models to predict proteolysis with an R^2 of 92% and 97% for Mozzarella (Fig. 3) and Cheddar (Fig. 4), respectively. Alternatively, any of the input features can be adjusted to achieve a desired level of proteolysis. This information could then be used to optimize cheese production and to enhance inventory management planning, ensuring a particular Mozzarella functionality or Cheddar flavor and texture profile.

While the developed models hold significant promise, they may not capture the full range of conditions observed in industrial cheese production and could be extended with industrial data. Predictions made by machine learning models are known to be restricted to the design range of the training dataset and the models developed here rely on data available in the literature, which is mostly sourced from laboratory or pilot scale testing. Data obtained from the industry could also consider the buildup of microflora in commercial cheese plants (Possas et al., 2021), which could significantly affect cheese proteolysis over the course of a manufacturing day. Incorporating additional industry data could enhance model robustness and applicability, although manufacturing data may inherently be less variable, due to the restricted specifications used within manufacturing, so the combination of literature and industry data may be more productive than data replacement.

The current models may be extended to predict the quality and functionality of Mozzarella and Cheddar cheeses, by incorporating additional input features related to sensory attributes, chemical composition and physical properties (i.e., stretchability and meltability). In this way a more comprehensive model could be developed to assess overall cheese desirability, assisting manufacturers to tailor their products for specific markets, consumer preferences and regulatory requirements.

The methods and findings obtained from this study have broader potential for application beyond Mozzarella and Cheddar cheeses. For instance, similar modeling approaches could be employed to predict proteolysis in different types of cheeses or dairy products. The knowledge gained from this study can also serve as a foundation for predicting proteolysis in other protein-based foods (e.g., meat (Lana and Zolla, 2016), fish (Singh and Benjakul, 2018) and fermented beverages (Liu et al., 2022)) or even non-food systems (e.g., protein-based pharmaceuticals (Benchabane et al., 2008; Nejadnik et al., 2018)), where protein degradation plays a significant role. By adapting and extending these methods, researchers may be able to better understand and control proteolysis in a wider range of materials and systems.

4. Conclusion

The developed gradient boosting algorithm can accurately predict the level of soluble nitrogen (as an index for proteolysis) in Mozzarella and Cheddar cheese, outperforming random forest and support vector regression methods and achieving an R^2 of 92% for a Mozzarella dataset and 97% for a Cheddar dataset, with a low MAPE of 13% and 9%, respectively. All three machine learning algorithms developed outperformed multilinear regression due to the presence of linearly correlated features and the existence of a non-linear relationship between most input features and soluble nitrogen. Storage time was found to be the most significant feature impacting the level of proteolysis of both cheese types, followed by other factors, such as the coagulating enzyme concentration and calcium content for Mozzarella and the fat and moisture content for Cheddar. The models developed use readily available input feature data (composition, storage time and manufacturing conditions) and this approach could be used by manufacturers to both predict and optimize production and storage processes, assisting in inventory management, particularly when supplemented with industry data. This could be useful in ensuring consistent quality and functionality in Mozzarella cheese, where proteolysis plays a crucial role in determining functional shelf life and in determining the quality characteristics of Cheddar, such as flavor and texture, which are key factors in determining overall desirability and consumer acceptance.

Declaration of Competing Interest

The authors declare no conflict of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fbp.2024.01.009.

References

Adedeji, A., Ngadi, M., Liu, L., Manzano, S.G., Kumar Basak, J., Gamage, B., Madhavi, K., Paudel, B., Kim, N.E., Kim, H.T., 2022. Prediction of total soluble solids and pH of strawberry fruits using RGB, HSV and HSL colour spaces and machine learning models, 2022 Foods 11 (14), 2086. https://doi.org/10.3390/FOODS11142086.

Afshari, R., Pillidge, C.J., Dias, D.A., Osborn, A.M., Gill, H., 2020. Microbiota and metabolite profiling combined with integrative analysis for differentiating cheeses of varying ripening ages. Front. Microbiol. *11*. https://doi.org/10.3389/FMICB.2020.592060/FULL.

Alizadeh, M., Hamedi, M., Khosroshahi, A., 2006. Modeling of proteolysis and lipolysis in Iranian white brine cheese. Food Chem. 97 (2), 294–301. https://doi.org/10.1016/j.foodchem.2005.05.009.

Awan, F.M., Saleem, Y., Minerva, R., Crespi, N., 2020. A comparative analysis of machine/deep learning models for parking space availability prediction. Sensors 20 (1), 322. https://doi.org/10.3390/S20010322.

Aydin, O., Yassikaya, M.Y., 2022. Validity and reliability analysis of the plotdigitizer software program for data extraction from single-case graphs. Perspect. Behav. Sci. 45 (1), 239–257. https://doi.org/10.1007/s40614-021-00284-0.

Barbano, D.M., Yun, J.J., Kindstedt, P.S., 1994. Mozzarella cheese making by a stirred curd, no brine procedure. J. Dairy Sci. Vol. 77 (Issue 9), 2687–2694. https://doi.org/10.3168/jds.S0022-0302(94)77211-8.

Benchabane, M., Goulet, C., Rivard, D., Faye, L., Gomord, V., Michaud, D., 2008. Preventing unintended proteolysis in plant protein biofactories. Plant Biotechnol. J. Vol. 6 (Issue 7), 633–648. https://doi.org/10.1111/j.1467-7652.2008.00344.x.

Ben-Hur, A., Ong, C.S., Sonnenburg, S., Schölkopf, B., Rätsch, G., 2008. Support vector machines and kernels for computational biology. PLoS Comput. Biol. 4 (10) https://doi.org/10.1371/journal.pcbi.1000173.

Borin, A., Ferrão, M.F., Mello, C., Maretto, D.A., Poppi, R.J., 2006. Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk. Anal. Chim. Acta 579 (1), 25–32. https://doi.org/10.1016/j.aca.2006.07.008.

Chan, J.Y. Le, Leow, S.M.H., Bea, K.T., Cheng, W.K., Phoong, S.W., Hong, Z.W., Chen, Y. L., 2022. Mitigating the multicollinearity problem and its machine learning approach: a review. *2022, Vol. 10, Page 1283* Mathematics 10 (8), 1283. https://doi.org/10.3390/MATH10081283.

Coker, C.J., Crawford, R.A., Johnston, K.A., Singh, H., Creamer, L.K., 2005. Towards the classification of cheese variety and maturity on the basis of statistical analysis of proteolysis data - a review. Int. Dairy J. 15 (6–9), 631–643. https://doi.org/10.1016/j.idairyj.2004.10.011.

Cortes, C., Vapnik, V., 1995. Support-vector networks editor. Mach. Learning 20, 273–297.

Costabel, L., Pauletti, M.S., Hynes, E., 2007. Proteolysis in Mozzarella cheeses manufactured by different industrial processes. J. Dairy Sci. 90 (5), 2103–2112. https://doi.org/10.3168/jds.2006-795.

Dave, R.I., McMahon, D.J., Oberg, C.J., Broadbent, J.R., 2003. Influence of coagulant level on proteolysis and functionality of mozzarella cheeses made using direct acidification. J. Dairy Sci. 86 (1), 114–126. https://doi.org/10.3168/jds.S0022-0302(03)73590-5.

Dave, R.I., Sharma, P., Muthukumarappan, K., 2003. Effects of starter culture and coagulating enzymes on viscoelastic behavior and melt of Mozzarella cheese. J. Food Sci. 68 (4), 1404–1410. https://doi.org/10.1111/j.1365-2621.2003.tb09657.x.

Elavarasan, D., Vincent, D.R., 2020. Reinforced XGBoost machine learning model for sustainable intelligent agrarian applications. J. Intell. Fuzzy Syst. 39 (5), 7605–7620. https://doi.org/10.3233/JIFS-200862.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., Tabona, O., 2021. A survey on missing data in machine learning. J. Big Data 8 (1), 1–37. https://doi.org/10.1186/S40537-021-00516-9.

Farah, J.S., Cavalcanti, R.N., Guimarães, J.T., Balthazar, C.F., Coimbra, P.T., Pimentel, T. C., Esmerino, E.A., Duarte, M.C.K.H., Freitas, M.Q., Granato, D., Neto, R.P.C., Tavares, M.I.B., Calado, V., Silva, M.C., Cruz, A.G., 2021. Differential scanning calorimetry coupled with machine learning technique: An effective approach to determine the milk authenticity. Food Control 121, 107585. https://doi.org/10.1016/J.FOODCONT.2020.107585.

Feeney, E.P., Fox, P.F., Guinee, T.P., 2001. Effect of ripening temperature on the quality of low moisture Mozzarella cheese: 1. Composition and proteolysis. Lait 81 (4), 463–474. https://doi.org/10.1051/lait:2001145.

Feeney, E.P., Guinee, T.P., Fox, P.F., 2002. Effect of pH and calcium concentration on proteolysis in Mozzarella cheese. J. Dairy Sci. 85 (7), 1646–1654. https://doi.org/10.3168/jds.S0022-0302(02)74237-9.

Fenelon, M.A., Guinee, T.P., 2000. Primary proteolysis and textural changes during ripening in Cheddar cheeses manufactured to different fat contents. Int. Dairy J. 10 (3), 151–158. https://doi.org/10.1016/S0958-6946(00)00040-6.

Fenelon, M.A., Ryan, M.P., Rea, M.C., Guinee, T.P., Ross, R.P., Hill, C., Harrington, D., 1999. Elevated temperature ripening of reduced fat Cheddar made with or without Lacticin 3147-producing starter culture. J. Dairy Sci. 82 (1), 10–22. https://doi.org/10.3168/jds.S0022-0302(99)75203-3.

Fenelon, M.A., O'Connor, P., Guinee, T.P., 2000. The effect of fat content on the microbiology and proteolysis in Cheddar cheese during ripening. J. Dairy Sci. 83 (10), 2173–2183. https://doi.org/10.3168/JDS.S0022-0302(00)75100-9.

Fife, R.L., McMahon, D.J., Oberg, C.J., 1996. Functionality of low fat Mozzarella cheese. J. Dairy Sci. 79 (11), 1903–1910. https://doi.org/10.3168/jds.S0022-0302(96)76559-1.

Fox, P.F., 1989. Proteolysis during cheese manufacture and ripening. J. Dairy Sci. 72 (6), 1379–1400. https://doi.org/10.3168/jds.S0022-0302(89)79246-8.

Fox, P.F., McSweeney, P.L.H., 1996. Proteolysis in cheese during ripening. Food Rev. Int. 12 (4), 457–509. https://doi.org/10.1080/87559129609541091.

Fox, P.F., Wallace, J.M., 1997. Formation of flavor compounds in cheese. In: Neidleman, S.L., Laskin, A.I. (Eds.), Adv. Appl. Microbiol. Vol. 45, 17–85. https://doi.org/10.1016/s0065-2164(08)70261-2.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232. https://doi.org/10.1214/aos/1013203451.

Giola, C., Danti, P., Magnani, S., Rojas, F., Herrera, L.J., Pomare, H., 2021. Learning curves: a novel approach for robustness improvement of load forecasting. *2021, Vol. 5, Page 38* Eng. Proc. 5 (1), 38. https://doi.org/10.3390/ENGPROC2021005038.

Golzarijalal, M., Ong, L., Harvie, D.J.E., Gras, S.L., 2024. Experimental investigation, numerical simulation and RSM modelling of the freezing and thawing of Mozzarella cheese. Food Bioprod. Process. 143, 143–157. https://doi.org/10.1016/j.fbp.2023.11.006.

Gomes, A.M.P., Silva, M.L.P.C., Malcata, F.X., Gomes, A.M.P., Luz, M., Silva, P.C., Malcata, F.X., 1998. Caprine cheese with probiotic strains: the effects of ripening temperature and relative humidity on proteolysis and lipolysis. In: *I*n Z Lebensm Unters Forsch A, Vol. 207. Springer-Verlag.

Gorzelany, J., Belcar, J., Kuźniar, P., Niedbała, G., Pentoś, K., 2022. Modelling of mechanical properties of fresh and stored fruit of large cranberry using multiple linear regression and machine learning. *2022, Vol. 12, Page 200* Agriculture 12 (2), 200. https://doi.org/10.3390/AGRICULTURE12020200.

Granitto, P.M., Gasperi, F., Biasioli, F., Trainotti, E., Furlanello, C., 2007. Modern data mining tools in descriptive sensory analysis: a case study with a Random forest approach. Food Qual. Prefer. 18 (4), 681–689. https://doi.org/10.1016/J.FOODQUAL.2006.11.001.

Guan, Y., Fu, G.H., 2022. A double-penalized estimator to combat separation and multicollinearity in logistic regression. Mathematics 10 (20). https://doi.org/10.3390/math10203824.

Guinee, Auty, M.A.E., Fenelon, M.A., 2000. The effect of fat content on the rheology, microstructure and heat-induced functional characteristics of Cheddar cheese. Int. Dairy J. 10 (4), 277–288. https://doi.org/10.1016/S0958-6946(00)00048-0.

Guinee, T.P., Wilkinson, M.G., Mulholland, E.O., Fox, P.F., 1991. Influence of ripening temperature, added commercial enzyme preparations and attenuated Mutant (Lac^-) lactococcus lactis starter on the proteolysis and maturation of Cheddar cheese. Ir. J. Food Sci. Technol. 15 (1/2), 27–52. ⟨http://www.jstor.org/stable/25558218⟩.

Guinee, T.P., Auty, M.A.E., Mullins, C., Corcoran, M.O., Mulholland, E.O., 2000. Preliminary observations on effects of fat content and degree of fat emulsification on the structure-functional relationship of Cheddar-type cheese. J. Texture Stud. 31 (6), 645–663. https://doi.org/10.1111/j.1745-4603.2000.tb01026.x.

Guinee, T.P., Harrington, D., Corcoran, M.O., Mulholland, E.O., Mullins, C., 2000. The compositional and functional properties of commercial mozzarella, Cheddar and analogue pizza cheeses. Int. J. Dairy Technol. 53 (2), 51–56. https://doi.org/10.1111/j.1471-0307.2000.tb02658.x.

Guinee, T.P., Feeney, E.P., Fox, P.F., 2001. Effect of ripening temperature on low moisture Mozzarella cheese: 2. Texture and functionality. Lait 81 (4), 475–485. https://doi.org/10.1051/lait:2001146.

Guinee, T.P., Feeney, E.P., Auty, M.A.E., Fox, P.F., 2002. Effect of pH and calcium concentration on some textural and functional properties of Mozzarella cheese. J. Dairy Sci. 85 (7), 1655–1669. https://doi.org/10.3168/jds.S0022-0302(02)74238-0.

Hannon, J.A., Wilkinson, M.G., Delahunty, C.M., Wallace, J.M., Morrissey, P.A., Beresford, T.P., 2003a. Use of autolytic starter systems to accelerate the ripening of Cheddar cheese. Int. Dairy J. 13 (4), 313–323. https://doi.org/10.1016/S0958-6946(02)00178-4.

Hannon, J.A., Wilkinson, M.G., Delahunty, C.M., Wallace, J.M., Morrissey, P.A., Beresford, T.P., 2003b. Use of autolytic starter systems to accelerate the ripening of Cheddar cheese. Int. Dairy J. 13 (4), 313–323. https://doi.org/10.1016/S0958-6946(02)00178-4.

Hasan, M.K., Alam, M.A., Roy, S., Dutta, A., Jawad, M.T., Das, S., 2021. Missing value imputation affects the performance of machine learning: a review and analysis of the literature (2010–2021). Inform. Med. Unlocked 27. https://doi.org/10.1016/J.IMU.2021.100799.

Henneberry, S., Wilkinson, M.G., Kilcawley, K.N., Kelly, P.M., Guinee, T.P., 2015. Interactive effects of salt and fat reduction on composition, rheology and functional properties of mozzarella-style cheese. Dairy Sci. Technol. 95 (5), 613–638. https://doi.org/10.1007/s13594-015-0231-8.

Hiabu, M., Meyer, J.T., Wright, M.N. , 2023. Unifying local and global model explanations by functional decomposition of low dimensional structures. Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS), 206. https://arxiv.org/abs/2208.06151v2.

Hosseinzadeh, A., Zhou, J.L., Altaee, A., Li, D., 2022. Machine learning modeling and analysis of biohydrogen production from wastewater by dark fermentation process. Bioresour. Technol. 343 (October 2021), 126111 https://doi.org/10.1016/j.biortech.2021.126111.

Hosseinzadeh, A., Zhou, J.L., Zyaie, J., AlZainati, N., Ibrar, I., Altaee, A., 2022. Machine learning-based modeling and analysis of PFOS removal from contaminated water by nanofiltration process. Sep. Purif. Technol. 289 (February), 120775 https://doi.org/10.1016/j.seppur.2022.120775.

Hou, J., Hannon, J.A., McSweeney, P.L.H., Beresford, T.P., Guinee, T.P., 2014. Effect of curd washing on cheese proteolysis, texture, volatile compounds, and sensory grading in full fat Cheddar cheese. Int. Dairy J. 34 (2), 190–198. https://doi.org/10.1016/j.idairyj.2013.08.008.

IDF 157, 2007. Determination of total milk-clotting activity of bovine rennets. International IDF Standard 157 (ISO 11815),, *Brussels, Belgium*.

İskenderoğlu, F.C., Baltacioğlu, M.K., Demir, M.H., Baldinelli, A., Barelli, L., Bidini, G., 2020. Comparison of support vector regression and random forest algorithms for estimating the SOFC output voltage by considering hydrogen flow rates. Int. J. Hydrog. Energy 45 (60), 35023–35038. https://doi.org/10.1016/j.ijhydene.2020.07.265.

Jelicic Kadic, A., Vucic, K., Dosenovic, S., Sapunar, D., Puljak, L., 2016. Extracting data from figures with software was faster, with higher interrater reliability than manual extraction. J. Clin. Epidemiol. 74, 119–123. https://doi.org/10.1016/j.jclinepi.2016.01.002.

Jeong, N., Chung, T.H., Tong, T., 2021. Predicting micropollutant removal by reverse osmosis and nanofiltration membranes: is machine learning viable? Environ. Sci. Technol. 55 (16), 11348–11359. https://doi.org/10.1021/acs.est.1c04041.

Kapoor, S., & Narayanan, A., 2022. Leakage and the Reproducibility Crisis in ML-based Science. Ml. ⟨http://arxiv.org/abs/2207.07048⟩.

Karoui, R., Mouazen, A.M., Dufour, É., Pillonel, L., Schaller, E., De Baerdemaeker, J., Bosset, J.O., 2006. Chemical characterisation of European Emmental cheeses by near infrared spectroscopy using chemometric tools. Int. Dairy J. 16 (10), 1211–1217. https://doi.org/10.1016/j.idairyj.2005.10.002.

Kruse, J., Schafer, B., & Witthaut, D. (2021). Exploring deterministic frequency deviations with explainable AI. 2021 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids, SmartGridComm 2021, 133–139. https://doi.org/10.1109/SMARTGRIDCOMM51999.2021.9632335.

Kuchroo, C.N., Fox, P.F., 1982. Soluble nitrogen in Cheddar cheese: comparison of extraction procedures. Milchwiss. Milk. Sci. Int. 37 (6), 331–335. ⟨https://agris.fao.org/agris-search/search.do?recordID=US201302191633⟩.

Lana, A., Zolla, L., 2016. Proteolysis in meat tenderization from the point of view of each single protein: a proteomic perspective. In: Journal of Proteomics, Vol. 147. Elsevier B.V.,, pp. 85–97. https://doi.org/10.1016/j.jprot.2016.02.011.

Lane, C.N., Fox, P.F., 1996. Contribution of starter and adjunct lactobacilli to proteolysis in Cheddar cheese during ripening. Int. Dairy J. 6 (7), 715–728. https://doi.org/10.1016/0958-6946(95)00067-4.

Lefevere, I., Dewettinck, K., Huyghebaert, A., 2000. Cheese fat as driving force in cheese flow upon melting. Milchwissenschaft 55 (10), 563–566.

Li, J., Pan, L., Suvarna, M., Tong, Y.W., Wang, X., 2020. Fuel properties of hydrochar and pyrochar: prediction and exploration with machine learning. Appl. Energy 269 (May), 115166. https://doi.org/10.1016/j.apenergy.2020.115166.

Li, M., Zhao, J., Yang, X., 2021. Building a new machine learning-based model to estimate county-level climatic yield variation for maize in Northeast China. Comput. Electron. Agric. 191, 106557 https://doi.org/10.1016/J.COMPAG.2021.106557.

Lin, J.C.C., Jeon, I.J., Roberts, H.A., Milliken, G.A., 1987. Effects of commercial food grade enzymes on proteolysis and textural changes in granular Cheddar cheese. J. Food Sci. 52 (3), 620–625. https://doi.org/10.1111/j.1365-2621.1987.tb06688.x.

Lin, Y., Ma, J., Wang, Q., Sun, D.W., 2022. Applications of machine learning techniques for enhancing nondestructive food quality and safety detection. Crit. Rev. Food Sci. Nutr. https://doi.org/10.1080/10408398.2022.2131725.

Liu, S., Kerr, E.D., Pegg, C.L., Schulz, B.L., 2022. Proteomics and glycoproteomics of beer and wine. In: Proteomics, Vol. 22. John Wiley and Sons Inc. https://doi.org/10.1002/pmic.202100329

de Llano González, D., Ramos, M., Martin-Alvarez, P.J., 1993. Comparative study of different methods for evaluating proteolysis in blue cheese. Z. Für Lebensm. -Unters. Und -Forsch. 196 (1), 29–31. https://doi.org/10.1007/BF01192980.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 4765–4774. ⟨https://github.com/slundberg/shap⟩.

Lundberg, S.M., Erion, G.G., Lee, S.- I , 2018. Consistent Individualized Feature Attribution for Tree Ensembles. ⟨https://arxiv.org/abs/1802.03888⟩.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. *2020 2:1* Nat. Mach. Intell. 2 (1), 56–67. https://doi.org/10.1038/s42256-019-0138-9.

Mastromatteo, M., Conte, A., Faccia, M., Del Nobile, M.A., Zambrini, A.V., 2014. Combined effect of active coating and modified atmosphere packaging on prolonging the shelf life of low-moisture Mozzarella cheese. J. Dairy Sci. 97 (1), 36–45. https://doi.org/10.3168/jds.2013-6999.

McCarthy, C.M., Wilkinson, M.G., Kelly, P.M., Guinee, T.P., 2016. Effect of salt and fat reduction on proteolysis, rheology and cooking properties of Cheddar cheese. Int. Dairy J. 56, 74–86. https://doi.org/10.1016/j.idairyj.2016.01.001.

McCarthy, C.M., Wilkinson, M.G., Guinee, T.P., 2017a. Effect of calcium reduction on the properties of half-fat Cheddar-style cheeses with full-salt or half-salt. Int. Dairy J. 73, 38–49. https://doi.org/10.1016/j.idairyj.2017.04.006.

McCarthy, C.M., Wilkinson, M.G., Guinee, T.P., 2017b. Effect of coagulant type and level on the properties of half-salt, half-fat Cheddar cheese made with or without adjunct starter: improving texture and functionality. Int. Dairy J. 75, 30–40. https://doi.org/10.1016/j.idairyj.2017.07.006.

McSweeney, P.L.H., Fox, P.F., 1997. Chemical methods for the characterization of proteolysis in cheese during ripening. Lait 77 (1), 41–76. https://doi.org/10.1051/lait:199713.

Muir, D.D., 1996. The shelf-life of dairy products: 4. Intermediate and long life dairy products. Int. J. Dairy Technol. 49 (4), 119–124. https://doi.org/10.1111/j.1471-0307.1996.tb02503.x.

Nawar, S., Mouazen, A.M., 2017. Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon. Sens. (Switz.) 17 (10). https://doi.org/10.3390/s17102428.

Nejadnik, M.R., Randolph, T.W., Volkin, D.B., Schöneich, C., Carpenter, J.F., Crommelin, D.J.A., Jiskoot, W., 2018. Postproduction handling and administration of protein pharmaceuticals and potential instability issues. J. Pharm. Sci. 107 (8), 2013–2019. https://doi.org/10.1016/j.xphs.2018.04.005.

Nguyen, H., Vu, T., Vo, T.P., Thai, H.T., 2021. Efficient machine learning models for prediction of concrete strengths. Constr. Build. Mater. 266 https://doi.org/10.1016/j.conbuildmat.2020.120950.

O'Keeffe, R.B., Fox, P.F., Daly, C., 1975. Proteolysis in Cheddar cheese: Influence of the rate of acid production during manufacture. J. Dairy Res. 42 (1), 111–122. https://doi.org/10.1017/S0022029900015144.

Ogutu, J.O., Piepho, H.P., Schulz-Streeck, T., 2011. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc. 5 (SUPPL. 3), 1–5. https://doi.org/10.1186/1753-6561-5-S3-S11/FIGURES/2.

Ong, L., Shah, N.P., 2009. Probiotic cheddar cheese: Influence of ripening temperatures on proteolysis and sensory characteristics of cheddar cheeses. J. Food Sci. 74 (5) https://doi.org/10.1111/j.1750-3841.2009.01164.x.

Patange, A.D., Pardeshi, S.S., Jegadeeshwaran, R., Zarkar, A., Verma, K., 2022. Augmentation of decision tree model through hyper-parameters tuning for monitoring of cutting tool faults based on vibration signatures. J. Vib. Eng. Technol. https://doi.org/10.1007/s42417-022-00781-9.

Phelan, J.A., Guiney, J., Fox, P.F., 1973. Proteolysis of β-casein in Cheddar cheese. J. Dairy Res. 40 (1), 105–112. https://doi.org/10.1017/S0022029900014400.

Possas, A., Bonilla-Luque, O.M., Valero, A., 2021. From cheese-making to consumption: exploring the microbial safety of cheeses through predictive microbiology models. Foods 10 (2), 355. https://doi.org/10.3390/foods10020355.

Povak, N.A., Hessburg, P.F., McDonnell, T.C., Reynolds, K.M., Sullivan, T.J., Salter, R.B., Cosby, B.J., 2014. Machine learning and linear regression models to predict catchment-level base cation weathering rates across the southern Appalachian Mountain region, USA. Water Resour. Res. 50 (4), 2798–2814. https://doi.org/10.1002/2013WR014203.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9 (2), 181–199. https://doi.org/10.1007/s10021-005-0054-1.

Rank, T.C., Grappin, R., Olson, N.F., 1985. Secondary proteolysis of cheese during ripening: a review. J. Dairy Sci. Vol. 68 (Issue 4), 801–805. https://doi.org/10.3168/jds.S0022-0302(85)80895-X.

Rocha, R.S., Calvalcanti, R.N., Silva, R., Guimarães, J.T., Balthazar, C.F., Pimentel, T.C., Esmerino, E.A., Freitas, M.Q., Granato, D., Costa, R.G.B., Silva, M.C., Cruz, A.G., 2020. Consumer acceptance and sensory drivers of liking of Minas Frescal Minas cheese manufactured using milk subjected to ohmic heating: performance of machine learning methods. LWT 126, 109342. https://doi.org/10.1016/J.LWT.2020.109342.

Rodgers, J.L., & Nicewander, W.A., 1988. Thirteen Ways to Look at the Correlation Coefficient (Vol. 42, Issue 1).

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71, 804–818. https://doi.org/10.1016/J.OREGEOREV.2015.01.001.

Rudan, M.A., Barbano, D.M., Kindstedt, P.S., 1998. Effect of fat replacer (Salatrim®) on chemical composition, proteolysis, functionality, appearance, and yield of reduced fat mozzarella cheese. J. Dairy Sci. Vol. 81 (Issue 8), 2077–2088. https://doi.org/10.3168/jds.S0022-0302(98)75782-0.

Rudan, M.A., Barbano, D.M., Yun, J.J., Kindstedt, P.S., 1999. Effect of fat reduction on chemical composition, proteolysis, functionality, and yield of Mozzarella cheese. J. Dairy Sci. Vol. 82 (Issue 4), 661–672. https://doi.org/10.3168/jds.S0022-0302(99)75282-3.

Sergent, M., Mathieu, D., Phan-Tan-Luu, R., Drava, G., 1995. Correct and incorrect use of multilinear regression. Chemom. Intell. Lab. Syst. Vol. 27.

Shapley, L., 1953. A value for n-person games. In: Kuhn, H.T.A. (Ed.), Contributions to the Theory of Games II. Princeton University Press, pp. 307–317. https://doi.org/10.1515/9781400881970-018.

Sheehan, J.J., Guinee, T.P., 2004. Effect of pH and calcium level on the biochemical, textural and functional properties of reduced-fat Mozzarella cheese. Int. Dairy J. Vol. 14 (Issue 2), 161–172. https://doi.org/10.1016/S0958-6946(03)00167-5.

Sheng, T., Shi, S., Zhu, Y., Chen, D., Liu, S., 2022. Analysis of protein and fat in milk using multiwavelength gradient-boosted regression tree. IEEE Trans. Instrum. Meas. 71 https://doi.org/10.1109/TIM.2022.3165298.

Singh, A., Benjakul, S., 2018. Proteolysis and its control using protease inhibitors in fish and fish products: a review. Compr. Rev. Food Sci. Food Saf. 17 (2), 496–509. https://doi.org/10.1111/1541-4337.12337.

Slob, N., Catal, C., Kassahun, A., 2021. Application of machine learning to improve dairy farm management: a systematic literature review. In: Preventive Veterinary Medicine, Vol. 187. Elsevier B.V. https://doi.org/10.1016/j.prevetmed.2020.105237

Smola, A.J., Scholkopf, B., Scholkopf, S., 2004. A tutorial on support vector regression. Stat. Comput. 14, 199–222.

Stangierski, J., Weiss, D., Kaczmarek, A., 2019. Multiple regression models and Artificial Neural Network (ANN) as prediction tools of changes in overall quality during the storage of spreadable processed Gouda cheese. Eur. Food Res. Technol. 245 (11), 2539–2547. https://doi.org/10.1007/s00217-019-03369-y.

Stine, R.A., 1995. Graphical interpretation of variance inflation factors. Am. Stat. 49 (1), 53–56. https://doi.org/10.1080/00031305.1995.10476113.

Sun, Y., Zhang, Y., Lu, L., Wu, Y., Zhang, Y., Kamran, M.A., Chen, B., 2022. The application of machine learning methods for prediction of metal immobilization remediation by biochar amendment in soil. Sci. Total Environ. 829 https://doi.org/10.1016/j.scitotenv.2022.154668.

Tabayehnejad, N., Castillo, M., Payne, F.A., 2012. Comparison of total milk-clotting activity measurement precision using the Berridge clotting time method and a proposed optical method. J. Food Eng. 108 (4), 549–556. https://doi.org/10.1016/j.jfoodeng.2011.09.009.

Taylor, R., 1990. Interpretation of the Correlation Coefficient: A Basic Review. ⟨Http://Dx.Doi.Org/10.1177/875647939000600106⟩⟨,⟩*6*(1), 35–39. https://doi.org/10.1177/875647939000600106.

To, C.M., Kerkaert, B., Bossier, S., Van Gaver, D., Van der Meeren, P., Guinee, T.P., 2022. Variations in the biochemical and functional properties of commercial low-moisture part-skim mozzarella during 3 months of storage at 4 °C. Int. Dairy J. Vol. 128 https://doi.org/10.1016/j.idairyj.2022.105320.

Upreti, P., Metzger, L.E., Hayes, K.D., 2006. Influence of calcium and phosphorus, lactose, and salt-to-moisture ratio on Cheddar cheese quality: proteolysis during ripening. J. Dairy Sci. 89 (2), 444–453. https://doi.org/10.3168/jds.S0022-0302(06)72108-7.

Uyanık, G.K., Güler, N., 2013. A study on multiple linear regression analysis. Procedia - Soc. Behav. Sci. 106, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027.

Viering, T., Loog, M., 2021. The shape of learning curves: a review. IEEE Trans. Pattern Anal. Mach. Intell. 45, 7799–7819. https://doi.org/10.1109/TPAMI.2022.3220744.

Walsh, E.A., Diako, C., Smith, D.M., Ross, C.F., 2020. Influence of storage time and elevated ripening temperature on the chemical and sensory properties of white Cheddar cheese. J. Food Sci. 85 (2), 268–278. https://doi.org/10.1111/1750-3841.14998.

Wang, L., Moraru, C.I., 2021. Structure and shelf stability of milk protein gels created by pressure-assisted enzymatic gelation. J. Dairy Sci. 104 (4), 3970–3979. https://doi.org/10.3168/jds.2020-19484.

Wilkinson, M.G., Guinee, T.P., O'Callaghan, D.M., Fox, P.F., 1992. Effects of commercial enzymes on proteolysis and ripening in Cheddar cheese. Le. Lait. 72 (5), 449–459. https://doi.org/10.1051/LAIT:1992533.

You, J., Tulpan, D., Malpass, M.C., Ellis, J.L., 2022. Using machine learning regression models to predict the pellet quality of pelleted feeds. Anim. Feed Sci. Technol. 293, 115443 https://doi.org/10.1016/J.ANIFEEDSCI.2022.115443.

Zhou, L., Fujita, H., Ding, H., Ma, R., 2021. Credit risk modeling on data with two timestamps in peer-to-peer lending by gradient boosting. Appl. Soft Comput. 110, 107672 https://doi.org/10.1016/j.asoc.2021.107672.

Zhu, Z., Zhu, X., Guo, W., 2022. Quantitatively determining the somatic cell count of raw milk using dielectric spectra and support vector regression. J. Dairy Sci. 105 (1), 772–781. https://doi.org/10.3168/jds.2021-20828.